Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

PREDICTING LYOPHILIZED BIOLOGIC COLLAPSE TEMPERATURE USING FORMULATION, THERMAL, FREEZING, AND MOISTURE DATA

James Walker[1]*, Olivia Harris[1]

1. *Department of Computational Pharmacy, Faculty of Medicine and Health Sciences, University of Edinburgh, Edinburgh, United Kingdom.*

ARTICLE INFO

Received:
27 April 2025
Received in revised form:
07 July 2025
Accepted:
09 July 2025
Available online:
28 August 2025

Keywords: Lyophilization, Collapse temperature, Biologics, Machine learning, Formulation design, Thermal analysis

ABSTRACT

Collapse of a lyophilized cake is a severe failure mode for biologic products because it compromises structure, appearance, and downstream usability. Collapse temperature is difficult to measure experimentally and is governed by interacting formulation and process variables. Formulation scientists lack a rapid quantitative tool for predicting collapse temperature from routinely collected development information. Current decisions often rely on empirical rules, limited thermal measurements, and small numbers of freeze-drying microscopy observations. The objective of this predictive model article is to describe a machine learning framework for estimating collapse temperature in lyophilized biologic formulations. The model is intended to use formulation composition, thermal analysis, freezing-process descriptors, and residual-moisture information. A gradient-boosted regression framework is proposed for learning relationships among protein concentration, excipient composition, glass-transition behavior, cooling history, annealing conditions, and residual moisture. The workflow emphasizes curated inputs, physicochemically meaningful feature engineering, uncertainty-aware prediction, and interpretability. Conceptually, the model would be expected to support collapse-temperature prediction across protein–excipient combinations without replacing experimental confirmation. It could rank formulation variables by their influence on collapse risk and guide in-silico screening before targeted freeze-drying microscopy. A predictive collapse-temperature tool could accelerate biologic lyophilization development by connecting formulation design with process-risk assessment. Such a model is aligned with quality-by-design thinking because it converts scattered development observations into actionable formulation and cycle-design knowledge.

To Cite This Article: Walker J, Harris O. Predicting Lyophilized Biologic Collapse Temperature Using Formulation, Thermal, Freezing, and Moisture Data. Pharmacophore. 2025;16(4):1-10. https://doi.org/10.51847/ALiZV6iSqX

Introduction

Lyophilization remains central to the manufacture of unstable biologic drug products because water removal can improve storage stability while preserving the activity of sensitive macromolecules [1]. However, the process is vulnerable to macroscopic collapse when the product temperature exceeds the structurally sustainable limit during drying, leading to poor cake elegance, altered reconstitution behavior, and potential stability concerns [2]. Because acceptable appearance is not merely cosmetic but also linked to manufacturability and product perception, collapse risk is treated as a critical formulation and process-development consideration [2]. The need to avoid collapse becomes especially important for high-concentration monoclonal antibody and vaccine formulations in which excipient selection, solids content, and freezing history interact strongly [3].

Current practice commonly relies on freeze-drying microscopy to observe collapse directly or on differential scanning calorimetry to estimate glass-transition behavior in the maximally freeze-concentrated phase [4]. For some amorphous systems, collapse temperature may be interpreted in relation to Tg′, but this relationship depends on formulation composition, crystallization behavior, and the thermal history imposed during freezing and drying [5]. Freeze-drying microscopy can be labor-intensive and instrument-dependent, while DSC-based thermal markers may not fully account for interactions among sugars, proteins, polyols, buffers, and surfactants [6]. These limitations restrict the ability to screen many candidate biologic formulations during early development, when rapid decisions about cryoprotectants and bulking agents are often required [7].

Corresponding Author: James Walker; Department of Computational Pharmacy, Faculty of Medicine and Health Sciences, University of Edinburgh, Edinburgh, United Kingdom. E-mail: james.walker@gmail.com

The increasing availability of formulation, thermal, process, and quality-attribute data from development batches creates an opportunity for predictive modelling in lyophilization. Machine learning has already been used to accelerate biopharmaceutical formulation decisions by learning complex relationships between excipients and product behavior [8]. Related computational formulation work has shown that supervised models can extract useful signals from antibody–excipient interaction data, supporting data-driven excipient selection in biologic formulation development [9]. In lyophilization itself, modelling studies and automated product-inspection workflows suggest that structured process and product data can be used to support drying optimization and quality assessment [10].

This article proposes a predictive modelling framework in which formulation composition, DSC-derived thermal descriptors, freezing parameters, and residual-moisture data are used to predict collapse temperature for lyophilized biologics. The approach is motivated by the broader movement toward model-based lyophilization design, where product temperature, endpoint behavior, and drying robustness are increasingly described through formal computational frameworks [11]. A machine learning model would not replace freeze-drying microscopy, but it could prioritize the most promising formulations for experimental confirmation and identify combinations that are more likely to collapse under aggressive cycles [12]. By combining prediction with interpretability, such a model could provide formulation guidance while remaining compatible with quality-by-design expectations for biologic process development [13].

