2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822009

Predicting Antibody Stability pH Values from Amino Acid Sequences: Leveraging Protein Language Models for Formulation Optimization

Takuya Tsutaoka
Bio Science & Engineering Laboratory
FUJIFILM Corporation
Kanagawa, Japan
takuya.tsutaoka@fujifilm.com

Noriji Kato
Bio Science & Engineering Laboratory
FUJIFILM Corporation
Tokyo, Japan
noriji.kato@fujifilm.com

Toru Nishino
Bio Science & Engineering Laboratory
FUJIFILM Corporation
Tokyo, Japan
toru.nishino@fujifilm.com

Yuanzhong Li
Bio Science & Engineering Laboratory
FUJIFILM Corporation
Kanagawa, Japan
yuanzhong.li@fujifilm.com

Masahito Ohue
School of Computing
Institute of Science Tokyo
Kanagawa, Japan
ohue@c.titech.ac.jp

***Abstract*—Monoclonal antibodies (mAbs) offer significant therapeutic benefits; however, their formulation requires careful optimization to prevent instability. Standard practices for determining optimal formulation conditions rely on time-consuming and costly wet lab experiments. We developed a machine learning-based approach to predict the optimal pH value for stabilizing mAbs using only their amino acid sequences by leveraging a protein language model. Due to the absence of directly relevant methods, we established a baseline by comparing various combinations of elements. We also conducted feature engineering to enhance the predictive performance by incorporating structural information and descriptors. Our approach achieved a high Pearson correlation coefficient of 0.88 on the test folds from 10-fold cross-validation, highlighting its potential to complement wet lab experiments and increase the efficiency of mAb formulation.**

***Index Terms*—Biopharmaceutical, Antibody Formulation Prediction, Protein Language Model**

I. Introduction

Monoclonal antibodies (mAbs) represent a significant advancement in therapeutic medicine, offering targeted treatment options for a variety of diseases [1]. However, their inherent instability poses challenges for long-term storage. To ensure mAb therapies remain effective and safe, formulation conditions must be optimized to enhance stability, particularly concerning excipients and the pH value of the solution.

Some studies have investigated factors contributing to the instability of antibodies and have proposed policies for their formulation [1]. However, the mechanisms underlying these instabilities are complex and not fully understood. Therefore, optimizing formulation conditions generally relies on time-consuming and expensive wet lab experiments, which include multiple assays to determine the most stable conditions.

Recent advances have introduced dry laboratory approaches that use machine learning to predict aspects of antibody stability, such as hydrophobicity [2], [3], and thermal stability [4]. Although these methods offer valuable insights, they are mainly intended for screening purposes and do not lead to the optimal formulation conditions. There have been prior study that have used machine learning to predict the pH value at which mAbs are stable [5], but the pH values addressed were defined based on conformation and were usually more acidic (approximately pH 2–3.5) than neutral (approximately pH 4.8–8), which is often used in the final formulation [6]. Our method aimed to provide an accurate prediction of the pH value within this general range.

The proposed method employs machine learning techniques to predict the optimal pH value for stabilizing mAbs. First, the amino acid sequence of the mAb was processed using a protein language model (pLM) [7]–[9]. These features were then aggregated and input into a regression model trained to correlate extracted features with known pH values associated with mAb formulations. To enhance the predictive performance, the method emphasizes extracting features from residues which are crucial for stability [10]. Additionally, existing descriptors [11] were integrated with the pLM features, improving the predictive performance.

II. Related Work

A. Descriptors

To effectively handle protein amino acid sequences in machine learning and statistical analyses, various descriptors have been proposed. These descriptors were designed using expert knowledge and provide numerical representations reflecting the biophysical or structural properties of amino acids within a sequence. Common examples include T-scale [12], FASGAI [13], Cruciani Properties [14], and ProtFP [15].

979-8-3503-8622-6/24/$31.00 ©2024 IEEE

B. Protein Language Models

Protein Language models (pLMs) are machine learning models designed to understand and predict various properties of proteins based on their amino acid sequences. Unlike traditional descriptors, pLMs automatically learn vector representations of proteins. These models are utilized in various downstream tasks, such as protein function prediction.

In recent years, several types of pLMs have been proposed. **General pLMs:** These models are trained on a broad range of protein sequences from various organisms. Examples include models, such as ESM-1b [7] and ESM-2 [8], which use large-scale sequence data to learn representation that can be applied to different protein-related tasks.

Antibody-Specific pLMs: These models, such as AbLang [9] are designed to extract antibody-specific features by training on datasets limited to antibodies. Despite having less training data than general pLMs, these models show better performance on antibody-related tasks, such as B-cell classification.

