# Automation and Active Learning for the Multi-Objective Optimization of Antibody Formulations

*D. Christopher Radford*[1,#], *Matthew Tamasi*[1,#], *Elena Di Mare*[1], *Adam J. Gormley*[1]

[1] Department of Biomedical Engineering, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA

**KEYWORDS:** Bayesian optimization, Biologics, GRAS excipients, Machine learning, Protein formulation

## ABSTRACT

Over the last forty years, biologics such as monoclonal antibodies have become an increasingly important therapeutic agent in the treatment of numerous diseases. Between 1986 and 2025, over 200 antibody-based treatments have been approved globally, most of which are manufactured as preformulated solutions for subsequent administration to patients. However, bioformulation of complex proteins is a difficult engineering challenge; formulations must be tailored to individual therapies, necessitating time- and material-intensive campaigns to select a combination of excipients to simultaneously optimize an array of design criteria. These many interacting additives complicate formulation design with unintuitive and non-linear relationships, thus creating a vast and multidimensional design space that is intrinsically difficult to optimize using traditional techniques. To address this challenge, we investigated a high-throughput discovery pipeline using machine learning to model and predict formulation behavior of Generally Recognized as Safe (GRAS) excipients on a model antibody. This was supported by automation-assisted “on-demand” formulation to produce dozens of uniquely formulated antibody solutions with high reproducibility for downstream evaluation and biophysical characterization. This pipeline was then integrated into an iterative closed-loop cycle of automated Design-Build-Test-Learn (DBTL), where new rounds of experiments are designed by the ML model. The process yielded both optimized formulations as well as highly accurate predictive models of formulation behavior. This validates the utility of this technique to both map the underlying property-function landscape and effectively guide formulation development while balancing multiple competing design requirements.

## 1. INTRODUCTION

Antibody-based therapeutics are an increasingly important part of the clinical armamentarium.[1] The first monoclonal antibody treatment (Muromonab) was approved by the FDA for the US market in 1986.[2] Since then, over 200 antibody-based therapies have been approved globally and an additional 178 are currently in late-stage clinical trials as of December 2024.[3] These represent numerous key breakthroughs and serve as frontline therapies in the fields of oncology, immunology, and infectious disease.[4, 5] Antibody-based therapies are typically administered via intravenous or subcutaneous routes.[6] As such, the majority are manufactured and provided to patients or providers as pre-formulated solutions rather than in solid dosage forms.[7] However, the process of formulating biologics such as antibodies presents numerous challenges. As large, complex macromolecules, antibodies are prone to denaturation and aggregation.[8] These leave a fraction of the product in an inactive form, reducing the effective dose the patient receives. Failure modes can also alter epitope presentation, presenting an immunogenic risk to the patient.[9] As such, formulated antibodies must exhibit strong thermal and colloidal stability to ensure the active agent maintains a reasonable shelf life. Additional considerations might also arise based on the delivery route. For example, subcutaneous administrations inherently limit dosage volume, necessitating formulation at high antibody concentrations (typically >100 mg/mL) to achieve the requisite therapeutic dose.[10] At these concentrations, where amphiphilic proteins now occupy substantial volume, intermolecular interactions dominate and drive exponential increases in solution viscosity and poor injectability.[11-13] This necessitates a formulation strategy that explicitly accounts for viscosity properties of the solution.

Clinical antibody formulations typically rely on combinations of so called “Generally Recognized as Safe” (GRAS) excipients to achieve favorable properties while ensuring safety.[7] This class of excipients includes components used to maintain pH (buffers), osmolarity (salts and sugars), viscosity (amino acids), and solubility (surfactants).[14] However, combinations of excipients are reported to have unintuitive, non-linear, and interactive effects.[15-18] This complicates the ability to map the underlying property-function relationships, intuitively predict performance, and efficiently select formulations via a rational design-based approach. This is further complicated by the need for formulations to simultaneously satisfy multiple design criteria (e.g., thermal and colloidal stability, low viscosity, etc.), where excipients beneficial to one property might degrade performance of another. For example, high antibody charge density (influenced by formulation pH) can both improve colloidal stability (by increasing repulsive forces between individual macromolecules) and reduce thermal stability (inducing unfolding due to increased intramolecular repulsions).[16] This, in turn, leads to cross-pressures for individual formulation decisions driven by competing objectives. Therefore, successful antibody formulations must include combinations of excipients that balance these trade-offs between metrics to have acceptable overall performance.

Furthermore, differences in physiochemical properties between individual antibodies limits the ability to extrapolate formulation performance from one to another.[19-21] This lack of a “one-size-fits-all” approach is observed in the diverse array formulations utilized across the current landscape of approved antibody-based therapies.[7] As such, formulations often need to be tailored for that specific therapy. This becomes even more acute for complex antibody-based derivative biologics such as bi-specifics and antibody-drug conjugates, which can present novel stability challenges or introduce new formulation requirements (e.g., drug linker stability).[22] As such, *de novo* formulation development for a novel antibody has historically relied on human experience and training, rational design, and high-throughput screening, which has in turn necessitated time- and material-intensive formulation campaigns.[23]

In contrast, recent efforts have demonstrated the potential for machine learning (ML) to effectively navigate these types of complex, multidimensional design spaces and map the underlying property-function relationships.[24, 25] These are often paired with techniques such as Bayesian optimization (BO) where a surrogate model representing these relationships is used to suggest designs for subsequent experimentation.[26] Significant success has been seen in the fields of medicinal chemistry[27-29], pharmaceutics[30-32], and biotechnology[33-36], where BO has been applied to guide the development of therapies. In parallel, automation platforms have become increasingly important engines for drug discovery[37, 38] and material design[39, 40], drastically increasing experimental throughput while also improving accuracy and precision. Previously, we successfully applied these tools for the development of novel polymer excipients for protein and drug stabilization.[41-43] Likewise, the pairing these technologies is well-suited to accelerate the process of formulation discovery.

Narayanan, et al.[44] previously reported the feasibility of applying BO for formulating biologic drugs, utilizing ML to simultaneously optimize thermal and interfacial stability. Herein, we demonstrate the potential of this approach to be further accelerated with automation, while extending the scope of the multi-objective optimization problem to higher-dimensional objective spaces. This, in turn, establishes the scalability of this optimization strategy to significantly more complex formulation challenges faced in clinical product development. Towards this end, we developed an automation-assisted formulation discovery pipeline driven by BO. Liquid handling robots were used for on-demand, low-volume, high-throughput formulation in 96 well plate formats. This was paired with active learning to intelligently guide experimentation and testing toward high-value formulations within the vast design space. Through multiple rounds of active learning, we evaluated the ability of this system to simultaneously optimize parameters associated with thermal and colloidal stability, alongside high concentration viscosity. Collectively, this demonstrates the potential of this strategy to leverage modest, experimentally feasible datasets to efficiently map a complex formulation design space and navigate to optimized designs.

## 2. RESULTS AND DISCUSSION

### 2.1 Data-driven discovery pipeline for antibody formulation

To evaluate a machine learning driven approach for antibody formulation, we iterated formulation designs leveraging a Build-Test-Learn-Design (DBTL)[45] cycle (Figure 1) to optimize a model bovine immunoglobulin G (bIgG) antibody across three objectives: maximization of diffusivity, maximization of thermal stability ($T_m$), and minimization of viscosity. These objectives were chosen in-line with needs for antibodies which must remain colloidally and thermally stable while retaining low viscosity when highly concentrated. Each cycle consisted of four primary steps: 1) automation-assisted formulation and sample preparation; 2) performing biophysical characterization measurements; 3) training machine learning models to predict $T_m$, diffusivity, and viscosity directly from formulation features; 4) proposing 24 new formulations to optimize target objectives using active machine learning. These newly proposed candidates were then formulated, characterized, and provided as training data to update models and begin the cycle again. Our approach here utilized two complete cycles to demonstrate the feasibility of improving each objective.

**Initial Seed Library & Formulation Space:** To generate initial data for our DBTL cycles, we performed an initial sampling of our formulation space. Clinical antibody formulations typically draw from a diverse combination of excipients to alter intermolecular interactions and enhance biophysical properties of formulations.[7] We attempted to capture this complexity within a streamlined formulation design space of GRAS excipients (Supplementary Table S1). Each formulation consists of a combination of buffer (defined by buffering system and pH), L-arginine, sucrose, and NaCl, each at an independently variable concentration within the specified range. Ranges for each excipient were selected to ensure typical concentrations used for clinical formations were accessible within the boundaries. Even with this streamlined design space, this yielded ~800,000,000 unique formulations within a high-dimensional design space when limiting continuous parameters to a step size of ±1% of their respective range. Exploring this design space at high resolution is experimentally intractable, necessitating selection of a subset of formulations that can cover the space and provide downstream machine learning models with a strong foundational information dataset to guide subsequent active learning.[46] Towards this end, 24 formulations were selected by Latin hypercube sampling (LHS)[47], a design of experiments (DOE) approach that efficiently samples the six-dimensional space. This seed library provided a modest, synthetically feasible number of formulations that nevertheless provided a diverse array of excipient combinations.