Background

Lyophilization Process and Collapse Mechanisms

Lyophilization consists of freezing, primary drying, and secondary drying, and each stage influences the physical structure of the final dried cake [14]. During freezing, solute concentration in the unfrozen fraction determines the viscosity and thermal properties of the maximally freeze-concentrated matrix, while primary drying exposes this matrix to stress as ice sublimes [4]. Collapse occurs when the dried or partially dried matrix lacks sufficient viscosity to maintain pore structure, producing shrinkage, loss of cake elegance, and potential changes in drying resistance [15]. Because cake appearance is connected to process robustness and patient-facing product quality, collapse is often treated as a practical failure mode even when chemical degradation is not immediately observed [2].

Measurement and Prediction of Collapse Temperature

Collapse temperature is most directly assessed through freeze-drying microscopy, whereas DSC provides supporting thermal descriptors such as Tg′ and eutectic melting events [4]. The relationship between Tg′ and collapse temperature is useful but not universal because multi-component biologic systems can display overlapping thermal transitions, excipient crystallization, and concentration-dependent mobility [5]. Calorimetric investigations of amorphous lyophilized solids show that relaxation behavior provides information about molecular mobility, which is relevant to collapse and long-term physical stability [5]. Therefore, a predictive model should treat FDM-derived Tc as the reference output while using DSC-derived descriptors as informative but not fully deterministic inputs [16].

Formulation Effects on Thermal Properties

Formulation composition governs the thermal behavior of lyophilized biologics because proteins, sugars, polyols, buffers, and surfactants alter water distribution, glass formation, crystallization, and molecular mobility [6]. Sucrose and trehalose are often used to stabilize proteins in amorphous matrices, whereas mannitol and other polyols may contribute bulking effects but can introduce crystallization-dependent thermal complexity [1]. Dextran-containing systems illustrate that excipients can affect thermal properties, product quality attributes, and monoclonal antibody stability in ways that are not captured by a single composition variable [7]. Consequently, collapse-temperature prediction requires features that represent both excipient identity and concentration, rather than relying only on total solids or protein load [8].

Freezing Parameters and Residual Moisture

Freezing parameters influence the ice-crystal network, dried-layer resistance, and local solute concentration, which in turn affect the apparent robustness of the lyophilized cake [14]. Controlled ice nucleation and annealing can change pore structure and drying behavior, while residence time in the freeze-concentrated phase may influence protein stability before drying is complete [17]. Annealing and nucleation conditions have also been shown to affect the properties of high-concentration monoclonal antibody formulations, making them relevant inputs for any collapse-risk model [18]. Residual moisture after secondary drying further modifies glassy-state mobility and apparent Tg, so moisture measurements can connect drying history with post-drying collapse susceptibility and stability [16].

Table 1 summarizes how freezing-process variables and residual moisture measurements can be translated into mechanistically meaningful inputs for a collapse-temperature prediction model.

Table 1. Compact Input–Mechanism Map for Collapse-Temperature Prediction in Lyophilized Biopharmaceutical Formulations

Input domain	Example variables	Mechanistic relevance to collapse risk	Model value

Freezing rate and shelf temperature history	cooling rate, minimum shelf temperature, hold duration	Shapes ice-crystal size, pore network, and freeze-concentrated matrix structure	Helps distinguish formulations with similar composition but different drying resistance
Controlled ice nucleation	nucleation temperature, induction method, nucleation uniformity	Influences initial ice distribution and cake microstructure	Captures process-driven variation that may affect apparent robustness
Annealing conditions	annealing temperature, annealing duration, number of cycles	Promotes ice-crystal growth and modifies dried-layer resistance	Links deliberate thermal conditioning to predicted collapse behavior
Freeze-concentrated phase exposure	time above critical low-temperature thresholds, residence time before primary drying	May increase molecular mobility and stress exposure before drying is complete	Adds temporal context beyond static formulation descriptors
Residual moisture after secondary drying	Karl Fischer moisture, gravimetric moisture, target moisture range	Alters glassy-state mobility, apparent Tg, and post-drying structural stability	Connects drying history with final cake susceptibility to collapse or instability

Machine Learning in Lyophilization and Biopharmaceutical Development

Machine learning has been used in biopharmaceutical formulation design to accelerate the selection of excipients and predict formulation behavior from structured experimental inputs [8]. In lyophilization, automated computer-vision approaches have been proposed for inspecting continuously freeze-dried products and classifying product defects, demonstrating that data-driven models can extract quality-relevant patterns from visual or process-derived information [10]. Deep learning methods for freeze-dried defect classification further suggest that predictive analytics can complement conventional lyophilization development workflows when carefully linked to product attributes [19]. However, a dedicated model focused specifically on predicting collapse temperature from formulation, thermal, freezing, and moisture inputs remains conceptually distinct from models aimed at drying endpoints, image defects, or general formulation optimization [20].

Model Development Overview

High-Level Prediction Framework

The proposed framework assembles each biologic formulation into a single feature vector containing composition, thermal analysis, freezing-process descriptors, and moisture-related inputs. Composition features would capture protein type, protein concentration, stabilizer identity, cryoprotectant and bulking-agent fractions, buffer system, and surfactant presence, reflecting the types of variables used in machine learning-guided formulation development [21]. Thermal features would include Tg′, Tg, eutectic or melting events, and heat-capacity changes where available, because calorimetric markers provide a mechanistic bridge between formulation composition and collapse behavior [5]. The model output would be a continuous predicted collapse temperature with an associated uncertainty estimate so that the prediction can be used conservatively in lyophilization cycle design [12].

