CODEN: WJAPAC Impact Factor: 5.48 ISSN: 3049-3013

World Journal of Advance Pharmaceutical Sciences

Volume 3, Issue 8, Page: 84-90 Research Article www.wjaps.com

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) FOR PREDICTING PROTEIN AGGREGATION AND VISCOSITY IN HIGH-CONCENTRATION LYOPHILIZED BIOLOGICS DURING LONG TERM STORAGE

***Dr. Akhilesh Kumar**

Kalra Hospital SRCNC Private Limited, New Delhi, India.

How to cite this Article: *Dr. Akhilesh Kumar. (2026). EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) FOR PREDICTING PROTEIN AGGREGATION AND VISCOSITY IN HIGH-CONCENTRATION LYOPHILIZED BIOLOGICS DURING LONG TERM STORAGE. World Journal of Advance Pharmaceutical Sciences, 3(8), 84-90.

Article Info

Article Received: 01 July 2026,

Article Revised: 21 July 2026,

Article Accepted: 11 August 2026.

***Corresponding author:**

Dr. Akhilesh Kumar

Kalra Hospital SRCNC Private Limited, New Delhi, India.

DOI: https://doi.org/10.5281/zenodo.22014156

ABSTRACT

High-concentration biologics are increasingly developed for subcutaneous administration because they enable smaller injection volumes and improve patient convenience. However, increasing protein concentration can substantially increase viscosity, protein-protein interactions, self-association, and aggregation. Lyophilization can improve the storage stability of protein therapeutics, but freezing, dehydration, residual moisture, glass-transition behaviour, reconstitution, and long-term storage may introduce additional degradation risks. Conventional stability development is largely dependent on experimental screening and accelerated studies, which can be time-consuming and may not adequately capture complex nonlinear interactions. This study proposes an Explainable Artificial Intelligence (XAI) framework for predicting aggregation and viscosity in high-concentration lyophilized biologics during long-term storage. The framework integrates protein molecular descriptors, formulation composition, lyophilization parameters, storage conditions, and analytical stability measurements. Random Forest and gradient-boosting models are proposed for prediction, while SHapley Additive exPlanations (SHAP) are used to identify the variables responsible for individual predictions. Protein concentration, pH, ionic strength, stabilizer-to-protein ratio, residual moisture, glass-transition temperature, storage temperature, storage duration, and molecular interaction characteristics are considered key predictive variables. The framework separately models aggregation and viscosity while identifying shared molecular and formulation drivers. Existing evidence indicates that concentrated protein viscosity is strongly affected by intermolecular interactions, while lyophilized stability depends on formulation composition, freezing history, residual moisture, and storage conditions. The proposed XAI approach can convert complex multidimensional stability data into interpretable predictions and experimentally testable formulation hypotheses. It provides a transparent computational strategy for accelerating formulation development and improving long-term stability-risk assessment.

KEYWORDS: Explainable artificial intelligence; protein aggregation; high-concentration biologics; viscosity; lyophilization; stability prediction.

1. INTRODUCTION

Biological medicines, particularly monoclonal antibodies and other protein therapeutics, have become essential components of modern pharmaceutical therapy. The increasing demand for patient-friendly administration has encouraged the development of high-concentration formulations suitable for subcutaneous delivery. However, increasing protein concentration creates

substantial formulation challenges, including elevated viscosity, protein-protein interactions, self-association, aggregation, particle formation, and reduced injectability.[1–4]

Viscosity is particularly important because highly viscous formulations can create difficulties during manufacturing, filling, syringe ability, and administration. Protein concentration is a major determinant, but different proteins can exhibit markedly different viscosities at equivalent concentrations. Molecular size, shape, charge distribution, hydrophobicity, intermolecular attraction, and formulation composition contribute to this variability.[2–4]

Aggregation is another major quality concern. Protein molecules can aggregate because of conformational instability, hydrophobic association, disulfide formation, chemical modification, or other intermolecular mechanisms. Aggregation may reduce product quality and potentially affect safety and immunogenicity.