**Build:** The proposed 24 formulations were then prepared via automation assisted on-demand formulation using a Hamilton MLSTARlet liquid handing robot. Importantly, this approach facilitated accurate, reproducible, combinatorial excipient mixing in a high-throughput 96 well plate format. This provided sufficient experimental material for downstream characterization while simultaneously minimizing antibody consumption (approximately 10 mg of antibody for $T_m$ and diffusivity studies and 85 mg for viscosity study per formulation replicate).

**Test:** After automation-assisted preparation in 96-well plates, formulation $T_m$, diffusivity, and

https://doi.org/10.26434/chemrxiv-2025-rxq1t **ORCID:** https://orcid.org/0000-0002-2884-725X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

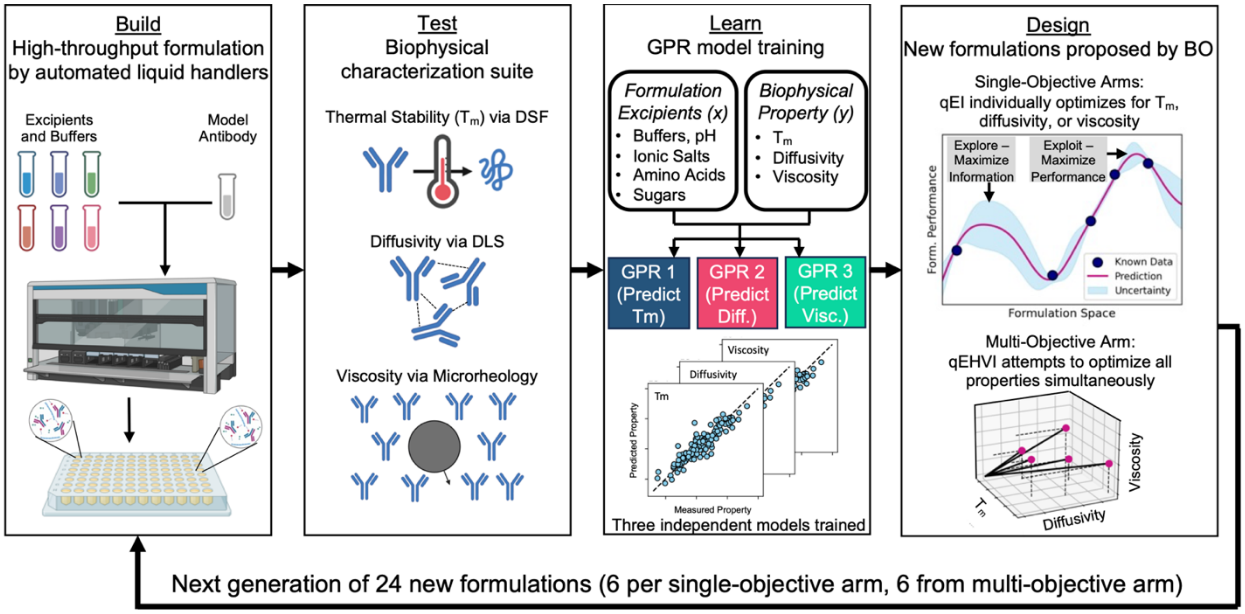
Build
High-throughput formulation by automated liquid handlers
Excipients and Buffers
Model Antibody
Test
Biophysical characterization suite
Thermal Stability (Tm) via DSF
Diffusivity via DLS
Viscosity via Microrheology
Learn
GPR model training
Formulation Excipients (x)
• Buffers, pH
• Ionic Salts
• Amino Acids
• Sugars
Biophysical Property (y)
• Tm
• Diffusivity
• Viscosity
GPR 1 (Predict Tm)
GPR 2 (Predict Diff.)
GPR 3 (Predict Visc.)
Viscosity
Diffusivity
Tm
Predicted Property
Measured Property
Three independent models trained
Design
New formulations proposed by BO
Single-Objective Arms: qEI individually optimizes for Tm, diffusivity, or viscosity
Explore – Maximize Information
Exploit – Maximize Performance
Form. Performance
Known Data
Prediction
Uncertainty
Formulation Space
Multi-Objective Arm: qEHVI attempts to optimize all properties simultaneously
Tm
Diffusivity
Viscosity
Next generation of 24 new formulations (6 per single-objective arm, 6 from multi-objective arm)

**Figure 1. Overview of study.** Build) Automated-assisted formulation of bIgG by a Hamilton MLSTARlet liquid handling robot in 96-well plates. Test) Biophysical characterization of bIgG formulations is performed to acquire thermal stability ($T_m$) by differential scanning fluorimetry (DSF), diffusivity by dynamic light scattering (DLS), and viscosity data by high-throughput microrheology. Learn) After initialization with seed data, objective-specific Gaussian process regressors (GPRs) surrogate models are trained to predict $T_m$, Diffusion, and Viscosity of a given formulation. Design) Models are paired with expected improvement acquisition functions to optimize formulations for single and multiple objectives.

viscosity were then collected through a suite of high-throughput characterization techniques. Differential scanning fluorometry (DSF) was used to collect information on thermal stability of the formulation by quantifying its $T_m$ at formulation concentrations. In parallel, dynamic light scattering (DLS) was used to determine diffusivity of the antibody in solution to obtain insight into changes in intermolecular interactions and colloidal stability. Finally, DLS was also utilized to capture viscosity at high antibody concentrations (120 mg/mL) utilizing micro-rheology techniques. Experimental details of all techniques are described in Section 4.

**Learn:** To efficiently predict the diffusivity, thermal stability, and viscosity of new formulations, we trained three independent Gaussian process regressors (GPRs)[48] directly from formulation feature vectors (see Section 4). Models provided forward-looking predictions (μ) and uncertainty estimates (σ) of formulation behavior on any combinations of excipients and buffers in our streamlined design space.

**Design:** Trained GPRs were then used to launch four independent BO campaigns, each targeting a different objective for formulation improvement. Three of the campaigns aimed to optimize a single biophysical parameter without regard for the other two (i.e., $T_m$ only). In contrast, the fourth arm attempted to consider all three parameters in order to select balanced formulations across multiple objectives. For each objective, six novel formulations that had not previously been evaluated were proposed by maximizing single objective or multi-objective expected improvement (EI) acquisition functions (see Section 4). This batch of formulations was suggested by the function to facilitate two goals: 1) leveraging the model's understanding to select highly-performant formulations, and 2) evaluating areas of the design space where model uncertainty is elevated, thus providing additional data to improve the model's understanding.

### 2.2 ML models guided selection of improved formulations via Bayesian optimization

The 24 seed formulations proposed by LHS successfully produced formulations with a wide range of melt temperatures, diffusivities, and viscosities across our set of formulation parameters. Across all four optimization arms, BO utilizing surrogate GPR models yielded new generations of candidates whose target properties significantly improved on the seed library (Figure 2A). With only two rounds of active learning, the diffusion campaign increased the mean diffusivity by 22.2%, the viscosity campaign lowered mean viscosity by 13.7%, and $T_m$ campaign raised mean melt temperature by 1.43°C, each change reaching statistical significance ($p < 0.005$). The expected-improvement acquisition functions utilized in our approach (qEI[49] for single objectives, qEHVI for HV) explicitly balance exploitation of high performance regions with exploration of high uncertainty regions; the broad inter-quartile ranges of the dashed-box "predicted" distributions in Figure 2A confirm that both facets were represented in every 6-formulation batch. Importantly, utilizing this approach, each arm was also able to identify individual high-performing formulations with properties that exceed those in the seed library. Specifically, lead candidates from the $T_m$, diffusivity, and viscosity-optimizing arms were able to improve their target objective by 0.75%, 4.64%, and -7.67%, respectively, over the best-performing seed formulation.

