ImmUQBench: A Benchmark on Uncertainty Quantification of Protein Immunogenicity Prediction

Alif Bin Abdul Qayyum[1*] **Amir Hossein Rahmati**[1*]
Xiaoning Qian[1,2,3] **Byung-Jun Yoon**[1,3]
[1]Department of Electrical and Computer Engineering, Texas A&M University
[2]Department of Computer Science and Engineering, Texas A&M University
[3]Computing and Data Sciences, Brookhaven National Laboratory

Abstract

Discovering antigen proteins, capable of eliciting desired immune responses, is of paramount importance in developing immunogenic therapeutics for combating various diseases, particularly autoimmune disorders, infectious diseases, as well as cancers. Accurate and generalizable immunogenicity prediction with recent AI/ML advancements that can guide antigen design has emerged as a crucial subject in computational therapeutic discovery. However, due to insufficient labeled data, existing approaches tend to be overly simple. Many immunogenicity prediction models do not generalize well, making their predictions unreliable. Uncertainty Quantification (UQ) approaches are commonly used to address the aforementioned challenges when applying AI/ML methods with limited training data, aiming to reduce the risk of catastrophic errors. In developing AI/ML immunogenicity prediction models, these errors may lead to significant waste in cost and time for consequent therapeutic development for new immunogenic antigen proteins. We here present **ImmUQBench**, a benchmark for evaluating different well-known UQ methods for antigen immunogenicity prediction. Our work has the potential to facilitate more effective and reliable therapeutic antigen design, by providing insights into the efficacy of different UQ methods on immunogenicity predictions.

1 Introduction

Immunogenicity refers to the ability of a pathogen to provoke host immune responses. Identifying which pathogen proteins are likely to trigger an immune response is an absolutely vital step when developing new protein-based immunogenic therapeutics, e.g. vaccines [12, 1]. This process, often called *immunogenicity prediction*, is a key task that helps scientists anticipate and manage potential problems before a therapeutic even reaches clinical trials. Deep learning models have significantly advanced this task by enabling scalable and accurate predictions [33, 59]. However, their effectiveness is often hindered by the scarcity of labeled data and a mismatch between the task complexity and model assumptions, leading to suboptimal performance and limited generalizability, particularly when designing for broad viral efficacy.

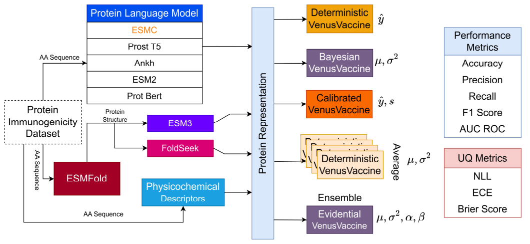

Figure 1: Uncertainty Quantification (UQ) in protein immunogenicity prediction

The advent of Large Language Models (LLMs) marks a pivotal advancement in natural language processing (NLP), fundamentally reshaping its

[*]equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Structured Probabilistic Inference & Generative Modeling: Probabilistic Inference in the Era of Large Foundation Models.

capabilities [50, 10, 6]. This progress has, in turn, facilitated the emergence of general-purpose computational tools within the field of biology. In particular, the adaptation of language modeling techniques to proteins has led to the emergence of powerful protein language models (PLMs), which have demonstrated strong performance on a variety of downstream tasks [57, 15, 17]. These models frequently surpass traditional approaches and offer improved generalization capabilities.

Despite their empirical success, both traditional ML prediction models and PLM integrated DL models designed for downstream tasks often exhibit overconfident predictions and are prone to generating hallucinated outputs [19, 48], raising concerns about their reliability and trustworthiness in sensitive applications, e.g. safety and efficacy related therapeutic design. To mitigate these limitations, the machine learning community has increasingly turned to uncertainty quantification (UQ) techniques. Broadly, UQ methods fall into two categories: Bayesian approaches [8, 11, 18], which provide a principled probabilistic framework but can be computationally intensive or impractical, and non-Bayesian approaches [2, 9, 28], which are often more tractable and performant but lack strong theoretical guarantees.

In this work, we introduce **ImmUQBench**, a benchmark designed to systematically evaluate a diverse set of UQ methods in the context of immunogenicity prediction. Specifically, we compare both Bayesian and non-Bayesian methods across various experimental scenarios, including in-distribution and out-of-distribution settings, to assess their predictive accuracy, uncertainty estimation quality, and robustness to distributional shifts. Additionally, we conduct an ablation study to evaluate the performance of these methods when part of the input information is missing, comparing them against their deterministic counterparts. We also examine the impact of different protein sequence encoding schemes, highlighting robustness to alternate encodings that express the same underlying sequence information.

To the best of our knowledge, **ImmUQBench** is the first comprehensive benchmark to assess UQ methods for immunogenicity prediction on three distinct immunogenic protein data sources, an essential step in therapeutic design, including vaccine development. Followings are the brief summaries of our contributions:

- *Pioneering Benchmark in Immunogenicity*: We introduce **ImmUQBench**, a benchmark for uncertainty quantification in immunogenicity prediction.
- *Extensive Evaluation of UQ across Several Data Distributions*: We systematically evaluate a wide range of Bayesian and non-Bayesian UQ approaches on three distinct immunogenic data sources, both across in-distribution and out-of-distribution scenarios.
- *Evaluation of Various Data Representation and Model Ablation*: We provide insights through extensive experiments and ablation studies to support antigen design and broadly effective therapeutic development.

2 UQ for Deep Neural Networks

As Deep Learning models have become widely adopted across a broad range of tasks, their trustworthiness and reliability of their prediction have become essential and critically important, as they struggle to distinguish between in and out-of-distribution datasets [22, 54] as well as being sensitive to domain shift [40]. This is particularly important in safety critical tasks where data is often scarce, wrong predictions can lead to severe consequences. Hence, it is essential that these models express their uncertainty when confronting out-of-distribution data. Different approaches have been developed and utilized for uncertainty quantification that can be broadly categorized into Bayesian methods [18, 29, 20, 43, 36, 32, 45, 8, 11] and non-Bayesian methods [61, 30, 28, 2, 4, 9, 3].

In this work, for each category of UQ methods, we consider widely used and representative approaches. Finally, as all the UQ methods ultimately aim to produce well-calibrated and reliable predictive distributions, we also include **Temperature Scaling (TS)** [22] as a baseline for comparison.

3 Experiments

In ImmUQBench, we focus on investigating UQ approaches in identifying whether proteins—originating from humans, bacteria, or viruses—are immunogenic. This task can be cast as a binary classification problem, where the model is trained to predict whether a given protein (or peptide segment) is an immunogenic antigen.

Table 1: In-Distribution immunogenicity prediction and UQ results.

Dataset	Model	Accuracy(↑)	Precision(↑)	Recall(↑)	F1 Score(↑)	AUC ROC(↑)	ECE(↓)	NLL(↓)	Brier Score(↓)
Virus	Ensemble	**0.9345**	**0.9479**	**0.9249**	**0.9363**	**0.9809**	0.0238	**0.1850**	**0.0521**
	Deterministic	0.9055	0.9082	0.9059	0.9071	0.9647	0.0413	0.2808	0.0711
	SWAG	0.9219	0.9380	0.9108	0.9242	0.9728	0.0743	0.2422	0.0667
	DROPOUT	0.9055	0.9082	0.9059	0.9071	0.9647	0.0411	0.2799	0.0710
	DVBLL	0.9030	0.9107	0.8995	0.9051	0.9621	0.0346	0.2882	0.0729
	LA	0.8904	0.9256	0.8674	0.8956	0.9633	**0.0108**	0.2498	0.0750
	EDL	0.9005	0.9305	0.8803	0.9047	0.9643	0.0152	0.2537	0.0743
	TS	0.9055	0.9082	0.9059	0.9071	0.9647	0.0311	0.2604	0.0699
	SVDKL	0.9156	0.9305	0.9058	0.9180	0.9577	0.1058	0.3050	0.0801
Bacteria	Ensemble	**0.8327**	0.6897	0.8054	0.7430	**0.8883**	0.0685	0.4788	0.1302
	Deterministic	0.8145	0.6897	0.7595	0.7229	0.8702	0.1332	0.7672	0.1526
	SWAG	**0.8327**	**0.7414**	0.7725	**0.7566**	0.8780	**0.0133**	**0.4046**	**0.1251**
	DROPOUT	0.8125	0.6897	0.7547	0.7207	0.8699	0.1349	0.7654	0.1525
	DVBLL	0.8185	0.6782	0.7763	0.7239	0.8743	0.0824	0.5112	0.1375
	LA	0.8246	0.7356	0.7574	0.7464	0.8675	0.0752	0.5251	0.1386
	EDL	0.7944	0.6494	0.7338	0.6890	0.8318	0.1259	0.5237	0.1739
	TS	0.8145	0.6897	0.7595	0.7229	0.8702	0.1192	0.6425	0.1475
	SVDKL	0.8246	0.6437	**0.8175**	0.7203	0.8689	0.1513	0.4991	0.1577
Tumor	Ensemble	0.7500	0.7869	0.6486	0.7111	0.8483	0.0413	0.4701	0.1599
	Deterministic	0.7436	0.7869	0.6400	0.7059	0.8336	0.0298	0.4908	0.1659
	SWAG	**0.7692**	0.7541	0.6866	0.7188	0.8537	**0.0189**	**0.4610**	**0.1527**
	DROPOUT	0.7500	0.7869	0.6486	0.7111	0.8330	0.0360	0.4908	0.1659
	DVBLL	0.7628	0.5574	0.7727	0.6476	**0.8585**	0.0669	0.4743	0.1574
	LA	**0.7692**	0.5574	**0.7907**	0.6538	0.8564	0.0724	0.4975	0.1639
	EDL	**0.7692**	**0.8033**	0.6712	**0.7313**	0.8552	0.0700	0.4867	0.1600
	TS	0.7436	0.7869	0.6400	0.7059	0.8336	0.0298	0.4908	0.1659
	SVDKL	0.6154	0.3770	0.5111	0.4340	0.5770	0.1132	0.6920	0.2494