Figure 1 presents the proposed end-to-end predictive modeling architecture linking biologic formulation composition, DSC-derived thermal descriptors, freezing-process variables, and residual-moisture data to uncertainty-aware collapse-temperature estimates and QbD-oriented experimental confirmation.

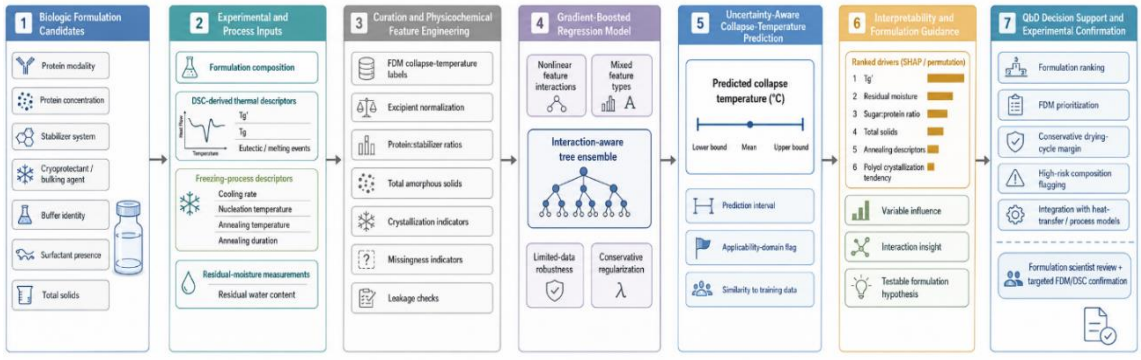

Figure 1. Predictive Collapse-Temperature Modeling Workflow for Lyophilized Biologics Using Formulation, Thermal, Freezing, and Moisture Data.

Core Input Features

Core input features should include the mass fraction of protein and key excipients, especially disaccharides, polyols, buffers, and surfactants, because excipient identity and concentration can shift thermal behavior and product quality attributes [7]. Measured Tg′ values of complete formulations should be prioritized over component-level estimates, although component descriptors may still be useful when complete DSC data are unavailable [16]. Freezing rate, annealing temperature, annealing duration, and nucleation temperature should be represented because these variables affect ice morphology, product resistance,

and final cake structure [18]. Residual moisture content measured after secondary drying should be encoded as a post-process descriptor that helps link drying severity with glassy-state mobility and collapse risk [22].

Design Principles

The model should be interpretable, robust to incomplete development data, and constrained by formulation-science knowledge rather than optimized as a purely statistical exercise. Small biopharmaceutical formulation datasets can be vulnerable to overfitting, so model development should favor cross-validation strategies, regularization, and conservative uncertainty estimates consistent with the limited-data setting of formulation design [9]. Missing data should be handled explicitly because moisture or thermal measurements may not be available for every formulation at early screening stages, while tree-based models can support practical missing-value strategies [23]. The prediction should remain continuous and uncertainty-aware so it can guide safety margins without implying that computational estimates replace FDM confirmation [4].

Data Sources and Feature Engineering

Compilation of a Lyophilization Formulation Dataset

A curated collapse-temperature dataset would combine FDM-derived Tc values with corresponding composition, thermal, freezing, and moisture descriptors from peer-reviewed literature, development reports, and structured freeze-drying databases. Literature sources should be harmonized carefully because studies may differ in sample preparation, thermal history, imaging criteria, and interpretation of cake collapse [4]. Model-based lyophilization studies can provide useful process descriptors such as product temperature, primary-drying strategy, and endpoint behavior, but these inputs must be separated from the intrinsic Tc target to avoid leakage [11]. FDM should remain the reference measurement whenever available, while DSC-derived Tg′ and related events should be treated as predictive features rather than substitutes for the target value [16].

Encoding Formulation Composition and Thermal Features

Formulation composition should be encoded using normalized excipient concentrations, protein-to-stabilizer ratios, total solids, buffer identity, surfactant presence, and indicators for crystalline or amorphous excipient behavior. Because excipients such as dextran, sucrose, trehalose, and mannitol can alter thermal properties and stability through different mechanisms, the feature representation should preserve both identity and concentration information [1]. DSC-derived Tg′, Tg, eutectic melting temperature, and relaxation-related descriptors should be encoded as numeric features because calorimetry captures mobility and phase behavior that are relevant to collapse [5]. For mixtures without a single clear Tg′, pseudo-component descriptors or multiple thermal-event features can represent the complexity without forcing the model to rely on one oversimplified transition [22]. Freezing-process descriptors should also capture the distinction between intrinsic formulation collapse susceptibility and process-driven structural variation. Controlled ice nucleation has been reviewed as a way to reduce variability in nucleation temperature and improve consistency during freeze-drying, supporting its inclusion as a model feature rather than as an uncontrolled batch effect [24]. Practical lyophilization design guidance further indicates that controlled ice nucleation can reduce dry-layer resistance and batch variability in amorphous formulations, making nucleation strategy relevant to both collapse risk and drying performance [25]. In addition, studies of protein formulations dried above Tg′ suggest that collapse risk should not be interpreted from Tg′ alone, because acceptable cake structure may depend on the relationship among formulation composition, protein concentration, product temperature, and the true collapse threshold [26].