The multi-objective optimization (MOO) arm, steered by the q-expected hyper-volume improvement (qEHVI)[49], registered the largest gain; a 73.3% gain in mean normalized hyper-volume (HV), illustrating that the joint surrogate captured the Pareto structure of the problem from the initial 24 formulations in the seed dataset. Furthermore, a lead candidate was identified that improved the metric by 21.9% over the best seed formulation. As hypervolume represents the product of the three scaled objectives, the qEHVI function attempts to balance performance between them. To achieve this, the MOO explored multiple optimization strategies, selecting mutually non-dominated formulations along the pareto front. In practice, this led to diverse formulation designs that prioritized one or two properties while attempting to limit performance losses in the remainder. These strategies yielded varying levels of success. When biophysical properties of the MOO formulations were characterized and used to calculate associated HVs, the batch displayed inconsistent performance, leading to a wide range of HVs across the candidates. Interestingly, deeper inspection reveals that the composite metric preferentially favored the diffusion/viscosity axis at the expense of thermal stability. This bias emerges as formulations optimized for diffusion and viscosity demonstrated strong cross-over performance in both objectives (Supplementary Figure S1). qEHVI therefore rewarded candidate formulations that drove the dual colloidal objectives aggressively while maintaining moderate thermal stability, yielding HV improvements. This result suggests the need to articulate explicit objective floors in future multi-objective campaigns.

Interestingly, while the first generation of BO candidates yielded significant improvement over the seed library, minimal further improvement was observed after the second round of BO. For diffusion and viscosity, Gen 2 medians were statistically indistinguishable from Gen 1, and no

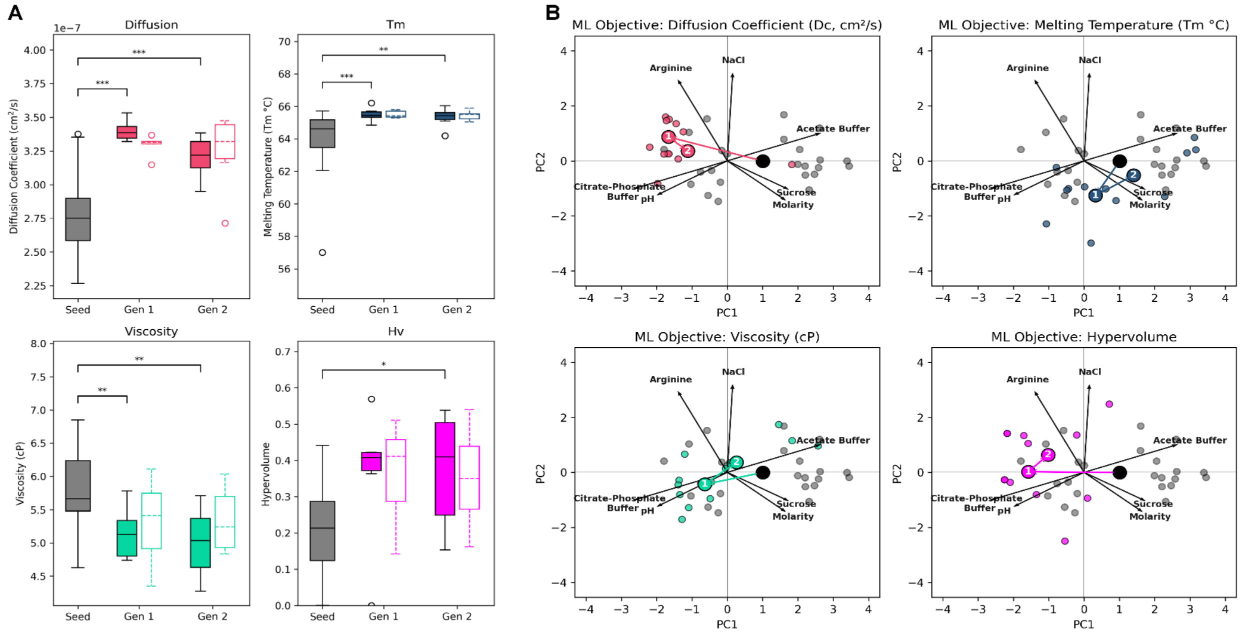

**Figure 2. ML guides the design of bIgG formulation based on optimization objective.** A) Measurements (solid plots) and predictions (dashed plots) acquired for formulation $T_m$, $D_c$, viscosity, and hypervolume (HV) after a single round (Gen 1) and two rounds (Gen 2) of active learning loops compared to seed formulations. Formulation predictions (μ) were determined from the GPR used by BO to propose that batch. Statistical significance was assessed between measured values for seed formulations and ML proposed formulations and determined by single tailed t-test. *($p < 0.05$), **($p < 0.005$), ***($p < 0.0005$), unlabeled pairs are not significantly different. B) Principal component analysis (PCA) of ML guided bIgG formulations based on optimization objective demonstrating unique pathing through formulation space. Large numbered points are calculated centroids of proposed samples (colored points) from Gen 1 and Gen 2 formulations for each objective. All instances originate from the centroid of the Seed formulations (grey points).

Gen 2 formulation exceeded the best Gen 1 candidate in any objective besides viscosity (Figure 2A). *In silico* projections of a hypothetical Gen 3 likewise failed to predict further improvement (Supplementary Figure S2), indicating that the exploitable region of design space was largely exhausted. This rapid saturation is characteristic of BO in smoothly varying physicochemical spaces: once qEI has exploited the principal gradient, remaining unexplored niches confer diminishing marginal utility. This observation suggests that one or two DBTL rounds may suffice to reach near-optimal performance, greatly reducing experimental burden relative to classical factorial and high-throughput screens.

To better understand how the model optimized each of these target features with respect to the underlying formulation parameters, principal component analysis (PCA) was used to visualize the locations of formulations within the design space (Figure 2B). Centroids were calculated for each generation to understand pathing through the design space by BO for each of the four respective optimization goals. Interestingly, each arm of the BO led to proposed formulations centered in distinct formulation regions. For example, optimization of diffusivity yielded formulations largely in Quadrant II of design space, which represent formulations with high arginine, low sucrose, and low osmolarity. In contrast, the $T_m$ arm proposed formulations in Quadrant IV, representing low arginine and high sucrose. The fact that diffusivity-optimized and $T_m$-optimized centroids identified by BO occupy opposite regions of the design space demonstrates the inherent tradeoff when optimizing both parameters simultaneously via MOO. The multi-objective optimization arm yielded formulations with the largest spread throughout the design space, with individual candidates in each of the four quadrants. This likely arises from the multiple strategies explored by the MOO to balance the three competing objectives, as well as the higher uncertainty inherent in the hypervolume parameter. However, the majority of the MOO candidates are in Quadrant II and III, which reflects the dominant strategy of favoring candidates with better diffusivity and viscosity.

### 2.3 Exploring property-function relationships of individual formulation components

As objective-specific BO was generally successful in identifying generation-on-generation improvements in formulation behavior, we sought to understand the formulation designs that underlie objective-specific optimizations. Explainable artificial intelligence (xAI) techniques are essential for building trust and ensuring that models make predictions based on sound formulation principles rather than spurious correlations.[50] Using SHAP (Shapley Additive Explanations)[51, 52], we probed the objective-specific GPR models for $T_m$, diffusivity, and viscosity to validate their reasoning and extract formulation insights. Here, positive SHAP values indicate positive contributions to $T_m$, diffusivity, and viscosity, while negative SHAP values suggest negative contributions, and we use the mean absolute SHAP value of a feature as a proxy for its overall