Lyophilization is commonly used to improve protein stability by removing water and reducing molecular mobility. The process consists principally of freezing, primary drying, and secondary drying. However, freezing can produce concentrated solute environments, while drying can alter protein-excipient interactions. Residual moisture and glass-transition behaviour can further influence long-term stability.[5–7]

Long-term storage adds additional complexity. Aggregation kinetics may be nonlinear and may differ between accelerated and real-time conditions. Consequently, prediction based solely on accelerated stability data may not always accurately represent long-term behaviour.[8]

These challenges generate complex datasets containing molecular, formulation, manufacturing, analytical, and storage variables. Conventional regression methods may not adequately capture nonlinear interactions among these variables.

Machine learning can model complex relationships within high-dimensional datasets. Recent studies have demonstrated the potential of machine learning for predicting viscosity in high-concentration monoclonal antibodies.[9] However, a prediction without an explanation has limited value for pharmaceutical formulation development.

Explainable Artificial Intelligence provides a solution by identifying the contribution of individual variables to model predictions. SHAP, for example, can explain both global feature importance and individual formulation predictions.[10]

The objective of this study is therefore to develop an interpretable XAI framework for predicting protein aggregation and viscosity in high-concentration lyophilized biologics during long-term storage.

2. METHODOLOGY

2.1 Study Framework

The proposed framework integrates four major data domains:

Protein characteristics → Formulation → Lyophilization → Storage

The primary prediction targets are:

1. protein aggregation;
2. post-reconstitution viscosity.

Each observation represents a specific protein-formulation-process combination measured at a defined storage time and temperature.

2.2 Molecular Features

Protein-level descriptors should include:

- molecular weight;
- sequence length;
- theoretical pI;
- net charge;
- hydrophobicity;
- charged-residue fraction;
- exposed hydrophobic surface;
- aggregation-prone regions;
- secondary-structure characteristics;
- cysteine content;
- molecular shape;
- charge-patch characteristics.

These descriptors capture intrinsic differences in protein stability and intermolecular interactions.

2.3 Formulation Features

The formulation dataset should include:

- protein concentration;
- pH;
- buffer type and concentration;
- ionic strength;
- sodium chloride concentration;
- sucrose or trehalose concentration;
- mannitol;
- arginine;
- surfactant concentration;
- stabilizer-to-protein ratio;
- osmolality.

These variables are important because formulation composition can simultaneously affect viscosity, aggregation and protein stability.

2.4 Lyophilization and Storage Features

Process variables should include:

- freezing rate;
- nucleation conditions;
- shelf and product temperature;
- chamber pressure;

- primary-drying duration;
- secondary-drying conditions;
- residual moisture;
- glass-transition temperature;
- collapse temperature.

Storage variables should include:
- storage temperature;
- storage duration;
- temperature excursions;
- container-closure characteristics.

2.5 Analytical Measurements
The framework can incorporate:
- size-exclusion chromatography;
- dynamic light scattering;
- viscosity/rheology;
- subvisible particle analysis;
- residual moisture;
- glass-transition measurements;
- circular dichroism;
- FTIR;
- interaction parameters such as the Huggins coefficient.

Aggregation should preferably be represented as percentage aggregate or percentage monomer loss. Viscosity should be measured under standardized temperature and shear conditions.

2.6 Machine-Learning Models
Random Forest and gradient-boosting models are proposed as the principal algorithms because they can model nonlinear relationships and feature interactions in heterogeneous tabular datasets.[11,12]

Linear regression should be used as a baseline model.

Separate models should initially be developed for aggregation and viscosity because these outcomes have overlapping but distinct mechanisms.

Model performance should be evaluated using:
- MAE;
- RMSE;
- R^2;
- sensitivity;
- specificity;
- F1-score;
- ROC-AUC, where classification is used.

To avoid data leakage, all measurements belonging to the same protein-formulation combination should remain within the same training, validation, or test group.

