SUPPLEMENTARY INFORMATION

S1: Methodology for refolding experimentation

Chemicals including urea, D-sorbitol, L-glutathione oxidized (GSSG), glacial acetic acid, dithiothreitol (DTT), hydrochloric acid (HCL), trifluoracetic acid (TFA) and acetonitrile (ACN) were purchased from Sisco Research Laboratories Pvt. Ltd. (SRL) and used in buffer preparations along with L-Arginine (Molychem) and Tri-base from Merck Life Science Pvt. Ltd, India. Reduction was performed using 4 mM DTT for 30 minutes under stirring. SIB samples were generated after centrifugation at 12,000 rpm for 20 minutes. 5 mM GSSG (oxidized) was added 5-15 minutes prior to SIB addition to the refolding buffer (50 mM Tris, 5 % Sorbitol, 0.7 M Arginine, pH 9.0-11.0). Drip dilution was utilized for SIB dissolution, diluting them 30 times. Refolding was conducted in a beaker (500ml) equipped with a magnetic stirrer having fixed frequency (250 rpm) and different probes were inserted directly into the beakers for data acquisition during the process as shown in Figure 3a. Eutech pH 510 probe spanned from 0 to 14 and the temperature probe from 0 to 100°C. Each probe was standardized, calibrated, and tested against a standard solution. For all the recorded data, the corresponding value of yield and product concentration was calculated using Reversed Phase HPLC. Here, C-8 RP-HPLC column (Zorbax 300SB, 4.6 X 150 mm) (Agilent Technologies) was used on an Agilent 1260 HPLC system. Mobile phases A and B consisted of deionized water with 0.1 % (v/v) TFA and ACN with 0.1 % TFA (v/v), respectively. 30 % mobile phase B was utilized for column equilibration and a linear gradient of 30 to 38 % B carried out for elution. Method was observed at a flow rate of 1.2 mL/min and 80 °C oven temperature. The detection was performed at 215 nm wavelength to achieve the chromatogram.

Screening of process parameters

Foremost, the importance of refolding time was analyzed. It was seen through RP-HPLC analysis, that from the 8^{th} hour of refolding, the reaction exhibited formation of the native Ranibizumab molecule. Samples at various time points were analyzed using process analytical tools to validate the observation and the time points were confirmed. Based on this, the refolding time was divided in 2 segments i.e., 0 to 8hr and 8hr to total refolding time (24 hrs in the present case). In this study, segment based experimental technique is adopted for acquiring data for characterization space over conventional refolding technique as it results in higher yield [Figure S1]. For the first segment, it was seen that temperature, concentration, and pH play a critical role whereas in second segment, temperature and pH were deemed important.

Once all the relevant parameters were established after FMEA analysis, the next step was to establish the parameters relevant for each section and determine the permissible range. Two approaches were used to shortlist parameters step by step as shown in Table (S1). JMP® statistical software (SAS Institute, Cary, NC) was used for the analysis.

