Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6dd00035e

Spectroscopy-assisted Bayesian optimization for efficient refolding of inclusion body proteins

Florian Gisperg, †[ab] Robert Klausser, †[ab] Matthias Kierein, [a] Eva Prada Brichtova, [ab] Mohamed Elshazly, [ab] Julian Kopp [ab] and Oliver Spadiut *[ab]

The production of recombinant proteins in *Escherichia coli* often yields insoluble inclusion bodies, which require denaturation and refolding to obtain the native product. The protein refolding step usually represents a major bottleneck. Conventional development and optimization typically rely on sequential design of experiments with high-performance liquid chromatography readouts. This approach is slow, labor-intensive, and requires an established chromatographic method as well as purified protein standards. At the beginning of process development, these prerequisites may not be met—especially for proteins that can only be expressed as inclusion bodies. We introduce a more efficient, data-driven workflow that pairs Bayesian optimization with a rapid, in-line readout from intrinsic tryptophan fluorescence. Using a disulfide-bonded single-chain variable fragment, we explored a five-dimensional design space of refolding buffer composition (dithiothreitol, oxidized glutathione, dilution factor, pH, and final urea concentration) guided by two spectroscopy-derived objectives. We showed that the spectral shift correlates with chromatographic yields, supporting its use as a fit-for-purpose sensor to guide process development with 25 experiments. Bayesian optimization identified conditions that delivered a refolded protein concentration of 1.29 ± 0.06 g L^{-1} at 58.7 ± 1.3% refolding yield with a dilution factor of 3.14, whereas a three-stage design of experiments with more than 60 experiments concluded at 0.37 ± 0.02 g L^{-1} and 61.4 ± 3.1% with a dilution factor of 11.39. Thus, the presented workflow achieved roughly 3.5-fold higher product concentration at comparable yield, while operating at substantially higher protein concentrations. Therefore, spectroscopy-assisted Bayesian optimization was found to be a practical, sample-efficient tool for refolding optimization that is especially valuable in early development stages.

Received 22nd January 2026
Accepted 14th May 2026

DOI: 10.1039/d6dd00035e

rsc.li/digitaldiscovery

1 Introduction

Inclusion bodies (IBs) are a common outcome of fast recombinant protein expression in *Escherichia coli*.[1,2] While historically viewed as a liability, the perception has shifted: the formation of IBs represents an attractive production strategy compared to soluble expression, as it can increase space-time yields, simplify primary recovery, and protect products from proteolysis.[2,3] However, IB formation also creates the need for efficient solubilization and refolding procedures to obtain a functional protein. For many targets, these downstream processing steps remain the main bottlenecks in processing of IBs and become further complicated with disulfide bonds or cofactors.[4,5] Development and optimization of protein refolding processes is inherently multi-dimensional. Factors such as denaturant dilution, redox environment, pH, protein concentration, and chemical additives, determine the balance between protein folding and aggregation in a highly interactive manner. This interplay leads to nonlinear protein-specific responses that are difficult to predict in advance.[6,7] This complexity is especially problematic in early process development. The material is scarce, constructs evolve, and fit-for-purpose reference standards for analytics are often unavailable.[8,9] Producing such a reference material itself may require workable refolding conditions.

In practice, protein refolding process development has relied on Design of Experiments (DoE), coupled with off-line or at-line chromatographic protein quantification.[10–14] DoE is an attractive approach because it offers efficiency over one-factor-at-a-time approaches, allows the discovery of interactions, and provides interpretable response surfaces aligning with Quality-by-Design (QbD) documentation.[12] However, this approach still faces some limitations. Fractional designs can become experiment-intensive as the number of factors and interactions grows, and low-order polynomial models may be too rigid to

[a]*Research Division Integrated Bioprocess Development, Institute of Chemical, Environmental and Bioscience Engineering, Technische Universität Wien, Gumpendorferstraße 1A, Vienna 1060, Austria. E-mail: oliver.spadiut@tuwien.ac.at*
[b]*Christian Doppler Laboratory for Inclusion Body Processing 4.0, Institute of Chemical, Environmental and Bioscience Engineering, Technische Universität Wien, Gumpendorferstraße 1A, Vienna 1060, Austria*
† These authors contributed equally to this work.

capture protein- and buffer-specific response landscapes.[15] Furthermore, HPLC is quantitative, but still mostly limited to at-line application. Depending on the specific method, it is often slow, labor-intensive, and expensive when columns need to be replaced frequently. Additionally, a purified protein standard is expended for evaluation.

Therefore, Process Analytical Technology (PAT) implementation for protein refolding is still an exception rather than a rule.[1] Recent work has begun to address this gap with smarter data acquisition and inference layered on top of standard analytics, for example, particle-filter-based state estimation using chromatography signals to track refolding progress despite delayed measurement and noise.[16] In parallel, intrinsic tryptophan fluorescence has emerged as a promising label-free spectroscopic PAT tool for protein refolding, reflecting changes in the local chemical environment of tryptophan (Trp) and tyrosine (Tyr) residues during the protein's transition from the denatured to the native state.[17] Several studies show that fluorescence features can act as soft sensors, enabling online assessment of the refolding state and even feedback control concepts.[18,19] Our group previously demonstrated a soft sensor based on simple correlations of intrinsic Trp and Tyr fluorescence spectral metrics, and also enforced physical constraints using a particle filter that enabled rapid assessment of all folding states.[19] Relying on the same principle of Trp fluorescence in this study, we used the total delta of fluorescence wavelength shift at the endpoint of the folding reaction as a proxy response for the refolding yield (schematically depicted in Fig. 1). We used this proxy response to guide refolding process development in a Bayesian optimization (BO) campaign. Bayesian optimization has recently gained traction as a sample-efficient, model-based strategy for experimental design in chemical and bioprocess engineering.[20–25] BO couples a probabilistic surrogate, often a Gaussian process (GP), with an acquisition function that balances exploration of uncertain regions and exploitation of promising ones.[26] This paradigm is well-suited for optimizing black-box systems with limited prior knowledge of the underlying input–output relationship, particularly when experiments are noisy, as well as material-, time-, and labor-intensive.[27] Relying on such a flexible GP surrogate, BO can capture nonlinear and interacting responses more flexibly than low-order response-surface models.[26] This flexibility is relevant for protein refolding, where multiple process variables jointly affect folding, misfolding, and aggregation in a protein-specific manner.[28] Related advantages of nonlinear machine-learning enhanced DoE over conventional DoE analysis have also been reported for ranibizumab refolding.[29] Furthermore, in bioprocess and chemical engineering optimization tasks, BO has been reported to identify high-performing conditions with fewer experiments than conventional DoE workflows.[30–32]

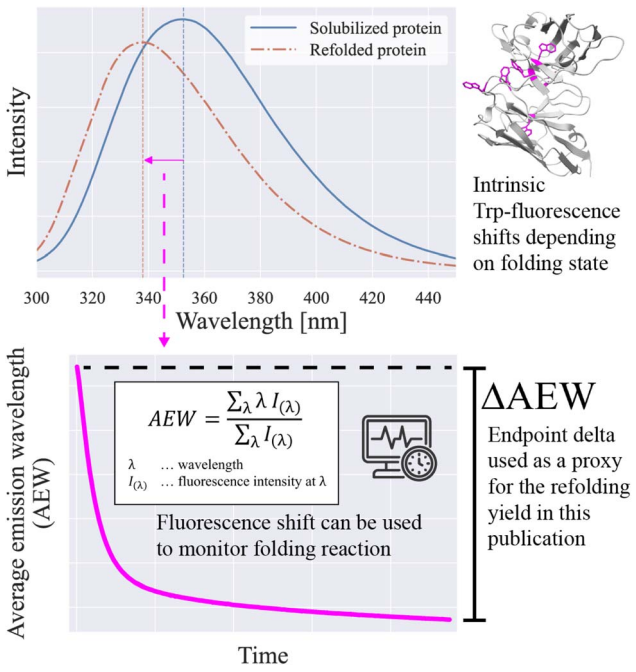

Fig. 1 Schematic depiction of the Trp fluorescence measurement principle used to monitor protein folding reactions with a soft sensor in a previous publication.[19] In this study, the reaction endpoint ΔAEW was used as a proxy response for the refolding yield to guide refolding process optimization.

By combining these advances in PAT and experimental design, we present a novel approach to protein refolding process development. We integrated intrinsic Trp and Tyr fluorescence spectroscopy with BO to optimize the buffer composition and dilution factor for the refolding of a single-chain variable fragment (scFvM) expressed as IBs in *E. coli*.[33] We benchmarked this spectroscopy-assisted workflow against a traditional sequential DoE/HPLC approach, and found that BO identified conditions with higher yield and total refolded protein using only about one-third of the experiments. The contributions of this work are twofold: (i) intrinsic fluorescence spectroscopy can serve as a generalizable PAT tool for protein refolding monitoring, even under strongly varying buffer compositions, and (ii) coupling spectroscopic proxies with BO provides a practical, sample-efficient route to optimize protein refolding despite inherent measurement uncertainty. The combined approach supports rapid, informed iteration consistent with QbD/PAT principles.