In this study, we leverage **ImmunoDB**, the largest curated immunogenicity database with 7,216 labeled antigens from bacteria, viruses, and humans, providing a rigorous benchmark for model evaluation. As the backbone, we adopt VenusVaccine [34], a multimodal deep learning model that integrates sequence, structural, and physicochemical features through a hierarchical cross-attention framework. For our experiments, we follow VenusVaccine's training settings with slight architectural modifications—adding probabilistic layers to enable stable training and robust uncertainty estimation. Details are provided in the appendix.

3.1 ID (In-Distribution) Results

Table 1 presents the performance of various models on three in-distribution immunogenic datasets. In-distribution evaluation refers to the scenario where the train and test sets both originate from the same immunogenic data source, e.g. virus, bacteria or tumor (test data source = train data source). For each of the three immunogenic datasets, the majority of UQ methods showed superior results compared to the deterministic model across nearly all performance and uncertainty metrics, underscoring the benefits of uncertainty-aware modeling.

For Immuno-Virus dataset, the ensemble model outperforms other models in terms of all metrics except ECE where LA yields the best calibrated predictions. In evaluating the Immuno-Bacteria dataset, SWAG stands out for its dominant uncertainty quantification performance, achieving the best results across all relevant metrics. While it also performs better than most models in terms of predictive accuracy, SVDKL and Ensemble achieve slightly higher performance in Recall and AUC ROC, respectively. For the Immuno-Tumor dataset, SWAG consistently demonstrates the most effective uncertainty quantification, achieving the highest scores across all uncertainty metrics. However, in terms of predictive performance, it is often outperformed by LA or EDL, depending on the specific metric considered.

3.2 OoD (Out-of-Distribution) Results

Apart from the evaluation of models on test datasets, trained on respective data sources; we also report the out-of-distribution evaluation results based on the following three settings (test data source $\neq$ train data source). OoD results, when the model is trained on Bacteria and Tumor datasets, are provided in the appendix.

Generalization Performance of Models Trained on Immuno-Virus Dataset: Table 2 shows the evaluation of models on Immuno-Bacteria and Immuno-Tumor datasets, while they were trained on Immuno-Virus dataset. Although the ensemble model outperformed other models at the in-distribution scenario (Immuno-Virus test dataset), as mentioned in the first block of table 1, its performance degraded at the out-of-distribution scenario. The ensemble model performed best in the Immuno-Bacteria dataset in terms of the performance metrics, however it lagged behind in terms of the UQ

Table 2: Results on Immuno-Bacteria and Immuno-Tumor datasets of models trained on Immuno-Virus dataset.

Dataset	Model	Accuracy(↑)	Precision(↑)	Recall(↑)	F1 Score(↑)	AUC ROC(↑)	ECE(↓)	NLL(↓)	Brier Score(↓)
Bacteria	Ensemble	**0.7117**	**0.1954**	**0.9189**	**0.3223**	0.7026	0.1429	0.6879	**0.2110**
	Deterministic	0.6875	0.1437	0.8065	0.2439	0.6319	0.2817	2.0780	0.2934
	SWAG	0.6895	0.1379	0.8571	0.2376	**0.7265**	0.1863	0.6885	0.2277
	DROPOUT	0.6875	0.1437	0.8065	0.2439	0.6324	0.2814	2.0630	0.2932
	DVBLL	0.6774	0.1322	0.7188	0.2233	0.6682	0.2875	2.4999	0.2988
	LA	0.6915	0.1494	0.8387	0.2537	0.6934	0.1921	0.7689	0.2339
	EDL	0.6653	0.1782	0.5741	0.2719	0.6165	0.2254	0.8691	0.2743
	TS	0.6875	0.1437	0.8065	0.2439	0.6319	0.2702	1.6484	0.2875
	SVDKL	0.6935	0.1724	0.7895	0.2830	0.5950	**0.0916**	**0.6395**	0.2204
Tumor	Ensemble	0.5577	0.3770	0.4259	0.4000	0.5346	0.1072	0.7398	0.2632
	Deterministic	0.5449	0.5246	0.4324	0.4741	0.5381	0.2894	1.3243	0.3525
	SWAG	0.6538	**0.6393**	0.5493	**0.5909**	**0.6481**	**0.0789**	**0.6637**	**0.2345**
	DROPOUT	0.5449	0.5246	0.4324	0.4741	0.5377	0.2882	1.3166	0.3520
	DVBLL	0.5833	0.5574	0.4722	0.5113	0.6283	0.2880	1.5885	0.3346
	LA	0.5000	0.4426	0.3803	0.4091	0.5103	0.2354	0.8796	0.3095
	EDL	**0.6667**	0.5410	**0.5789**	0.5593	0.6388	0.1616	0.7663	0.2545
	TS	0.5449	0.5246	0.4324	0.4741	0.5381	0.2630	1.1369	0.3354
	SVDKL	0.6026	0.3607	0.4889	0.4151	0.5774	0.1320	0.7249	0.2592

metrics. Particularly, SWAG performed best in terms of AUC ROC, while SVDKL outperformed other models concerning ECE and NLL. The SWAG model performed best in terms of most of the performance and UQ metrics in the Immuno-Tumor dataset. To summarize, when the models are trained on Immuno-Virus dataset, both Ensemble and SWAG demonstrate superior performance compared to other models.

In summary, the deterministic model was outperformed by uncertainty-aware methods across almost all evaluation metrics on both datasets, highlighting its limited generalizability and reliability.

3.3 Discussion

All the abovementioned experimental results consistently show that uncertainty-aware models outperform deterministic baselines in immunogenicity prediction, offering gains in both predictive accuracy and calibration. These improvements were observed across all in-distribution datasets and persisted across different protein language model embeddings, suggesting that the benefits of UQ are largely independent of the upstream sequence representation. Enhanced calibration—most notably achieved by SWAG and SVDKL—has particular relevance in high-stakes biomedical applications, where overconfident errors can result in costly experimental misallocation or safety risks.

Performance differences became more nuanced across the out-of-distribution evaluation scenarios. SVDKL and SWAG demonstrated strong calibration robustness under distribution shift, whereas LA and Ensembles often achieved higher predictive performance in specific scenarios. This indicates a trade-off between reliability and sharpness that should be aligned with downstream objectives.

Overall, empirical findings establish UQ as a means of achieving more reliable and effective immunogenicity prediction and provide actionable guidance for selecting model–task configurations in both experimental and clinical settings.

4 Conclusion

In this study, we have introduced **ImmUQBench**, a new benchmark for evaluating a range of uncertainty quantification (UQ) methods on the task of immunogenicity prediction. This benchmark demonstrates that UQ methods deliver benefits extending beyond calibration, consistently enhancing predictive performance across diverse datasets and embedding strategies. By leveraging multimodal information and incorporating it through a state-of-the-art backbone model, our benchmark enables comprehensive evaluation under both in-distribution and out-of-distribution settings. Our results demonstrate that most UQ methods consistently outperform the deterministic baseline across various metrics. In particular, Ensemble, SWAG, and Laplace Approximation (LA) exhibit superior performance in terms of predictive accuracy, uncertainty estimation, and generalization. While some UQ methods occasionally underperform relative to the deterministic model, overall, UQ-based approaches yield more robust and calibrated predictions, especially in out-of-distribution scenarios. To further examine the performance of each method, we have also conducted ablation studies by removing structural information and analyzing the impact of different PLM encodings. **ImmUQBench**, the first UQ benchmark for immunogenicity prediction to the best of our knowledge, offers a valuable resource for developing more reliable and uncertainty-aware antigen design tools.

Funding Statement

This work has been supported by the Advanced Research Projects Agency for Health (ARPA-H) Award 1AY1AX000053-01.

References

[1] J. Adu-Bobie, B. Capecchi, D. Serruto, R. Rappuoli, and M. Pizza. Two years into reverse vaccinology. *Vaccine*, 21(7-8):605–610, Jan. 2003.

[2] A. Amini, W. Schwarting, A. Soleimany, and D. Rus. Deep evidential regression, 2020. URL `https://arxiv.org/abs/1910.02600`.

[3] A. N. Angelopoulos, S. Bates, M. Jordan, and J. Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=eNdiU_DbM9`.

