

# ImmUQBench: A Benchmark on Uncertainty Quantification of Protein Immunogenicity Prediction

Anonymous Authors

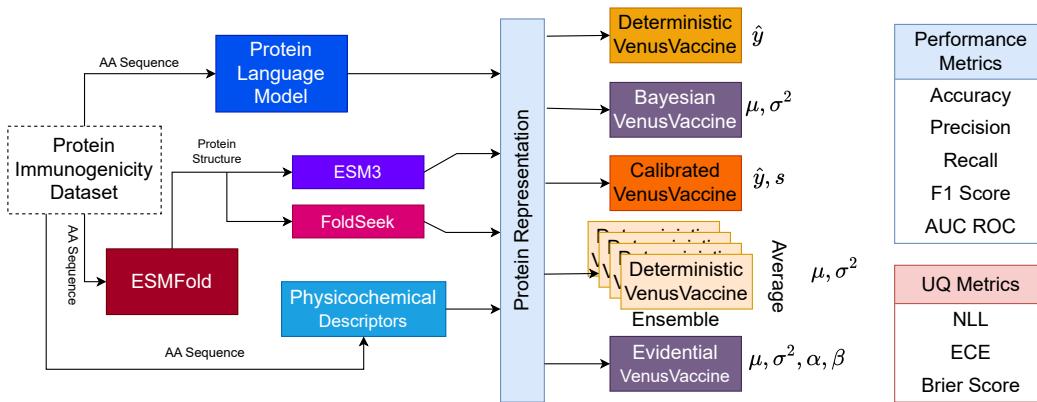


Figure 1: Benchmarking Uncertainty Quantification (UQ) in protein immunogenicity prediction

## 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.

## CCS CONCEPTS

• Applied computing → Bioinformatics.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## KEYWORDS

Uncertainty quantification, immunogenicity prediction.

## ACM Reference Format:

Anonymous Authors. 2025. ImmUQBench: A Benchmark on Uncertainty Quantification of Protein Immunogenicity Prediction. In . ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 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 [1, 12]. 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.

The advent of Large Language Models (LLMs) marks a pivotal advancement in natural language processing (NLP), fundamentally reshaping its capabilities [6, 10, 50]. 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 [15, 17, 57]. 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

117 tasks often exhibit overconfident predictions and are prone to generating  
 118 hallucinated outputs [19, 48], raising concerns about their reliability  
 119 and trustworthiness in sensitive applications, e.g. safety and  
 120 efficacy related therapeutic design. To mitigate these limitations,  
 121 the machine learning community has increasingly turned to un-  
 122 certainty quantification (UQ) techniques. Broadly, UQ methods fall  
 123 into two categories: Bayesian approaches [8, 11, 18], which provide  
 124 a principled probabilistic framework but can be computationally  
 125 intensive or impractical, and non-Bayesian approaches [2, 9, 28],  
 126 which are often more tractable and performant but lack strong  
 127 theoretical guarantees.

128 In this work, we introduce **ImmUQBench**, a benchmark de-  
 129 signed to systematically evaluate a diverse set of UQ methods in  
 130 the context of immunogenicity prediction. Specifically, we compare  
 131 both Bayesian and non-Bayesian methods across various experi-  
 132 mental scenarios, including in-distribution and out-of-distribution  
 133 settings, to assess their predictive accuracy, uncertainty estimation  
 134 quality, and robustness to distributional shifts. Additionally, we  
 135 conduct an ablation study to evaluate the performance of these  
 136 methods when part of the input information is missing, comparing  
 137 them against their deterministic counterparts. We also examine  
 138 the impact of different protein sequence encoding schemes, high-  
 139 lighting robustness to alternate encodings that express the same  
 140 underlying sequence information.

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

- 146 • *Pioneering Benchmark in Immunogenicity*: We introduce **Im-**  
   **mUQBench**, a benchmark for uncertainty quantification in  
   immunogenicity prediction.
- 147 • *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 sce-  
   narios.
- 148 • *Evaluation of Various Data Representation and Model Ablation*:  
   We provide insights through extensive experiments and abla-  
   tion studies to support antigen design and broadly effective  
   therapeutic development.

149 The remainder of this paper is organized as follows: Section 2  
 reviews existing UQ approaches for deep neural network models;  
 150 Section 3 presents the setup of ImmUQBench; Section 4 describes  
 151 the experimental settings and results; Section 5 reviews related  
 152 works; and Section 6 concludes the study.

## 153 2 UQ FOR DEEP NEURAL NETWORKS

154 As Deep Learning models have become widely adopted across  
 155 a broad range of tasks, their trustworthiness and reliability of  
 156 their prediction have become essential and critically important,  
 157 as they struggle to distinguish between in and out-of-distribution  
 158 datasets [22, 55] as well as being sensitive to domain shift [40].  
 159 This is particularly important in safety critical tasks where data is  
 160 often scarce, wrong predictions can lead to severe consequences.  
 161 Hence, it is essential that these models express their uncertainty

162 when confronting out-of-distribution data. Different approaches  
 163 have been developed and utilized for uncertainty quantification  
 164 that can be broadly categorized into Bayesian methods [8, 11, 18],  
 165 [20, 29, 32, 36, 43, 45] and non-Bayesian methods [2–4, 9, 28, 30, 61].