2 Experimental

2.1 Materials

If not stated otherwise, all chemicals and consumables were purchased from Carl Roth GmbH (Karlsruhe, Germany). Centrifugation steps were carried out using a 3–18 KS centrifuge (Sigma Laborzentrifugen, Osterode am Harz, Germany).

2.2 Protein expression and preparation of IBs

Production of the scFvM IBs was performed using an *Escherichia coli* BL21(DE3) strain as described in detail elsewhere.[3] Cells were lysed by two cycles of high pressure homogenization at 650 bar and IBs were subsequently washed by three cycles of recursive high pressure homogenization at 650 bar (described in detail elsewhere[34]) before freezing IB aliquots at −20 °C until further use.

2.3 Solubilization and refolding of IBs

Previously frozen pellets of scFvM IBs were thawed at room temperature immediately before use. Pellets were solubilized at 100 g L^{-1} wet IB concentration (leading to 8.5 g L^{-1} of solubilized scFvM) by resuspension and mixing in a solubilization buffer (8 M urea, 50 mM glycine, pH 10) with a pipette. After the pellets were dissolved (approximately one minute of pipetting up and down), 1 M dithiothreitol (DTT) stock solution was added to the solubilized protein, to reach the final concentration of DTT for the respective experiment setting. The solubilized protein was incubated on a PMR-30 platform rocker shaker (Grant Instruments, Royston, UK) set to 20 rpm at 10 °C for 30 min, and subsequently centrifuged for 20 min at 10 °C and 20 000 rcf. After taking a sample for SDS-PAGE analysis, the clarified supernatant was used for refolding. Refolding was initiated by adding an appropriate volume of solubilized protein to the refolding buffer, followed by immediate mixing with an IKA Vortex 2 vortex mixer (IKA, Staufen, Germany). Refolds were incubated at 10 °C for at least 17 h prior to quantification of the refolded product by HIC-HPLC.

2.4 SDS-PAGE of solubilized protein

Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) was used to analyze solubilized IBs. After solubilization, 10 μL of the solubilized protein solution was diluted 1 : 10 in solubilization buffer, and 20 μL of this dilution was mixed with 20 μL reducing LDS sample buffer (494 mM Tris–HCl, 1 mM ethylenediaminetetraacetic acid, 4% lithium dodecyl sulphate, 20% glycerol, 0.44% Coomassie blue, 0.35 mM phenol red, 200 mM DTT, pH 8.5). Precast Mini-Protean TGX Stain-Free Gels 4–15% (Bio-Rad Laboratories, Feldkirchen, Germany) were used with a tris-glycine running buffer (25 mM Tris, 192 mM glycine, 0.1% sodium dodecyl sulfate, pH 8.3). After heat denaturation at 95 °C for 5 min, 4 μL of each sample or 8 μL of standard was loaded. Precision Plus Protein Unstained Standard (Bio-Rad) was used as a molecular weight ladder. Electrophoresis was conducted at 180 V for 32 min, and bands were visualized with a ChemiDoc Imager. Quantification was performed using Image Lab software with protein standards in the concentration range of 400 mg L^{-1} to 1000 mg L^{-1}.

2.5 Fluorescence spectroscopy

2.5.1 Equipment and data acquisition. Intrinsic Trp fluorescence was monitored using an FP-8550 spectrofluorometer (Jasco, Tokyo, Japan) equipped with a multi-cuvette holder with a magnetic stirrer and temperature control. Unless noted otherwise, the following settings were used: excitation wavelength 280 nm; excitation bandwidth 1 nm; emission bandwidth 10 nm; emission scan range 300 nm to 450 nm; scan speed 200 nm min^{-1}; response time 1 s; data interval 0.5 nm. The sample temperature was maintained at 10 °C. Before sample measurement, the auto-zero function was executed on the corresponding buffer (background correction). The detector gain was set to the instrument's "Low" sensitivity preset.

2.5.2 Refolding measurements. For BO experiments, measurements were performed based on the method published here.[19] Four refolding reactions were monitored in parallel in 3 mL quartz cuvettes (Hellma; 10 mm path length) with magnetic stirring at 600 rpm. Refolding buffers were temperature-equilibrated in the instrument for at least 10 min. The process was initiated by adding the solubilized protein, after which cuvettes were lidded to minimize evaporation. Full emission spectra were acquired from each cuvette approximately once per minute over the ~17 h process, generating a time series per process. While spectra were recorded continuously to track the process progress, only the initial spectrum ($t = 0$) and the endpoint spectrum ($t \geq 17$ h) were used to compute the optimization objectives.

For each spectrum, the intensity-weighted average emission wavelength (AEW) was computed over $\lambda \in [300\ \text{nm}, 450\ \text{nm}]$ as

$$\text{AEW} = \frac{\sum_{\lambda} \lambda I(\lambda)}{\sum_{\lambda} I(\lambda)}, \tag{1}$$

where $I(\lambda)$ denotes the background-corrected fluorescence intensity at wavelength λ. No spectral smoothing was applied.

2.5.3 Spectroscopic objectives used in Bayesian optimization. Two objectives were derived from the start and endpoint spectra for each condition. The spectral shift was defined as

$$\Delta\text{AEW} = \text{AEW}(t = 0) - \text{AEW}(t \geq 17\ \text{h}) \tag{2}$$

In units of nm.

A volumetric titer proxy was computed as

$$P_{\text{proxy}} = \Delta\text{AEW}\,\frac{C_{\text{sol}}}{\text{DF}} \tag{3}$$

where C_{sol} is the initial concentration of solubilized protein determined for each run from SDS-PAGE, and DF is the dilution factor used to set the final concentration for the respective refolding experiment. Bayesian optimization simultaneously maximized ΔAEW and P_{proxy}. No purified protein standard or absolute calibration assay was required for the optimization.

2.5.4 Protein denaturation monitored by AEW. To contextualize AEW as a readout of protein folding during refolding experiments, an equilibrium series was measured with a purified scFvM standard. Nineteen samples spanning from 0 M to 9 M urea in 50 mM glycine (pH 8) were prepared by a 1 : 50 dilution to a final protein concentration of 4 mg L^{-1}. Triplicate emission spectra were recorded for each urea concentration under the same instrument settings as above, using a fresh sample for each measurement to avoid artifacts from photobleaching. AEW (eqn (1)) was computed for each spectrum and plotted *versus* urea to investigate the observable dynamic range between native-like (low urea) and unfolded (high urea) conditions (Fig. 4).

2.6 HPLC quantification of refolded protein

As a reference measurement, refolded protein was quantified by hydrophobic interaction HPLC (HIC-HPLC). All samples were analyzed on an UltiMate 3000 HPLC system (Thermo

Fisher Scientific) equipped with a fluorescence detector using a MabPacTM HIC-20 HPLC column (Thermo Fisher Scientific) with a particle size of 5 μm, a pore size of 1000 Å, a length of 100 mm and a diameter of 4.6 mm. The binding buffer (buffer A) consisted of 15.75 g L^{-1} $Na_2HPO_4 \cdot 7H_2O$, 5.8 g L^{-1} $NaH_2PO_4 \cdot H_2O$ and 199 g L^{-1} $(NH_4)_2SO_4$ dissolved in distilled water with the final pH set to 7. The same buffer system but without $(NH_4)_2SO_4$ was used for elution (buffer B: 15.75 g L^{-1} $Na_2HPO_4 \cdot 7H_2O$, 5.8 g L^{-1} $NaH_2PO_4 \cdot H_2O$, final pH 7). To condition the column after storage, a minimum of 200 μL of 2 g L^{-1} bovine serum albumin (analytical standard, Sigma-Aldrich) dissolved in ultrapure water was injected onto the column in 50 μL steps, before measuring samples. Refolding samples were centrifuged for 5 min at 10 °C and 20 000 rcf before analysis. An external calibration, with at least three concentrations of scFvM standard ranging from 100 mg L^{-1} to 1000 mg L^{-1} was used to quantify the amount of protein present in a sample. The flow rate was 1 mL min^{-1} at 30 °C. For both standard and samples, an injection volume of 5 μL was used. Gradient elution was performed using the following steps:

- Binding: 25% B for 2 min.
- Linear elution: 25 to 80% B over 10 min.
- Cleaning: 100% B for 4 min.
- Re-equilibration: 25% B for 5 min.

2.7 Experimental design methods

2.7.1 Design space and parameters. Both optimization strategies, DoE and BO, were conducted in the same five-dimensional parameter space, encompassing both solubilization and refolding conditions. The parameter ranges were selected based on prior knowledge and exploratory trials to ensure that relevant folding behavior was captured. Table 1 summarizes the variables and their respective ranges.

