molecular pharmaceutics

pubs.acs.org/molecularpharmaceutics

Article

Design of Biopharmaceutical Formulations Accelerated by Machine Learning

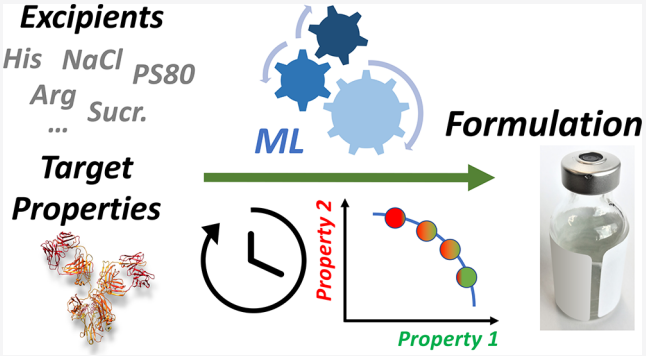
Harini Narayanan, Fabian Dingfelder, Itzel Condado Morales, Bhargav Patel, Kristine Enemærke Heding, Jais Rose Bjelke, Thomas Egebjerg, Alessandro Butté, Michael Sokolov, Nikolai Lorenzen, and Paolo Arosio*

Cite This: *Mol. Pharmaceutics* 2021, 18, 3843−3853

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In addition to activity, successful biological drugs must exhibit a series of suitable developability properties, which depend on both protein sequence and buffer composition. In the context of this high-dimensional optimization problem, advanced algorithms from the domain of machine learning are highly beneficial in complementing analytical screening and rational design. Here, we propose a Bayesian optimization algorithm to accelerate the design of biopharmaceutical formulations. We demonstrate the power of this approach by identifying the formulation that optimizes the thermal stability of three tandem single-chain Fv variants within 25 experiments, a number which is less than one-third of the experiments that would be required by a classical DoE method and several orders of magnitude smaller compared to detailed experimental analysis of full combinatorial space. We further show the advantage of this method over conventional approaches to efficiently transfer historical information as prior knowledge for the development of new biologics or when new buffer agents are available. Moreover, we highlight the benefit of our technique in engineering multiple biophysical properties by simultaneously optimizing both thermal and interface stabilities. This optimization minimizes the amount of surfactant in the formulation, which is important to decrease the risks associated with corresponding degradation processes. Overall, this method can provide high speed of converging to optimal conditions, the ability to transfer prior knowledge, and the identification of new nonlinear combinations of excipients. We envision that these features can lead to a considerable acceleration in formulation design and to parallelization of operations during drug development.



KEYWORDS: *formulation, machine learning, artificial intelligence, biopharmaceuticals, antibodies, developability, stability, Bayesian optimization*

1. INTRODUCTION

Biotherapeutics represent an important class of drugs that have proven successful for treating diseases such as certain types of cancers and autoimmune and inflammatory disorders.[1] This success can be largely attributed to high specificity, high efficacy, lower toxicity, and reduced side effects. In addition to activity and safety, the translation of a candidate molecule into a successful biotherapeutic drug requires consistent manufacturing at the highest possible standards as well as stability during storage, transportation, and administration.[2−5] These requirements are challenging to achieve with complex molecules such as proteins since they are amenable to a variety of chemical and physical degradation pathways under the different conditions encountered during the entire life cycle.[6,7] These risks can be reduced by developing molecules with a variety of suitable biophysical properties that are globally indicated as "developability" of the product.[8−11] These suitable properties are system-specific and can be optimized by modulating either the protein primary sequence or the formulation of the molecule.[5]

Liquid formulations currently represent the most common administration route for monoclonal antibodies.[12] These formulations contain a variety of excipients to maintain pH and tonicity and to increase protein stability and preservation.[13−15] Typical categories of excipients include buffering agents, tonicity modifiers, thermal stabilizers, surfactants, and amino acids.[15−18] This strategy offers many degrees of freedom since in principle an infinite number of different excipients and

Received: June 10, 2021
Revised: August 22, 2021
Accepted: August 23, 2021
Published: September 14, 2021

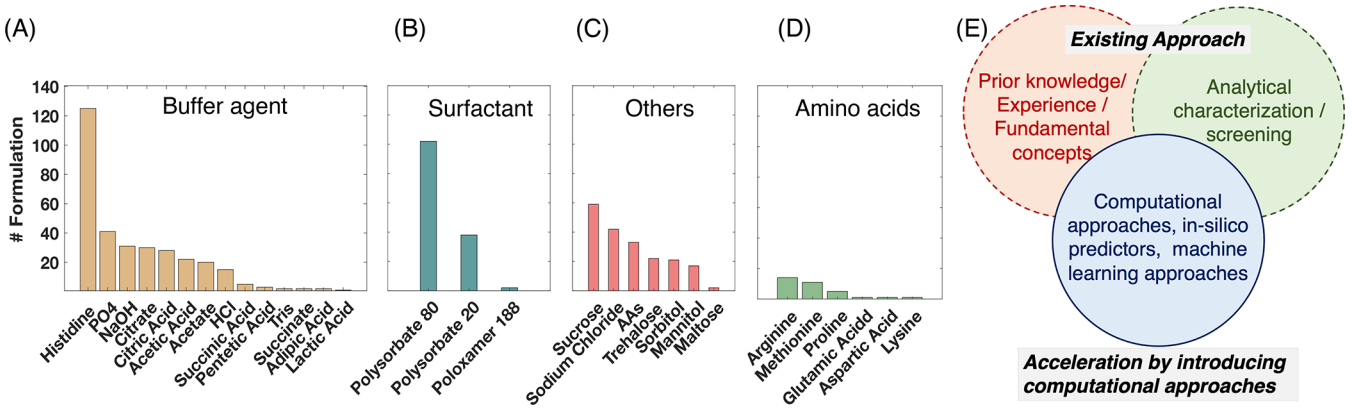


Figure 1. (A–D) Frequency of excipient appearance in the antibody formulations marketed since 2015. PO_4: phosphates, NaOH: sodium hydroxide, and HCl: hydrochloric acid. (E) In silico predictors and approaches from machine learning[21,22] can accelerate formulation design by complementing prior knowledge and analytical characterization.

their combinations can be selected to simultaneously optimize multiple properties of a protein. The introduction of novel excipients, however, is largely prevented by regulatory considerations and corresponding increases in the approval timelines.[17,19,20] As a consequence, excipients are typically selected from the approved list of molecules documented under the FDA's inactive substance database, which are generally regarded as safe (GRAS).[16]

Figure 1 summarizes the excipients that are most commonly used in marketed antibody formulations based on the data available in the PharmaCircle database. In the last 5 years, histidine has been the most common buffering agent in the marketed antibody formulations followed by phosphates, sodium hydroxide, citrate (citric acid), and acetate (acetic acid) (Figure 1A). Polysorbates (Polysorbate 80 or 20) are the most common surfactants, with Polysorbate 80 appearing more often than Polysorbate 20 (Figure 1B). Among other excipients, shown in Figure 1C, sucrose and sodium chloride (NaCl) are the most predominant, followed by trehalose, mannitol, sorbitol, and amino acids (AAs). AAs are commonly added also to prevent issues with aggregation and high viscosity.[23] Arginine and methionine appear often, followed by proline, while other AAs are less common (Figure 1D). In particular, sodium chloride (NaCl) and arginine, often in the form of its salt ArgHCl, are increasingly popular excipients for high-concentration antibody formulations often used for subcutaneous administration, as they often can reduce viscosity and occasionally also aggregation propensity.[23,24] Methionine is introduced due to its antioxidant properties.[24]

Formulation design is currently largely driven by previous knowledge and experience, assisted by extensive analytical characterization.[25−27] Most of the reported studies analyze the effect of individual excipients on a variety of biophysical properties.[28−30] However, different excipients could have a synergistic effect, and their combinations may lead to drastic improvements in the performance of molecules, in particular because multiple properties must be simultaneously optimized.

