PolUQBench: A Benchmark Study on Uncertainty Quantification of Polymer Property Prediction

Alif Bin Abdul Qayyum[1] **Byung-Jun Yoon**[1,2]
[1]Department of Electrical and Computer Engineering, Texas A&M University
[2]Computing and Data Sciences, Brookhaven National Laboratory

Abstract

Large Language Model(LLM)s have demonstrated remarkable capabilities to tackle multidomain challenges, a capability often lacking in conventional machine learning methods. This makes them particularly promising for understanding the complex relationship between a material's composition and its properties, which can significantly accelerate materials design, especially for polymers. Leveraging the hidden states of domain-specific pretrained LLMs for downstream tasks like property prediction has gained significant traction. This approach is now widely used for small molecules and proteins, along with recent efforts also extending to polymers. In addition to achieving superior predictive performance, Uncertainty Quantification (UQ) is another crucial aspect for enhancing the reliability of machine learning models used as property predictors. This is particularly important for high-stakes applications like the discovery of new functional polymers. We introduce **Pol**ymer Property Predictor **U**ncertainty **Q**uantification **Bench**mark, a pioneering study that evaluates the effectiveness of embeddings extracted from a Polymer Language Model for representing polymer data and assesses the performance of several different UQ methods for reliable polymer property prediction.

1 Introduction

Language model embeddings are transforming how complex data is represented and used in science and engineering. These embedding techniques, which originated in Natural Language Processing to model linguistic relationships in a vector space, have demonstrated surprising utility across a wide range of disciplines. In bioinformatics, protein and genomic language models use sequence data to create embeddings that encode biological properties, enabling accurate predictions of protein function and structure [26, 37, 22]. Models like the ESM series [38, 23, 11, 6], have been shown to capture evolutionary and structural relationships directly from protein sequences, providing a powerful alternative to traditional feature engineering.

In materials science and chemistry, language models trained on chemical data (like SMILES) or text can produce embeddings that predict material properties, guide drug discovery, and accelerate the design of new compounds [44, 13]. These embeddings can encapsulate complex chemical knowledge, such as periodic table trends or structure-property relationships, in a vectorized format [44]. As a result, language model embeddings have unlocked new avenues for discovery and innovation by providing a unified and powerful representation for sequential and compositional data across a variety of domains.

The complex and diverse structures of polymers create unique challenges, despite their critical role in medical devices, energy storage, and aerospace engineering [9, 45, 25]. Accurate prediction of diverse polymer properties is essential for tailoring polymers to meet specific functional requirements [18]. In contrast to time and resource-intensive traditional approaches for polymer property evaluation,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat: AI for Accelerated Materials Design.

machine learning enables rapid, scalable, and cost-effective prediction of material properties, accelerating the development of advanced polymers for diverse applications [15, 52, 34, 30, 31, 49, 29]. By integrating data-driven insights with predictive accuracy, machine learning can streamline polymer discovery, optimize design, and transform polymer informatics [43].

Even after the tremendous success of ML polymer property predictors, relying solely on a single, deterministic prediction, also known as point estimate, is insufficient and potentially dangerous for high-stakes applications like polymer material design [40]. This critical limitation underscores the necessity for Uncertainty Quantification (UQ), a discipline dedicated to characterizing and quantifying the sources of uncertainty in predictive models [7]. Uncertainty Quantification is a critical enabler for trustworthy AI and scientific computing, as it provides the necessary confidence intervals that allow agents, human or automated, to make informed decisions, evaluate potential risks, and determine the credibility of a given prediction [53]. The two primary types of uncertainty that UQ methods aim to capture are *aleatoric* and *epistemic* uncertainty [42]. Aleatoric uncertainty is an irreducible property of the data generation process, arising from its inherent stochasticity. Conversely, epistemic uncertainty results from insufficient knowledge, often attributable to a model's limited exposure to certain areas of the input space during training. This form of uncertainty is reducible in principle through the acquisition of more comprehensive training data [14].

Although machine learning has greatly advanced polymer property prediction, with a few works on benchmarking uncertainty quantification [41, 40], pioneering UQ benchmarking in the field still relies on traditional representations like Morgan fingerprints [32]. Consequently, there remains a critical gap in the literature regarding a systematic evaluation of modern polymer language model (PLM) embeddings for uncertainty quantification.

In this work, we introduce **Pol**ymer property predictor **U**ncertainty **Q**uantification **Bench**mark (**PolUQBench**), an extensive benchmark for uncertainty quantification of ML polymer property predictors with Polymer Language Model embeddings as polymer representations. We utilize Polymer Prediction 2025 Dataset [24]. To the best of our knowledge, this is the most recent and extensive publicly available dataset for labeled polymer properties.

2 Problem Statement

The central problem addressed in this work is quantification of predictive uncertainty for machine learning polymer property predictors. Consider s_i be the SMILES representation of a polymer. Given a pretrained Polymer Language Model p_ϕ, the hidden state embedding representation from the Polymer Language model is $x_i = p_\phi(s_i)$. We define a property predictor as a function $f_\theta(x_i)$ that maps a polymer representation x_i to a predicted property $\hat{y}_i$. The goal of Uncertainty Quantification (UQ) is to complement this point prediction with an estimate of the probability distribution $p(y_i|x_i, \mathcal{D})$ over the possible true property values y_i, given the training data $\mathcal{D}$. This distribution can be summarized by its moments, most importantly the *Predictive Mean*, $\mu(x_i)$ and *Predictive Variance*, $\sigma^2(x_i)$:

$$\mu(x_i) = \mathbb{E}_{\theta\sim p(\theta|\mathcal{D})} f_\theta(x_i) \approx \hat{y}(x_i) \tag{1}$$

$$\sigma^2(x_i) = Var_{\theta\sim p(\theta|\mathcal{D})} f_\theta(x_i) \tag{2}$$

The variance $\sigma^2(x_i)$ captures the total predictive uncertainty.