For the DoE, chromatography-based scFvM yield and concentration were used as objectives. Furthermore, following sequential DoE iterations used a smaller design space, as described in the next section. For the BO, spectroscopy-derived proxies for the former have been used.

2.7.2 Design of experiments. The DoE approach was implemented using the MODDE software package (Sartorius, Göttingen, Germany). The initial design was based on a D-optimal criterion (22 total experiments) with the full 5-dimensional space shown in Table 1. Due to the high number of parameters and enforced constraint of physically possible final urea concentrations, only two possible D-optimal designs were suggested by the software, and the one with higher predictive power was selected.

Table 1 Design space used in optimization

Phase	Parameter	Range
Solubilization	DTT [mM]	0–25
Refolding buffer	pH [−]	8–11
	Dilution factor [−]	2–40
	GSSG [mM]	0–2.5
Final mixture	Urea [M]	0–6

The subsequent optimization DoE (DoE2) was similarly based on a D-optimal criterion, but with a reduced design space that was later extended with beyond-boundary points (32 total experiments) as depicted in Table 2. For the final optimization DoE (DoE3) only pH and final urea concentration were varied (14 total experiments), again selecting the D-optimal criterion due to the aforementioned reasons, maximizing the chance of detecting significant effects. In summary, this led to a total of 68 experiments conducted for the entire DoE campaign.

2.7.3 Bayesian optimization. Bayesian optimization was performed using a surrogate model and acquisition function implemented in the `BoTorch` and `GPyTorch` libraries. A GP model with a Matérn 5/2 kernel and automatic relevance determination was used to approximate the objective functions across the five-dimensional parameter space defined in Table 1.

The optimization was formulated as a multi-objective problem, aiming to simultaneously maximize two spectroscopic proxies: (i) the spectral shift (ΔAEW) and (ii) the volumetric titer proxy (P_{proxy}). Batch optimization was performed using the q-noisy expected hypervolume improvement (qNEHVI) acquisition function, which proposed four new experiments per iteration.[35] After each iteration, the GP model was updated with all previously collected data. Initial sampling was performed using a maximin Latin hypercube design (LHD) comprising of nine experiments. In total, four BO iterations were conducted after the initial design, resulting in 25 experiments. This experimental budget was defined *a priori* to allow for a direct comparison of efficiency against the more extensive DoE campaign, which comprised over 60 experiments in total, of which 22 were included in the first DoE1.

A physical constraint on the relationship between the final urea concentration and the dilution factor was imposed to ensure valid buffer compositions. Specifically, the concentration of urea in the refolding buffer stock ($C_{urea,ref}$) had to remain positive:

$$C_{urea,ref} = \frac{C_{urea,final} \times DF - 8}{DF - 1} > 0 \quad (4)$$

where $C_{urea,final}$ is the target urea concentration in the final mixture, DF is the dilution factor, and 8 M is the urea concentration in the solubilized protein solution.

Table 2 Reduced parameter ranges used in the optimization DoE. Parameters kept constant are denoted with a *. Parameters in {curly brackets} denote ranges extended by beyond-boundary points

Optimization DoE	Parameter	Range
DoE2	DTT [mM]	0–12.5
	pH [−]	9–10.5 {11}
	Dilution factor [−]	2–24
	GSSG [mM]	0*
	Urea [M]	0–4 {5}
DoE3	DTT [mM]	6.5*
	pH [−]	9–11
	Dilution factor [−]	11.39*
	GSSG [mM]	0*
	Urea [M]	4–6

Suggested points were post-processed using an iterative adjustment strategy. Infeasible points were corrected by alternating steps of decreasing the final urea concentration (by 0.1 M per step) and increasing the dilution factor (by 0.5 M per step), until the constraint was satisfied or a predefined maximum number of adjustments was reached. This ensured all evaluated conditions were physically valid.

3 Results and discussion

3.1 Refolding development using traditional DoE and HPLC

Similarly to a recent example by Sharma *et al.*,[13] a sequential DoE development was done as a benchmark by first performing a screening DoE including the entire design space (DoE1), followed by two subsequent optimization DoEs with reduced factor ranges (DoE2 and DoE3). To develop a productive and efficient process, the refolded scFvM concentration and the refolding yield were chosen as equally weighted optimization targets. Process analytics were based on SDS-PAGE (for solubilized scFvM) and HPLC measurements (for refolded scFvM) using a protein standard produced and purified in-house (see the SI).

The R^2, Q^2, and p-values for lack of fit of all three DoE optimization runs and the final validation results are summarized in Table 3. While the predicted optimum of DoE1 was 0.78 g L^{-1} refolded scFvM with a refolding yield of 74.1%, a validation in triplicate did not match the prediction (0.36 $\pm$ 0.03 g L^{-1} of refolded scFvM with 42.6 $\pm$ 2.0% refolding yield). This showed the lack of predictive power of the initial screening DoE (also reflected in a low $Q^2 = 0.6$ for the refolded scFvM). After two rounds of reducing the parameter ranges of the more influential factors while keeping the less influential ones at a constant value, a less favorable final optimum of 0.30 g L^{-1} and a refolding yield of 57.4% was predicted. However, this prediction was confirmed by a validation in triplicate (Table 3). The increased predictive power could also be inferred from a progressive reduction of the difference between R^2 and Q^2, especially for the refolded scFvM concentration. These results highlight a problem with the sequential screening and optimization DoE approach for complex processing steps that include many critical parameters. The parameter space must be reduced to obtain a model with good predictive power. However, the decisions of which factors to fix and which ranges to narrow are made based on suboptimal models, thereby potentially removing productive regions from the design space. In the case of refolding, overlooking lower-impact components becomes likely when highly influential parameters, like the dilution factor, are part of the optimization.

To illustrate this issue, we outline the narrowing of the parameter ranges in detail, starting with the redox system. The predicted parameter effects on both responses are shown in Fig. 2. Of the redox system components DTT and GSSG, only DTT was found to be a significant model term in DoE1, and only for the refolded scFvM concentration. With lower DTT concentrations leading to higher scFvM concentrations, the range was reduced to the lower half of the initial design space (0 mM to 12.5 mM) for DoE2. In DoE2, the DTT concentration was found to be an insignificant factor for both responses. However, since the scFvM contains two disulfide bonds, a low DTT concentration of 6.5 mM in the middle of that range was chosen for the final process conditions. A decision like this relies on operator expertise and cannot be fully automated. In the context of miniaturized process development this is another downside of the sequential DoE approach.

In DoE1, GSSG concentration showed no significant effect on either response. Given that costs for GSSG exceed the second most expensive buffer component by a factor of 10 (see SI Table S1), we judged that in an industrial setting it would only be included in a buffer composition if it had a significantly positive effect. Eliminating GSSG also reduces the parameter space, facilitating optimization of the remaining factors in subsequent DoEs. We therefore fixed GSSG at 0 mM for all further DoEs. For both redox agents, these decisions were based on early models with a significant lack of fit and required the operators to make assumptions. Although the dilution factor had the highest variable importance score for refolded scFvM (Fig. 3), it was not a significant model factor for the refolding yield in DoE1. Unsurprisingly, increasing the dilution factor decreased the refolded scFvM concentration. However, as a reduction in yield can be expected past a certain point of reducing the dilution factor, but we still had little information to estimate the exact location of that point, we decided to reduce the investigated design space of DoE2 to the lower half of the initial range (DF 2–24). With the reduced factor ranges in DoE2, the opposing

Table 3 Summary of the main statistics of the three DoE rounds and comparison of the predicted responses at the calculated optimal conditions for equal weighing of concentration and yield. Experimental validation was performed in triplicate for the optimum conditions of the first and third DoE. The p-value for the lack of fit is calculated in MODDE based on the variance explained by the pure error (replicates) and the variance not explained by the model. A p-value below the threshold of 0.05 (indicated with a red value) marks a statistically significant likelihood of a lack of fit (LOF)

Model	Response	R^2	Q^2	p-value* (LOF)	Predicted optimum	Validation experiment
DoE1	Refolded scFvM	0.86	0.60	0.029	0.78 g L^{-1}	0.36 $\pm$ 0.03 g L^{-1}
	Refolding yield	0.94	0.87	0.406	74.1%	42.6 $\pm$ 2.0%
DoE2	Refolded scFvM	0.70	0.57	0.000	0.34 g L^{-1}	n.a
	Refolding yield	0.63	0.49	0.000	27.1%	n.a
DoE3	Refolded scFvM	0.89	0.84	0.106	0.3 g L^{-1}	0.37 $\pm$ 0.02 g L^{-1}
	Refolding yield	0.89	0.84	0.106	57.4%	61.4 $\pm$ 3.1%

