UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Bioquímica y Ciencias Biológicas

Tesis para la obtención del Grado Académico de

Doctor en Ciencias Biológicas

"Desarrollo de nuevas herramientas quimiométricas y su aplicación en la Tecnología Analítica de Procesos"

Lic. Fabricio Alejandro Chiappini

Dr. Héctor Casimiro Goicoechea
Director de Tesis

Dra. Ángela Guillermina Forno
Co-director de Tesis

Laboratorio de Desarrollo Analítico y Quimiometría (LADAQ)
Cátedra de Química Analítica I
Facultad de Bioquímica y Ciencias Biológicas
Universidad Nacional del Litoral

-2021-

AGRADECIMIENTOS

A la Universidad Nacional del Litoral.

A la Facultad de Bioquímica y Ciencias Biológicas.

A las instituciones que financiaron este trabajo: el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET); la empresa Zelltek SA; la Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT); la *European Society for Animal Cell Technology* (ESACT).

A mi familia, por inculcarme los valores que hoy defiendo incansablemente, por transmitirme el gusto por el estudio y por brindarme el privilegio de estudiar en la Universidad pública. A papá, mamá, Ceci, a mis sobrinos Emanuele y Gianluca, y a mis abuelos. A la memoria de mi nono Santiago.

A mi mejor amigo, Ale.

A mis amigos de la Facultad y a mis amigos de la música.

A mis amigos de hoy y de siempre. A los que están cerca y a los que están lejos.

A mi Director, Héctor, por haberme invitado a sumarme a su equipo y por abrirme las puertas hacia un mundo nuevo y desconocido de conocimiento que me colmó de grandes satisfacciones y que terminó de consolidar mi vocación científica. Por su gran generosidad y compañerismo.

A mi Codirectora, Guillermina y a todo el personal de Zelltek, por la confianza, por la predisposición y por abrir las puertas de la empresa para que podamos llevar adelante esta investigación.

Al Dr. Alejandro Olivieri, por haber contribuido invaluablemente a mi formación científica y personal, porque representa el ideal de científico al que aspiro: talentoso, con una vocación de trabajo incansable y, por sobre todo, un gran ser humano, generoso y humilde.

A los integrantes del Laboratorio de Desarrollo Analítico y Quimiometría, quienes brindaron su conocimiento y ayuda con la mejor predisposición en mis primeros pasos del doctorado, con quienes comparto esta hermosa tarea de investigar y de quienes aprendo todos los días un poco.

A los compañeros y amigos del Departamento de Matemática de la Facultad, con quienes tengo el privilegio de compartir la maravillosa y difícil misión de enseñar.

A mis primeros padrinos en la ciencia, Julio y Virginia.

A Stella, por la confianza y por ser la primera persona que me motivó a embarcarme en la carrera del doctorado.

¡Gracias!

PUBLICACIONES

Publicaciones de la tesis:

- FA Chiappini, S Azcarate, MR Alcaraz, HC Goicoechea (2021) Prospective inference of bioprocess cell viability through chemometric modelling of fluorescence multiway data. *Biotechnol Progress*, e3173.
- FA Chiappini, HC Goicoechea, AC Olivieri (2020) MVC1_GUI: a MATLAB graphical user interface for first-order multivariate calibration. An upgrade including artificial neural networks modelling. *Chemom Intell Lab Syst* 206: artículo 104162.
- FA Chiappini, F Allegrini, HC Goicoechea, AC Olivieri (2020) Sensitivity for multivariate calibration based on multilayer perceptron artificial neural networks. *Anal Chem* 92 (18): 12265-12272.
- FA Chiappini, CM Teglia, AG Forno, HC Goicoechea (2020) Modelling of bioprocess non-linear fluorescence data for at-line prediction of etanercept based on artificial neural networks optimized by response surface methodology. *Talanta* 210, artículo 120664.
- FA Chiappini, MR Alcaraz, HC Goicoechea, AC Olivieri (2019) A graphical user interface as a new tool for scattering correction in fluorescence data. *Chemom Intell Lab Syst* 193, artículo 103810.
- FA Chiappini, MR Alcaraz, HC Goicoechea (2019) An improved signal-conservative methodology to cope with Rayleigh and Raman signals in fluorescence landscapes. *Chemom Intell Lab Syst* 187: 6-10.

Publicaciones por trabajos en colaboración durante la carrera de doctorado:

- FA Chiappini, MR Alcaraz, GM Escandar, HC Goicoechea, AC Olivieri (2021) Chromatographic applications in the multi-way calibration field. *Molecules* 26, 6357.
- FA Chiappini, F Gutierrez, HC Goicoechea, AC Olivieri (2021) Achieving the analytical second-order advantage with non-bilinear second-order data. *Anal Chim Acta* 1181, artículo 338911.
- FA Chiappini, F Gutierrez, HC Goicoechea, AC Olivieri (2021) Interference-free calibration with first-order instrumental data and multivariate curve resolution. When and why? *Anal Chim Acta* 1161, artículo 338465.
- ML Senovieski, SA Gegenschatz, FA Chiappini, CM Teglia, MJ Culzoni, HC Goicoechea (2020) In-syringe dispersive liquid-liquid microextraction vs. solid phase extraction: a comparative analysis for the liquid chromatographic determination of three neonicotinoids in cotyledons. *Microchem J* 158, artículo 105181.

ÍNDICE

ABREVIATURAS

Aclaración sobre el lenguaje utilizado: la gran mayoría de las abreviaturas utilizadas en esta tesis corresponden a definiciones en inglés, debido a que las mismas son de uso extensivo en las áreas estadística, quimiometría y química analítica. A continuación, se listan las abreviaturas utilizadas (en orden de aparición en el texto principal), junto con sus definiciones en español y/o inglés, según corresponda.

Abreviatura	Definición
QbT	*Quality-by-Testing*
FDA	*Food and Drug Administration*
QbD	*Quality-by-Design*
PAT	Tecnología Analítica de Procesos (*Process Analytical Technology*)
CPP	Parámetro crítico de proceso (*critical process parameter*)
CQA	Atributo crítico de calidad (*critical quality attribute*)
DoE	Diseño experimental (*design of experiments*)
RSM	Metodología de la superficie de respuesta (*response surface methodology*)
UV	Ultravioleta (*ultraviolet*)
IR	Infrarrojo (*infrared*)
NIR	Infrarrojo cercano (*near infrared*)
HPLC	Cromatografía líquida de alto rendimiento (*high-performance liquid chromatography*)
DAD	Detector de arreglo de diodos (*diode array detector*)
FSFD	Detector de fluorescencia de barrido rápido (*fast-scanning fluorescence detector*)
EEM	Matriz de excitación-emisión (*excitation-emission matrix*)
AFOM	Cifra analítica de mérito (*analytical figure of merit*)
IUPAC	Unión Internacional de Química Pura y Aplicada (*International Union of Pure and Applied Chemistry*)
PCA	Análisis de componentes principales (*principal component analysis*)
PLS	Regresión en cuadrados mínimos parciales (*partial least-squares regression*)
ADN	Ácido desoxirribonucleico

TNF	Factor de necrosis tumoral (*tumor necrosis factor*)
Fc	Fragmento cristalizable de una inmunoglobulina
CHO	Ovario de hámster chino (*chinese hamster ovary*)
IPC	Control intraproceso (*in-process control*)
GUI	Interfaz gráfica de usuario (*graphical user interface*)
PARAFAC	Análisis paralelo de factores (*parallel factor analysis*)
MCR-ALS	Resolución multivariada de curvas con cuadrados mínimos alternantes (*multivariate curve resolution – alternating least squares*)
PLS-DA	cuadrados mínimos parciales con análisis discriminante (*PLS discriminant analysis*)
ANN	Red neuronal artificial (*artificial neural network*)
UA	Unidades arbitrarias
Exi	Excitación
Emi	Emisión
Int. de Fl.	Intensidad de fluorescencia
OFL	Ofloxacina
RES	resorrufina
p.a.	Calidad pro análisis
PMT	Fotomultiplicador (*photomultiplier*)
PMH	Altura máxima de pico (*peak maximum high*)
PMW	Mitad del ancho de pico (*peak medium width*)
U-PCA	PCA desdoblado o vectorizado (*unfolded-PCA*)
U-PLS	PLS desdoblado o vectorizado (*unfolded-PLS*)
LV	Variable latente (*latent variable*)
PC	Componente principal (*principal component*)
SVD	Descomposición en valores singulares (*singular value decomposition*)
NIPALS	Cuadrados mínimos parciales iterativos no lineales (*non-linear iterative partial least squares*)
PVA	Porcentaje acumulado de varianza explicada
CV	Validación cruzada (*cross-validation*)

PRESS	Suma de los cuadrados de los residuos de predicción (*predicted residual error sum of squares*)
RMSECV	Raíz del error cuadrático medio de validación cruzada (r*oot mean square error in cross validation*)
LOF	Falta de ajuste (*lack of fit*)
LOO	Dejar una muestra fuera (*leave-one-out*)
MLP	Perceptrón multicapa (*multilayer perceptron*)
SP	Perceptrón simple (*simple perceptron*)
FFD	Diseño factorial completo (*full factorial design*)
CCD	Diseño central compuesto (*central composite design*)
BBD	Diseño de Box y Behnken (*Box-Behnken design*)
iPLS	PLS por intervalos (*interval-PLS*)
RMSEC	Raíz del error cuadrático medio de calibración (*root mean square error in calibration*)
APaRP	Gráfica de residuos parciales aumentados (*augmented partial residuals plot*)
RMSEP	Raíz del error cuadrático medio de predicción (*root mean square error in prediction*)
REP%	Error relativo porcentual de predicción (*percentage relative error of prediction*)
EJCR	Región de confianza elíptica (*elliptic joint confidence región*)
RS	Región seleccionada
ANOVA	Análisis de la varianza (*analysis of variance*)
RBF	Funciones de base radial (*radial basis function*)
RMSEM	Raíz del error cuadrático medio de monitoreo (*root mean square error in monitoring*)
LOD	Límite de detección (*limit of detection*)
LOQ	Límite de cuantificación (*limit of quantitation*)
NAS	Señal neta de analito (*net analyte signal*)

SÍMBOLOS, UNIDADES Y NOTACIÓN MATEMÁTICA

Con respecto a la notación matemática: los escalares se representan con letra minúscula itálica; las matrices, en letra mayúscula negrita; los vectores, en minúscula negrita. El superíndice T indica matriz transpuesta; el superíndice -1, matriz inversa; el superíndice $+$, matriz pseudoinversa. El símbolo || || representa la norma euclídea de un vector. El subíndice A indica matriz truncada o reconstruida con A componentes principales o variables latentes; el subíndice aum se refiere a una matriz aumentada. Las derivadas ordinarias se indican en notación de Leibniz; las derivadas parciales se denotan con la d de Jacobi; el acento circunflejo indica estimación.

Ry	Señal de dispersión de Rayleigh
Rm	Señal de dispersión de Raman
Ry1	Dispersión de Rayleigh de primer orden
Ry2	Dispersión de Rayleigh de segundo orden
Rm1	Dispersión de Raman de primer orden
Rm2	Dispersión de Raman de segundo orden
λ	Longitud de onda
nm	nanómetro
MΩ	mega-ohmio
cm	centímetro
µm	micrómetro
mg	miligramo
L	litro
mL	mililitro
min	minuto
V	voltio
n	Orden de señal de dispersión; también, número de componentes en datos simulados
$\mathbf{X}_r$	Matriz de excitación-emisión de fluorescencia cruda (sin preprocesamiento)
tol	Tolerancia
μ_{em}	Posición media esperada de una señal de dispersión en un espectro de emisión
$\mu_{\mathrm{ex},i}$	i-ésima longitud de onda de excitación
wnr	Diferencia en número de onda entre la posición del Ry y del Rm en una EEM (del inglés, *wavenumber for Raman*)
$\mathbf{X}$	Matriz de datos; matriz de datos espectrales (ya sean de segundo orden, EEMs, o de primer orden, EEMs desdobladas)
I	Número total de muestras o patrones de calibración (entrenamiento)

J	Número de variables predictoras o sensores en un modo instrumental
A	Número óptimo de variables latentes (o componentes principales); también, número de neuronas en la capa de entrada de un MLP
$\mathbf{T}$	Matriz de *scores* de PCA/PLS en un conjunto de datos de entrenamiento (calibrado)
$\mathbf{P}$	Matriz de *loadings* de PCA/PLS en un conjunto de datos de entrenamiento (calibrado)
$\mathbf{t}$	Vector de *scores* de PCA/PLS
$\mathbf{p}$	Vector de *loadings* de PCA/PLS
K	Número de sensores en un segundo modo instrumental (dato de segundo orden)
x_{ij}	Elemento de $\mathbf{X}$ que corresponde a la i-ésima muestra y el j-ésimo sensor espectral
$\mathbf{C}$	Matriz de perfiles de MCR en el modo excitación
$\mathbf{S}$	Matriz de perfiles de MCR en el modo emisión
$\mathbf{E}$	Matriz error
$\mathbf{y}$	Vector que contiene los N valores de la variable respuesta
y_i	i-ésimo elemento del vector $\mathbf{y}$
$\mathbf{b}$	Vector de coeficientes de regresión
$\mathbf{e}$	Vector error
$\mathbf{x}_{test}$	Vector de variables predictoras (espectro) de una muestra *test*
$\mathbf{t}_A$	Vector de scores de una muestra test proyectada en el espacio PLS de A variables latentes
$\mathbf{W}$	Matriz de *loadings* de PLS
g	Número de clases en un problema de clasificación
Sn_g	Sensibilidad de un clasificador asociada a la clase g
c_{gg}	Cantidad de muestras de la clase g correctamente clasificadas
N_g	Cantidad de muestras de la clase g
Pr_g	Precisión de un clasificador para la clase g
n_g	Cantidad de muestras totales asignadas a la clase g
Sp_g	Especificidad de un clasificador para la clase g
N_k	Cantidad de muestras de la clase k
c_{kg}	Cantidad de muestras de la clases g que fueron asignadas a la clase k
Acc	Exactitud de clasificación
NER	Tasa de no error de clasificación (*non-error-rate*)
DO	Oxígeno disuelto
h	hora
°C	grado centígrado
mm	milímetro
M	molar

Q	Residuo Q de PCA (suma de los cuadrados de los residuos)
T^2	Residuo T^2 de Hotelling
u_i	Suma ponderada de un perceptrón para la i-ésima muestra de entrenamiento
w	Peso sináptico
N	Número de neuronas en la capa oculta (*hidden layer*) de un MLP
O	Número de neuronas en la capa de salida (*output layer*) de un MLP
$\mathbf{W}$	Matriz de pesos sinápticos
∇e	Gradiente de la función error e
μ	Velocidad de aprendizaje
D	Función deseabilidad de Derringer
F	Estadístico F
R^2	Coeficiente de determinación
x	Señal analítica univariada
β_0	Coeficiente constante asociado a un modelo lineal univariado (poblacional)
β_1	Coeficiente de la variable predictora y asociado a un modelo lineal univariado (poblacional)
ε	Término de error del modelo lineal univariado (poblacional)
SEN	Sensibilidad
σ_x	Desvío estándar; incertidumbre asociada a la señal instrumental x
σ_y	Desvío estándar; incertidumbre asociada a la concentración y
γ	Sensibilidad analítica
$x(j)$	Señal de una muestra a la longitud de onda j
$s_n(j)$	Señal del n-ésimo componente puro a la longitud de onda j
y_n	Concentración del n-ésimo componente puro
$\mathrm{SEN}_{\mathrm{MC}}$	Sensibilidad estimada mediante simulación Montecarlo
$\mathrm{SEN}_{\mathrm{BS}}$	Sensibilidad estimada mediante método *bootstrap*
a_y	Parámetro de escalado del MLP para una dada función de transferencia
b_y	Parámetro de escalado del MLP para una dada función de transferencia
E	Esperanza matemática
σ^2	Varianza
$\sum_{\mathrm{x}}$	Matriz de covarianza del error
h	Leva de la muestra

RESUMEN

La Tecnología Analítica de Procesos (PAT) constituye una serie de lineamientos metodológicos impulsados por la *Food and Drug Administration* (FDA) en el año 2004 que se enmarcan en el nuevo paradigma de control de calidad de la industria farmacéutica, conocido como *Quality-by-Design* (QbD), cuyo propósito principal es el de aumentar la flexibilidad, eficiencia y consistencia de los procesos productivos. Para ello, la PAT promueve el desarrollo e implementación de estrategias analíticas de alto rendimiento que permitan un diseño, monitoreo y control de proceso más eficientes. Los desarrollos analíticos en el contexto de la PAT se basan principalmente en el uso de técnicas analíticas multivariadas, en donde la quimiometría juego un rol fundamental para el modelado de datos.

En este trabajo se tomó como caso de estudio un bioproceso estandarizado, en su etapa de fermentación, para la producción de una proteína recombinante de uso terapéutico, utilizando células animales como plataforma biológica. El objetivo de la investigación fue desarrollar estrategias analíticas novedosas para el monitoreo de variables críticas del bioproceso, cuyas características sigan los lineamientos de la PAT. Para ello, se recabaron datos de variables de proceso y se generaron datos multidimensionales de fluorescencia (matrices de excitación-emisión, EEMs), que luego fueron tratados mediante diferentes herramientas quimiométricas, con el fin extraer información relevante sobre aspectos críticos de la fermentación, permitiendo el reconocimiento de patrones y la elaboración de modelos predictivos.

En este sentido, se utilizó una gran variedad de métodos quimiométricos que implicaron diferentes enfoques de análisis, tales como preprocesamiento de datos, análisis exploratorio, diseño y optimización experimental, calibración y clasificación multivariadas. En particular, respecto al preprocesamiento, se desarrolló una estrategia novedosa para la corrección computacional de señales de dispersión en EEMs, cuyo rendimiento fue analizado en forma comparativa con otras metodologías previamente reportadas en la literatura. Por otra parte, se aplicaron métodos quimiométricos para el análisis exploratorio multivariado que permitieron la caracterización del proceso respecto a la información contenida en los datos generados (variables de proceso y EEMs). Además, se desarrolló una estrategia de PAT cualitativa para el monitoreo *at-line* de la viabilidad celular, basada en el uso de un algoritmo de clasificación. Por otro lado, a partir de la información espectral, se desarrolló una estrategia de PAT cuantitativa para el monitoreo *at-line* de la concentración del producto de la fermentación, basado en el uso del algoritmo perceptrón multicapa (MLP), que constituye un tipo específico de red neuronal artificial (ANN). Finalmente y dado que

estos métodos de regresión no paramétricos se encuentran menos caracterizados desde el punto de vista estadístico, se desarrollaron y validaron ecuaciones para el cálculo de la sensibilidad y la sensibilidad analítica del MLP, basadas en la teoría de propagación de errores. Ambos parámetros constituyen dos Cifras Analíticas de Mérito (AFOMs) de gran relevancia para la caracterización, validación y comparación de las potencialidades de los métodos analíticos. Esta contribución permitió calcular la sensibilidad del método de calibración PAT desarrollado y compararla con la de la técnica cromatográfica univariada de referencia.

En todas sus etapas, la complejidad que presentó el sistema en estudio desde el punto de vista analítico motivó no solamente la aplicación de métodos quimiométricos conocidos, sino que también se lograron desarrollar elementos teóricos y metodologías quimiométricas novedosas para dar respuesta a los interrogantes planteados, que además resultan extensibles a otros contextos analíticos.

SUMMARY

Process Analytical Technology (PAT) constitutes a methodological framework promoted by Food and Drug Administration (FDA) in 2004, in the context of Quality-by-Design (QbD) paradigm for pharmaceutical industry, that aims to improve the flexibility, efficiency and consistency of their productive processes. For this purpose, PAT guidelines encourage the development and implementation of high-performance analytical techniques, increasing the efficiency in process design, monitoring and control. PAT strategies are mainly based on the used of multivariate analytical methodologies, where chemometrics plays a fundamental role for data modelling.

In this work, the fermentation step of a standard mammalian cell-based bioprocess for the production of a recombinant protein for therapeutic use was considered as a case study. The aim of this investigation was to develop novel analytical strategies for the monitoring of critical process variables, in agreement with the principles of PAT. In this sense, data from process variables were recorded and multi-way fluorescence data were generated (excitation-emission matrices, EEMs). These data were analysed by means of diverse chemometric methods with the aim of extracting relevant information regarding critical aspects of the fermentation. Multivariate data analysis was implemented for pattern recognition and the obtention of predictive models.

In this context, a great variety of chemometric methods were utilized, with different analysis approaches, such as data pre-processing, exploratory analysis, experiment design and optimization and multivariate calibration and classification. In particular, regarding EEM pre-processing, an alternative methodology for digital correction of scattering signals was developed. Its performance was comparatively assessed with other previously reported strategies. On the other hand, chemometric algorithms were implemented for the characterization of the bioprocess, regarding the information contained in the generated data (process variables and EEMs), through multivariate exploratory analysis. Besides, a qualitative PAT strategy was proposed for the at-line monitoring of process cell viability, based on the use of a classification algorithm. Moreover, multivariate spectral information enabled the development of a quantitative PAT strategy for the at-line monitoring of the fermentation product, based on the use of multilayer perceptron (MLP) algorithm, which represents a specific type of artificial neural network (ANN). Finally, and owing to the fact that these non-parametric regression methods are less characterized from the statistical point of view, equations for the theoretical estimation of MLP sensitivity and analytical sensitivity were derived, based on error propagation theory. Both parameters constitute two Analytical Figures of Merit (AFOMs) of great relevance for the characterization, validation and comparison of the

performance of analytical methodologies. This contribution enabled the estimation of the analytical sensitivity of the developed calibration PAT strategy, and hence, it was possible to make a comparison with the sensitivity of the reference univariate chromatographic method.

During all stages, the complexity of the system under study from the analytical perspective motivated not only the use of well-stablished chemometric methods, but also the development of both theoretical and methodological contributions for fundamental chemometrics. All these achievements are also applicable to other analytical contexts.

Introducción

1. Tecnología Analítica de Procesos (PAT)

1.1. PAT: definición, origen y fundamento en el contexto de la industria biofarmacéutica

Durante los últimos veinte años, se ha producido en la industria farmacéutica un cambio importante de paradigma en lo que respecta al control y aseguramiento de la calidad. Tradicionalmente, la calidad de un producto es demostrada a partir de un conjunto de exhaustivas pruebas que se le realizan una vez que es liberado de la línea de producción, es decir, que bajo el paradigma clásico, la calidad se evalúa posmanufactura. Esta manera de concebir a la calidad, habitualmente se conoce como paradigma *Quality-by-Testing* (QbT), el cual ha demostrado ser ampliamente satisfactorio para garantizar la calidad y seguridad de los productos farmacéuticos [1]. Sin embargo, los avances científicos y tecnológicos y las nuevas demandas del mercado motivan continuamente la innovación por parte de las industrias, lo cual conlleva necesariamente una mejora continua en el desarrollo de nuevos productos y procesos productivos. No obstante, este progreso genera una industria cada vez más compleja que necesariamente, trae aparejado nuevas problemáticas que obligan a repensar los criterios de calidad.

Esta situación se manifestó fuertemente a principios de 2000, cuando las agencias regulatorias de la industria farmacéutica comenzaron a evidenciar nuevos problemas de trazabilidad y consistencia de los procesos productivos. Esta situación puso en evidencia que, para analizar y demostrar la calidad de sus productos, las empresas no sólo debían enfocarse en un control intensivo posmanufactura, sino en implementar nuevas estrategias que les permitieran diseñar y monitorear los procesos de producción de manera más integrada y eficiente. En este contexto es que la agencia regulatoria estadounidense *Food and Drug Administration* (FDA) ha impulsado durante las últimas dos décadas un nuevo paradigma conocido como *Quality-by-Design* (QbD). Este establece que la calidad debe ser concebida de manera integrada al proceso, desde su desarrollo temprano, utilizando el conocimiento científico acerca del producto y su proceso de manufactura, junto con los análisis de riesgo. Así, este paradigma entiende a la calidad como una variable más del proceso y no como una salida de este, de manera que debe ser diseñada, monitoreada y controlada en tiempo real durante toda la vida del producto [2,3].

Una vez desarrollado, optimizado y aprobado un proceso, la nueva manera de concebir a la calidad en el marco de la QbD establece como objetivo de máxima que las industrias sean capaces de implementar estrategias de monitoreo y control en línea de variables críticas que permitan, por un lado, mejorar la eficiencia en la toma de decisiones para optimizar recursos y, por otro lado, incrementar continuamente el

conocimiento acerca del proceso de manera tal de garantizar la seguridad y eficacia de sus productos, explicar las variaciones o desviaciones observadas y mejorar la flexibilidad ante las agencias regulatorias para la implementación de cambios posaprobación [1].