Table 2defines the proposed physicochemical feature architecture by distinguishing formulation, thermal, freezing, moisture, derived, and reference-outcome variables according to their mechanistic relevance and modeling role.

Table 2. Physicochemical Feature Architecture for Predicting Collapse Temperature in Lyophilized Biologic Formulations

Feature domain	Representative variables	Mechanistic relevance to collapse temperature	Preferred encoding strategy	Key modeling caution
Biologic identity and load	Protein modality, protein concentration, monoclonal antibody or vaccine class, total biologic solids	Protein concentration affects the freeze-concentrated matrix, viscosity, glass formation, and interaction with stabilizers during drying [3].	Encode protein modality categorically; encode concentration as continuous mass or molar fraction; include protein-to-excipient ratios.	Avoid assuming that all biologics respond similarly to the same excipient system.
Stabilizing sugars	Sucrose fraction, trehalose fraction, sugar-to-protein ratio, total disaccharide content	Disaccharides stabilize amorphous matrices and may raise the structural resistance of the freeze-concentrated phase [1].	Encode individual sugar identity and concentration separately; derive sugar-to-protein and sugar-to-total-solids ratios.	Do not collapse all sugars into a single “stabilizer” variable because excipient identity matters.
Bulking agents and polyols	Mannitol, glycine, dextran, polyol fraction, crystallization tendency	Bulking agents can improve cake structure but may introduce crystallization-dependent thermal behavior and	Use identity-specific categorical indicators, concentration values, and crystallization-risk flags.	A crystalline bulking effect should not be interpreted as equivalent to

		heterogeneous collapse risk [7].		amorphous stabilization.
Buffer and surfactant system	Buffer identity, buffer strength, pH class, surfactant presence, surfactant type	Buffers and surfactants influence local chemistry, protein stability, and phase behavior during freezing and drying [6].	Encode buffer and surfactant identity categorically; include pH or ionic strength when available.	Sparse categories may create unstable model splits in small datasets.
DSC-derived thermal descriptors	Tg′, Tg, eutectic or melting events, heat-capacity change, relaxation descriptors	Thermal transitions provide mechanistic markers of mobility, phase behavior, and collapse susceptibility [5].	Treat complete-formulation Tg′ as a priority numeric feature; retain multiple thermal events when transitions overlap.	Do not use Tg′ as a direct substitute for FDM-derived collapse temperature [16]
Freezing-process descriptors	Cooling rate, nucleation condition, nucleation temperature, annealing temperature, annealing duration, post-freezing hold	Freezing history affects ice morphology, pore structure, dried-layer resistance, and the physical robustness of the cake [17, 18].	Encode continuous process parameters and binary indicators for controlled nucleation or annealing.	Separate formulation-intrinsic collapse susceptibility from equipment-specific drying dynamics.
Residual-moisture descriptors	Karl Fischer moisture, residual water content, water-sorption behavior, secondary-drying severity	Residual water modifies glassy-state mobility, apparent Tg, and post-drying structural vulnerability [16, 22].	Encode moisture as continuous percentage or mass fraction; add missingness indicators when unavailable.	Missing moisture values should not be treated as zero moisture.
Derived physicochemical features	Protein-to-sugar ratio, total amorphous solids, effective water content, crystallization indicators, formulation similarity score	Derived descriptors translate raw composition into formulation-science logic and reduce dimensionality in small datasets [9].	Engineer interpretable ratios, domain-informed interaction terms, and applicability-domain metrics.	Derived variables must remain chemically plausible and should not introduce leakage from the target.
Reference outcome label	FDM-derived collapse temperature, observed cake collapse threshold, measurement context	FDM remains the reference method for directly observing collapse and should define the supervised learning target [4].	Encode collapse temperature as a continuous °C outcome with source and measurement metadata.	DSC-derived values should support prediction but should not replace the reference collapse-temperature label.

Freezing and Moisture Feature Representation

Freezing features should include nominal cooling rate, controlled or spontaneous nucleation conditions, annealing temperature, annealing duration, and any post-freezing hold that affects ice-crystal growth [17]. These features are important because the freezing step can change product resistance and drying structure even when the formulation composition remains unchanged [14]. Moisture descriptors should include residual water content after secondary drying and, where available, water-sorption behavior that reflects how strongly the dried matrix interacts with water [16]. Encoding these variables allows the model to connect pre-drying thermal vulnerability with post-drying glassy-state mobility, which is central to predicting collapse risk and storage robustness [22].

Predictive Model Architecture

Model Choice – Gradient-Boosted Trees

Gradient-boosted tree models are suitable for this conceptual framework because they can learn nonlinear interactions among formulation composition, thermal markers, freezing variables, and moisture descriptors without requiring the strong parametric assumptions of simpler regression models. This is particularly relevant for biologic formulations, where excipient effects may depend on protein concentration, total solids, and thermal history rather than acting independently [8]. Tree-based approaches also provide practical advantages for small, heterogeneous datasets because they can accommodate mixed feature types and support missing-value handling more naturally than many conventional algorithms [23]. In a lyophilization setting, the model should be selected not only for predictive utility but also for its ability to support interpretation and formulation decisions [20].