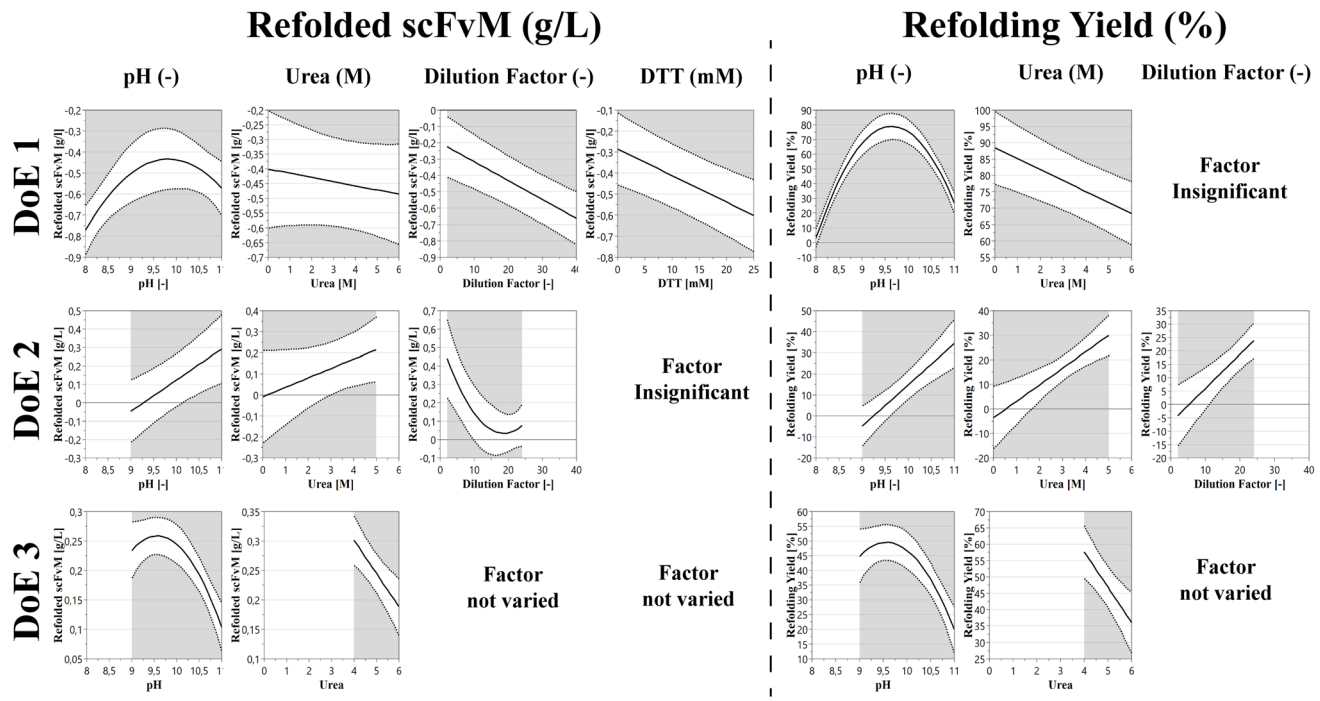

Fig. 2 Effects of the investigated factors on both optimized responses (refolded scFvM concentration and refolding yield) across three DoE rounds. GSSG was omitted as it had no significant coefficient in any model. For DoE1, the refolded scFvM concentration was transformed using a negative logarithmic transformation to improve model fit. Solid lines represent model predictions, while dotted lines indicate the 95% confidence interval. Factors labeled as "insignificant" or "not varied" were excluded from the respective models.

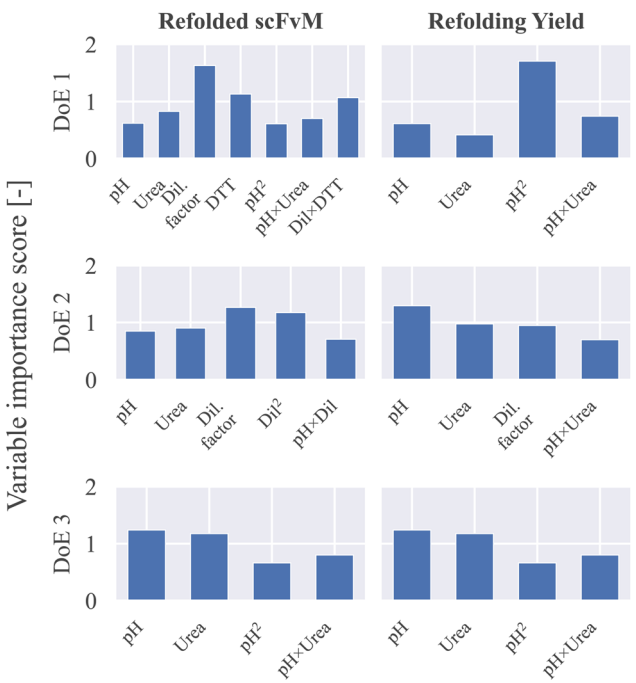

Fig. 3 Variable importance scores as calculated in MODDE 12, ordering the factors from largest to smallest impact on the respective response. The left column of plots shows the variable importance for the refolded scFvM concentration, the right column of plots shows the variable importance for the refolding yield. Each column of plots is ordered according to the sequential DoE rounds (top to bottom: DoE1, DoE2, DoE3).

effects of the dilution factor on the refolding yield and refolded scFvM concentration were depicted in the model as the theory would suggest. However, the effects of the urea concentration and pH contradicted the previous results of DoE1, with optimum conditions at the upper edge of the design space. To address this, beyond-boundary points with pH values of up to pH 11 and final urea concentrations of up to 5 M were added to the design. These new upper boundaries were picked based on the information that a combination of pH 11 and 6 M urea seemed to fully prevent refolding in DoE1 experiment no. 5 (2.8% refolding yield). Therefore we concluded that an optimum might lie between the upper corner of DoE2 (pH 10.5 and 4 M urea and this extreme value. However, even with the inclusion of these extra experiments, the resulting model did not clearly identify an optimum of pH and urea. To clarify the impact of these two factors, the dilution factor was kept constant for DoE3 at the predicted optimum of DoE2's model (DF 11.39, weighing refolded scFvM concentration and refolding yield equally).

Due to the constant dilution factor and constant DTT concentration enabling the use of pooled solubilized protein, the optimization resulted in the models for the two responses being linear transformations of each other, explaining the identical statistical descriptors for both responses. Maintaining three of the five parameters constant left a two-dimensional, narrowed down, design space for the final DoE, optimizing the pH and urea concentration with a statistically sound model ($R^2 = 0.89$ *vs.* $Q^2 = 0.84$, unlikely lack of fit). However, by using the sequential DoE approach, only two out of five parameters

could be determined by high quality models. Having optimized the pH and urea concentration, the other parameters could be investigated in further sequential DoEs, while maintaining these two parameters constant. However, this would further increase the already very high number of experiments.

3.2 Development and validation of a generalizable spectroscopic proxy for refolding

To enable screening experiments without the prerequisite of purified standard material, we aimed to establish a rapid, in-line readout that can guide BO across widely varying buffer compositions. However, relying on the fluorescence intensity and wavelength shift as done in the previously mentioned approach[19] proved to be inconsistent in the present, broader design space: changes in pH, urea, and redox agents may have altered the fluorophore microenvironments in a way that a general intensity-based response is not directly representative of the native protein concentration. We therefore turned solely to the AEW, the intensity-weighted spectral center (eqn (1)). Because AEW is a ratio of weighted sums, any uniform spectrum scaling cancels in the numerator and denominator. AEW is thus primarily sensitive to changes in spectral position and shape (*i.e.*, the relative distribution of emission across wavelengths) that reflect the Trp microenvironment.

To first confirm that AEW reports on the conformational state of scFvM, we measured an equilibrium protein standard denaturation series in urea. As shown in Fig. 4, AEW undergoes a sigmoidal transition from a native-like value at low urea to a red-shifted value at high urea, demonstrating increased solvent exposure of Trp residues. This provides direct physical evidence that AEW tracks conformational change. Fluorescence-monitored equilibrium denaturation curves, using AEW/center-of-mass or related wavelength-shift metrics, are widely used to quantify unfolding transitions.[36–38]

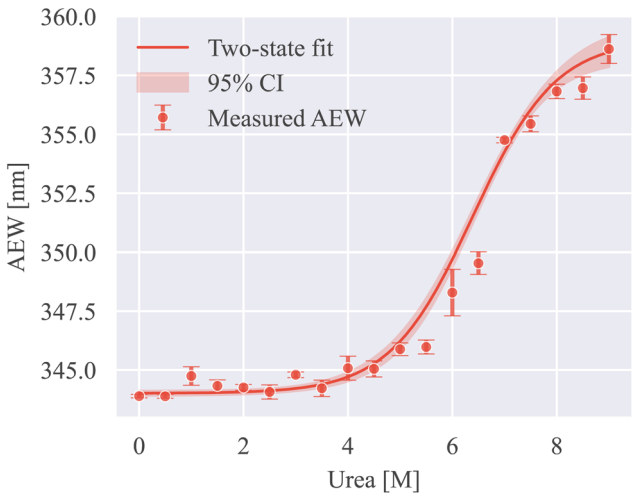

Fig. 4 Urea-induced equilibrium unfolding of scFvM standard monitored by intrinsic Trp and Tyr fluorescence. AEW exhibits a sigmoidal transition from folded (0 M urea) to unfolded (9 M urea), indicating sensitivity to conformational state. The solid line is a two-state fit.