En el contexto de la QbD, la FDA publicó en el año 2004 el documento "*Guidance for Industry. PAT: a framework for innovative pharmaceutical development, manufacturing and quality assurance*" [4]. Este documento introduce el concepto de Tecnología Analítica de Procesos (*Process Analytical Technology*, PAT) y establece un marco metodológico que aspira a alcanzar los principios de la QbD. Según este documento, la PAT se define como un sistema para diseñar, analizar y controlar el proceso de manufactura a través de mediciones en tiempo real de los parámetros críticos de proceso (CPPs)[1] que afectan a su rendimiento y/o a los atributos críticos de calidad del producto final (CQAs)[2]. Esencialmente, la PAT promueve el desarrollo y la implementación de estrategias analíticas de alto rendimiento que permitan un diseño, monitoreo y control más eficiente del proceso, asegurando su consistencia y trazabilidad y la calidad y seguridad del producto final [5,6,2].

Resulta evidente que las industrias farmacéuticas en las cuales se utilizan células animales como plataforma para la producción de proteínas terapéuticas (biofarmacéuticas o industrias de biosimilares o de biofármacos[3]) son un blanco natural para la implementación de los principios de la QbD y la PAT, lo cual reviste un interés creciente que se evidencia por las publicaciones científicas realizadas durante las últimas décadas [7,8].

Los cultivos de células animales constituyen en la actualidad la plataforma de elección para la producción de biofármacos debido a su capacidad para expresar y excretar proteínas de elevado peso molecular y con las características fisicoquímicas adecuadas [9]. Sin embargo, a diferencia de otros sistemas de expresión, las células de mamífero presentan algunas desventajas tales como una velocidad de crecimiento lenta, elevados requerimientos nutricionales y niveles de expresión relativamente bajos. Estas características hacen que los bioprocesos basados en estas plataformas celulares resulten comparativamente más costosos y menos eficientes. Asimismo, es frecuente que, debido a la complejidad intrínseca que poseen los sistemas biológicos, estos

[1] CPP es cualquier atributo del proceso cuya variabilidad tiene un impacto sobre algún parámetro crítico de calidad del proceso/producto y, por lo tanto, debe ser monitoreado o controlado para garantizar su calidad (incluye materias primas, variables operativas, variables fisicoquímicas y biológicas) [1].
[2] CQA son las propiedades o características físicas, químicas, biológicas o microbiológicas que definen la calidad del producto y, por lo tanto, deben mantenerse dentro de un determinado rango de valores aceptables. En el contexto de la QbD, los CQA se relacionan con la seguridad, eficacia y estabilidad de un producto [1].
[3] Se define como biofármaco a todo tipo de reactivo de uso clínico, vacuna o droga producida mediante biotecnología moderna con fines diagnósticos, profilácticos o terapéuticos [9].

procesos experimenten problemas de consistencia y trazabilidad, aun bajo estrictas condiciones estándares de operación. En este contexto, la aplicación de estrategias de PAT resulta sumamente deseable en el sentido de que permiten obtener información acerca de la calidad del proceso en tiempo real, reduciendo los tiempos de producción, y mejorando la eficiencia en el monitoreo y control de los CPPs. Esto último, además, implica la capacidad de detectar fallas o desvíos de manera prematura, de manera tal de mejorar la toma de decisiones y posibilitar la aplicación de acciones correctivas. Es evidente que todos estos aspectos impactan de manera directa sobre la optimización de los recursos y, por lo tanto, sobre la rentabilidad del proceso.

1.2. Objetivos y niveles de aplicación de las herramientas de PAT. Relación con la quimiometría

El documento de la FDA establece que los objetivos específicos que persigue la PAT son:

- reducir el tiempo de producción;
- prevenir el rechazo, descarte o reprocesamiento de lotes;
- incrementar la automatización para mejorar la seguridad de los operadores y reducir los errores humanos;
- incrementar la capacidad de proceso, optimizando el uso de materiales y energía;
- facilitar el régimen de proceso continuo para mejorar su eficiencia y el manejo de las fuentes de variabilidad.

Tal como establece su definición, la PAT implica un conjunto de herramientas para diseñar, monitorear y controlar un proceso. En este sentido entonces, las estrategias PAT se pueden agrupar y describir en términos metodológicos, de acuerdo a estos tres aspectos [4]:

1. *Nivel 1 o PAT para diseño del proceso:*

   Desde el punto de vista físicoquímico y biológico, los bioprocesos constituyen sistemas complejos multifactoriales. En este sentido, es bien sabido que los experimentos de optimización univariados no tienen en cuenta las interacciones entre los factores. Por lo tanto, el diseño óptimo del entorno experimental (límites aceptables de las variables del proceso) debe realizarse desde un enfoque multivariado. Durante el desarrollo inicial del proceso (escala laboratorio o planta piloto), el análisis de riesgo inicial permite la identificación de los CPPs, los cuales deben ser optimizados en relación con los CQAs. Las técnicas de diseño (DoE) y optimización de experimentos (metodología de la superficie de respuesta, RSM) constituyen una de las herramientas más potentes

para alcanzar este objetivo. Otras estrategias incluyen simulación de procesos y técnicas estadísticas de reconocimiento de patrones.

2. *Nivel 2 o PAT para monitoreo o análisis del proceso:*

Este tipo de estrategias han sido ampliamente desarrolladas y descriptas en la bibliografía, fenómeno que se ve favorecido por los avances en técnicas instrumentales capaces de ser automatizadas y que permiten la recolección de gran cantidad de datos del proceso y en tiempos muy breves. Además, muchas de estas técnicas resultan no destructivas para el material del proceso y son capaces de brindar información sobre atributos biológicos del sistema.

En el nivel PAT de monitoreo, en general, se pretende monitorear una o más variables críticas de proceso (CPPs) en tiempo real. Dependiendo cada dispositivo o técnica instrumental, la adquisición de datos para el monitoreo en el contexto de la PAT se puede efectuar de las siguientes maneras:

- *at-line*: se aísla una muestra del proceso (generalmente en forma manual) y esta es analizada próxima a la unidad de proceso en la que se generó;
- *on-line*: se adquieren datos en tiempo real luego de un muestreo automático (la muestra puede o no ser devuelta a la línea de proceso);
- *in-line*: el dato se adquiere en tiempo real a través de un dispositivo ubicado directamente en el seno de la línea de proceso (no se realiza muestreo).

Estas tres maneras de adquisición de datos en un proceso caracterizan a una estrategia de monitoreo de tipo PAT-QbD. Las técnicas instrumentales utilizadas deben presentar algunas características particulares, tales como, robustez, capacidad de automatización, bajo costo y realizar la adquisición de datos de manera rápida. Asimismo, para las mediciones *in-* y *on-line* se espera que sean no destructivas y poco invasivas. Por su parte, para las mediciones *at-line* es deseable que se requiera un volumen pequeño de muestra, con un mínimo o nulo pretratamiento de la misma, de manera de acortar lo más posible el tiempo de análisis. Las formas de adquisición de datos en línea se distinguen de la forma tradicional de medición fuera de línea (*off-line*), en la cual la muestra es analizada en una unidad de proceso diferente a la que le dio origen (por ejemplo, el laboratorio de control de calidad). En general, la adquisición de datos *off-line* es una característica importante del modelo de calidad QbT y suele ser realizada a través de técnicas analíticas sofisticadas y validadas por entes regulatorios. Asimismo, bajo esta metodología de monitoreo y dependiendo de la complejidad

de la técnica involucrada, el tiempo de retardo (*delay*) entre el muestreo y la obtención del resultado analítico puede ser considerable. En este sentido, la PAT persigue acortar dicho lapso de tiempo para así aumentar la eficiencia en el monitoreo del proceso.

3. *Nivel 3 o PAT para control del proceso:*

Estas herramientas constituyen el nivel más alto y sofisticado de un desarrollo de PAT, ya que integra toda la información y conocimiento generado por los niveles anteriores, junto con metodologías específicas de control automatizado de procesos. Una estrategia PAT de control se alimenta de la información proporcionada por el monitoreo en línea de los CPPs para generar acciones correctivas automáticas y en tiempo real de manera de mantener al proceso dentro de los límites establecidos en la etapa de diseño y garantizar así su consistencia. Debido a su complejidad, este nivel de PAT ha sido menos reportado en la literatura.

Para alcanzar sus objetivos, las estrategias de PAT implican necesariamente la generación de una gran cantidad de datos. Estos deben ser apropiadamente procesados y modelados para que brinden información relevante, lo cual se realiza frecuentemente mediante una gran diversidad de técnicas estadísticas multivariadas [10]. Asimismo y como se explicará en las siguientes secciones, muchas de las metodologías analíticas empleadas para las estrategias de PAT están basadas en técnicas espectroscópicas [11]. De esta manera, resulta evidente la relación entre la PAT y la quimiometría, ya que el objetivo principal de esta disciplina apunta al modelado de datos químicos para la extracción de información [12-14]. En la siguiente sección se definirán conceptos fundamentales propios del área de la quimiometría que servirán de soporte teórico a lo largo de todo el trabajo.

2. Quimiometría: definición y conceptos fundamentales

2.1. Definición y subáreas de la quimiometría

La quimiometría se define como una disciplina de la química analítica que tiene como objetivo extraer información relevante de sistemas químicos mediante el modelado de datos, a través de una combinación de técnicas matemáticas, estadísticas y computacionales [15,16]. Si bien originalmente fue concebida para resolver problemas químicos, esta disciplina se ha extendido ampliamente a contextos bioquímicos y biológicos por la versatilidad de sus técnicas y su naturaleza interdisciplinaria.

Dentro de esta área se pueden distinguir tres subáreas o especialidades:

a. Diseño y optimización de experimentos: por un lado, el diseño experimental (DoE) consiste en estudiar, modelar y predecir la importancia e interacción entre los factores que influyen en una dada respuesta experimental, minimizando el número de experimentos. Por otra parte, la optimización consiste en encontrar las condiciones experimentales (niveles de los factores) que optimizan una o varias respuestas experimentales. La optimización experimental también se denomina como RSM [17].
b. Reconocimiento de patrones: consiste en modelar las propiedades de un sistema con el objetivo de analizar su estructura y características subyacentes, a fin de estudiar similitudes o diferencias de un conjunto de muestras. Esto permite luego establecer agrupamientos o clases entre las mismas [18].
c. Calibración: consiste en modelar la relación entre una variable o propiedad de interés de un sistema (respuesta) y una o más variables medibles experimentalmente (predictoras), con el objetivo de efectuar predicciones de la variable de interés a partir de futuros valores de las predictoras [19].

2.2. Dimensión de los datos

En quimiometría, se entiende como "orden" a la propiedad de los datos instrumentales medidos para una muestra experimental y caracteriza la clase de objeto matemático que puede construirse a partir de dichos datos. De esta manera, por ejemplo, si se obtiene una lectura de absorbancia a una determinada longitud de onda, los datos son de orden cero. Si se mide un espectro, las lecturas podrán acomodarse en un vector que matemáticamente tiene una dimensión y el dato se denomina de primer orden. Si, por el contrario, para cada muestra se tiene la capacidad instrumental de generar información en dos modos instrumentales, entonces se obtiene una matriz de números, que constituye un dato de segundo orden, puesto que matemáticamente una matriz tiene dos dimensiones. De la misma manera, aumentando el número de modos instrumentales, se pueden generar datos de mayor orden. En general, los datos de segundo orden o mayor se consideran de orden superior [20].

Los datos de orden cero son propios de la estadística univariada, mientras que datos de orden uno o mayor, forman parte de los métodos de análisis multivariados. Si bien no hay una separación estricta, podemos decir que la quimiometría como tal surge a partir del desarrollo y aplicación de modelos multivariados para la resolución de problemas analíticos, mientras que los métodos univariados representan una antesala a la disciplina. En este trabajo se utilizará la denominación de "método quimiométrico" en referencia a métodos basados en datos de orden mayor o igual a uno. En particular,

los datos de orden superior también se suelen denominar “multi-modo” o “multi-vía” (*multi-way*) y su estudio constituye actualmente un área fuerte de investigación y desarrollo dentro de la disciplina [21,22].

En quimiometría es frecuente utilizar la palabra “orden” para referirse al orden instrumental, razón por la cual, el orden de un dato dependerá del instrumento utilizado y de su capacidad de generar información analítica en modos instrumentales diferentes. Es importante señalar que el incremento en el orden instrumental puede lograrse también por el acoplamiento de técnicas instrumentales en tándem. Por otro lado, el número de modos o vías caracteriza al arreglo matemático que puede construirse a partir de la yuxtaposición o concatenamiento de datos provenientes de diferentes muestras. Esto naturalmente permite aumentar la dimensión de un arreglo, pero no implica el incremento en el orden de los datos. Así, por ejemplo, la superposición de datos provenientes de un conjunto de muestras de orden cero, uno o dos, permite la obtención de arreglos de tipo vectoriales, matriciales y tridimensionales, respectivamente. Estos conceptos se resumen gráficamente en la Figura 1.

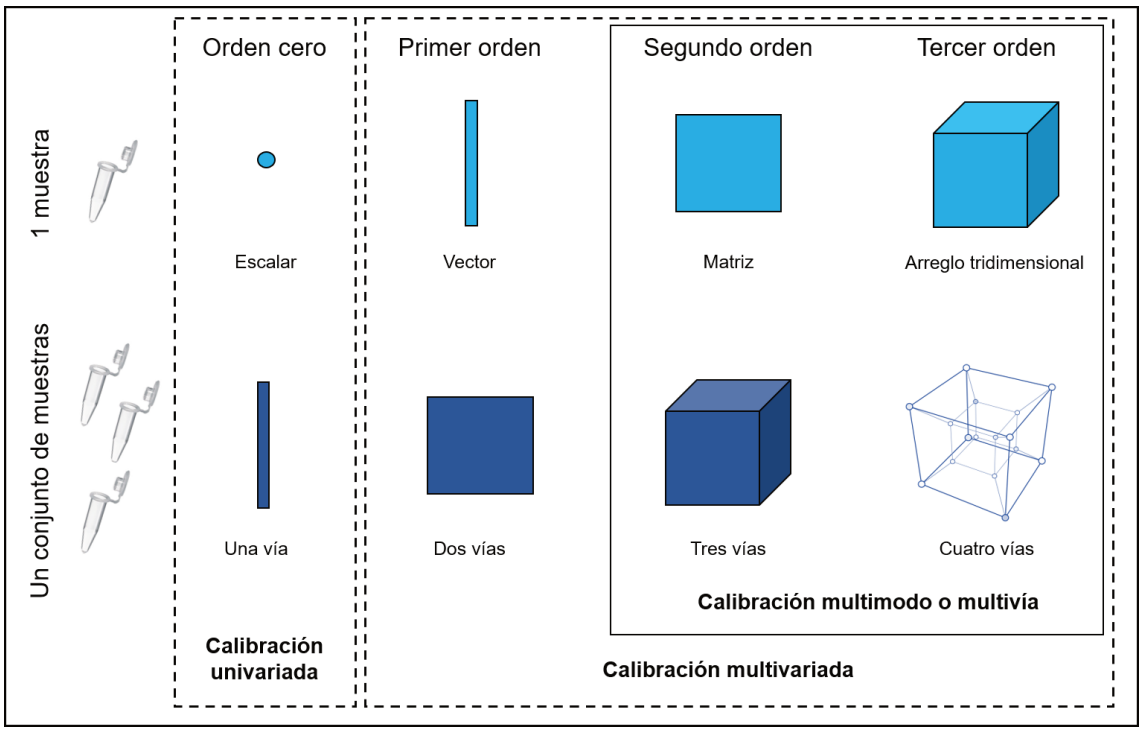


Figura 1. Dimensión de datos: tipos de datos de acuerdo al orden y los modos para una muestra (arriba) o un conjunto de muestras (abajo).

Tal como se ilustra en numerosas fuentes bibliográficas, el salto del mundo uni- al multivariado ha representado en los últimos años una revolución en el campo de la química analítica, no sólo desde el punto de vista científico, sino también por sus aplicaciones e impacto tecnológico y ambiental [23,24]. La mayoría de los métodos

quimiométricos de primer y segundo orden se basan en el uso de datos de origen espectral. En este sentido, las técnicas instrumentales más habituales por su accesibilidad y robustez incluyen: métodos espectroscópicos, tales como la absorción UV-visible, la espectroscopía infrarroja (IR) o infrarroja cercana (NIR), la fluorescencia y la espectroscopía Raman; y métodos cromatográficos, tales como la cromatografía líquida de alto rendimiento (HPLC) acoplada a diferentes sistemas de detección espectral como el arreglo de diodos (DAD) o el detector de fluorescencia de barrido rápido (FSFD) [25]. Todas estas técnicas se pueden encontrar en la bibliografía afín al tema de investigación de esta tesis [26,8,11].

En particular, este trabajo se basa esencialmente en el uso de datos de segundo orden. Ejemplos típicos de datos de segundo orden son las denominadas matrices de excitación-emisión de fluorescencia (EEMs) y las matrices espectro-tiempo de elución obtenidas por HPLC-DAD o HPLC-FLD. Estos dos ejemplos constituyen dos de los tipos de datos de segundo orden que más han sido reportados en la literatura en los últimos años. Esto se debe, por un lado, a que resulta cada vez más accesible su generación en el laboratorio y, por otro lado, a que cumplen con un supuesto matemático fundamental sobre el que están basados muchos de los métodos quimiométricos de segundo orden, y es la denominada propiedad de "bilinealidad de rango pequeño" que será tratada en los capítulos siguientes.

Por otro lado, una característica interesante que presentan los datos de orden superior, es la capacidad de construir diferentes tipos de arreglos matemáticos a partir de un conjunto de datos obtenidos para un número dado de muestras. En particular, para un conjunto de datos segundo orden, es posible generar un arreglo de tres vías, mediante el apilamiento de matrices. Por otra parte, las matrices también se pueden disponer una a continuación de la otra, tanto en el sentido de las filas como de las columnas. Se obtiene de esta manera, un arreglo que se denomina de dos vías aumentado en filas o aumentado en columnas, según el caso. Finalmente, si cada una de las matrices es desdoblada, de manera que cada fila o columna se dispone una a continuación de la otra, se rompe la estructura de dato de segundo orden y este pasa a ser un dato de primer orden (en la jerga quimiométrica se suele decir que las matrices son "linealizadas" o "vectorizadas"). El concatenamiento de matrices desdobladas en vectores, genera un arreglo de dos vías. Debido a que las diferentes maneras en que se pueden arreglar los datos de segundo orden permite la obtención de objetos matemáticos con diferentes propiedades, esta tarea resulta crucial a la hora de seleccionar el tipo de método quimiométrico. Los tipos de arreglos mencionados aquí se resumen gráficamente en la Figura 2.

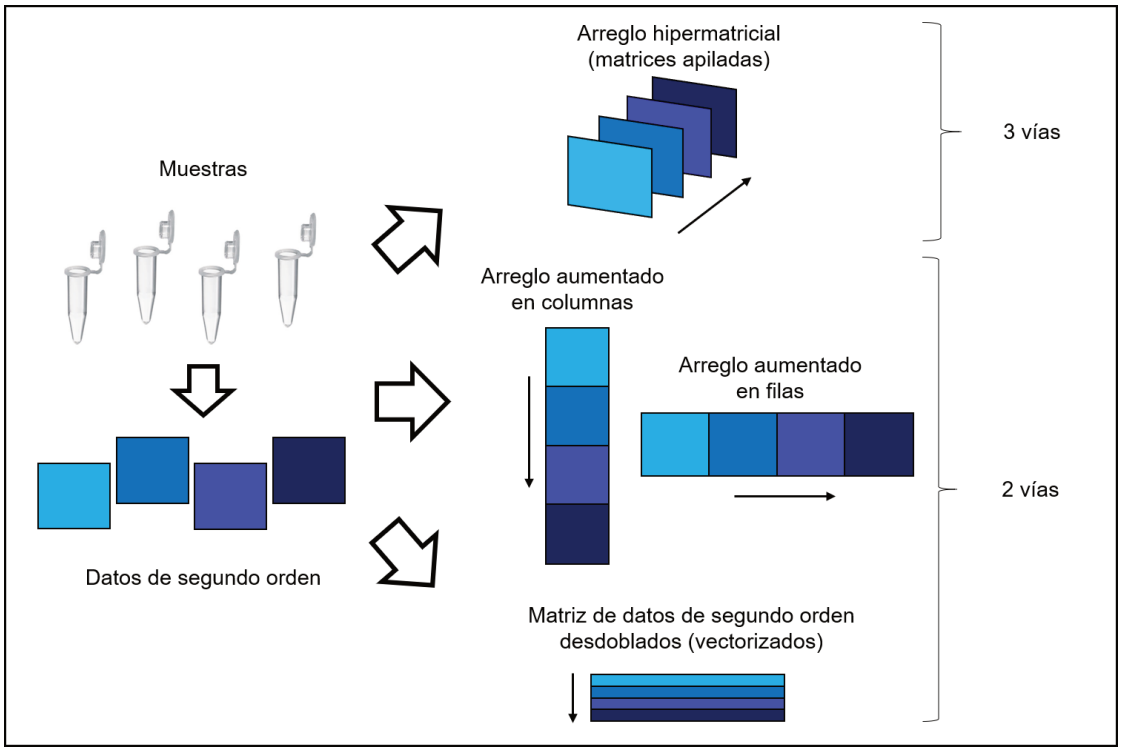


Figura 2. Resumen de los tipos de arreglos que se pueden obtener a partir de un conjunto de datos de segundo orden (matrices) para el posterior modelado quimiométrico. Las flechas negras indican el modo muestral.

2.3. Taxonomía de los métodos quimiométricos

En primer lugar, es importante establecer una diferencia conceptual entre los términos "método", "modelo" y "algoritmo", a pesar de que más adelante en el texto, en ocasiones, estos puedan ser utilizados de manera equivalente. Como su palabra lo indica, "método" hace referencia a un enfoque metodológico, a una manera de proceder. En quimiometría, muchos métodos están basados en la formulación de un modelo explícito que, por lo general, responde a modelos estadísticos. En este contexto, la palabra "modelo" refiere entonces a una formulación matemática que incluye hipótesis estocásticas y que tiene como objetivo representar matemáticamente la generación de un conjunto de datos. Finalmente, la palabra "algoritmo" se refiere a una secuencia de operaciones sistemáticas de cálculo para la obtención de un resultado, que puede o no, involucrar un modelo explícito.

Los avances tecnológicos en materia de técnicas analíticas instrumentales han permitido la generación cada vez más accesible de una diversidad importante de datos de complejidad y dimensionalidad crecientes. No es llamativo entonces que, ante este panorama, hayan proliferado y sigan siendo continuamente desarrollados una gran variedad de métodos quimiométricos. Resulta interesante, entonces, establecer algunos criterios de clasificación que orienten y faciliten su selección frente a un problema quimiométrico particular. Según los conceptos presentados en las secciones anteriores,

dos criterios de clasificación útiles tienen que ver con el objetivo del análisis (diseño y optimización experimental, reconocimiento de patrones o calibración) y con la dimensión de los datos (orden cero, primer orden, segundo orden, etc.). Así, por ejemplo, existen métodos de reconocimiento de patrones de primer y segundo orden y métodos de calibración de segundo y tercer orden, entre otros.

Existen varios otros criterios que han sido heredados de la estadística multivariada. Uno de ellos, por ejemplo, es aquel que los agrupa en métodos supervisados y no supervisados. En un método no supervisado, se busca analizar la estructura subyacente de un conjunto de observaciones experimentales, pero sin introducir ninguna información *a priori* sobre el sistema durante el entrenamiento o ajuste de los datos. Contrariamente, en un método supervisado, los objetos muestrales son "etiquetados" por el operador de acuerdo a algún criterio y dicha información se incorpora al entrenamiento del modelo. Estos conceptos permiten establecer que los métodos quimiométricos de diseño y optimización experimental y de calibración son, por definición, métodos supervisados. Para el caso de los métodos de reconocimiento de patrones, existen metodologías tanto no supervisadas (que suelen denominarse métodos de agrupamiento o clusterización[4]), como supervisadas (que suelen denominarse métodos de clasificación).

Otro criterio importante para describir un método quimiométrico está relacionado con la idea de si este se basa en un modelo explícito o no. El primer caso corresponde a la mayoría de los métodos quimiométricos clásicos, los cuales están basados en modelos lineales de la estadística paramétrica. Por otro lado, resulta cada vez más frecuente encontrar en la literatura, aplicaciones quimiométricas que se basan en el uso de estrategias no paramétricas, como el caso de los algoritmos de aprendizaje maquinal (*machine learning*) para la resolución de problemas que, debido a su complejidad, no se ajustan a los supuestos de los modelos paramétricos.