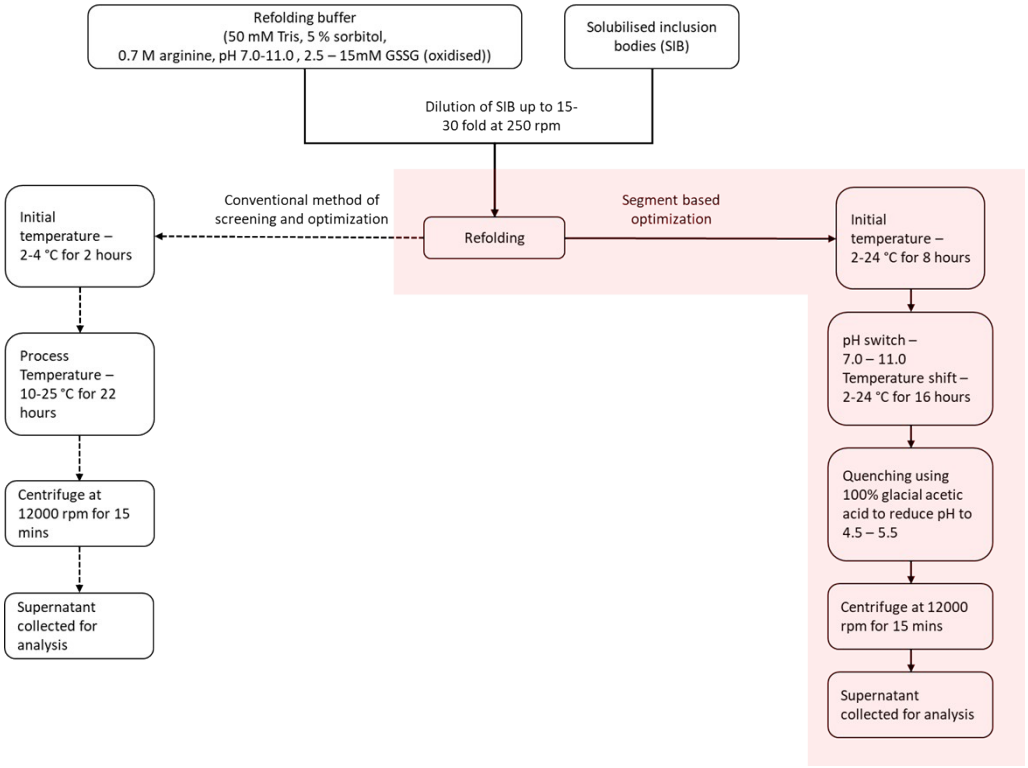

Figure S1: Segment based methodology adopted for refolding of protein in the present study for data acquisition

Table S1: Summary of the parameters shortlisted based on fractional factorial screening and OFAT approaches; X indicate the rejected parameters and ⚑ indicate the parameters considered for next step

S. No.	Parameters	Segment 1		Segment 2	
		Fractional Factorial screening	**OFAT screening**	**Full factorial screening**	**OFAT screening**
1	Concentration IB	⚑	⚑	**s**	**X**
2	Concentration of urea	**X**	**X**	**X**	**X**
3	Concentration of DTT	**X**	**X**	**X**	**X**
4	Concentration of GSSG	**X**	**X**	**X**	**X**
5	pH of solubilization buffer	**X**	**X**	**X**	**X**
6	pH of refolding buffer	⚑	⚑	⚑	⚑
7	Refolding time	⚑	⚑	⚑	⚑
8	Refolding temperature;	⚑	⚑	⚑	⚑
9	Sequence of additives addition	**X**	**X**	**X**	**X**

S2: Mathematical formulation of linear regression technique

In linear regression, the model assumes linear relationship between the dependent and regressor variables and can be written as

$$y_i = b_0 + b_1x_{i1} + \ldots\ldots\ldots + b_px_{ip} + \varepsilon_i = x_i^T b + \varepsilon_i \quad \ldots\ldots\ldots(s1)$$

Where, i varies from 1 to n and ^{T} denotes the transpose. In matrix form, equation 2 can be expressed as

$$Y = Xb + \varepsilon \quad \ldots\ldots\ldots(s2)$$

Where, Y is $[y_1, y_2, \ldots\ldots\ldots y_n]^T$, X is $[x_1^T, x_2^T, \ldots\ldots\ldots, x_n^T]^T$, b is $[b_0, b_1, \ldots\ldots\ldots, b_n]^T$, and ε is $[\varepsilon_1, \varepsilon_2, \ldots\ldots\ldots, \varepsilon_n]^T$. Numerous techniques, differing in terms of computation simplicity, have been developed for parameter estimation. One of the most common techniques is error minimization wherein square of error, ε, is minimized to estimate the regression coefficient b. The technique assumes model prediction is $Y_i \approx b_0 + \sum_{j=1}^{m} b_j x_j^i$ with x_i being $[x_1^i, x_2^i, \ldots\ldots\ldots, x_n^i]$ and b being $[b_0, b_1, \ldots\ldots\ldots, b_n]$. The objective function is formulated to determine the value of b such that $\varepsilon^2 = \arg\left(\sum_{i=1}^{n} (b.x_i - y_i)^2\right)$ is minimum.

S3: Mathematical formulation of Support vector regression (SVR)

For a finite dataset represented as F = (x_i,y_i) for i varying from 1 to n, obtained from an unknown relation y = g(x), support vector regression technique aims to determine a function y = f(x) such that f is as close to g as possible. This function f is derived using data from the set F. For linear regression, linear relation between y and x is assumed to be linear with x as feature vector having 'm' dimensions. Mathematically, it can be written as

$$y = g(x,w,b) = w.x + b = \sum w_j x_j + b \quad \text{.....(s3)}$$

Where, j varies from 1 to m. Once the relationship is finalised, the next step is to determine the hyperplane equation. It is given as :

$$y = f(x,\alpha,b) = \sum \alpha_k y_k x_k.x + b \quad \text{.....(s4)}$$

Where, x_k are support vectors with corresponding labels as y_k.

Support vectors and the parameters (α, b) can be determined to build a model. However, for nonlinear systems, x data points are mapped on to new feature space. In this new space, the relation between the new feature vectors and y data points is assumed to be linear. The hyperplane equation is modified as :

$$y = f(x,\alpha,b) = \sum \alpha_k y_k K(x_k.x) + b \quad \text{.....(s5)}$$

Where, K is kernel function having inner product of feature vectors.