Modern formulations, however, tend to keep the composition as simple as possible.[15] Among the 1758 excipients approved by FDA,[31] only 30 excipients appear in the marketed antibody formulations since 2015 and only 18 appear in more than 10 products (Figure 1). This is also due to the fact that performing screening campaigns of a larger number of excipient combinations may not be feasible under the inherent time pressure in drug development. A classical experimental design has been applied,[32−35] however mostly limited to a reduced number of excipients identified from prior screenings.

This limitation motivates the development of new theoretical and experimental methods to identify optimal combinations of multiple excipients with minimal resource utilization. In the context of this highly dimensional optimization problem, advanced algorithms from the domain of machine learning and mathematical optimization would be extremely beneficial to complement rational design based on analytical screenings and prior knowledge or experience[36] (Figure 1E) (see refs 21 and 22, and references herein).

Bayesian optimization[37,38] has become popular for the optimization of "black-box" functions wherein the underlying relationship between the input and output is unknown and points in the input−output space can be determined only experimentally. In brief, Bayesian optimization suggests experiments sequentially using a surrogate model that mimics the system under study based on the experiments observed previously. Due to this property of adaptively sampling locations, the algorithm is capable of achieving optimal conditions faster and with reduced number of overall experiments.

Bayesian optimization has been applied to various fields such as the tuning of the hyperparameters of machine learning algorithms,[39] robotics,[40] circuit optimization,[41] synthetic gene design,[42] directed-evolution of proteins,[43] and more recently also in materials science.[44]

In this work, we demonstrate the potential of the Bayesian optimization algorithm to optimize the biophysical property of a target protein by identifying the ideal formulation composition within a complex design space. Specifically, we optimized the thermal stability, described via the melting temperature (T_m), which is one of the most important quality attributes of biologics. We applied our approach to three different variants of a tandem single-chain variable fragment (scFv) derived from the antibody Humira. Eight factors (pH, sodium chloride, L-arginine, L-lysine, L-proline, trehalose, mannitol, and Tween 20) were considered as independent variables to maximize T_m. We show that using this technique the optimal combination that maximizes T_m is achieved within 25 experiments. This number is at least 3-fold lower compared to the experiments that would be required by a classical DoE method and several orders of magnitude smaller compared to the experimental screening of the entire combinatorial space, also called the full screening method. Moreover, we

https://doi.org/10.1021/acs.molpharmaceut.1c00469
Mol. Pharmaceutics 2021, 18, 3843−3853

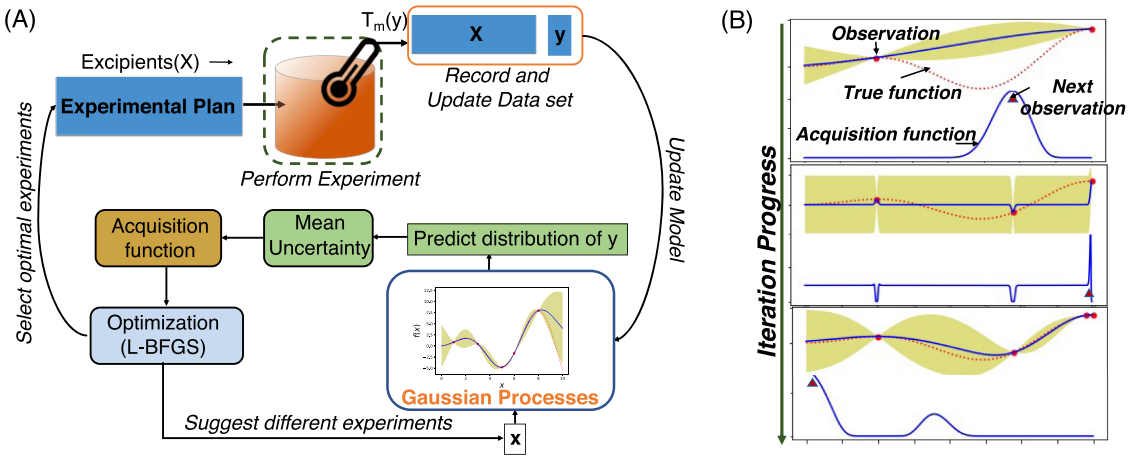


Figure 2. (A) Schematic illustration of a sequential experimental design using Bayesian optimization. (B) Illustration of the unknown response surface (red-dashed line), distribution of the response as predicted by the model (mean—black line, uncertainty—green bands), corresponding acquisition function (blue), and the resulting sampling (red triangle) in the different iteration.

demonstrate the advantage of this method over conventional approaches to efficiently transfer historical information as prior knowledge for the development of new biologics or when new excipients are available.

Finally, we highlight that the use of such techniques is even more powerful when multiple properties of a molecule must be simultaneously optimized. To this end, we simultaneously optimize both the T_m and the stability of proteins toward hydrophobic interfaces measured with a nanoparticle-based assay recently developed in our laboratory. This operation is important for instance to minimize the amount of surfactant in the formulation, decreasing the risks associated with corresponding degradation processes.[45,46]

2. MATERIALS AND METHODS

2.1. Sequential Design of Experiments Using Bayesian Optimization. Screening all possible combinations of excipients at different concentrations (referred to as the grid search or full screening) results in millions of experiments since the number of experiments increases exponentially with the number of factors and the number of compositions per factor to be tested, therefore rendering it impractical. In principle, the number of experiments can be reduced by testing one factor at a time and fixing all of the others constant. This approach, called the one factor at a time (OFAT), however, results in suboptimal conditions since it does not probe synergistic interactions between multiple factors. Thus, it becomes important to strategically plan experiments to explore the interactions between the different factors while minimizing the experimental effort.

To this aim, several statistical DoEs plan experiments in a structured manner, as summarized in the Supporting Information (Figure S1).

However, a common drawback of all of these methodologies is that they provide standard designs with fixed number of experiments[47] and allocate equal resources for both the "good" and the "bad" performers, or, in other words, they are nonadaptive.[37] This is due to the fact that classical DoE treats the design of experiment, modeling, and optimization as independent blocks. Additionally, these designs require a correct assumption about the response surface. In a real application, the underlying relationship between the input and output could be unknown, and the only possibility is to make point evaluation through experimental measurements. In such cases, the Bayesian optimization algorithm (BO) has proven to be a very powerful method.[38,48]

In contrast to the static designs of classical DoE methods, BO applies a sequential procedure in which a surrogate model of the actual system suggests the next experiment(s) based on the data already acquired, as schematically represented in Figure 2A. BO plans the next experiment(s) by optimizing a trade-off between "exploration" and "exploitation". First, it samples in areas that have not been explored before (based on distances from experiments already performed). Second, it samples in the region that has interesting behavior such as the maximum value of response observed until that moment. BO incorporates two main components:

2.1.1. Surrogate Modeling via Gaussian Processes (GPs). GPs are specified by a mean function ($m(x)$) and the covariance function ($k(x,x')$).

$$y_i = f(x_i) + \epsilon_i \tag{1}$$

$$f(x) \sim GP(m(x),\ k(x, x')) \tag{2}$$

The covariance function is referred to as the kernel, which indicates the closeness of two points x and x' obtained from the design space. Matern kernel was used in this work since it is a flexible, smooth kernel. A Matern kernel is represented by the following equation

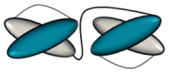
$$k_{Matern}(x,\ x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\,|x-x'|}{\theta}\right)^{\nu} \mathrm{K}_{\nu}\left(\frac{\sqrt{2\nu}\,|x-x'|}{\theta}\right) \tag{3}$$

The $|x - x'|$ indicates the distance between two points. The kernel value decreases as the distance increases (i.e., the correlation between the points decreases) over the length scale, θ. The value of θ is a hyperparameter that is obtained by fitting the surrogate model to the training data set. The parameter ν controls the smoothness, with smaller values of ν indicating less smoothness. Typically used values of ν are 1.5 and 2.5. Γ is the γ function, and K_ν is the modified Bessel function.