Finalmente, se puede realizar una distinción general entre modelos lineales (que asumen algún tipo de linealidad en las propiedades de los datos) y no lineales. En general, es habitual que los enfoques no paramétricos sean empleados para el tratamiento de problemas no lineales.

2.4. Quimiometría, química analítica y calibración

Es evidente que la química analítica y la quimiometría guardan una estrecha relación y que ambas se han beneficiado mutuamente debido a desarrollos que tienen que ver tanto con la quimiometría pura como con las aplicaciones analíticas. En este

[4] Anglicismo castellanizado proveniente de la palabra en inglés *clustering*, que significa agrupamiento.

contexto, un método quimiométrico puede utilizarse con un fin cualitativo exploratorio o cualitativo/cuantitativo predictivo. En el primer caso, se busca analizar las propiedades de los datos y sus cualidades subyacentes, que describen la relación entre variables y objetos muestrales. En el contexto de los datos químicos espectroscópicos multivariados, esta tarea implica, por ejemplo, utilizar las huellas espectrales para analizar similitudes y diferencias en un conjunto de muestras. En particular, en el caso de los datos de orden superior, el análisis de perfiles espectrales también permite la identificación de las especies químicas que originan las señales en una dada muestra y sus contribuciones relativas. Por otra parte, en el caso de los métodos predictivos, se busca modelar las propiedades químicas de un sistema, con el objetivo de predecir alguna propiedad de interés que sea difícil de medir, pero que se pueda inferir a partir de una medición instrumental indirecta, que resulte más conveniente por su accesibilidad, rapidez y/o costo. En particular, los métodos quimiométricos de clasificación y calibración multivariada son predictivos por excelencia. La clasificación se trata de un enfoque cualitativo, en el que se busca predecir si una dada muestra pertenece o no a un grupo o clase determinada. Por el contrario, la calibración es de índole cuantitativa y permite conocer cuánto de una propiedad de interés se encuentra presente en una dada muestra (en general, se desea conocer la concentración de una especie química de interés).

En el presente trabajo de tesis se realizaron desarrollos y aplicaciones quimiométricos que abarcan sus tres especialidades. Sin embargo, una buena parte de la investigación se dedicó al área de la calibración multivariada, dada su estrecha relación con la PAT del nivel 2 (monitoreo de procesos). En este sentido, resulta de interés definir una serie de conceptos referidos a la calibración que serán luego de uso habitual en el resto del texto.

En el contexto de la calibración analítica, se define como analito a aquella especie química o componente de una muestra que se desea medir o cuantificar. El conjunto de todos los componentes de una muestra, menos el analito, es lo que se conoce como matriz. Cuando todos los componentes de la matriz de una muestra son perfectamente conocidos, se dice que la matriz es químicamente definida, mientras que, en caso contrario, se trata de una matriz indefinida o compleja. Un componente de la matriz que puede generar una señal similar a la del analito y por lo tanto enmascarar su detección, es lo que habitualmente se conoce como interferente [27].

El proceso de obtención de un método de calibración suele abarcar las siguientes etapas: (i) etapa de calibración propiamente dicha, en la que se busca modelar matemáticamente la relación entre la/las variables predictoras (la señal instrumental) y la variable de interés (concentración del analito), mediante el uso de soluciones de

concentración conocida (muestras o patrones de calibración); (ii) etapa de validación, en la que se evalúa la capacidad predictiva del modelo enfrentándolo a un conjunto de muestras conocidas que no fueron utilizadas para calibrar el modelo. Los resultados de las predicciones permiten calificar cuantitativamente la *performance* analítica del modelo a través del cálculo de las denominadas cifras analíticas de mérito (AFOMs, del inglés *analytical figures of merit*) [28]; (iii) etapa de predicción, en la que el modelo obtenido y apropiadamente validado se utiliza para cuantificar muestras incógnita reales (de concentración del analito desconocida).

La calibración univariada se basa en el uso de datos de orden cero y sólo permite determinar una única especie. En este escenario resulta fundamental que la señal instrumental en la que se basa el modelo de calibración sea selectiva para el analito de interés. La Unión Internacional de Química Pura y Aplicada (IUPAC) define a la selectividad como "la extensión en la que un método puede utilizarse para determinar analitos particulares en mezclas o matrices sin interferencias de otros componentes con un comportamiento similar" [29]. Cuando esto no ocurre, es necesario proceder a la separación física del analito de la matriz para liberarlo de los interferentes, mediante el uso de exhaustivos procedimientos experimentales.

Por el contrario, el cambio revolucionario que establece la calibración multivariada consiste en la posibilidad de modelar y cuantificar más de una especie química en forma simultánea (varios analitos en presencia de uno o más interferentes), dada la naturaleza multivariada de los datos, que permite lidiar con los interferentes de forma matemática [20]. En particular, en calibración de primer orden, es posible modelar la señal del analito en presencia de los interferentes que son esperables en las muestras de predicción, incluyendo muestras de composición similar en la etapa de calibración. Asimismo, si una muestra de predicción en particular presenta una composición diferente (por la presencia de un interferente no modelado en la calibración), el modelo es capaz de etiquetarla como anómala, explotando una propiedad que se conoce como "ventaja analítica de primer orden" [20]. Esta ventaja permite, en caso de que se detecten muestras atípicas que impiden la cuantificación del analito, recalibrar el modelo, incluyendo nuevas muestras de calibración, evitando la separación física de los interferentes.

Por otra parte, el uso de datos de segundo orden para calibración reviste de algunas ventajas adicionales. En primer lugar, un aumento importante en la cantidad de información química que alimenta al modelo permite mejorar notablemente su rendimiento analítico, esencialmente, en términos de selectividad y sensibilidad. Asimismo, el segundo orden permite explotar la denominada "ventaja de segundo orden", que implica la posibilidad de cuantificar un analito en una muestra de predicción,

sin necesidad de modelar las interferencias durante la etapa de calibración. Esto reduce notablemente el número de muestras de calibración, respecto del primer orden. En este sentido, la ventaja de segundo orden es posible gracias a que el aumento en el orden de los datos, permite descomponer la señal del analito en una muestra de predicción, aún en presencia de interferentes no modelados en el calibrado, aumentando considerablemente la selectividad del modelo [20].

Finalmente, resulta fundamental mencionar que el desarrollo y aplicación de métodos analíticos basados en calibración multivariada ha tenido un impacto ambiental muy positivo. Desde sus inicios, las ciencias químicas han sido siempre grandes generadoras de residuos y la química analítica no es una excepción. En particular, muchas técnicas analíticas univariadas requieren del uso extensivo de reactivos y solventes tanto para las etapas de pretratamiento de la muestra como de generación de la señal, lo cual representa una problemática ambiental y económica a la hora de gestionar su descarte. En este contexto, la calibración multivariada representa un hecho revolucionario no solamente por sus ventajas analíticas, sino también porque presenta una serie ventajas operativas, tales como: disminución de los tiempos de análisis, reducción o eliminación total de técnicas de pretratamiento, reemplazo de métodos separativos por técnicas espectrales, entre otras. Estas ventajas permiten reducir drásticamente el uso de solventes orgánicos y, por lo tanto, han contribuido fuertemente a la revolución de la química verde [24].

3. Antecedentes del tema y caso de estudio

Tal como se puede evidenciar en la literatura especializada, en las últimas décadas, se han producido avances muy importantes en cuanto al desarrollo de métodos quimiométricos, que posibilitan el modelado exhaustivo de datos multidimensionales para las más diversas aplicaciones. En particular, en el caso del desarrollo y aplicación de estrategias de PAT para la industria biofarmacéutica, se evidencia un interés creciente debido a las bondades que ofrecen estas herramientas, de acuerdo a lo descripto en las secciones anteriores. A pesar de que se han reportado numerosos trabajos al respecto, esta área no ha sido completamente explorada ya que, naturalmente, una mayor sofisticación tecnológica en los procesos trae aparejados nuevos desafíos analíticos.

Desde el lanzamiento del documento sobre la PAT por parte de la FDA, numerosas han sido las publicaciones respecto a esta temática, como puede verse en los trabajos de revisión de Rathore y col. (2010) [8], Glassey y col. (2011) [6], Streefland y col. (2013) [30] y Mercier y col. (2014) [1]. En líneas generales, en el contexto de los

cultivos celulares, las aplicaciones de PAT han permitido: (i) definir nuevos criterios de clasificación entre lotes de productos; (ii) estudiar las interacciones entre variables para poder interpretar mejor la variabilidad observada en un proceso; y (iii) generar modelos predictivos para predecir variables de proceso de manera más rápida y económica. Según establecen Mercier y col. (2014), las contribuciones respecto del nivel 3 de PAT son menos frecuentes debido a la complejidad que reviste su desarrollo e implementación. En este sentido, la mayoría de las investigaciones publicadas se relacionan con el análisis multivariado para la detección de fallas y la comparabilidad de productos [31,32] (nivel 1) y el uso de técnicas espectroscópicas para el monitoreo en línea de CPPs (nivel 2) [33-36]. En particular, se puede ver que los enfoques de PAT para clasificación son menos frecuentes. Algunos ejemplos son los trabajos reportados por Kirdar y col. (2009) [37] y por Li y col. (2011) [38] para el control de calidad de materias primas. Más recientemente, en el artículo publicado por Jin y col. (2019) se propone una estrategia para el diagnóstico temprano del crecimiento y productividad celular en un proceso discontinuo [39]. En todos los casos se puede apreciar que las técnicas espectroscópicas son las de preferencia para el desarrollo de PAT.

Por otro lado, con respecto a las herramientas quimiométricas utilizadas en los trabajos previos, se destacan fuertemente el análisis de componentes principales (PCA) y la regresión en cuadrados mínimos parciales (PLS). Ambas técnicas constituyen herramientas poderosas y de gran versatilidad para el modelado de datos de primer orden, cuyos fundamentos serán expuestos en los capítulos siguientes. Por el contrario, el uso de algoritmos para datos de orden superior así como las estrategias de *machine learning* en el contexto de la PAT para bioprocesos se encuentran menos explorados. En particular, debido a la naturaleza químicamente compleja de las muestras provenientes de un bioproceso, la aplicación de algoritmos para datos de orden superior resulta una herramienta interesante no sólo porque permite analizar cuali- y cuantitativamente muestras cuya composición es indefinida, sino porque además, los modelos contemplan el tratamiento de diversos problemas analíticos tales como interferencias, efecto matriz, ruido de fondo, entre otros [40].

La exploración de las potencialidades y limitaciones de los métodos quimiométricos de primer y segundo orden aplicados a datos multidimensionales generados a partir de muestras de un bioproceso que resulten en una mejora de sus capacidades operativas para el monitoreo en el contexto de la PAT, es el interés del presente trabajo de tesis. Esto implica no sólo una motivación desde el punto de vista científico, sino que además, reviste un interesante potencial tecnológico para el sector productivo.

Este trabajo se ha realizado en colaboración con la empresa santafesina de base biotecnológica Zelltek SA, perteneciente al grupo AMEGA Biotech (http://www.amegabiotech.com/). La misma produce los principios activos para una variedad de biofármacos, mediante tecnología de ADN recombinante, utilizando principalmente células animales como plataforma celular. En particular, se ha seleccionado como caso de estudio la etapa de fermentación (*upstream*[5]) del proceso estándar para la producción del biosimilar etanercept, el cual se encuentra aprobado para su uso en humanos desde hace varios años. Este producto es una proteína recombinante de uso terapéutico, que se indica para el tratamiento de enfermedades autoinmunes, tales como la artritis reumatoidea. Estructuralmente consiste en una proteína de fusión, formada por el dominio extracelular soluble de unión a ligando del receptor 2 del factor de necrosis tumoral (TNF), unido a la región constante (Fc) de la inmunoglobulina G1 humana [41]. El etanercept se produce actualmente en una plataforma celular establecida, derivada de células de ovario de hámster chino (CHO).

En la empresa Zelltek, durante la etapa de *upstream*, el etanercept se produce a partir del cultivo de células CHO modificadas genéticamente, adaptadas al crecimiento en suspensión, en medio libre de suero fetal bovino. Las células se cultivan en biorreactores de 4,5 L (planta piloto) y 100-500 L (planta industrial) de volumen de trabajo, mediante el régimen de cultivo continuo con retención celular (perfusión). La duración promedio de una fermentación ronda alrededor de los 20-30 días, aunque puede variar dependiendo la escala de trabajo. En cada fermentación se registran y controlan de manera automática una serie de variables fisicoquímicas, tales como la temperatura, el pH, la velocidad de agitación/aireación, tasa de perfusión, entre otras, mediante el uso de sensores específicos. Además, diariamente se realiza la toma manual de un pequeño volumen de medio de cultivo del reactor para el registro de las variables bioquímicas y biológicas del proceso, que resultan críticas para el monitoreo de su *performance*. Tales variables incluyen, esencialmente, la densidad de células totales y viables, la concentración de los metabolitos glucosa y lactato y la concentración de etanercept.

Durante la realización de este trabajo se recolectaron las muestras diarias y los datos de las determinaciones de variables de proceso de diferentes lotes de fermentación de etanercept, en dos escalas de producción (planta piloto e industrial). A partir de dichas muestras, se generaron EEMs de fluorescencia como datos de segundo

[5] Del inglés, "corriente arriba": se refiere al conjunto de operaciones unitarias que abarcan la preparación del inóculo celular, el proceso fermentativo propiamente dicho (amplificación de las células), cosecha y almacenamiento del sobrenadante de cultivo. La etapa de *upstream* es la que antecede a la de purificación del producto, conocida como etapa de *downstream* ("corriente abajo").

orden. De esta manera, los desarrollos y aportes realizados en esta tesis se basaron en el modelado quimiométrico integral de todos los datos generados.

En este punto es importante aclarar que tanto las variables registradas de manera automática (operativas y fisicoquímicas) como las medidas a partir del muestreo diario (variables bioquímicas y biológicas) constituyen CPPs del proceso. En particular, dependiendo la bibliografía, a las variables bioquímicas y biológicas que en general son medidas por técnicas *off-line* también se las suele denominar como controles intraproceso o IPCs (por sus siglas en inglés) para distinguirlas de aquellas variables que se registran de manera automatizada. Asimismo, existe una diferencia conceptual entre ambos tipos de variables ya que, en general, las variables operativas y/o fisicoquímicas son ajustables, mientras que los IPCs son no ajustables, puesto que son altamente dependientes de la biología del sistema. No obstante, todas las variables mencionadas se consideran CPPs del proceso, ya que en mayor o menor medida, impactan sobre la calidad del cultivo celular y, por lo tanto, del producto final.

En relación al modelado quimiométrico efectuado en este trabajo, las variables fisicoquímicas mencionadas anteriormente no fueron tenidas en cuenta, ya que al estar controladas en torno a valores de *setpoint* no se pueden considerar como variables aleatorias en el sentido estadístico estricto. De esta manera y para simplificar la nomenclatura a lo largo del texto, se hará un uso indistinto de los términos CPP o "variables de proceso" para hacer referencia únicamente a las variables bioquímicas y biológicas de forma genérica.

4. Panorama general de la metodología empleada y de las contribuciones desarrolladas en esta tesis

El presente trabajo de tesis está enfocado, en particular, en el nivel de PAT 2 (monitoreo) y se basa en el modelado integral de datos de segundo orden de fluorescencia (EEMs), en conjunto con datos de variables de proceso o CPPs.

Debido a las características propias de las EEMs, el Capítulo 1 está destinado al preprocesamiento de estos datos, es decir, a la etapa que antecede al modelado quimiométrico, en la cual los datos se someten a correcciones o transformaciones computacionales para mejorar su calidad. Con frecuencia, las EEMs son afectadas por señales espurias de origen instrumental, debidas a fenómenos de dispersión de la luz. En particular, en esta parte del trabajo se analizaron en forma comparativa estrategias de corrección previamente reportadas en la bibliografía. Además, se desarrolló un algoritmo novedoso, junto con una interfaz gráfica de usuario (GUI) para la implementación integrada de alguna de las metodologías descriptas.

En el Capítulo 2 se describe detalladamente el proceso de fermentación de etanercept y la generación de los datos. Asimismo, se planteó un análisis exploratorio para la caracterización del proceso respecto de los datos de CPPs y de fluorescencia generados. Para ello se utilizaron los algoritmos exploratorios PCA y resolución multivariada de curvas mediante cuadrados mínimos alternantes (MCR-ALS). Además, se desarrolló una estrategia de PAT cualitativa para el monitoreo *at-line* de la viabilidad celular, basada en el uso de PLS con análisis discriminante (PLS-DA) como algoritmo de clasificación.

En el Capítulo 3 se muestra el desarrollo de una estrategia de PAT cuantitativa para el monitoreo *at-line* de la concentración de etanercept. Para ello se utilizó, en primer lugar, la regresión PLS como algoritmo típico de calibración de primer orden. Este modelo permitió evidenciar una relación no lineal entre la señal de fluorescencia y la concentración de la proteína recombinante. En virtud de ello, se desarrolló una metodología de calibración basada en el uso de un tipo específico de red neuronal artificial (ANN). La optimización de este modelo de calibración se llevó a cabo mediante herramientas de diseño y optimización de experimentos.

Finalmente, el Capítulo 4 está destinado al estudio de la sensibilidad y la sensibilidad analítica de modelos de calibración de primer orden basados en ANNs. En este sentido, se desarrollaron y validaron ecuaciones para el cálculo de la sensibilidad y la sensibilidad analítica en un tipo particular de ANN, cuya estimación permitió caracterizar la sensibilidad del método desarrollado en el Capítulo 3 y compararlo con la de la técnica cromatográfica univariada de referencia.

En todas sus etapas, la complejidad que presentó el sistema en estudio desde el punto de vista analítico motivó no solamente la aplicación de métodos quimiométricos conocidos, sino que también se lograron desarrollar elementos teóricos y metodologías quimiométricas novedosas para dar respuesta a los interrogantes planteados. Tal escenario es el que da sentido al título de este trabajo.

Objetivos

Objetivo general

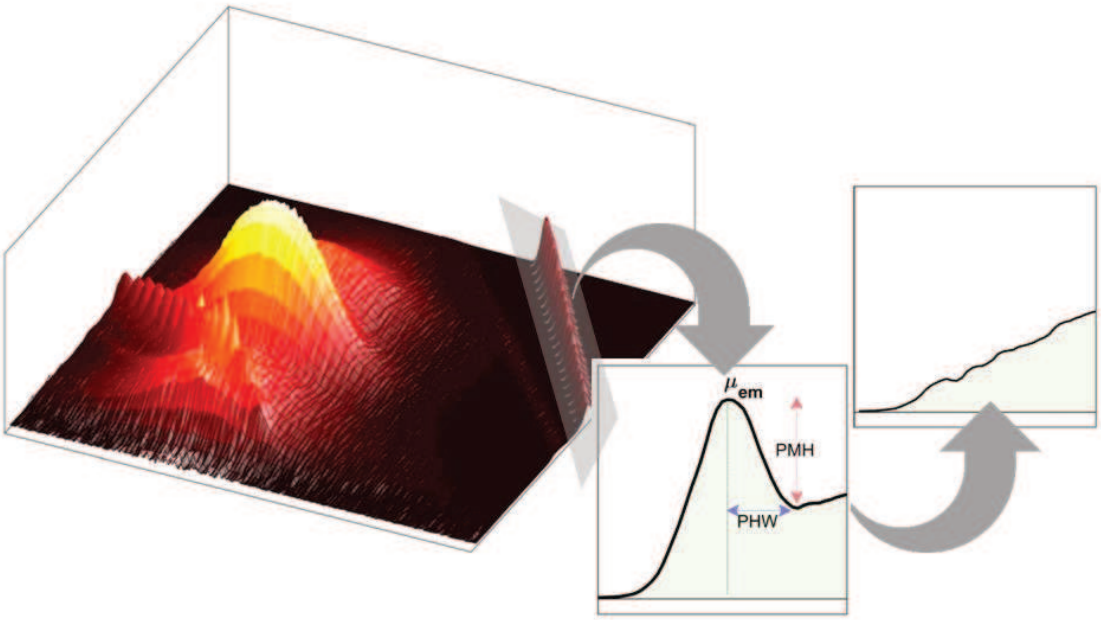
Desarrollar estrategias analíticas alternativas, basadas en la aplicación de herramientas quimiométricas, que permitan el monitoreo de variables críticas de un bioproceso estandarizado, en el contexto de la PAT.

Objetivos específicos

1. Generar conjuntos de datos analíticos de segundo orden a partir de muestras de un bioproceso estándar en su etapa de fermentación mediante espectroscopía de fluorescencia.
2. Mejorar la calidad de los datos mediante el desarrollo y aplicación de algoritmos de preprocesamiento de la señal analítica de fluorescencia.
3. Desarrollar métodos analíticos de monitoreo en línea de variables críticas, en el contexto de la PAT, a través del modelado conjunto de la información espectral y las variables de proceso utilizando diferentes herramientas de análisis multivariado de datos.
4. Validar las metodologías propuestas mediante el cálculo de cifras analíticas de mérito.
5. Implementar algoritmos de análisis multivariado de datos mediante el desarrollo y optimización de herramientas computacionales.

CAPÍTULO 1

Preprocesamiento de datos multivariados de fluorescencia



1.1. Introducción

Una vez finalizada la etapa de adquisición de datos y previamente al procesamiento quimiométrico, es frecuente que estos requieran de una etapa de preprocesamiento. Esta se refiere al conjunto de correcciones y/o transformaciones matemáticas/computacionales que se realiza sobre los datos crudos[6] para mejorar su calidad, lo cual puede tener diferentes implicancias respecto de la etapa de procesamiento posterior. Se debe tener en cuenta que los datos que son originados por fuentes instrumentales presentan no solamente variaciones aleatorias en las lecturas que ocasionan el denominado "ruido", sino que también son susceptibles a sufrir efectos sistemáticos, deformaciones y/o incluir artificios que también son de origen instrumental. De esta manera, algunos de los principales objetivos que se persiguen en la etapa de preprocesamiento de los datos son [42]:

- disminuir la variabilidad de los datos y, por lo tanto, mejorar la interpretabilidad y la capacidad predictiva del método quimiométrico que será luego utilizado;
- mejorar la relación señal-ruido, cuando la señal debida a las especies químicas de interés se encuentra desfavorecida;
- corregir los datos para que se ajusten a los supuestos que asume un método quimiométrico en particular;
- eliminar efectos instrumentales sistemáticos, tales como líneas de base o corrimientos espectrales.

La decisión de implementar o no un preprocesamiento se relaciona directamente con el objetivo quimiométrico que se busca con los datos. Asimismo, la selección del o los métodos de preprocesamiento depende fuertemente de la calidad o complejidad de los datos crudos y del origen instrumental de estos. En cualquier caso, se debe tener en cuenta que tanto un dato crudo como un dato mal preprocesado pueden llevar a resultados erróneos o a una mala interpretación de los resultados que se obtienen luego de la implementación de un método quimiométrico, razón por la cual, el preprocesamiento puede resultar una etapa crítica. Si bien se han desarrollado numerosas herramientas computacionales que facilitan y automatizan esta tarea, es fundamental que la misma esté orientada por el buen criterio y conocimiento del sistema en estudio por parte del analista.

Los desarrollos quimiométricos realizados en esta tesis están basados en el uso de EEMs. Por lo tanto, la primera etapa del trabajo se orientó hacia el preprocesamiento de este tipo de datos de segundo orden.

[6] Datos que se obtienen directamente del instrumento analítico.

La generación de una EEM se realiza en un instrumento conocido como espectrofluorómetro o fluorímetro y consiste en adquirir una señal de fluorescencia en un rango específico de longitudes de onda de emisión, excitando la muestra a varias longitudes de onda de excitación, en un rango espectral particular [43]. En combinación con la quimiometría, las EEMs son herramientas valiosas para la resolución de sistemas multicomponentes. Sin embargo, las EEMs presentan señales espurias particulares que deben ser removidas o corregidas durante la etapa de preprocesamiento. Las mismas son artificios de origen instrumental debidas a las denominadas dispersiones de Rayleigh (Ry) y de Raman (Rm). En conjunto, estos efectos suelen denominarse de manera genérica como señales de dispersión o *scattering* [44,45].