3 PolUQBench

In this work, we focus on investigating UQ approaches in polymer property prediction tasks. This task can be cast as a regression problem, where the model is trained to predict some physical properties related to polymer. We first briefly describe the dataset we utilized in Section 3.1, the hidden state embedding extraction from Polymer Language models in Section 3.2, experimental setup in Section 3.3 and experimental results in Section 3.4.

3.1 Dataset

We utilize Polymer Prediction 2025 Dataset [24], a large-scale open-source dataset. This dataset provides a polymer's structure as a simple text string (SMILES) [46, 48, 47]. The set of labeled properties include: (i) *Density*, (ii) *Heat Thermal Conductivity* (Tc), (iii) *Glass Transition Temperature* (Tg), (iv) fundamental molecular size and packing efficiency *Radius of Gyration* (Rg) and (v) *Fractional Free Volume* (FFV). The ground truth for each property is averaged from multiple runs of molecular dynamics simulation.

3.2 Data Processing

We used two different Polymer Language Models (PLM) to extract embeddings that we utilize as input data for our property predictors: (i) *PolyNC* [35] and *PolyBERT* [19]. The PLM tokenizes the non-Hydrogen atoms and pass the tokens into the model. Considering length of the SMILES representation of the polymer is N, the size of the hidden state embedding of the PLM is $N \times D$ where D is the hidden state dimension. We averaged over the length of the hidden state dimension, resulting the polymer embedding dimension to be D. The value of D are 768 and 600 for *PolyNC* and *PolyBERT* accordingly.

3.3 Experimental Setup

Architectural Setup: We trained different models for each polymer property. For each property we conducted five-fold cross-validation experiments and report the evaluations on those five validation sets. We utilized the *LightningUQBox* [21] package to design and train all the different probabilistic models. For the backbone model, we employed a Fully Connected ResNet which is available in the *LightningUQBox* [21] package, with a hidden dimension of 256 and a depth of 8. All models were trained a total of 500 epochs, with a learning rate of 0.0001.

Model: Apart from deterministic model, we also evaluated the following set of probabilistic models in this benchmark: (i) MVE (Mean Variance Estimation) [33, 39], (ii) DEL (Deep Evidential Learning) [1, 28], (iii) QREG (Quantile Regression) [17], (iv) LA (Laplace Approximation) [36, 2], (v) MCD (Monte-Carlo Dropout) [8], (vi) MFVI (Mean Field Variational Inference) [16], (vii) BNNVI (Bayesian Neural Network with Variational Inference) [3], (viii) SWAG (Stochastic Weight Averaging Gaussian) [27], (ix) DKL (Deep Kernel Learning) [51], (x) VBLL (Variational Bayesian Last Layer) [10], (xi) VBLL SNGP (Variational Bayesian Last Layer (VBLL) with SNGP Regression), (xii) Deep Ensemble [20, 50], (xiii) Mask Ensemble [4], and (xiv) Zigzag [5].

Evaluation Metrics: To evaluate the models, we selected 4 metrics: (i) MAE (Mean Absolute Error), (ii) PCC (Pearson Correlation Coefficient), (iii) CRPS (Continuous Ranked Probability Score) [12] and (iv) CA (Calibration Area) [40]. The first 2 metrics evaluate the predictive performance of the model, and CRPS and CA are the UQ metrics.

3.4 Results

We report the evaluation metric values averaged over the five cross-validation splits for all polymer properties, and for both Polymer Language Models.

Observations: Table 1 shows the metric values for different models. For all properties, the probabilistic models obtained lower MAE than deterministic models, except for Tg with PolyNC embeddings. Unlike the case of MAE, the deterministic model obtained higher PCC values than probabilistic models for quite a lot of polymer properties. For most properties, the Mask Ensemble model obtained lower CRPS than other models. The VBLL SNGP model also obtained lowest CRPS value in some of the evaluation scenarios. For most properties, the MVE model obtained lowest CA than other models. Almost at all scenarios and for every metrics, models trained with PolyNC hidden state embeddings obtained better performances compared to models trained with PolyBERT hidden state embeddings.

Discussion: Experimental results indicate that a clear winner among the probabilistic models could not be identified, as no single model consistently outperformed the others across every metric. The probabilistic models consistently achieved a lower MAE than the deterministic baseline, signifying a general improvement in predictive accuracy. Additionally, a distinct performance advantage was observed for models trained with PolyNC hidden state embeddings over those trained with PolyBERT hidden state embeddings.

4 Conclusion

This work presented a pioneering benchmark study evaluating the efficacy of Polymer Language Model (PLM) embeddings for uncertainty quantification (UQ) in polymer property prediction. Diverging from prior UQ investigations that largely relied on conventional molecular fingerprints, such as Morgan fingerprints, our study uniquely leveraged the rich, contextual representations extracted from Protein Language Models. We evaluated a diverse set of probabilistic models across multiple polymer properties, providing a comprehensive characterization of predictive accuracy alongside robust uncertainty estimates.

Table 1: Evaluation Results.

MAE (↓)