166 In this work, for each category of UQ methods, we consider  
 167 widely used and representative approaches.

### 168 Bayesian Methods:

169 We begin with **Monte Carlo (MC)-Dropout** [18], which in-  
 170 terprets dropout as a form of approximate Bayesian inference in  
 171 deep Gaussian Processes. This interpretation allows dropout to cap-  
 172 ture epistemic uncertainty by maintaining stochasticity at test time.  
 173 Specifically, model predictions are obtained by performing multiple  
 174 stochastic forward passes with dropout enabled and computing the  
 175 predictive mean and variance from these passes. Formally, the pre-  
 176 dictive mean is approximated by averaging outputs over  $T$  stochas-  
 177 tic passes, while the variance is estimated as the sample variance  
 178 plus a model precision term  $\tau^{-1}$ , where  $\tau$  depends on the dropout  
 179 rate, dataset size, and weight decay. This enables tractable Bayesian  
 180 approximation without incurring additional test-time complexity.

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

188 We also include **Stochastic Variational Deep Kernel Learn-  
 189 ing (SVDKL)** [54], which synergizes the representation power of  
 190 deep neural networks with the non-parametric flexibility of Gauss-  
 191 ian Processes. SVDKL extends Deep Kernel Learning (DKL) to clas-  
 192 sification and multi-task learning using a scalable variational infer-  
 193 ence framework. This allows for training on large-scale datasets  
 194 via stochastic gradients, and supports more expressive covariance  
 195 structures compared to prior DKL models.

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

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

### 216 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.

Finally, as all the aforementioned methods ultimately aim to produce well-calibrated and reliable predictive distributions, we also include **Temperature Scaling (TS)** [22] as a baseline for comparison. Temperature scaling is a post-hoc calibration technique that adjusts the confidence of a classifier by optimizing a single non-negative scalar parameter  $T > 0$  on a validation set, typically using negative log-likelihood as the objective. By dividing the logits by  $T$  before applying the softmax, this method effectively increases the entropy of the predictive distribution, leading to better-calibrated output probabilities without altering the model's accuracy.

### 3 IMMUQBENCH

In this work, 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 a protective antigen.

#### 3.1 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 protective antigens that are capable of eliciting immune responses resulting in the formation of memory cells within the host organism [1, 34].

Researchers are increasingly focused on fast and precise prediction of protective 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 calculate 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 physicochemical 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.

#### 3.2 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, 25, 44]. This adaptation—hence modeling of protein sequences—was enabled by equating words with amino acids and interpreting the entire protein sequences as sentences [42, 52]. 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 [25, 44], 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.

#### 3.3 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  $f_{\theta}$  is a deterministic classifier parameterized by  $\theta$ .

**Bayesian Methods:** In this work, for evaluating Bayesian methods, we treat  $\theta$  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], DVBL [23], SVDKL [54] 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,  $E_{\text{seq}}$ , for different model in the ensemble comes from different PLMs. The 5 PLMs used in this work are: ESM-Cambrian [15], ProstT5 [26], Ankh [13], ESM-2 [35] and Prot-Bert [14].

**EDL:** In a binary classification, EDL models the class probability as a Beta distribution,  $\text{Beta}(\alpha_1, \alpha_2)$ . 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  $\sigma$  denotes the softmax function. TS adjusts confidence levels without affecting the model's accuracy, making it a simple yet effective post-hoc calibration method.

## 4 EXPERIMENTS

### 4.1 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 [22, 38],

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

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 [5, 38]. 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,

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

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.

### 4.2 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.

### 4.3 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

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

Dataset	Model	Accuracy( $\uparrow$ )	Precision( $\uparrow$ )	Recall( $\uparrow$ )	F1 Score( $\uparrow$ )	AUC ROC( $\uparrow$ )	ECE( $\downarrow$ )	NLL( $\downarrow$ )	Brier Score( $\downarrow$ )
Virus	Ensemble	<b>0.9345</b>	<b>0.9479</b>	<b>0.9249</b>	<b>0.9363</b>	<b>0.9809</b>	0.0238	<b>0.1850</b>	<b>0.0521</b>
	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	<b>0.0108</b>	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
	Ensemble	<b>0.8327</b>	0.6897	0.8054	0.7430	<b>0.8883</b>	0.0685	0.4788	0.1302
Bacteria	Deterministic	0.8145	0.6897	0.7595	0.7229	0.8702	0.1332	0.7672	0.1526
	SWAG	<b>0.8327</b>	<b>0.7414</b>	0.7725	<b>0.7566</b>	0.8780	<b>0.0133</b>	<b>0.4046</b>	<b>0.1251</b>
	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	<b>0.8175</b>	0.7203	0.8689	0.1513	0.4991	0.1577
	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
Tumor	SWAG	<b>0.7692</b>	0.7541	0.6866	0.7188	0.8537	<b>0.0189</b>	<b>0.4610</b>	<b>0.1527</b>
	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	<b>0.8585</b>	0.0669	0.4743	0.1574
	LA	<b>0.7692</b>	0.5574	<b>0.7907</b>	0.6538	0.8564	0.0724	0.4975	0.1639
	EDL	<b>0.7692</b>	<b>0.8033</b>	0.6712	<b>0.7313</b>	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

compresses amino acid-level features into a protein-level vector by highlighting key regions, which is used for final binary classification of immunogenicity.

#### 4.4 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 5, 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**.