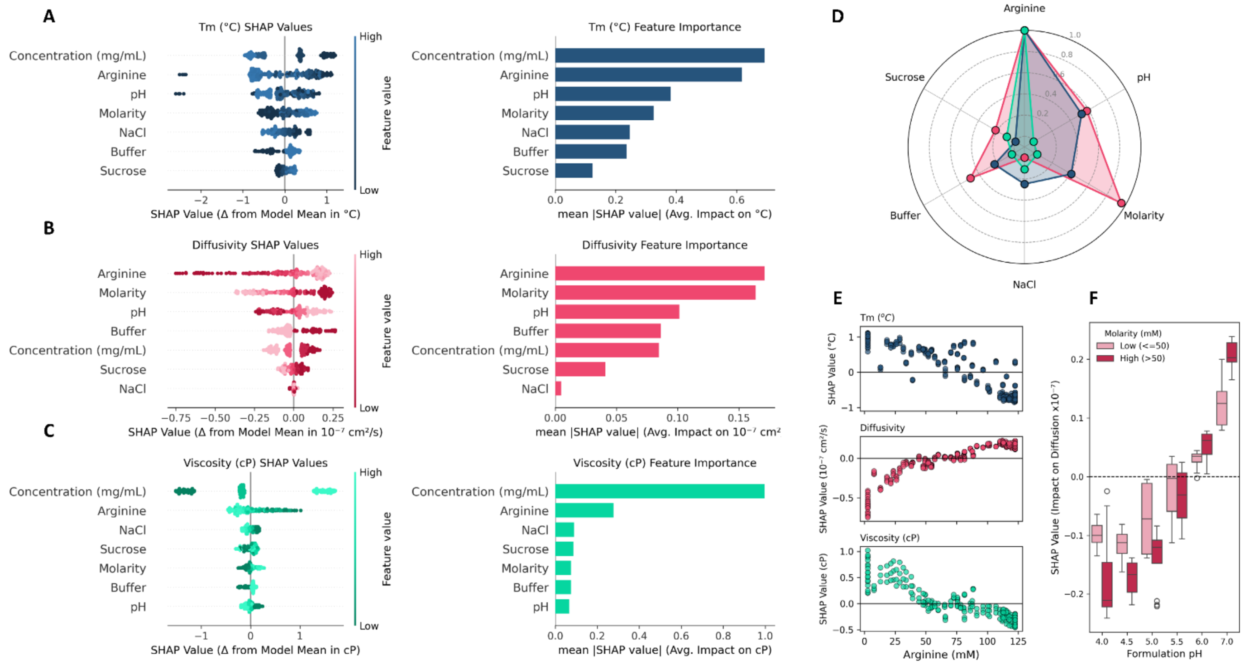

**Figure 3. xAI analysis reveals distinct priorities in formulation composition for each objective.** A-C) Summary of SHAP values for GPR models calculated from available data after all rounds of active ML cycles. (Left) Each point corresponds to a uniquely evaluated formulation, and the point's position along the X-axis shows the impact of a feature on predicted $T_m$, diffusivity, and viscosity respectively. (Right) SHAP derived feature importance values across all model features for $T_m$, diffusivity, and viscosity respectively. SHAP derived feature importance are taken as the mean(|SHAP value|) to demonstrate the average impact of a feature on model predictions for each target. D) Normalized mean absolute SHAP values calculated for $T_m$, diffusivity, and viscosity for each model to quantify relative feature importance. E) $T_m$, Diffusivity, and viscosity SHAP values for arginine content. F) Diffusivity SHAP values for formulation pH.

importance to model predictions. In order to evaluate these trends over the entire dataset, SHAP values were computed for all 72 formulations across all three single objective models (Figure 3A-C). This summary analysis reveals key drivers of formulation performance for each objective. Comparing SHAP summary plots across $T_m$, diffusivity, and viscosity we observe trends that are largely in line with reported literature.[17, 53] Diffusivity is impacted most by arginine and charge state of the antibody (which is largely controlled by buffer parameters). Arginine is well known to reduce attractive interactions in protein formulations by transiently interacting with hydrophobic regions on the protein and increasing net repulsive forces, reducing self-association and improving diffusivity.[54] While the "ideal" buffer pH will vary significantly from antibody to antibody based on the isoelectric point (pI), the relatively acidic pI of the model bIgG antibody used herein leads to improved diffusivity with increasing pH, as this will increase the net repulsive forces between individual antibodies in solution. Interestingly, a negative correlation is observed between buffer molarity and diffusivity, which is likely due to the fact that buffering salts may also produce charge-shielding effects, reducing the strength of intermolecular repulsive forces.[55] Additionally, we observe that SHAP values for viscosity are dominated by antibody concentration and arginine content. As with diffusivity, arginine's ability to reduce intermolecular interactions – particularly hydrophobic interactions that increasingly dominate at higher protein concentrations where macromolecules are forced into closer proximity – contributes to its ability to significantly reduce viscosity of formulated antibody.[54] Interestingly, $T_m$ SHAP analysis suggests concentration, arginine, and pH as the most impactful features. While classically, sucrose might be anticipated to be the largest driver of mAb thermal stability due to its ability to modify the protein hydration layer[53], SHAP analysis reveals that the model bIgG is likely highly charge-sensitive requiring careful balance of pH and ionic shielding. This protein-specific identification of dominate formulation features is a primary benefit of utilizing explainable AI approaches.

Furthermore, through SHAP analysis, nuanced biophysical phenomena can be observed. While diffusivity improves with increasing pH due to the antibody's low pI, this effect is enhanced with increasing buffer molarity (Figure 3F). This is likely due to the fact that the antibody itself as well as other formulation components (i.e., arginine) can also influence overall pH of the solution. As such, increasing molarity of the core buffering agent allows it to dominate. In contrast, low pH buffers see a negative influence of buffer molarity on diffusivity. This is likely due to the increasingly dominating effect of the buffer driving the pH of the system closer to the pI of the antibody[17], as well as the increased charge shielding effects inherent with higher salt content[55], both of which is collectively increase potential for attractive intermolecular interactions and decrease diffusivity.

When these mean absolute SHAP values are scaled relative to arginine (Figure 3D), its overarching dominance across all three objectives is readily apparent. While the diffusion model assigns a similar weight to buffer molarity (~0.95), arginine far surpasses all other features like pH (~0.6) and excipients such as sucrose (~0.2) or NaCl (<0.1). Weights are even more lopsided in favor of arginine for the $T_m$ and diffusion models. This positions arginine as a key excipient in our formulation space and with regard to our model bIgG antibody. However, arginine demonstrates both synergistic and antagonistic trade-offs depending on a given optimization objective (Figure 3E). For example, increasing arginine concentration correlates with increasingly positive SHAP values for diffusivity (middle panel) and increasingly negative SHAP values for viscosity (lower panel), promoting favorable diffusivity and rheological properties simultaneously. However, this comes at the expense of $T_m$ depression (upper panel), where similar increases in arginine concentration are predicted to decrease the thermal stability of bIgG significantly. These findings by SHAP largely explain the results of our two DBTL cycles in which formulations that were optimized for diffusion or viscosity generally performed well in both objectives, but often demonstrated low thermal stability (Supplementary Figure S1). In contrast, formulations that were optimized for $T_m$ generally demonstrated lower diffusivity and higher viscosities. In the multi-objective hypervolume (HV) arm of the study, most designs intended to maximize our qEHVI acquisition function localized to an area of PCA space similar to the diffusion- and viscosity-optimized formulations (Figure 1B). This was a synergistic design space where high performance for two objectives could be simultaneously obtained. In contrast, the distant high-$T_m$ space satisfied high-performance for only a single objective, leading to the MOO arm clearly biasing formulation selection away from this area.

### 2.4 Post-hoc analysis of model learning

After two DBTL cycles, each GPR model trained on the full set of 72 formulations achieved high predictive fidelity (Supplementary Figure S3). The viscosity model demonstrated the highest accuracy of $R^2 = 0.87$, while the diffusion and $T_m$ models demonstrated $R^2$ values of 0.80 and 0.70 respectively. This performance underscores the capability of GPRs to model the complex, nonlinear structure-activity relationships inherent to antibody formulations. To better understand this learning capacity, the learning rate of the three models – their ability to translate additional data into improved model performance – was assessed (Figure 4A). Here, to mimic the experimental design of the active learning, a modified leave-one-out cross validation approach (LOOCV) was implemented. The “left out” formulation was predicted using a model progressively trained on more data – from the initial seed library (24 formulations) to the entire library minus the held-out formulation (71 formulations). By repeating this process and holding out a different formulation from Generation 1 or 2 each time, the entire non-seed library (48 formulations) was predicted and aggregated to compute overall $R^2$ and mean absolute error (MAE) for each model as a function of training set size.

Impressively, even using solely the seed library for training, the $T_m$ and diffusivity models demonstrated reasonable $R^2$ values of 0.68 and 0.59, respectively, while the viscosity model demonstrated a strong $R^2$ of 0.80. The latter is likely due to the uniquely dominating influence of antibody concentration on viscosity vis-a-vis other formulation features (Figure 3C), enabling the model to make reasonably accurate predictions without needing to fully understand the nuances of the formulation itself. As the models receive additional formulation data in their training set, these performance metrics consistently increase. When given the full training set, $R^2$ scores for all models exceed 0.82. Alongside improvements in $R^2$ scores, error is significantly reduced for all GPRs as more data is provided to the model, with 35.6%, 43.4%, and 20.8% reductions in MAE for the $T_m$, diffusivity, and viscosity models, respectively. Notably, the final MAE for $T_m$ (0.50 °C) and viscosity (0.35 cP) is comparable to the average experimental error (0.48 °C and 0.34 cP, respectively), suggesting the models are approaching the noise floor of the measurements. In contrast, the diffusivity MAE ($1.23 \times 10^{-8}$ cm²/s) remains higher than the