Model	Density		Tc		Tg		Rg		FFV	
	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT
BNN VI	0.0183	0.0216	0.0271	0.0292	0.1033	0.1132	0.0445	0.0456	0.0086	0.0100
Deep Ensemble	0.0194	0.0381	0.0278	0.0474	0.1035	0.1189	0.0443	0.0659	0.0222	0.0236
DEL	0.0280	0.0429	0.0354	0.0445	0.1274	0.1351	0.0585	0.0670	0.0129	0.0175
DKL	0.0294	0.0190	0.0306	0.0289	0.1229	0.1261	0.0494	0.0417	0.0079	0.0073
LA	0.0162	0.0194	0.0257	0.0274	**0.1000**	0.1050	0.0421	0.0400	0.0094	0.0112
Mask Ensemble	0.0165	**0.0183**	**0.0247**	**0.0260**	0.1126	0.1229	0.0412	**0.0395**	0.0089	0.0091
MC Dropout	0.0184	0.0194	0.0257	0.0277	0.1002	0.1070	0.0424	0.0403	0.0094	0.0112
MFVI	0.0178	0.0213	0.0275	0.0290	0.1245	0.1433	0.0445	0.0453	0.0196	0.0210
MVE	0.0195	0.1206	0.0271	0.0631	0.1041	0.1077	0.0418	0.0899	0.0162	0.0520
QREG	**0.0151**	0.0196	0.0256	0.0278	0.1018	0.1104	0.0411	0.0411	0.0079	0.0120
SWAG	0.0168	0.0190	0.0254	0.0271	0.1020	0.1086	0.0415	0.0421	0.0090	0.0104
VBLL	0.0445	0.0655	0.0375	0.0636	0.1151	0.1314	0.0583	0.0750	0.0088	0.0097
VBLL SNGP	0.1351	0.0899	0.0716	0.1134	0.1226	**0.1048**	0.1377	0.1807	**0.0060**	**0.0069**
ZigZag	0.0164	0.0186	0.0255	0.0271	0.1017	0.1076	**0.0410**	0.0416	0.0085	0.0108
Deterministic	0.0162	0.0194	0.0257	0.0274	**0.1000**	0.1050	0.0421	0.0400	0.0094	0.0112

PCC (↑)

Model	Density		Tc		Tg		Rg		FFV	
	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT
BNN VI	0.9102	0.8970	0.9019	0.8925	0.8013	0.7481	0.8448	0.8419	0.8918	0.8451
Deep Ensemble	0.9004	0.7085	0.9011	0.8037	0.8026	0.7153	0.8438	0.7419	0.7446	0.4683
DEL	0.9010	0.8474	0.8826	0.8664	0.7066	0.6715	0.8307	0.8208	0.7938	0.6763
DKL	0.8866	**0.9174**	0.8769	0.8884	0.7187	0.6906	0.8406	0.8446	0.9088	0.9229
LA	**0.9265**	0.9156	0.9109	**0.9017**	0.8069	0.7814	0.8606	**0.8694**	0.8697	0.8093
Mask Ensemble	0.9072	0.9006	0.9075	0.8982	0.7610	0.6904	0.8512	0.8624	0.8316	0.8551
MC Dropout	0.9223	0.9155	0.9104	0.8989	**0.8078**	0.7757	0.8600	0.8674	0.8690	0.8077
MFVI	0.9180	0.8955	0.9002	0.8890	0.7169	0.6342	0.8444	0.8370	0.6124	0.5734
MVE	0.9060	0.2897	0.9089	0.6502	0.7994	0.7684	0.8628	0.6049	0.7285	0.3875
QREG	0.9254	0.9068	0.9046	0.8957	0.8004	0.7599	0.8627	0.8637	0.8838	0.7893
SWAG	0.9250	0.9145	**0.9109**	0.9004	0.8073	0.7794	0.8604	0.8634	0.8722	0.8253
VBLL	0.7881	0.4882	0.8913	0.6933	0.7722	0.6866	0.8050	0.6161	0.9119	0.8708
VBLL SNGP	0.5392	0.7400	0.6827	0.5918	0.6397	0.7763	0.5622	0.5518	**0.9409**	**0.9282**
ZigZag	0.9205	0.9140	0.9095	0.9017	0.8073	**0.7816**	**0.8643**	0.8653	0.8980	0.8506
Deterministic	**0.9265**	0.9156	0.9109	**0.9017**	0.8069	0.7814	0.8606	**0.8694**	0.8697	0.8093

CRPS (↓)

Model	Density		Tc		Tg		Rg		FFV	
	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT
BNN VI	0.0169	0.0192	0.0256	0.0272	0.1011	0.1059	0.0429	0.0425	0.0077	0.0084
Deep Ensemble	0.0162	0.0299	0.0221	0.0324	0.0875	0.0905	0.0365	0.0466	0.0129	0.0201
DEL	0.1286	0.1443	0.1297	0.1403	0.1645	0.1775	0.1415	0.1539	0.0993	0.1187
DKL	0.0213	0.0149	0.0230	0.0219	0.0920	0.0999	0.0368	0.0326	0.0068	0.0064
LA	0.0158	0.0191	0.0256	0.0271	0.0994	0.1017	0.0420	0.0396	0.0094	0.0111
Mask Ensemble	**0.0119**	**0.0134**	**0.0181**	**0.0193**	0.0834	0.0930	**0.0301**	**0.0287**	0.0064	0.0066
MC Dropout	0.0140	0.0143	0.0194	0.0206	0.0770	0.0784	0.0315	0.0304	0.0069	0.0083
MFVI	0.0206	0.0234	0.0251	0.0274	0.0972	0.1116	0.0397	0.0400	0.2406	0.2451
MVE	0.0164	0.1276	0.0207	0.0982	0.0736	0.0766	0.0312	0.0983	0.1397	0.2399
QREG	0.0117	0.0152	0.0189	0.0213	**0.0738**	0.0807	0.0304	0.0306	0.0061	0.0093
SWAG	0.0152	0.0178	0.0247	0.0261	0.0996	0.1042	0.0406	0.0399	0.0087	0.0098
VBLL	0.1212	0.1353	0.0688	0.1311	0.1917	0.1590	0.1099	0.1419	0.0081	0.0081
VBLL SNGP	0.0982	0.0578	0.0478	0.0734	0.0913	**0.0765**	0.0981	0.1155	**0.0046**	**0.0052**
ZigZag	0.0147	0.0172	0.0228	0.0253	0.0919	0.0951	0.0364	0.0397	0.0076	0.0096

CA (↓)