[4] W. Bao, Q. Yu, and Y. Kong. Evidential deep learning for open set action recognition, 2021. URL `https://arxiv.org/abs/2107.10161`.

[5] G. W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1, Jan. 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020. URL `https://arxiv.org/abs/2005.14165`.

[7] J. J. A. Calis, M. Maybeno, J. A. Greenbaum, D. Weiskopf, A. D. De Silva, A. Sette, C. Keşmir, and B. Peters. Properties of mhc class i presented peptides that enhance immunogenicity. *PLOS Computational Biology*, 9(10):1–13, 10 2013. doi: 10.1371/journal.pcbi.1003266. URL `https://doi.org/10.1371/journal.pcbi.1003266`.

[8] T. Chen, E. B. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo, 2014. URL `https://arxiv.org/abs/1402.4102`.

[9] S. Cortes-Gomez, C. M. Patiño, Y. Byun, S. Wu, E. Horvitz, and B. Wilder. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=iOMnn1hSBO`.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL `https://arxiv.org/abs/1810.04805`.

[11] N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven. Bayesian sampling using stochastic gradient thermostats. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3203–3211, Cambridge, MA, USA, 2014. MIT Press.

[12] N. Doneva, I. Doytchinova, and I. Dimitrov. Predicting immunogenicity risk in biopharmaceuticals. *Symmetry*, 13(3), 2021. ISSN 2073-8994. doi: 10.3390/sym13030388. URL `https://www.mdpi.com/2073-8994/13/3/388`.

[13] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44 (10):7112–7127, Oct. 2022.

[14] A. Elnaggar, H. Essam, W. Salah-Eldin, W. Moustafa, M. Elkerdawy, C. Rochereau, and B. Rost. Ankh: Optimized protein language model unlocks general-purpose modelling, 2023. URL `https://arxiv.org/abs/2301.06568`.

[15] ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. https://evolutionaryscale.ai/blog/esm-cambrian, Dec. 2024. EvolutionaryScale Website.

[16] X. Fan, S. Zhang, K. Tanwisuth, X. Qian, and M. Zhou. Contextual dropout: An efficient sample-dependent dropout module, 2021. URL https://arxiv.org/abs/2103.04181.

[17] N. Ferruz, S. Schmidt, and B. Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.*, 13(1):4348, July 2022.

[18] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

[19] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. X. Zhu. A survey of uncertainty in deep neural networks, 2022. URL https://arxiv.org/abs/2107.03342.

[20] E. Goan and C. Fookes. *Bayesian Neural Networks: An Introduction and Survey*, page 45–87. Springer International Publishing, 2020. ISBN 9783030425531. doi: 10.1007/978-3-030-42553-1_3. URL http://dx.doi.org/10.1007/978-3-030-42553-1_3.

[21] K. P. Greenman, A. P. Amini, and K. K. Yang. Benchmarking uncertainty quantification for protein engineering. *PLoS Comput. Biol.*, 21(1):e1012639, Jan. 2025.

[22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks, 2017. URL https://arxiv.org/abs/1706.04599.

[23] J. Harrison, J. Willes, and J. Snoek. Variational bayesian last layers, 2024. URL https://arxiv.org/abs/2404.11599.

[24] 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.

[25] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, and B. Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(1):723, Dec. 2019.

[26] M. Heinzinger, K. Weissenow, J. Sanchez, A. Henkel, M. Mirdita, M. Steinegger, and B. Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 11 2024. ISSN 2631-9268. doi: 10.1093/nargab/lqae150. URL https://doi.org/10.1093/nargab/lqae150.

[27] P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, and A. G. Wilson. Subspace inference for bayesian deep learning, 2019. URL https://arxiv.org/abs/1907.07504.

[28] S. Jain, G. Liu, J. Mueller, and D. Gifford. Maximizing overall diversity for improved uncertainty estimates in deep ensembles, 2020. URL https://arxiv.org/abs/1906.07380.

[29] A. Kristiadi, M. Hein, and P. Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks, 2020. URL https://arxiv.org/abs/2002.10118.

[30] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017. URL https://arxiv.org/abs/1612.01474.

[31] C. H. Lee, J. Huh, P. R. Buckley, M. Jang, M. P. Pinho, R. A. Fernandes, A. Antanaviciute, A. Simmons, and H. Koohy. A robust deep learning workflow to predict CD8 + t-cell epitopes. *Genome Med.*, 15(1):70, Sept. 2023.

[32] J. Lee, M. Humt, J. Feng, and R. Triebel. Estimating model uncertainty of neural networks in sparse information form, 2020. URL https://arxiv.org/abs/2006.11631.

[33] G. Li, B. Iyer, V. B. S. Prasath, Y. Ni, and N. Salomonis. Deepimmuno: deep learning-empowered prediction and generation of immunogenic peptides for t-cell immunity. *Briefings in Bioinformatics*, 22(6):bbab160, 05 2021. ISSN 1477-4054. doi: 10.1093/bib/bbab160. URL https://doi.org/10.1093/bib/bbab160.

[34] S. Li, Y. Tan, S. Ke, L. Hong, and B. Zhou. Immunogenicity prediction with dual attention enables vaccine target selection, 2025. URL https://arxiv.org/abs/2410.02647.

[35] 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.

[36] D. J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 05 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.448. URL https://doi.org/10.1162/neco.1992.4.3.448.

[37] W. Maddox, T. Garipov, P. Izmailov, D. Vetrov, and A. G. Wilson. A simple baseline for bayesian uncertainty in deep learning, 2019. URL https://arxiv.org/abs/1902.02476.

[38] K. P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL http://probml.github.io/book1.

[39] H. O'Brien, M. Salm, L. T. Morton, M. Szukszto, F. O'Farrell, C. Boulton, L. King, S. K. Bola, P. D. Becker, A. Craig, M. Nielsen, Y. Samuels, C. Swanton, M. R. Mansour, S. R. Hadrup, and S. A. Quezada. A modular protein language modelling approach to immunogenicity prediction. *PLoS Comput. Biol.*, 20(11):e1012511, Nov. 2024.

[40] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, 2019. URL https://arxiv.org/abs/1906.02530.

[41] A. H. Rahmati, S. Jantre, W. Zhang, Y. Wang, B.-J. Yoon, N. M. Urban, and X. Qian. C-lora: Contextual low-rank adaptation for uncertainty estimation in large language models, 2025. URL https://arxiv.org/abs/2505.17773.

[42] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=fylclEqgvgd.

[43] 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.

[44] 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.

[45] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 880–887, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390267. URL https://doi.org/10.1145/1390156.1390267.

[46] H. Schellekens. Bioequivalence and the immunogenicity of biopharmaceuticals. *Nat. Rev. Drug Discov.*, 1(6):457–462, June 2002.

[47] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty, 2018. URL https://arxiv.org/abs/1806.01768.

[48] O. Shorinwa, Z. Mei, J. Lidard, A. Z. Ren, and A. Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions, 2025. URL https://arxiv.org/abs/2412.05563.

[49] Y. Tan, M. Li, Z. Zhou, P. Tan, H. Yu, G. Fan, and L. Hong. PETA: evaluating the impact of protein transfer learning with sub-word tokenization on downstream applications. *J. Cheminform.*, 16(1):92, Aug. 2024.

[50] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

[51] M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. M. Gilchrist, J. Söding, and M. Steinegger. Fast and accurate protein structure search with foldseek. *Nat. Biotechnol.*, 42(2):243–246, Feb. 2024.

[52] J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani. Bertology meets biology: Interpreting attention in protein language models, 2021. URL https://arxiv.org/abs/2006.15222.

[53] Y. Wang, H. Shi, L. Han, D. Metaxas, and H. Wang. Blob: Bayesian low-rank adaptation by backpropagation for large language models, 2025. URL https://arxiv.org/abs/2406.11675.

[54] A. G. Wilson and P. Izmailov. Bayesian deep learning and a probabilistic perspective of generalization, 2022. URL https://arxiv.org/abs/2002.08791.

[55] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Stochastic variational deep kernel learning, 2016. URL https://arxiv.org/abs/1611.00336.

[56] J. Wu, W. Wang, J. Zhang, B. Zhou, W. Zhao, Z. Su, X. Gu, J. Wu, Z. Zhou, and S. Chen. DeepHLApan: A deep learning approach for neoantigen prediction considering both HLA-peptide binding and immunogenicity. *Front. Immunol.*, 10:2559, Nov. 2019.

[57] M. Xu, Z. Zhang, J. Lu, Z. Zhu, Y. Zhang, C. Ma, R. Liu, and J. Tang. Peer: A comprehensive and multi-task benchmark for protein sequence understanding, 2022. URL https://arxiv.org/abs/2206.02096.

[58] A. X. Yang, M. Robeyns, X. Wang, and L. Aitchison. Bayesian low-rank adaptation for large language models, 2024. URL https://arxiv.org/abs/2308.13111.

[59] X. Yang, L. Zhao, F. Wei, and J. Li. DeepNetBim: deep learning model for predicting HLA-epitope interactions based on network analysis by harnessing binding and immunogenicity information. *BMC Bioinformatics*, 22(1):231, May 2021.