Although Fig. 4 characterizes unfolding, the same mechanism underlies refolding: as hydrophobic Trp residues tend to be buried in the protein core during folding, the emission band blue-shifts (lower AEW). In the successful refolding process, AEW decreases over time.[17] As defined in eqn (2), we formalized that the refolding-induced change is positive ($\Delta AEW > 0$) for a net blue shift, indicating productive protein folding.

AEW reports the spectral position of Trp fluorescence emission, which depends on local polarity and is modulated by buffer-dependent quenching and energy-transfer processes across all emitting residues.[38] Consequently, absolute AEW values are condition-dependent: different buffer conditions can yield the same refolding outcome, but produce different absolute AEW values (and *vice versa*). Nonetheless, we hypothesized that ΔAEW could serve as a condition-dependent surrogate for the refolding yield—useful, but potentially with limitations in accuracy. We therefore pre-validated the proxy on a different set of conditions spanning the same design space. For each run, we measured endpoint ΔAEW and HIC-HPLC yield. The two quantities showed a clear linear relationship ($R^2 = 0.71$; Fig. 5).

Accordingly, we chose two standard-free objectives (requiring no purified product standard or calibration curve) to be optimized in the BO: (i) the spectral shift ΔAEW defined in eqn (2) as a proxy for yield, and (ii) the volumetric titer proxy P_{proxy} defined in eqn (3), where C_{sol} is the initial concentration of solubilized protein determined by SDS-PAGE for each run, and DF is the dilution factor of the respective experiment.

3.3 Spectroscopy-guided optimization using Bayesian optimization

To optimize identical parameters within the same design space boundaries as in the DoE, BO was run in four iterations of four suggestions each, initiated by a 9-point Latin hypercube design (LHD) (total $n = 25$). The algorithm optimized two spectroscopy-derived objectives: the spectral shift ΔAEW and a volumetric titer proxy P_{proxy}.

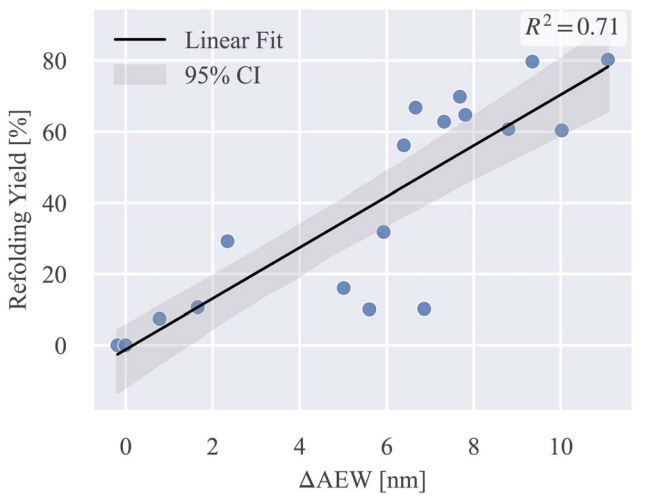

Fig. 5 Independent validation of ΔAEW as a proxy for refolding yield. Linear fit (black) with 95% confidence band (gray); data were not used in BO training.

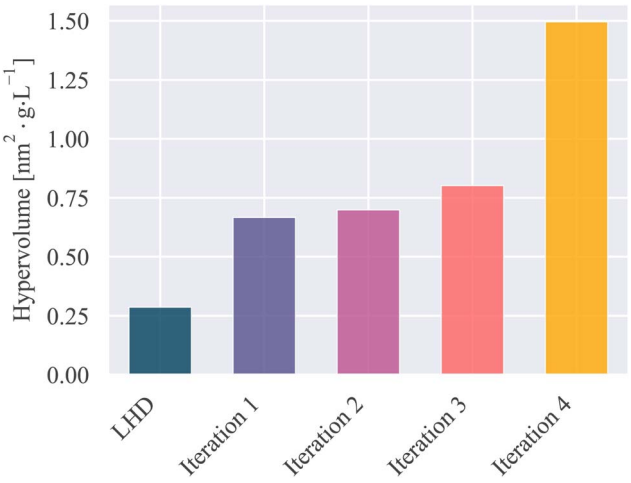

Fig. 6 Dominated hypervolume in the objective space across the initial Latin hypercube design (LDH) and subsequent BO iterations. Hypervolume increases monotonically, the largest gain occurs in iteration 4.

Fig. 6 shows the progression of the dominated hypervolume across iterations. The hypervolume is the size of the objective space dominated by the current Pareto set relative to a fixed reference point. Across iterations 1–4 the hypervolume grew steadily, with a pronounced jump from iteration 3 to 4, indicating successful navigation to the regions with jointly improved proxy responses under the qNEHVI acquisition.

Fig. 7A visualizes the evolution of the proxy frontier by iteration. Points move away from the baseline over time, and by iteration 4 the frontier is effectively defined by two candidates: one LHD candidate with high ΔAEW, and one iteration-4 point with a remarkably high P_{proxy} and the second-highest ΔAEW. In comparison, Fig. 7B displays the corresponding HPLC objectives (titer *vs.* yield), which were measured only for validation and not used for the optimization. Compared with the proxy space, the best-by-proxy condition appears overestimated. Nevertheless, it remains Pareto optimal under HPLC. By design, HPLC is species-resolved and calibrated, yielding tighter, more specific estimates of titer and yield. In contrast, P_{proxy} is multiplicative in measured quantities (ΔAEW, C_{sol}, and DF), and therefore measurement noise can propagate and especially inflate the most extreme point estimates. Despite this imperfect correspondence, the proxies provided a sufficient signal to steer BO into the correct region of the design space, and the HPLC frontier confirms the outward shift across iterations. Importantly, the spectroscopy-derived objectives track the chromatographic evaluation for complementary targets (titer *vs.* yield). P_{proxy} aligned strongly with the HPLC-determined refolding titer (Spearman $\rho \approx 0.91$; Pearson $R^2 \approx 0.68$), while ΔAEW aligned with HPLC determined refolding yield (Spearman $\rho \approx 0.88$; Pearson $R^2 \approx 0.72$). These associations are not driven solely by factor settings: controlling for process factors (pH, GSSG, DTT, final urea, dilution factor), the correlation between ΔAEW and HPLC yield remains very strong (partial $r = 0.842$, $p = 1.3 \times 10^{-7}$), and the correlation between P_{proxy} and HPLC titer remains substantial (partial $r = 0.696$, $p = 1.1 \times 10^{-4}$).

Fig. 8 shows that the search is progressively concentrated around a coherent operating window. Early iterations still sampled broadly, but the suggested conditions moved toward alkaline pH, with iterations 3–4 concentrating mostly near the upper bound (≈10.5–11). In parallel, the redox environment shifted toward strongly oxidizing conditions: GSSG moved from a wide initial sweep to values near the high end of the range (≈2 mM to 2.5 mM). DTT showed no benefit at high levels and