Model	Density		Tc		Tg		Rg		FFV	
	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT	PolyNC	PolyBERT
BNN VI	0.0635	0.1315	0.1510	**0.0213**	0.2896	0.3011	0.1461	0.0735	0.0858	0.0513
Deep Ensemble	0.1663	0.2641	**0.0230**	0.2215	0.1611	0.1034	0.0528	0.1329	0.4517	0.4577
DEL	0.4170	0.3884	0.3971	0.3798	0.2019	0.2092	0.3457	0.3362	0.4487	0.4428
DKL	0.1860	0.0856	0.1818	0.2191	0.2899	0.3256	0.2223	0.2119	**0.0394**	0.0462
LA	0.4687	0.4705	0.4717	0.4728	0.4693	0.4557	0.4747	0.4691	0.4750	0.4744
Mask Ensemble	0.1932	0.1437	0.2151	0.1911	0.2422	0.2953	0.2080	0.2146	0.1297	0.1234
MC Dropout	0.1035	0.1650	0.2075	0.1930	0.3005	0.2130	0.2708	0.2165	0.1970	0.1637
MFVI	0.1823	0.1595	0.0668	0.0911	**0.0451**	**0.0627**	**0.0521**	**0.0517**	0.4652	0.4637
MVE	**0.0398**	0.1205	0.0827	0.1760	0.1906	0.1919	0.1044	0.1207	0.2964	0.4181
QREG	0.0977	**0.0790**	0.1710	0.1130	0.1620	0.1596	0.1554	0.1350	0.0912	0.0698
SWAG	0.4085	0.4405	0.4607	0.4610	0.4642	0.4594	0.4618	0.4498	0.4577	0.4443
VBLL	0.3565	0.3174	0.2869	0.3164	0.2722	0.1834	0.2945	0.3002	0.0698	**0.0228**
VBLL SNGP	0.0741	0.1447	0.1732	0.2197	0.1241	0.1573	0.1188	0.2109	0.1095	0.1502
ZigZag	0.3765	0.4226	0.4112	0.4329	0.4307	0.4139	0.3985	0.4547	0.4185	0.4116

References

[1] A. Amini, W. Schwarting, A. Soleimany, and D. Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

[2] E. Daxberger, E. Nalisnick, J. U. Allingham, J. Antorán, and J. M. Hernández-Lobato. Bayesian deep learning via subnetwork inference. In *International Conference on Machine Learning*, pages 2510–2521. PMLR, 2021.

[3] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR, 2018.

[4] N. Durasov, T. Bagautdinov, P. Baque, and P. Fua. Masksembles for uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13539–13548, 2021.

[5] N. Durasov, N. Dorndorf, H. Le, and P. Fua. Zigzag: Universal sampling-free uncertainty estimation through two-step inference. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QSvb6jBXML.

[6] ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning, 2024. URL https://evolutionaryscale.ai/blog/esm-cambrian.

[7] Y. Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[8] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[9] P. L. Golas and K. Matyjaszewski. Marrying click chemistry with polymerization: expanding the scope of polymeric materials. *Chem. Soc. Rev.*, 39(4):1338–1354, 2010. ISSN 1460-4744. doi: 10.1039/b901978m. URL http://dx.doi.org/10.1039/B901978M.

[10] J. Harrison, J. Willes, and J. Snoek. Variational bayesian last layers. In *International Conference on Learning Representations (ICLR)*, 2024.

[11] T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. A. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido, and A. Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736): 850–858, 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/abs/10.1126/science.ads0018.

[12] H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, Oct. 2000. ISSN 1520-0434. doi: 10.1175/1520-0434(2000)015<0559:dotcrp>2.0.co;2. URL http://dx.doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

[13] C. Jin, S. Guo, S. Zhou, and J. Guan. Effective and explainable molecular property prediction by chain-of-thought enabled large language models and multi-modal molecular information fusion. *Journal of Chemical Information and Modeling*, 65(11):5438–5455, 2025. doi: 10.1021/acs.jcim.5c00577. URL https://doi.org/10.1021/acs.jcim.5c00577. PMID: 40392109.

[14] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[15] C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad. Polymer genome: A data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C*, 122(31):17575–17585, July 2018. ISSN 1932-7455. doi: 10.1021/acs.jpcc.8b02913. URL http://dx.doi.org/10.1021/acs.jpcc.8b02913.

[16] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33, Jan. 1978. ISSN 0012-9682. doi: 10.2307/1913643. URL http://dx.doi.org/10.2307/1913643.

[18] D. V. Krevelen. *Properties of Polymers; Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions*. Elsevier, 1997. ISBN 9780444828774. doi: 10.1016/c2009-0-05459-2.

[19] C. Kuenneth and R. Ramprasad. polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nature Communications*, 14(1), July 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-39868-6. URL http://dx.doi.org/10.1038/s41467-023-39868-6.

[20] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[21] N. Lehmann, N. M. Gottschling, J. Gawlikowski, A. J. Stewart, S. Depeweg, and E. Nalisnick. Lightning uq box: Uncertainty quantification for neural networks. *Journal of Machine Learning Research*, 26(54):1–7, 2025. URL http://jmlr.org/papers/v26/24-2110.html.

[22] S. Li, Y. Tan, S. Ke, L. Hong, and B. Zhou. Immunogenicity prediction with dual attention enables vaccine target selection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hWmwL9gizZ.

[23] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

[24] G. Liu, J. Xu, E. Inae, Y. Zhu, Y. Li, T. Luo, M. Jiang, Y. Yan, W. Reade, S. Dane, A. Howard, and M. Cruz. Neurips - open polymer prediction 2025. https://kaggle.com/competitions/neurips-open-polymer-prediction-2025, 2025. Kaggle.

[25] Y. Liu, H. Du, L. Liu, and J. Leng. Shape memory polymers and their composites in aerospace applications: a review. *Smart Materials and Structures*, 23(2):023001, Jan. 2014. ISSN 1361-665X. doi: 10.1088/0964-1726/23/2/023001. URL http://dx.doi.org/10.1088/0964-1726/23/2/023001.

[26] S. Madan, M. Lentzen, J. Brandt, D. Rueckert, M. Hofmann-Apitius, and H. Fröhlich. Transformer models in biomedicine. *BMC Medical Informatics and Decision Making*, 24(1), July 2024. ISSN 1472-6947. doi: 10.1186/s12911-024-02600-5. URL http://dx.doi.org/10.1186/s12911-024-02600-5.

