Comparison of Machine Learning-enhanced dynamic hybrid models for a nanobody scorpion antivenom production with Escherichia coli

Irene Martínez Menéndez, Juan Camilo Acosta-Pavas, David Camilo Corrales Munoz, Susana María Alonso Villela, Balkiss Bouhaouala-Zahar, Georgios Georgakilas, Konstantinos Mexis, Stefanos Xenios, Theodore Dalamagas, Antonis Kokosis, et al.

Comparison of Machine Learning-enhanced dynamic hybrid models for a nanobody scorpion antivenom production with *Escherichia coli*

Irene Martínez Menéndez[1][0009-0005-4913-1591], Juan C. Acosta-Pavas[1][0000-0003-0398-7712], David Camilo Corrales[2][0000-0003-4717-3040], Susana María Alonso Villela[1][0000-0001-8290-1635], Balkiss Bouhaouala-Zahar[3,4][0000-0003-3147-245X], Georgios K. Georgakilas[5][0000-0003-1160-5753], Konstantinos Mexis[6][0009-0005-5605-3499], Stefanos Xenios[6][0009-0000-5688-363X], Theodore Dalamagas[7][0000-0002-5002-7901], Antonis Kokosis[6][0000-0002-8074-2818], Michael O'donohue[1][0000-0003-4246-3938], Luc Fillaudeau[1][0000-0002-6389-5441], Nadia Boukhelifa[8,9][0000-0002-0541-8022], Alberto Tonda[8,9][0000-0001-5895-4809] , César A. Aceves-Lara[1][0000-0001-6291-3655]

[1]TBI, Université de Toulouse, CNRS, INRAE, INSA, Toulouse, France
{martinez-men,aceves}@insa-toulouse.fr
[2]INRAE, UMS (1337) TWB, 135 Avenue de Rangueil, 31077 Toulouse, France
[3]Laboratoire des Biomolécules, Venins et Applications Théranostiques (LBVAT), Institut Pasteur de Tunis, 13 Place Pasteur, BP-74, 1002 Le Belvédère, Tunis, Tunisia
[4]Faculté de Médecine de Tunis, Université Tunis El Manar, 15 rue Djebel Lakhdhar, 10007 Bab Saadoun Tunis, Tunisia
[5]Laboratory of Genetics, Section of Genetics, Cell Biology and Development, Department of Biology, University of Patras, 26504, Patras, Greece
[6]Department of Process Engineering, NTUA, Iroon Politechniou 6 Zografou, Athens, Greece
[7]Athena Research Center, Marousi, Greece
[8] INRAE, AgroParisTech, Université Paris-Saclay UMR 518 MIA-PS, 91120, Palaiseau, France
[9]Institut des Systèmes Complexes de Paris Île-de-France (ISC-PIF), UAR 3611 CNRS, Paris, France

Abstract. This study compares several hybrid dynamic models built with offline data to optimize a nanobody production in *Escherichia coli*. These models combine the insights of mechanistic knowledge and Machine Learning algorithms (ML). Four ML algorithms were tested: Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) and K-Nearest Neighbors (KNN). The efficacy of each model was analyzed with three metrics: mean absolute error (MAE), root mean squared error (RMSE), normalized root mean squared error (NRMSE). Results demonstrate that the gaussian SVM model was the best performing, with a NRMSE between 0.063 and 0.665. The insights gained are pivotal for advancing computational modeling techniques in biomanufacturing, with a specific emphasis on the production of recombinant therapeutic proteins, such as nanobodies, using bacterial hosts.

Keywords: Hybrid model, Machine learning, Bioprocess, *Escherichia coli*, Nanobody, Antivenom.

1 Introduction

Molecularly targeted biologics have become the new generation of therapies due to their precise specificity, affinity and their effective therapeutic performance [1]. Nanobodies are single domain antibody fragments significantly smaller than conventional antibodies, which allows for better tissue penetration and targeting of cryptic epitopes that are inaccessible to larger molecules [2]. They are more thermostable and the production can be carried out with microbial cell hosts which represent a cost-effective alternative to monoclonal antibodies [3]. The bacteria *Escherichia coli* has been engineered as the preferred host for the efficient production of the antivenom proteins, leveraging its scalability and cost-effectiveness [4, 5].

The Food and Drug Administration (FDA) has promoted the digitalization of the biomanufacturing industry in recent years, through knowledge-aided methodologies [6, 7]. The tendency towards reducing the costs improving the production and quality of therapeutic molecularly targeted biologics goes beyond legislative measures. Enhancing cell culture productivity within existing bioreactors has been proven an effective strategy [8].