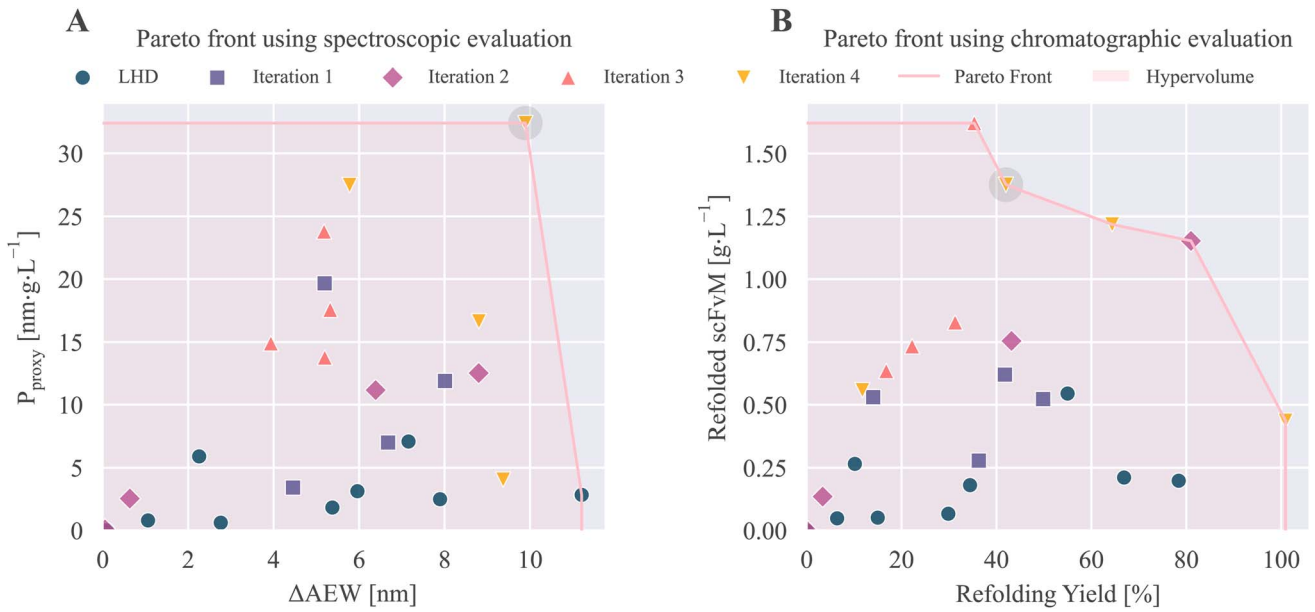

Fig. 7 Comparison of spectroscopy- and chromatography-based Pareto fronts during Bayesian optimization. (A) Proxy space (ΔAEW *vs.* P_{proxy}) and (B) chromatographic space (HPLC yield *vs.* refolded scFvM titer) colored by BO iteration. In both panels, the pink shaded region indicates the hypervolume dominated by the current Pareto optimal experiments. Furthermore, over the optimization, points progressively move away from the origin. The final operating point is highlighted by a gray shaded circle.

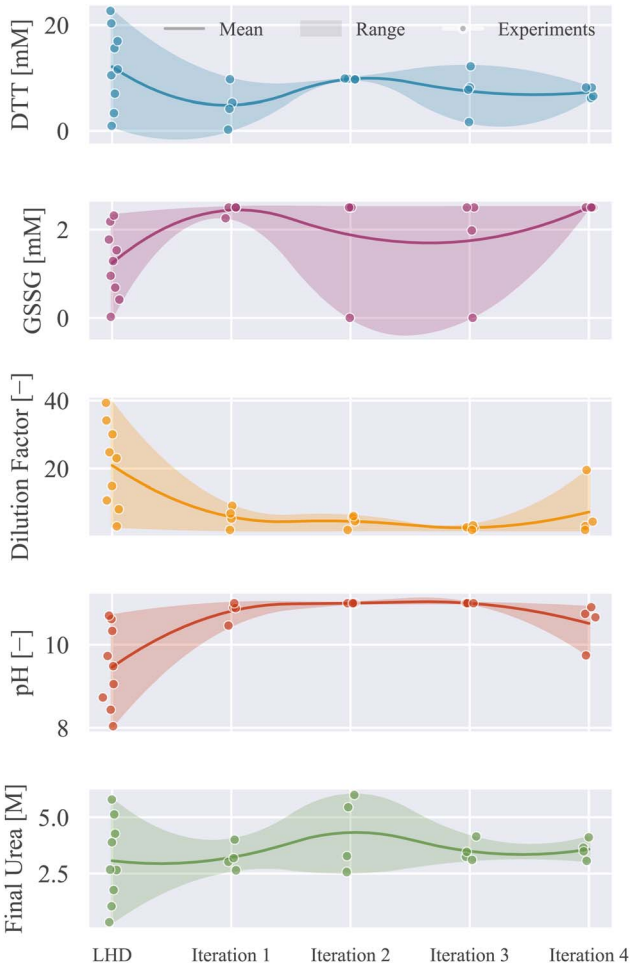

Fig. 8 Convergence of BO proposals for each factor by iteration. Dots indicate individual experiments, the line is the iteration mean, and the shaded band spans the sampled range (min–max). Across iterations, proposals concentrate toward alkaline pH (≈10.5–11), high GSSG (≈2 mM to 2.5 mM), moderate dilution (between DF ≈ 3–6), and mid-range urea (≈3 M to 4 M). DTT trends downward toward low-to-moderate levels, consistent with the oxidizing regime associated with the high-P_{proxy} candidates in iteration 4.

drifted downward over iterations, with late iterations favoring low-to-moderate values, aligning with the oxidizing GSSG trend. The dilution factor range focused from a broad LHD spread to moderate values (between 3 and 6), which raises the effective protein concentration and thus the volumetric proxy, while still exploring variations at high protein concentrations. Urea likewise converged from the full range to a mid-range plateau (≈3 M to 4 M), balancing sufficient chaotrope to avoid aggregation with enough dilution of the denaturant to allow refolding. Taken together, BO steered the campaign toward an alkaline, oxidizing, moderately concentrated, mid-urea regime; the design space region that produced the high-P_{proxy} candidates in iteration 4 and from which the final operating point was selected is indicated in Fig. 7 (grey circle). The final operating point was afterwards validated in triplicate and compared with the DoE results.

3.4 Comparison of the optima

The final validation conditions for the BO based workflow were selected from the proxy Pareto and were set to be a balanced trade-off between ΔAEW and P_{proxy}. The selection was based only on the spectroscopic BO objectives, with HPLC data obtained afterwards for validation. For the DoE-based workflow, the final conditions were based on the scalarized optimization of the second-order response surface model.

For both workflows, the optimum ended up with a refolding yield of around 60% (see Table 4). However, the spectroscopically assisted BO achieved this yield at a much lower dilution factor, leading to 1.29 ± 0.06 g L^{-1} compared to only 0.37 ± 0.02 g L^{-1} refolded scFvM for the DoE-based optimization. The corresponding independent triplicate statistics are provided in the SI: the BO condition increased the titer by 0.927 g L^{-1} relative to the final DoE conditions (95% CI 0.808 g L^{-1} to 1.046 g L^{-1}), whereas the yield difference was not significant within the triplicate validation uncertainty. Meanwhile, only 25 experiments were conducted for the BO-based optimization, compared to a total of 68 experiments across all DoEs. Still, it should be noted that individual experiments with high protein concentrations were also observed in the DoEs, but they showed generally very low refolding yields (for example 1.48 g L^{-1} with 33.6%).

The obvious main difference between the two found optima is the dilution factor. Due to the operational difficulty in precisely controlling the protein concentration after solubilization, the easily controllable dilution factor is commonly

Table 4 Summary of the final process parameters and results for the optimization targets from validation experiments. Validation experiments were performed as independent process triplicates, each starting from a separate IB aliquot and including independent solubilization, refolding, sample preparation, and HIC-HPLC analysis. Raw replicate values, 95% confidence intervals, and statistical comparisons are provided in the SI. For the validation experiment, the quantification of refolded scFvM by HIC-HPLC was additionally confirmed with a secondary HPLC method (cation exchange chromatography). The corresponding data can be found in the data repository available at https://doi.org/10.5281/zenodo.18270843

Parameter/target	DoE + chromatography	BO + Trp fluorescence
pH (–)	9.9	10.7
Urea (M)	4.03	3.64
Dilution factor (–)	11.39	3.14
GSSG (mM)	0.0	2.5
DTT (mM)	6.5	6.13
Refolded scFvM (g L^{-1})	0.37 ± 0.02	1.29 ± 0.06
Refolding yield (%)	61.4 ± 3.1	58.7 ± 1.3