2.1.2. Trade-Off Encoded in the Acquisition Function. The trained GP model can now predict the distribution of response values at each point in the design space. Owing to the surrogate model being Gaussian processes, this prediction is a

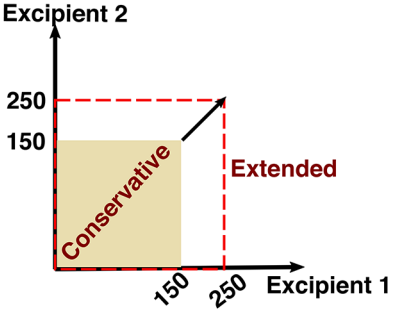
https://doi.org/10.1021/acs.molpharmaceut.1c00469
Mol. Pharmaceutics 2021, 18, 3843−3853

(A)



Variant	T_m (°C)
Variant 1	52.40
Variant 2	59.60
Variant 3	55.60

(B)



(C)

Excipient Names	pH	NaCl (mM)	Trehalose (w/v %)	Mannitol (w/v %)	L- Arginine (mM)	L-Lysine (mM)	L-Proline (mM)	Tween 20 (v/v %)
Lower Bound (LB)	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00
Upper Bound (UB)	6.5	140.0	8.0	4.0	150.0	150.0	150.0	0.01
Extended UB	6.5	140.0	10.0	5.0	250.0	250.0	250.0	0.01
Type/Level	Disc. / 4	Cont.	Disc./ 10	Disc./ 5	Cont.	Cont.	Cont.	Disc./2

Figure 3. (A) T_m values of the three variants in the starting reference formulation. (B) Schematic representation of the conservative design space (yellow region) and the extended design space (highlighted in red) demonstrated for a combination of two excipients. (C) Summary of the parameters and the corresponding boundaries considered for formulation optimization.

normal distribution that can be characterized by a mean $(\mu(x))$ and uncertainty $(\sigma(x))$. This mean and uncertainty is used to build the acquisition function, which is optimized by a local optimization solver LBFGS[49]—with multiple shootings to suggest the next experiment. The acquisition function is formulated to encode the trade-off between exploration (high uncertainty) and exploitation (high mean). In this wok, the LCB acquisition function is used to maximize the worst case predicted by the surrogate model. It is important to draw this trade-off efficiently to avoid getting trapped in local optimum or exhausting resources in exploring the design space. This can be ensured by a careful selection of the acquisition function and its parameters to allocate appropriate weights between exploration and exploitation. Another crucial aspect that could lead to local optimality is improper hyperparameter tuning of the kernels, as described in ref 50. This can be avoided by some specific additions to the acquisition function such as in ref 50. This problem can also be addressed by performing an initial space-filling design that provides the model with a sufficiently good overview of the space to robustly learn the hyperparameters. In this work, we followed the latter approach for efficient hyperparameter learning. The sampling of Bayesian optimization based on the acquisition function is illustrated in Figure 2B. Several software packages exist that provide the implementation of Bayesian optimization. The current work is based on the implementation of "skopt" and "gpflowopt" packages in python. All of the technical details about the Bayesian optimization approach are provided in the Supporting Information.

A traditional Bayesian optimization (BO) approach is entirely sequentially, meaning that only one experiment is suggested in every iteration. However, often it is more pragmatic to perform multiple experiments in parallel, as in our study. Under such circumstances, it is more convenient to apply the batch Bayesian optimization (Batch BO) approach, which allows for several experiments to be designed in parallel at each iteration. Multiple strategies have been proposed in the literature to perform batch BO such as in refs 51−54. Here, we adopt the simplest "constant-liar" approach as proposed in ref 55.

2.2. Experimental Part. *2.2.1. Tandem scFv Variants.* Expression plasmids for production of tandem scFv formats of Humira were purchased from Geneart/Life technologies. The variants were designed with a C-terminal HPC4 tag for purification purposes. All compounds were expressed by transient transfection of Expi293 cells following the instructions of the manufacturer (Expi system, Life technologies), and the resulting compounds were purified from cell culture supernatants using the anti-HPC4 immunoaffinity purification step followed by a Superdex200 size-exclusion purification step. The size-exclusion running buffer, in which the compounds was initially formulated prior to the formulation optimization study, was 20 mM Hepes, 150 mM NaCl, pH 7.4. The resulting compound preparations were analyzed by SDS-PAGE/Coomassie and SE-HPLSC analysis, and the identity of each compound was verified by intact mass LC-MS analysis. Protein concentrations were determined based on UV280 measurements.

2.2.2. Nano Differential Scanning Fluorimetry (nanoDSF). The melting temperature (T_m) of each variant was measured via nano differential scanning fluorimetry (nanoDSF) on a Prometheus NT.48 system (NanoTemper, PR001). Samples were prepared by filling standard capillaries (NanoTemper, PR-C006) with 10 μL of 1 mg/mL protein solution. The intrinsic tryptophan fluorescence was measured at 330 and 350 nm, while heating the sample from 20 to 90 °C using a 1 °C/min temperature ramp. The data is analyzed using NanoTempers PR Control software. An exemplary unfolding curve of the three different variants is shown in the Supporting Information (Figure S2). T_m was derived from the maximum of the first derivative of the fluorescence ratio at 350 and 330 nm.[56]

2.2.3. Nanoparticle Assay for Interface Instability. The stability of the different variants against hydrophobic surfaces was evaluated following a recently developed accelerated assay based on nanoparticles.[57−59] Briefly, the variants were mixed in 1:1 ratio with the nanoparticle solution in 1.5 mL reaction tubes (Eppendorf) to reach a final antibody concentration of 0.5 mg/mL and a final protein to nanoparticle surface ratio of 25:1. After incubation for 30 min, aggregates were precipitated

https://doi.org/10.1021/acs.molpharmaceut.1c00469
Mol. Pharmaceutics 2021, 18, 3843−3853

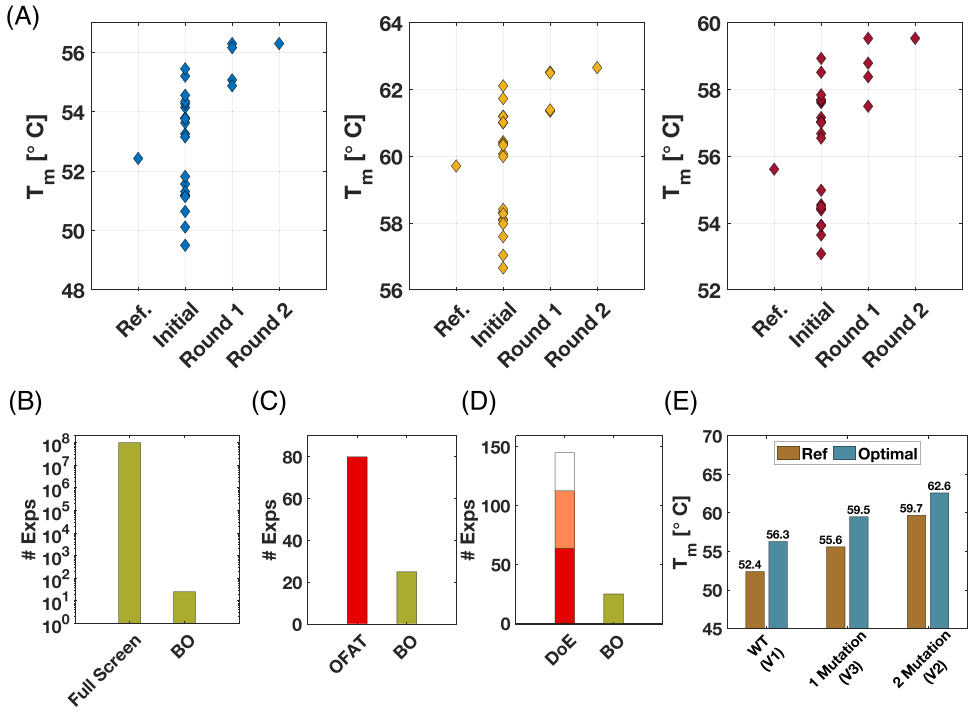


Figure 4. (A) Evolution of T_m values during the different rounds of sequential planning, starting from the Initial round of 20 experiments, followed by "Round 1" of 4 experiments and "Round 2" of the final experiment. The blue, yellow, and red colors indicate Variant 1, Variant 2, and Variant 3, respectively. (B) Comparison between the number of experiments required for a traditional full screen compared to the Bayesian optimization approach (BO) to identify near-optimal formulation. (C) Comparison between the number of experiments required for a one-factor-at-a-time (OFAT) approach that will lead to suboptimal formulation selection (red) and BO that leads to near-optimal formulation (green). (D) Comparison between the number of experiments required for DoE approaches that can lead to optimal formulation with further techniques such as response surface methodology and BO that leads to near-optimal formulation (green). The minimum number of experiments required by the DoE (with not sufficient resolution) is indicated in red. The orange and transparent segments respectively indicate the minimum and maximum number of experiments that are required considering different DoEs (with sufficient resolution). (E) Comparison of increase in T_m values caused by mutation and formulation optimization. WT, 1 Mutation, and 2 Mutation correspond to Variant 1 (V1), Variant 3 (V3), and Variant 2 (V2), respectively.