The use of Machine Learning to complement mechanistic models is a promising tool to enhance upstream bioprocess understanding and control [9–11]. By analyzing the data generated during protein production, ML can identify complex protein-related patterns and relationships that traditional methods might miss [12, 13]. This approach has been preliminarily explored in various studies, showcasing its potential in improving process efficiency and product quality [14–16].

One of the challenges associated with offline datasets in protein production lies in the small size and heterogeneous nature of the experimental data. This requires the testing of several ML-enhanced hybrid models that can perform when datasets are not big enough to deploy artificial neural networks [17].

This study aims to identify the best ML enhanced hybrid dynamic model for the described biopharmaceutical production among Decision Trees (DT), K-Nearest Neighbors (KNN), Random Forests (RF) and Support Vector Machine (SVM).

The paper is organized as follows:

- Section 2 outlines the methodology, detailing the development of the hybrid models, the experimental setup, and the specific ML algorithms employed.
- Section 3 presents the results and discussion from the comparative analysis, focusing on the practical implications and the efficiency of each model in the specific context of nanobody production.
- Section 4 concludes with a summary of the findings and suggests future directions for the application of ML-enhanced hybrid models in biotechnological processes.

2 Materials and Methods

2.1 Experimental Data

The experimental data were obtained from a previous work [18] and are summarized in Table 1. Two recombinant strains were used: WK6 *E. coli* NbF12-10 and CH10-12 scorpion anti-toxin bispecific nanobody and its humanized format highly neutralizing whole venom [19, 20]. A total of seven experiments were carried out in a working volume of 2L at a biomass production temperature (T) of 37°C for the batch and fed-batch modes. Dissolved oxygen ($pO2$) was kept over 15 %, and the pH was regulated at 7 for the entire experiment. Induction was carried out in each reactor at the temperature displayed in Table 1. Nanobody concentration was monitored over time for each experiment, with a number of data points, between 15 and 28. Following NiNTA downstream purification step, the specific recombinant protein concentration was determined using SDS-PAGE and densitometry [21]. The bioreactor was equipped with a BioPAT® MFCS (Sartorius) software, capable of measuring every 5 seconds different online variables such as pH, agitation, pO2 and temperature.

Table 1. Protein instances measured at each experiment [18].

Experiment	*E. coli* Strain	Induction temperature (°C)	Experimental offline measurements
1	CH10-12	28	17
2		30	21
3		33	28
4		37	27
5		29	28
6		32	21
7	NbF12-10	29	25

2.1. Mechanistic Model

The bioreactor is assumed to be an infinitely mixed culture without any transfer limitation Corrales. The mass balances for biomass, glucose, proteins and volume are given by Eq. (1) – (4).

$$\frac{dX}{dt} = r_X - X * \frac{F_{in}}{V} \tag{1}$$

$$\frac{dS}{dt} = \frac{F_{in}}{V}(S_{in}) - r_S - S * \frac{F_{in}}{V} \tag{2}$$

$$\frac{dP}{dt} = r_P - P * \frac{F_{in}}{V} \tag{3}$$

$$\frac{dV}{dt} = F_{in} \qquad (4)$$

Where *X* and *S* are the concentrations of biomass and glucose, in g/L, and *P* is the concentration of nanobody, in mg/L. F_{in} is the glucose flowrate, in L/h, *V* is the volume of the liquid phase in the bioreactor, in L, S_{in} is the concentration of glucose in the feed, in g/L, and r_X is the production rate of biomass, in g/h, and defined by Eq. (5). r_S is the consumption rate of glucose, in g/h, as defined by Eq. (6). AND RP

$$r_X = \mu * X \qquad (5)$$

$$r_S = -\frac{1}{Y_{SX}} * \mu * X \qquad (6)$$

With µ as the specific growth rate, in 1/h, as defined by a Monod kinetics in Eq. (7); Y_{SX} as the yield coefficient of substrate, in g cdw/Gs and g_{wcdw} is the wet cell dry weight.

$$\mu = \frac{\mu_{max} * S}{K_S + S} \qquad (7)$$

$$\mu_{max} = \mu'_{max} * T + b \qquad (8)$$

μ_{max} is the maximum specific growth rate in 1/h and K_S is the substrate saturation coefficient, in g/L. µ'max is a temperature-dependent factor in 1/h, T is the temperature of the bioreactor, in °C, and b is a temperature-independent factor, in 1/h.