#### 4.5 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.

#### 4.6 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).

**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

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

Dataset	Model	Accuracy( $\uparrow$ )	Precision( $\uparrow$ )	Recall( $\uparrow$ )	F1 Score( $\uparrow$ )	AUC ROC( $\uparrow$ )	ECE( $\downarrow$ )	NLL( $\downarrow$ )	Brier Score( $\downarrow$ )
Bacteria	Ensemble	<b>0.7117</b>	<b>0.1954</b>	<b>0.9189</b>	<b>0.3223</b>	0.7026	0.1429	0.6879	<b>0.2110</b>
	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	<b>0.7265</b>	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	<b>0.0916</b>	<b>0.6395</b>	0.2204
	Ensemble	0.5577	0.3770	0.4259	0.4000	0.5346	0.1072	0.7398	0.2632
Tumor	Deterministic	0.5449	0.5246	0.4324	0.4741	0.5381	0.2894	1.3243	0.3525
	SWAG	0.6538	<b>0.6393</b>	0.5493	<b>0.5909</b>	<b>0.6481</b>	<b>0.0789</b>	<b>0.6637</b>	<b>0.2345</b>
	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	<b>0.6667</b>	0.5410	<b>0.5789</b>	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
	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
Virus	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	0.5466	0.7593	0.5378	0.6296	0.6308	0.2812	1.0047	0.3309
	LA	<b>0.6297</b>	0.7717	<b>0.6062</b>	<b>0.6790</b>	<b>0.7236</b>	0.1630	0.7336	0.2491
	EDL	0.5403	<b>0.8685</b>	0.5287	0.6573	0.5305	<b>0.0581</b>	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	<b>0.6860</b>	<b>0.2461</b>
	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
Bacteria	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	<b>0.6026</b>	0.4262	<b>0.4906</b>	<b>0.4561</b>	<b>0.5453</b>	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	<b>0.4754</b>	0.3973	0.4328	0.4901	<b>0.1000</b>	<b>0.7187</b>	<b>0.2619</b>

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

Dataset	Model	Accuracy( $\uparrow$ )	Precision( $\uparrow$ )	Recall( $\uparrow$ )	F1 Score( $\uparrow$ )	AUC ROC( $\uparrow$ )	ECE( $\downarrow$ )	NLL( $\downarrow$ )	Brier Score( $\downarrow$ )
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	<b>0.6297</b>	0.7717	<b>0.6062</b>	<b>0.6790</b>	<b>0.7236</b>	0.1630	0.7336	0.2491
	EDL	0.5403	<b>0.8685</b>	0.5287	0.6573	0.5305	<b>0.0581</b>	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	<b>0.6860</b>	<b>0.2461</b>
	Ensemble	0.5513	0.0984	0.2857	0.1463	0.4751	0.1992	0.9358	0.2976
Tumor	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	<b>0.6026</b>	0.4262	<b>0.4906</b>	<b>0.4561</b>	<b>0.5453</b>	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	<b>0.4754</b>	0.3973	0.4328	0.4901	<b>0.1000</b>	<b>0.7187</b>	<b>0.2619</b>

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

Dataset	Model	Accuracy( $\uparrow$ )	Precision( $\uparrow$ )	Recall( $\uparrow$ )	F1 Score( $\uparrow$ )	AUC ROC( $\uparrow$ )	ECE( $\downarrow$ )	NLL( $\downarrow$ )	Brier Score( $\downarrow$ )
Virus	Ensemble	<b>0.5932</b>	<b>0.3945</b>	0.6681	<b>0.4961</b>	<b>0.6576</b>	0.1279	0.7179	<b>0.2499</b>
	Deterministic	0.5277	0.0868	<b>0.8333</b>	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	<b>0.0105</b>	<b>0.6931</b>	0.2500
	Ensemble	0.4335	<b>0.6839</b>	0.3449	<b>0.4586</b>	<b>0.5125</b>	0.2584	0.8812	0.3257
Bacteria	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	<b>0.5867</b>	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	<b>0.0888</b>	0.7370	0.2682
	SVDKL	0.5403	0.3736	<b>0.3533</b>	0.3631	0.4765	0.1484	<b>0.6929</b>	<b>0.2499</b>

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

704 **Generalization Performance of Models Trained on Immuno-**  
**Bacteria Dataset:** Table 3 shows the evaluation of models on  
 705 Immuno-Virus and Immuno-Tumor datasets, while they were trained  
 706 on Immuno-Bacteria dataset. Even though SWAG performed best  
 707 in majority of the metrics at in-distribution scenario, as reported in  
 708 the second block of table 1, LA turned out to be a more consistent  
 709 performer at the out-of-distribution datasets, specially according  
 710 to the performance metrics. SVDKL performed best for majority of  
 711 the UQ metrics in the Immuno-Virus and Immuno-Tumor datasets.  
 712

713 **Generalization Performance of Models Trained on Immuno-**  
**Tumor Dataset:** Table 4 shows the evaluation of models on Immuno-  
 714 Virus and Immuno-Bacteria datasets, while they were trained on  
 715 Immuno-Tumor dataset. Models trained on Immuno-Tumor dataset  
 716 performed in the most random manner among the three. However,  
 717 the SVDKL model performed better than all other counterparts ac-  
 718 cording to UQ metrics and ensemble performed better than others  
 719 according to performance metrics, even though it lagged behind  
 720 others at in-distribution scenario as reported in the third block of  
 721 Table 1.

722 **Summary on OoD Performance:** In summary, while SVDKL  
 723 and SWAG showed greater consistency in uncertainty quantifi-  
 724 cation, no method clearly outperformed the others under OOD  
 725 conditions. Also, as expected, the deterministic model was outper-  
 726 formed by uncertainty-aware methods across almost all evalua-  
 727 tion metrics on both datasets, highlighting its limited generalizability  
 728 and reliability.