La señal de Ry es una dispersión de luz elástica que es producida por sustancias presentes en el seno de la solución cuyo radio es mucho menor a la longitud de onda de la luz incidente. Estas partículas oscilan a la misma frecuencia y, por lo tanto, emiten una señal de idéntica longitud de onda. Esto se conoce, habitualmente, como dispersión de Ry de primer orden (Ry1). La señal de Ry1 aparece en el espectro de emisión, a la longitud de onda que coincide con la de excitación a la que se adquirió dicho espectro. Por otro lado, la señal de Raman (Rm) es una forma de dispersión inelástica y dependiente del solvente, que se origina por las sustancias del seno de la solución que causan una pérdida constante de energía de la luz incidente. Asimismo, señales de dispersión de órdenes mayores se pueden observar producto de la difracción de la luz en los monocromadores del instrumento [46]. Debido a que estas señales se originan por la interacción entre las moléculas de la solución y la luz incidente, las mismas no contienen información acerca de las características químicas del sistema en estudio. Ejemplos gráficos de espectros de emisión de fluorescencia y de una EEM en presencia de señales de dispersión de tipo Ry y Rm se muestran en la Figura 1.1.

Por otra parte, es importante mencionar que, por definición, una EEM tiene la particularidad de cumplir con un supuesto matemático fundamental para los modelos quimiométricos de segundo orden y que es el de ser una matriz bilineal de rango pequeño (*low-rank bilinearity*). Según Wilson y Kowalski (1989) [47], esta propiedad puede ser concebida según dos criterios matemáticos, considerando la matriz generada para una sustancia pura libre de ruido. En este sentido, una matriz es bilineal de rango pequeño si, por un lado, su rango es igual a 1 y, por otro lado, si la información comprendida en los dos modos instrumentales (excitación y emisión) puede ser explicada por dos perfiles vectoriales individuales e independientes[7].

[7] Matemáticamente, el rango de una matriz se define como el número de filas (o columnas) linealmente independientes. Por otra parte, en quimiometría suele hacerse una distinción entre el rango matemático de un dato matricial y su rango químico. Este último se refiere al número de componentes independientes que aportan a la señal analítica. Idealmente, para un sistema químico de A componentes independientes en

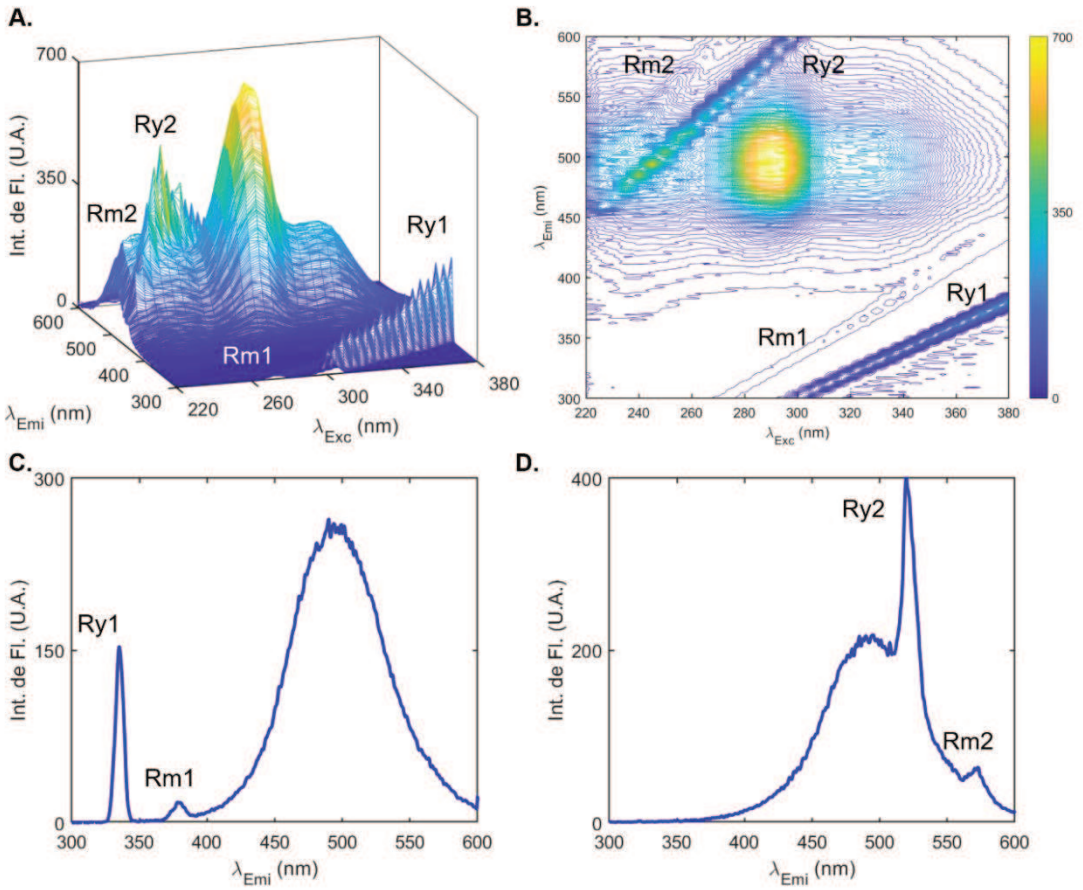


Figura 1.1. Datos de fluorescencia en presencia de señales de dispersión de tipo Rayleigh (Ry) y Raman (Rm). **A-B.** EEM de ofloxacina (OFL) en su representación tridimensional (izquierda) y como diagrama de contorno (derecha), en la que se muestran los patrones diagonales que forman las señales de dispersión Ry de primer (Ry1) y segundo orden (Ry2) y Rm de primer (Rm1) y segundo orden (Rm2). **C.** Espectro de emisión de OFL adquirido a una excitación de 335 nm en presencia de las señales de *scattering* de primer orden. Se observa que la posición media en nm de la señal de Ry1 coincide con la longitud de onda de excitación del espectro. **D.** Espectro de emisión de OFL adquirido a una excitación de 260 nm en presencia de las señales de *scattering* de segundo orden. Se observa que la posición media en nm de la señal de Ry2 coincide aproximadamente con el doble de la longitud de onda de excitación del espectro. En todos los casos, las señales de Rm aparecen a longitudes de onda mayores que la de Ry, debido a la pérdida constante de energía propia de este fenómeno.

Siguiendo el mismo criterio, en ausencia de fenómenos de interacción entre fluoróforos tales como efecto de filtro interno o apagamiento (*quenching*), un dato de tres vías generado por apilamiento de EEMs cumple con el supuesto de trilinealidad de rango pequeño. Ambos supuestos son de vital importancia cuando se utilizan para el modelado quimiométrico, algoritmos basados en la descomposición bi- o trilineal de datos, tales como MCR-ALS y PARAFAC, respectivamente [21]. Sin embargo, en

todos los modos instrumentales, los rangos químico y matemático son iguales a A. Este concepto se aplica de manera directa para datos libres de ruido. Sin embargo, en la práctica, los datos instrumentales poseen una incertidumbre asociada (ruido instrumental). En este contexto, se prefiere utilizar el término "pseudo rango" para caracterizar al número de componentes independientes que aportan de manera significativa a la señal analítica. En términos estadísticos, el pseudo rango está asociado al número de componentes que capturan un porcentaje significativo de la variabilidad del dato.

presencia de señales de Ry o Rm, las mismas aparecen en las EEMs formando patrones diagonales que rompen con los supuestos de bi- y trilinealidad de rango pequeño [48]. En el resto de este trabajo, las propiedades bilinealidad y trilinealidad de rango pequeño serán denominadas simplemente como bilinealidad y trilinealidad, para simplificar la nomenclatura.

Hasta la actualidad, la mayoría de los algoritmos de segundo orden clásicos son incapaces de modelar satisfactoriamente datos de dos y tres vías en donde estos supuestos no se cumplen. Por esta razón, es que las señales de *scattering* deben ser cuidadosamente removidas o corregidas, sin ocasionar pérdida de información analítica o distorsionar las señales propias de los analitos de interés.

En la literatura es posible hallar un número importante de estrategias que han sido desarrolladas para mitigar los fenómenos de dispersión en las señales de fluorescencia. No obstante, no existe un acuerdo en la comunidad científica sobre cuál es la mejor estrategia para una situación determinada [49]. En este sentido, si las señales de dispersión no se solapan con las señales de los analitos, entonces la región donde se observa el *scattering* puede ser omitida o removida de manera directa del dato [50-52]. Por el contrario, cuando las señales de Ry o Rm se solapan total o parcialmente con las señales de interés, tales señales de dispersión deben ser adecuadamente tratadas para evitar sesgos en la resolución quimiométrica e interpretaciones erróneas de la información espectral.

Entre los procedimientos disponibles para este objetivo, las metodologías más frecuentemente reportadas en bibliografía se basan en correcciones digitales de los datos, mediante algoritmos matemáticos/computacionales específicos, muchos de los cuales requieren de habilidades matemáticas y de programación [49,53]. Este último aspecto no es menor, en el sentido de que puede resultar una limitación para el uso de las herramientas por parte de usuarios no expertos o de áreas no quimiométricas. En este sentido, una manera conveniente para facilitar el acceso y la implementación de las metodologías de corrección de manera amigable consiste en el uso de interfaces gráficas de usuario (GUI, por sus siglas en inglés).

1.2. Objetivos específicos del capítulo

En este capítulo se plantearon los siguientes objetivos específicos:

- realizar un relevamiento, descripción y comparación mediante el tratamiento de datos generados a partir de sistemas modelo, de las metodologías de corrección digital de señales de dispersión en datos de fluorescencia más frecuentemente citadas en la literatura;

- proponer una estrategia novedosa de corrección que mejore y optimice algunas de las cualidades de las estrategias analizadas en el objetivo anterior y proceder a la evaluación de su *performance*, en comparación con los métodos previamente reportados;
- desarrollar una GUI de acceso libre para la implementación integrada y amigable de algunas de las metodologías de corrección de *scattering* estudiadas.

1.3. Materiales y métodos

A continuación, se describe la metodología empleada para la generación de EEMs a partir de sistemas de fluorescencia modelo para el estudio sistemático y comparativo de estrategias de corrección de señales de dispersión.

1.3.1. Reactivos y soluciones

La ofloxacina (OFL) fue provista por Sigma Aldrich (Steinheim, Alemania) y la resorrufina (RES) fue gentilmente provista por el Laboratorio de Dispositivos Moleculares (INQUIMAE-CONICET, Buenos Aires, Argentina). El acetato de sodio trihidrato p.a. fue adquirido de Anedra (La Plata, Argentina), el ácido acético glacial y el hidróxido de sodio p.a. fueron adquiridos de Cicarelli (San Lorenzo, Argentina). Todas las disoluciones de base acuosa fueron preparadas con agua ultrapura obtenida mediante ósmosis inversa (resistividad 18,2 MΩ.cm) y filtración (0,22 μm), empleando un equipo Milli Q de Millipore (Bedford, EEUU).

Se preparó un buffer acético/acetato 0,05 mol L^{-1} pesando y disolviendo una cantidad apropiada de acetato de sodio en agua ultrapura, ajustando el pH a 5,9 con ácido acético glacial y completando el volumen a 100 mL con agua ultrapura.

Se prepararon soluciones madre de patrones de cada analito disolviendo una masa apropiada de cada droga en buffer acetato pH 5,0 para OFL y agua ultrapura alcalinizada (pH>10) para la RES. Las soluciones de trabajo se prepararon transfiriendo una alícuota adecuada de la solución madre a frascos volumétricos de 5,0 mL, completando el volumen con buffer acetato o agua ultrapura, según cada caso.

1.3.2. Instrumento

Todos los experimentos fueron realizados en un espectrómetro de fluorescencia LS-55 (Perkin Elmer, Waltham, Massachussetts, EEUU), equipado con el software FL WinLab (Perkin Elmer, Waltham, Massachussetts, EEUU) para el control del instrumento

y la adquisición de los datos. Todas las mediciones se realizaron a temperatura ambiente en una cubeta de cuarzo de 1 cm de paso óptico.

Las mediciones de pH se realizaron con un potenciómetro 410A de Orion (Massachusetts, EEUU), equipado con un electrodo combinado de vidrio Boeco BA17 (Hamburgo, Alemania).

1.3.3. Generación de los datos

Sistema A – OFL

Las EEMs fueron registradas en el rango espectral de emisión de 300,0 a 600,5 nm, con un intervalo de 0,5 nm, variando las excitaciones en el rango de 200,0 a 400,0 nm, cada 5,0 nm, y utilizando una velocidad de barrido de 800 nm min^{-1}. Los anchos de rendijas (*slit*) del monocromador de excitación y de emisión se fijaron ambos en 5 nm y se utilizó un voltaje del tubo fotomultiplicador (PMT) de 750 V. De esta manera, se registraron matrices de 41×602 puntos para los modos de excitación y emisión, respectivamente.

Sistema B – RES

Las EEMs fueron registradas en el rango espectral de emisión de 400,0 a 700,5 nm, con un intervalo de 0,5 nm, variando las excitaciones en el rango de 350,0 a 600,0 nm, cada 5,0 nm, y utilizando una velocidad de barrido de 1000 nm min^{-1}. Los anchos de rendijas (*slit*) del monocromador de excitación y de emisión se fijaron ambos en 5 nm y se utilizó un voltaje del tubo fotomultiplicador (PMT) de 900 V. De esta manera, se registraron matrices de 51×602 puntos para los modos de excitación y emisión, respectivamente.

1.3.4. Software

Todas las matrices de datos fueron utilizadas tal cual fueron obtenidas del instrumento, es decir, no se implementó ninguna estrategia de preprocesamiento adicional. El tratamiento de los datos para la corrección de señales de dispersión se realizó con el software MATLAB R2017b, mediante el uso de rutinas específicas. La rutina *Eemscat* se encuentra disponible on-line en el sitio http://www.models.life.ku.dk/EEM_correction [45]. Las rutinas para las estrategias *Cleanscan* [54] y aquella basada en ajuste gaussiano [55] fueron gentilmente provistas por los respectivos autores.

1.4. Resultados y discusión

Los sistemas fluorescentes modelo A y B fueron seleccionados para esta parte del trabajo considerando sus diferencias en los corrimientos de Stokes y el rango de emisión espectral: el sistema A presenta un corrimiento de Stokes particularmente importante en la región espectral UV, mientras que el sistema B muestra un corrimiento muy pequeño en la región espectral visible. Una vez generados los datos, se procedió a realizar un estudio sistemático y comparativo de seis metodologías de corrección diferentes.

1.4.1. Estado del arte de las metodologías de corrección de scattering en EEMs basadas en algoritmos de corrección digital mayormente citadas en la bibliografía

Metodología 1 (M1): restado del blanco

La estrategia más simple y que ha sido tradicionalmente utilizada para la corrección de señales de dispersión consiste en la sustracción de la señal obtenida a partir de una señal blanco apropiada [50,51]. Esencialmente, una muestra blanco tiene una composición idéntica a la de la muestra, pero no contiene los compuestos fluoróforos de interés. Así, el blanco puede revelar la presencia de señales de Ry y/o Rm, además de la presencia de impurezas. Para que el blanco resulte representativo de la muestra de trabajo y la corrección resulte adecuada, la señal del blanco debe adquirirse exactamente en las mismas condiciones experimentales e instrumentales.

Una desventaja importante de esta metodología es que la señal del blanco no siempre se encuentra disponible. Esto puede representar una limitación importante para el uso de M1 especialmente en el caso de sistemas químicamente complejos en donde, en ocasiones, no es posible obtener una muestra libre de los analitos fluoróforos de interés. Por otro lado, debe tenerse en cuenta la posición e intensidad de las señales de Ry y Rm en la EEM. En particular, para el caso del Ry, en el caso en que las señales de dispersión se solapan con la señal de interés, el restado del blanco puede no resultar apropiado debido a que se puede perder información debido a las diferencias de intensidad de las señales de Ry entre la muestra y el blanco. Sin embargo, esta estrategia resulta muy útil para la eliminación de la señal de Rm, ya que su posición e intensidad depende exclusivamente de la composición del medio.

En síntesis, si bien M1 es una estrategia de muy fácil implementación y que no requiere habilidades matemáticas/computacionales, su aplicación puede no resultar trivial, debido al hecho de que se requiere información previa sobre el sistema, la cual no siempre se encuentra disponible.

Metodología 2 (M2): inserción de valores faltantes (*missing values*)

Desde el punto de vista químico, es sabido que en fluorescencia no es esperable ninguna señal de emisión a longitudes de onda que se encuentren por debajo de la longitud de onda de excitación, ya que esto representaría una emisión de mayor energía que la que la provoca. Por lo tanto, la emisión registrada por debajo de las longitudes de onda de excitación es esencialmente cero. En este sentido, cuando se registra una EEM, se obtiene una región triangular sin información química por debajo del patrón diagonal del Ry1.

Basada en este hecho, una estrategia común para eliminar las señales de Ry y Rm consiste en obviar las regiones en las que se espera señal cero, ya sea recortando las matrices o bien, ajustando los valores de señal a cero. Asimismo, si bien este no es el caso de las emisiones más allá del *scattering* de segundo orden, el mismo procedimiento suele emplearse para la señal registrada por encima del Ry2. La primera desventaja que se puede observar es de tipo matemática, ya que la inserción de grandes regiones con ceros rompe la estructura de bilinealidad de la EEM. Para evitar este inconveniente, la estrategia implementada suele ser la inserción de valores faltantes[8]. Esta estrategia ha sido ampliamente utilizada cuando se utiliza PARAFAC como método quimiométrico. Sin embargo, se ha demostrado que cuando se utilizan datos con gran cantidad de valores faltantes, se pueden experimentar problemas de convergencia de este algoritmo [56-58].

Finalmente, esta estrategia resulta útil en sistemas con un corrimiento de Stokes grande, de manera que las señales de dispersión no se solapan con la señal de interés ya que, de otra manera, el reemplazo por valores faltantes en zonas de gran solapamiento, ocasiona necesariamente pérdida de información espectral. Asimismo, es importante destacar que no todos los métodos quimiométricos admiten valores faltantes en los datos, lo cual representa otra limitación de esta metodología.

Metodología 3 (M3): *cleanscan* o método de interpolación bidimensional

Uno de los primeros algoritmos de corrección específicos desarrollados para remover las señales de Ry y Rm fue el publicado por Zepp y col. en 2004 [54]. Esta metodología está basada en la interpolación bidimensional, mediante el método de Triangulación de Delaunay. La misma requiere de una serie de parámetros que deben ser fijados por el usuario.

El fundamento de esta metodología consiste en remover las señales de *scattering* y luego realizar una interpolación para rellenar la/s regiones removidas. Para la

[8] En la jerga estadística y computacional, *missing-values.* También, NaN en notación de MATLAB.

identificación de cada señal de dispersión, se proponen criterios matemáticos. Así, la posición en nm de la señal de Ry sobre cada espectro de emisión en la EEM sigue una relación lineal con la longitud de onda de excitación (λ_{ex}), es decir, $\mu_{\mathrm{em}} = n\,\lambda_{\mathrm{ex},i}$, donde μ_{em} representa la posición media esperada del pico de Ry en un espectro de emisión, adquirido a la i-ésima longitud de excitación $\lambda_{\mathrm{ex},i}$, y n es el orden del *scattering.* Por otra parte, la posición media para la señal de Rm se modela mediante una función polinomial, cuyos parámetros deben ser ajustados por el usuario para la identificación de las señales. Además, el usuario debe fijar parámetros de tolerancia (tol), que representan la mitad del ancho de pico en su base, para así proceder a identificar la zona o ventana del espectro en la que se espera caiga la señal de dispersión completa. Así, en cada espectro de la EEM, la posición de cada señal de dispersión completa es identificada en nm mediante el cálculo $\mu_{\mathrm{em}} \pm tol$. Una vez identificadas las señales de dispersión, las mismas son reemplazadas por valores faltantes para proceder luego a la interpolación bidimensional, en la que se tiene en cuenta no sólo la información del espectro sobre el cual se está efectuando la interpolación, sino, además, puntos localizados en espectros vecinos.

Como puede evidenciarse, esta estrategia está basada sólo en principios matemáticos y puede no resultar accesible para algunos tipos de usuarios. Asimismo, la selección de los parámetros por parte del operador no resulta intuitiva y debe realizarse de manera cuidadosa para que la corrección sea satisfactoria. No obstante, la eficiencia de esta metodología ha sido probada en varias aplicaciones, en la que se observa que se trata de una herramienta versátil para diferentes tipos de datos y que posee la habilidad de corregir simultáneamente una gran cantidad de matrices, utilizando el mismo conjunto de parámetros [59-61].

Metodología 4 (M4): *eemscat* o método de interpolación unidimensional

Esta metodología fue propuesta por Bahram y col. en 2006 [45] y ha sido ampliamente utilizada desde su publicación. La misma se basa en la interpolación unidimensional realizada sobre los espectros de emisión individuales de una EEM. Se trata de un tipo de interpolación simple que evita, en principio, el sobreajuste de los datos. En este sentido, los autores mencionan que el método de interpolación basado en una función propia de MATLAB, denominada *shape-preserving piece wise cubic polynomial interpolation* permite obtener los mejores resultados, aunque la rutina admite la libre elección del método de interpolación por parte del usuario.

En esta metodología, la posición del Ry se define de la misma manera que la descripta para M3. Sin embargo, en el caso del Rm, el criterio es diferente al propuesto por Zepp y se basa en una fórmula empírica de fundamento físico, en la que la posición

media de la señal de Rm μ_{em} en el espectro de emisión adquirido a la i-ésima longitud de excitación $\lambda_{\mathrm{ex},i}$ está dada por la expresión:

$$\mu_{em} = \frac{10^7 \lambda_{ex,i}}{10^7 - wnr \times \lambda_{ex,i}} \tag{1.1}$$

donde wnr es un parámetro que establece la diferencia en número de onda entre la posición del Ry y del Rm en la EEM. Por ejemplo, cuando se trabaja en medio acuoso, la señal de Rm aparece a una frecuencia 3600 cm^{-1} menor que la frecuencia a la que aparece la señal de Ry [44,46]. Así, el parámetro wnr puede ser modificado en función de la naturaleza del solvente de la muestra.

Una vez más, el usuario debe proporcionar valores de tol para cada tipo de *scattering*. Una vez que las señales de dispersión son identificadas, las mismas se reemplazan por valores faltantes y luego se sigue con la interpolación unidimensional sobre cada espectro individual.

Metodología 5 (M5): método basado en ajuste gaussiano

En 2014, Eilers y Kroonenberg presentaron una estrategia interesante para la corrección de *scattering* basado en el ajuste de curvas, en lugar de la interpolación [55]. La premisa que fundamenta este método consiste en el hecho de que frecuentemente las señales de Ry y Rm pueden ser modeladas por curvas de tipo gaussianas. Por lo tanto, estas señales pueden ser modeladas como tales mediante regresión no lineal y restadas de los espectros de emisión. Si los picos de Ry y Rm no pueden ser modelados como curvas gaussianas, entonces este método no resulta satisfactorio.

Para la corrección, el usuario debe ajustar una serie de parámetros tanto para la identificación como para el ajuste gaussiano de las señales de dispersión. Si bien esta metodología resulta, en principio, de gran interés en el sentido de que preserva los atributos originales de las señales espectrales, el ajuste de los parámetros resulta tedioso y poco intuitivo. Por lo tanto, la implementación de esta rutina requiere de un dominio avanzado de programación y de conceptos estadísticos, para la comprensión de los códigos computacionales. Estos aspectos representan una limitación importante para la implementación de esta metodología. Por las razones expuestas no fue posible su implementación para los datos generados en este trabajo y, por lo tanto, no se pudieron corregir las EEMs generadas mediante esta estrategia, razón por la cual no se incluirán los resultados correspondientes. Sin embargo, la filosofía de no interpolación que propone este método ha motivado el desarrollo de un algoritmo novedoso, que se describe en la siguiente sección.

1.4.2. Scscat: una estrategia novedosa basada en el principio de conservación de la señal

En base al estudio sistemático de las metodologías presentadas y en virtud de las ventajas y limitaciones identificadas en cada caso, en la presente tesis se desarrolló una estrategia novedosa de corrección de *scattering* que optimiza y/o mejora algunas de las características de las estrategias descriptas anteriormente.

La metodología aquí desarrollada, referida como M6, ha sido denominada *scscat*, en donde las letras '*sc*' refieren a la idea de conservación de la señal (*signal-conservative*). El *core* del algoritmo consiste en la corrección mediante el principio propuesto por Eilers y Kroonenberg [55], en combinación con algunas ideas implementadas en el algoritmo de Bahram y col. [45]. Así, el algoritmo se estructura en tres etapas: identificación, modelado y sustracción de la señal de *scattering*. Esta estrategia propone la corrección de cada tipo de señal de dispersión de manera independiente, y sobre los espectros de emisión individuales de una EEM.