[27] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

[28] N. Meinert, J. Gawlikowski, and A. Lavin. The unreasonable effectiveness of deep evidential regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9134–9142, Jun. 2023. doi: 10.1609/aaai.v37i8.26096. URL https://ojs.aaai.org/index.php/AAAI/article/view/26096.

[29] L. A. Miccio and G. A. Schwartz. From chemical structure to quantitative polymer properties prediction through convolutional neural networks. *Polymer*, 193:122341, 2020. ISSN 0032-3861. doi: https://doi.org/10.1016/j.polymer.2020.122341. URL https://www.sciencedirect.com/science/article/pii/S0032386120301786.

[30] L. A. Miccio and G. A. Schwartz. Localizing and quantifying the intra-monomer contributions to the glass transition temperature using artificial neural networks. *Polymer*, 203:122786, 2020. ISSN 0032-3861. doi: https://doi.org/10.1016/j.polymer.2020.122786. URL https://www.sciencedirect.com/science/article/pii/S0032386120306145.

[31] L. A. Miccio and G. A. Schwartz. Mapping chemical structure–glass transition temperature relationship through artificial intelligence. *Macromolecules*, 54(4):1811–1817, 2021. doi: 10.1021/acs.macromol.0c02594. URL https://doi.org/10.1021/acs.macromol.0c02594.

[32] H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2): 107–113, May 1965. ISSN 1541-5732. doi: 10.1021/c160017a018. URL http://dx.doi.org/10.1021/c160017a018.

[33] D. Nix and A. 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, 1994. doi: 10.1109/ICNN.1994.374138.

[34] G. Pilania, C. N. Iverson, T. Lookman, and B. L. Marrone. Machine-learning-based predictive modeling of glass transition temperatures: A case of polyhydroxyalkanoate homopolymers and copolymers. *Journal of Chemical Information and Modeling*, 59(12):5013–5025, 2019. doi: 10.1021/acs.jcim.9b00807. URL https://doi.org/10.1021/acs.jcim.9b00807. PMID: 31697891.

[35] H. Qiu, L. Liu, X. Qiu, X. Dai, X. Ji, and Z.-Y. Sun. Polync: a natural and chemical language model for the prediction of unified polymer properties. *Chemical Science*, 15(2):534–544, 2024. ISSN 2041-6539. doi: 10.1039/d3sc05079c. URL http://dx.doi.org/10.1039/D3SC05079C.

[36] H. Ritter, A. Botev, and D. Barber. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skdvd2xAZ.

[37] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

[38] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

[39] L. Sluijterman, E. Cator, and T. Heskes. Optimal training of mean variance estimation neural networks. *Neurocomputing*, 597:127929, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2024.127929. URL https://www.sciencedirect.com/science/article/pii/S0925231224007008.

[40] H. Tang, T. Yue, and Y. Li. Assessing uncertainty in machine learning for polymer property prediction: A benchmark study. *Journal of Chemical Information and Modeling*, 65(13):6585–6598, 2025. doi: 10.1021/acs.jcim.5c00550. URL https://doi.org/10.1021/acs.jcim.5c00550. PMID: 40560148.

[41] L. Tao, V. Varshney, and Y. Li. Benchmarking machine learning models for polymer informatics: An example of glass transition temperature. *Journal of Chemical Information and Modeling*, 61(11):5395–5413, Oct. 2021. ISSN 1549-960X. doi: 10.1021/acs.jcim.1c01031. URL http://dx.doi.org/10.1021/acs.jcim.1c01031.

[42] F. Tavazza, B. DeCost, and K. Choudhary. Uncertainty prediction for machine learning models of material properties. *ACS omega*, 6(48):32431–32440, 2021.

[43] H. Tran, R. Gurnani, C. Kim, G. Pilania, H.-K. Kwon, R. P. Lively, and R. Ramprasad. Design of functional and sustainable polymers assisted by artificial intelligence. *Nature Reviews Materials*, 9(12):866–886, Aug. 2024. ISSN 2058-8437. doi: 10.1038/s41578-024-00708-8. URL http://dx.doi.org/10.1038/s41578-024-00708-8.

[44] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, July 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1335-8. URL http://dx.doi.org/10.1038/s41586-019-1335-8.

[45] B. D. Ulery, L. S. Nair, and C. T. Laurencin. Biomedical applications of biodegradable polymers. *Journal of Polymer Science Part B: Polymer Physics*, 49(12):832–864, May 2011. ISSN 1099-0488. doi: 10.1002/polb.22259. URL http://dx.doi.org/10.1002/polb.22259.

[46] D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, Feb. 1988. ISSN 1520-5142. doi: 10.1021/ci00057a005. URL http://dx.doi.org/10.1021/ci00057a005.

[47] D. Weininger. Smiles. 3. depict. graphical depiction of chemical structures. *Journal of Chemical Information and Computer Sciences*, 30(3):237–243, Aug. 1990. ISSN 1520-5142. doi: 10.1021/ci00067a005. URL http://dx.doi.org/10.1021/ci00067a005.

[48] D. Weininger, A. Weininger, and J. L. Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, May 1989. ISSN 1520-5142. doi: 10.1021/ci00062a008. URL http://dx.doi.org/10.1021/ci00062a008.

[49] B. K. Wheatle, E. F. Fuentes, N. A. Lynd, and V. Ganesan. Design of polymer blend electrolytes through a machine learning approach. *Macromolecules*, 53(21):9449–9459, 2020. doi: 10.1021/acs.macromol.0c01547. URL https://doi.org/10.1021/acs.macromol.0c01547.

[50] A. G. Wilson and P. Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

[51] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR, 2016.

[52] S. Wu, Y. Kondo, M.-a. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa, and R. Yoshida. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Computational Materials*, 5(1), June 2019. ISSN 2057-3960. doi: 10.1038/s41524-019-0203-2. URL http://dx.doi.org/10.1038/s41524-019-0203-2.