[60] C. Yu, X. Fang, and H. Liu. A unified cross-attention model for predicting antigen binding specificity to both hla and tcr molecules, 2025. URL https://arxiv.org/abs/2405.06653.

[61] J. Zhang, B. Kailkhura, and T. Y.-J. Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning, 2020. URL https://arxiv.org/abs/2003.07329.

A UQ for Deep Neural Networks

Bayesian Methods: We begin with **Monte Carlo (MC)-Dropout** [18], which interprets dropout as a form of approximate Bayesian inference in deep Gaussian Processes. This interpretation allows dropout to capture epistemic uncertainty by maintaining stochasticity at test time. Specifically, model predictions are obtained by performing multiple stochastic forward passes with dropout enabled and computing the predictive mean and variance from these passes. Formally, the predictive mean is approximated by averaging outputs over T stochastic passes, while the variance is estimated as the sample variance plus a model precision term τ^{-1}, where τ depends on the dropout rate, dataset size, and weight decay. This enables tractable Bayesian approximation without incurring additional test-time complexity.

Next, we consider **Variational Bayesian Last Layer (VBLL)** [23], which provides an efficient sampling-free approach to Bayesian modeling by maintaining a posterior only over the final layer of a neural network. By casting the training objective as a deterministic variational bound, VBLL introduces minimal overhead and is easily integrated into existing architectures, yielding principled uncertainty estimates without requiring stochastic forward passes.

We also include **Stochastic Variational Deep Kernel Learning (SVDKL)** [55], which synergizes the representation power of deep neural networks with the non-parametric flexibility of Gaussian Processes. SVDKL extends Deep Kernel Learning (DKL) to classification and multi-task learning using a scalable variational inference framework. This allows for training on large-scale datasets via stochastic gradients, and supports more expressive covariance structures compared to prior DKL models.

SWAG [37] (Stochastic Weight Averaging-Gaussian) is another Bayesian approximation technique that constructs a Gaussian posterior over network weights by leveraging the trajectory of stochastic gradient descent (SGD). It estimates the posterior mean via the running average of SGD iterates and approximates the covariance using both a low-rank approximation based on recent deviations from the mean and a diagonal component derived from the second moment. This results in a scalable approach to uncertainty estimation that enables Bayesian model averaging through weight sampling at test time.

Finally, we use the **Laplace Approximation (LA)** [36], which approximates the posterior distribution over model weights with a Gaussian centered at the maximum a posteriori (MAP) estimate. This is achieved by performing a second-order Taylor expansion of the log-posterior, resulting in a Gaussian with covariance given by the inverse Hessian of the log-posterior evaluated at the MAP point. Formally, the posterior is approximated as $p(\mathbf{w} \mid \mathcal{D}) \approx \mathcal{N}(\hat{\mathbf{w}}, \mathbf{H}^{-1})$, where $\mathbf{H}$ is the Hessian of the negative log posterior and $\hat{\mathbf{w}}$ is the MAP. This method provides a fast and principled estimate of uncertainty without requiring sampling during inference.

Non-Bayesian Methods: Among non-Bayesian approaches, we consider **Deep Ensembles** [30], which train an ensemble of M neural networks with different initializations. Each model outputs a probabilistic prediction, and the ensemble prediction is computed by averaging these outputs. This ensemble captures both model and data uncertainty and has been shown to outperform many Bayesian approximations in terms of calibration and robustness. Each network is typically trained using proper scoring rules such as negative log-likelihood to ensure meaningful probabilistic outputs.

We also evaluate **Evidential Deep Learning (EDL)** [47], which explicitly models predictive uncertainty by placing a Dirichlet distribution over class probabilities. Rather than producing point estimates via softmax, the network outputs non-negative "evidence" values that parameterize the Dirichlet. This allows the model to represent both aleatoric and epistemic uncertainty in a unified framework. The loss function combines the Bayes risk (under an L2 norm) with a KL-divergence term that regularizes the model to prevent overconfident predictions, enabling uncertainty-aware classification from a single forward pass.

B Immunogenicity

Immunogenicity is linked to the therapeutic use of proteins and can result in serious clinical outcomes, including reduced treatment effectiveness or potentially life-threatening complications. Naturally, determining the cause of immunogenicity in biologic therapies is a necessary pursuit [46]. Particularly, immunogenicity prediction has become a central component in reverse vaccinology aiming to identify antigens that are capable of eliciting immune responses resulting in the formation of memory cells within the host organism [34, 1].

Researchers are increasingly focused on fast and precise prediction of immunogenic antigens for vaccine development, as this approach minimizes costs and associated risks, while supporting safe and effective responses to infectious disease threats. [7] use a simple linear scoring function to calculates immunogenicity score. DeepImmuno [33] introduces two deep learning models aimed at modeling T-cell immunity, which is crucial for the development of cancer immunotherapies and vaccines. Specifically, DeepImmuno-CNN predicts immunogenicity, while DeepImmuno-GAN generates immunogenic peptides. TRAP [31] presents a robust deep learning framework for predicting CD8 + T-cell epitopes from both pathogenic and self-peptides. It also estimates the immunogenic potential of MHC-I peptides by providing a prediction score along with a confidence measure. Some current methods also consider using physiochemical properties of amino acids for immunogenicity prediction.

As our core model to integrate and evaluate different UQ approaches for immunogenicity prediction, we adopt VenusVaccine [34], a cutting-edge multi-modal deep learning framework. Leveraging a dual-attention mechanism, VenusVaccine integrates sequence, structural, and physicochemical information to effectively interpret immunogenicity.

C Protein Language Models (PLMs)

The adaptation of LLMs—the advent of which marked a major shift in natural language processing (NLP)—to protein sequences has resulted in the emergence of advanced protein language models (PLMs) [17, 44, 25]. This adaptation—hence modeling of protein sequences—was enabled by equating words with amino acids and interpreting the entire protein sequences as sentences [52, 42]. Via self-supervised learning, generic PLMs are often pre-trained on large datasets of amino acid sequences, which then due to learning contextual residue representations [44, 25], they can serve as feature extractors for a wide-range of protein tasks, such as prediction of structure, binding residues, sub-cellular localization, and fold classification.

D Problem Setup

Here, we formalize the problem, incorporating multi-modal information from protein sequences, structures, and physicochemical properties. Following [34], sequence and structure embeddings are extracted from pre-trained protein language models (PLMs). These embeddings are passed through the dual-attention module of VenusVaccine, which summarizes them into a unified representation:

$$H = \text{DualAtt}(E_{\text{seq}}, E_{\text{strc}}),$$

where

$$E_{\text{seq}} = \text{PLM}_{\text{seq}}(\mathbf{x}) \in \mathbb{R}^{L \times d}, \quad E_{\text{strc}} = \text{PLM}_{\text{strc}}(\mathbf{x}) \in \mathbb{R}^{L \times d}$$

denote the sequence and structure embeddings of the amino acid sequence $\mathbf{x}$ of length L, respectively, and d is the embedding dimension. The attention output H, along with the sequence and physicochemical features, is concatenated and passed to a classifier:

$$\hat{y} = f_\theta(Z), \quad Z = \text{concat}(E_{\text{seq}}, E_{\text{pc}}, H), \quad \hat{y} \in \{0, 1\},$$

where f_θ is a deterministic classifier parameterized by θ.

Bayesian Methods: In this work, for evaluating Bayesian methods, we treat θ as a random variable to enable uncertainty estimation and to evaluate the performance of different uncertainty quantification methods. Thus, the predictive distribution is given by:

$$\hat{y} = \mathbb{E}_{\theta \sim p_\Theta(\theta)}[f_\theta(Z)].$$

In ImmUQBench, we have implemented MC-Dropout [18], SWAG [37], DVBLL [23], SVDKL [55] and LA [36].

As **Non-Bayesian** methods employ varied and often method-specific mechanisms for uncertainty estimation, a general predictive formulation analogous to the Bayesian case is not readily available. Hence, we briefly outline the evaluation formulation for Deep Ensembles and EDL. In addition, we describe TS which is a widely-used calibration technique for adjusting predicted probabilities.

Deep Ensemble: By training M neural networks independently, we estimate the uncertainty. Specifically, each model outputs a prediction, and the ensemble predictive distribution is computed as the average,

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} f_{\theta^m}(Z).$$

The diversity among the members captures the uncertainty. Instead of training the same model with different initialization or data shuffling, we employed different data representation for different models in the ensemble. Each ensemble consists of 5 models, each with similar architecture following VenusVaccine [34]. However, the amino acid sequence level encoding, E_{seq}, for different model in the ensemble comes from different PLMs. The 5 PLMs used in this work are: ESM-Cambrian [15], ProstT5 [26], Ankh [14], ESM-2 [35] and Prot-Bert [13].

EDL: In a binary classification, EDL models the class probability as a Beta distribution, Beta(α_1, α_2). Considering the network outputs the evidence parameters, $\alpha_i = e_i + 1, i \in \{1, 2\}$, where $e_i > 0$, the predictive probability for class i is given by,

$$\hat{y} = \mathbb{E}[p] = \frac{\alpha_i}{\alpha_1 + \alpha_2}.$$

Uncertainty is then captured through the variance of the Beta distribution.