## 729 4.7 PLM Embedding Comparison

730 Table 5 shows the comparative performance among models with  
 731 different PLMs for extracting amino acid sequence embeddings.

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

## 748 4.8 Ablation Studies

749 We have further conducted ablation studies on how protein struc-  
 750 tural information may affect immunogenicity prediction and UQ

751 performances by comparing the results from the VenusVaccine  
 752 model architecture with and without the derived protein structural  
 753 features. Table 6 shows the ablation results for all models for the  
 754 Immuno-Virus dataset. The table supports three principal conclu-  
 755 sions. **First**, structural information offers predictive advantages  
 756 in methods such as Ensembles, SWAG, and SVDKL, but fails to  
 757 demonstrate consistent utility across all approaches. For instance,  
 758 its effect is minimal in Deterministic, Dropout, and EDL, and is  
 759 detrimental in LA and DVBL. Thus, its overall contribution to  
 760 predictive performance appears method-specific and inconclusive.  
 761 **Second**, the inclusion of structural information generally enhances  
 762 uncertainty quantification, as evidenced by improved performance  
 763 across uncertainty metrics in Table 6. **Third**, uncertainty-aware  
 764 approaches consistently outperform the deterministic baseline in  
 765 nearly all cases, with the only exception being the precision metric  
 766 in the absence of structural information, where the deterministic  
 767 model exhibits a slight advantage. These observations highlight the  
 768 robustness of uncertainty-aware models and their superiority in  
 769 capturing both predictive accuracy and well-calibrated predictions.  
 770

## 771 4.9 Discussion

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

783 Performance differences became more nuanced across the out-of-  
 784 distribution evaluation scenarios. SVDKL and SWAG demonstrated  
 785 strong calibration robustness under distribution shift, whereas LA  
 786 and Ensembles often achieved higher predictive performance in  
 787 specific scenarios. This indicates a trade-off between reliability and  
 788 sharpness that should be aligned with downstream objectives. The  
 789 inclusion of structural information produced method-dependent  
 790 effects—enhancing both accuracy and calibration for SWAG, En-  
 791 sembles, and SVDKL, while being negligible or even detrimental  
 792 for some methods—likely due to the speed–accuracy trade-offs in-  
 793 herent in the ESMFold predictor used in this study.

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

## 798 5 RELATED WORKS

799 **Protein Language Models.** Advances in deep learning and the  
 800 emergence of large language models, including specialized protein  
 801 language models (PLMs), have revolutionized computational biol-  
 802 ogy by offering accurate, generalizable, and scalable solutions to  
 803 complex downstream tasks such as vaccinology, drug discovery, im-  
 804 munogenicity prediction, and therapeutic design. DeepNetBim [59]  
 805 employs a hybrid architecture combining convolutional neural net-  
 806 works with attention mechanisms to integrate sequence features  
 807

**Table 5: 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	<b>0.0215</b>	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	<b>0.1058</b>	0.1150			
	DVBLL	0.0346	0.0628	0.0550	0.0891	<b>0.0182</b>	0.0824	0.1463	0.1520	0.1353	0.1297	0.0669	0.1218	<b>0.0094</b>	0.1735	<b>0.0310</b>			
	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	<b>0.0108</b>	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	<b>0.0255</b>	<b>0.0143</b>	0.0528	0.0412	<b>0.0133</b>	0.1344	0.0587	<b>0.0269</b>	<b>0.0041</b>	<b>0.0189</b>	0.1159	0.0483	0.1276	0.0491			
	EDL	0.0152	0.0477	0.0409	0.0367	0.0260	0.1259	0.0991	<b>0.0537</b>	0.0337	0.0717	0.0700	<b>0.1129</b>	0.1580	0.1630	0.0944			
	TS	0.0311	0.0588	0.0592	0.0223	0.0266	0.1192	<b>0.0935</b>	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	<b>0.2443</b>	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	<b>0.4580</b>	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	<b>0.5528</b>	0.5083			
	SWAG	<b>0.2422</b>	<b>0.2317</b>	<b>0.2172</b>	0.2505	<b>0.2207</b>	<b>0.4046</b>	0.6300	<b>0.4420</b>	0.4299	<b>0.4196</b>	<b>0.4610</b>	0.5865	0.4784	0.5988	<b>0.4825</b>			
	EDL	0.2537	0.3013	0.3017	0.2796	0.2480	0.5237	<b>0.4966</b>	0.4924	<b>0.4224</b>	0.4517	0.4867	<b>0.5852</b>	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	<b>0.0645</b>	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	<b>0.1550</b>	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	<b>0.1302</b>	0.1639	0.2518	0.1993	<b>0.1784</b>	0.1713			
	SWAG	<b>0.0667</b>	<b>0.0665</b>	<b>0.0608</b>	0.0725	<b>0.0623</b>	<b>0.1251</b>	0.1611	<b>0.1303</b>	0.1363	0.1316	<b>0.1527</b>	<b>0.1915</b>	0.1583	0.1890	<b>0.1609</b>			
	EDL	0.0743	0.0907	0.0877	0.0822	0.0714	0.1739	<b>0.1549</b>	0.1546	<b>0.1303</b>	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			