Input Feature Vector and Pre-processing

The input feature vector should combine continuous variables such as protein concentration, sucrose fraction, Tg′, cooling rate, annealing duration, and residual moisture with categorical variables such as protein modality, buffer identity, and surfactant type. Continuous inputs may be normalized to support stable modelling, while categorical formulation descriptors can be encoded in a way that preserves excipient identity and avoids implying artificial numerical order [21]. Missing moisture or thermal measurements should be handled through explicit imputation strategies or missingness indicators so that the model

can distinguish unavailable data from true low values [9]. Pre-processing should also include checks for chemically implausible combinations, because a statistical model trained on formulation data should remain aligned with known physicochemical constraints [7].

Output: Collapse Temperature and Prediction Interval

The model output should be a point prediction of collapse temperature expressed in degrees Celsius, accompanied by a calibrated prediction interval that communicates uncertainty rather than overstating confidence. This uncertainty is essential because cycle design decisions depend on maintaining product temperature below a safe limit, and model-based lyophilization control strategies require conservative interpretation of product-temperature and quality-risk estimates [12]. Prediction intervals could be generated conceptually through bootstrapping, quantile-style modelling, or ensemble variability, provided that the resulting bounds are evaluated without claiming unsupported numerical performance [15]. The output should therefore function as a decision-support estimate for formulation screening and process design, not as an experimental replacement for FDM in final confirmation [4].

Handling Small Datasets and Physicochemical Constraints

Dealing With Limited Lyophilization Data

Lyophilization datasets for biologics are usually small, heterogeneous, and biased toward formulations that were considered worth developing, so the model should be designed for conservative generalization rather than aggressive fitting. Leave-one-batch-out or leave-one-formulation-family-out validation would be appropriate because it tests whether the model can extrapolate across related but nonidentical development conditions [27]. Bayesian regularization or constrained loss terms could incorporate prior knowledge, such as the expected relationship between Tg′ and collapse risk, without forcing an overly rigid empirical rule [5]. A useful conceptual formulation is $\hat{T}c = f\phi(C_{protein}, C_{sugar}, C_{polyol}, Tg', R_{cool}, A_{time}, A_{temp}, M_{res})$where the learned function maps formulation, thermal, freezing, annealing, and residual-moisture features to a predicted collapse temperature.

Incorporating Domain Knowledge as Feature Engineering

Feature engineering should translate raw formulation variables into physicochemically meaningful descriptors, such as protein-to-sugar ratio, total amorphous solids, effective water content, and indicators of likely crystallization. These derived features reduce dimensionality while preserving formulation logic, which is important when the number of candidate variables exceeds the number of well-characterized lyophilized formulations [9]. For example, the effect of a sugar concentration may depend on protein load and residual water rather than acting as an isolated linear contributor to Tc [1]. Similarly, studies of amorphous-based lyophilized cakes show that process conditions and formulation properties jointly influence collapse temperature and drying behavior, supporting interaction-aware feature construction [28].

Evaluating Confidence on New Formulations

Predictions for new formulations should be accompanied by an applicability-domain or similarity assessment so that the model can flag chemically distant inputs as exploratory. This is especially important when a candidate formulation contains an excipient class, protein modality, or lipid-associated biologic system that is poorly represented in the training data, as shown by recent modelling work on mRNA lipid nanoparticle lyophilization where process parameters and formulation context must be interpreted together [29]. A similarity score could compare the new feature vector with the distribution of known formulations, helping users distinguish interpolation from extrapolation. Such a safeguard would make the model more useful in development settings because uncertainty would be treated as a decision variable rather than hidden behind a single predicted Tc [30].

Model Interpretability and Formulation Guidance

Shap Analysis to Rank Formulation Factors

Model interpretability should focus on identifying which formulation and process features drive predicted collapse temperature, rather than presenting the model as a black-box replacement for thermal analysis. SHAP or permutation-style importance methods could be used to rank Tg′, sugar-to-protein ratio, residual moisture, total solids, and annealing descriptors according to their contribution to the predicted Tc [8]. If Tg′ emerges as a dominant factor, this would be consistent with the known relevance of glass-transition behavior, but the model could also reveal interactions in which moisture or freezing history modifies the practical collapse risk [22]. Automated structural-defect assessment of freeze-dried products further supports the value of interpretable links between measurable product attributes and process decisions [31].

Translating SHAP into Formulation Recommendations

Interpretability should be translated into formulation guidance through controlled “what-if” comparisons that vary one formulation or process input while holding the remaining variables constant. For instance, the model could conceptually evaluate whether increasing trehalose relative to protein, reducing residual moisture, or altering annealing conditions would be expected to increase the predicted Tc for a given formulation family [7]. These recommendations should be framed as

hypotheses for experimental confirmation because excipient effects may depend on crystallization, protein–excipient interactions, and the thermal history of the frozen concentrate [6]. In this role, the model becomes a formulation decision-support tool rather than an autonomous optimizer.

Integration Into Lyophilization Process Design

Pre-Screening before Freeze-Drying Microscopy

Before freeze-drying microscopy, formulators could use the model to rank candidate formulations by predicted collapse temperature and uncertainty. This would allow FDM resources to be focused on formulations that appear promising, borderline, or scientifically informative, rather than on every composition generated during screening [4]. The approach would be especially valuable when many excipient ratios must be compared under development pressure, because machine learning-guided space-filling designs can support efficient exploration of formulation space [30]. Experimental Tc measurements would still be needed for confirmation, but the model could make the screening workflow more targeted and knowledge-driven.