Para iniciar el algoritmo, se requieren de cinco entradas, tres de las cuales corresponden al dato experimental, es decir la matriz cruda $(\mathbf{X}_r)$, y los vectores de longitudes de onda de excitación y emisión. Los restantes dos parámetros son valores de tolerancias que debe establecer el usuario (únicos parámetros ajustables del algoritmo). Dichas tolerancias representan los puntos de control del método para permitir una identificación y modelado apropiados del pico de *scattering.* El primer parámetro de tolerancia $tol1$ denota el valor esperado de la mitad del ancho del pico en su base (PHW), el cual controla el tipo de corrección a emplear. El segundo parámetro de tolerancia $tol2$ representa un umbral de intensidad, por debajo del cual, no se realiza ninguna corrección. De esta manera, sólo son corregidas aquellas señales que presentan una altura máxima de pico (PMH) mayor que $tol2$.

Una vez ajustados los parámetros, se procede a la identificación de las posiciones medias μ_{em} de las señales de dispersión, utilizando para ello, los mismos criterios físicos establecidos en el método M4.

Posteriormente, se calcula el valor de PMH como la diferencia entre la intensidad máxima y mínima que se obtiene en la ventana $\mu_{em} \pm tol1$. Dicho valor se compara con el parámetro $tol2$. Este paso resulta crítico en aquellas regiones en donde la señal de *scattering* se encuentra totalmente embebida por la señal del analito y no se necesita ninguna corrección. Contrariamente, en caso de que se evidencie que $\mathrm{PMH} > tol2$, se confirma la presencia de *scattering* y el pico es aislado para proceder a su modelado y corrección.

Para efectuar la corrección, la región espectral que contiene a la señal de *scattering* es temporalmente escindida del espectro. Luego, para estimar la ventana

espectral sobre la cual se realizará la corrección, es decir el ancho del pico, se utiliza el parámetro $tol1$. Posteriormente, se utiliza un segundo punto de control que define el tipo de procedimiento para la corrección del pico. En este sentido es importante mencionar que $tol1$ resulta sencillo de conocer a partir de los parámetros instrumentales utilizados en la adquisición del dato. Idealmente, las mitades izquierda y derecha a cada lado de la posición media del pico resultan idénticas si este se comporta como una curva gaussiana. Sin embargo, en la práctica es frecuente que las señales de dispersión sean asimétricas o incluso, incompletas. Las señales incompletas se observan en los extremos de las EEMs y, en consecuencia, estas señales no pueden ser consideradas como curvas gaussianas. Para distinguir entre picos completos e incompletos, en el algoritmo se propone el siguiente criterio: un pico es considerado como completo cuando la mitad del ancho de pico tanto a la izquierda como a la derecha de su posición media resulta igual o mayor al 80% de $tol1$. Si este criterio no es alcanzado, el pico se considera incompleto y la corrección se realiza de una manera similar a la propuesta en la rutina de M4, es decir, mediante interpolación unidimensional. Pero si el criterio se cumple, entonces, se procede a la corrección mediante ajuste gaussiano.

La corrección basada en ajuste gaussiano procede mediante las siguientes etapas (Figura 1.2):

i. se estima una línea de base de la señal de dispersión cruda, mediante el método de suavizado basado en cuadrados mínimos asimétricos de Eilers [62] (Figura 1.2.1 y 1.2.2);
ii. se resta la línea de base a la señal de dispersión (Figura 1.2.3);
iii. los puntos que definen la señal obtenida en ii. se someten a ajuste no lineal con un modelo gaussiano, en el que se utilizan los valores de PMH y PHW como estimadores iniciales;
iv. la curva gaussiana de ajuste se resta a la señal obtenida en ii. para obtener una nueva región (Figura 1.2.4). Esta contiene los residuos de la regresión gaussiana y el ruido instrumental inherente al espectro. Es importante destacar en este punto que, si el ajuste no lineal no es bueno, se introducen artefactos indeseables a la señal espectral (esto ocurre cuando la señal de *scattering* no sigue un comportamiento gaussiano);
v. finalmente, para recuperar la orientación original del espectro, la línea de base calculada en el punto i. es sumada a la región obtenida en la etapa iv. (Figura 1.2.5) y el resultado es reincorporado al espectro completo de la EEM. De esta manera, se obtiene un espectro de emisión libre de señal de dispersión.

Es importante destacar que la secuencia de funciones y parámetros necesarios para llevar a cabo los pasos de estimación de línea de base y ajuste no lineal han sido optimizados y estandarizados con el objetivo de facilitar la implementación de la rutina. El paquete de *scripts* para la ejecución del método en MATLAB se encuentra disponible para su descarga en la página web del Laboratorio de Desarrollo Analítico y Quimiometría (LADAQ) (https://fbcb.web1.unl.edu.ar/laboratorios/ladaq/download/).

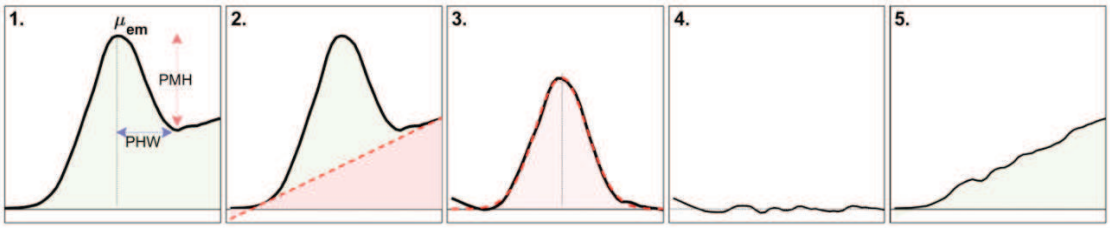


Figura 1.2. Etapas de corrección de una señal de dispersión mediante ajuste gaussiano. **1.** Región escindida del espectro de emisión original que contiene la señal de Ry solapada con la señal del analito de interés. **2.** Señal original (línea blanca) y línea de base estimada (línea roja discontinua). **3.** Señal espectral luego del restado de la línea de base (línea negra) y curva gaussiana ajustada (línea roja discontinua). **4.** Señal resultante luego de la sustracción de la curva gaussiana de ajuste a la señal de dispersión sin línea de base. **5.** Señal obtenida luego de la adición de la línea de base a la señal obtenida luego de la sustracción de la curva gaussiana ajustada a la señal de dispersión. Los colores verde y rojo representan, respectivamente, las áreas remanentes y sustraídas en cada etapa.

1.4.3. Análisis comparativo de las estrategias estudiadas a partir de la corrección de datos generados con dos sistemas modelo

Los dos sistemas modelo considerados se muestran en la Figura 1.3. Como puede verse, en el caso del sistema A (OFL), las señales de dispersión de primer orden aparecen lejos de la señal del analito, mientras que las de segundo orden, están totalmente solapadas. Por el contrario, para el sistema B (RES), se observan señales de primer orden totalmente solapadas con la señal del analito y no se observa *scattering* de segundo orden en la región espectral considerada para este sistema.

Con el objetivo de comparar el desempeño en términos espectrales cualitativos de las seis metodologías descriptas para la corrección de *scattering*, ambos sistemas fueron sometidos a la corrección con cada una de las estrategias. Posteriormente, se seleccionó de cada sistema, un espectro en una región de la EEM libre de dispersión de Ry y Rm como referencia para efectuar comparaciones con los espectros corregidos. Este análisis comparativo se realizó, para cada sistema, en tres regiones de las EEMs diferentes: (1) región de bajo solapamiento entre la señal de dispersión y el espectro de interés; (2) región de solapamiento parcial y (3) región de solapamiento total.

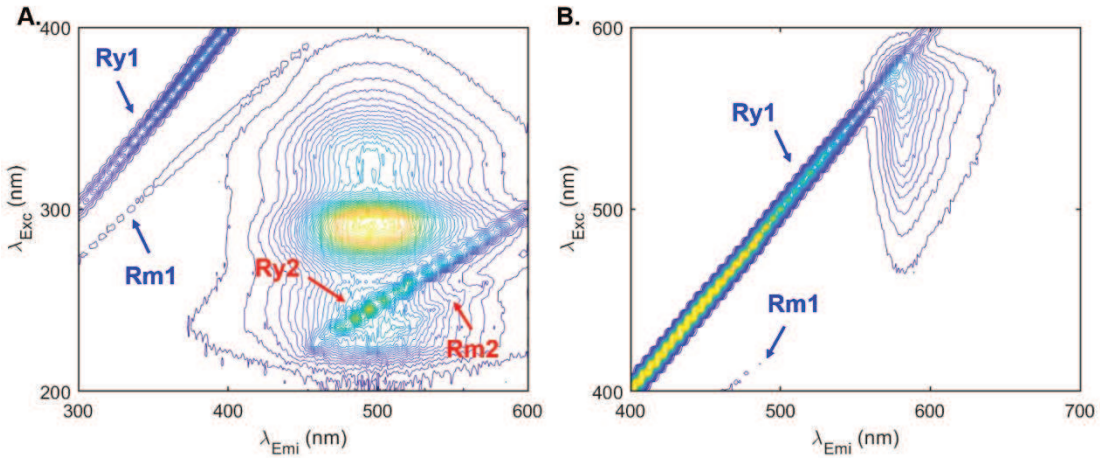


Figura 1.3. Mapas de contorno de las EEMs generadas para **A.** sistema A (OFL) y **B.** sistema B (RES), indicando las señales de dispersión de Ry y Rm de primer (azul) y segundo orden (rojo).

En la Figura 1.4 se muestran los resultados obtenidos para el análisis comparativo de las metodologías M1 a M5 (es decir, las reportadas en bibliografía), mientras que en la Figura 1.5, la corrección efectuada con la estrategia M6, desarrollada en este trabajo. En cada caso, las regiones (1), (2) y (3) seleccionadas para el análisis corresponden a espectros de emisión adquiridos a excitaciones de 225 nm, 250 nm y 290 nm, respectivamente para el sistema A y 550 nm, 565 nm y 575 nm, respectivamente para el sistema B.

Para llevar a cabo M1, una EEM de buffer o agua alcalinizada (según el sistema) fue generada en las mismas condiciones que las muestras, y posteriormente cada EEM blanco fue directamente restada a la EEM de la muestra correspondiente. Para el caso de M2, se reemplazaron por valores faltantes los datos originales a la izquierda del Rm1 y a la derecha del Ry2 en emisión, incluyendo las señales de dispersión. Como puede observarse en la Figura 1.4, el restado del blanco ocasiona una corrección completa y exacta de la señal de Rm en todos los casos, conservando la forma espectral y el patrón de ruido del dato. Sin embargo, para el caso de las señales de Ry, se produce una distorsión importante del espectro. Es esperable que en una solución homogénea y límpida, en ausencia de analito fluoróforo, la señal de Ry sea considerablemente mayor que en presencia del mismo. Por lo tanto, esta diferencia notable de intensidad del Ry ocasiona una depresión en la señal de interés, luego de la sustracción directa del blanco a las muestras, siendo el efecto mucho más pronunciado en regiones donde la señal de Ry se encuentra poco o muy solapada con la señal de interés (por ejemplo, Figura 1.4.A.1 y 1.4.A.2). Más aún, en casos de solapamiento, cuanto mayor es la señal del analito, más débil suele ser la señal del Ry. Por lo tanto, el efecto indeseable luego del restado del blanco, es más pronunciado. Esto se observa para el caso del sistema B (Figura 1.4.B).

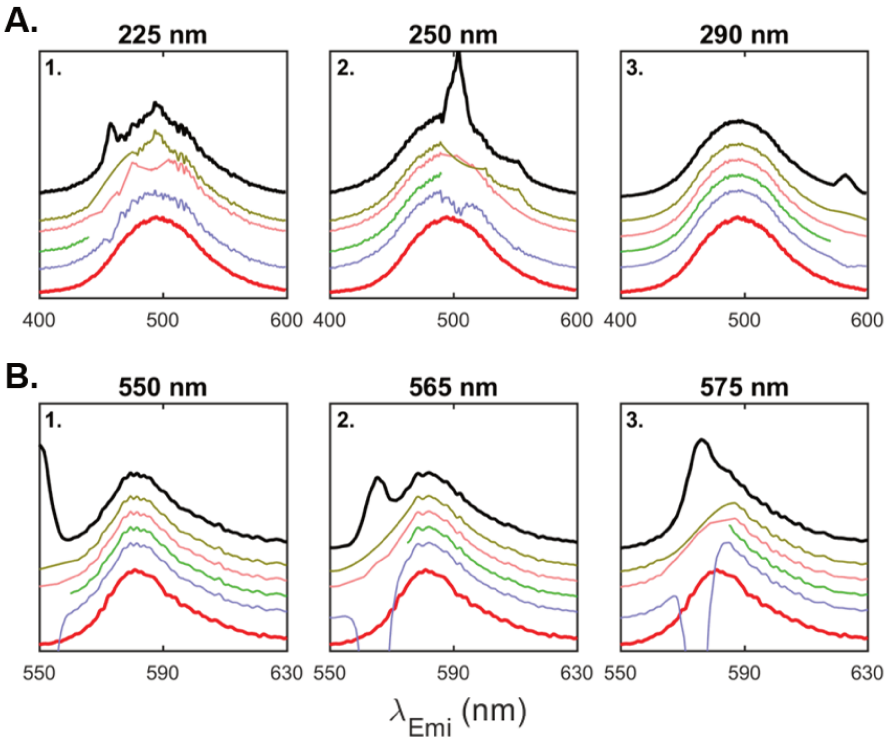


Figura 1.4. A. Espectros de emisión de OFL extraídos de una EEM a las longitudes de onda de excitación de **1.** 225 nm, **2.** 250 nm y **3.** 290 nm. **B.** Espectros de emisión de RES extraídos de una EEM a las longitudes de onda de excitación de **1.** 550 nm, **2.** 565 nm y **3.** 575 nm. En ambos casos, las tres longitudes de onda de excitación corresponden a las regiones de solapamiento bajo (1), parcial (2), y total (3) entre la señal de dispersión y la señal del analito. En todos los casos, se representa en líneas negras al espectro de emisión antes de la corrección. Además, los espectros corregidos con las metodologías M1 a M5 se muestran, respectivamente, en azul, verde, rosado y marrón. En líneas rojas se muestran los espectros tomados como referencia (longitud de onda de excitación de 305 nm y 535 nm para OFL y RES, respectivamente).

En el caso de M2, una clara desventaja tiene que ver con la pérdida significativa de información. En las Figura 1.4.A.1 y 1.4.B.3 se muestran los espectros respectivos de OFL y RES, en los que las regiones hacia la derecha del espectro de OFL y hacia la izquierda en el de RES desaparecen completamente, lo cual ocurre por el alto grado de solapamiento de las señales de dispersión en estos sistemas. En particular, este efecto es crítico en sistemas con corrimiento de Stokes pequeño como el caso del sistema B.

Por otro lado, para el caso de M3, los espectros corregidos muestran una alta similitud con los de referencia. Resultados preliminares evidenciaron que el resultado de la corrección empeora en casos de señales de dispersión más anchas (resultados no mostrados)[9]. El aspecto más interesante de este método es que no sólo se preserva la forma espectral, sino también existe una conservación parcial del patrón de ruido, debido a que la interpolación bidimensional tiene en cuenta información de los espectros

[9] Es bien conocido que el ancho promedio de las señales de dispersión está directamente relacionado con parámetros instrumentales. Concretamente, con los anchos de rendija (*slit*) de los monocromadores de excitación y emisión. A mayor ancho de rendija, las señales de *scattering* aumentan en ancho e intensidad.

vecinos en la EEM. No obstante, en el caso de un alto grado de solapamiento, se observan algunas distorsiones, especialmente para el caso del sistema B. Por otra parte, en el caso M4, la interpolación se realiza de manera unidimensional, con lo cual, la eficiencia en la corrección, depende fuertemente del ancho del *scattering* y del grado de solapamiento. Más aún, los resultados sugieren que la corrección es mejor cuando se trata de espectros anchos y simétricos, como el caso de la OFL, en lugar de espectros angostos y asimétricos como el de la RES. Otra cuestión importante a destacar es que el nivel de ruido instrumental puede representar una limitación para esta metodología. En este sentido, cuando los datos tienen un determinado nivel de ruido, debido a que la porción de espectro que es incorporada por la interpolación es libre de ruido, el patrón de ruido del espectro original se ve interrumpido. Además, cuando la señal de dispersión se solapa con el máximo de intensidad del espectro, se observa la introducción de artefactos, debido a que el método de interpolación no tiene en cuenta ningún aspecto espectral. En síntesis, para el caso de señales de dispersión altamente solapadas y anchas y en presencia de un nivel considerable de ruido, este método resulta poco eficiente. Finalmente, es necesario aclarar que, si bien la rutina de los autores admite la introducción de códigos adicionales, la misma fue implementada en su formato original, la cual no contempla la corrección de las señales de Rm2. Asimismo, para el caso de M5 no se muestra ningún resultado ya que no fue posible encontrar un conjunto de parámetros adecuado para la corrección de las EEMs y, por lo tanto, la estrategia no pudo ser implementada.

En la Figura 1.5 se observan los resultados tras el uso de M6. Es importante destacar que, debido a la filosofía de no interpolación de este método, el espectro corregido contiene el resultado de la diferencia entre la señal de dispersión y la curva gaussiana ajustada, junto con el ruido original del dato. En este sentido, resulta importante reiterar que el resultado de la corrección será tanto más bueno, cuanto mejor sea el ajuste no lineal. De lo contrario, se introducen artefactos en el espectro corregido.

Se puede observar en la Figura 1.5.B.1 que, en el caso de las señales incompletas, la corrección es satisfactoria debido a la aplicación de la interpolación unidimensional similar a la de M4. Por otro lado, en el caso de señales de *scattering* altamente solapadas, particularmente cuando los espectros son angostos o asimétricos (por ejemplo, en Figura 1.5.B.3), ninguna de las metodologías evaluadas resulta satisfactoria. Este último caso constituye una de las situaciones más dificultosas para el tratamiento de este tipo de datos y, hasta el momento, no ha sido publicada ninguna metodología capaz de corregir satisfactoriamente esta situación.

Respecto de las ventajas de M6, en todos los casos puede observarse una mejora notoria en la conservación del ruido instrumental del dato. En particular, la estrategia

desarrollada demuestra ser altamente eficiente en este sentido para datos con un nivel considerable de ruido (por ejemplo, Figura 1.5.A.1). Para enfatizar este último resultado, en la Figura 1.6 se muestra la corrección de un espectro de emisión generado en el laboratorio con un nivel considerable de ruido, en forma comparativa con el algoritmo basado en interpolación M4.

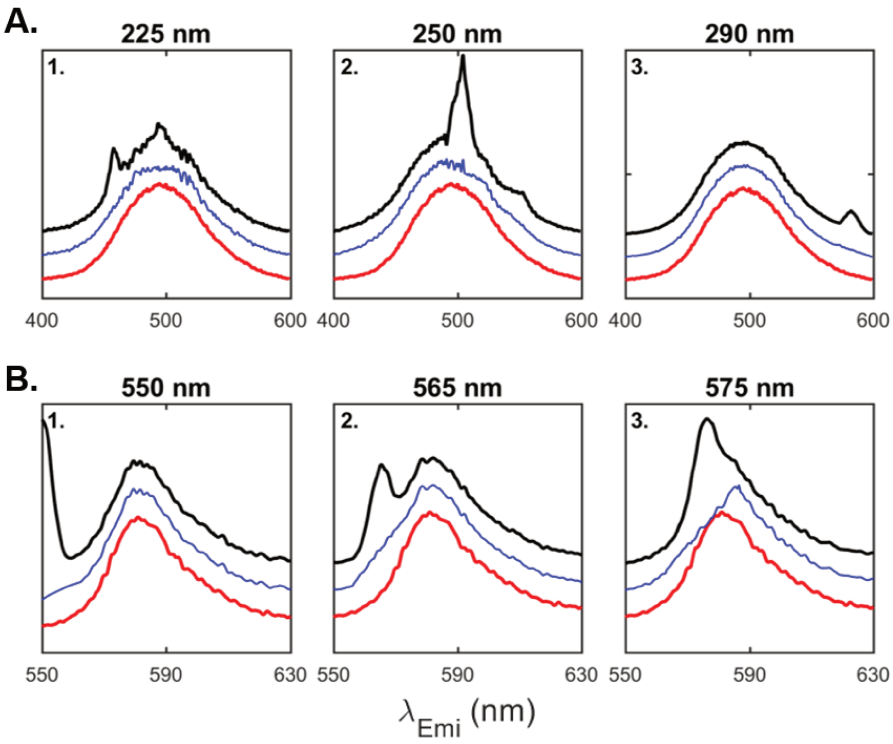


Figura 1.5. A. Espectros de emisión de OFL extraídos de una EEM a las longitudes de onda de excitación de **1.** 225 nm, **2.** 250 nm y **3.** 290 nm. **B.** Espectros de emisión de RES extraídos de una EEM a las longitudes de onda de excitación de **1.** 550 nm, **2.** 565 nm y **3.** 575 nm. En ambos casos, las tres longitudes de onda de excitación corresponden a las regiones de solapamiento bajo (1), parcial (2), y total (3) entre la señal de dispersión y la señal del analito. En todos los casos, se representa en líneas negras al espectro de emisión antes de la corrección. Además, los espectros corregidos con la metodología M6 se muestran en azul. En líneas rojas se muestran los espectros tomados como referencia (longitud de onda de excitación de 305 nm y 535 nm para OFL y RES, respectivamente).

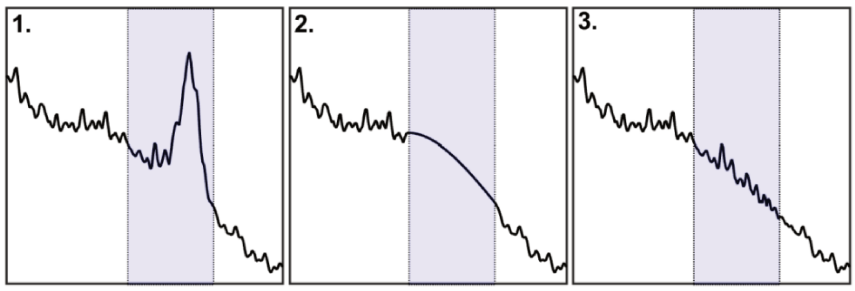


Figura 1.6. Región espectral de un espectro de emisión extraído de **1.** una EEM cruda; **2.** la misma EEM corregida mediante M4 y **3.** la misma EEM corregida mediante M6. Los rectángulos celestes indican la ventana espectral considerada para la corrección.

1.4.4. EEM_corr: una interfaz gráfica de usuario para la implementación amigable de tres estrategias de corrección

La última tarea de esta parte del trabajo estuvo abocada al desarrollo de una GUI, denominada "EEM_corr", para la implementación amigable de tres de las metodologías estudiadas: M1 (restado del blanco), M4 (*eemscat*) y M6 (*scscat*). Las mismas fueron integradas en un entorno gráfico sencillo e intuitivo, en el que no se requiere de ninguna habilidad en programación para su utilización. La versión publicada de la GUI en 2019 fue desarrollada en MATLAB R2015b, aunque luego fue actualizada para versiones de MATLAB más recientes, pero sin cambios sustanciales en sus códigos. Una vez que es ejecutada, la GUI muestra una ventana principal como la que se ilustra en la Figura 1.7.

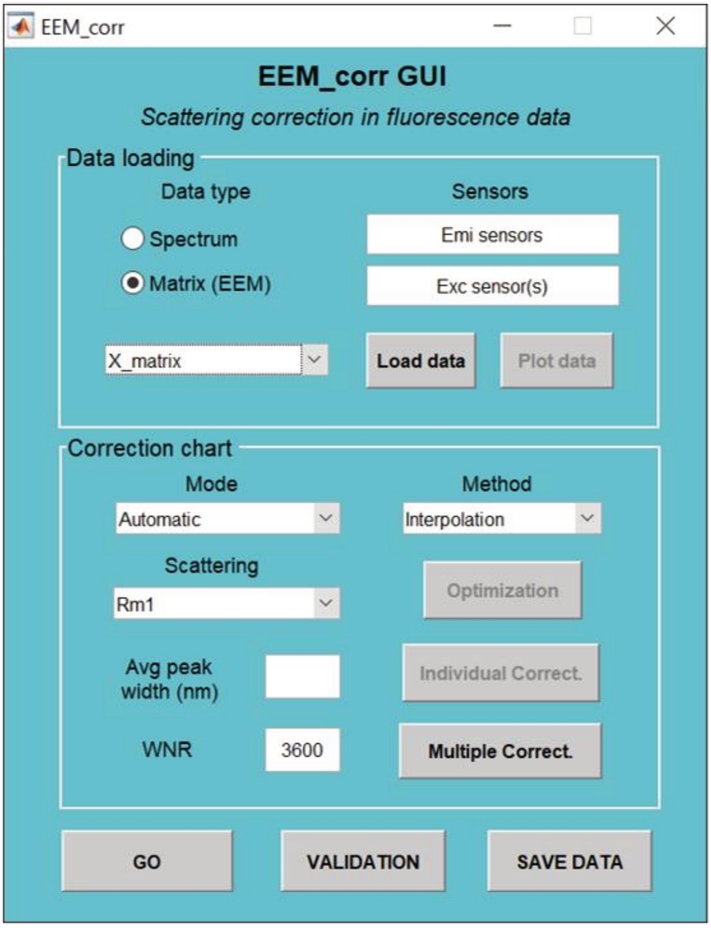


Figura 1.7. Ventana principal de la GUI EEM_corr para la implementación amigable de las metodologías M1, M4 y M6 de corrección de *scattering* en datos de fluorescencia. Panel superior (*Data loading*): módulo de carga y visualización de datos crudos; panel inferior (*Correction chart*): módulo de selección del método de corrección y ajuste de parámetros.