by $MgCl_2$ addition. The mixture was transferred to filter plates (Corning 96-well filter plates) and centrifuged for 30 min (2500 rpm, 20 °C). The absorbance at 280 nm of the filtrates was then measured in duplicate per sample in a Nanodrop (ThermoScientific). For each condition, the experiment was conducted in duplicate and two control samples were analyzed in the absence of nanoparticles. Control solutions of nanoparticles without protein were also measured in duplicate.

3. RESULTS AND DISCUSSION

In this work, we optimized the thermal stability of tandem scFv variants based on sequences retrieved from the public domain of the marketed antibody Humira. The tandem scFv variants consist of two identical scFvs connected with an additional 25-residue-long glycine−serine peptide linker to provide flexibility for bivalent antigen engagement. Three tandem scFv variants were used in this study including the wild type derived from the sequence of Humira (Variant 1) and two additional variants with improved thermal stability—Variant 2 with two-point mutation in each scFv domain (R16G_D30S) and Variant 3 with one-point mutations in each scFv domain (R16G). Full sequences are given in Table S2. Each individual scFv moiety includes a 21-residue-long glycine−serine linker between VH and VL domains. These variants originate from the internal T_m optimization campaign.

These three variants showed different T_m values spanning a range from 52 to 60 °C in a reference formulation condition (10 mM L-histidine, 140 mM NaCl, pH 6) as indicated in Figure 3A. Our goal was to optimize the composition of the formulation to maximize T_m. Typical formulations consist of a buffering agent, a tonicity modifier, a stabilizer, and a surfactant (Figure 1). In our study, we considered single or multiple options under each category, as tabulated in Figure 3C, which shows the design space of our study. The upper bounds for the different excipients were chosen within the values of currently marketed formulations (Table S1).

Initially, a conservative upper limit for the amino-acid stabilizers and the sugars (Trehalose and mannitol) was chosen (column UB in Figure 3C). First, we obtained an optimal formulation in this design space by applying our Bayesian optimization framework. Next, the design boundaries were extended to the maximum possible values (Extended UB), as schematically shown in Figure 3B, to assess the optimal formulation outside of the explored design space. Using the data and surrogate model generated in the conservative design space as the prior knowledge, further iterations of experiments were suggested using the same framework to identify the optimal formulation in the extended design space.

3.1. Initial Screening Experiments. To first obtain the model parameters described in Section 2, we designed an initial set of 20 training experiments ("initial round") using the conservative design space indicated with the upper bound in Figure 3C. A space-filling design called Latin hypercube sampling[60] was used for continuous factors, namely, salt and

https://doi.org/10.1021/acs.molpharmaceut.1c00469
Mol. Pharmaceutics 2021, 18, 3843−3853

amino acids. pH was considered as a discrete factor with 4 levels (5, 5.5, 6, 6.5), while trehalose, mannitol, and Tween 20 were considered discrete factors with two levels (0, upper bound). A random design was used for all of the discrete factors.

The bivariate distribution of the different excipient compositions suggested by the design for the "initial round" is summarized in Figure S3. From the bivariate plots, it can be seen that the initial design chose a combination of continuous variables (indicated by the crimson dots in Figure S3) to cover the mutual design space. Moreover, the bivariate plots of a continuous factor against a discrete factor (green dots in Figure S3) show that the design was spread across the range of continuous factors at each level of the discrete factors. Finally, the bivariate plot of the discrete factors (gray dots in Figure S3) indicates the presence of combinations of the different levels of the discrete factors. Such a design allows the surrogate model to learn much more efficiently the multivariate interactions of the different factors, which are potentially nonlinear, within the design space.

The label "Initial" in Figure 4A showcases the distribution of T_m measured under the formulation conditions chosen by the space-filling design for the three different variants. The corresponding density distribution plots are provided in Figure S4. As a result of the space-filling design, these initial 20 experiments provided a broad distribution of T_m. A total of 12 (out of the 20) formulation conditions led to a T_m higher than the reference formulation. The maximum observed T_m in this round was 55.4, 62.1, and 58.9 °C for the three variants, respectively, corresponding to a 3, 2.5, and 3.3 °C increment with respect to the values measured in the reference formulation conditions.

3.2. Optimal Formulation Design: Conservative Design Space. As the next step, these "initial round" experiments were used as a training data set to define the Gaussian process (the surrogate model in our study). We noted that we trained a different surrogate model for each of the variants since we performed the optimization of formulation conditions independently for the different variants. However, since the initial design is independent of the response and was designed with the purpose of filling the design space, a common design was used for all of the variants.

By applying the Bayesian optimization framework (Figure 2A) to these trained models, we defined the next sets of formulation conditions, measured the corresponding T_m, updated the model with the new experiments, and iterated until meeting the objective.

Given the considered conservative design space, Bayesian optimization could converge to the condition with maximum T_m within two rounds of experiments, using a total of five additional experiments. Figure 4A shows the evolution of T_m in the different rounds for the three variants, with Round 2 indicating the converged optimal value of T_m. Bayesian optimization could identify the formulation conditions resulting in a T_m of 56.3, 62.7, and 59.5 °C, corresponding to an increase of 1, 0.6, and 0.6 °C with respect to the values observed from the initial experiments. Altogether, we could identify formulation conditions that provided an increase of the T_m of 4, 3, and 4 °C for the three variants with respect to the reference formulation.

The excipient composition corresponding to the maximum T_m for each of the three variants is tabulated in the Supporting Information (Table S3). Since the optimal conditions in this case lie at the boundaries for Variant 1 and Variant 3, it could be argued that any classical DoE methodology such as the fractional factorial design could have identified the optimal formulation. However, applying the formula shown in Figure S1C for eight factors, we estimated that fractional factorial, central composite, and Box−Behnken designs would require 128 (or 64, less precise), 145 (or 81, less precise), and 113 experiments, respectively, to arrive at the same optimal T_m. Bayesian optimization could achieve the same result with only 25 experiments, using at least three times less resources and correspondingly speeding up the design procedure. We compared the number of experiments required by BO with the full screening approach (Figure 4B) and with the suboptimal OFAT (Figure 4C) or classical DoE that could lead to optimal formulation when coupled with other methods such as response surface modeling (RSM) (Figure 4D). We note that classical DoE additionally provides information about the main and interaction effects of the factors on the target variable. This analysis of the influence of each factor and the corresponding nonlinearity is specific to the linear or quadratic model (which are still linear in parameter). In our approach, the influence of specific factors on the target can be inferred by analyzing the length scales of the surrogate GP model.