Supporting QbD-Based Control Strategies

Within a quality-by-design framework, the predicted collapse-temperature landscape could help define formulation and process regions where product temperature is expected to remain below a conservative safety threshold. Model-based primary drying optimization has already shown how computational frameworks can support lyophilization control strategies by linking process settings with product-temperature behavior [12]. First-principles primary-drying models used in process design and transfer provide a complementary foundation, because they address heat and mass transfer while the proposed machine learning model focuses on the product-specific collapse limit [32]. Together, these approaches could support a design space in which formulation composition, thermal risk, and drying-cycle aggressiveness are considered jointly.

Evaluation Strategy

Table 3 presents a validation and deployment framework that connects predictive accuracy, uncertainty calibration, applicability-domain assessment, interpretability, and QbD decision use for the proposed collapse-temperature model.

Table 3. Validation, Interpretability, and QbD Deployment Framework for the Collapse-Temperature Prediction Model

Evaluation or deployment layer	Purpose	Recommended approach	Decision-use output	Risk controlled
Reference-label validation	Confirm that predicted values correspond to experimentally observed collapse behavior.	Compare predicted collapse temperature with FDM-derived Tc values using RMSE, MAE, calibration plots, and error stratification by formulation family.	Evidence that the model can support screening decisions before targeted FDM confirmation.	Prevents the model from appearing accurate only because it learned DSC proxies rather than collapse behavior.
Leave-one-family-out validation	Test whether the model generalizes across related but nonidentical biologic formulation groups.	Hold out formulation families, protein classes, or excipient systems during validation.	Conservative estimate of model transferability to new biologic candidates.	Reduces overconfidence caused by highly similar training and test formulations.
Temporal validation	Evaluate whether the model remains useful as development practices, materials, or batches change over time.	Train on earlier development data and test on later formulation batches or later process campaigns.	Evidence of robustness under realistic development progression.	Controls drift from material, equipment, or procedural changes.
Multi-site validation	Determine whether model performance depends on laboratory, equipment, or measurement practice.	Validate across sites, freeze-dryer types, DSC/FDM protocols, and development environments [32].	Classification of the model as local, transferable, or requiring site-specific recalibration.	Prevents inappropriate transfer of a model trained under one measurement context.
Prediction-interval calibration	Ensure uncertainty bounds are meaningful for conservative process decisions.	Use bootstrapping, quantile-style modeling, or ensemble variability, then assess interval coverage.	Predicted Tc with lower-confidence bound for cycle-margin decisions.	Prevents a single point estimate from being treated as a verified thermal limit.
Applicability-domain assessment	Identify whether a new formulation lies within the chemical and process space	Use similarity scoring, distance-to-training-domain measures, or flagging of novel excipient classes [30].	"Interpolation," "borderline," or "exploratory" prediction status.	Prevents extrapolated predictions from being used as if they were validated estimates.

	represented in training data.			
SHAP or feature-importance analysis	Explain which formulation and process variables drive predicted collapse temperature.	Rank Tg′, residual moisture, sugar-to-protein ratio, total solids, annealing descriptors, and crystallization indicators [8].	Interpretable factor ranking for formulation scientists.	Reduces black-box use and supports mechanistic review.
What-if formulation testing	Translate model explanations into experimentally testable formulation hypotheses.	Vary one input at a time, such as trehalose fraction, residual moisture, or annealing duration, while holding other variables constant.	Prioritized formulation modifications for DSC and FDM confirmation.	Prevents model output from becoming an unsupported autonomous recommendation.
FDM prioritization utility	Test whether the model improves experimental screening efficiency.	Compare conventional heuristic screening with model-guided ranking of promising, borderline, and high-risk formulations.	Smaller, more informative FDM testing set without eliminating confirmation experiments.	Keeps the model aligned with real formulation-development decisions.
QbD integration	Link predicted collapse susceptibility with formulation design space and drying-cycle conservatism.	Combine predicted Tc, uncertainty bounds, and product-temperature/process models to define conservative operating regions [12, 13].	QbD-oriented formulation and cycle-design guidance.	Prevents collapse-temperature prediction from being used without heat-transfer and process-context modeling.
Human-review boundary	Maintain scientific and regulatory accountability.	Require formulation scientist review before using predictions to select formulations or modify drying cycles.	Reviewer-approved experimental plan and confirmation strategy.	Prevents automated model output from replacing expert judgment or experimental evidence.

Prediction Accuracy

Prediction accuracy should be evaluated with metrics such as RMSE and MAE, but the manuscript should report them only after independent validation rather than assuming performance in advance. Evaluation should compare predicted Tc with reference FDM-derived values and stratify results conceptually by biologic modality, such as monoclonal antibody, fusion protein, vaccine, or lipid-associated biologic formulation [3]. The purpose of this evaluation would be to determine whether the model can support screening decisions with appropriate uncertainty, not to claim that a single metric proves universal reliability. Graphical design-space thinking for primary drying reinforces that predictive tools are most useful when their outputs are interpreted in relation to process limits and safety margins [15].