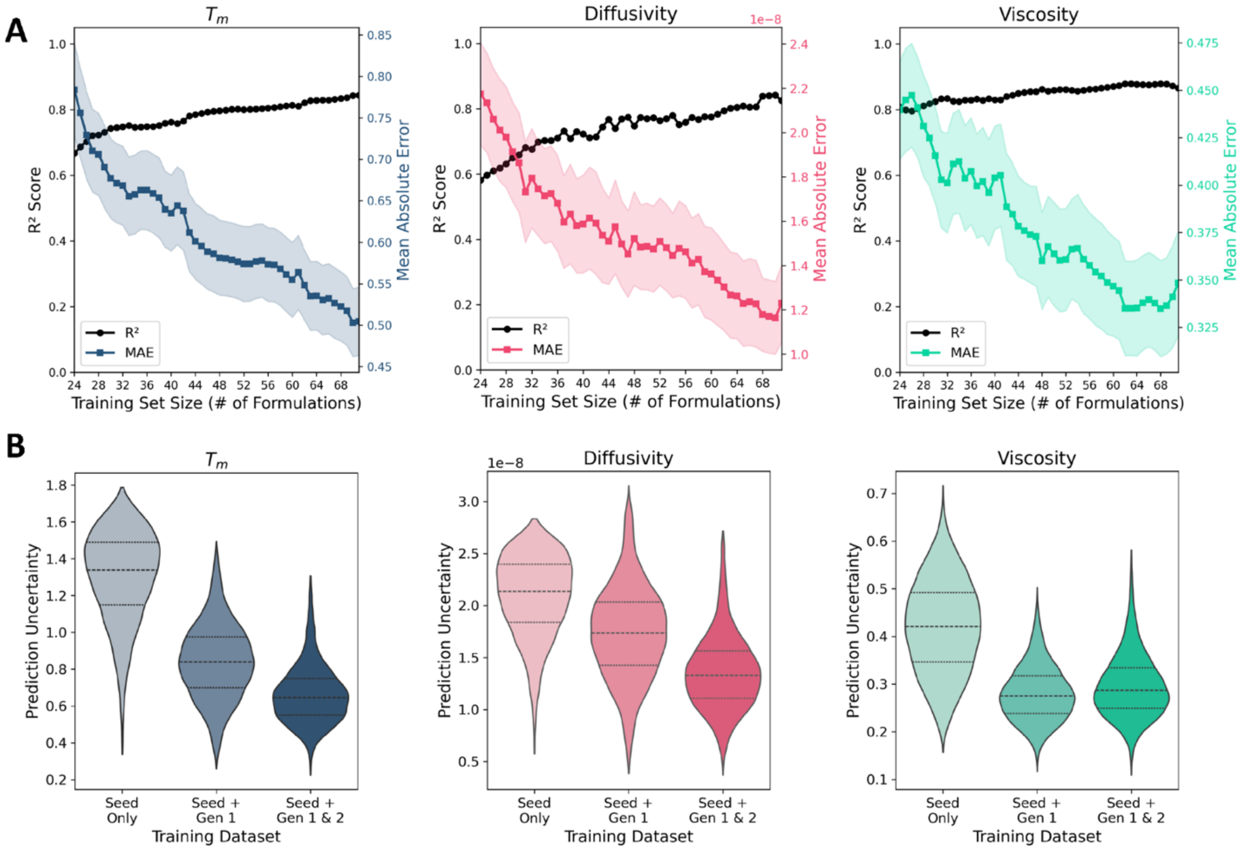

**Figure 4. ML models demonstrate high accuracy and prediction confidence in low data regimes.** A) Learning rate analysis for the three models, tracking improvement in predictive accuracy in response to additional training data. MAE mean is presented with standard error of values across the 48 arms of LOOCV. B) Gen-on-gen improvement in model uncertainty (standard deviation of posterior probability density function) when predicting over the entire design space (represented by a survey of 10,000 formulations *in silico*).

experimental error ($6.83 \times 10^{-9}$ cm²/s), likely reflecting the exceptionally low relative noise of that technique. Overall model error reflects both aleatoric (intrinsic) and epistemic (model-related) uncertainty.[56] As such, convergence of model MAE to the experimental noise indicates that the models have largely captured the learnable patterns in the data, with residual error dominated by irreducible measurement noise.

Ultimately, the value of these models lies in their ability to make accurate and confident predictions for novel formulations across the entire design space. To assess this, we quantified the generation-on-generation reduction in global model uncertainty by predicting the properties of 10,000 formulations *in silico* (Figure 4B). The analysis revealed a significant decrease in the mean predictive uncertainty (the standard deviation of the GPR posterior) for all three objectives: 49.1%

for $T_m$, 34.9% for diffusivity, and 29.5% for viscosity over the course of the campaign (Supplementary Table S2). This global reduction demonstrates that the active learning strategy, by balancing exploration of high-uncertainty regions with exploitation of high-performance regions, successfully builds a more confident and generalizable model of the entire formulation landscape. As the model receives more data, its probability density function (PDF) shifts. Ideally, predictions for novel formulations outside of the training data should not only become more accurate, but also more confident as the model develops a more precise understanding of the relationship between the design features and target objective values (Supplementary Figure S4). Further, this improvement is not merely localized to the sampled points but is propagated across the design space via the GPR's covariance kernel, validating the model's utility for future *in silico* screening and optimization. The observed plateau in uncertainty reduction for the viscosity model after the first generation is consistent with its rapid convergence to the noise floor, indicating that it reached a state of diminishing returns on additional data sooner than the other models.

## 3. CONCLUSION

Efficient optimization of antibody formulations remains a critical bottleneck in biologics development. Next-generation antibody therapies often require concentrations ≥100 mg/mL for subcutaneous delivery, but at such concentrations antibodies can exhibit extreme viscosities and stability issues, creating formidable hurdles in development, manufacturing, and administration. Conventional formulation campaigns rely on trial-and-error excipient screening to mitigate aggregation and viscosity, yet the complexity of protein-excipient interactions makes this process time-consuming and inefficient. Identifying a stable, low-viscosity formulation can demand extensive experimentation and material, slowing the path to the clinic and straining production resources. These challenges underscore the need for more data-driven, accelerated formulation design strategies.

In this work, we demonstrate a machine learning- and automation-assisted pipeline to streamline antibody formulation development. The platform implements a Design-Build-Test-Learn framework, integrating robotics-assisted formulation preparation, high-throughput biophysical characterization, and GPR surrogate models to navigate the antibody formulation design space. Over two iterative DBTL rounds guided by BO, the system was able to successfully target key formulation objectives: raising thermal unfolding temperature, increasing protein diffusivity, and lowering viscosity. Additionally, the multi-objective optimization acquisition function used in this approach demonstrated the ability to optimize these parameters simultaneously, intelligently negotiating between trade-offs to maximize overall performance. Remarkably, robust surrogate models trained on only 72 formulations captured the underlying property-function relationships. This outcome highlights the power of active learning to rapidly converge on optimal formulations from modest datasets, echoing recent evidence that ML-guided experimentation can dramatically improve data efficiency in pharmaceutical development. Beyond performance gains, a pivotal advance of this study lies in moving beyond prediction to mechanistic understanding through the application of explainable AI. By leveraging SHAP, we deconstructed the GPR model to assign a quantitative contribution of each excipient to each training feature, thereby elucidating the distinct synergistic and antagonistic roles of specific formulation components. These mechanistic insights validate known biophysical trends (for example, arginine-mediated suppression of self-association) and illuminate the complex interplay of factors in a multi-objective formulation. Such explainable models move beyond black-box predictions, providing a rational foundation to guide excipient choices and balance trade-offs in antibody formulation.

Looking forward, this data-efficient approach, guided by machine learning and automation could be further expanded to broader excipient spaces and applied across multiple antibody candidates to derive generalizable formulation design rules. This strategy could also be extended to ultra-high protein concentrations (>200 mg/mL) that form the current frontier of bioformulation challenges and/or integrate additional optimization objectives, further accelerating the development of stable, low-viscosity biologics. Finally, this work represents a step toward the realization of fully autonomous self-driving laboratories (SDLs)[45], which promise to execute the entire scientific method – from hypothesis generation to experimental validation and learning – with minimal human intervention. Such systems have the potential to not only dramatically compress drug development timelines and enhance research reproducibility but also unlock the capacity to rapidly engineer novel and more effective medicines to meet urgent global health needs.

## 4. EXPERIMENTAL SECTION

**Materials:** All reagents were purchased from Sigma-Aldrich unless specified otherwise. Bovine IgG (bIgG) was purchased from MP Biomedicals.