The specific productivity connects the ML models with the mechanistic dynamic model thanks to the nanobody production rate r_p:

$$r_P = \begin{cases} 0 & if\ there\ is\ not\ induction \\ q_P * X & if\ there\ is\ an\ induction \end{cases} \qquad (9)$$

Where X is the biomass in g/L.

2.2 Machine Learning Algorithms

Four different machine learning algorithms were explored to build the hybrid models: Decision Tree CART (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and gaussian SVM. All the algorithms were trained using the Scikit-learn Python library 1.3.0. A fivefold cross-validation with three repetitions was performed without a training/test split given the reduced quantity of experimental data.

The training dataset is the experimental data provided by Alonso Villela *et al.* (2021) for both recombinant strains NF12-10 and CH10-12 and was described in Table 1. The ML algorithms estimate q_P based on two variables:

$$q_P = f(T, \mu) \tag{10}$$

Where T is the temperature in ^{o}C, µ the growth rate in 1/h and q_p is the specific productivity of the nanobody production $mg_{protein}/(g_{wcdw} \cdot h)$

Here, $mg_{protein}$ refers to the nanobody produced mass in milligrams; and h the production time in hours.

2.3 Statistical Analysis

The performance of the predictive models for produced nanobody concentration was analyzed comparing with the experimental values at the same times. The metrics used were MAE, RMSE and NRMSE:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y - \hat{Y}\right| \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y - \hat{Y}\right)^2} \tag{12}$$

$$NRMSE = \frac{\frac{1}{n}\sqrt{\sum_{i=0}^{n}(Y_i - \hat{Y}_i)^2}}{Y_{max} - Y_{min}} \tag{13}$$

With n the number of instances, Y the testing label set for the protein concentration, $\hat{Y}$ the predicted value of the protein concentration using the soft sensors, and $\bar{Y}$ the mean value of Y.

3 Results

3.1 Validation of the construction of the model

Figure 1 shows the scorpion venom specific nanobody concentration at different temperatures for both strains. Four ML-enhanced hybrid models are displayed, in comparison with the experimental data points. In the three last graphs, the model Random Forest diverges from experimental data compared to the others. Also, SVM is the best fitted model for the majority of experiments. Finally, DT and KNN have a similar behavior between each other showing a lower fitting to the data