Temporal and Multi-Site Validation

Temporal validation should test whether a model trained on earlier development batches remains useful for formulations manufactured later, after changes in materials, equipment, or process practice. Multi-site validation would be similarly important because freeze-dryer geometry, shelf behavior, chamber pressure control, and measurement practices can vary across development and manufacturing environments [32]. Digital-twin approaches for lyophilization highlight the need to connect model predictions with process context, especially when transferring biologics from laboratory to manufacturing scale [13]. A Tc model that performs acceptably only within one laboratory setting should be treated as a local screening aid rather than a broadly transferable predictive tool.

Utility in Formulation Development

The model's practical utility should be evaluated by simulating how it would change formulation-development decisions before FDM confirmation. In such a study, the model could rank candidate formulations, identify high-risk compositions, and recommend a smaller subset for experimental collapse-temperature measurement, while the final decision would remain experimentally grounded [10]. The evaluation should compare the decision pathway created by the model against the pathway that would have been followed using conventional heuristic screening, without inventing numerical savings or unvalidated performance claims. This decision-focused assessment would align the model with real formulation workflows rather than treating prediction accuracy as the only success criterion [21].

Limitations

Data Scarcity for Novel Excipients and Complex Biologics

The proposed model would be limited by the availability and diversity of well-annotated collapse-temperature data. Formulations containing novel cryoprotectants, nontraditional bulking agents, lipid nanoparticles, or complex multi-component biologics may fall outside the training domain and therefore carry large prediction uncertainty [29]. Excipients that modify thermal behavior through crystallization, molecular mobility, or phase separation may also challenge feature

representations that were developed for simpler sugar–protein systems [22]. For these cases, model predictions should be treated as exploratory hypotheses requiring targeted FDM and DSC confirmation.

Influence of Freeze-Dryer Geometry and Scale

Collapse temperature is often treated as a product-specific property, but observed collapse during drying is also influenced by heat transfer, vial location, chamber pressure, dried-layer resistance, and equipment scale. Product-temperature and endpoint modelling studies show that lyophilization performance depends on process dynamics, not only on the intrinsic thermal limit of the formulation [11]. Therefore, a Tc prediction model should be integrated with process models rather than used alone to define a complete drying cycle [32]. The model's scope is best understood as predicting the formulation's collapse susceptibility, while equipment-specific drying kinetics require complementary modelling and experimental verification.

Conclusion

A machine learning model for predicting collapse temperature could connect formulation composition, thermal analysis, freezing history, and residual-moisture information into a single decision-support framework. Such a model would help formulation scientists estimate collapse risk before committing to extensive freeze-drying microscopy. Its value would come from organizing complex development data into interpretable predictions rather than replacing experimental characterization. The strongest advantage of the proposed framework is rapid in-silico screening across biologic formulation candidates. By combining predicted Tc with feature-importance methods, the model could identify which formulation variables most strongly influence collapse risk. This interpretability would make the tool more useful for formulation scientists because it would explain why certain compositions appear more robust than others.

Important challenges remain before such a model could be used routinely. Small training datasets, inconsistent measurement methods, missing thermal or moisture values, and extrapolation to novel excipient systems could all limit reliability. Integration with freeze-dryer heat-transfer models would also be necessary because collapse during drying depends on both formulation properties and process conditions.

The long-term opportunity is a shared, well-curated lyophilization dataset that links formulation composition, thermal events, freezing parameters, moisture data, and experimentally measured collapse temperature. Collaboration among industry, academic laboratories, and technology developers would make such datasets more diverse and useful. With careful validation, uncertainty reporting, and experimental confirmation, AI-assisted lyophilization cycle design could become a routine part of biologic formulation development.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Haeuser C, Goldbach P, Huwyler J, Friess W, Allmendinger A. Excipients for room temperature stable freeze-dried monoclonal antibody formulations. J Pharm Sci. 2020;109(1):807-17.
2. Patel SM, Nail SL, Pikal MJ, Geidobler R, Winter G, Hawe A, et al. Lyophilized drug product cake appearance: what is acceptable? J Pharm Sci. 2017;106(7):1706-21.
3. Najarian J, Metsi-Guckel E, Renawala HK, Grosse D, Sims A, Walter A, et al. Optimizing lyophilization primary drying: a vaccine case study with experimental and modeling techniques. Int J Pharm. 2024;659:124168.
4. Ohori R, Yamashita C. Effects of temperature ramp rate during the primary drying process on the properties of amorphous-based lyophilized cake, Part 1: Cake characterization, collapse temperature and drying behavior. J Drug Deliv Sci Technol. 2017;39:131-9.
5. Groël S, Menzen T, Winter G. Calorimetric investigation of the relaxation phenomena in amorphous lyophilized solids. Pharmaceutics. 2021;13(10):1735.
6. Haeuser C, Goldbach P, Huwyler J, Friess W, Allmendinger A. Be aggressive! Amorphous excipients enabling single-step freeze-drying of monoclonal antibody formulations. Pharmaceutics. 2019;11(11):616.
7. Haeuser C, Goldbach P, Huwyler J, Friess W, Allmendinger A. Impact of dextran on thermal properties, product quality attributes, and monoclonal antibody stability in freeze-dried formulations. Eur J Pharm Biopharm. 2020;147:45-56.
8. Narayanan H, Dingfelder F, Condado Morales I, Patel B, Heding KE, Bjelke JR, et al. Design of biopharmaceutical formulations accelerated by machine learning. Mol Pharm. 2021;18(10):3843-53.