III. Materials and Methods

A. Dataset

We constructed an original dataset by collecting information on 56 commercially available FDA-approved mAb drugs, proposed between 1999 and 2022, from the web. In particular, the amino acid sequences and domain information (variable regions and Fab regions) were retrieved from the IMGT-DB [16], a publicly accessible web database. The pH values of these mAbs were extracted from FDA documents available online. The dataset comprised both the amino acid sequences and their corresponding pH values, which were used for training and evaluating the machine learning models.

B. Baseline Construction

We compared various combinations of the following:

- **Antibody Domains:** We tested different antibody domains, including the variable region (V), Fab region, and full antibody (Whole), as the choice of domain can influence the predictive performance.
- **Protein Language Models:** We used general pLMs, such as ESM-1b and ESM-2, as well as the antibody-specific pLM, AbLang.
- **Heavy and Light Chain Features:** After the features extraction of each chain by the pLM, how to integrate them into the features of the entire mAb is not obvious. We compared the averaging (Mean) and concatenation (Concat) of the features of both chains.
- **Regression Models:** We selected models commonly used in scenarios with limited data, including Lasso [17] and Support Vector Regression (SVR) [18].

C. Feature Engineering

We explored several feature engineering strategies to enhance the predictive performance:

- **Using Structural Information:** We aimed to enhance the predictive performance by incorporating structural information into pLM features. Considering novel mAbs and the computing resources at runtime, we utilized ESMFold [19], a fast structure prediction method. ESMFold will be compared with AlphaFold2 [20] in the subsection of Ablation Study.
  To utilize structural information, we considered two key aspects: solvent-accessible surface areas (SASA) and loop regions. SASA of residues were calculated using mkdssp [21] based on the predicted structures, and the features were calculated by the SASA-weighted average of the pLM vectors of each residue. Loop regions were detected by analyzing the 3D coordinates of amino acid chains. We defined nine residues around the vertices as loop regions, and the features were calculated from only the residues within the loop regions. An example of the loop region detection results is shown in Fig. 1. We assessed the effect of varying the number of residues considered in the loop region (which we defined as 'loop width') on performance as shown in the subsection of Ablation Study.
- **Combining Descriptors:** To enhance prediction performance, we aimed to combine descriptors that are thought to be related to biophysical quantities with the pLM. We tested all available descriptors from the R package 'Peptides' [11] and explored their combination with the pLM features, which are represented as high-dimensional vectors, in two ways: addition and multiplication. The predictive performances of individual descriptors will be discussed in the subsection of Ablation Study.

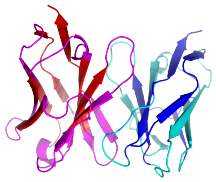

Fig. 1. Example of loop region detection. The red and blue represent heavy and light chains, and magenta and cyan represent the respective detected loop regions.

D. Evaluation Method

We used a 10-fold cross-validation to divide the data into training, validation, and test sets for model training, hyperparameter tuning, and evaluation. The evaluation metric was the Pearson correlation coefficient. Hyperparameter tuning was performed using Optuna [22].

IV. Results

A. Baseline Construction

The results, summarized in Table I, indicated that the optimal combination for predicting the pH value achieved a Pearson correlation coefficient of 0.75. The best-performing combination was to use the variable region, perform feature extraction using AbLang, average the feature values of the heavy and light chains, and apply Lasso. We decided on this combination as the baseline configuration.

TABLE I
COMPARATIVE EXPERIMENTAL RESULTS FOR BASELINE CONSTRUCTION.

Protein Language Model		Lasso			SVR		
		Whole	Fab	V	Whole	Fab	V
ESM-1b	Concat	0.4305	0.4488	0.3317	0.5169	0.2988	0.3136
	Mean	0.4929	0.4247	0.4386	0.4204	0.4289	0.3278
ESM-2	Concat	0.3492	0.2749	0.0728	0.1941	0.1239	0.2168
	Mean	0.3830	0.3903	0.3697	0.2701	0.2393	0.2892
AbLang	Concat	0.5453	0.5203	0.5493	0.6696	0.6166	0.6798
	Mean	0.4471	0.3196	**0.7518**	0.3336	0.5046	0.6770