[53] N. Zhan and J. R. Kitchin. Uncertainty quantification in machine learning and nonlinear least squares regression models. *AIChE Journal*, 68(6), Nov. 2021. ISSN 1547-5905. doi: 10.1002/aic.17516. URL http://dx.doi.org/10.1002/aic.17516.

A Evaluation Metrics

Mean Absolute Error (MAE): Considering the trained property predictor model f_θ, the polymer representation x, and the ground truth property value y, the MAE metric is evaluated as:

$$\textbf{MAE}(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} |f_\theta(x_i) - y_i| \tag{3}$$

Pearson Correlation Coefficient (PCC): Considering the trained property predictor model f_θ, the test set polymer representation $\mathbf{x}$, and the ground truth property value of test set $\mathbf{y}$, and the predicted property values of test set $\hat{\mathbf{y}}$, the PCC metric is evaluated as:

$$\textbf{PCC}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\text{Cov}(\hat{\mathbf{y}}, \mathbf{y})}{\text{Var}(\mathbf{y})\text{Var}(\hat{\mathbf{y}})} \tag{4}$$

Continuous Ranked Probability Score (CRPS): Considering the trained property predictor model f_θ, the polymer representation x, and the ground truth property value y, the CRPS metric is evaluated as:

$$\textbf{CRPS}(f_\theta(x_i), y_i) = \int_{-\infty}^{\infty} [\Phi(f_\theta(x_i)) - 1_{f_\theta(x_i) \geq y_i}]^2 df_\theta(x_i), \tag{5}$$

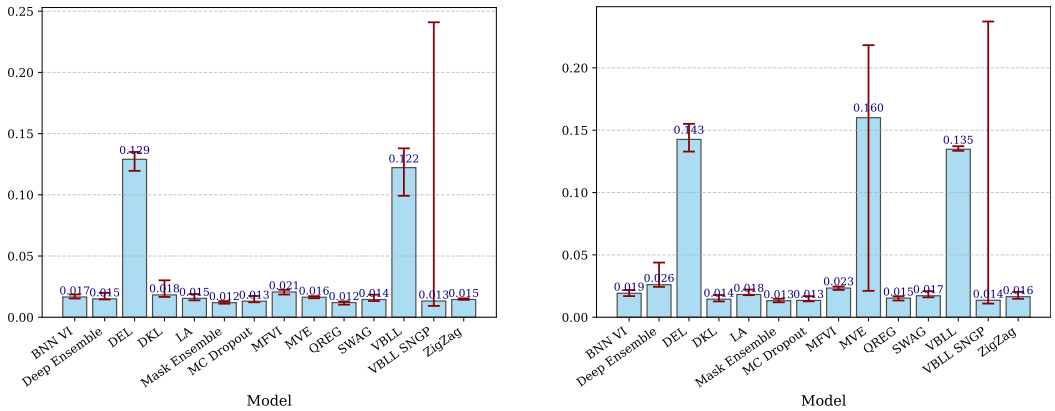
Here $\Phi(f_\theta(x_i))$ is the cumulative distribution function of the predictive distribution $f_\theta(x_i)$. To evaluate CRPS over the full test set $(\mathbf{x}, \mathbf{y})$, we take the average over the CRPS for each sample in the test set:

$$\textbf{CRPS}(f_\theta, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i} \textbf{CRPS}(f_\theta(x_i), y_i) \tag{6}$$

Calibration Area (CA): We follow the definition of CA metric defined in Tang et al. [40]. The process for evaluating this metric involves constructing confidence intervals based on uncertainty estimates and then determining the proportion of true values that lie within these intervals. Plotting the observed confidence against the expected confidence yields the calibration curve. A well-calibrated model's curve will align with the diagonal, while deviations reveal overconfidence (below the diagonal) or under-confidence (above the diagonal). The CA metric is the area between the calibration curve and the diagonal.

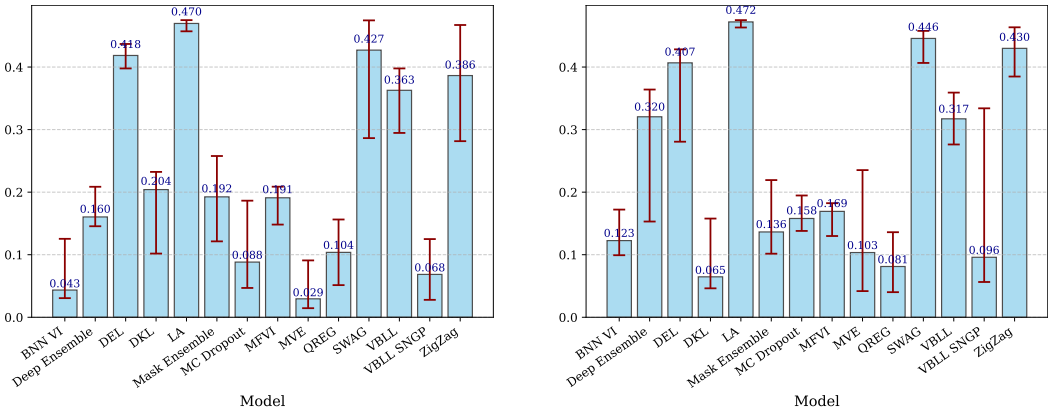
B Additional Results

To show the variability of our results, we've created bar plots for all models and polymer properties, illustrating the range of metric values over the five cross-validation splits. The value on each bar represents the median value among the metrics for the cross-validation splits.