2.7 Explainable AI
SHAP should be used as the primary XAI method.[10]
For each prediction, SHAP identifies whether a feature increases or decreases the predicted outcome.
Global analysis can rank variables according to their mean absolute SHAP contribution.
Local analysis can explain individual high-risk formulations.

For example, a high aggregation prediction may be associated with:
- elevated storage temperature;
- high residual moisture;
- low stabilizer-to-protein ratio;
- high protein concentration.

A high viscosity prediction may be associated with:
- high protein concentration;
- unfavourable charge distribution;
- strong intermolecular interactions;
- formulation-specific ionic strength;
- insufficient viscosity-reducing excipient.

SHAP interaction analysis can additionally evaluate relationships such as concentration × pH, concentration × ionic strength, and residual moisture × glass-transition temperature.

3. RESULTS
3.1 Predictive Framework
Because an experimentally generated dataset was not supplied for this study, numerical model-performance values are not fabricated. The proposed framework defines the experimental and computational results that should be obtained following validation with real stability data.

Tree-based models are expected to outperform simple linear models because protein viscosity and aggregation involve nonlinear molecular and formulation interactions.

The expected performance structure is:

Model	Expected capability
Linear regression	Baseline
Random Forest	Strong nonlinear prediction
Gradient boosting	Strong nonlinear and interaction modeling
Neural network	Potentially high performance with large datasets

3.2 Expected Aggregation Drivers

The most important aggregation predictors are expected to include:

1. storage temperature;
2. storage duration;
3. residual moisture;
4. stabilizer-to-protein ratio;
5. glass-transition margin;
6. protein concentration;
7. pH;
8. freezing conditions;
9. intrinsic aggregation propensity.

Temperature and time are expected to become increasingly important during long-term storage. Residual moisture may demonstrate nonlinear behaviour because both excessive moisture and excessive drying can influence protein stability.

3.3 Expected Viscosity Drivers

Protein concentration is expected to be the strongest general predictor of viscosity. However, molecular characteristics are expected to explain differences among proteins at the same concentration.

Important variables are expected to include:

- protein concentration;
- molecular charge distribution;
- pH;
- ionic strength;
- hydrophobicity;
- molecular architecture;
- protein-protein interaction parameters;
- excipient composition.

This expectation is consistent with experimental studies showing that monoclonal antibodies can exhibit substantially different viscosities at similar concentrations.[2–4]

3.4 XAI Interpretation

The principal advantage of the proposed framework is its ability to convert predictions into interpretable explanations.
For example:

High aggregation risk

- storage temperature → positive contribution;
- storage time → positive contribution;
- residual moisture → positive contribution;
- stabilizer ratio → negative contribution;
- glass-transition margin → negative contribution.

High viscosity risk

- protein concentration → strong positive contribution;
- attractive intermolecular interactions → positive contribution;
- ionic strength → formulation-dependent contribution;
- arginine → potentially negative contribution;
- pH → protein-dependent contribution.

These represent the intended interpretation structure and must be replaced with actual SHAP values following model training.

4. DISCUSSION

4.1 Viscosity and Protein Concentration

High protein concentration increases molecular crowding and protein-protein interactions. However, concentration alone does not explain viscosity differences between biologics. Previous studies have demonstrated substantial molecule-specific variation in viscosity, highlighting the importance of molecular charge, shape and interaction characteristics.[2–4]

XAI can identify these molecule-specific drivers and therefore provide information that conventional concentration-based models cannot provide.

4.2 Aggregation During Long-Term Storage

Aggregation is affected by both intrinsic protein stability and environmental conditions. Temperature and time are expected to be major predictors, but accelerated studies cannot always reproduce real-time aggregation mechanisms.[8]

A time- and temperature-aware model is therefore preferable:
[A=f(X,t,T)]

where (A) is aggregation, (X) represents molecular and formulation variables, (t) is storage duration and (T) is temperature.

This approach allows the model to learn formulation-specific stability patterns instead of assuming a universal degradation equation.

4.3 Role of pH and Ionic Strength

pH influences protein charge, conformation and intermolecular attraction. Ionic strength can alter electrostatic interactions and may either increase or decrease viscosity depending on the protein.