TS: To improve the calibration of predicted probabilities, TS introduces a scalar temperature parameter $T > 0$, which is optimized on a validation set by minimizing the negative log-likelihood. This adjustment rescales the logit outputs to produce softer probability distributions. Specifically, given the logit vector $\mathbf{z}$, the calibrated probabilities are computed as:

$$q = \sigma\left(\frac{\mathbf{z}}{T}\right),$$

where σ denotes the softmax function. TS adjusts confidence levels without affecting the model's accuracy, making it a simple yet effective post-hoc calibration method.

E Uncertainty Evaluation Metrics

We assess uncertainty quantification using three established metrics: Expected Calibration Error (ECE), negative log-likelihood (NLL), and the Brier score, which have been commonly employed in the literature. ECE and Brier scores are considered as calibration metrics while NLL is mostly regarded as an indicator of overconfidence.

A calibrated model is the one that its predicted probabilities match the empirical frequency of the output [38]. A well-calibrated model can prevent wrong decisions in case of high uncertainty. ECE is used to assess calibration. Particularly by partitioning predictions into M equally-spaced bins based on their prediction confidence, ECE can be calculated as [38, 22],

$$ECE = \sum_{m=1}^{M} \frac{B_m}{N} |\text{acc}(B_m) - \text{conf}(B_m)|,$$

with N indicating the size of the dataset, and $\text{acc}(B_m) = 1/|B_m| \sum_{i \in B_m} \mathbb{I}(\hat{y}_i = y_i)$ and $\text{conf}(B_m) = 1/|B_m| \sum_{i \in B_m} P(\hat{y}_i)$ the average accuracy and confidence in bin B_m with size $|B_m|$ respectively.

Calibration can also be evaluated by the Brier score, which is a proper scoring rule and a widely accepted tool in the context of uncertainty quantification due to its ability in assessing the quality of probabilistic predictions [38, 5]. Especially, it captures how correct a model is and if it expresses proper confidence levels, by measuring mean squared difference between predicted probabilities and predictions. For a binary class, it is,

$$Brier = \frac{1}{N} \sum_{i=1}^{N} (y_i - P(\hat{y}_i))^2$$

On the other hand, NLL is often used to detect overconfidence. It is computed as the negative log-probability assigned to the true label,

$$NLL = \sum_{i=1}^{N} -\log P(\hat{y}_i = y_i)$$

When a model is overconfident in an incorrect prediction, it assigns a high probability to the wrong class. Therefore, the log loss becomes very large, that results in a high NLL.

F Dataset

In this study, we use **ImmunoDB**, an immunogenicity database comprising 7,216 labeled antigens derived from three distinct sources: bacteria, viruses, and humans. Each antigen is labeled as either immunogenic (positive) or non-immunogenic (negative). The dataset is constructed through a combination of literature curation, database mining, and bioinformatics filtering, with most positive samples originating from previously published studies. To ensure quality, redundant sequences and samples from tail regions were filtered out. This process resulted in three curated subsets: **Immuno-Virus**, **Immuno-Bacteria**, and **Immuno-Tumor**. Owing to its rigorous quality control, diverse species coverage, and comprehensive sourcing, **ImmunoDB** provides a valuable benchmark for evaluating the robustness and generalizability of immunogenicity prediction models. To the best of our knowledge, at the time this study was conducted, it represents the most extensive labeled antigen resource available for this task.

G Backbone Architecture

In this work, we use **VenusVaccine** [34], a supervised deep learning model for immunogenicity prediction, as the backbone in our experiments. The model integrates sequence, structural, and physicochemical information using a dual attention mechanism. It encodes protein sequences with pretrained PLMs and represents structures at both atomic and peptide levels using FoldSeek [51] and ESM-3 [24], respectively. Handcrafted physicochemical descriptors are also included to enhance biological relevance.

The model employs a hierarchical cross-attention framework that fuses sequence and structure representations at multiple scales, enabling rich interaction across modalities. Attention pooling then compresses amino acid-level features into a protein-level vector by highlighting key regions, which is used for final binary classification of immunogenicity.

H Related Works

Protein Language Models. Advances in deep learning and the emergence of large language models, including specialized protein language models (PLMs), have revolutionized computational biology by offering accurate, generalizable, and scalable solutions to complex downstream tasks such as vaccinology, drug discovery, immunogenicity prediction, and therapeutic design. DeepNetBim [59] employs a hybrid architecture combining convolutional neural networks with attention mechanisms to integrate sequence features and network centrality metrics for predicting HLA–peptide binding affinity and immunogenicity. ImmugenX [39] introduces a modular PLM-based pipeline to predict immunogenic CD8+ epitopes, a task central to personalized immunotherapy. DeepHLApan [56] uses bi-directional GRUs with attention to jointly model HLA–peptide binding and immunogenicity for neoantigen discovery. UnifyImmun [60] adopts a transformer-based framework with dual encoders and cross-attention to simultaneously model HLA–peptide and peptide–TCR interactions, offering improved generalization and interpretability.

Uncertainty Quantification. Despite the success of deep learning and large language models (LLMs) across numerous domains, these models often suffer from overconfident predictions and, in the case of LLMs, hallucinations. This has motivated the development of uncertainty quantification (UQ) techniques to assess the reliability of model outputs. Subspace Inference [27] constructs Bayesian posteriors in low-dimensional subspaces of model parameters, enabling efficient inference and calibrated uncertainty estimates. Contextual Dropout [16] learns data-dependent dropout probabilities, offering both improved predictive performance and uncertainty estimation. Laplace-LoRA [58] applies Laplace approximation over low-rank adaptation parameters in a post-hoc manner, allowing for efficient posterior estimation after fine-tuning. BLoB [53] formulates a Bayesian low-rank adaptation framework by jointly estimating mean and covariance during fine-tuning. Contextual LoRA [41] further extends this by incorporating contextual uncertainty modules that dynamically adjust aleatoric uncertainty on a per-sample basis.

Existing Benchmarks. Several benchmark studies have been developed to evaluate the performance of PLMs across a wide range of biological tasks, offering insights into their generalization and transfer learning capabilities. However, few benchmarks have explicitly investigated their behavior under uncertainty or assessed their reliability in critical biomedical applications. PEER [57] provides a comprehensive multi-task evaluation framework across protein function, localization, structure, and molecular interaction tasks, comparing traditional methods and PLMs. PETA [49] evaluates 13 PLMs with varying vocabulary sizes and tokenization strategies across 15 downstream tasks, shedding light on the impact of subword vs. amino-acid-level tokenization on PLM performance. The authors of [21] benchmarked multiple UQ methods on protein fitness regression tasks under various distributional shifts, revealing key trade-offs between calibration, accuracy, and data efficiency.

I Experimental Settings

For training the models, we followed the similar techniques and hyperparameters adopted in VenusVaccine [34]. For DVBLL, LA and SVDKL, we modified the last MLP segment of the original VenusVaccine architecture by adding an extra linear layer and converted this extra linear layer as the probabilistic segment. This was aimed at promoting stable training while maintaining reasonable computational costs. The additional linear layer has dimension 64 for both LA and DVBLL models, and 16 for the SVDKL model.

For BNNs, we obtained 64 MC sample predictions. All reported results in this work, except table 8, utilize protein sequence embeddings derived from the ESM-Cambrian protein language model [15]. It should be emphasized that the ensemble model also uses sequence embeddings extracted from all 5 different PLMs including ESM-Cambrian.

J Additional OoD Experimental Results

Table 3: Results on Immuno-Virus and Immuno-Tumor datasets of models trained on Immuno-Bacteria dataset.

Dataset	Model	Accuracy(↑)	Precision(↑)	Recall(↑)	F1 Score(↑)	AUC ROC(↑)	ECE(↓)	NLL(↓)	Brier Score(↓)
Virus	Ensemble	0.4698	0.5931	0.4819	0.5317	0.5332	0.2294	0.8228	0.2969
	Deterministic	0.4811	0.6948	0.4921	0.5761	0.5802	0.4449	2.0371	0.4526
	SWAG	0.5781	0.7469	0.5637	0.6425	0.6718	0.1510	0.7136	0.2563
	DROPOUT	0.4798	0.6923	0.4912	0.5747	0.5802	0.4458	2.0299	0.4523
	DVBLL	0.5466	0.7593	0.5378	0.6296	0.6308	0.2812	1.0047	0.3309
	LA	**0.6297**	0.7717	**0.6062**	**0.6790**	**0.7236**	0.1630	0.7336	0.2491
	EDL	0.5403	**0.8685**	0.5287	0.6573	0.5305	**0.0581**	0.7345	0.2621
	TS	0.4811	0.6948	0.4921	0.5761	0.5802	0.4234	1.6522	0.4330
	SVDKL	0.5793	0.8015	0.5598	0.6592	0.6420	0.0705	**0.6860**	**0.2461**
Tumor	Ensemble	0.5513	0.0984	0.2857	0.1463	0.4751	0.1992	0.9358	0.2976
	Deterministic	0.4359	0.1967	0.2353	0.2143	0.3570	0.4844	2.8196	0.5100
	SWAG	0.5385	0.2623	0.3721	0.3077	0.4575	0.1627	0.7999	0.2849
	DROPOUT	0.4359	0.1967	0.2353	0.2143	0.3577	0.4840	2.8094	0.5096
	DVBLL	0.5000	0.2459	0.3191	0.2778	0.3902	0.3209	1.4177	0.3792
	LA	**0.6026**	0.4262	**0.4906**	**0.4561**	**0.5453**	0.1754	0.9290	0.2931
	EDL	0.4167	0.2787	0.2656	0.2720	0.4049	0.2672	0.8398	0.3002
	TS	0.4359	0.1967	0.2353	0.2143	0.3570	0.4627	2.2597	0.4914
	SVDKL	0.5128	**0.4754**	0.3973	0.4328	0.4901	**0.1000**	**0.7187**	**0.2619**

Generalization Performance of Models Trained on Immuno-Bacteria Dataset: Table 3 shows the evaluation of models on Immuno-Virus and Immuno-Tumor datasets, while they were trained on Immuno-Bacteria dataset. Even though SWAG performed best in majority of the metrics at in-distribution scenario, as reported in the second block of table 1, LA turned out to be a more consistent performer at the out-of-distribution datasets, specially according to the performance metrics. SVDKL performed best for majority of the UQ metrics in the Immuno-Virus and Immuno-Tumor datasets.