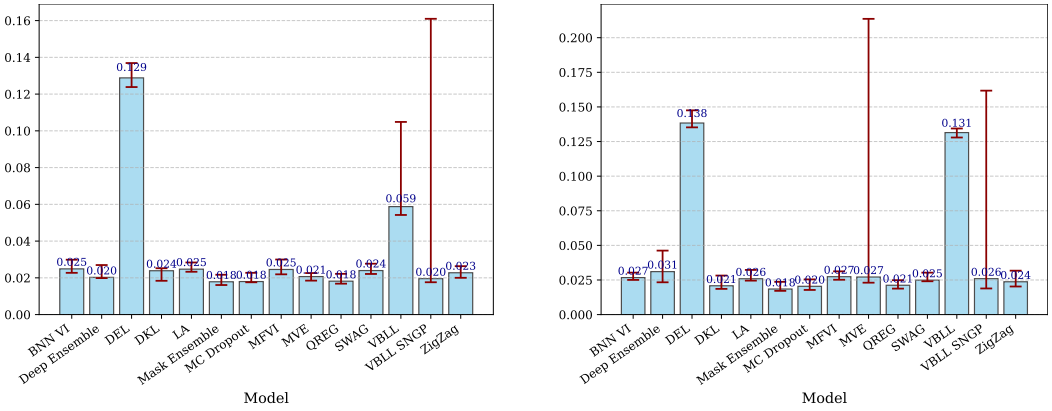
(a) PolyNC (b) PolyBERT

Figure 1: CRPS values for Density property of polymers.



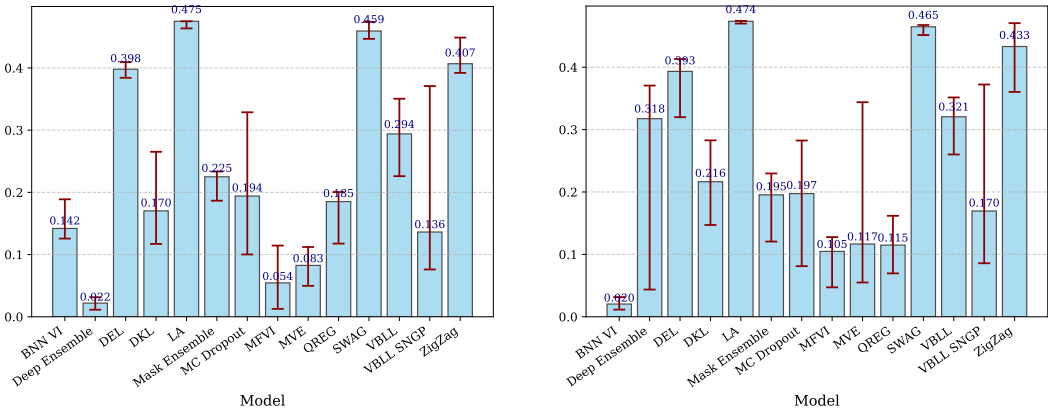
(a) PolyNC (b) PolyBERT

Figure 2: CA values for Density property of polymers.



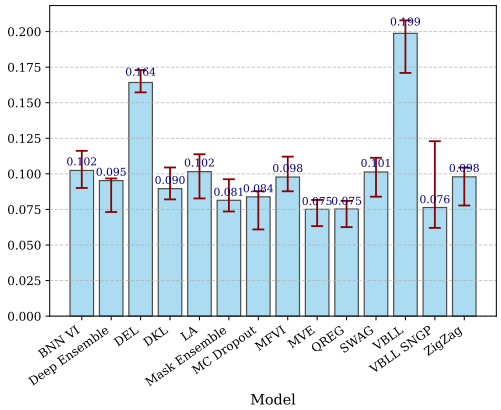
(a) PolyNC (b) PolyBERT

Figure 3: CRPS values for Tc property of polymers.

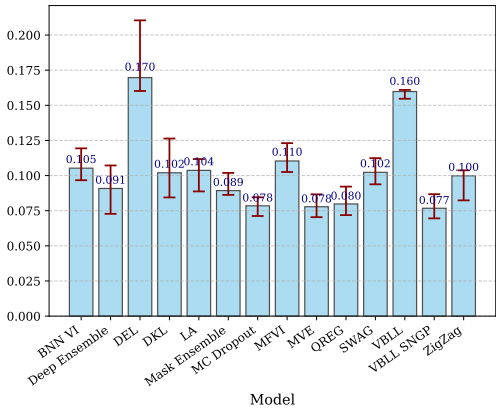


(a) PolyNC (b) PolyBERT

Figure 4: CA values for Tc property of polymers.

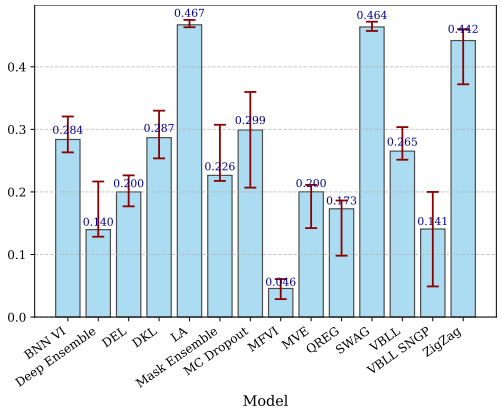


(a) PolyNC

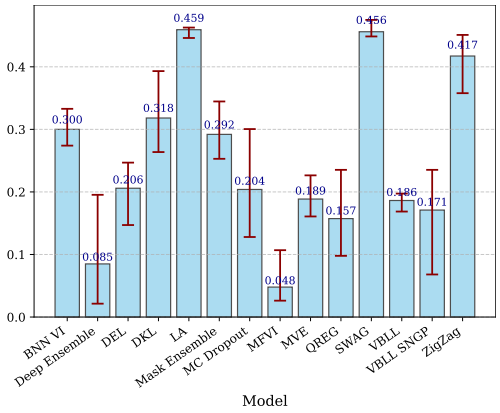


(b) PolyBERT

Figure 5: CRPS values for Tg property of polymers.