**Automation-assisted on-demand antibody formulation:** To facilitate automated and on-demand formulation of bIgG, a Hamilton MLSTARlet liquid handler was utilized to perform formulation preparation and necessary dilutions. Stock solutions for each formulation component were prepared in deionized water and filtered using a 0.2 μm PES filter prior to use. For $T_m$ and diffusivity studies, stock solutions of bIgG were first prepared in Picopure water at 30 mg/mL, triple-filtered using 0.05 μm PS filters to remove aggregates, then diluted to 20 mg/mL. Reagent and bIgG stock solutions were then loaded into the Hamilton ML STARlet liquid handler to prepare formulated bIgG in 96 well plates. Custom Python software was used to convert formulations compositions into actionable transfer steps that were then implemented automatically. Each excipient was available at 2-3 stock concentrations to facilitate formulation with high precision and accuracy while simultaneously ensuring the entire formulation design space remained accessible via a legitimate series of transfer steps. For each combination of excipients to be evaluated, working formulation stocks were first prepared from excipient stocks at 3x of the final intended concentration. These 3x formulation stocks were then combined with the bIgG stocks and additional deionzied water as needed to produce formulated antibody solutions at the correct concentrations of both antibody and excipient. For viscosity studies, stock solutions of bIgG were prepared at 200 mg/mL. 200 nm polystryene beads were then added to the solution to yield a 180 mg/mL solution with 0.1 wt% bead. This solution was then used to prepare formulated antibodies at high concentrations. Antibody was formulated at 2.5, 5, 10, and 15 mg/mL for $T_m$ and diffusivity characterization studies, and at 72, 90, and 120 mg/mL for viscosity characterization studies. At each concentration, antibody formulations were prepared in triplicate for $T_m$ and diffusivity measurements, and in quadruplicate for viscosity measurements. bIgG stock solutions for the entire study were prepared immediately prior to seed library preparation and analysis. These same stocks were used for preparation for each round of DBTL to eliminate batch-to-batch variability. Likewise, stocks of individual excipients were prepared once at the start of the study and used for all rounds.

**High-throughput characterization:**

*Thermal Stability via differential scanning fluorometry (DSF):* Thermal stability of the formulated antibody solutions was evaluated by differential scanning fluorometry (DSF) using a QuantStudio 3 (Thermo Fisher Scientific) real-time PCR system and established techniques.[57] A Sypro Orange solution was first prepared in Picopure water and filtered using a 0.45 μm hydrophilic PTFE filter prior to use. Samples were then prepared by addition of 45 μL of the formulated antibody solutions were transferred to a PCR plate, followed by the addition of 5μL of the Sypro Orange solution. The plate was then loaded into the instrument and formulations were subjected to a heat ramp from 25°C to 95°C at a rate of 1°C/min. Fluorescence ($\lambda_{Ex}$ = 520±10, $\lambda_{Em}$ = 558±11) was measured over the course of the ramp. Automated analysis was used to calculate the first protein melting event by applying a Savgol filter to pre-process data and then curve fitting the Boltzmann sigmoid equation (Equation 1). Melt temperature ($T_m$) was defined as the inflection point of the curve fit. In cases where multiple thermal transitions were observed, likely corresponding to denaturation of different domains of the antibody, the first transition was used for $T_m$ calculation to provide a more conservative estimate of thermal stability. Each formulation was evaluated at four antibody concentrations (2.5, 5, 10, and 15 mg/mL), each with three experimental replicates prepared and measured.

$$y(x) = A_2 + \frac{A_1 - A_2}{1 + \exp((x - x_0)/dx)} \qquad \textit{Equation 1}$$

*Diffusivity measurements via dynamic light scattering (DLS):* DLS measurements[58] were performed on a DyanaPro Plate Reader III (Wyatt Technologies) and analyzed using the accompanying Dynamics 7.10 software package. 96-well plates containing the formulated antibodies were centrifuged at 2000 g for 5 mins at room temperature to remove microbubbles prior to analysis. For each sample, eight repeat acquisitions with 8 second acquisition time were collected at 25°C. A data filter was applied to all data to assess data quality, following the Wyatt Technology default parameters: minimum amplitude = 0, maximum amplitude = 1, baseline limit = 1+/- 0.01, polydispersity =< 40%. $D_c$ measurements for that formulation replicate were then calculated from the average of the cumulant autocorrelation function. Each formulation was evaluated at four antibody concentrations (2.5, 5, 10, and 15 mg/mL) with each three experimental replicates prepared and measured per concentration.

*Microrheology for viscosity measurements:* Viscosity measurements of antibody formulations were performed by using a high-throughput DLS-based technique to track the diffusion of 200 nm diameter polystyrene beads added as tracer particles.[59] 384-well plates containing the formulated antibodies along with 200 nm microspheres were first centrifuged at 2000 g for 5 mins at room temperature to remove microbubbles prior to analysis. For each well, ten consecutive 20s acquisitions were collected at 25°C using optimized laser power and attenuation settings. A data filter was applied to all data to assess data quality, following the Wyatt Technology default parameters: minimum amplitude = 0, maximum amplitude = 1, baseline limit = 1+/- 0.01, polydispersity =< 40%. Regularization was used to obtain the apparent radius of the microsphere. Viscosity of the solution was then calculated using the Stokes-Einstein equation (Equation 2) to relate measured radius ($R_{h\ app}$) of the tracer bead to the known radius ($R_{\mathrm{h\ Bead}}$) and determine the shift in viscosity at fixed temperature (Equation 3). Viscosity was measured 24 h after initial formulation to avoid transient viscosity effects arising from network rearrangement. Each formulation was evaluated at three antibody concentrations (72, 90, and 120 mg/mL), with each four experimental replicates prepared and measured per concentration.

$$R_h = \frac{k_B T}{6\pi\eta D} \qquad \textit{Equation 2}$$

$$\eta_2 = \frac{\eta_1 R_{h\,app}}{R_{h\,Bead}} \qquad \textit{Equation 3}$$

*Machine-Learning Surrogate Models:* Antibody formulations were encoded as seven-dimensional physicochemical vectors comprising Molarity, NaCl, Sucrose, Arginine, pH, buffer identity, and protein concentration. Continuous features entered the model unaltered, whereas the categorical buffer was label-encoded (1 = acetate, 2 = citrate) with discrete values. Objective-specific MinMaxScalers from Sci-Kit Learn subsequently mapped each feature to [0, 1]; distinct scalers were retained for melting temperature ($T_m$), diffusion coefficient ($D_c$), and viscosity (η) so that the scale of one objective could not influence another. Targets were scaled analogously and stored alongside formulation identifiers that are later used to enforce group-wise splits in cross-validation. For each objective the relationships between formulations and biophysical characterization measurements was modeled using GPR to capture nontrivial, nonlinear mapping and to support active learning as GPR naturally provide both mean μ and uncertainty $\sigma^2$ on predicted points. The GPR surrogate was instantiated with BoTorch's SingleTaskGP in which covariance is modelled by an isotropic squared-exponential kernel:

$$k\left(\vec{x}, \overrightarrow{x'}\right) = \sigma^2 \exp\left(-\frac{\left\|\vec{x}-\overrightarrow{x'}\right\|^2}{2l^2}\right) + \sigma_n^2 \qquad \textit{Equation 4}$$

Where $\vec{x}$ is the feature vector of the formulation, and $\sigma, l$, $\sigma_n$ are kernel hyperparameters. These parameters were optimized by maximizing the exact log-marginal likelihood. Predictive performance was assessed with 10-fold group cross-validation in which all repeats from the same "Formulation ID" reside in the same fold; a choice made to buffer against overfitting due to sampling multiple concentrations of the same formula. Predictions across all folds were collected and model performance was then assessed via coefficient of determination ($R^2$ value) and mean-squared error (MSE), calculated between predicted and measured values.

*Candidate Formulation Generation:* BO was employed to navigate the seven-dimensional formulation domain and prioritize new experimental queries. Each property—melting temperature ($T_m$), diffusion coefficient (D), and viscosity (η)—was first optimized in isolation using its dedicated Gaussian-process surrogate. All coordinates were expressed in objective-specific [0, 1] space obtained by Min–Max scaling of the original experimental limits; the transformation is invertible, guaranteeing that proposed designs remain within chemically feasible bounds. A fixed batch size of q = 6 for each objective was imposed to align with plate-based synthesis throughput. During each round, single-objective optimizers maximized the q-exclusive expected improvement (qEI) acquisition function[49], defined as:

$$qEI(\boldsymbol{X}) = E\left[\max_{1\le i\le q}(f_i - f^+)_+\right] \qquad \textit{Equation 5}$$

where $f^+$ denotes the best observed response. A quasi-Monte Carlo (QMC) sampler[60] was used to approximate the expected improvement by drawing samples from the model's multivariate normal posterior distribution using Sobol sequences[61], improving accuracy of acquisition values. For viscosity, the training targets were stored as (1 – η) so that all objectives share a maximization convention; the same transformation was applied to $f^+$ when evaluating qEI. qEI was evaluated at 500 uniformly spaced Sobol seeds and the best local optimum across restarts was taken as the final batch. These six points were inverse-scaled to experimental units, their protein concentration field overwritten to match the training distribution (15 $\text{mg mL}^{-1}$ for $T_m$ and $D_c$; 120 $\text{mg mL}^{-1}$ for η), and screened to ensure valid buffer/pH pairs are proposed before automation-assisted formulation.