9. Cloutier TK, Sudrik C, Mody N, Sathish HA, Trout BL. Machine learning models of antibody–excipient preferential interactions for use in computational formulation design. Mol Pharm. 2020;17(9):3589-99.
10. Herve Q, Ipek N, Verwaeren J, De Beer T. Automated particle inspection of continuously freeze-dried products using computer vision. Int J Pharm. 2024;664:124629.
11. Juckers A, Knerr P, Harms F, Strube J. Model-based product temperature and endpoint determination in primary drying of lyophilization processes. Pharmaceutics. 2022;14(4):809.
12. Vanbillemont B, Nicolaï N, Leys L, De Beer T. Model-based optimisation and control strategy for the primary drying phase of a lyophilisation process. Pharmaceutics. 2020;12(2):181.
13. Klepzig LS, Juckers A, Knerr P, Harms F, Strube J. Digital twin for lyophilization by process modeling in manufacturing of biologics. Processes. 2020;8(10):1325.
14. Juckers A, Knerr P, Harms F, Strube J. Effect of the freezing step on primary drying experiments and simulation of lyophilization processes. Processes. 2023;11(5):1404.
15. Srinivasan JM, Sacha GA, Nail SL. The graphical design space for the primary drying phase of freeze drying: factors affecting the dried product layer resistance. Int J Pharm. 2023;630:122417.
16. Clavaud M, Lema-Martinez C, Roggo Y, Bigalke M, Guillemain A, Hubert P, et al. Near-infrared spectroscopy to determine residual moisture in freeze-dried products: model generation by statistical design of experiments. J Pharm Sci. 2020;109(1):719-29.
17. Fang R, Bogner RH, Nail SL, Pikal MJ. Stability of freeze-dried protein formulations: contributions of ice nucleation temperature and residence time in the freeze-concentrate. J Pharm Sci. 2020;109(6):1896-904.
18. Wang J, Searles JA, Torres E, Tchessalov SA, Young AL. Impact of annealing and controlled ice nucleation on properties of a lyophilized 50 mg/ml MAB formulation. J Pharm Sci. 2022;111(9):2639-44.
19. Herve Q, Ipek N, Verwaeren J, De Beer T. A deep learning approach to perform defect classification of freeze-dried product. Int J Pharm. 2025;670:125127.
20. Vanbillemont B, Greiner AL, Ehrl V, Menzen T, Friess W, Hawe A. A model-based optimization strategy to achieve fast and robust freeze-drying cycles. Int J Pharm X. 2023;5:100180.
21. Vidal-Henriquez E, Holder T, Lee NF, Pompe C, Teese MG. Machine learning driven acceleration of biopharmaceutical formulation development using Excipient Prediction Software (ExPreSo). Comput Struct Biotechnol J. 2025;27:4517-25.
22. Groël S, Menzen T, Winter G. Prediction of unwanted crystallization of freeze-dried protein formulations using α-relaxation measurements. Pharmaceutics. 2023;15(2):703.
23. Gupta D, Biswas AA, Sahu RC, Arora S, Kumar D, Agrawal AK. Advancing pharmaceutical intelligence via computationally prognosticating the in-vitro parameters of fast disintegration tablets using machine learning models. Eur J Pharm Biopharm. 2024;204:114508.
24. Geidobler R, Winter G. Controlled ice nucleation in the field of freeze-drying: fundamentals and technology review. Eur J Pharm Biopharm. 2013;85(2):214-22.
25. Tchessalov S. Practical advice on scientific design of freeze-drying process. Pharmaceutics. 2023;15(11):2547.
26. Depaz RA, Pansare SK, Patel SM. Freeze-drying above the glass transition temperature in amorphous protein formulations while maintaining product quality and improving process efficiency. J Pharm Sci. 2016;105(1):40-9.
27. Pansare SK, Patel SM. Lyophilization process design and development: a single-step drying approach. J Pharm Sci. 2019;108(4):1423-33.
28. Ohori R, Akita T, Yamashita C. Effect of temperature ramp rate during the primary drying process on the properties of amorphous-based lyophilized cake, Part 2: Successful lyophilization by adopting a fast ramp rate during primary drying in protein formulations. Eur J Pharm Biopharm. 2018;130:83-95.
29. Mizogaki I, Suzuki T, Ohori R, Sano S, Hara S, Miyazaki T, et al. Evaluating the impact of lyophilization process parameters on mRNA encapsulated lipid nanoparticles using machine learning. J Drug Deliv Sci Technol. 2025;107573.
30. Chitre A, Semochkina D, Woods DC, Lapkin AA. Machine learning-guided space-filling designs for high throughput liquid formulation development. Comput Chem Eng. 2025;195:109007.
31. Müller P, Sack A, Dümler J, Heckel M, Wenzel T, Siegert T, et al. Automated tomographic assessment of structural defects of freeze-dried pharmaceuticals. AAPS PharmSciTech. 2024;25(6):143.
32. Tchessalov S, Latshaw II D, Nulu S, Bentley M, Tharp T, Ewan S, et al. Application of first principles primary drying model to lyophilization process design and transfer: case studies from the industry. J Pharm Sci. 2021;110(2):968-81.