S4: Model performance evaluation metrics

Goodness of fit indices: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

AIC and BIC indices enable model section with highest explanatory power. These criteria are implemented for measurement of model fit trade off with model complexity and are calculated as

$$AIC = -2\log L + 2K \quad \text{……..(s6)}$$

$$BIC = -2 \log L + K \log N \quad \text{……(s7)}$$

Wherein L represents likelihood, K represents the number of model parameters, and N is the number of data points used to train a model. Lowest values of AIC and BIC are preferred. Of the two, BIC is preferred over AIC, as it severely penalizes the model for having multiple parameters. AIC is suitable for determining the best model for prediction whereas BIC is more apt for selecting the correct model.

Coefficient of determination (R^2)

R^2 is the most commonly used indicator for determining the accuracy of a regression algorithm with a range of [0, 1]. It is estimated by generating model predictions for the data not utilized in model training. The biggest advantage it offers is that it can be used to compare models generated using different datasets. It is given as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad \text{…….(s8)}$$

Where, $\bar{y}_i$ is the mean value of y_i. When the R^2 is equal to 1, it means that the regression model makes 100% accurate predictions. The greater the R^2 number, the better the fitting result is in general.

p-value

The p-value indicates the probability of finding a particular set of data under null hypothesis. It determines that whether the relationship that is observed in the dataset exists for larger population or not. The p value tests null hypothesis for each variable. The null hypothesis considers if there is no correlation between the independent and dependent variable. No correlation indicates no association between the changes in independent variable and shift in

dependent variable. If p value is less than threshold value then it is said that enough data is provided by the sample dataset to reject null hypothesis and there is a non-zero correlation between the variables whereas if the p value is more then it indicates that sufficient evidence is not there to conclude a non-zero correlation. So, the independent variables having p value less than threshold are statistically significant and are considered for the analysis. It is calculated by employing a statistical test on the data. Commonly, $p < 0.05$ is considered as a threshold. This means that there is a chance of 5% that the null hypothesis holds true.

S5: Statistical analysis

Table S2: Statistical analysis of the linear regression model with different interaction terms for segment 1

Parameters	Standard error of the coefficients	t_{Stat}	SumSq	DF	MeanSq	F	p value
For model Eq. 4							
C_1	0.7456	0.8005	6.466104	1	6.466104	0.475579	0.50355
pH_1	17.9211	1.1145	269.2814	1	269.2814	19.80552	0.000792
T_1	3.1367	-0.4847	152.429	1	152.429	11.21108	0.005798
C_1* pH_1	0.0737	-0.5069	3.494656	1	3.494656	0.25703	0.621355
C_1*T_1	0.0105	0.8311	9.391397	1	9.391397	0.690733	0.422144
pH_1 * T_1	0.3122	1.0174	14.0743	1	14.0743	1.035158	0.329025
C_1^2	0.0014	-2.4575	82.11668	1	82.11668	6.039643	0.030172
pH_1^2	0.8443	-1.0879	16.0935	1	16.0935	1.18367	0.297979
T_1^2	0.0379	-3.1914	138.4849	1	138.4849	10.1855	0.007754
For model Eq. 5							
C_1	0.4134	3.2005	6.157486	1	6.157486	0.503321	0.488255
pH_1	0.8326	4.2175	217.6091	1	217.6091	17.78764	0.000654
C_1^2	0.0096	-3.1163	39.09562	1	39.09562	3.195725	0.092785
T_1^2	0.0067	-4.3024	226.4565	1	226.4565	18.51084	0.000548
C_1^3	$6.0775e^{-5}$	2.9414	105.8504	1	105.8504	8.652343	0.009579

Table S3: Statistical analysis of the linear regression model with different interaction terms for segment 2

Parameters	Standard error of the coefficients	t_{Stat}	SumSq	DF	MeanSq	F	p value
For model eq. 6							
pH	238.3222	2.2687	83.68399	1	83.68399	5.781002	0.033252
Temp	0.5909	3.9109	108.71	1	108.71	7.509835	0.017918
pH^2	25.8822	-2.1979	65.62344	1	65.62344	4.533355	0.054636
$Temp^2$	0.0193	-3.3861	165.9782	1	165.9782	11.466	0.005406
pH^3	0.9316	2.1132	64.64578	1	64.64578	4.465817	0.056206

Supplementary S6: Comparison performance of SVR and GPR algorithm

	SEGMENT 1		SEGMENT 2	
Parameters	GPR	SVR	GPR	SVR
	Ardexponential kernel (sigma = 0.011208)	RBFkernel (kernel scale= 0.01)	Ardexponential kernel (Sigma = 1.67)	RBF kernel (kernel scale= 0.1)
R^2	0.97	0.98	0.99	0.93
Adjusted R^2	0.93	0.99	0.97	0.87
RMS training error	0.95	1.15	0.53	0.87
RMS cross val. error	1.81	2.52	5.51	1.9