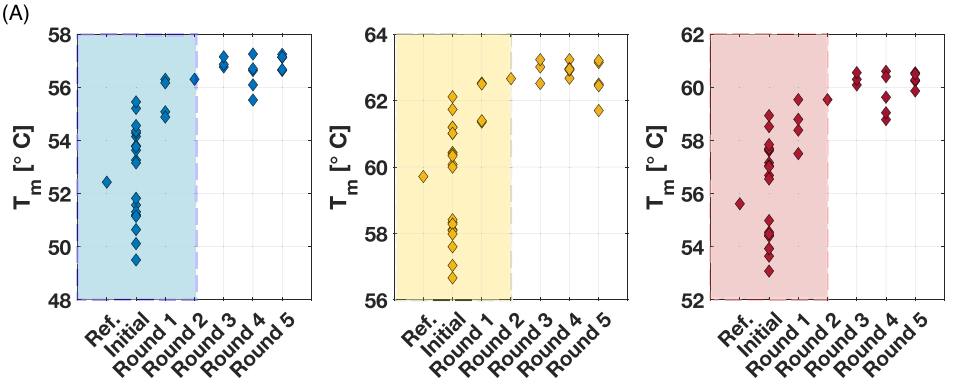
In addition to formulation, the thermal stability of a molecule can be improved by mutations of the sequence. As shown in Figure 4E, in this study, the T_m of the wild-type variant (Variant 1) of 52.4 °C could be increased to 55.6 °C by performing a single-point mutation (Variant 3). With our approach, we could achieve a comparable increase of the T_m (56.3 °C) by changing the buffer formulation. It is important to note that this approach also opens the opportunity of synergistic effects between mutational studies and formulation design. In the optimal formulation conditions, Variant 2 (which exhibits two mutations compared to the wild type) exhibits a T_m as high as 62.7 °C, corresponding to an increase of 10 °C with respect to the wild-type variant in the reference buffer condition. We envision that even higher T_m could be achieved by co-engineering the mutations together with the formulations based on such a Bayesian optimization approach.

These considerations demonstrate that machine learning can not only assist experimental design and accelerate current procedures but also open opportunities to change the current workflow to drastically improve product quality. Specifically, it opens possibilities to parallelize protein engineering, developability, and formulation since the early stages of drug development.

3.3. Optimal Formulation Design: Extended Design Space. A key feature of the Bayesian approach is the use of prior belief about the system as starting point and the continuous update of this belief using new experiments. The prior belief could be based on expert knowledge or based on historical experiments performed in the same system or similar systems. We expect that this ability to transfer knowledge within and across similar systems to aid experimental design will be very impactful in the biopharmaceutical industry, where typically multiple products are present in the pipeline. We demonstrate the advantage of this approach by extending the design space considered previously to the extended UB indicated in Figure 3C. For a classical DoE, this operation would require redetermining the design from scratch, as now one of the levels ("high" in this case) has changed due to the re-definition of the design boundaries. All of the data and the

https://doi.org/10.1021/acs.molpharmaceut.1c00469
Mol. Pharmaceutics 2021, 18, 3843−3853

(A)



(B)

Excipient Names	pH	NaCl (mM)	L- Arginine (mM)	L-Lysine (mM)	L-Proline (mM)	Trehalose (w/v %)	Mannitol (w/v %)	Tween 20 (v/v %)	T_m (°C)
Variant 1	6.5	0.0	0.0	250.0	181.5	10.0	5.0	0.0	57.24
Variant 2	6.5	111.1	0.0	74.1	201.9	10.0	5.0	0.0	63.23
Variant 3	6.5	0.0	0.0	250.0	123.6	8.0	5.0	0.0	60.55

Figure 5. (A) Evolution of the T_m values during the different additional rounds, "Round 3", "Round 4" and "Round 5", of the sequential planning with the extended design space (indicated as Extended UB in Figure 3C). The blue, yellow, and red colors indicate Variant 1, Variant 2, and Variant 3, respectively, and the experiments performed with the conservative design space is represented in the shaded region. (B) Optimal composition of the excipients resulting in maximal T_m for the three different variants.

information there-in cannot be used anymore even if only such a subtle modification is imposed.

In contrast, with the Bayesian optimization strategy, this adjustment can be done in a straightforward manner. The posterior distribution obtained from the study performed with conservative bounds can be used as the prior belief for the extended design space. Following the same sequential design procedure, new experiments in the extended design space can be generated with the same objective to maximize the T_m, as demonstrated in Figure 5A ("Round 3", "Round 4", "Round 5").

With the extended boundaries, the T_m values could be increased further to reach values of 57.24, 63.23, and 60.55 °C for the three variants, respectively. These values are, respectively, ∼5, 3.5, and 5 °C higher than the T_m values of the three variants in the reference formulation. Additionally, the optimal formulation could be obtained using only 13 additional experiments when knowledge (in the form of priors) from conservative design space could be transferred through the surrogate model. Furthermore, in contrast with the conservative design space, in the extended design space, the optimal formulation (tabulated in Figure 5B) does not lie along the boundaries of the different excipients. Thus, classical DoE such as full/fractional-factorial designs would not be able to identify these conditions.

Given the exploration−exploitation search of Bayesian optimization, during the sequential experimental campaign, we could identify multiple formulations that had T_m values similar to the observed maximum T_m value for the respective variant (Table S4). This indicates that there could be multiple combinations of the excipients that could lead to desired objective (maximum T_m), signifying that the objective function has a nonconvex behavior in the design space. This situation can change when multiple properties must be simultaneously optimized, as discussed in Section 3.4.

3.4. Trade-Offs with Multiple Objectives. So far, we could demonstrate that the Bayesian optimization approach can considerably accelerate the identification of optimal formulations to maximize a single biophysical property. In practice, the development of a biological drug requires the optimization of multiple biophysical properties. The use of such algorithms to find optimal formulations becomes even more valuable with the increasing number of target properties.

Indeed, when considering multiple target properties, it might not be possible to achieve the individual optimality of each target property simultaneously. For instance, formulation conditions resulting in the highest T_m might not necessarily represent the ideal solutions to protect against other stresses. Often a trade-off needs to be established. In such cases, multiple optimal solutions (also called the pareto solutions) lie on a curve known as the pareto front, as shown in Figure S5A. When the underlying response surface for each objective is known, determining the pareto front is straightforward. However, for our system, where the mathematical representation of underlying response surface is unknown, the optimization algorithm must converge to these pareto solutions while simultaneously learning the underlying response function and ensuring minimal utilization of resources. In this context, Bayesian optimization is a very suitable method for such task due to its exploration−exploitation property.

Similar to the case of single objective optimization (e.g., maximization of T_m), Bayesian optimization can also be used for multiobjective optimization to traverse the design space strategically such that the pareto front can be obtained quickly and with minimal requirement of experiments. To illustrate this concept, we selected Variant 2 and simultaneously monitored two biophysical properties, adding to the T_m the stability of the proteins toward hydrophobic interfaces, which was measured with a nanoparticle-based assay recently developed in our laboratory.[57,58] This assay evaluates the %

https://doi.org/10.1021/acs.molpharmaceut.1c00469
Mol. Pharmaceutics 2021, 18, 3843−3853

of monomer loss in the presence of a negatively charged hydrophobic interface, with a higher value of monomer in solution indicating higher stability. A subset of excipients (pH, sodium chloride, trehalose, mannitol, and Tween 20) were considered within the ranges indicated in Figure S5B. Following the same procedure adopted earlier for the maximization of T_m, we performed an initial design of experiments for 15 formulation conditions using a space-filling strategy. Subsequently, the multiobjective Bayesian optimization was applied to iteratively determine the next batches of experiments to learn the pareto front. Figure 6 plots the value of the two properties, i.e., T_m and % monomer in solution, for the formulation conditions tested in the different iterations.

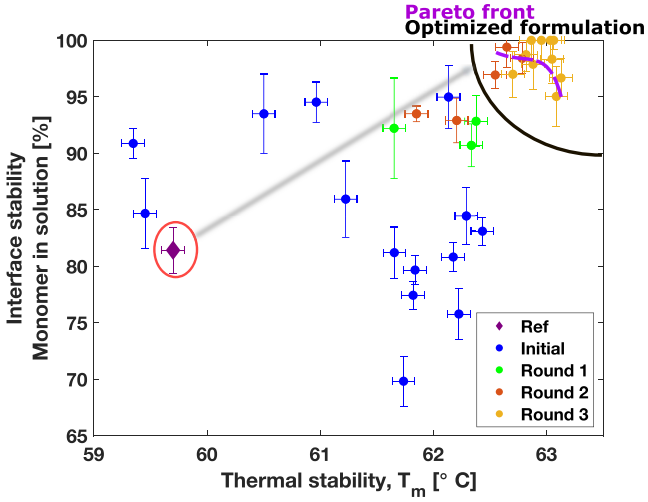


Figure 6. Trade-off between two biophysical properties, thermal stability (T_m) and interface stability (expressed as the % monomer in solution from nano-particle assay), for Variant 2. The violet dot line indicates the pareto front obtained after Round 3. "Ref" indicates the initial reference formulation.