Therefore, the model should not assume that a particular pH or salt concentration is universally beneficial.

XAI is useful because it can show whether the effect of pH or ionic strength is positive, negative or conditional for a particular biologic.

4.4 Role of Excipients

Sugars such as sucrose and trehalose can stabilize proteins during freezing and drying but may also increase solution viscosity after reconstitution.[13]

Arginine and related excipients may reduce viscosity in selected high-concentration formulations, but their effects depend on protein identity and formulation conditions.

This creates a formulation trade-off between:
structural stability ↔ viscosity ↔ processability ↔ injectability.

An XAI model can identify which variables appear to control each quality attribute and thereby guide targeted experimental studies.

4.5 Lyophilization Effects

Freezing rate, residual moisture and glass-transition behaviour are particularly relevant to lyophilized biologics.

Freezing can redistribute protein and excipients into concentrated regions, potentially affecting subsequent aggregation. Residual moisture can increase molecular mobility, whereas aggressive drying may alter protein-excipient interactions.

Therefore, residual moisture should be modeled as a continuous variable rather than simply categorized as "acceptable" or "unacceptable."

5. COMPARISON WITH PRIOR STUDIES

Previous research has examined individual components of the proposed problem.

High-concentration antibody studies have demonstrated that viscosity is influenced by concentration, molecular interactions, charge and molecular architecture.[2–4]

Lyophilization research has demonstrated that freezing conditions, stabilizer composition and drying conditions influence protein stability.[5–8]

Sequence-based approaches have investigated intrinsic aggregation propensity, while high-throughput formulation studies have evaluated viscosity and stability experimentally.

More recent research has applied machine learning to high-concentration antibody viscosity prediction.[9]

The proposed framework differs by integrating:
molecular characteristics + formulation + lyophilization + storage + XAI
into one interpretable prediction system.

Approach	Main strength	Limitation
Experimental screening	Direct evidence	Time and material intensive
Regression	Simple interpretation	Limited nonlinear modeling
Molecular prediction	Captures intrinsic risk	Limited formulation context
ML viscosity prediction	Strong predictive capability	May lack interpretability
Proposed XAI framework	Prediction + explanation	Requires high-quality longitudinal data

6. PRACTICAL SIGNIFICANCE

The proposed system could support pharmaceutical development through:

- early identification of high-viscosity formulations;
- prediction of aggregation risk;
- prioritization of stability experiments;
- identification of critical formulation variables;
- optimization of lyophilization parameters;
- interpretation of unexpected stability failures;
- long-term stability-risk assessment.

A practical workflow would be:
Characterization → Formulation → Lyophilization → Stability Testing → Machine Learning → XAI → Formulation Hypothesis → Experimental Validation
This creates a closed-loop approach in which model predictions guide experiments and experimental results subsequently improve the model.

7. LIMITATIONS

The framework has several limitations.
First, model quality depends on the quantity and quality of experimental data.
Second, analytical differences among laboratories and instruments may introduce measurement bias.

Third, correlated variables can complicate interpretation of SHAP values.

Fourth, model explanations describe statistical relationships rather than proving biological causality.

Fifth, models trained primarily on monoclonal antibodies may not generalize to other biologic modalities.

Finally, no numerical experimental results are claimed in this manuscript because an original stability dataset was not supplied. Actual model performance, SHAP values and statistical significance should be reported only after experimental validation.

8. FUTURE DIRECTIONS

Future research should focus on:

1. **Multimodal AI:** integration of protein sequence, structural and formulation data.
2. **Uncertainty quantification:** reporting prediction confidence alongside point estimates.
3. **Active learning:** selecting the most informative next formulation experiments.
4. **Causal AI:** distinguishing correlation from experimentally supported causation.

5. **Digital stability models:** continuously updating shelf-life predictions as real-time stability data accumulate.
6. **Prospective validation:** testing predictions using previously unseen protein molecules and formulations.