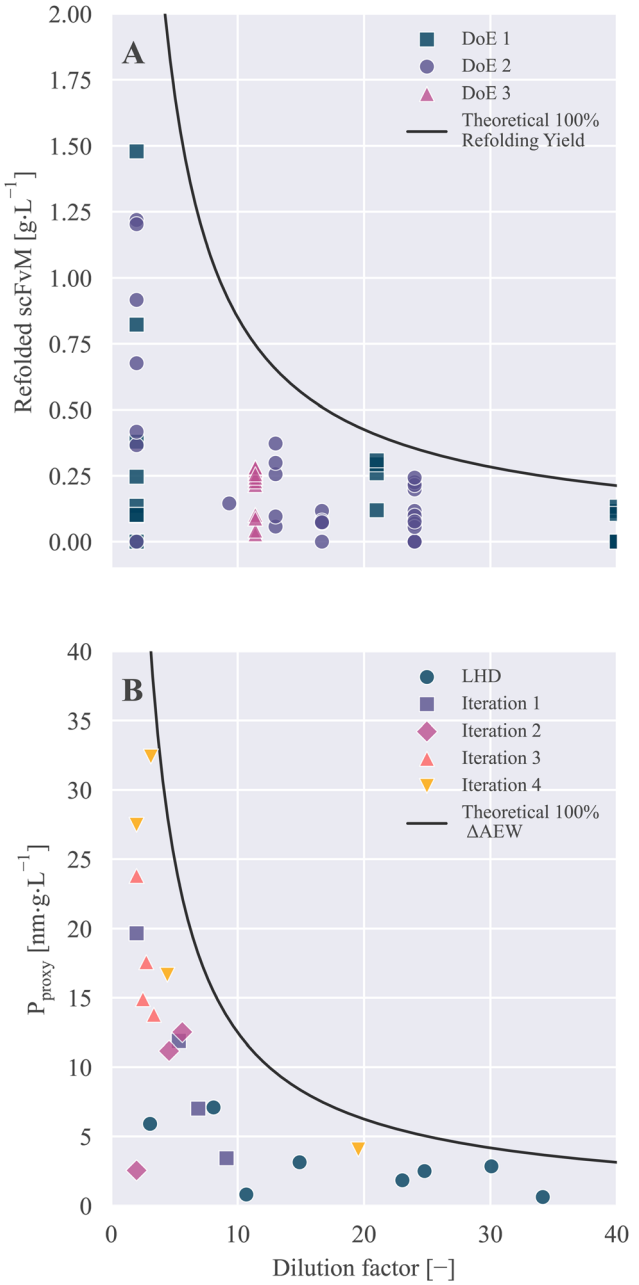

Fig. 9 (A) Overview of refolded scFvM concentrations for all individual experiments of the DoE optimization. Colored symbols represent the different DoE series. The plotted function shows the corresponding theoretical scFvM concentration at 100% refolding yield, assuming 8.5 g L^{-1} of solubilized product. (B) Equivalent plot for the volumetric titer proxy P_{Proxy} optimized in the BO experiments. Colored symbols represent the subsequent optimization iterations. The black line marks the theoretical P_{Proxy} at 100% calculated based on the correlation of ΔAEW and refolding yield shown in Fig. 5 assuming a maximum ΔAEW shift of 14.7 nm (corresponding to 100% refolding yield in the correlation).

varied instead.[13,39–41] However, this crucial factor can be difficult to optimize by DoE due to the issue of conflicting optimization goals. Targeting only the product concentration or biological activity can lead to very inefficient processes like the single experiment with 33.6% yield mentioned above. However, optimizing only the refolding yield means that a higher dilution factor will almost always be beneficial. This would lead to either the highest dilution factor,[13] or some arbitrary best trade-off being chosen by the operator.[39,41] Hence, we chose to target both responses, weighing them equally in this test case. This was handled naturally by the BO approach, which used a hypervolume-improvement acquisition to expand the non-dominated region over titer and yield. By contrast, in a multi-phase DoE the next design region must be chosen *via* a scalarized objective (*e.g.*, desirability weights) and manual redefinition of the search space—choices that are complex under the opposing DF effects.

Furthermore, the dilution factor affects the total protein concentration in a geometric $1/x$ function. This means that very small variations in dilution factor affect the protein concentration greatly at low dilution factors, visualized in Fig. 9A as the function line representing a theoretical 100% refolding yield. For experiments near dilution factor 2, this leads to a far greater variance of the protein concentration compared to higher dilution factors. This effectively leads to heteroscedasticity of the data, violating the base assumptions of statistically sound DoE design. Furthermore, the dilution factor variation serves as an example for the situation when the factor response relationship is not accurately captured by lower-order polynomials. Trying to investigate a wide design space in the initial screening DoE1 led to a large gap in experiments within the dilution factor range of 2–24. However, there is no *a priori* reasoning that can be used to perfectly set this range for the next DoE, causing many wasted experiments at unsuitable settings. Even after three rounds of sequential DoEs, the space between dilution factors 2–10 was left poorly explored. In contrast, the GP model adapted well to the protein folding at such unusually high protein concentrations. As shown in Fig. 9B, the majority of experimental effort (17 out of 25 experiments) was spent in the range of dilution factors 2–10. Meanwhile, the full parameter range was still available for exploration. This allowed for gathering information on the impact of GSSG at a lesser total experimental variance in later optimization rounds, while it was deemed an insignificant factor in the DoE-based optimization. As previously described by Lapierre *et al.*,[31] apparent early insignificance can lead to premature factor removal for sequential DoEs. In this experimental case, the high cost of GSSG combined with its apparent insignificance led to a suboptimal operator decision of entirely removing it from the subsequent DoEs. This was confirmed by running the BO optimized process without GSSG and monitoring the online Trp fluorescence shift compared to the Pareto-optimal candidate (SI Fig. S1). As the residual DTT from the solubilizate has to oxidize before disulfide bonds can form, the reaction kinetics seemed to be highly delayed. This was likely the cause of the experiment without GSSG resulting in only half the refolding yield. As the dilution factor determines how much DTT is carried over to the refold, we hypothesize that this effect is only significant at low dilution factors.

The advantage of the BO workflow lies in the sensibility of the experimental allocation. When objectives or factors pull in opposite directions (*e.g.*, dilution factor increases yield but

depresses titer *via* 1/DF), BO advances the Pareto frontier directly rather than committing to a fixed scalarization. When several factors interact and their importance is uncertain, BO updates factor weights and effective ranges through the surrogate, avoiding mis-specified polynomial structure and range choices that can lock a DoE into uninformative regions. In this case, BO concentrated evaluations where marginal information and improvement were highest (DF 2–10) while keeping the full space open, revisiting the GSSG effect without manual redesign. This is why the spectroscopy-assisted BO workflow delivered the same ~60% refolding yield at markedly higher titers with fewer runs: the method selects more meaningful experiments under both conflicting-objective and many-factor conditions.

Finally, the scope of the spectroscopy-derived proxies should be considered. The workflow relies on a folding-sensitive intrinsic fluorescence response of the target protein, primarily from Trp and Tyr residues. It is therefore best suited to proteins with a sufficiently strong and target-specific fluorescence signal. Proteins lacking those are not directly suitable for this method. In addition, the fluorescence signal should be dominated by the target protein. Substantial amounts of other fluorescent (host-cell) proteins or other co-refolding impurities could contribute to the AEW shift and weaken the relationship between the proxy and target refolding. More generally, AEW is a conformational, but not species-resolved, readout. It cannot distinguish a correctly folded monomer from misfolded, aggregated, fragmented, or impurity-derived species with the specificity of chromatographic analysis. Moreover, since the sequential DoE approach relies on operator decisions about how to constrain the subsequent design spaces, we cannot exclude the possibility that different choices could have led to a different optimum with comparable performance to the one from spectroscopy-assisted BO. However, we want to emphasize that not having to make these choices to constrain the design space is one of the main advantages of BO over DoE, especially for parameter-rich optimizations.

4 Conclusion

This work establishes a spectroscopy-assisted Bayesian optimization workflow as a practical alternative to sequential DoE/HPLC for protein refolding process development. By integrating intrinsic Trp fluorescence with multi-objective BO over a five-factor space, we located a high-performing operating window in 25 experiments – roughly one third of the DoE effort – while delivering ~3.5-fold higher product concentration at comparable yield. Besides the higher product concentration, the almost 4-fold lower dilution factor also reduces the water and urea consumption in a direct process comparison. Validation by HPLC confirmed that the spectroscopy-derived objectives provided sufficient directional signal to steer the search toward the true HPLC Pareto region. Thereby, process optimization was possible without the requirements to expend purified protein standards, or a fully validated chromatographic method. Importantly, keeping the entire experimental space open for exploration, BO revealed a positive influence of GSSG, which the sequential DoE path had excluded.

There are clear opportunities to further broaden the scope and robustness. If a standard is available, multi-fidelity schemes that occasionally confirm by HPLC while primarily optimizing on spectroscopy would be a logical and attractive extension. Another avenue is explicit handling of costs and constraints for reagents, run time, and material consumption. We also see promise in transfer and meta-learning across proteins. Finally, this method also holds potential for integration into high-throughput systems, enabling closed-loop experimentation with BO and fluorescence-based online analytics. In the shown test-case, this method delivered higher titers at similar yields with far fewer experiments compared to a traditional sequential DoE approach.

Acknowledgements

Author contributions

Florian Gisperg: writing – original draft, writing – review and editing, conceptualization, methodology, investigation, formal analysis, visualization, data curation, validation; Robert Klausser: writing – original draft, writing – review and editing, conceptualization, methodology, investigation, formal analysis, visualization, data curation, validation; Matthias Kierein: investigation, validation; Eva Prada Brichtova: writing – review and editing, conceptualization, methodology, investigation; Mohamed Elshazly: writing – review and editing, resources; Julian Kopp: writing – review and editing, conceptualization, methodology; Oliver Spadiut: writing – review and editing, conceptualization, methodology, supervision, project administration, funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Data availability

All generated data and code used in this work are openly available at https://doi.org/10.5281/zenodo.18270843.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d6dd00035e.