**Table 6: Experimental results for evaluation on effect of protein structural information (PSI = Protein Structural Information) 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	✓	<b>0.9345</b>	<b>0.9479</b>	<b>0.9249</b>	<b>0.9363</b>	<b>0.9809</b>	0.0238	<b>0.1850</b>	<b>0.0521</b>
	✗	0.9181	0.9380	0.9043	0.9208	0.9767	<b>0.0146</b>	0.1994	0.0593
Deterministic	✓	<b>0.9055</b>	0.9082	<b>0.9059</b>	0.9071	0.9647	<b>0.0413</b>	0.2808	<b>0.0711</b>
	✗	0.9030	<b>0.9479</b>	0.8721	<b>0.9084</b>	<b>0.9721</b>	0.0463	<b>0.2542</b>	0.0713
SWAG	✓	<b>0.9219</b>	<b>0.9380</b>	<b>0.9108</b>	<b>0.9242</b>	<b>0.9728</b>	0.0743	<b>0.2422</b>	<b>0.0667</b>
	✗	0.9005	0.8933	0.9091	0.9011	0.9682	<b>0.0569</b>	0.2586	0.0737
DROPOUT	✓	<b>0.9055</b>	0.9082	<b>0.9059</b>	0.9071	0.9647	<b>0.0411</b>	0.2799	<b>0.0710</b>
	✗	0.9018	<b>0.9454</b>	0.8719	0.9071	<b>0.9720</b>	0.0468	<b>0.2533</b>	0.0714
DVBLL	✓	0.9030	0.9107	0.8995	0.9051	0.9621	<b>0.0346</b>	0.2882	0.0729
	✗	<b>0.9118</b>	<b>0.9181</b>	<b>0.9091</b>	<b>0.9136</b>	<b>0.9739</b>	0.0482	<b>0.2551</b>	<b>0.0655</b>
LA	✓	0.8904	0.9256	0.8674	0.8956	0.9633	<b>0.0108</b>	<b>0.2498</b>	0.0750
	✗	<b>0.9055</b>	0.9256	<b>0.8923</b>	<b>0.9086</b>	<b>0.9668</b>	0.0626	0.3317	<b>0.0745</b>
EDL	✓	0.9005	0.9305	<b>0.8803</b>	0.9047	0.9643	<b>0.0152</b>	0.2537	0.0743
	✗	<b>0.9018</b>	<b>0.9429</b>	0.8736	<b>0.9069</b>	<b>0.9709</b>	0.0404	<b>0.2441</b>	<b>0.0683</b>
TS	✓	<b>0.9055</b>	0.9082	<b>0.9059</b>	0.9071	0.9647	<b>0.0311</b>	0.2604	0.0699
	✗	0.9030	<b>0.9479</b>	0.8721	<b>0.9084</b>	<b>0.9721</b>	0.0372	<b>0.2381</b>	<b>0.0698</b>
SVDKL	✓	<b>0.9156</b>	<b>0.9305</b>	<b>0.9058</b>	<b>0.9180</b>	<b>0.9577</b>	<b>0.1058</b>	<b>0.3050</b>	<b>0.0801</b>
	✗	0.9081	0.9156	0.9044	0.9100	0.9572	0.3113	0.5392	0.1743

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.

## 6 CONCLUSION & LIMITATIONS

**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 multi-modal 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**—to the best of our knowledge, the first UQ benchmark for immunogenicity prediction—offers a valuable resource for developing more reliable and uncertainty-aware antigen design tools.

**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.

## REFERENCES

- [1] Jeannette Adu-Bobie, Barbara Cepcchi, Davide Serruto, Rino Rappuoli, and Mariagrazia Pizza. 2003. Two years into reverse vaccinology. *Vaccine* 21, 7–8 (Jan. 2003), 605–610.
- [2] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep Evidential Regression. arXiv:1910.02600 [cs.LG] <https://arxiv.org/abs/1910.02600>
- [3] Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. 2021. Uncertainty Sets for Image Classifiers using Conformal Prediction. In *International Conference on Learning Representations*. [https://openreview.net/forum?id=eNdiU\\_DbM9](https://openreview.net/forum?id=eNdiU_DbM9)
- [4] Wentao Bao, Qi Yu, and Yu Kong. 2021. Evidential Deep Learning for Open Set Action Recognition. arXiv:2107.10161 [cs.CV] <https://arxiv.org/abs/2107.10161>
- [5] Glenn W. Brier. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 78, 1 (Jan. 1950), 1. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- [6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] <https://arxiv.org/abs/2005.14165>
- [7] Jorg J. A. Calis, Matt Maybeno, Jason A. Greenbaum, Daniela Weiskopf, Aruna D. De Silva, Alessandro Sette, Can Kesmir, and Bjoern Peters. 2013. Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLOS Computational Biology* 9, 10 (10 2013), 1–13. <https://doi.org/10.1371/journal.pcbi.1003266>
- [8] Tianqi Chen, Emily B. Fox, and Carlos Guestrin. 2014. Stochastic Gradient Hamiltonian Monte Carlo. arXiv:1402.4102 [stat.ME] <https://arxiv.org/abs/1402.4102>
- [9] Santiago Cortes-Gomez, Carlos Miguel Patiño, Yewon Byun, Steven Wu, Eric Horvitz, and Bryan Wilder. 2025. Utility-Directed Conformal Prediction: A Decision-Aware Framework for Actionable Uncertainty Quantification. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=IOmnn1hSBO>
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] <https://arxiv.org/abs/1810.04805>
- [11] Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D. Skeel, and Hartmut Neven. 2014. Bayesian sampling using stochastic gradient thermostats. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) (NIPS'14). MIT Press, Cambridge, MA, USA, 3203–3211.
- [12] Nikolae Doneva, Irini Doytchinova, and Ivan Dimitrov. 2021. Predicting Immunogenicity Risk in Biopharmaceuticals. *Symmetry* 13, 3 (2021). <https://doi.org/10.3390/sym13030388>
- [13] Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkardawy, Charlotte Rochereau, and Burkhard Rost. 2023. Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. arXiv:2301.06568 [cs.LG] <https://arxiv.org/abs/2301.06568>
- [14] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinberger, Debsindhu Bhowmik, and Burkhard Rost. 2022. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 10 (Oct. 2022), 7112–7127.
- [15] ESM Team. 2024. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. <https://evolutionaryscale.ai/blog/esm-cambrian>. EvolutionaryScale Website.
- [16] Xinjie Fan, Shujian Zhang, Korawat Tanwisuth, Xiaoning Qian, and Mingyu Zhou. 2021. Contextual Dropout: An Efficient Sample-Dependent Dropout Module. arXiv:2103.04181 [cs.LG] <https://arxiv.org/abs/2103.04181>
- [17] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 2022. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* 13, 1 (July 2022), 4348.
- [18] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 1050–1059. <https://proceedings.mlr.press/v48/gal16.html>
- [19] Jakob Gawlikowski, Cedrique Rovile Njieutche Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. 2022. A Survey of Uncertainty in Deep Neural Networks. arXiv:2107.03342 [cs.LG] <https://arxiv.org/abs/2107.03342>
- [20] Ethan Goan and Clinton Fookes. 2020. *Bayesian Neural Networks: An Introduction and Survey*. Springer International Publishing, 45–87. [https://doi.org/10.1007/978-3-030-42553-1\\_3](https://doi.org/10.1007/978-3-030-42553-1_3)
- [21] Kevin P Greenman, Ava P Amini, and Kevin K Yang. 2025. Benchmarking uncertainty quantification for protein engineering. *PLoS Comput. Biol.* 21, 1 (Jan. 2025), e1012639.