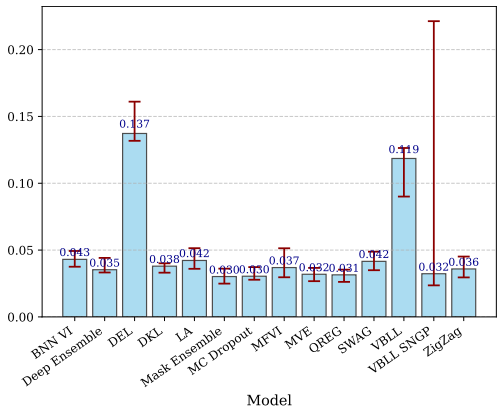


(a) PolyNC

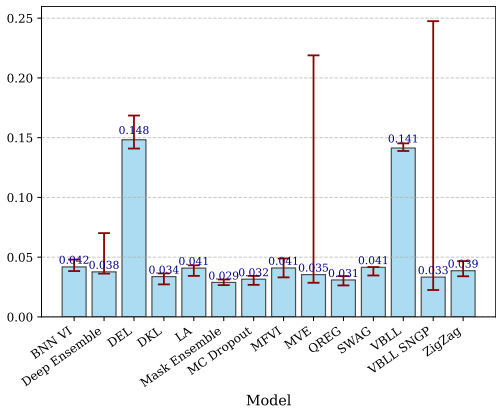


(b) PolyBERT

Figure 6: CA values for Tg property of polymers.

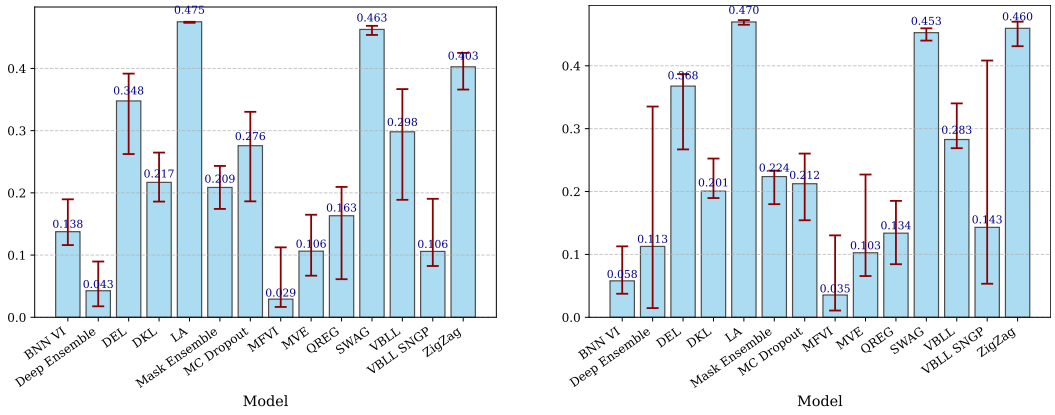


(a) PolyNC



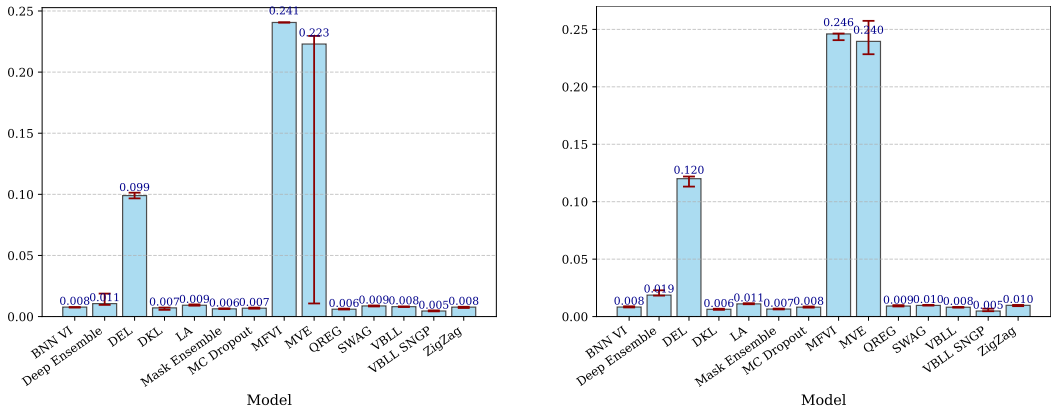
(b) PolyBERT

Figure 7: CRPS values for Rg property of polymers.



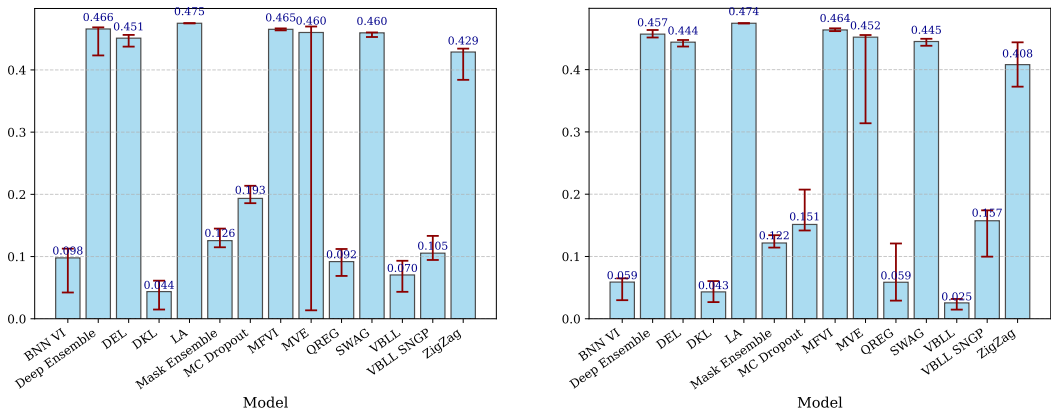
(a) PolyNC (b) PolyBERT

Figure 8: CA values for Rg property of polymers.



(a) PolyNC (b) PolyBERT

Figure 9: CRPS values for FFV property of polymers.



(a) PolyNC (b) PolyBERT

Figure 10: CA values for FFV property of polymers.