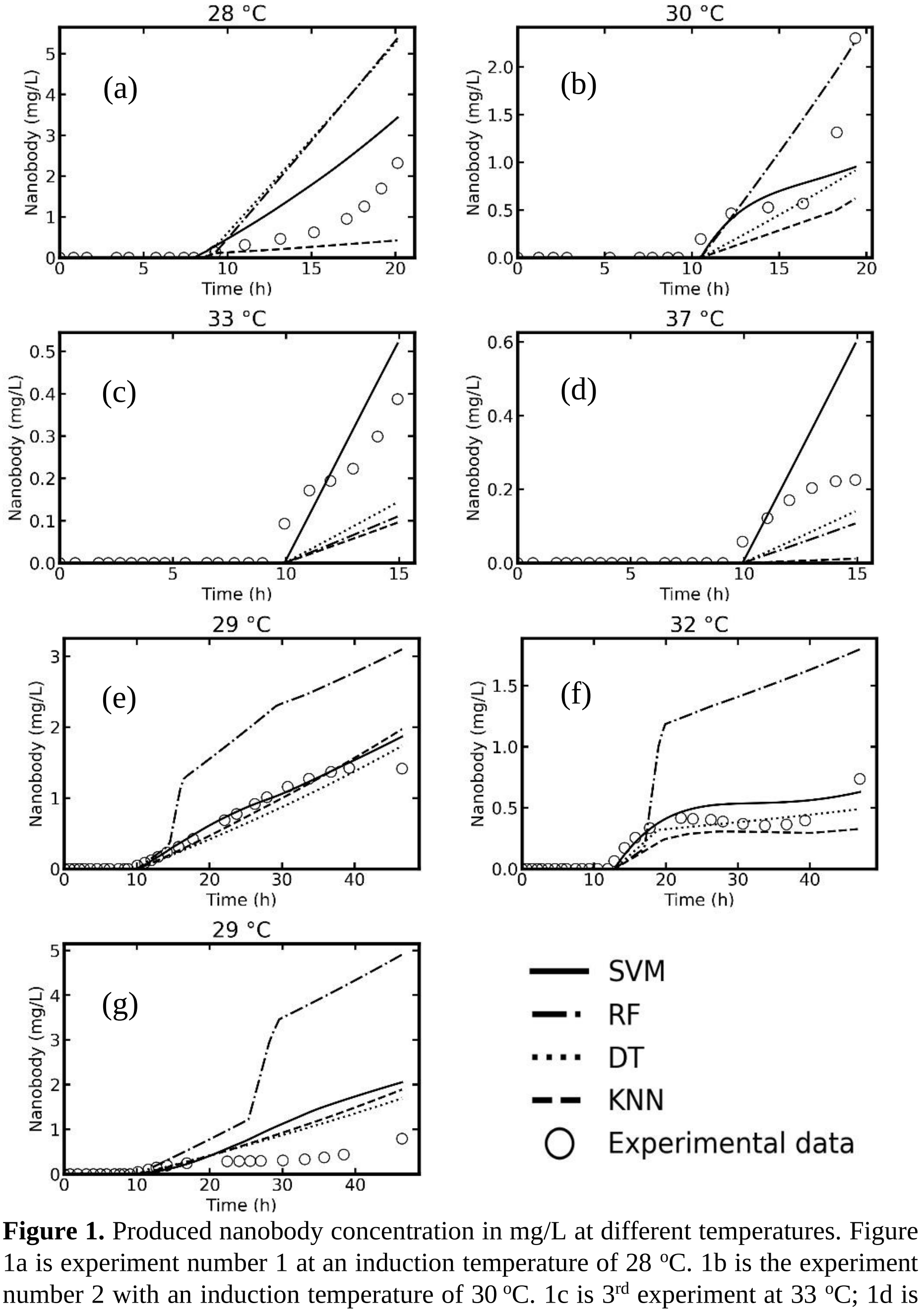

Figure 1. Produced nanobody concentration in mg/L at different temperatures. Figure 1a is experiment number 1 at an induction temperature of 28 ºC. 1b is the experiment number 2 with an induction temperature of 30 ºC. 1c is 3rd experiment at 33 ºC; 1d is the 4th at 37 ºC; 1e is the 5th at 29 ºC; 1f is the 6th at 32 ºC; 1g is the 7th experiment at 29 ºC with Nb1012.

Table 2 shows the different metrics to compare the performance of the four models. Concretely, MAE, RMSE and NRMSE for each model and in each experiment. The data suggests that the SVM model generally offers robust performance across a range of metrics, particularly fitting for experiments 2, 3, and 5. This consistent low-error rate indicates that the SVM model has high predictive accuracy under certain conditions, validating its utility in scenarios similar to these experiments. However, the notable drop in the model’s predictions performance for protein concentration for experiments 4 and 7 highlighting potential areas for improvement, particularly in terms of overfitting and generalization, which are critical for the model's applicability to a wider range of bioprocess scenarios.

The RF model demonstrates diverse fitness to the data, with fitter results in 2 but significantly poorer outcomes in 6 and 7. The insights gained from these observations are invaluable for refining the model-building process, particularly in understanding the parameters under which the RF model is likely to perform well or poorly.

The DT model shows a consistent and competitive performance in certain experiments, notably 5 and 6, suggesting that DT models can effectively capture and predict complex bioprocess dynamics under specific conditions. However, the increased model simulation errors to fit experiments 1 and 7 could indicate limitations that must be addressed during the model construction phase, for example, using online data or other offline variables as estimators for the ML model.

The KNN model's performance in experiment 1, followed by more variable fitness in other experiments, suggests that while it can be effective, its applicability may be limited by dataset specificity. This variability highlights the need for careful consideration of the KNN model's parameters and the characteristics of the dataset during the model construction process.

Table 2. MAE, $RMSE$, and $NRMSE$ for each ML-enhanced hybrid model in each experiment compared to the experimental data.

Experiment	1	2	3	4	5	6	7
Induction T (°C)	28	30	33	37	33	32	29
SVM							
MAE	0.474	0.157	0.025	0.045	0.034	0.050	0.302
RMSE	0.762	0.375	0.051	0.108	0.090	0.083	0.527
NRMSE	0.329	0.163	0.132	0.478	0.090	0.083	0.665
RF							
MAE	0.988	0.148	0.049	0.032	0.462	0.360	0.924
RMSE	1.646	0.308	0.097	0.062	0.462	0.600	0.924
NRMSE	0.710	0.134	0.251	0.274	0.740	0.816	2.143
DT							
MAE	1.005	0.170	0.044	0.027	0.082	0.036	0.213
RMSE	1.646	0.394	0.087	0.052	0.127	0.064	0.369
NRMSE	0.710	0.171	0.224	0.231	0.089	0.087	0.466
KNN							
MAE	0.323	0.234	0.051	0.046	0.059	0.061	0.239
RMSE	0.627	0.501	0.102	0.091	0.121	0.107	0.420
NRMSE	0.270	0.218	0.102	0.404	0.085	0.146	0.530