The initial screening of formulation conditions resulted in a broad range of values for both properties. The T_m spanned from 59 to 62 °C, while the % monomer loss varied from 5 to 35%. The subsequent iterations led to a combination of target values that approached the pareto front. Finally, the formulations tested in "Round 3" and (partially) "Round 2" provided the set of optimal solutions resulting in the pareto front, shown with the violet dashed line in Figure 6. With the optimized formulation, T_m could be improved by about 3.2 °C (from 59.7 to 62.9 °C) while the % monomer loss from the nanoparticle assay could be reduced by 15.5% (from 18.1 to 2.5%) in comparison to the reference formulation (Figure S5C).

Only 33 experiments were required to obtain the optimal formulation that simultaneously optimized both properties. As before, this number is ∼3 times smaller than the traditional DoE methods. Additionally, the experimental conditions converging to the pareto region (Table S5) do not lie at the boundaries or at the center point of the design space. Thus, with a classical DoE, an additional optimization step must be performed (also called the response surface methodology or RSM) to arrive at the optimum inside the design space. However, the RSM will be able to identify these conditions only when the underlying assumption of quadratic response surface holds. Moreover, in multiobjective optimization, classical DoE methods cannot identify the full pareto front but only a single objective derived from the weighted sum of the multiple objectives, which may converge to a single point on the pareto front.

We finally note that the T_m is optimized by formulations that lack surfactant, while some amounts of polysorbate is required for interface stability. Our method identifies the minimum amount of surfactant that represents the best compromise. This is important to optimize different biophysical properties while minimizing the risks associated with polysorbate degradation.[45,46]

4. DISCUSSION AND CONCLUSIONS

Formulation design, currently performed by analytical screening and prior knowledge, can largely benefit from computational tools based on machine learning. These are crucial not only to extract patterns from large sets of data but also to guide experimentation.[22] In this work, we present a surrogate model-based sequential optimization, called Bayesian optimization, to identify a formulation that optimizes the thermal stability of three different tandem single chains. We could demonstrate that the framework is capable of converging to the optimal condition using only 25 experiments. This is approximately one-third of the experimental burden required by classical DoE and orders of magnitude lower in comparison to the millions of experiments required under the full screening or grid search approach.

We additionally demonstrate how data or information generated in a previous campaign can serve as an efficient starting point (so-called "prior-belief") for a new campaign, something which is not feasible with the traditional DoE.

Overall, these two combined aspects, the high speed of converging to optimal conditions and the ability to incorporate and transfer prior knowledge across campaigns, lead to a considerable acceleration in the development time scale and also minimize the amount of resources required. For instance, in this case, we required only 125 μg of sample in comparison to the ∼500 g of sample required for full screening study (to execute 10^8 experiments, c.f. Figure 4B) or 400 μg for OFAT study or around 600 μg for DoE study. We expect that the qualitative trend of the number of experiments required by the full-screen, OFAT, classical DoE, and Bayesian optimization will remain the same independently of the molecule and the number and type of excipients, while the actual numbers of experiments may strongly differ.

These aspects become even more important with increasing the number of properties of the molecule to be optimized. This not only drastically increases the design space but also requires the identification of optimal formulations that guarantee a delicate trade-off among multiple properties. In this work, we demonstrated this concept through an exemplary study that considered both thermal and interface stabilities, which highlighted the need of compromises in the final formulation and minimized the amount of required surfactant.

We expect that Bayesian optimization for multiobjective optimization will be very useful to strategically plan experiments and reach trade-offs between properties. In analogy with protein engineering, where machine learning methods can accelerate directed evolution of proteins[43,61−63] and antibodies,[64] Bayesian optimization can complement rational design of formulations to accelerate their optimization.

As next steps, further robustness can be added into the surrogate model by incorporating physically relevant con-

https://doi.org/10.1021/acs.molpharmaceut.1c00469
Mol. Pharmaceutics 2021, 18, 3843−3853

straints (e.g., physiological osmolality) and information gained by experience (Figure 1E). Such approaches are gaining popularity in bioprocessing under the so-called "hybrid modeling" framework,[65−67] which combines knowledge-based and data-driven models. Extending this concept to formulation optimization will be a very interesting direction of future works.

Moreover, since the success of the Bayesian optimization framework depends on the feedback from experiments, the approach will highly benefit from advances in emerging high-throughput screening (HTS) technologies, such as methods based on microfluidics[68] and robotic platforms that can ensure rapid inflow of data. The coupling of advances in experimental methods, algorithms, and in silico predictors[69] is expected to further accelerate biotherapeutic formulation screening[22] (Figure 1E).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.1c00469.

Principles of design of experiments and Bayesian optimization and details and plots of auxiliary results and tests (PDF)

Experimental data for the multiobjective optimization (Table S1) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Paolo Arosio − *Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, Swiss Federal Institute of Technology, 8093 Zurich, Switzerland;* orcid.org/0000-0002-2740-1205; Email: paolo.arosio@chem.ethz.ch

Authors

Harini Narayanan − *Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, Swiss Federal Institute of Technology, 8093 Zurich, Switzerland;* orcid.org/0000-0003-4545-4885

Fabian Dingfelder − *Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, Swiss Federal Institute of Technology, 8093 Zurich, Switzerland; Department of Biophysics and Injectable Formulation, Global Research Technologies, Novo Nordisk A/S, Måløv 2760, Denmark*

Itzel Condado Morales − *Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, Swiss Federal Institute of Technology, 8093 Zurich, Switzerland; Department of Biophysics and Injectable Formulation, Global Research Technologies, Novo Nordisk A/S, Måløv 2760, Denmark;* orcid.org/0000-0001-7592-4556

Bhargav Patel − *Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, Swiss Federal Institute of Technology, 8093 Zurich, Switzerland*

Kristine Enemærke Heding − *Department of Biophysics and Injectable Formulation, Global Research Technologies, Novo Nordisk A/S, Måløv 2760, Denmark*

Jais Rose Bjelke − *Department of Purification Technologies, Global Research Technologies, Novo Nordisk A/S, Måløv 2760, Denmark*

Thomas Egebjerg − *Department of Mammalian Expression, Global Research Technologies, Novo Nordisk A/S, Måløv 2760, Denmark*

Alessandro Butté − *DataHow AG, 8600 Dübendorf, Switzerland*

Michael Sokolov − *DataHow AG, 8600 Dübendorf, Switzerland;* orcid.org/0000-0001-8396-4099

Nikolai Lorenzen − *Department of Biophysics and Injectable Formulation, Global Research Technologies, Novo Nordisk A/S, Måløv 2760, Denmark*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.molpharmaceut.1c00469

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We want to thank Annette Juhl Gajhede for help with thermal stability measurements. F.D. and I.C.M were supported by the STAR program at Novo Nordisk A/S.

■ REFERENCES

(1) Walsh, G. Biopharmaceutical Benchmarks 2018. *Nat. Biotechnol.* **2018**, *36*, 1136−1145.

(2) Elgundi, Z.; Reslan, M.; Cruz, E.; Sifniotis, V.; Kayser, V. The State-of-Play and Future of Antibody Therapeutics. *Adv. Drug Delivery Rev.* **2017**, *122*, 2−19.

(3) Chi, E. Y.; Krishnan, S.; Randolph, T. W.; Carpenter, J. F. Physical Stability of Proteins in Aqueous Solution: Mechanism and Driving Forces in Nonnative Protein Aggregation. *Pharm. Res.* **2003**, *20*, 1325−1336.

(4) Randolph, T. W.; Carpenter, J. F. Engineering Challenges of Protein Formulations. *AIChE J.* **2007**, *53*, 1902−1907.