9. CONCLUSION

High-concentration biologics provide important advantages for patient-centric drug delivery but introduce substantial viscosity and aggregation challenges. Lyophilization can improve stability but creates additional variables related to freezing, drying, residual moisture and molecular mobility.

The proposed XAI framework integrates molecular, formulation, lyophilization and storage variables to predict protein aggregation and viscosity while providing interpretable explanations.

Protein concentration is expected to be a major viscosity determinant, whereas storage temperature, time, residual moisture, stabilizer ratio and intrinsic molecular characteristics are expected to contribute strongly to aggregation. pH, ionic strength and excipients may affect both outcomes in a protein-specific manner.

The principal advantage of XAI is its ability to move beyond prediction toward explanation. Instead of simply identifying a formulation as high risk, the framework can identify the variables contributing to that risk and generate experimentally testable hypotheses.

The approach can potentially reduce formulation-screening requirements, improve stability-risk assessment and support data-driven pharmaceutical development. However, AI predictions should complement rather than replace laboratory stability studies. Prospective validation using independent biologic molecules, real-time storage data and controlled formulation experiments is essential before regulatory or commercial application.

Overall, explainable artificial intelligence represents a promising approach for transforming complex biologic stability data into transparent and actionable information and may contribute significantly to the development of robust high-concentration lyophilized biologics.

ACKNOWLEDGEMENTS

The author acknowledges the scientific literature and research community whose work on protein formulation, aggregation, viscosity, lyophilization and artificial intelligence contributed to this research framework.

FUNDING

No external funding was received for this work.

CONFLICT OF INTEREST

The author declares no conflict of interest.

DATA AVAILABILITY

No proprietary experimental dataset was generated or analysed for this manuscript. The framework requires validation using experimentally generated stability data.

ETHICAL APPROVAL

Not applicable.

REFERENCES

1. Páramo T, et al. High-concentration protein formulations: How high is high? Eur J Pharm Biopharm, 2017; 119: 95–105.
2. Liu J, Nguyen MDH, Andya JD, Shire SJ. Reversible self-association increases the viscosity of a concentrated monoclonal antibody in aqueous solution. J Pharm Sci., 2005; 94: 1928–1940.
3. Yadav S, Shire SJ, Kalonia DS. Factors affecting the viscosity in high concentration solutions of different monoclonal antibodies. J Pharm Sci., 2010; 99: 4812–4829.
4. Yadav S, Shire SJ, Kalonia DS. Viscosity behavior of high-concentration monoclonal antibody solutions. J Pharm Sci., 2012; 101: 998–1011.
5. Ó'Fágáin C, Colliton K. Storage and lyophilization of pure proteins. Methods Mol Biol., 2017; 1485: 159–190.
6. Franks F. Freeze-drying: from empiricism to predictability. The significance of glass transitions. Dev Biol Stand, 1992; 74: 9–18.
7. Miller MA, Rodrigues MA, Glass MA, Singh SK, Johnston KP, Maynard JA. Frozen-state storage stability of a monoclonal antibody: aggregation is impacted by freezing rate and solute distribution. J Pharm Sci., 2013; 102: 1194–1208.
8. Accelerated aggregation studies of monoclonal antibodies: considerations for storage stability. J Pharm Sci., 2020.
9. Kalejaye LA, et al. Accelerating high-concentration monoclonal antibody development with large-scale viscosity data and ensemble deep learning. mAbs., 2025; 17.
10. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, 2017.
11. Breiman L. Random forests. Mach Learn, 2001; 45: 5–32.
12. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016; 785–794.
13. Effect of sugar molecules on the viscosity of high concentration monoclonal antibody solutions. Pharm Res., 2011.
14. Woldeyes MA, Qi W, Razinkov VI, Furst EM, Roberts CJ. How well do low- and high-concentration protein interactions predict solution viscosities of monoclonal antibodies? J Pharm Sci., 2019; 108: 142–154.
15. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any

classifier. In: Proceedings of NAACL-HLT., 2016; 97–101.
16. International Council for Harmonisation. ICH Q5C: Stability Testing of Biotechnological/Biological Products. ICH.