- 1045 [22] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration  
1046 of Modern Neural Networks. arXiv:1706.04599 [cs.LG] <https://arxiv.org/abs/1706.04599>
- 1047 [23] James Harrison, John Willes, and Jasper Snoek. 2024. Variational Bayesian Last  
1048 Layers. arXiv:2404.11599 [cs.LG] <https://arxiv.org/abs/2404.11599>
- 1049 [24] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay,  
1050 Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert,  
Rohil Badkundri, Irfhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina,  
1051 Neil Thomas, Yousfat A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie,  
Matthew Nemeth, Patrick D. Hsu, Tom Seraru, Salvatore Candido, and Alexander Rives.  
1052 2025. Simulating 500 million years of evolution with a language  
model. *Science* 387, 6736 (2025), 850–858. <https://doi.org/10.1126/science.ads0018>  
arXiv:<https://www.science.org/doi/pdf/10.1126/science.ads0018>
- 1053 [25] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii  
1054 Nechaev, Florian Matthes, and Burkhard Rost. 2019. Modeling aspects of the  
1055 language of life through transfer-learning protein sequences. *BMC Bioinformatics*  
1056 20, 1 (Dec. 2019), 723.
- 1057 [26] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez,  
1058 Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost.  
1059 2024. Bilingual language model for protein sequence and structure.  
*NAR Genomics and Bioinformatics* 6, 4 (11 2024), lqae150. <https://doi.org/10.1093/nargab/lqae150> arXiv:<https://academic.oup.com/nargab/article-pdf/6/4/lqae150/6077754/lqae150.pdf>
- 1060 [27] Pavel Izmailov, Wesley J. Maddox, Polina Kirichenko, Timur Garipov, Dmitry  
1061 Vetrov, and Andrew Gordon Wilson. 2019. Subspace Inference for Bayesian Deep  
1062 Learning. arXiv:1907.07504 [cs.LG] <https://arxiv.org/abs/1907.07504>
- 1063 [28] Siddhartha Jain, Ge Liu, Jonas Mueller, and David Gifford. 2020. Maximizing  
1064 Overall Diversity for Improved Uncertainty Estimates in Deep Ensembles.  
arXiv:1906.07380 [cs.LG] <https://arxiv.org/abs/1906.07380>
- 1065 [29] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. 2020. Being  
1066 Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks.  
arXiv:2002.10118 [stat.ML] <https://arxiv.org/abs/2002.10118>
- 1067 [30] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple  
1068 and Scalable Predictive Uncertainty Estimation using Deep Ensembles.  
arXiv:1612.01474 [stat.ML] <https://arxiv.org/abs/1612.01474>
- 1069 [31] Chloe H Lee, Jaesung Huh, Paul R Buckley, Myeongjun Jang, Mariana Pereira  
1070 Pinho, Ricardo A Fernandes, Agne Antanaviciute, Alison Simmons, and Hashem  
Koohey. 2023. A robust deep learning workflow to predict CD8 + T-cell epitopes.  
*Genome Med.* 15, 1 (Sept. 2023), 70.
- 1071 [32] Jongseok Lee, Matthias Humt, Jianxiang Feng, and Rudolph Triebel. 2020. Estimating  
1072 Model Uncertainty of Neural Networks in Sparse Information Form.  
arXiv:2006.11631 [cs.LG] <https://arxiv.org/abs/2006.11631>
- 1073 [33] Guangyuan Li, Balaji Iyer, V B Surya Prasath, Yizhao Ni, and Nathan  
1074 Salomonis. 2021. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity.  
*Briefings in Bioinformatics* 22, 6 (05 2021), bbab160. <https://doi.org/10.1093/bib/bbab160> arXiv:<https://academic.oup.com/bib/article-pdf/22/6/bbab160/41087451/bbab160.pdf>
- 1075 [34] Song Li, Yang Tan, Song Ke, Liang Hong, and Bingxin Zhou. 2025. Immunogenicity Prediction with Dual Attention Enables Vaccine Target Selection.  
arXiv:2410.02647 [cs.LG] <https://arxiv.org/abs/2410.02647>
- 1076 [35] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu,  
1077 Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos  
Costa, Maryam Fazel-Zarandi, Tom Seraru, Salvatore Candido, and Alexander  
1078 Rives. 2023. Evolutionary-scale prediction of atomic-level protein structure with  
a language model. *Science* 379, 6637 (2023), 1123–1130. <https://doi.org/10.1126/science.adc2574> arXiv:<https://www.science.org/doi/pdf/10.1126/science.adc2574>
- 1079 [36] David J. C. MacKay. 1992. A Practical Bayesian Framework for Backpropagation  
1080 Networks. *Neural Computation* 4, 3 (05 1992), 448–472. <https://doi.org/10.1162/neuro.1992.4.3.448> arXiv:<https://direct.mit.edu/neuro/article-pdf/4/3/448/812348/neuro.1992.4.3.448.pdf>
- 1081 [37] Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon  
1082 Wilson. 2019. A Simple Baseline for Bayesian Uncertainty in Deep Learning.  
arXiv:1902.02476 [cs.LG] <https://arxiv.org/abs/1902.02476>
- 1083 [38] Kevin P. Murphy. 2022. *Probabilistic Machine Learning: An introduction*. MIT  
1084 Press. <http://probml.github.io/book1>
- 1085 [39] Hugh O'Brien, Max Salm, Laura T Morton, Maciej Szukszto, Felix O'Farrell,  
1086 Charlotte Boulton, Laurence King, Supreet Kaur Bola, Pablo D Becker, Andrew  
Craig, Morten Nielsen, Yardena Samuels, Charles Swanton, Marc R Mansour,  
Sine Reker Hadrup, and Sergio A Quezada. 2024. A modular protein language  
1087 modelling approach to immunogenicity prediction. *PLoS Comput. Biol.* 20, 11  
(Nov. 2024), e1012511.
- 1088 [40] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin,  
1089 Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can You Trust  
1090 Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset  
Shift. arXiv:1906.02530 [stat.ML] <https://arxiv.org/abs/1906.02530>
- 1091 [41] Amir Hossein Rahmati, Sanket Janatre, Weifeng Zhang, Yucheng Wang, Byung-  