Generalization Performance of Models Trained on Immuno-Tumor Dataset: Table 4 shows the evaluation of models on Immuno-Virus and Immuno-Bacteria datasets, while they were trained on Immuno-Tumor dataset. Models trained on Immuno-Tumor dataset performed in the most random

Table 4: Results on Immuno-Virus and Immuno-Bacteria datasets of models trained on Immuno-Tumor dataset.

Dataset	Model	Accuracy(↑)	Precision(↑)	Recall(↑)	F1 Score(↑)	AUC ROC(↑)	ECE(↓)	NLL(↓)	Brier Score(↓)
Virus	Ensemble	**0.5932**	**0.3945**	0.6681	**0.4961**	**0.6576**	0.1279	0.7179	**0.2499**
	Deterministic	0.5277	0.0868	**0.8333**	0.1573	0.4927	0.2912	0.9634	0.3414
	SWAG	0.4849	0.1861	0.4808	0.2683	0.5132	0.2176	0.8532	0.3073
	DROPOUT	0.5277	0.0868	0.8333	0.1573	0.4930	0.2910	0.9627	0.3413
	DVBLL	0.4899	0.1141	0.4894	0.1851	0.5155	0.2229	0.8623	0.3096
	LA	0.4219	0.2357	0.3862	0.2928	0.4902	0.3208	0.9821	0.3435
	EDL	0.5063	0.1514	0.5495	0.2374	0.4843	0.3321	1.1036	0.3783
	TS	0.5277	0.0868	0.8333	0.1573	0.4927	0.2571	0.8827	0.3210
	SVDKL	0.4987	0.1737	0.5185	0.2602	0.5126	**0.0105**	**0.6931**	0.2500
Bacteria	Ensemble	0.4335	**0.6839**	0.3449	**0.4586**	**0.5125**	0.2584	0.8812	0.3257
	Deterministic	0.5101	0.2241	0.2653	0.2430	0.3672	0.0952	0.7543	0.2739
	SWAG	0.5081	0.2529	0.2785	0.2651	0.4183	0.1353	0.7678	0.2784
	DROPOUT	0.5000	0.2241	0.2566	0.2393	0.3669	0.1053	0.7542	0.2739
	DVBLL	**0.5867**	0.1149	0.2817	0.1633	0.4611	0.0982	0.7382	0.2607
	LA	0.5625	0.1264	0.2529	0.1686	0.4168	0.1704	0.8447	0.2923
	EDL	0.4637	0.2126	0.2229	0.2176	0.3624	0.3066	0.9684	0.3504
	TS	0.5101	0.2241	0.2653	0.2430	0.3672	**0.0888**	0.7370	0.2682
	SVDKL	0.5403	0.3736	**0.3533**	0.3631	0.4765	0.1484	**0.6929**	**0.2499**

manner among the three. However, the SVDKL model performed better than all other counterparts according to UQ metrics and ensemble performed better than others according to performance metrics, even though it lagged behind others at in-distribution scenario as reported in the third block of Table 1.

K Additional Experimental Results

Table 5, 6 and 7 demonstrates the results on all three immunogenic datasets for models with protein sequence embeddings extracted from PLMs except ESM-Cambrian.

Table 5: Results on Immuno-Virus dataset.

PLM	Model	Accuracy(↑)	Precision(↑)	Recall(↑)	F1 Score(↑)	AUC ROC(↑)	ECE(↓)	NLL(↓)	Brier Score(↓)
-	Ensemble	0.9345	0.9479	0.9249	0.9363	0.9809	0.0238	0.1850	0.0521
ProstT5	Deterministic	0.9081	0.9330	0.8910	0.9115	0.9649	0.0657	0.3949	0.0788
	SWAG	**0.9156**	**0.9429**	**0.8962**	**0.9190**	**0.9675**	**0.0255**	**0.2317**	**0.0665**
	DROPOUT	0.9081	0.9330	0.8910	0.9115	0.9649	0.0655	0.3934	0.0787
	DVBLL	0.8967	0.8834	0.9105	0.8967	0.9572	0.0628	0.3946	0.0834
	LA	0.8841	0.8784	0.8917	0.8850	0.9571	0.0666	0.3416	0.0869
	EDL	0.8778	0.9206	0.8509	0.8844	0.9570	0.0477	0.3013	0.0907
	TS	0.9081	0.9330	0.8910	0.9115	0.9649	0.0588	0.3377	0.0765
	SVDKL	0.9055	0.9330	0.8868	0.9093	0.9456	0.0344	0.2757	0.0760
Ankh	Deterministic	0.9131	0.9380	0.8957	0.9164	0.9582	0.0656	0.4375	0.0730
	SWAG	**0.9232**	0.9305	0.9191	**0.9248**	**0.9714**	**0.0143**	**0.2172**	**0.0608**
	DROPOUT	0.9131	0.9380	0.8957	0.9164	0.9583	0.0654	0.4349	0.0730
	DVBLL	0.9194	0.9280	0.9144	0.9212	0.9625	0.0550	0.4004	0.0698
	LA	0.9055	0.8983	0.9141	0.9061	0.9580	0.0756	0.4801	0.0825
	EDL	0.8904	0.9032	0.8835	0.8933	0.9541	0.0409	0.3017	0.0877
	TS	0.9131	**0.9380**	0.8957	0.9164	0.9582	0.0592	0.3644	0.0711
	SVDKL	0.9194	0.9256	**0.9165**	0.9210	0.9567	0.1424	0.3282	0.0862
ESM2	Deterministic	0.8715	0.8685	0.8772	0.8728	0.9556	**0.0215**	0.2772	0.0861
	SWAG	0.9043	0.9231	0.8921	0.9073	**0.9683**	0.0528	0.2505	0.0725
	DROPOUT	0.8715	0.8685	0.8772	0.8728	0.9555	0.0216	0.2772	0.0861
	DVBLL	0.8929	0.8933	0.8955	0.8944	0.9629	0.0891	0.5418	0.0935
	LA	0.9194	**0.9479**	0.8988	0.9227	0.9662	0.0536	0.3829	0.0716
	EDL	0.8917	0.9032	0.8856	0.8943	0.9598	0.0367	0.2796	0.0822
	TS	0.8715	0.8685	0.8772	0.8728	0.9556	0.0223	0.2770	0.0860
	SVDKL	**0.9244**	0.9380	**0.9153**	**0.9265**	0.9660	0.0318	**0.2443**	**0.0645**
Prot Bert	Deterministic	0.9106	0.8908	0.9301	0.9100	0.9669	0.0362	0.2911	0.0712
	SWAG	**0.9244**	**0.9305**	0.9214	**0.9259**	**0.9716**	0.0412	**0.2207**	0.0623
	DROPOUT	0.9106	0.8908	**0.9301**	0.9100	0.9670	0.0361	0.2900	0.0711
	DVBLL	0.8992	0.8710	0.9261	0.8977	0.9643	**0.0182**	0.2465	0.0735
	LA	0.8929	0.9132	0.8804	0.8965	0.9525	0.0708	0.2999	0.0890
	EDL	0.9068	0.9206	0.8983	0.9093	0.9650	0.0260	0.2480	0.0714
	TS	0.9106	0.8908	0.9301	0.9100	0.9669	0.0266	0.2658	**0.0701**
	SVDKL	0.9005	0.8859	0.9154	0.9004	0.9429	0.0246	0.2916	0.0812

L PLM Embedding Comparison

Table 8 shows the comparative performance among models with different PLMs for extracting amino acid sequence embeddings.

Table 6: Results on Immuno-Bacteria dataset.