Acknowledgements

The authors acknowledge the TU Wien Bibliothek for financial support through its Open Access Funding Program. The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

Notes and references

1 D. Humer and O. Spadiut, *World J. Microbiol. Biotechnol.*, 2018, **34**, 158.

2 E. García-Fruitós, E. Vázquez, C. Díez-Gil, J. L. Corchero, J. Seras-Franzoso, I. Ratera, J. Veciana and A. Villaverde, *Trends Biotechnol.*, 2012, **30**, 65–70.
3 M. Elshazly, B. Leeb, E. P. Brichtova, F. Gisperg, R. Klausser, S. Vijayakumar, B. Lendl, M. Voigtmann, M. Berkemeyer, O. Spadiut and J. Kopp, *J. Biotechnol.*, 2025, **405**, 182–190.
4 R. Klausser, J. Kopp, E. Prada Brichtova, F. Gisperg, M. Elshazly and O. Spadiut, *Front. Bioeng. Biotechnol.*, 2023, **11**, 1249196.
5 L. Buscajoni, M. C. Martinetz, M. Berkemeyer and C. Brocard, *Biotechnol. Adv.*, 2022, **61**, 108050.
6 C. Slouka, J. Kopp, O. Spadiut and C. Herwig, *Appl. Microbiol. Biotechnol.*, 2019, **103**, 1143–1153.
7 L. Li, A. Kantor and N. Warne, *Protein Sci.*, 2013, **22**, 1118–1123.
8 International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), Q6B: Specifications: Test Procedures and Acceptance Criteria for Biotechnological/Biological Products, ICH Harmonised Tripartite Guideline, 1999.
9 S. Prior, S. E. Hufton, B. Fox, T. Dougall, P. Rigsby and A. Bristow, *mAbs*, 2018, **10**, 129–142.
10 P. D. Bade, S. P. Kotu and A. S. Rathore, *J. Sep. Sci.*, 2012, **35**, 3160–3169.
11 D. M. Boyle, J. J. Buckley, G. V. Johnson, A. Rathore and M. E. Gustafson, *Biotechnol. Appl. Biochem.*, 2009, **54**, 85–92.
12 V. Dechavanne, N. Barrillat, F. Borlat, A. Hermant, L. Magnenat, M. Paquet, B. Antonsson and L. Chevalet, *Protein Expression Purif.*, 2011, **75**, 192–203.
13 R. Sharma, A. Anupa, N. Kateja and A. S. Rathore, *Biochem. Eng. J.*, 2022, **187**, 108601.
14 S. Ughade, S. Rana, M. Nadeem, R. Kumthekar, S. Mahajani and R. Bhambure, *ACS Omega*, 2024, **9**, 3204–3216.
15 F. Gisperg, R. Klausser, M. Elshazly, J. Kopp, E. P. Brichtová and O. Spadiut, *Biotechnol. Bioeng.*, 2025, **122**, 1313–1325.
16 J. N. Pauk, C. L. Igwe, C. Herwig and J. Kager, *Chem. Eng. Sci.*, 2024, **287**, 119774.
17 C. L. Igwe, D. F. Müller, F. Gisperg, J. N. Pauk, M. Kierein, M. Elshazly, R. Klausser, J. Kopp, O. Spadiut and E. Přáda Brichtová, *Anal. Bioanal. Chem.*, 2024, **416**, 3019–3032.
18 R. Sharma, N. G. Jesubalan and A. S. Rathore, *Biochem. Eng. J.*, 2024, **204**, 109252.
19 C. L. Igwe, F. Gisperg, M. Kierein, E. P. Brichtová, O. Spadiut and D. F. Müller, *Comput. Chem. Eng.*, 2024, **187**, 108734.
20 A. Martens, M. Neufang, A. Butté, M. v. Stosch, A. d. R. Chanona and L. M. Helleckes, Holistic Bioprocess Development Across Scales Using Multi-Fidelity Batch Bayesian Optimization, *arXiv*, 2025, preprint, arXiv:2508.10970 [q-bio], DOI: 10.48550/arXiv.2508.10970.
21 H. Narayanan, J. A. Hinckley, R. Barry, B. Dang, L. A. Wolffe, A. Atari, Y.-Y. Tseng and J. C. Love, *Nat. Commun.*, 2025, **16**, 6055.
22 S. S. Rosa, D. Nunes, J. G. visualisation, D. M. F. Prazeres, A. M. Azevedo, D. G. Bracewell and M. P. C. Marques, *Sep. Purif. Technol.*, 2025, **367**, 132881.
23 R. Siedentop, M. Siska, J. Hermes, S. Lütz, E. von Lieres and K. Rosenthal, *ChemCatChem*, 2025, **17**, e202400777.
24 E. Claes, T. Heck, K. Coddens, M. Sonnaert, J. Schrooten and J. Verwaeren, *Biotechnol. Bioeng.*, 2024, **121**(5), 1569–1582.
25 M. Fitzner, A. Šošić, A. V. Hopp, M. Müller, R. Rihana, K. Hrovatin, F. Liebig, M. Winkel, W. Halter and J. G. Brandenburg, *Digital Discovery*, 2025, **4**, 1991–2000.
26 S. Greenhill, S. Rana, S. Gupta, P. Vellanki and S. Venkatesh, *IEEE Access*, 2020, **8**, 13937–13948.
27 M. Siska, E. Pajak, K. Rosenthal, A. del Rio Chanona, E. von Lieres and L. M. Helleckes, *Biotechnol. Bioeng.*, 2026, **123**, 805–830.
28 K. Tsumoto, D. Ejima, I. Kumagai and T. Arakawa, *Protein Expression Purif.*, 2003, **28**, 1–8.
29 S. Nikita, R. Sharma, J. Fahmi and A. S. Rathore, *React. Chem. Eng.*, 2023, **8**, 592–603.
30 H. Narayanan, F. Dingfelder, I. Condado Morales, B. Patel, K. E. Heding, J. R. Bjelke, T. Egebjerg, A. Butté, M. Sokolov, N. Lorenzen and P. Arosio, *Mol. Pharm.*, 2021, **18**, 3843–3853.
31 F. M. Lapierre, P. Mattaliano, D. Raith, M. Castillo-Cota, J. Bermeitinger and R. Huber, *J. Chem. Technol. Biotechnol.*, 2025, **100**, 1571–1583.
32 S. S. Rosa, D. Nunes, L. Antunes, D. M. F. Prazeres, M. P. C. Marques and A. M. Azevedo, *Biotechnol. Bioeng.*, 2022, **119**, 3127–3139.
33 S. Edwardraja, R. Neelamegam, V. Ramadoss, S. Venkatesan and S.-G. Lee, *Biotechnol. Bioeng.*, 2010, **106**, 367–375.
34 R. Klausser, L. Veiter, J. Kopp, N. Hammerschmidt, T. Frierss, F. Gisperg, M. Elshazly, E. P. Brichtova, M. Martinetz, M. Voigtmann and O. Spadiut, *J. Biotechnol.*, 2025, **409**, 183–194.
35 M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson and E. Bakshy, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020, pp. 21524–21538.
36 K. L. Maxwell, D. Wildes, A. Zarrine-Afsar, M. A. De Los Rios, A. G. Brown, C. T. Friel, L. Hedberg, J.-C. Horng, D. Bona, E. J. Miller, A. Vallée-Bélisle, E. R. Main, F. Bemporad, L. Qiu, K. Teilum, N.-D. Vu, A. M. Edwards, I. Ruczinski, F. M. Poulsen, B. B. Kragelund, S. W. Michnick, F. Chiti, Y. Bai, S. J. Hagen, L. Serrano, M. Oliveberg, D. P. Raleigh, P. Wittung-Stafshede, S. E. Radford, S. E. Jackson, T. R. Sosnick, S. Marqusee, A. R. Davidson and K. W. Plaxco, *Protein Sci.*, 2005, **14**, 602–616.
37 S. V. S. R. K. Pulavarti, J. B. Maguire, S. Yuen, J. S. Harrison, J. Griffin, L. Premkumar, E. A. Esposito, G. I. Makhatadze, A. E. Garcia, T. M. Weiss, E. H. Snell, B. Kuhlman and T. Szyperski, *J. Phys. Chem. B*, 2022, **126**, 1212–1231.
38 *Protein Supersecondary Structures: Methods and Protocols*, ed. A. E. Kister, Springer New York, New York, NY, 2019, vol. 1958.
39 F. Wang, Y. Fang, J. Yu, X. Zhao, Y. Liu, X. Jing, J. Wang, S. Wang, S. Wang, J. Jiang and S. Zhang, *Protein Expression Purif.*, 2025, **236**, 106806.
40 K. N. Mihooliya, N. Nitika, R. Bhambure and A. S. Rathore, *Biotechnol. J.*, 2023, **18**, 2200505.
41 X. Bao, L. Xu, X. Lu and L. Jia, *Protein Expression Purif.*, 2016, **117**, 59–66.