TABLE II
FEATURE ENGINEERING RESULTS OF COMBINING DESCRIPTORS WITH PROTEIN LANGUAGE MODEL FEATURES

	Baseline With Loop Region-Aware Features					
mAb Domains for Descriptors	Whole		Fab		V	
Combining methods	add	multiply	add	multiply	add	multiply
T2 (T-scale)	0.8535	**0.8839**	0.8525	0.8625	0.8451	0.7634
PP1 (Cruciani Properties)	0.8470	0.8821	0.8371	0.8193	0.8427	0.7557
VHSE2 (VHSE Scales)	0.8570	0.8817	0.8574	0.8772	0.8593	0.8721
F5 (FASGAI)	0.8482	0.8755	0.8427	0.7678	0.8815	0.2998
ProtFP4 (ProtFP)	0.8560	0.8741	0.8511	0.8738	0.8427	0.7940

B. Feature Engineering

Feature Engineering was confirmed to improve the predictive performance:

- **Using Structural Information:** Incorporating features based on SASA and loop regions led to improvements over the baseline. Features based on SASA yielded the test correlation coefficient of 0.80, while focusing on loop regions resulted in a substantial increase to 0.85.
- **Combining Descriptors:** Combining pLM features with existing descriptors also improved the performance. Due to the large number of descriptors, only the top five descriptors are shown in Table II. In contrast to the pLM, the use of the whole mAb was confirmed to be the best in the calculation of descriptors. The best configuration in this paper was the loop region-aware features multiplied by T2, achieving a test correlation coefficient of 0.88.

The scatter plot of the prediction results of the best configuration in this paper is shown in Fig. 2. The predictions of our method were found to correlate well with the ground truth.

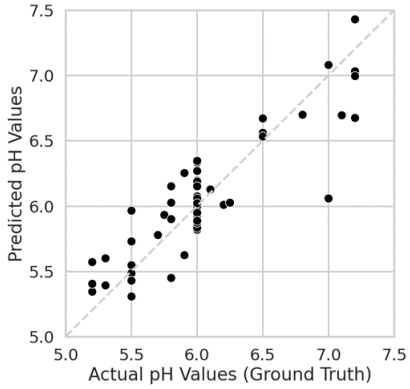

Fig. 2. Scatter plots of the test folds obtained from the best configuration in this paper.

V. DISCUSSION

A. Baseline Construction

The best-performing combination, which achieved a test correlation coefficient of 0.75, likely benefited from several key factors. The use of the Lasso regression model was particularly effective due to its feature selection capabilities for the multi-dimensional pLM features. Using the variable region of the antibody also proved advantageous. This region is known to exhibit more variation between different antibodies. Moreover, the antibody-specific language model, AbLang, may have effectively captured these differences, leading to improved feature characterization.

B. Feature Engineering: Using Structural Information

The improvement in predictive performance by focusing on features extracted from residues with high solvent accessibility or residues within loop regions suggests that these residues may markedly contribute to antibody stability. This finding is consistent with that of related research [2], [3], [10], supporting the importance of these structural information. SASA and flexible loop regions may play crucial roles in the stability of antibodies by interacting with their environment. In addition, pLMs might lack some structural information.

C. Feature Engineering: Combining Descriptors

The enhancement in performance by combining pLM features with descriptors indicates that existing pLMs may not explicitly capture all relevant biophysical information. Descriptors provided supplementary biophysical insights that contributed to the overall predictive performance of the model.

D. Ablation Study

To validate the robustness and reliability of our proposed method, we conducted an ablation study to assess the impact of various factors on predictive performance.

TABLE III
COMPARISON OF LOOP WIDTHS.

Width	3	5	7	9	11	13	15
Corr.	0.73	0.81	0.85	**0.85**	0.83	0.80	0.81

Effect of Different Fold Splits: We tested 10 different fold splits. The baseline configuration had an average correlation coefficient of 0.7557 ± 0.0169. The performance remained stable and close to the result of 0.75, suggesting that the choice of splits does not markedly affect the performance, indicating the robustness of our approach.

Impact of Loop Width: We assessed the influence of varying the loop width on predictive performance. Although changing the width led to some variation in performance, widths that were not extremely small consistently resulted in better performance than the baseline and the SASA (Table III). This result supports the effectiveness of focusing on loop regions for enhancing the predictive performance.

Comparison of ESMFold with AlphaFold2: We compared ESMFold and AlphaFold2 for loop region detection. For AlphaFold2, we evaluated all five default structural predictions, and the result was 0.84 ± 0.01. Although the AlphaFold2 results were slightly worse than ESMFold, they still outperformed the baseline. The performance with ESMFold may be comparable to AlphaFold2 because of the relative simplicity of the structural features required to detect loop regions.

Performance with Descriptors Only: We assessed the performance using only descriptors (i.e., without the pLM features). The maximum test correlation coefficient achieved was approximately 0.4. This finding implies that the pLM provides the high-quality features needed for pH value prediction while the descriptors may provide supplementary information.