PLM	Model	Accuracy(↑)	Precision(↑)	Recall(↑)	F1 Score(↑)	AUC ROC(↑)	ECE(↓)	NLL(↓)	Brier Score(↓)
-	Ensemble	**0.8327**	**0.6897**	**0.8054**	**0.7430**	**0.8883**	**0.0685**	**0.4788**	**0.1302**
ProstT5	Deterministic	0.8024	0.5747	**0.8065**	0.6711	0.8622	0.1078	0.5953	0.1603
	SWAG	0.8085	**0.6954**	0.7423	0.7181	0.8590	0.1344	0.6300	0.1611
	DROPOUT	0.8024	0.5747	**0.8065**	0.6711	0.8623	0.1078	0.5948	0.1603
	DVBLL	**0.8165**	0.6782	0.7712	**0.7217**	0.8586	0.1463	1.0648	0.1703
	LA	0.7782	0.6667	0.6905	0.6784	0.8485	0.1516	0.7059	0.1753
	EDL	0.8085	0.6092	0.7970	0.6906	**0.8644**	**0.0991**	**0.4966**	**0.1549**
	TS	0.8024	0.5747	**0.8065**	0.6711	0.8622	0.0935	0.5348	0.1558
	SVDKL	0.8044	0.6782	0.7421	0.7087	0.8316	0.1165	0.5048	0.1608
Ankh	Deterministic	0.8306	0.6954	**0.7961**	0.7423	0.8616	0.1134	0.6872	0.1485
	SWAG	0.8246	0.7241	0.7636	0.7434	0.8689	0.0587	0.4420	0.1303
	DROPOUT	0.8306	0.6954	**0.7961**	0.7423	0.8617	0.1132	0.6855	0.1485
	DVBLL	0.8226	**0.7701**	0.7363	**0.7528**	0.8662	0.1520	1.1560	0.1619
	LA	0.8125	0.7586	0.7213	0.7395	0.8552	0.1313	0.7498	0.1571
	EDL	0.8105	0.7241	0.7326	0.7283	0.8451	**0.0537**	**0.4924**	0.1546
	TS	**0.8306**	0.6954	**0.7961**	0.7423	0.8616	0.0974	0.5863	**0.1449**
	SVDKL	0.8286	0.7414	0.7633	0.7522	0.8530	0.1792	0.5236	0.1679
ESM2	Deterministic	0.8085	0.6379	0.7762	0.7003	0.8749	0.1073	0.6592	0.1513
	SWAG	0.7984	0.6839	0.7256	0.7041	0.8684	**0.0269**	0.4299	0.1363
	DROPOUT	0.8085	0.6379	0.7762	0.7003	**0.8750**	0.1071	0.6581	0.1512
	DVBLL	0.8145	0.7241	0.7412	0.7326	0.8720	0.1353	0.9366	0.1558
	LA	**0.8306**	0.7759	0.7500	**0.7627**	0.8748	0.1182	0.8484	0.1487
	EDL	0.8266	**0.7874**	0.7366	0.7611	0.8777	0.0337	**0.4224**	**0.1303**
	TS	0.8085	0.6379	0.7762	0.7003	0.8749	0.0904	0.5730	0.1471
	SVDKL	0.8085	0.7069	0.7365	0.7214	0.8579	0.2716	0.6446	0.2258
Prot Bert	Deterministic	0.8206	0.6897	0.7742	0.7295	0.8775	0.1200	0.6543	0.1484
	SWAG	0.8266	0.7241	0.7683	0.7456	0.8708	**0.0041**	**0.4196**	0.1316
	DROPOUT	0.8206	0.6897	0.7742	0.7295	0.8774	0.1199	0.6527	0.1483
	DVBLL	0.8367	0.7586	0.7719	**0.7652**	0.8736	0.1297	1.2571	**0.1477**
	LA	0.8185	0.6954	0.7658	0.7289	0.8750	0.0453	0.4279	**0.1302**
	EDL	0.8145	**0.8103**	0.7050	0.7540	0.8717	0.0717	0.4517	0.1399
	TS	0.8206	0.6897	0.7742	0.7295	**0.8775**	0.1037	0.5634	0.1431
	SVDKL	**0.8407**	0.7529	**0.7844**	**0.7683**	0.8525	0.2675	0.5999	0.2038

Table 7: Results on Immuno-Tumor dataset.

PLM	Model	Accuracy(↑)	Precision(↑)	Recall(↑)	F1 Score(↑)	AUC ROC(↑)	ECE(↓)	NLL(↓)	Brier Score(↓)
-	Ensemble	0.7500	0.7869	0.6486	0.7111	0.8483	0.0413	0.4701	0.1599
ProstT5	Deterministic	0.7115	0.7213	0.6111	0.6617	0.8098	0.2400	1.2505	0.2487
	SWAG	0.7179	0.7049	0.6232	0.6615	0.7938	0.1159	0.5865	0.1915
	DROPOUT	0.7115	0.7213	0.6111	0.6617	0.8098	0.2399	1.2477	0.2487
	DVBLL	0.7244	0.7705	0.6184	0.6861	0.7727	0.1218	0.6729	0.2046
	LA	**0.7372**	**0.8033**	**0.6282**	**0.7050**	0.8014	0.2350	1.6140	0.2518
	EDL	0.7244	0.7541	0.6216	0.6815	0.7883	**0.1129**	**0.5852**	**0.1967**
	TS	0.7115	0.7213	0.6111	0.6617	**0.8098**	0.2251	1.0160	0.2389
	SVDKL	0.4936	0.6230	0.4043	0.4903	0.4989	0.1147	0.6962	0.2515
Ankh	Deterministic	0.7692	**0.8852**	0.6506	0.7500	**0.8607**	0.0725	0.4661	0.1562
	SWAG	0.7500	0.6721	0.6833	0.6777	0.8507	0.0483	0.4784	0.1583
	DROPOUT	**0.7756**	**0.8852**	**0.6585**	**0.7552**	0.8595	0.0723	0.4660	0.1562
	DVBLL	0.7692	0.7049	0.7049	0.7049	0.8490	**0.0094**	**0.4580**	**0.1550**
	LA	0.7051	0.8033	0.5904	0.6806	0.7520	0.0794	0.5761	0.1993
	EDL	0.6090	0.0000	0.0000	0.0000	0.6626	0.1580	0.7024	0.2527
	TS	0.7692	**0.8852**	0.6506	0.7500	**0.8607**	0.0716	0.4652	0.1559
	SVDKL	0.6410	0.0984	0.8571	0.1765	0.7165	0.1386	0.6911	0.2490
ESM2	Deterministic	0.7436	0.6721	0.6721	0.6721	0.8149	0.2236	2.1182	0.2280
	SWAG	0.7628	0.7213	0.6875	0.7040	0.8269	0.1276	0.5988	0.1890
	DROPOUT	0.7436	0.6721	0.6721	0.6721	0.8154	0.2232	1.5745	0.2278
	DVBLL	**0.7756**	**0.8033**	**0.6806**	**0.7368**	0.8110	0.1735	1.3219	0.2139
	LA	0.7436	0.7705	0.6438	0.7015	**0.8302**	**0.1110**	**0.5528**	**0.1784**
	EDL	0.7372	0.7541	0.6389	0.6917	0.8207	0.1630	0.6568	0.2086
	TS	0.7436	0.6721	0.6721	0.6721	0.8150	0.2121	1.2658	0.2219
	SVDKL	0.5513	0.3934	0.4211	0.4068	0.5032	0.1058	0.6926	0.2497
Prot Bert	Deterministic	0.7179	0.7869	0.6076	0.6857	0.7991	0.0741	0.5529	0.1868
	SWAG	0.7436	0.7705	0.6438	0.7015	**0.8409**	0.0491	0.4825	0.1609
	DROPOUT	0.7179	0.7869	0.6076	0.6857	0.7988	0.0739	0.5521	0.1866
	DVBLL	**0.7885**	0.7541	**0.7188**	**0.7360**	0.8311	**0.0310**	**0.5197**	0.1705
	LA	0.7564	**0.8197**	0.6494	0.7246	0.8364	0.0745	0.5083	0.1713
	EDL	0.7500	0.8033	0.6447	0.7153	0.8312	0.0944	0.5361	0.1779
	TS	0.7179	0.7869	0.6076	0.6857	0.7991	0.0551	0.5417	0.1840
	SVDKL	0.5385	**0.9344**	0.4560	0.6129	0.7027	0.1150	0.6919	0.2494

Table 8: Uncertainty quantification performance across datasets and models employing various PLMs. Best-performing models per PLM are shown in bold.