Este programa permite la corrección de datos de fluorescencia, tanto de espectros individuales, como de EEMs. En todos los casos, los tipos de *scattering* son tratados individualmente y se pueden optar por estrategias más o menos automatizadas, es decir, con menor o mayor grado de intervención sobre el método por parte del usuario.

Para la validación de los resultados que se obtienen, la interfaz contempla tanto herramientas gráficas para la inspección visual de espectros y EEMs, como herramientas estadísticas.

El paquete para correr la GUI dentro del entorno de MATLAB, junto con el manual de usuario y datos ejemplos se encuentran disponibles para su libre descarga en la página web del LADAQ (https://fbcb.web1.unl.edu.ar/laboratorios/ladaq/download/).

1.5. Conclusiones del capítulo

M1, basada en el restado del blanco, resulta apropiada para la corrección de señales de Rm, pero en el caso de las señales tipo Ry, se producen distorsiones importantes en la señal de interés, debido a las diferencias de intensidades entre las señales de Ry de la muestra y del blanco. Por otro lado, si bien se trata de una metodología simple y que no requiere el uso de algoritmos computacionales, la falta de disponibilidad de una muestra blanco adecuada puede resultar una limitación para su uso en algunas ocasiones.

M2, basada en la inserción de datos faltantes, puede ser aplicada sin problemas en sistemas con un corrimiento de Stokes grande, en donde no se produce un solapamiento importante entre la señal de interés y las de dispersión. En el caso contrario, puede ocasionar una significativa pérdida de información espectral. Asimismo, se debe tener en cuenta el algoritmo de procesamiento quimiométrico posterior, ya que puede no admitir datos faltantes, o bien, experimentar problemas de convergencia.

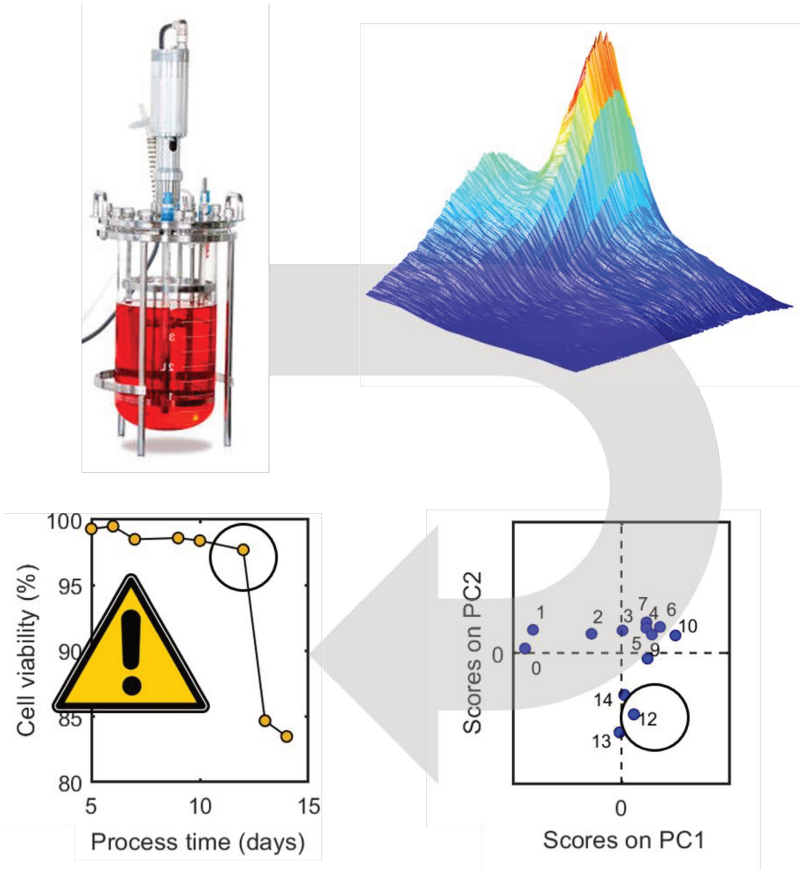
M3 y M4: ambas arrojan resultados similares, aunque la implementación computacional de M4 resulta comparativamente bastante más sencilla. M3, además, demostró ser capaz de conservar parcialmente el patrón de ruido debido a la estrategia de interpolación bidimensional.

M6 se propone como herramienta novedosa que optimiza y mejora algunas de las cualidades de las metodologías previas. Su ventaja principal radica en el hecho de que, cuando las señales de dispersión siguen un comportamiento gaussiano, esta estrategia es capaz de conservar satisfactoriamente el patrón de ruido original del dato, debido a la filosofía de corrección basada en ajuste gaussiano.

La GUI “EEM_corr” desarrollada, constituye una herramienta simple y versátil para la implementación en forma gráfica y amigable de las metodologías M1, M4 y M6, sin requerir habilidades de programación.

CAPÍTULO 2

Desarrollo de una estrategia de PAT cualitativa para el monitoreo de la viabilidad celular



La imagen del biorreactor corresponde a una vista parcial de una fotografía tomada de https://www.labwrench.com/equipment/21988/sartorius-group-biostat-a

2.1. Introducción

La selección de un método quimiométrico depende esencialmente del objetivo del análisis (agrupamiento, clasificación o calibración) y de las propiedades de los datos. Por ejemplo, el conjunto de variables de proceso (datos de orden cero) medidas sobre un conjunto de muestras de fermentación (dato de una vía) se pueden disponer en un arreglo matricial (dato de dos vías) y modelarlo mediante PCA. Este método constituye una de las herramientas más populares para el reconocimiento no supervisado de patrones [63]. Esta técnica facilita la visualización e interpretación de datos de primer orden, que se pueden obtener a partir de un conjunto de datos de orden cero tomados para una muestra, o bien, de señales instrumentales de tipo vectoriales, tales como espectros UV o fluorescencia, voltamperogramas, cromatogramas, etc.

PCA permite explorar de manera conjunta las correlaciones entre variables y evaluar similitudes o diferencias entre las muestras, lo que se denomina habitualmente como análisis de agrupamiento. Si se detectan grupos o *clusters*, las muestras pueden ser asignadas entonces a una determinada clase y luego, es posible generar modelos predictivos de clasificación mediante la descomposición supervisada de los datos. En este sentido, uno de los algoritmos más utilizados para la clasificación multivariada es el PLS-DA [64].

Por otro lado, los datos espectrales de segundo orden también pueden ser analizados mediante PCA y PLS-DA si son desdoblados (vectorizados), y los espectros concatenados resultantes que se obtienen para un conjunto de muestras, son apilados en un nuevo arreglo matricial. En este caso, los datos de segundo orden son tratados como datos de primer orden y en la literatura especializada, suele utilizarse para la denominación de los métodos PCA y PLS-DA con datos de segundo orden desdoblados las abreviaturas U-PCA y U-PLS-DA, haciendo referencia al desdoblamiento (*unfolding*) de los datos. Sin embargo, en este trabajo, aunque se hayan usado matrices desdobladas, se utilizarán en todos los casos las denominaciones PCA y PLS por simplicidad.

Por otra parte, si se desean conservar ambos modos instrumentales, es posible descomponer la información espectral mediante la estrategia de MCR-ALS [65]. Este algoritmo es capaz de descomponer la información espectral y extraer los perfiles de señal y las contribuciones relativas de los componentes individuales presentes en el sistema estudiado. Entre otros métodos quimiométricos de segundo orden, MCR-ALS ha demostrado ser una herramienta poderosa, no solamente para el desarrollo de métodos de calibración, sino también con fines exploratorios, debido a su versatilidad y a la interpretabilidad química de sus resultados. Esto implica, por ejemplo, que los

perfiles espectrales que se obtienen por MCR-ALS pueden ser asociados con componentes reales presentes en las muestras analizadas.

En este trabajo se ha tomado como caso de estudio la etapa de fermentación de un bioproceso estándar para la producción de etanercept en células CHO. En este proceso, las células son cultivadas en régimen de perfusión. Tal como ha sido extensamente probado en la literatura, este régimen de cultivo permite maximizar la productividad de la proteína de interés en alta densidad celular, mediante el suplemento de medio de cultivo fresco y la remoción continua de medio metabolizado [66].

El bioproceso analizado presenta una particularidad que es inherente a la forma de crecimiento que tiene el clon productor utilizado. Hacia el final de cada fermentación, se observa una caída importante del parámetro porcentaje de células viables que no responde a ningún cambio en las variables operativas del proceso. Además, a pesar de que la fermentación se realiza bajo condiciones operativas estándares, el momento en que esta situación comienza a ocurrir resulta impredecible. Debido a que el monitoreo de la viabilidad celular resulta crítica para evaluar el desempeño del bioproceso, esta parte del trabajo de tesis se ha orientado hacia el desarrollo de una estrategia de PAT cualitativa para el seguimiento de este parámetro.

En el presente capítulo se realiza una descripción detallada del bioproceso y de la metodología empleada para la generación y análisis cualitativo de los datos, tanto exploratorios como predictivos. En este sentido, se utilizaron los dos tipos de datos obtenidos a partir de muestras diarias de fermentación: datos de orden cero, representados por las variables de proceso (medidas a través de técnicas univariadas de referencia); y datos de segundo orden, representados por las EEMs obtenidas directamente a partir de muestras de sobrenadante de cultivo.

Con todos los datos generados se planteó, en primer lugar, realizar una caracterización de la fermentación y análisis de agrupamiento, respecto de las variables de proceso y de la información espectral generada, utilizando para ello los métodos quimiométricos exploratorios PCA y MCR-ALS. Posteriormente, el desarrollo de una estrategia de PAT para el monitoreo de la viabilidad celular se basó en el uso de PLS-DA como algoritmo de clasificación.

En las siguientes subsecciones se detallan algunos fundamentos de las técnicas quimiométricas empleadas. Es importante señalar que, en todos los casos, los fundamentos teóricos y algoritmos específicos han sido extensamente reportados en la bibliografía específica, por lo que en este trabajo sólo se presentarán sus fundamentos básicos.

2.1.1. PCA como método quimiométrico de reconocimiento no supervisado de patrones para datos de primer orden

El PCA constituye una poderosa técnica estadística multivariada clásica que ha sido apropiada por diversas disciplinas debido a su versatilidad y aplicaciones. En particular, la quimiometría hace un uso extensivo de esta técnica debido a las bondades matemáticas que presentan muchos de los datos químicos instrumentales. Desde el punto de vista quimiométrico, el PCA se considera un método de reconocimiento de patrones no supervisado de primer orden. Como tal, utiliza datos de dos vías (matriciales) como *input*[10]. Esta técnica se basa en la compresión de los datos mediante una descomposición matricial, generando combinaciones lineales de las variables originales que maximizan la variabilidad capturada por el modelo, eliminando ruido experimental e información redundante (altamente correlacionada) [63]. De esta manera, el método facilita la interpretación de un conjunto de datos multivariados, permitiendo explorar las correlaciones entre variables y las similitudes entre las muestras.

Dado un dato matricial $\mathbf{X}$ de I observaciones (muestras de entrenamiento o calibración) × J variables predictoras (por ejemplo, variables de proceso o sensores espectrales), el método de PCA efectúa una proyección ortogonal de las variables experimentales (variables reales) en un espacio de A variables latentes (LVs) vectoriales en la dirección de máxima variabilidad del dato, de acuerdo al siguiente modelo [67]:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \tag{2.1}$$

donde $\mathbf{T}$ es de tamaño $I \times A$ y se denomina matriz de *scores* y $\mathbf{P}^{\mathrm{T}}$ es la matriz de *loadings* de $A \times J$. Cada vector de *loadings* $\mathbf{p}^{\mathrm{T}}$está asociado a un vector de *scores* $\mathbf{t}$ y cada uno de los A pares $\mathbf{t}\mathbf{p}^{\mathrm{T}}$ constituye una componente principal (PC) del modelo[11].

El primer paso del método consiste en el cálculo de la matriz de *loadings*, que matemáticamente son los autovectores asociados a la matriz cuadrada $\mathbf{X}\mathbf{X}^{\mathrm{T}}$. Existen diferentes métodos auxiliares para llevar adelante esta operación, aunque los más habituales son la descomposición en valores singulares (SVD) y el método de cuadrados mínimos parciales iterativos no lineales (NIPALS) [68]. Los vectores filas $\mathbf{p}^{\mathrm{T}}$ de la matriz de *loadings* tienen la propiedad de ser ortonormales (ortogonales y de magnitud unitaria) y representan las LVs del modelo (combinaciones lineales de las variables originales). Una vez calculada $\mathbf{P}^{\mathrm{T}}$, es posible resolver la Ec. 2.1 en $\mathbf{T}$ para hallar la matriz de *scores.* En este sentido, la proyección de $\mathbf{X}$ en el espacio de los *loadings* es lo que permite la reducción real de la dimensión de los datos, siempre que $J \gg A$. En general, se espera

[10] *Input* (entrada): dato que alimenta al modelo.
[11] En la bibliografía especializada, los términos “componente principal” y “variable latente” suelen utilizarse como sinónimos.

para un modelo PCA que, si las variables están altamente correlacionadas, un número relativamente pequeño de PCs expliquen un porcentaje alto de la variabilidad de los datos, lo cual puede estimarse a partir del cálculo de un estadístico asociado al modelo que se conoce como varianza explicada. Sea $\mathbf{X}_A$ la matriz reconstruida que resulta del producto $\mathbf{T}_A\mathbf{P}_A^{\mathrm{T}}$ para un modelo PCA con A PCs, entonces el porcentaje acumulado de varianza explicada (PVA) se calcula como:

$$\mathrm{PVA} = \frac{\|\mathbf{X}_A\|^2}{\|\mathbf{X}\|^2} 100 \tag{2.2}$$

Si se efectúa el mismo cálculo para cada una de las PCs (desde 1 hasta A), entonces es posible conocer su contribución relativa a la varianza explicada total. Si las variables experimentales tienen un alto grado de correlación, la inspección de las PCs permite establecer una diferencia entre componentes significativas (capturan información relevante para describir a la matriz $\mathbf{X}$) y no significativas (asociadas con el ruido experimental). Esto conlleva la necesidad de determinar un número óptimo A de PCs para la construcción del modelo. En este sentido, si el número de PCs es demasiado bajo, entonces las matrices de *loadings* y *scores* no serán representativas del total de información relevante contenida en la matriz $\mathbf{X}$ (modelo subajustado). Por el contrario, si el número A es demasiado grande, el modelo capturará información poco relevante que puede afectar y/o entorpecer la interpretación de la solución (modelo sobreajustado), en especial, si los resultados del modelo se utilizan posteriormente con fines predictivos. La sub- o sobreestimación de los parámetros ajustables de un modelo se relacionan directamente con una máxima fundamental del aprendizaje estadístico conocida como "Principio de parsimonia de Ockham"[12]. Este principio establece que un modelo estadístico óptimo debe ser parsimonioso, en el sentido de que la simplicidad está asociada directamente con varios de sus atributos, tales como la robustez, la interpretabilidad y la capacidad predictiva o de generalización.

Numerosos son los criterios que han sido descriptos para la determinación del número óptimo de PC, con mayor o menor grado de sofisticación estadística. Entre ellos, los más utilizados en quimiometría (y que serán de uso común en este trabajo) son los siguientes:

- Inspección visual de *loadings*: en el caso de datos espectroscópicos, los *loadings* significativos guardan cierta reminiscencia con los perfiles espectrales de los analitos que generan la señal, mientras que aquellos *loadings* poco significativos muestran un comportamiento que refleja el ruido instrumental.

[12] *Frustra fit per plura quod potest fieri per pauciora* ("es inútil hacer con más lo que se puede hacer con menos").

- Contribución relativa de cada PC al PVA del modelo: en general, aquellas PCs que contribuyen poco a la variabilidad total capturada por el modelo (por ejemplo, menor a 1%) se consideran poco significativas.
- Validación cruzada o *cross-validation* (CV): constituye una técnica estadística de uso muy común en análisis multivariado y en general sirve para evaluar la autoconsistencia y capacidad predictiva de un modelo. En el caso de la determinación del número óptimo de PC, la CV consiste en dejar fuera un grupo pequeño de muestras de entrenamiento y predecirlas durante el ajuste del modelo, calculando los errores de predicción mediante un proceso iterativo. La estadística asociada a los residuos que se obtiene con esta técnica permite establecer criterios para la selección del número de PC. Por ejemplo, si se colectan los residuos de las predicciones de las muestras dejadas fuera en cada iteración, es posible calcular el estadístico suma de los cuadrados de los residuos de predicción (PRESS) como:

$$\mathrm{PRESS}_A = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(e_{ij}{}^{(A)}\right)^2 \tag{2.3}$$

donde $e_{ij}{}^{(A)}$ es el residuo de la muestra i y la variable j para el ciclo de CV que corresponde a A PCs. A partir del cálculo de PRESS, es posible calcular la raíz cuadrada del error cuadrático medio de la validación cruzada (RMSECV) como:

$$\mathrm{RMSECV}_A = \sqrt{\frac{\mathrm{PRESS}_A}{IJ}} \tag{2.4}$$

Así, un criterio para la selección del número óptimo de A es seleccionar aquel número que minimiza el valor de RMSECV. La manera en que se particiona iterativamente el conjunto de muestras de entrenamiento es lo que da lugar a una familia de métodos de CV, cuya elección depende esencialmente de la cantidad y calidad de los datos disponibles. Los detalles sobre la implementación de los diferentes métodos de CV empleados en este trabajo (tanto para PCA como para otros métodos quimiométricos) serán descriptos en las secciones de Materiales y métodos de cada capítulo, según corresponda.

Una vez obtenido un modelo PCA satisfactorio, diversas son las utilidades que se le pueden dar a las matrices de *loadings* y *scores.* En líneas generales, la inspección visual de la matriz de *scores* permite establecer similitudes y diferencias entre conjuntos de muestras. Esto tiene implicancia para la exploración de valores atípicos (*outliers*) y para los análisis de agrupamiento, ya que el conjunto de *scores* significativos que caracteriza a una dada muestra se encuentra directamente asociado con la idea de distancia geométrica en el espacio de PCA. Asimismo, la matriz de *scores* puede

utilizarse como *input* de un segundo método quimiométrico, por ejemplo, de calibración o de clasificación. Finalmente, la matriz de *loadings* captura información valiosa de índole cualitativa, que permite por un lado conocer el peso relativo que tienen las variables experimentales en cada PC y, por otro lado, interpretar la influencia de las variables reales en el agrupamiento o diferenciación de las muestras.

2.1.2. MCR-ALS como método quimiométrico no supervisado para la descomposición de datos espectrales de segundo orden

MCR-ALS constituye uno de los métodos quimiométricos más frecuentemente reportados para el modelado de datos bilineales de segundo orden [69]. Tal como se describió anteriormente, una EEM libre de señales de dispersión cumple con el supuesto de bilinealidad de rango pequeño [70]. De esta manera, MCR-ALS permite descomponer la señal matricial de una muestra individual, de acuerdo al siguiente modelo [71]:

$$\mathbf{X} = \mathbf{C}\,\mathbf{S}^{\mathrm{T}} + \mathbf{E} \tag{2.5}$$

donde $\mathbf{X}$ es una matriz de $J \times K$ que contiene J y K sensores en cada modo espectral. $\mathbf{C}$ y $\mathbf{S}^{\mathrm{T}}$ representan, respectivamente, submatrices de tamaño $J \times A$ y $A \times K$ que capturan los perfiles espectrales en cada modo instrumental (excitación y emisión) de un número A de componentes químicos del sistema que generan las señales. $\mathbf{E}$ es la matriz error del modelo.

La interpretabilidad química que posee este modelo se debe al hecho de que el algoritmo que permite hallar su solución óptima, incluye una serie de restricciones que se fundamentan en aspectos químicos propios del sistema que se desea modelar. En este sentido, la descomposición de un dato bilineal $\mathbf{X}$ por MCR se efectúa mediante la optimización de los perfiles de concentración y espectros puros a través de un algoritmo iterativo conocido como ALS, bajo una serie de restricciones adecuadas. Para iniciar el algoritmo es necesario proporcionar una estimación inicial, ya sea de los perfiles espectrales o de concentración. La función de minimización general de ALS puede ser expresada como:

$$\min \left\| \mathbf{X} - \mathbf{C}\mathbf{S}^{\mathrm{T}} \right\| \tag{2.6}$$

donde || || representa la norma euclídea. La filosofía de los cuadrados mínimos alternantes radica en el hecho de que las matrices $\mathbf{C}$ y $\mathbf{S}^{\mathrm{T}}$ son calculadas en cada iteración de manera alternada.

Cuando se desean modelar EEMs obtenidas para un conjunto de muestras (por ejemplo, para las muestras diarias de una fermentación), es posible formular un MCR-ALS extendido como [71]:

$$\mathbf{X}_{\mathrm{aum}} = \mathbf{C}_{\mathrm{aum}}\,\mathbf{S}^{\mathrm{T}} + \mathbf{E}_{\mathrm{aum}} \tag{2.7}$$

donde I EEMs de $J \times K$ son dispuestas en un arreglo $\mathbf{X}_{\mathrm{aum}}$ aumentado en filas o en columnas (de tamaño $J \times KI$ o $JI \times K$, respectivamente según el caso), disponiendo las EEMs una a continuación de la otra (Figura 2 de la Introducción). En este caso, el arreglo $\mathbf{X}_{\mathrm{aum}}$ se caracteriza por presentar un modo aumentado y uno no aumentado.

En líneas generales, la implementación del método MCR-ALS implica las siguientes tres etapas:

a. Determinación del número de componentes: se refiere al número de especies químicas que generan la señal espectral. Cuando se trata de un sistema químicamente definido, este número suele ser conocido, aunque se debe tener en cuenta que, si dos analitos presentan exactamente el mismo comportamiento espectral en ambos modos, no es posible resolverlos por este método. Por otra parte, en el caso de sistemas de composición indefinida, este número puede estimarse mediante un análisis auxiliar. En este sentido, el método de estimación de componentes más habitual consiste en explorar la magnitud relativa de los autovalores de la matriz $\mathbf{X}$, luego de una descomposición mediante PCA, ya que MCR asume que la variabilidad de $\mathbf{X}$ puede ser descripta mediante un número de contribuciones A igual al número de componentes principales significativas.
b. Inicialización: para inicializar el algoritmo ALS, es necesario proveer una estimación de $\mathbf{C}$ o de $\mathbf{S}^{\mathrm{T}}$. Para ello, se prefieren estimadores que guarden algún tipo de relación con las propiedades de los perfiles que se desean modelar. Por esta razón, la inicialización con valores al azar no se prefiere en este método. Por el contrario, la inicialización puede basarse en el conocimiento previo del sistema (por ejemplo, mediante el uso de espectros de referencia), o bien, en métodos auxiliares. Entre los métodos de inicialización mayormente utilizados en sistemas con datos espectrales, se destaca la técnica denominada de las "variables más puras" (*purest variables*) [72].
c. Iteración mediante ALS, bajo restricciones. En cada iteración se calculan dos cifras de mérito asociadas al ajuste del modelo, las cuales se utilizan habitualmente como criterios de convergencia. Las mismas son el porcentaje de falta de ajuste (LOF):

$$\%\mathrm{LOF} = 100 \sqrt{\frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} x_{ij}^2}} \tag{2.8}$$

y el PVA, cuya definición es análoga a la que establece la Ec. 2.2. En la Ec. 2.8, e_{ij} es el residuo que se obtiene a partir de la reconstrucción MCR del elemento asociado x_{ij} en el dato original (para la i-ésima muestra y el j-ésimo sensor espectral). Habitualmente, la convergencia es alcanzada cuando la diferencia en

el ajuste del modelo entre dos iteraciones consecutivas no difiere significativamente (por ejemplo, menos del 0,1% de diferencia entre los %LOF entre dos iteraciones sucesivas).