(5) Gentiluomo, L.; Svilenov, H. L.; Augustijn, D.; El Bialy, I.; Greco, M. L.; Kulakova, A.; Indrakumar, S.; Mahapatra, S.; Morales, M. M.; Pohl, C.; Roche, A.; Tosstorff, A.; Curtis, R.; Derrick, J. P.; Nørgaard, A.; Khan, T. A.; Peters, G. H. J.; Pluen, A.; Rinnan, Å.; Streicher, W. W.; Van Der Walle, C. F.; Uddin, S.; Winter, G.; Roessner, D.; Harris, P.; Frieß, W. Advancing Therapeutic Protein Discovery and Development through Comprehensive Computational and Biophysical Characterization. *Mol. Pharm.* **2020**, *17*, 426−440.

(6) Krause, M. E.; Sahin, E. Chemical and Physical Instabilities in Manufacturing and Storage of Therapeutic Proteins. *Curr. Opin. Biotechnol.* **2019**, *60*, 159−167.

(7) Jiskoot, W.; Randolph, T. W.; Volkin, D. B.; Middaugh, C. R.; Schöneich, C.; Winter, G.; Friess, W.; Crommelin, D. J. A.; Carpenter, J. F. Protein Instability and Immunogenicity: Roadblocks to Clinical Application of Injectable Protein Delivery Systems for Sustained Release. *J. Pharm. Sci.* **2012**, *101*, 946−954.

(8) Liu, Y.; Caffry, I.; Wu, J.; Geng, S. B.; Jain, T.; Sun, T.; Reid, F.; Cao, Y.; Estep, P.; Yu, Y.; Vásquez, M.; Tessier, P. M.; Xu, Y. High-Throughput Screening for Developability during Early-Stage Antibody Discovery Using Self-Interaction Nanoparticle Spectroscopy. *MAbs* **2014**, *6*, 483−492.

(9) Wolf Pérez, A. M.; Sormanni, P.; Andersen, J. S.; Sakhnini, L. I.; Rodriguez-Leon, I.; Bjelke, J. R.; Gajhede, A. J.; De Maria, L.; Otzen, D. E.; Vendruscolo, M.; Lorenzen, N. In Vitro and in Silico Assessment of the Developability of a Designed Monoclonal Antibody Library. *MAbs* **2019**, *11*, 388−400.

(10) Jain, T.; Sun, T.; Durand, S.; Hall, A.; Houston, N. R.; Nett, J. H.; Sharkey, B.; Bobrowicz, B.; Caffry, I.; Yu, Y.; et al. Biophysical Properties of the Clinical-Stage Antibody Landscape. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 944−949.

(11) Rabia, L. A.; Desai, A. A.; Jhajj, H. S.; Tessier, P. M. Understanding and Overcoming Trade-Offs between Antibody Affinity, Specificity, Stability and Solubility. *Biochem. Eng. J.* **2018**, *137*, 365−374.
(12) Uchiyama, S. Liquid Formulation for Antibody Drugs. *Biochim. Biophys. Acta* **2014**, *1844*, 2041−2052.
(13) Wang, W. Instability, Stabilization, and Formulation of Liquid Protein Pharmaceuticals. *Int. J. Pharm.* **1999**, *185*, 129−188.
(14) Zbacnik, T. J.; Holcomb, R. E.; Katayama, D. S.; Murphy, B. M.; Payne, R. W.; Coccaro, R. C.; Evans, G. J.; Matsuura, J. E.; Henry, C. S.; Manning, M. C. Role of Buffers in Protein Formulations. *J. Pharm. Sci.* **2017**, *106*, 713−733.
(15) Falconer, R. J. Advances in Liquid Formulations of Parenteral Therapeutic Proteins. *Biotechnol. Adv.* **2019**, *37*, No. 107412.
(16) Ionova, Y.; Wilson, L. Biologic Excipients: Importance of Clinical Awareness of Inactive Ingredients. *PLoS One* **2020**, *15*, No. e0235076.
(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. *Adv. Drug Delivery Rev.* **2011**, *63*, 1118−1159.
(18) Chi, E. Y. Excipients Used in Biotechnology Products. In *Pharmaceutical Excipients: Properties, Functionality, and Applications in Research and Industry*; John Wiley & Sons, Inc., 2016; pp 145−198.
(19) Rayaprolu, B. M.; Strawser, J. J.; Anyarambhatla, G. Excipients in Parenteral Formulations: Selection Considerations and Effective Utilization with Small Molecules and Biologics. *Drug Dev. Ind. Pharm.* **2018**, *44*, 1565−1571.
(20) Kingwell, K. Excipient Developers Call for Regulatory Facelift. *Nat. Rev. Drug Discovery* **2020**, *19*, 823−824.
(21) Kamerzell, T. J.; Middaugh, C. R. Prediction Machines: Applied Machine Learning for Therapeutic Protein Design and Development. *J. Pharm. Sci.* **2021**, *110*, 665−681.
(22) Narayanan, H.; Dingfelder, F.; Butté, A.; Lorenzen, N.; Sokolov, M.; Arosio, P. Machine Learning for Biologics: Opportunities for Protein Engineering, Developability, and Formulation. *Trends Pharmacol. Sci.* **2021**, *42*, 151−165.
(23) 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. *Mol. Pharm.* **2015**, *12*, 4478−4487.
(24) Strickley, R. G.; Lambert, W. J. A Review of Formulations of Commercially Available Antibodies. *J. Pharm. Sci.* **2021**, *110*, 2590−2608.
(25) *Rational Design of Stable Protein Formulations*, 1st ed.; In Carpenter, J. F.; Manning, M. C., Eds.; Springer US, 2002.
(26) Kumar, S.; Singh, S. K. *Developability of Biotherapeutics: Computational Approaches*; CRC Press, 2015.
(27) 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. *J. Pharm. Sci.* **2020**, *109*, 584−594.
(28) Thakkar, S. V.; Joshi, S. B.; Jones, M. E.; Sathish, H. A.; Bishop, S. M.; Volkin, D. B.; Middaugh, C. R. Excipients Differentially Influence the Conformational Stability and Pretransition Dynamics of Two IgG1 Monoclonal Antibodies. *J. Pharm. Sci.* **2012**, *101*, 3062−3077.
(29) Manikwar, P.; Majumdar, R.; Hickey, J. M.; Thakkar, S. V.; Samra, H. S.; Sathish, H. A.; Bishop, S. M.; Middaugh, C. R.; Weis, D. D.; Volkin, D. B. Correlating Excipient Effects on Conformational and Storage Stability of an IgG1 Monoclonal Antibody with Local Dynamics as Measured by Hydrogen/Deuterium-Exchange Mass Spectrometry. *J. Pharm. Sci.* **2013**, *102*, 2136−2151.
(30) Cheng, W.; Joshi, S. B.; He, F.; Brems, D. N.; He, B.; Kerwin, B. A.; Volkin, D. B.; Middaugh, C. R. Comparison of High-Throughput Biophysical Methods to Identify Stabilizing Excipients for a Model IgG2 Monoclonal Antibody: Conformational Stability and Kinetic Aggregation Measurements. *J. Pharm. Sci.* **2012**, *101*, 1701−1720.
(31) FDA-Inactive Ingredient Database. https://www.fda.gov/drugs/drug-approvals-and-databases/inactive-ingredients-database-download (accessed January 2, 2021).
(32) Wang, W.; Ohtake, S. Science and Art of Protein Formulation Development. *Int. J. Pharm.* **2019**, *568*, No. 118505.
(33) Falconer, R. J. Biotechnology Advances. *Biotechnol. Adv.* **2019**, *37*, 993.
(34) Roessl, U.; Humi, S.; Leitgeb, S.; Nidetzky, B. Design of Experiments Reveals Critical Parameters for Pilot-Scale Freeze-and-Thaw Processing of L-Lactic Dehydrogenase. *Biotechnol. J.* **2015**, *10*, 1390−1399.
(35) Chavez, B. K.; Agarabi, C. D.; Read, E. K.; Boyne, M. T.; Khan, M. A.; Brorson, K. A. Improved Stability of a Model IgG3 by DoE-Based Evaluation of Buffer Formulations. *BioMed Res. Int.* **2016**, *2016*, No. 2074149.
(36) Manning, M. C.; Liu, J.; Li, T.; Holcomb, R. E. *Rational Design of Liquid Formulations of Proteins*, 1st ed.; Elsevier Inc., 2018; Vol. *112*.
(37) Greenhill, S.; Rana, S.; Gupta, S.; Vellanki, P.; Venkatesh, S. Bayesian Optimization for Adaptive Experimental Design: A Review. *IEEE Access* **2020**, *8*, 13937−13948.
(38) Brochu, E.; Cora, V. M.; de Freitas, N. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning, 2010. arXiv:1012.2599, https://arxiv.org/abs/1012.2599v1.
(39) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. *Adv. Neural Inf. Process. Syst.* **2012**, *4*, 2951−2959.
(40) Berkenkamp, F.; Krause, A.; Schoellig, A. P.; Apr, R. O. *Bayesian Optimization with Safety Constraints: Safe and Automatic Parameter Tuning in Robotics*, 2015, arXiv:1602.04450, https://arxiv.org/abs/1602.04450.
(41) Lyu, W.; Xue, P.; Yang, F.; Yan, C.; Hong, Z.; Zeng, X.; Zhou, D.; Member, S. An Efficient Bayesian Optimization Approach for Automated Optimization of Analog Circuits. *IEEE Trans. Circuits Syst.* **2018**, *65*, 1954−1967.
(42) González, J.; Longworth, J.; James, D. C.; Lawrence, N. D. Bayesian Optimization for Synthetic Gene Design, 2015. arXiv:1505.01627, https://arxiv.org/abs/1505.01627.
(43) Romero, P. A.; Krause, A.; Arnold, F. H. Navigating the Protein Fitness Landscape with Gaussian Processes. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 1−9.
(44) Lookman, T.; Balachandran, P. V.; Xue, D.; Yuan, R. Active Learning in Materials Science with Emphasis on Adaptive Sampling Using Uncertainties for Targeted Design. *npj Comput. Mater.* **2019**, *5*, No. 21.
(45) Kerwin, B. A. Polysorbates 20 and 80 Used in the Formulation of Protein Biotherapeutics: Structure and Degradation Pathways. *J. Pharm. Sci.* **2008**, *97*, 2924−2935.
(46) Grabarek, A. D.; Bozic, U.; Rousel, J.; Menzen, T.; Kranz, W.; Wuchner, K.; Jiskoot, W.; Hawe, A. What Makes Polysorbate Functional? Impact of Polysorbate 80 Grade and Quality on IgG Stability During Mechanical Stress. *J. Pharm. Sci.* **2020**, *109*, 871−880.
(47) Cox, D. R.; Reid, N. *The Theory of the Design of Experiments*; Chapman & Hall/CRC: Boca Raton, 2000.
(48) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; et al. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Chin. J. Evidence-Based Med.* **2014**, *14*, 1270−1275.
(49) Liu, D. C.; Jorge, N. On the Limited Memory BFGS Method for Large Scale Optimization. *Math. Program.* **1989**, *45*, 503−528.
(50) Berkenkamp, F.; Schoellig, A. P.; Krause, A. No-Regret Bayesian Optimization with Unknown Hyperparameters. *J. Mach. Learn. Res.* **2019**, *20*, 1−24.
(51) González, J.; Dai, Z.; Hennig, P.; Lawrence, N. In *Batch Bayesian Optimization via Local Penalization*, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, 2016; pp 648−657.
(52) Azimi, J.; Fern, A.; Fern, X. Z. In *Batch Bayesian Optimization via Simulation Matching*, Advances in Neural Information Processing