Metric	Model	Dataset														
		Virus					Bacteria					Tumor				
		ESMC	ProstT5	Ankh	ESM2	Prot Bert	ESMC	ProstT5	Ankh	ESM2	Prot Bert	ESMC	ProstT5	Ankh	ESM2	Prot Bert
ECE(↓)	Deterministic	0.0413	0.0657	0.0656	**0.0215**	0.0362	0.1332	0.1078	0.1134	0.1073	0.1200	0.0298	0.2400	0.0725	0.2236	0.0741
	SVDKL	0.1058	0.0344	0.1424	0.0318	0.0246	0.1513	0.1165	0.1792	0.2716	0.2675	0.1132	0.1147	0.1386	**0.1058**	0.1150
	DVBLL	0.0346	0.0628	0.0550	0.0891	**0.0182**	0.0824	0.1463	0.1520	0.1353	0.1297	0.0669	0.1218	**0.0094**	0.1735	**0.0310**
	DROPOUT	0.0411	0.0655	0.0654	0.0216	0.0361	0.1349	0.1078	0.1132	0.1071	0.1199	0.0360	0.2399	0.0723	0.2232	0.0739
	LA	**0.0108**	0.0666	0.0756	0.0536	0.0708	0.0752	0.1516	0.1313	0.1182	0.0453	0.0724	0.2350	0.0794	0.1110	0.0745
	SWAG	0.0743	**0.0255**	**0.0143**	0.0528	0.0412	**0.0133**	0.1344	0.0587	**0.0269**	**0.0041**	**0.0189**	0.1159	0.0483	0.1276	0.0491
	EDL	0.0152	0.0477	0.0409	0.0367	0.0260	0.1259	0.0991	**0.0537**	0.0337	0.0717	0.0700	**0.1129**	0.1580	0.1630	0.0944
	TS	0.0311	0.0588	0.0592	0.0223	0.0266	0.1192	**0.0935**	0.0974	0.0904	0.1037	0.0298	0.2251	0.0716	0.2121	0.0551
NLL(↓)	Deterministic	0.2808	0.3949	0.4375	0.2772	0.2911	0.7672	0.5953	0.6872	0.6592	0.6543	0.4908	1.2505	0.4661	2.1182	0.5529
	SVDKL	0.3050	0.2757	0.3282	**0.2443**	0.2916	0.4991	0.5048	0.5236	0.6446	0.5999	0.6920	0.6962	0.6911	0.6926	0.6919
	DVBLL	0.2882	0.3946	0.4004	0.5418	0.2465	0.5112	1.0648	1.1560	0.9366	1.2571	0.4743	0.6729	**0.4580**	1.3219	0.5197
	DROPOUT	0.2799	0.3934	0.4349	0.2772	0.2900	0.7654	0.5948	0.6855	0.6581	0.6527	0.4908	1.2477	0.4660	1.5745	0.5521
	LA	0.2498	0.3416	0.4801	0.3829	0.2999	0.5251	0.7059	0.7498	0.8484	0.4279	0.4975	1.6140	0.5761	**0.5528**	0.5083
	SWAG	**0.2422**	**0.2317**	**0.2172**	0.2505	**0.2207**	**0.4046**	0.6300	**0.4420**	0.4299	**0.4196**	**0.4610**	0.5865	0.4784	0.5988	**0.4825**
	EDL	0.2537	0.3013	0.3017	0.2796	0.2480	0.5237	**0.4966**	0.4924	**0.4224**	0.4517	0.4867	**0.5852**	0.7024	0.6568	0.5361
	TS	0.2604	0.3377	0.3644	0.2770	0.2658	0.6425	0.5348	0.5863	0.5730	0.5634	0.4908	1.0160	0.4652	1.2658	0.5417
Brier Score(↓)	Deterministic	0.0711	0.0788	0.0730	0.0861	0.0712	0.1526	0.1603	0.1485	0.1513	0.1484	0.1659	0.2487	0.1562	0.2280	0.1868
	SVDKL	0.0801	0.0760	0.0862	**0.0645**	0.0812	0.1577	0.1608	0.1679	0.2258	0.2038	0.2494	0.2515	0.2490	0.2497	0.2494
	DVBLL	0.0729	0.0834	0.0698	0.0935	0.0735	0.1375	0.1703	0.1619	0.1558	0.1477	0.1574	0.2046	**0.1550**	0.2139	0.1705
	DROPOUT	0.0710	0.0787	0.0730	0.0861	0.0711	0.1525	0.1603	0.1485	0.1512	0.1483	0.1659	0.2487	0.1562	0.2278	0.1866
	LA	0.0750	0.0869	0.0825	0.0716	0.0890	0.1386	0.1753	0.1571	0.1487	**0.1302**	0.1639	0.2518	0.1993	**0.1784**	0.1713
	SWAG	**0.0667**	**0.0665**	**0.0608**	0.0725	**0.0623**	**0.1251**	0.1611	**0.1303**	0.1363	0.1316	**0.1527**	**0.1915**	0.1583	0.1890	**0.1609**
	EDL	0.0743	0.0907	0.0877	0.0822	0.0714	0.1739	**0.1549**	0.1546	**0.1303**	0.1399	0.1600	0.1967	0.2527	0.2086	0.1779
	TS	0.0699	0.0765	0.0711	0.0860	0.0701	0.1475	0.1558	0.1449	0.1471	0.1431	0.1659	0.2389	0.1559	0.2219	0.1840

Across all three immunogenic datasets, and for nearly every performance metric considered, deterministic models consistently demonstrated inferior performance compared to other models. The only exception to this trend was observed on the Immuno-Virus dataset when protein sequences were embedded using the ESM-2 PLM, specifically concerning the ECE metric. Among the various models investigated, SWAG consistently emerged as a strong performer. It outperformed the majority of other models across all three immunogenic datasets. While SWAG demonstrated overall strong performance, EDL also exhibited competitive results, particularly on the Immuno-Bacteria dataset, suggesting its potential for accurate immunogenicity prediction in this specific domain. To summarize, advantages over uncertainty-aware models over deterministic models persist irrespective of the protein language model used, underscoring compatibility with advances in protein representation learning.

M Ablation Studies

We have further conducted ablation studies on how protein structural information may affect immunogenicity prediction and UQ performances by comparing the results from the VenusVaccine model architecture with and without the derived protein structural features. Table 9 shows the ablation results for all models for the Immuno-Virus dataset. The table supports three principal conclusions.

Table 9: Experimental results for evaluation on effect of protein structural information (**PSI** = **P**rotein **S**tructural **I**nformation) on immunogenic virus dataset. For each method the superior performance between the two settings is shown in blue.

Model	PSI	Accuracy(↑)	Precision(↑)	Recall(↑)	F1 Score(↑)	AUC ROC(↑)	ECE(↓)	NLL(↓)	Brier Score(↓)
Ensemble	✓	0.9345	0.9479	0.9249	0.9363	0.9809	0.0238	0.1850	0.0521
	×	0.9181	0.9380	0.9043	0.9208	0.9767	0.0146	0.1994	0.0593
Deterministic	✓	0.9055	0.9082	0.9059	0.9071	0.9647	0.0413	0.2808	0.0711
	×	0.9030	0.9479	0.8721	0.9084	0.9721	0.0463	0.2542	0.0713
SWAG	✓	0.9219	0.9380	0.9108	0.9242	0.9728	0.0743	0.2422	0.0667
	×	0.9005	0.8933	0.9091	0.9011	0.9682	0.0569	0.2586	0.0737
DROPOUT	✓	0.9055	0.9082	0.9059	0.9071	0.9647	0.0411	0.2799	0.0710
	×	0.9018	0.9454	0.8719	0.9071	0.9720	0.0468	0.2533	0.0714
DVBLL	✓	0.9030	0.9107	0.8995	0.9051	0.9621	0.0346	0.2882	0.0729
	×	0.9118	0.9181	0.9091	0.9136	0.9739	0.0482	0.2551	0.0655
LA	✓	0.8904	0.9256	0.8674	0.8956	0.9633	0.0108	0.2498	0.0750
	×	0.9055	0.9256	0.8923	0.9086	0.9668	0.0626	0.3317	0.0745
EDL	✓	0.9005	0.9305	0.8803	0.9047	0.9643	0.0152	0.2537	0.0743
	×	0.9018	0.9429	0.8736	0.9069	0.9709	0.0404	0.2441	0.0683
TS	✓	0.9055	0.9082	0.9059	0.9071	0.9647	0.0311	0.2604	0.0699
	×	0.9030	0.9479	0.8721	0.9084	0.9721	0.0372	0.2381	0.0698
SVDKL	✓	0.9156	0.9305	0.9058	0.9180	0.9577	0.1058	0.3050	0.0801
	×	0.9081	0.9156	0.9044	0.9100	0.9572	0.3113	0.5392	0.1743

First, structural information offers predictive advantages in methods such as Ensembles, SWAG, and SVDKL, but fails to demonstrate consistent utility across all approaches. For instance, its effect is minimal in Deterministic, Dropout, and EDL, and is detrimental in LA and DVBLL. Thus, its overall contribution to predictive performance appears method-specific and inconclusive. **Second**, the inclusion of structural information generally enhances uncertainty quantification, as evidenced by improved performance across uncertainty metrics in Table 9. **Third**, uncertainty-aware approaches consistently outperform the deterministic baseline in nearly all cases, with the only exception being the precision metric in the absence of structural information, where the deterministic model exhibits a slight advantage. These observations highlight the robustness of uncertainty-aware models and their superiority in capturing both predictive accuracy and well-calibrated predictions.

N Limitations & Future Work

Our proposed **ImmUQBench** provides a targeted evaluation of selected Bayesian and non-Bayesian UQ methods. However, it does not explore the broader design space, including variations in backbone architectures or uncertainty propagation beyond the prediction head. Also, the current benchmark only addresses epistemic uncertainty; the consideration of uncertainty within protein representations and its effect on immunogenicity prediction still remains an open research endeavor.