VI. CONCLUSION

We proposed a novel machine learning approach for predicting optimal pH values to stabilize mAbs using only their amino acid sequences. To the best of our knowledge, this is the first application for such task. Through a comprehensive evaluation of various combinations of antibody domains, pLMs, and regression models, we established a robust baseline and improved the predictive performance with targeted feature engineering. Our method highlights the potential of enhancing the efficiency of mAb formulation by providing accurate predictions of pH values for mAb stabilization.

REFERENCES

[1] C. Mieczkowski, X. Zhang, D. Lee, K. Nguyen, W. Lv, Y. Wang, Y. Zhang, J. Way, and J.-M. Gries, "Blueprint for antibody biologics developability," *mAbs*, vol. 15, no. 1, p. 2185924, 2023.

[2] T. Jain, T. Boland, A. Lilov, I. Burnina, M. Brown, Y. Xu, and M. Vásquez, "Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning," *Bioinformatics*, vol. 33, no. 23, pp. 3758–3766, 2017.

[3] F. Waibl, M. L. Fernández-Quintero, F. S. Wedl, H. Kettenberger, G. Georges, and K. R. Liedl, "Comparison of hydrophobicity scales for predicting biophysical properties of antibodies," *Frontiers in Molecular Biosciences*, vol. 9, p. 960194, 2022.

[4] A. Harmalkar, R. Rao, Y. Richard Xie, J. Honer, W. Deisting, J. Anlahr, A. Hoenig, J. Czwikla, E. Sienz-Widmann, D. Rau *et al.*, "Toward generalizable prediction of antibody thermostability using machine learning on sequence and structure features," *mAbs*, vol. 15, no. 1, p. 2163584, 2023.

[5] A. C. King, M. Woods, W. Liu, Z. Lu, D. Gill, and M. R. Krebs, "High-throughput measurement, correlation analysis, and machine-learning predictions for ph and thermal stabilities of Pfizer-generated antibodies," *Protein Science*, vol. 20, no. 9, pp. 1546–1557, 2011.

[6] R. G. Strickley and W. J. Lambert, "A review of formulations of commercially available antibodies," *Journal of Pharmaceutical Sciences*, vol. 110, no. 7, pp. 2590–2608, 2021.

[7] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021.

[8] 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*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[9] T. H. Olsen, I. H. Moal, and C. M. Deane, "AbLang: an antibody language model for completing antibody sequences," *Bioinformatics Advances*, vol. 2, no. 1, p. vbac046, 2022.

[10] H.-J. Chang, J.-W. Jian, H.-J. Hsu, Y.-C. Lee, H.-S. Chen, J.-J. You, S.-C. Hou, C.-Y. Shao, Y.-J. Chen, K.-P. Chiu *et al.*, "Loop-sequence features and stability determinants in antibody variable domains by high-throughput experiments," *Structure*, vol. 22, no. 1, pp. 9–21, 2014.

[11] D. Osorio, P. Rondón-Villarreal, and R. Torres, "Peptides: A Package for Data Mining of Antimicrobial Peptides," *The R Journal*, vol. 7, no. 1, pp. 4–14, 2015.

[12] F. Tian, P. Zhou, and Z. Li, "T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides," *Journal of Molecular Structure*, vol. 830, no. 1-3, pp. 106–115, 2007.

[13] G. Liang, G. Chen, W. Niu, and Z. Li, "Factor analysis scales of generalized amino acid information as applied in predicting interactions between the human amphiphysin-1 SH3 domains and their peptide ligands," *Chemical Biology & Drug Design*, vol. 71, no. 4, pp. 345–351, 2008.

[14] G. Cruciani, M. Baroni, E. Carosati, M. Clementi, R. Valigi, and S. Clementi, "Peptide studies by means of principal properties of amino acids derived from MIF descriptors," *Journal of Chemometrics*, vol. 18, no. 3-4, pp. 146–155, 2004.

[15] G. J. van Westen, R. F. Swier, J. K. Wegner, A. P. IJzerman, H. W. van Vlijmen, and A. Bender, "Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets," *Journal of Cheminformatics*, vol. 5, pp. 1–11, 2013.

[16] V. Giudicelli, D. Chaume, and M.-P. Lefranc, "IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes," *Nucleic Acids Research*, vol. 33, no. suppl_1, pp. D256–D261, 2005.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[18] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer science & business media, 2013.

[19] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido *et al.*, "Language models of protein sequences at the scale of evolution enable accurate structure prediction," *bioRxiv*, 2022.

[20] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[21] W. Kabsch and C. Sander, "DSSP: definition of secondary structure of proteins given a set of 3D coordinates," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[22] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.