1092 Jun Yoon, Nathan M. Urban, and Xiaoning Qian. 2025. C-LoRA: Contextual  
1093 Low-Rank Adaptation for Uncertainty Estimation in Large Language Models.  
arXiv:2505.17773 [cs.LG] <https://arxiv.org/abs/2505.17773>
- 1094 [42] Roshan Rao, Joshua Meier, Tom Seraru, Sergey Ovchinnikov, and Alexander Rives.  
2020. Transformer protein language models are unsupervised structure learners.  
(Dec. 2020).
- 1095 [43] Hippolyt Ritter, Aleksandar Botev, and David Barber. 2018. A Scalable Laplace  
1096 Approximation for Neural Networks. In *International Conference on Learning  
Representations*. <https://openreview.net/forum?id=Skvd2xAZ>
- 1097 [44] Alexander Rives, Joshua Meier, Tom Seraru, Siddhartha Goyal, Zeming Lin, Jason  
1098 Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus.  
2021. Biological structure and function emerge from scaling unsupervised learning  
1099 to 250 million protein sequences. *Proceedings of the National Academy of  
Sciences* 118, 15 (2021), e2016239118. <https://doi.org/10.1073/pnas.2016239118>  
arXiv:<https://www.pnas.org/doi/pdf/10.1073/pnas.2016239118>
- 1100 [45] Ruslan Salakhutdinov and Andriy Mnih. 2008. Bayesian probabilistic matrix  
1101 factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International  
Conference on Machine Learning* (Helsinki, Finland) (ICML '08). Association for Computing  
1102 Machinery, New York, NY, USA, 880–887. <https://doi.org/10.1145/1390156.1390267>
- 1101 [46] Huub Schellekens. 2002. Bioequivalence and the immunogenicity of biopharmaceuticals.  
*Nat. Rev. Drug Discov.* 1, 6 (June 2002), 457–462.
- 1102 [47] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential Deep  
1103 Learning to Quantify Classification Uncertainty. arXiv:1806.01768 [cs.LG]  
1104 <https://arxiv.org/abs/1806.01768>
- 1105 [48] Ola Shorinwa, Zhihong Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar.  
2025. A Survey on Uncertainty Quantification of Large Language Models: Taxonomy,  
1106 Open Research Challenges, and Future Directions. arXiv:2412.05563 [cs.CL]  
1107 <https://arxiv.org/abs/2412.05563>
- 1108 [49] Yang Tan, Mingchen Li, Ziyi Zhou, Pan Tan, Huiqun Yu, Guisheng Fan, and  
1109 Liang Hong. 2024. PETA: evaluating the impact of protein transfer learning with  
1110 sub-word tokenization on downstream applications. *J. Cheminform.* 16, 1 (Aug.  
1111 2024), 92.
- 1112 [50] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne  
1113 Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambo, Faisal  
1114 Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lampe.  
2023. LLAMA: Open and Efficient Foundation Language Models.  
arXiv:2302.13971 [cs.CL] <https://arxiv.org/abs/2302.13971>
- 1113 [51] Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita,  
1114 Jeongjae Lee, Cameron L M Gilchrist, Johannes Söding, and Martin Steinegger.  
2024. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.*  
42, 2 (Feb. 2024), 243–246.
- 1115 [52] Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and  
1116 Nazneen Fatema Rajani. 2021. BERTology Meets Biology: Interpreting Attention  
1117 in Protein Language Models. arXiv:2006.15222 [cs.CL] <https://arxiv.org/abs/2006.15222>
- 1118 [53] Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. 2025.  
1119 BLO-B: Bayesian Low-Rank Adaptation by Backpropagation for Large Language  
1120 Models. arXiv:2406.11675 [cs.LG] <https://arxiv.org/abs/2406.11675>
- 1121 [54] Andrew Gordon Wilson, Zhitong Hu, Ruslan Salakhutdinov, and Eric P. Xing.  
2016. Stochastic Variational Deep Kernel Learning. arXiv:1611.00336 [stat.ML]  
1122 <https://arxiv.org/abs/1611.00336>
- 1123 [55] Andrew Gordon Wilson and Pavel Izmailov. 2022. Bayesian Deep Learning  
1124 and a Probabilistic Perspective of Generalization. arXiv:2002.08791 [cs.LG]  
1125 <https://arxiv.org/abs/2002.08791>
- 1126 [56] Jingcheng Wu, Wenzhe Wang, Jiucheng Zhang, Binbin Zhou, Wenyi Zhao, Zhixi  
1127 Su, Xun Gu, Jian Wu, Zhan Zhou, and Shuqing Chen. 2019. DeepHLApan: A  
1128 deep learning approach for neoantigen prediction considering both HLA-peptide  
binding and immunogenicity. *Front. Immunol.* 10 (Nov. 2019), 2559.
- 1129 [57] Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang  
1130 Ma, Runcheng Liu, and Jian Tang. 2022. PEER: A Comprehensive and Multi-Task  
1131 Benchmark for Protein Sequence Understanding. arXiv:2206.02096 [cs.LG]  
1132 <https://arxiv.org/abs/2206.02096>
- 1133 [58] Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison.  
2024. Bayesian Low-rank Adaptation for Large Language Models.  
arXiv:2308.13111 [cs.LG] <https://arxiv.org/abs/2308.13111>
- 1134 [59] Xiaoyun Yang, Liyuany Zhao, Fang Wei, and Jing Li. 2021. DeepNetBIM: deep  
1135 learning model for predicting HLA-epitope interactions based on network analysis  
1136 by harnessing binding and immunogenicity information. *BMC Bioinformatics*  
22, 1 (May 2021), 231.
- 1137 [60] Chengpeng Yu, Xing Fang, and Hui Liu. 2025. A unified cross-attention model  
1138 for predicting antigen binding specificity to both HLA and TCR molecules.  
arXiv:2405.06653 [q-bio.BM] <https://arxiv.org/abs/2405.06653>
- 1139 [61] Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. 2020. Mix-n-Match: Ensemble  
1140 and Compositional Methods for Uncertainty Calibration in Deep Learning.  
arXiv:2003.07329 [cs.LG] <https://arxiv.org/abs/2003.07329>