4 Conclusions

Four hybrid models with machine learning were compared to predict the specific production rate of a NbF12-10 and CH10-12 bispecific Nanobodies, expressed in recombinant WK6 *E.coli* strain [20] based on the Temperature and growth rate. SVM has the highest fit with the experimental data, showing low errors MAE between 0.025 and 0.474; RMSE between 0.051 and 0.762; and NRMSE between 0.063 and 0.665. The SVM-hybrid model has been highlighted for its ability to handle non-linear and small-sized datasets, specifically for predicting variables such as the specific production rate of the nanobody concentration from the temperature and biomass growth rate.

The next steps in leveraging these hybrid models and to feed them with new data involve their application across a broader spectrum of biomanufacturing conditions, coupled with the integration of real-time data acquisition technologies, and the development of soft sensors, advancing towards the concept of a digital twin for the bioprocess.

In summary, the shift towards hybrid computational models, supported by real-time protein production monitoring and digital twin technologies can be paradigm-changing if we are able to gather enough representative data. This approach promises not only to enhance the efficiency and yield of a specific recombinant protein bioprocess but also to pave the way for new innovations in the production of biologics and other complex biomolecules.

Acknowledgments. This work was funded (or co-funded) by the European Union under the Horizon Europe project Bioindustry 4.0, grant n. 101094287, the INRAE PhD scholarships and it has benefited from a State grant managed by the National Research Agency under the "Investissements d’Avenir" program with the reference ANR-18-EURE-0021

5 References

1. Lee, Y.T., Tan, Y.J., Oon, C.E.: Molecular targeted therapy: Treating cancer with specificity. Eur. J. Pharmacol. 834, 188–196 (2018). https://doi.org/10.1016/j.ejphar.2018.07.034.
2. Hamers-Casterman, C., Atarhouch, T., Muyldermans, S., Robinson, G., Hammers, C., Songa, E.B., Bendahman, N., Hammers, R.: Naturally occurring antibodies devoid of light chains. Nature. 363, 446–448 (1993). https://doi.org/10.1038/363446a0.
3. Muyldermans, S.: Nanobodies: Natural Single-Domain Antibodies. Annu. Rev. Biochem. 82, 775–797 (2013). https://doi.org/10.1146/annurev-biochem-063011-092449.
4. Alonso Villela, S.M., Kraïem-Ghezal, H., Bouhaouala-Zahar, B., Bideaux, C., Aceves Lara, C.A., Fillaudeau, L.: Production of recombinant scorpion antivenoms in E. coli: current state and perspectives. Appl. Microbiol. Biotechnol. 107, 4133–4152 (2023). https://doi.org/10.1007/s00253-023-12578-1.
5. Bouhaouala-Zahar, B., Ben Abderrazek, R., Hmila, I., Abidi, N., Muyldermans, S., El Ayeb, M.: Immunological Aspects of Scorpion Toxins: Current Status and Perspectives. Inflamm. Allergy - Drug Targets. 10, 358–368 (2011). https://doi.org/10.2174/187152811797200713.
6. U.S. Food and Drug Administration: Pharmaceutical Quality for the 21st Century: A Risk-Based Approach Progress Report., https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/pharmaceutical-quality-21st-century-risk-based-approach-progress-report, (2017).
7. Yu, L.X., Raw, A., Wu, L., Capacci-Daniel, C., Zhang, Y., Rosencrance, S.: FDA’s new pharmaceutical quality initiative: Knowledge-aided assessment & structured applications. Int. J. Pharm. X. 1, 100010 (2019). https://doi.org/10.1016/j.ijpx.2019.100010.
8. Rodrigues, M.E., Costa, A.R., Henriques, M., Azeredo, J., Oliveira, R.: Technological progresses in monoclonal antibody production systems. Biotechnol. Prog. 26, 332–351 (2010). https://doi.org/10.1002/btpr.348.
9. Nikita, S., Mishra, S., Gupta, K., Runkana, V., Gomes, J., Rathore, A.S.: Advances in bioreactor control for production of biotherapeutic products. Biotechnol. Bioeng. 120, 1189–1214 (2023). https://doi.org/10.1002/bit.28346.
10. Khanal, S.K., Tarafdar, A., You, S.: Artificial intelligence and machine learning for smart bioprocesses. Bioresour. Technol. 375, 128826 (2023). https://doi.org/10.1016/j.biortech.2023.128826.