Otro aspecto esencial en el modelado MCR es el uso de restricciones, lo cual se efectúa con dos objetivos: (i) introducir conocimiento químico sobre el sistema al modelo, de manera de darle interpretabilidad química a su solución; (ii) eliminar o minimizar los fenómenos de ambigüedad[13] [65,73]. Una restricción se define como cualquier propiedad matemática o química que está presente de manera sistemática en los perfiles de los componentes químicos del sistema [71]. Así, durante cada ciclo ALS, los perfiles calculados se modifican de manera tal que estos se ajusten a las restricciones previamente establecidas. Existen diferentes tipos y formas de aplicar restricciones y los detalles teóricos y metodológicos exceden a los objetivos de este trabajo. Algunas de las restricciones más habituales son, por ejemplo, la no negatividad (se fuerza a que los perfiles calculados sean estrictamente positivos) y la unimodalidad (fuerza a que los perfiles espectrales muestren un único máximo). La selección y uso de restricciones depende fuertemente del conocimiento previo sobre el sistema en estudio y reduce significativamente la ambigüedad de MCR, permitiendo obtener soluciones confiables con interpretabilidad química.

En líneas generales, la descomposición de datos de segundo orden por MCR-ALS permite, en primera instancia, efectuar una exploración cualitativa de los perfiles individuales de los compuestos que generan las señales instrumentales, lo cual resulta relevante *per se* para la caracterización de sistemas químicos complejos (por ejemplo, en la identificación de especies químicas en una muestra indefinida o en la caracterización espectral de especies de transición en estudios cinéticos). Asimismo, el área bajo la curva de los perfiles capturados por la matriz $\mathbf{C}$ están asociados con la abundancia relativa de los A componentes modelados. En este sentido, cuando se considera un arreglo de datos del tipo $\mathbf{X}_{\mathrm{aum}}$, luego de la convergencia, las áreas bajo la curva de los perfiles obtenidos en el modo aumentado, $\mathbf{C}_{\mathrm{aum}}$, acopladas a algún modelo de regresión, pueden utilizarse para la obtención de modelos de calibración.

En virtud de todas las características mencionadas, en su versión clásica, MCR-ALS puede considerarse como un método de segundo orden no supervisado con fines

[13] Debido a la flexibilidad del modelo MCR y la versatilidad del algoritmo de optimización ALS, la ambigüedad ocurre porque puede existir un conjunto más o menos extenso de combinaciones de matrices $\mathbf{C}$ y $\mathbf{S}^{\mathrm{T}}$ que describan adecuadamente la matriz $\mathbf{X}$ en términos de ajuste del modelo. Sin embargo, no todas las soluciones que se obtienen, resultan relevantes desde el punto de vista químico/analítico. Los fenómenos de ambigüedad de MCR constituyen un área activa de investigación dentro de la disciplina y no serán tratados en esta tesis.

exploratorios, o bien, acoplado a un modelo de regresión, como un auténtico método de calibración que permite explotar la ventaja de segundo orden.

2.1.3. PLS-DA como método quimiométrico de reconocimiento supervisado de patrones para datos de primer orden

El método PLS-DA es una variante de una técnica tradicional de regresión multivariada, conocida como PLS. A diferencia de los métodos anteriores, PLS constituye una metodología supervisada de primer orden, cuyos *inputs* son una matriz de datos $\mathbf{X}$ de I muestras $\times$ J variables, conocida coloquialmente como "bloque X" (o también "bloque de predictoras") y un vector de variables respuesta o *target* $\mathbf{y}$ de tamaño $I \times 1$ (también denominado como "bloque Y")[14], cuyos elementos y_i representan una propiedad que se desea regresar con el conjunto de predictoras. Una vez ajustado el modelo, como en cualquier técnica de regresión, el objetivo es poder predecir una propiedad de interés de un sistema, a partir de las predictoras medidas sobre una muestra incógnita.

El método PLS se basa en una descomposición lineal supervisada del bloque X con reducción de la dimensión, en un conjunto de A LVs que maximizan tanto la varianza explicada del bloque de predictoras como la covarianza con el bloque Y [74]. La formulación básica del modelo es equivalente a la presentada en la Ec. 2.1 aunque el cálculo de *loadings* y *scores* se realiza de una manera diferente[15]. El método PLS, además de la descomposición del bloque X, implica la resolución del problema de regresión inverso dado por [75]:

$$\mathbf{y} = \mathbf{T}_A \mathbf{b} + \mathbf{e} \tag{2.9}$$

donde $\mathbf{T}_A$ es la matriz $I \times A$ de *scores* en el espacio de A LVs de PLS, $\mathbf{b}$ es el vector de coeficientes de regresión (de tamaño $A \times 1$) y $\mathbf{e}$ el vector error. Análogamente al método PCA, el número A de LVs debe ser optimizado para garantizar la parsimonia del modelo, y por lo tanto, una buena capacidad predictiva. Los criterios de selección son similares a los que se utilizan para PCA. Los detalles de la implementación de métodos de CV se presentan en secciones posteriores.

Una vez estimado el vector de coeficientes de regresión, durante la etapa de predicción, el conjunto de J variables medidas para una muestra incógnita $\mathbf{x}_{test}$ es

[14] Si se construye un modelo PLS para la predicción de una única variable respuesta (el "bloque Y" es un vector), el modelo se denomina PLS1. Este método también admite la regresión con más de una variable respuesta a través del algoritmo conocido como PLS2. En este trabajo, se utilizó el método PLS1 en todos los casos y en el texto se lo denominará siempre como PLS, por simplicidad.

[15] Por simplicidad, se omiten los detalles específicos del método, los cuales pueden consultarse en la bibliografía específica.

proyectado en el espacio de PLS para obtener un vector de *scores* de la muestra $\mathbf{t}_A$, de acuerdo a:

$$\mathbf{t}_A = \left(\mathbf{W}_A^{\mathrm{T}}\mathbf{P}_A\right)^{-1}\mathbf{W}_A\mathbf{x}_{test} \tag{2.10}$$

donde $\mathbf{W}_A$ y $\mathbf{P}_A$ son las matrices de *loadings* de PLS truncadas. El vector $\mathbf{t}_A$ es luego empleado para calcular la propiedad o concentración de interés $\hat{y}$ como:

$$\hat{y} = \mathbf{b}^{\mathrm{T}}\mathbf{t}_A \tag{2.11}$$

Cuando las componentes del bloque Y representan una variable continua que refleja una propiedad de interés a ser estimada (por ejemplo, la concentración de un analito), entonces PLS opera como un modelo de calibración que, dadas sus características, permite explotar la ventaja de primer orden. Por el contrario, si el bloque Y es un vector de índices lógicos que representan información codificada de clases de objetos (por ejemplo 0 y 1 para un problema de dos clases), entonces a la etapa de regresión PLS se suma el análisis discriminante (DA) que permite obtener modelos de clasificación.

PLS-DA constituye uno de los algoritmos de discriminación de clases más frecuentemente reportados en la bibliografía. En este método, una vez que se obtienen las predicciones de las clases $\hat{y}$ del conjunto de entrenamiento mediante regresión PLS, la etapa de DA consiste en calcular una probabilidad de pertenencia a una clase dada para realizar la asignación, de acuerdo a una regla de clasificación. Existen diferentes métodos descriptos al respecto. Una de las técnicas más habituales consiste en calcular un umbral de clasificación, basado en teoría Bayesiana. Dicho umbral está representado por un hiperplano discriminante que minimiza el número de falsos negativos y falsos positivos. Este es utilizado luego en la etapa de predicción para la asignación de clases de muestras incógnitas [76,66].

Para la evaluación de la autoconsistencia y la caracterización de la capacidad predictiva de un modelo de clasificación como PLS-DA se calculan una serie de estimadores numéricos que se conocen en general como índices de clasificación [77].

Para calcular estos índices, se debe partir de una matriz de confusión, que se obtiene a partir de los resultados de las clases predichas de un conjunto de muestras conocidas. Para un problema de g clases, la matriz de confusión es un arreglo de tamaño $g \times g$ que ubica en la diagonal principal la cantidad de muestras correctamente clasificadas, y deja fuera de la misma, aquellas muestras que no fueron correctamente asignadas (falsos positivos y falsos negativos). Existen diferentes tipos de índices de clasificación. Por un lado, algunos índices reflejan la capacidad de discriminación del modelo de una clase en particular. Estos son la sensibilidad, la especificidad y la

precisión. La sensibilidad de una dada clase g (Sn_g) representa la habilidad del modelo de identificar correctamente muestras de la clase g y numéricamente se calcula como:

$$Sn_g = \frac{c_{gg}}{N_g} \tag{2.12}$$

donde c_{gg} es el número de muestras de la clase g clasificadas correctamente y N_g es el número total de muestras de la clase g. En segundo lugar, la precisión (Pr_g) se define como:

$$Pr_g = \frac{c_{gg}}{n_g} \tag{2.13}$$

donde n_g es el número de muestras asignadas a la clase g (predichas por el modelo). Finalmente, la especificidad (Sp_g) representa la habilidad del modelo de rechazar muestras de otra clase k, de acuerdo con:

$$Sp_g = \frac{\sum_{k=1,k\neq g}^{G}(N_k - c_{kg})}{I - N_g} \tag{2.14}$$

donde c_{kg} es el número de muestras de la clase g asignadas a la clase k y N_k es el número de muestras de la clase k.

Además de los índices relativos a una clase, existen indicadores globales. Los más importantes son la exactitud (Acc), que se define como el cociente entre el número de muestras clasificadas correctamente y el número total de predicciones, y la tasa de no error (NER), la cual se estima como la media aritmética de la Sn_g para todas las clases (también conocida como sensibilidad promedio).

2.2. Objetivos específicos del capítulo

En este capítulo se plantearon los siguientes objetivos específicos:

- recabar datos de monitoreo de los principales CPPs de diferentes lotes del proceso de fermentación de etanercept;
- generar conjuntos de datos analíticos de segundo orden a partir de muestras de fermentación, mediante espectroscopía de fluorescencia;
- realizar una caracterización del proceso y análisis de agrupamiento (*clustering*) en términos de los CPPs y de la información espectral, mediante el uso de algoritmos quimiométricos exploratorios;
- desarrollar una estrategia de PAT para el monitoreo de la viabilidad celular basada en el método PLS-DA de clasificación multivariada.

2.3. Materiales y métodos

2.3.1. Condiciones de cultivo celular

Las células CHO productoras de etanercept y adaptadas al crecimiento en suspensión fueron cultivadas en alta densidad en biorreactores operados en modo perfusión. Se utilizó un medio de cultivo libre de suero fetal bovino y con hidrógeno carbonato de sodio pH 7,2 como solución buffer. Todos los reactivos fueron de calidad para cultivo celular.

Para este estudio se consideraron dos tipos de bioprocesos (Sistema A y Sistema B). Si bien en ambos casos el tipo de plataforma celular empleado, el régimen de cultivo y el producto de interés son los mismos, las condiciones experimentales no fueron exactamente iguales. Por lo tanto, ambos sistemas han sido tratados de manera independiente a lo largo de la investigación. Además, se debe mencionar que la información respecto a los clones recombinantes de células CHO y la formulación de los medios de cultivos utilizados en cada caso son confidenciales.

El Sistema A consistió en un proceso de escala mediana (piloto), mientras que el Sistema B, en un proceso de escala grande (industrial). En cada caso, se utilizaron biorreactores tanques agitados de volumen de trabajo 4,5 L (Sartorius) y 100 L (New MBR), respectivamente, utilizando como sistema de retención celular un dispositivo interno de tipo *spin filter* de 8-14 µm de tamaño de poro. En cada sistema, el tipo de clon celular productor de etanercept fue el mismo, aunque el número de pasajes y la formulación del medio de cultivo fue diferente en cada caso. Además, dadas las características de los reactores empleados, resulta evidente que los sistemas no sólo se diferencian por la escala operativa, sino también por otras diferencias tales como la geometría del tanque y la velocidad de agitación.

En relación al control de las variables fisicoquímicas oxígeno disuelto (DO), pH, temperatura y tasa de perfusión (expresada como proporción de volumen de reactor intercambiado cada 24 h), se implementaron diferentes estrategias operativas de acuerdo a los requerimientos del cultivo. En particular, se utilizó el aireado continuo con una mezcla de aire, oxígeno y nitrógeno mediante una membrana de aireación/agitación para mantener el DO en un valor de *setpoint* de 60%. La presión interna del reactor se mantuvo constante mediante una válvula de escape automática. Por otro lado, el pH se mantuvo controlado en el rango 7,0-7,2. En este sentido y según necesidad, se bombeó dióxido de carbono para acidificar el medio, mientras que para aumentar el valor de pH se modificó la velocidad de perfusión como estrategia de control.

Con el objetivo de controlar el crecimiento celular, las variables temperatura y tasa de perfusión fueron manipuladas de acuerdo al siguiente esquema. En relación a la

temperatura, se aplicaron diferentes rampas luego de la estabilización del cultivo, en el rango de 32 a 37 °C. La temperatura fue regulada mediante circulación de agua caliente o fría a través de una camisa de agua. La tasa de perfusión fue modificada de acuerdo a la densidad celular y las concentraciones de glucosa y lactato. En este sentido, la cantidad de volumen perfundido se varió aproximadamente entre 0,2 y 1,0 volumen de reactor por día, durante todo el proceso y en ningún caso se realizó el sangrado del biorreactor.

2.3.2. Muestreo

En total se analizaron datos provenientes de seis fermentaciones independientes (cuatro del Sistema A y dos del Sistema B). En todos los casos, se tomaron muestras del seno del biorreactor en esterilidad para el monitoreo de las variables de proceso (CPPs) y para la generación de EEMs. Para el monitoreo del bioproceso, el muestreo fue realizado aproximadamente cada 24-36 horas hasta el final de la fermentación. En el caso de los datos espectrales, para el Sistema A, todas las muestras recolectadas para el monitoreo fueron procesadas para generar EEMs. En cuanto al Sistema B, si bien los CPPs fueron medidas a lo largo del proceso completo, la generación de datos de fluorescencia sólo se pudo realizar en intervalos de tiempo específicos ya que, debido a cuestiones logísticas propias de la empresa, no fue posible procesar la totalidad de muestras recolectadas. La identificación de cada lote de proceso analizado, su duración y el número total de muestras procesadas por lote para el registro de los CPPs y la adquisición de datos espectrales para cada sistema se resumen en la Tabla 2.1.

Tabla 2.1. Resumen de la identificación (ID) de cada lote de proceso analizado, su duración y el número total de muestras procesadas por lote para el registro de los CPPs y la adquisición de datos espectrales para los bioprocesos analizados.

Sistema	Escala	ID	Duración total (días)	Muestras para el monitoreo de CPPs[a]	Muestras para la generación de EEMs
A	Planta piloto (mediana)	A1	15	14	13
		A2	15	12	12
		A3	13	12	12
		A4	16	14	14
B	Industrial (grande)	B1	30	31	24 (día 1 a 24)
		B2	30	31	24 (día 1 a 24)

[a]En algunos casos, se contempla la muestra tomada al momento de la inoculación del biorreactor (día 0 del proceso).

2.3.3. Monitoreo de variables de proceso (CPPs) mediante técnicas analíticas univariadas de referencia

Tal como se mencionó anteriormente, los principales CPPs registrados para este estudio fueron la densidad de células viables y totales, el porcentaje de viabilidad, el título o concentración total de etanercept y las concentraciones de los metabolitos glucosa y lactato, los cuales representan metabolitos típicos indicadores de la calidad de crecimiento del cultivo. En particular, la disponibilidad de glucosa permite controlar la velocidad de crecimiento celular, mientras que la concentración de lactato debe ser monitoreada, ya que su acumulación resultan desfavorables tanto para el cultivo como para la calidad del producto [66].

Inmediatamente luego de la toma de muestra, se determinó la densidad de células viables y totales mediante el recuento en cámara de Neubauer y tinción con colorante vital (Tripan Blue). Posteriormente, las muestras fueron clarificadas mediante centrifugación para eliminar células y *debris* celular. Se tomaron alícuotas de los sobrenadantes para la cuantificación de los metabolitos glucosa y lactato, mediante el uso de técnicas colorimétricas (test rápidos MERCK). Asimismo, dos alícuotas de sobrenadante clarificado fueron almacenadas por separado a –15 °C para la posterior determinación de la concentración de etanercept y para la adquisición de EEMs.

El etanercept es una proteína que es secretada al entorno extracelular y por lo tanto, puede ser directamente detectada en sobrenadantes de cultivo mediante un método HPLC con interacción hidrofóbica. El método cromatográfico utilizado rutinariamente por la empresa consiste en una adaptación del protocolo compendiado establecido en las referencias [78,79]. Todos los experimentos se llevaron a cabo en un equipo HPLC Waters Alliance (Waters, Milford, EEUU) con un detector UV en línea. El método consistió en una cromatografía de interacción hidrofóbica, utilizando un gradiente reverso de sulfato de amonio en buffer fosfato de sodio pH 7,0. Se utilizó una columna Tosoh TSKgel butyl-NPR (35 x 4,6 mm). La fase móvil binaria consistió en buffer fosfato de sodio 0,1 M y pH 7,0 – sulfato de amonio 1.8 M (A) y fosfato de sodio 0,1 M pH 7,0 (B), circulando a una velocidad de 1 mL min^{-1}. El análisis comenzó con un 100% de A y 0% de B (2 minutos), seguido de un gradiente lineal hasta alcanzar secuencialmente 54% de A (8 min) y 14% A (28 minutos). La temperatura de la columna se mantuvo en 25 °C. El analito fue monitoreado mediante absorbancia UV a 214 nm. Se utilizó el software Empower2 (Waters, Milford, EEUU) para el control del instrumento, la adquisición y análisis de datos cromatográficos.

La concentración de etanercept fue calculada en mg L^{-1} mediante el uso de una curva de calibrado univariada estándar, preparada a partir de un material de referencia interno (patrón secundario obtenido a partir de un producto de alta pureza valorado). Es

importante mencionar que este protocolo se encuentra sometido periódicamente a una validación interna mediante un programa de control de calidad, diseñado de acuerdo a las recomendaciones establecidas en las referencias [78,79].

2.3.4. Generación de EEMs y preprocesamiento

La elección de la técnica espectral de fluorescencia radica esencialmente en el hecho de que se trata de una metodología robusta, que permite la obtención de datos de segundo orden bilineales (EEMs) de una manera accesible y relativamente rápida. Asimismo, constituye una técnica ampliamente reportada en la bibliografía para desarrollos de PAT debido a que las muestras de cultivos celulares (y de bioprocesos en general) exhiben una fluorescencia intrínseca que se atribuye a la presencia de fluoróforos biológicos tales como proteínas, vitaminas y cofactores enzimáticos.

Luego del descongelamiento de las muestras de fermentación recolectadas, las mismas fueron equilibradas térmicamente a temperatura ambiente durante varios minutos antes de su procesamiento. Las EEMs fueron registradas utilizando alícuotas de cada muestra de fermentación, sin implementar ningún paso de dilución. Asimismo, para el proceso de descongelamiento y lectura, el total de las muestras recolectadas fueron procesadas de manera aleatoria durante un período total de seis semanas [80].

Todos los experimentos se llevaron a cabo en una cubeta de cuarzo de 1 cm de paso óptico y a temperatura ambiente (25 °C), utilizando el equipo descripto en la Sección 1.3.2 del Capítulo 1. En todos los casos se registraron las emisiones en el rango espectral de 250,0 a 600,0 nm, con un intervalo de 0,5 nm, excitando las muestras en el rango espectral entre 225,0 y 495,0 nm, cada 5 nm, y utilizando una velocidad de barrido de 1000 nm min^{-1}. Los anchos de rendija de los monocromadores de excitación y emisión se fijaron ambos en 5 nm y el voltaje del detector (PMT) en 750 V. Es importante mencionar que los parámetros instrumentales seleccionados para la generación de los datos se basaron en experiencia previa del grupo de investigación y en la bibliografía afín a la temática del trabajo (por ejemplo Ref. [34]), priorizando el barrido de la mayor cantidad de información espectral posible y minimizando el tiempo de adquisición.

De esta manera, se generaron EEMs de 55×700 puntos, para los modos de excitación y emisión, respectivamente. El preprocesamiento de las EEMs para la corrección digital de las señales de dispersión de Ry y Rm se llevó a cabo mediante el uso de la GUI EEM_corr, a través de una combinación de las estrategias M4 [4] y M6, cuyos fundamentos fueron desarrollados en el capítulo anterior.

2.3.5. Implementación de los métodos quimiométricos PCA, MCR-ALS y PLS-DA

En primer lugar, se efectuaron análisis exploratorios de las variables de proceso de manera univariada. Por otra parte, se utilizó el método PCA con los objetivos de realizar la inspección de valores atípicos (*outliers*) desde un enfoque multivariado y el análisis de agrupamiento (*clustering*), tanto para los datos recabados de CPPs como para los datos de fluorescencia (EEMs). Debido a que los datos de CPPs constituyen datos de orden cero (ya que son medidos a través de técnicas univariadas), las variables de proceso medidas para los diferentes conjuntos de muestras (lotes) fueron dispuestos en un arreglo matricial, ubicando muestras y variables en filas y columnas, respectivamente. Por su parte, las EEMs fueron desdobladas en vectores fila que, a su vez, fueron luego dispuestos en un arreglo matricial. En ambos casos, en función del tipo de análisis, se construyeron diferentes arreglos, contemplando el modelado de lotes individuales o conjuntos. Debido a la formulación del modelo PCA (Ec. 2.1), todos los datos requirieron de ser centrados en la media (a cada muestra se le resta la media de cada variable predictora)[16]. Además, en el caso particular de los datos de CPPs, en virtud de que estos son medidos a través de diferentes técnicas y los valores resultan en unidades y escalas de distinta magnitud, los datos fueron escalados (el valor de cada muestra es dividido por la desviación estándar correspondiente a cada variable)[17]. En todos los casos, la determinación del número óptimo de PCs se basó en la inspección de la importancia relativa de cada PC y su contribución a la varianza total del modelo. Siguiendo el principio de parsimonia, la selección se basó en elegir un número de PCs pequeño que maximice la varianza explicada. Asimismo, para los análisis de agrupamiento (y también para los métodos de clasificación), la selección se basó en la minimización del RMSECV.

Tal como se mencionó anteriormente, existen diferentes procedimientos para llevar a delante una CV. En líneas generales, las técnicas pueden ser de dos tipos: dejando una muestra fuera (*leave-one-out,* LOO) o realizando particiones por bloque [81]. La estrategia de LOO consiste en dejar sólo una muestra fuera del conjunto de entrenamiento en cada ciclo de la CV, de manera que se garantiza que todas las muestras serán predichas durante el procedimiento. Este método resulta de utilidad cuando el tamaño muestral del conjunto de entrenamiento es pequeño. Por otra parte, en el caso de las particiones por bloque, el analista debe determinar *a priori* el número de particiones y la cantidad de muestras por partición que se deben apartar del conjunto de entrenamiento. A su vez, la obtención de bloques de muestras puede realizarse en

[16] A cada elemento de la *j*-ésima columna de la matriz de datos se resta la media de dicha columna, de manera que todas las variables poseen media igual a cero.
[17] Cada elemento de la *j*-ésima columna de la matriz de datos es dividido por la desviación estándar de dicha columna, de manera que todas las variables poseen desvío igual a uno.

n particiones aleatorias (*leave-k-out*) o en *k* bloques fijos (metodología denominada habitualmente como *k-folding*). Este último procedimiento admite dos variantes: seleccionar los índices de las muestras de manera continua en el conjunto de entrenamiento (*continuous blocks*) o bien, elegir muestras cuyos índices se encuentran a una determinada distancia fija (procedimiento conocido como "ciegos venecianos" o *venetian blinds*) [82]. En este trabajo se empleó para los análisis de agrupamiento y clasificación, la técnica de ciegos venecianos para la optimización de modelos PCA y PLS, generando en todos los casos entre 5 y 10 particiones. En este sentido, esta técnica de CV es conveniente cuando el conjunto de datos de entrenamiento se encuentra ordenado respecto al grupo o clase a la que pertenece cada muestra. Se evita de esta manera que, en un dado ciclo de CV, una clase o grupo quede subrepresentado. Por otra parte, para los procedimientos de CV para modelos de calibración con PLS, se utilizaron alternativamente las técnicas de *k-folding* en bloques continuos y de LOO.