## 1161 A ADDITIONAL EXPERIMENTAL RESULTS

1162 Table 7, 8 and 9 demonstrates the results on all three immunogenic datasets for models with protein sequence embeddings extracted from  
 1163 PLMs except ESM-Cambrian.

1165 **Table 7: Results on Immuno-Virus dataset.**

1167 PLM	1168 Model	Accuracy( $\uparrow$ )	Precision( $\uparrow$ )	Recall( $\uparrow$ )	F1 Score( $\uparrow$ )	AUC ROC( $\uparrow$ )	ECE( $\downarrow$ )	NLL( $\downarrow$ )	Brier Score( $\downarrow$ )
-	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

**Table 8: Results on Immuno-Bacteria dataset.**

PLM	Model	Accuracy( $\uparrow$ )	Precision( $\uparrow$ )	Recall( $\uparrow$ )	F1 Score( $\uparrow$ )	AUC ROC( $\uparrow$ )	ECE( $\downarrow$ )	NLL( $\downarrow$ )	Brier Score( $\downarrow$ )
-	Ensemble	0.8327	0.6897	0.8054	0.7430	0.8883	0.0685	0.4788	0.1302
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
ProstT5	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
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
Ankh	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
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
ESM2	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
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
Prot Bert	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 9: Results on Immuno-Tumor dataset.**

PLM	Model	Accuracy( $\uparrow$ )	Precision( $\uparrow$ )	Recall( $\uparrow$ )	F1 Score( $\uparrow$ )	AUC ROC( $\uparrow$ )	ECE( $\downarrow$ )	NLL( $\downarrow$ )	Brier Score( $\downarrow$ )
-	Ensemble	0.7500	0.7869	0.6486	0.7111	0.8483	0.0413	0.4701	0.1599
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
ProstT5	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
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
Ankh	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
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
ESM2	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
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
Prot Bert	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

1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508