11. Puranik, A., Dandekar, P., Jain, R.: Exploring the potential of machine learning for more efficient development and production of biopharmaceuticals. Biotechnol. Prog. 38, e3291 (2022). https://doi.org/10.1002/btpr.3291.
12. Gangadharan, N., Sewell, D., Turner, R., Field, R., Cheeks, M., Oliver, S.G., Slater, N.K.H., Dikicioglu, D.: Data intelligence for process performance prediction in biologics manufacturing. Comput. Chem. Eng. 146, 107226 (2021). https://doi.org/10.1016/j.compchemeng.2021.107226.
13. Corrales, D.C., Alonso Villela, S.M., Bouhaouala-Zahar, B., Cescut, J., Daboussi, F., O'Donohue, M., Fillaudeau, L., Aceves Lara, C.A.: Dinamic Hybrid Model for Nanobody-based Antivenom Production (scorpion antivenom) with E. coli CH10-12 and E. coli NbF12-10. Submitted. In: European Symposium on Computer Aided Process Engineering, ESCAPE 34, pp. 1–6 (2024)., (2024).
14. Nikita, S., Thakur, G., Jesubalan, N.G., Kulkarni, A., Yezhuvath, V.B., Rathore, A.S.: AI-ML applications in bioprocessing: ML as an enabler of real time quality prediction in continuous manufacturing of mAbs. Comput. Chem. Eng. 164, 107896 (2022). https://doi.org/10.1016/j.compchemeng.2022.107896.
15. Pinto, J., Mestre, M., Ramos, J., Costa, R.S., Striedner, G., Oliveira, R.: A general deep hybrid model for bioreactor systems: Combining first principles with deep neural networks. Comput. Chem. Eng. 165, 107952 (2022). https://doi.org/10.1016/j.compchemeng.2022.107952.
16. Pham, T.D., Manapragada, C., Sun, Y., Bassett, R., Aickelin, U.: A scoping review of supervised learning modelling and data-driven optimisation in monoclonal antibody process development. Digit. Chem. Eng. 7, 100080 (2023). https://doi.org/10.1016/j.dche.2022.100080.
17. Mondal, P.P., Galodha, A., Verma, V.K., Singh, V., Show, P.L., Awasthi, M.K., Lall, B., Anees, S., Pollmann, K., Jain, R.: Review on machine learning-based bioprocess optimization, monitoring, and control systems. Bioresour. Technol. 370, 128523 (2023). https://doi.org/10.1016/j.biortech.2022.128523.
18. Alonso Villela, S.M., Ghezal-Kraïem, H., Bouhaouala-Zahar, B., Bideaux, C., Aceves Lara, C.A., Fillaudeau, L.: Effect of temperature on the production of a recombinant antivenom in fed-batch mode. Appl. Microbiol. Biotechnol. 105, 1017–1030 (2021). https://doi.org/10.1007/s00253-021-11093-5.
19. Hmila, I., Cosyns, B., Tounsi, H., Roosens, B., Caveliers, V., Abderrazek, R.B., Boubaker, S., Muyldermans, S., El Ayeb, M., Bouhaouala-Zahar, B., Lahoutte, T.: Pre-clinical studies of toxin-specific Nanobodies: Evidence of in vivo efficacy to prevent fatal disturbances provoked by scorpion envenoming. Toxicol. Appl. Pharmacol. 264, 222–231 (2012). https://doi.org/10.1016/j.taap.2012.07.033.
20. Hmila, I., Saerens, D., Abderrazek, R.B., Vincke, C., Abidi, N., Benlasfar, Z., Govaert, J., Ayeb, M.E., Bouhaouala-Zahar, B., Muyldermans, S.: A bispecific nanobody to provide full protection against lethal scorpion envenoming. FASEB J. 24, 3479–3489 (2010). https://doi.org/10.1096/fj.09-148213.
21. Alonso Villela, S.M., Kraïem, H., Bouhaouala-Zahar, B., Bideaux, C., Aceves Lara, C.A., Fillaudeau, L.: A protocol for recombinant protein quantification by densitometry. MicrobiologyOpen. 9, 1175–1182 (2020). https://doi.org/10.1002/mbo3.1027.