Upon completing the three single-objective optimizations, $T_m$, $D_c$, and viscosity GPR models were utilized jointly to perform multi-objective optimization (MOO). For this, we maximized the q-expected hypervolume improvement (qEHVI) acquisition function[49]:

$$qEHVI(\boldsymbol{X}) = E_{\boldsymbol{F}\sim p(\boldsymbol{F}|\mathbb{D})}[\max!(H(\mathcal{P} \cup \boldsymbol{F}) - H(\mathcal{P}), 0)] \qquad \textit{Equation 6}$$

Where $\boldsymbol{X}$ is the batch of q candidates, $\boldsymbol{F}$ is the posterior samples at those points, $\mathcal{P}$ denotes the current Pareto set, and $H(\cdot)$ represents the dominated hyper-volume with respect to a fixed reference point r = (0, 0, 0). qEHVI quantifies the expected increase in dominated hyper-volume of the Pareto set when a batch of size six is added, with η again transformed to (1 – η) so that all objectives contribute positively while aligned with our objective of minimizing viscosity. The acquisition value and its gradient also estimated using Sobol-QMC sampling as described for single-objective optimization.

**Implementation of SHaply Additive Explanations (SHAP):** Calculation and visualization of SHAP values was implemented using the SHAP python package (v0.48.0). The full dataset collected from all 72 formulations was used as background to determine the base SHAP value for each GPR model. As the GPR models were trained using scaled inputs and outputs, SHAP values were first transformed from scaled into real units prior to plotting. This was achieved by calculating the scale factor originally used to transform the output (via the MinMaxScaler function) and then multiplying SHAP values by this factor, thus preserving their additive nature. The inverse transformation of the MinMaxScaler was then used to change the SHAP base value into native measurement units.

## ASSOCIATED INFORMATION

Supporting Information for "Automation and Active Learning for the Multi-Objective Optimization of Antibody Formulations"

## AUTHOR INFORMATION

### Corresponding Author

* Adam J. Gormley: adam.gormley@rutgers.edu

### Author Contributions

# D.C. Radford and M. Tamasi contributed equally.

### Funding Sources

This work was supported by funding from the National Institutes of Health (NIH) NIGMS R35GM138296 and NIBIB R01EB037022, as well as National Science Foundation (NSF) DMREF 2118860 and CBET 2309852

### Conflict of Interest Disclosure

MT and AJG co-founded Plexymer, Inc. which has licensed technology associated with this work.

The authors declare no other competing financial interest.

## REFERENCES

(1) Lu, R.-M.; Hwang, Y.-C.; Liu, I. J.; Lee, C.-C.; Tsai, H.-Z.; Li, H.-J.; Wu, H.-C. Development of therapeutic antibodies for the treatment of diseases. *Journal of Biomedical Science* **2020**, *27* (1).

(2) Marks, L. The birth pangs of monoclonal antibody therapeutics. *mAbs* **2014**, *4* (3), 403-412.

(3) Crescioli, S.; Kaplon, H.; Wang, L.; Visweswaraiah, J.; Kapoor, V.; Reichert, J. M. Antibodies to watch in 2025. *mAbs* **2024**, *17* (1).

(4) Lyu, X.; Zhao, Q.; Hui, J.; Wang, T.; Lin, M.; Wang, K.; Zhang, J.; Shentu, J.; Dalby, P. A.; Zhang, H.; et al. The global landscape of approved antibody therapies. *Antibody Therapeutics* **2022**, *5* (4), 233-257.

(5) Sharma, P.; Joshi, R. V.; Pritchard, R.; Xu, K.; Eicher, M. A. Therapeutic Antibodies in Medicine. *Molecules* **2023**, *28* (18).

(6) Sifniotis, V.; Cruz, E.; Eroglu, B.; Kayser, V. Current Advancements in Addressing Key Challenges of Therapeutic Antibody Design, Manufacture, and Formulation. *Antibodies* **2019**, *8* (2).

(7) Strickley, R. G.; Lambert, W. J. A review of Formulations of Commercially Available Antibodies. *Journal of Pharmaceutical Sciences* **2021**, *110* (7), 2590-2608.e2556.

(8) Wang, W.; Singh, S.; Zeng, D. L.; King, K.; Nema, S. Antibody Structure, Instability, and Formulation. *Journal of Pharmaceutical Sciences* **2007**, *96* (1), 1-26.

(9) Ratanji, K. D.; Derrick, J. P.; Dearman, R. J.; Kimber, I. Immunogenicity of therapeutic proteins: Influence of aggregation. *Journal of Immunotoxicology* **2013**, *11* (2), 99-109.

(10) Kollár, É.; Balázs, B.; Tari, T.; Siró, I. Development challenges of high concentration monoclonal antibody formulations. *Drug Discovery Today: Technologies* **2020**, *37*, 31-40.

(11) Wang, S.; Zhang, N.; Hu, T.; Dai, W.; Feng, X.; Zhang, X.; Qian, F. Viscosity-Lowering Effect of Amino Acids and Salts on Highly Concentrated Solutions of Two IgG1 Monoclonal Antibodies. *Molecular Pharmaceutics* **2015**, *12* (12), 4478-4487.

(12) Hung, J. J.; Dear, B. J.; Karouta, C. A.; Chowdhury, A. A.; Godfrin, P. D.; Bollinger, J. A.; Nieto, M. P.; Wilks, L. R.; Shay, T. Y.; Ramachandran, K.; et al. Protein–Protein Interactions of Highly Concentrated Monoclonal Antibody Solutions via Static Light Scattering and Influence on the Viscosity. *The Journal of Physical Chemistry B* **2019**, *123* (4), 739-755.

(13) Tomar, D. S.; Kumar, S.; Singh, S. K.; Goswami, S.; Li, L. Molecular basis of high viscosity in concentrated antibody solutions: Strategies for high concentration drug product development. *mAbs* **2016**, *8* (2), 216-228.

(14) Mieczkowski, C. A. The Evolution of Commercial Antibody Formulations. *Journal of Pharmaceutical Sciences* **2023**, *112* (7), 1801-1810.

(15) Zajac, J. W. P.; Muralikrishnan, P.; Heldt, C. L.; Perry, S. L.; Sarupria, S. Towards stable biologics: understanding co-excipient effects on hydrophobic interactions and solvent network integrity. *Molecular Systems Design & Engineering* **2025**, *10* (6), 432-446.

(16) Sahin, E.; Grillo, A. O.; Perkins, M. D.; Roberts, C. J. Comparative Effects of pH and Ionic Strength on Protein–Protein Interactions, Unfolding, and Aggregation for IgG1 Antibodies. *Journal of Pharmaceutical Sciences* **2010**, *99* (12), 4830-4848.

(17) Kamerzell, T. J.; Esfandiary, R.; Joshi, S. B.; Middaugh, C. R.; Volkin, D. B. Protein–excipient interactions: Mechanisms and biophysical characterization applied to protein formulation development. *Advanced Drug Delivery Reviews* **2011**, *63* (13), 1118-1159.

(18) Yang, D.; Walker, L. M. Synergistic Effects of Multiple Excipients on Controlling Viscosity of Concentrated Protein Dispersions. *Journal of Pharmaceutical Sciences* **2023**, *112* (5), 1379-1387.

(19) Lai, P.-K.; Fernando, A.; Cloutier, T. K.; Gokarn, Y.; Zhang, J.; Schwenger, W.; Chari, R.; Calero-Rubio, C.; Trout, B. L. Machine Learning Applied to Determine the Molecular Descriptors Responsible for the Viscosity Behavior of Concentrated Therapeutic Antibodies. *Molecular Pharmaceutics* **2021**, *18* (3), 1167-1175.

(20) Lai, P.-K.; Gallegos, A.; Mody, N.; Sathish, H. A.; Trout, B. L. Machine learning prediction of antibody aggregation and viscosity for high concentration formulation development of protein therapeutics. *mAbs* **2022**, *14* (1).

(21) Jain, T.; Sun, T.; Durand, S.; Hall, A.; Houston, N. R.; Nett, J. H.; Sharkey, B.; Bobrowicz, B.; Caffry, I.; Yu, Y. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences* **2017**, *114* (5), 944-949.

(22) Shim, H. Bispecific Antibodies and Antibody–Drug Conjugates for Cancer Therapy: Technological Considerations. *Biomolecules* **2020**, *10* (3).