Systems 23: 24th Annual Conference on Neural Information Processing Systems, 2010; pp 1−9.
(53) Chevalier, C.; Ginsbourger, D. Fast Computation of the Multi-Points Expected Improvement with Applications in Batch Selection. *Lect. Notes Comput. Sci.* **2013**, *7997*, 59−69.
(54) Azimi, J.; Jalali, A.; Fern, X. Z. In *Hybrid Batch Bayesian Optimization*, Proceedings of the 29th International Conference on Machine Learning (ICML 2012), 2012; pp 1215−1222.
(55) Ginsbourger, D.; Le Riche, R.; Carraro, L. Kriging Is Well-Suited to Parallelize Optimization. In *Computational Intelligence in Expensive Optimization Problems*; Springer: Berlin, 2010; pp 131−162.
(56) Wen, J.; Lord, H.; Knutson, N.; Wikström, M. Nano Differential Scanning Fluorimetry for Comparability Studies of Therapeutic Proteins. *Anal. Biochem.* **2020**, *593*, No. 113581.
(57) Kopp, M. R. G.; Capasso Palmiero, U.; Arosio, P. A Nanoparticle-Based Assay to Evaluate Surface-Induced Antibody Instability. *Mol. Pharm.* **2020**, *17*, 909−918.
(58) Kopp, M. R. G.; Wolf Pérez, A. M.; Zucca, M. V.; Capasso Palmiero, U.; Friedrichsen, B.; Lorenzen, N.; Arosio, P. An Accelerated Surface-Mediated Stress Assay of Antibody Instability for Developability Studies. *MAbs* **2020**, *12*, No. 1815995.
(59) Grigolato, F.; Colombo, C.; Ferrari, R.; Rezabkova, L.; Arosio, P. Mechanistic Origin of the Combined Effect of Surfaces and Mechanical Agitation on Amyloid Formation. *ACS Nano* **2017**, *11*, 11358−11367.
(60) McKay, M. D.; Beckman, R. J.; Conover, W. J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **2000**, *42*, 55−61.
(61) Romero, P. A.; Arnold, F. H. Exploring Protein Fitness Landscapes by Directed Evolution. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 866−876.
(62) Yang, K. K.; Wu, Z.; Arnold, F. H. Protein Engineering. *Nat. Methods* **2019**, *16*, 687−694.
(63) Wittmann, B. J.; Johnston, K. E.; Wu, Z.; Arnold, F. H. Advances in Machine Learning for Directed Evolution. *Curr. Opin. Struct. Biol.* **2021**, *69*, 11−18.
(64) Mason, D. M.; Friedensohn, S.; Weber, C. R.; Jordi, C.; Wagner, B.; Meng, S. M.; Ehling, R. A.; Bonati, L.; Dahinden, J.; Gainza, P.; Correia, B. E.; Reddy, S. T. Optimization of Therapeutic Antibodies by Predicting Antigen Specificity from Antibody Sequence via Deep Learning. *Nat. Biomed. Eng.* **2021**, *5*, 600−612.
(65) Narayanan, H.; Luna, M.; Sokolov, M.; Arosio, P.; Butté, A.; Morbidelli, M. Hybrid Models Based on Machine Learning and an Increasing Degree of Process Knowledge: Application to Capture Chromatographic Step. *Ind. Eng. Chem. Res.* **2021**, *60*, 10466−10478.
(66) Narayanan, H.; Seidler, T.; Luna, M. F.; Sokolov, M.; Morbidelli, M.; Butté, A. Hybrid Models for the Simulation and Prediction of Chromatographic Processes for Protein Capture. *J. Chromatogr. A* **2021**, *1650*, No. 462248.
(67) Narayanan, H.; Sokolov, M.; Morbidelli, M.; Butté, A. A New Generation of Predictive Models: The Added Value of Hybrid Models for Manufacturing Processes of Therapeutic Proteins. *Biotechnol. Bioeng.* **2019**, *116*, 2540−2549.
(68) Kopp, M. R. G.; Arosio, P. Microfluidic Approaches for the Characterization of Therapeutic Proteins. *J. Pharm. Sci.* **2018**, *107*, 1228−1236.
(69) Sormanni, P.; Aprile, F. A.; Vendruscolo, M. Third Generation Antibody Discovery Methods:: In Silico Rational Design. *Chem. Soc. Rev.* **2018**, *47*, 9137−9157.

https://doi.org/10.1021/acs.molpharmaceut.1c00469
Mol. Pharmaceutics 2021, 18, 3843−3853