(23) Zarzar, J.; Khan, T.; Bhagawati, M.; Weiche, B.; Sydow-Andersen, J.; Sreedhara, A. High concentration formulation developability approaches and considerations. *mAbs* **2023**, *15* (1).

(24) Bao, Z.; Bufton, J.; Hickman, R. J.; Aspuru-Guzik, A.; Bannigan, P.; Allen, C. Revolutionizing drug formulation development: The increasing impact of machine learning. *Advanced Drug Delivery Reviews* **2023**, *202*, 115108.

(25) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian process regression for materials and molecules. *Chemical reviews* **2021**, *121* (16), 10073-10141.

(26) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE* **2015**, *104* (1), 148-175.

(27) Rodríguez-Pérez, R.; Miljković, F.; Bajorath, J. Machine Learning in Chemoinformatics and Medicinal Chemistry. *Annual Review of Biomedical Data Science* **2022**, *5* (1), 43-65.

(28) Thompson, J.; Walters, W. P.; Feng, J. A.; Pabon, N. A.; Xu, H.; Maser, M.; Goldman, B. B.; Moustakas, D.; Schmidt, M.; York, F. Optimizing active learning for free energy calculations. *Artificial Intelligence in the Life Sciences* **2022**, *2*.

(29) Panteleev, J.; Gao, H.; Jia, L. Recent applications of machine learning in medicinal chemistry. *Bioorganic & Medicinal Chemistry Letters* **2018**, *28* (17), 2807-2815.

(30) Suriyaamporn, P.; Pamornpathomkul, B.; Patrojanasophon, P.; Ngawhirunpat, T.; Rojanarata, T.; Opanasopit, P. The Artificial Intelligence-Powered New Era in Pharmaceutical Research and Development: A Review. *AAPS PharmSciTech* **2024**, *25* (6).

(31) Arslan, A.; Yet, B.; Nemutlu, E.; Akdağ Çaylı, Y.; Eroğlu, H.; Öner, L. Celecoxib Nanoformulations with Enhanced Solubility, Dissolution Rate, and Oral Bioavailability: Experimental Approaches over In Vitro/In Vivo Evaluation. *Pharmaceutics* **2023**, *15* (2).

(32) Bannigan, P.; Aldeghi, M.; Bao, Z.; Häse, F.; Aspuru-Guzik, A.; Allen, C. Machine learning directed drug formulation development. *Advanced Drug Delivery Reviews* **2021**, *175*.

(33) Narayanan, H.; Hinckley, J. A.; Barry, R.; Dang, B.; Wolffe, L. A.; Atari, A.; Tseng, Y.-Y.; Love, J. C. Accelerating cell culture media development using Bayesian optimization-based iterative experimental design. *Nature Communications* **2025**, *16* (1).

(34) Claes, E.; Heck, T.; Coddens, K.; Sonnaert, M.; Schrooten, J.; Verwaeren, J. Bayesian cell therapy process optimization. *Biotechnology and Bioengineering* **2024**, *121* (5), 1569-1582.

(35) Bader, J.; Narayanan, H.; Arosio, P.; Leroux, J.-C. Improving extracellular vesicles production through a Bayesian optimization-based experimental design. *European Journal of Pharmaceutics and Biopharmaceutics* **2023**, *182*, 103-114.

(36) Khan, A.; Cowen-Rivers, A. I.; Grosnit, A.; Deik, D.-G.-X.; Robert, P. A.; Greiff, V.; Smorodina, E.; Rawat, P.; Akbar, R.; Dreczkowski, K.; et al. Toward real-world automated antibody design with combinatorial Bayesian optimization. *Cell Reports Methods* **2023**, *3* (1).

(37) Schneider, G. Automating drug discovery. *Nature Reviews Drug Discovery* **2017**, *17* (2), 97-113.

(38) Oyejide, A. J.; Adekunle, Y. A.; Abodunrin, O. D.; Atoyebi, E. O. Artificial intelligence, computational tools and robotics for drug discovery, development, and delivery. *Intelligent Pharmacy* **2025**, *3* (3), 207-224.

(39) Tom, G.; Schmid, S. P.; Baird, S. G.; Cao, Y.; Darvish, K.; Hao, H.; Lo, S.; Pablo-García, S.; Rajaonson, E. M.; Skreta, M.; et al. Self-Driving Laboratories for Chemistry and Materials Science. *Chemical Reviews* **2024**, *124* (16), 9633-9732.

(40) Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W. J.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J. R.; Curioni, A. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials* **2022**, *8* (1).

(41) Kosuri, S.; Borca, C. H.; Mugnier, H.; Tamasi, M.; Patel, R. A.; Perez, I.; Kumar, S.; Finkel, Z.; Schloss, R.; Cai, L.; et al. Machine-Assisted Discovery of Chondroitinase ABC Complexes toward Sustained Neural Regeneration. *Advanced Healthcare Materials* **2022**, *11* (10).

(42) Tamasi, M. J.; Patel, R. A.; Borca, C. H.; Kosuri, S.; Mugnier, H.; Upadhya, R.; Murthy, N. S.; Webb, M. A.; Gormley, A. J. Machine learning on a robotic platform for the design of polymer–protein hybrids. *Advanced Materials* **2022**, *34* (30), 2201809.

(43) Di Mare, E. J.; Punia, A.; Lamm, M. S.; Rhodes, T. A.; Gormley, A. J. Data-Driven Design of Novel Polymer Excipients for Pharmaceutical Amorphous Solid Dispersions. *Bioconjugate Chemistry* **2024**, *35* (9), 1363-1372.

(44) Narayanan, H.; Dingfelder, F.; Condado Morales, I.; Patel, B.; Heding, K. E.; Bjelke, J. R.; Egebjerg, T.; Butté, A.; Sokolov, M.; Lorenzen, N.; et al. Design of Biopharmaceutical Formulations Accelerated by Machine Learning. *Molecular Pharmaceutics* **2021**, *18* (10), 3843-3853.

(45) Tamasi, M. J.; Gormley, A. J. Biologic formulation in a self-driving biomaterials lab. *Cell Reports Physical Science* **2022**, *3* (9).

(46) Braham, E. J.; Davidson, R. D.; Al-Hashimi, M.; Arroyave, R.; Banerjee, S. Navigating the design space of inorganic materials synthesis using statistical methods and machine learning. *Dalton Trans* **2020**, *49* (33), 11480-11488.

(47) Loh, W.-L. On Latin hypercube sampling. *The Annals of Statistics* **1996**, *24* (5).

(48) Schulz, E.; Speekenbrink, M.; Krause, A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology* **2018**, *85*, 1-16.

(49) Balandat, M.; Karrer, B.; Jiang, D.; Daulton, S.; Letham, B.; Wilson, A. G.; Bakshy, E. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In 2020.

(50) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

(51) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.

(52) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* **2020**, *2* (1), 56-67.

(53) Svilenov, H. L.; Kulakova, A.; Zalar, M.; Golovanov, A. P.; Harris, P.; Winter, G. Orthogonal Techniques to Study the Effect of pH, Sucrose, and Arginine Salts on Monoclonal Antibody Physical Stability and Aggregation During Long-Term Storage. *Journal of Pharmaceutical Sciences* **2020**, *109* (1), 584-594.

(54) Ren, S. Effects of arginine in therapeutic protein formulations: a decade review and perspectives. *Antibody Therapeutics* **2023**, *6* (4), 265-276.

(55) Meyer, B. K.; Hu, B.; Ionescu, R.; Hamm, C.; Wang, N.; Mach, H.; Kirchmeier, M. J.; Sweeney, J. Opalescence of an IgG1 Monoclonal Antibody Formulation Mediated by Ionic Strength and Excipients. *BioPharm International* **2022**, *22* (4).

(56) Hüllermeier, E.; Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning* **2021**, *110* (3), 457-506.

(57) Huynh, K.; Partch, C. L. Analysis of protein stability and ligand interactions by thermal shift assay. *Current protocols in protein science* **2015**, *79* (1), 28.29. 21-28.29. 14.

(58) Bhirde, A. A.; Chiang, M.-J.; Venna, R.; Beaucage, S.; Brorson, K. High-throughput in-use and stress size stability screening of protein therapeutics using algorithm-driven dynamic light scattering. *Journal of Pharmaceutical Sciences* **2018**, *107* (8), 2055-2062.

(59) He, F.; Becker, G. W.; Litowski, J. R.; Narhi, L. O.; Brems, D. N.; Razinkov, V. I. High-throughput dynamic light scattering method for measuring viscosity of concentrated protein solutions. *Analytical biochemistry* **2010**, *399* (1), 141-143.

(60) Dick, J.; Pillichshammer, F. *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*; Cambridge University Press, 2010.

(61) Sobol, I. M. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics* **1967**, *7* (4), 86-112.