Una vez obtenidos los modelos PCA, la exploración de valores atípicos se basó en la inspección de los *scores* de las muestras en las primeras dos PCs y en el diagnóstico de los residuos. En particular, el PCA se caracteriza por calcular dos tipos de residuos asociados al modelo. En primer lugar, la suma de los cuadrados de los residuos de cada muestra constituye lo que habitualmente se denomina, en el contexto del PCA, como residuo Q. En este sentido, una muestras que presenta un residuo Q significativamente grande con respecto al promedio, quiere decir que no es adecuadamente explicada por el modelo. En segundo lugar, se define como residuo T^2 de Hotelling de la i-ésima muestra en un modelo PCA de A componentes a aquel dado por [63]:

$$T^2 = \frac{\mathbf{t}_i^{\mathrm{T}} \left(\mathbf{T}_A^{\mathrm{T}} \mathbf{T}_A\right)^{-1} \mathbf{t}_i}{I - 1} \tag{3.15}$$

Una muestra con un residuo T^2 significativamente grande constituye un valor atípico que se aleja del centroide del conjunto de datos de entrenamiento, pero que, a diferencia del residuo Q, la misma se encuentra en el mismo hiperplano del modelo. De esta manera, una forma conveniente de inspeccionar la presencia de *outliers* verdaderos consiste en graficar los residuos Q como función de los T^2, obteniendo lo que se conoce como diagrama de influencia de PCA. Asimismo, tanto para los *scores* como para los residuos, es frecuente calcular límites de confianza uni- o multiviariados para la comparación de valores, asumiendo que estos se distribuyen de manera normal.

Por otra parte, una vez descartados los valores atípicos, se procedió a utilizar las matrices de *scores* de cada modelo PCA para realizar análisis de agrupamiento. Tanto para el caso de datos de CPPs como espectrales, se graficaron los *scores* de las muestras en las PCs más significativas (2 o 3, según el caso). El análisis de

agrupamiento, asociado a los datos de viabilidad en función del tiempo para los diferentes lotes analizados, permitió identificar muestras de alta y baja viabilidad. De esta manera, estos resultados motivaron el desarrollo de un modelo predictivo basado en el método PLS-DA con datos espectrales. Debido al número insuficiente de muestras de los lotes del Sistema B, su estudio finalizó en la etapa exploratoria y la construcción del modelo de clasificación se realizó exclusivamente para los lotes del Sistema A. En este sentido, las muestras fueron asignadas a una de dos clases bien definidas, de acuerdo a los resultados del análisis de agrupamiento con datos espectrales.

Para la obtención del modelo de clasificación se consideraron dos grupos de muestras: de entrenamiento (*training*) y de validación (test). El conjunto de validación constituye un grupo de muestras conocidas que no son utilizadas para el entrenamiento del modelo. Estas muestras se utilizan para efectuar predicciones con un modelo ya entrenado y, de esta manera, evaluar su capacidad predictiva.

En primer lugar, se ajustó un modelo PLS-DA con las muestras pertenecientes a los lotes A1 a A3 (conjunto de entrenamiento), mientras que las muestras del lote A4 se utilizaron como conjunto de validación. Así, el tamaño muestral de los conjuntos de entrenamiento y validación fue de 75% y 25% del total de mediciones, respectivamente.

El conjunto de entrenamiento se sometió a una CV mediante la técnica de ciegos venecianos con 10 particiones. Posteriormente, se efectuaron las predicciones con el conjunto de validación. A partir de los resultados de las predicciones de muestras de entrenamiento, CV y validación externa, se construyeron las matrices de confusión respectivas para el posterior cálculo de los índices de clasificación descriptos más arriba.

Finalmente, el modelado quimiométrico de datos de primer orden (CPPs y EEMs desdobladas) se complementó con un análisis cualitativo de los datos de fluorescencia de segundo orden mediante el método MCR-ALS. Para ello, las EEMs obtenidas para diferentes grupos de muestras fueron dispuestas en un arreglo aumentado $\mathbf{X}_{\mathrm{aum}}$ en columnas, estableciendo la emisión y la excitación como modos no aumentado y aumentado, respectivamente. Se efectuó la determinación del número de componentes mediante PCA del arreglo $\mathbf{X}_{\mathrm{aum}}$. El algoritmo de optimización ALS se inicializó mediante el método de las variables más puras y su convergencia se monitoreó a partir del cálculo del $\%\mathrm{LOF}$. En todos los casos, la única restricción impuesta sobre el ALS fue la no negatividad en ambos modos espectrales. Los resultados del modelado MCR-ALS fueron utilizados para la identificación de posibles fluoróforos presentes en el medio de cultivo y para una caracterización preliminar de la variación de la señal espectral en función del tiempo de proceso.

En la Figura 2.1 se resumen de manera gráfica los métodos quimiométricos implementados en esta etapa del trabajo.

2.3.6. Software

El ensamblado de datos crudos, manejo, visualización de datos y ejecución de algoritmos específicos se llevó a cabo en MATLAB R2017b, utilizando para ellos rutinas caseras e interfaces gráficas de usuario. En particular, PCA y PLS-DA fueron respectivamente ejecutados a partir de las interfaces de MATLAB pca_gui [82] y classification_gui [64], ambas de acceso libre. Además, se utilizó la interfaz PLS Toolbox 8.7.1 (2019), en su versión de prueba gratuita disponible a través del enlace http://www.eigenvector.com (Eigenvector Research, Inc., Manson, WA USA 98831). El método MCR-ALS fue llevado a cabo en la interfaz de MATLAB libre MVC2 [83].

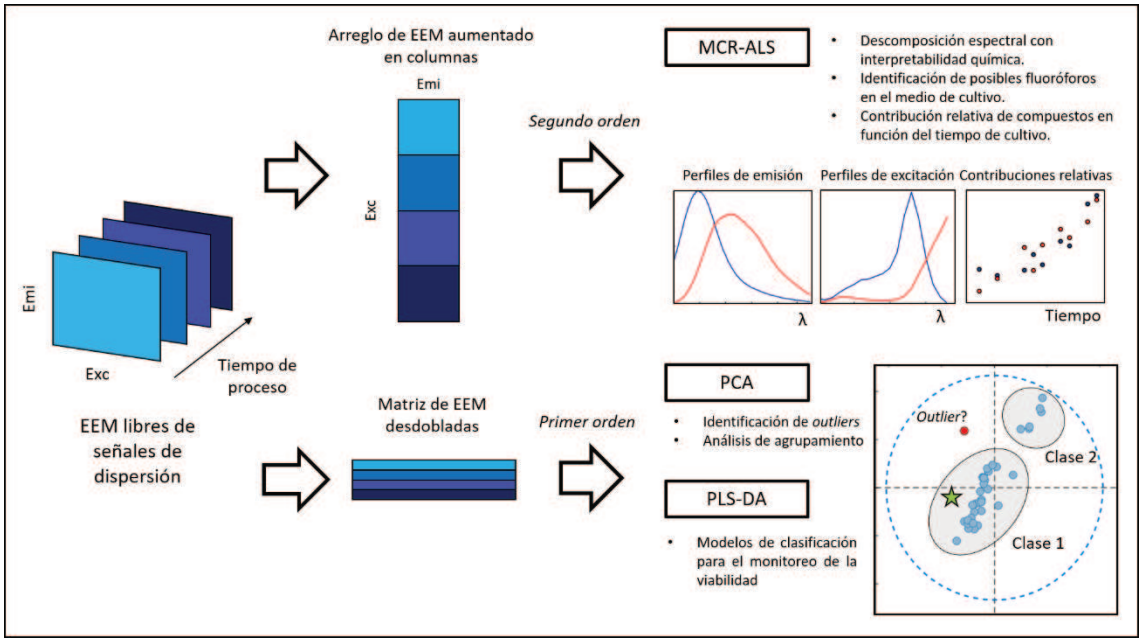


Figura 2.1. Representación de los métodos quimiométricos cualitativos implementados en esta etapa del trabajo. **Arriba:** descomposición bilineal por MCR-ALS de un arreglo aumentado en columnas a partir de EEMs (enfoque de modelado de segundo orden). Los resultados de MCR-ALS fueron utilizados para la identificación de posibles fluoróforos presentes en el medio de cultivo y una caracterización preliminar de la variación de la señal espectral en función del tiempo de proceso. **Abajo:** con las EEMs desdobladas se construyeron arreglos matriciales que fueron modelados mediante PCA y PLS-DA (enfoque de modelado de primer orden), para los objetivos de inspección de valores atípicos (*outliers*), análisis de agrupamiento y elaboración de modelos de clasificación.

2.4. Resultados y discusión

2.4.1. Análisis exploratorio de las variables de proceso (CPPs)

La inspección visual de los datos de variables de proceso permitió caracterizar la *performance* general de los bioprocesos estudiados, en ambas escalas (Sistemas A y B). En este contexto, vale la pena destacar que no se detectaron valores atípicos en los datos de CPPs mediante ninguna de las técnicas de análisis empleadas (exploración univariada y modelos PCA).

El número de células viables, la concentración de etanercept, el porcentaje de viabilidad celular y las concentraciones de glucosa y lactato como función del tiempo de cultivo para los cinco lotes estudiados se resumen gráficamente en la Figura 2.2. Como se mencionó anteriormente, estos parámetros constituyen las variables más importantes en términos de calidad y monitoreo del proceso en su etapa de *upstream*. En todos los casos, se observa la evolución típica de densidad de células viables y de la concentración de producto a lo largo de las fermentaciones. Las disimilitudes observadas entre cada sistema son esperables, en virtud de las diferencias entre las condiciones experimentales empleadas en cada caso.

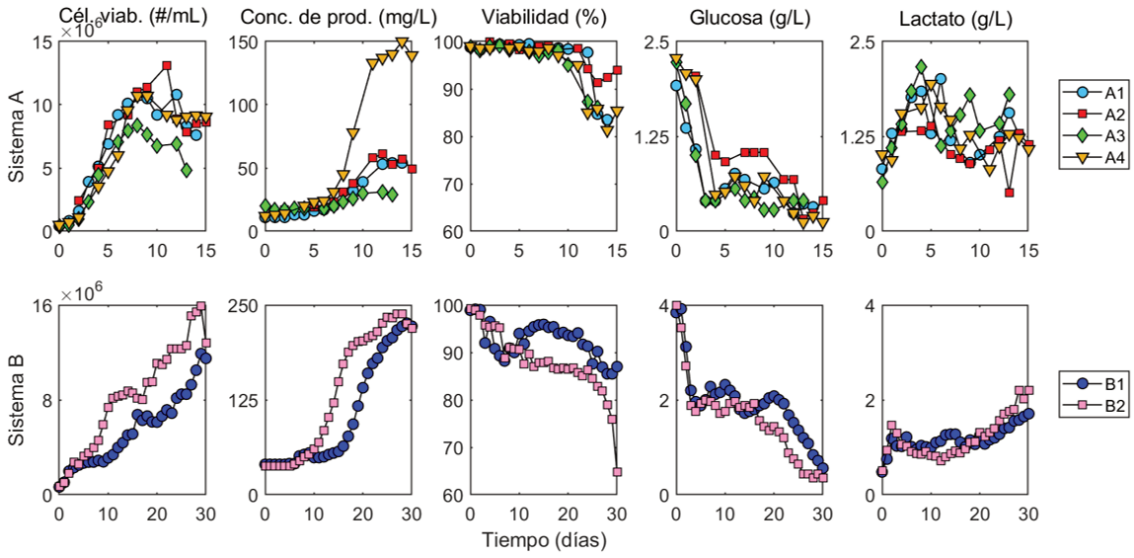


Figura 2.2. Evolución de las principales variables de proceso (CPPs): densidad de células viables, concentración de producto, porcentaje de viabilidad celular, concentración de glucosa y concentración de lactato, en función del tiempo de cultivo para cada uno de los procesos analizados del Sistema A (arriba, lotes A1 a A4) y del Sistema B (abajo, lotes B1 y B2).

En términos generales, se observa que la etapa de estabilización de las células (fase *lag*) es relativamente breve e incluso, inexistente en algunos casos. Las células experimentan un crecimiento exponencial y alcanzan densidades celulares elevadas, lo cual es favorecido por el régimen de perfusión. Durante esta fase de crecimiento, es

crítico el monitoreo de los metabolitos glucosa, que representa la principal fuente de carbono, y lactato, cuya acumulación excesiva resulta tóxica para el crecimiento celular. En este sentido, el cultivo es controlado mediante un conjunto de estrategias de proceso que esencialmente involucran el sistema de aireación/agitación, el control de la velocidad de perfusión y la aplicación de rampas de temperatura. Si bien esto ocasiona un comportamiento fluctuante de las concentraciones de glucosa y lactato, en líneas generales, la tendencia global en todos los casos es que la glucosa disminuya y el lactato se acumule conforme transcurre el proceso. Una vez que se alcanza la densidad celular óptima, la manipulación de los parámetros operativos del reactor apunta a inducir una fase de crecimiento estacionaria, caracterizada por una concentración de células viables aproximadamente constante. En este sentido, puede distinguirse que, para los lotes del Sistema A, el crecimiento celular se estaciona alrededor del día 6 de fermentación, mientras que para el Sistema B, este comportamiento ocurre entre los días 10 y 15. Además, para los lotes B1 y B2 también se observa que luego de la fase estacionaria, el crecimiento celular se acelera nuevamente.

Con respecto al producto, se puede ver que, durante la etapa de crecimiento exponencial, la concentración de etanercept oscila alrededor de un valor basal, mientras que luego de que la densidad de células viables se estaciona, la concentración de proteína recombinante se dispara debido a que las células redirigen fuertemente la energía metabólica a los procesos de síntesis, y en menor medida, a la duplicación celular. Para los procesos estudiados, esto ocurre aproximadamente alrededor del día 8 de cultivo para el Sistema A y del día 12-15, en el caso del Sistema B. Es importante destacar en este punto que las diferencias notables en productividad que se observan entre las dos escalas de producción se pueden atribuir principalmente a las diferencias respecto de la geometría del biorreactor, el número de pasajes del clon celular y la formulación del medio que fueron utilizados en cada caso. A su vez, en el caso del Sistema A, se observa que, en particular, el lote A4 presenta una productividad tres veces mayor en relación a los lotes A1-A3. Esto se debe a que para el lote A4 se utilizó un clon celular con un número menor de pasajes. Sin embargo, se puede evidenciar en virtud de las demás variables medidas que el lote A4 es comparable con los demás lotes, en términos de crecimiento celular y metabolismo, por lo que el mismo es incluido en este estudio, ya que el objetivo principal se centra en el monitoreo de la viabilidad celular.

Por otro lado, el porcentaje de viabilidad constituye una de las variables críticas que se toma como criterio para decidir sobre el fin del bioproceso. Para la fermentación de células productoras de etanercept, el cultivo se interrumpe cuando la viabilidad cae por debajo del 80-85%. Como ha sido descripto en la bibliografía, a pesar de que los

niveles de productividad permanecen altos hasta el final del proceso, una caída fuerte en la viabilidad celular puede afectar considerablemente la etapa de *downstream* y/o los CQAs del producto [84]. No obstante, como se observa en la Figura 2.2, el momento exacto en que la viabilidad comienza a decaer resulta bastante impredecible. Por ejemplo, para el caso del Sistema A, los cultivos experimentan una caída abrupta alrededor del día 10, aunque el día exacto no es reproducible y, por lo tanto, no resulta fácil anticiparlo. Por otra parte, se observan diferencias notables entre las curvas de viabilidad de los lotes del Sistema B. Ambos lotes han sido llevados a cabo bajo idénticas condiciones experimentales y con la misma duración total (30 días). Sin embargo, el lote B2 muestra una caída importante de la viabilidad alrededor del día 25, mientras que en el caso del lote B1, este efecto es menos notorio.

2.4.2. Análisis de agrupamiento mediante PCA a partir de datos de CPPs

Una vez realizada una inspección visual de las variables de proceso de manera individual, se procedió a realizar el análisis de agrupamiento. Para este fin, los datos recabados de los cinco CPPs fueron utilizados para construir modelos PCA para cada lote individual. Tomando como ejemplo los datos del lote A1, en la Figura 2.3 se muestran las gráficas más representativas que se obtienen en un modelo PCA, en relación a su optimización y uso en la inspección de valores atípicos. En este sentido, se muestran el PVA y autovalores en función del número de PCs (Figura 2.3.A), la variación del RMSECV en función del número de PCs (Figura 2.3.B), el diagrama de influencia (Figura 2.3.C), y los diagramas de *loadings* y de *scores* en las PCs significativas (Figura 2.3.D-F y G-I, respectivamente).

Debido a que el número de variables experimentales es relativamente pequeño y que las mismas no muestran un alto grado de correlación, son necesarios al menos 3 PCs para capturar una varianza explicada acumulada mayor al 90% (Figura 2.3.A). Asimismo, la variación del RMSECV indica un óptimo de PCs en 3, ya que este muestra un mínimo global para ese número (Figura 2.3.B). Por otra parte, el diagrama de influencia permite ver que todas las muestras presentan residuos Q y T^2 de Hotelling que caen dentro de los límites de confianza calculados para ambos tipos de residuos. Sólo una muestra aparece con un residuo Q levemente alejado de la nube de puntos principal, pero debido a que dicho valor no presenta un residuo excesivamente grande, el dato correspondiente no fue descartado del modelo. En este sentido, la inspección de las curvas que muestran la evolución de las variables de proceso presentan los comportamientos típicos y los valores de las variables medidas experimentalmente se encuentran dentro de los rangos esperados, por lo que no se evidencia la presencia de *outliers* experimentales. Podemos decir entonces que, para este sistema, todas las

muestras son adecuadamente modeladas por el modelo PCA. Esto también es coherente con el diagrama de *scores* (todas las muestras se encuentran dentro de la elipse de confianza en los planos formados por las PCs significativas).

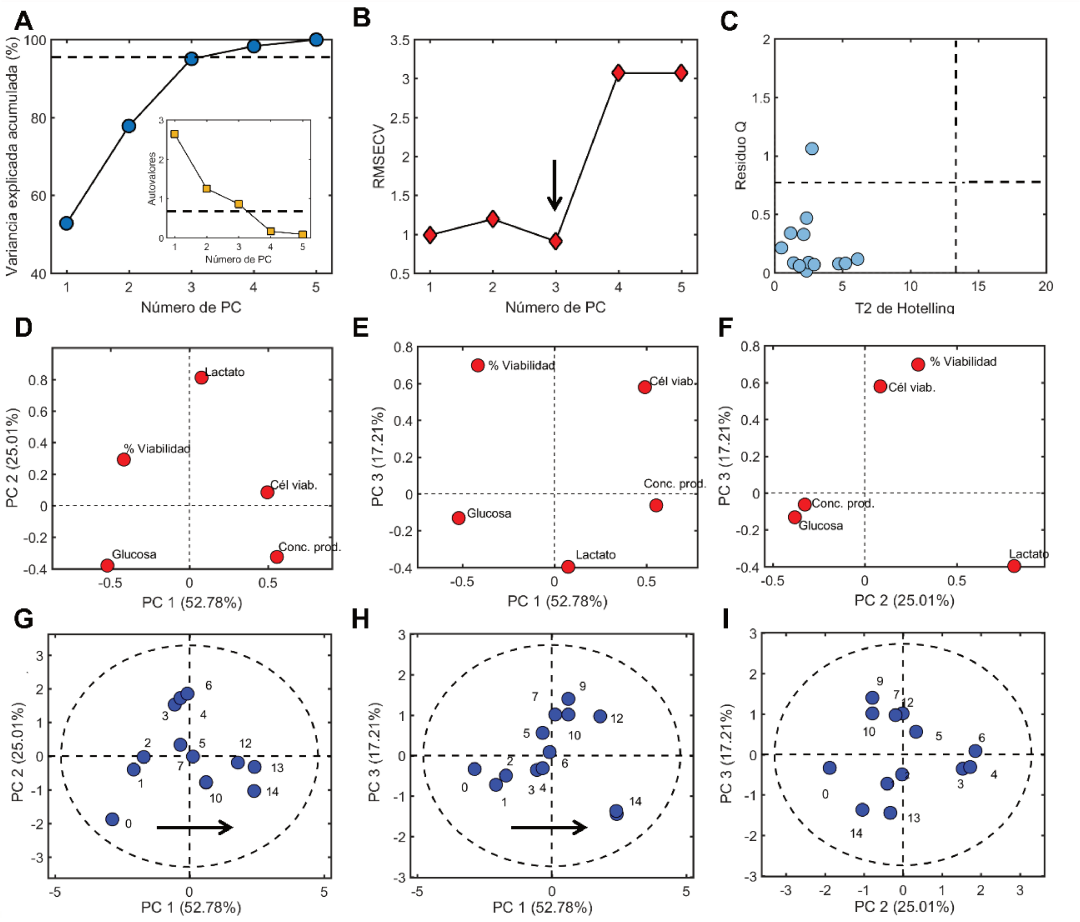


Figura 2.3. Salidas principales del modelo PCA para los datos de CPPs del lote A1: **A.** PVA y autovalores (figura incrustada) en función del número de PCs (en cada caso, las líneas discontinuas indican hasta qué PC se obtienen variaciones significativas en la varianza capturada y los autovalores significativos del modelo); **B.** RMSECV en función del número de PCs (la flecha indica el mínimo global); **C.** diagrama de influencia (las líneas discontinuas representan los límites de confianza estimados con un 0,05 de nivel de significancia); **D.** *loadings* del PC2 vs PC1; **E.** *loadings* del PC3 vs PC1; **F.** *loadings* del PC3 vs PC2; **G.** *scores* en PC2 vs PC1; **H.** *scores* en PC3 vs PC1; **I.** *scores* en PC3 vs PC2. En las gráficas G-I, las etiquetas representan el día de cultivo, mientras que la flecha indica el sentido de evolución del tiempo de proceso. Además, las líneas discontinuas representan los límites de confianza bivariados calculados con un nivel de significancia de 0,05.

Finalmente, los diagramas de *loadings* permiten, por un lado, evidenciar correlaciones entre variables y, por otro, explicar la posición de las muestras respecto de sus *scores.* En este sentido, se observa, por ejemplo, que las variables células viables y concentración de producto están correlacionadas de manera directa en los planos formados por PC1 y PC2 (Figura 2.3.D) y PC1 y PC3 (Figura 2.3.F), mientras que dichas variables se correlacionan de manera inversa con la viabilidad y la

concentración de glucosa. Asimismo, observando la evolución de los *scores* de las muestras en función del tiempo de cultivo (Figura 2.3.G y H) y su relación con las variables en los diagramas de *loadings*, puede corroborarse que hay una tendencia de las muestras a desplazarse en el sentido positivo de la PC1, que se caracteriza por una mayor densidad celular y concentración de producto. Todas estas observaciones son consistentes con el comportamiento esperado para las variables de proceso, lo cual también valida el modelo PCA.

El análisis de los resultados arrojados por los modelos construidos para cada uno de los lotes se realizó de la misma manera que el descripto en el párrafo anterior, obteniendo en todos los casos, resultados equivalentes. En este sentido, todos los modelos resultaron satisfactorios y no se detectaron valores atípicos en los datos de CPPs. Además, se puso particular énfasis en los diagramas de *scores.* Debido a que en todos los modelos PCA calculados las tres primeras PCs resultaron ser las más significativas, se procedió a graficar los *scores* de las muestras en las tres primeras PCs, para efectuar un análisis de agrupamiento. Los resultados se muestran en la Figura 2.4.

Debido a que el foco de este estudio estuvo puesto en el monitoreo de la viabilidad celular, la cual tiende a caer conforme transcurre la fermentación, el análisis de agrupamiento se orientó a identificar grupos de muestras que respondan a la evolución temporal de cada cultivo (es decir, no se analizaron aquellos grupos de muestras que pudieran observarse respecto a otros criterios). En este sentido, la inspección de los *scores* en la Figura 2.4 revela que, en el caso del Sistema A, las muestras tienden a agruparse hasta el día 10 aproximadamente y, a partir del momento en que se observa la caída de viabilidad, aparece un segundo grupo que corresponde a muestras en las que la viabilidad celular ya se encuentra en 90% o por debajo. Asimismo, resulta interesante observar que, para los lotes del Sistema B, los resultados obtenidos mediante PCA son consistentes con las observaciones realizadas a la luz de las variables de proceso. En particular, para el lote B2 se observan claramente los grupos de alta y baja viabilidad (luego del día 27), mientras que esto no ocurre para el caso del lote B1. Finalmente, en aquellos lotes en donde se produce la caída de viabilidad, se observa que, en general, las muestras tienden a agruparse en el sentido positivo de la PC1.

Hasta aquí se ha hecho hincapié en una descripción y exploración de las variables de proceso desde los enfoques uni- y multivariado. En las siguientes secciones, se demostrará cómo las huellas espectrales que son extraídas a partir del modelado quimiométrico de los datos de fluorescencia multidimensionales (EEMs), permiten establecer un nuevo criterio en relación al seguimiento de la viabilidad celular en los cultivos.

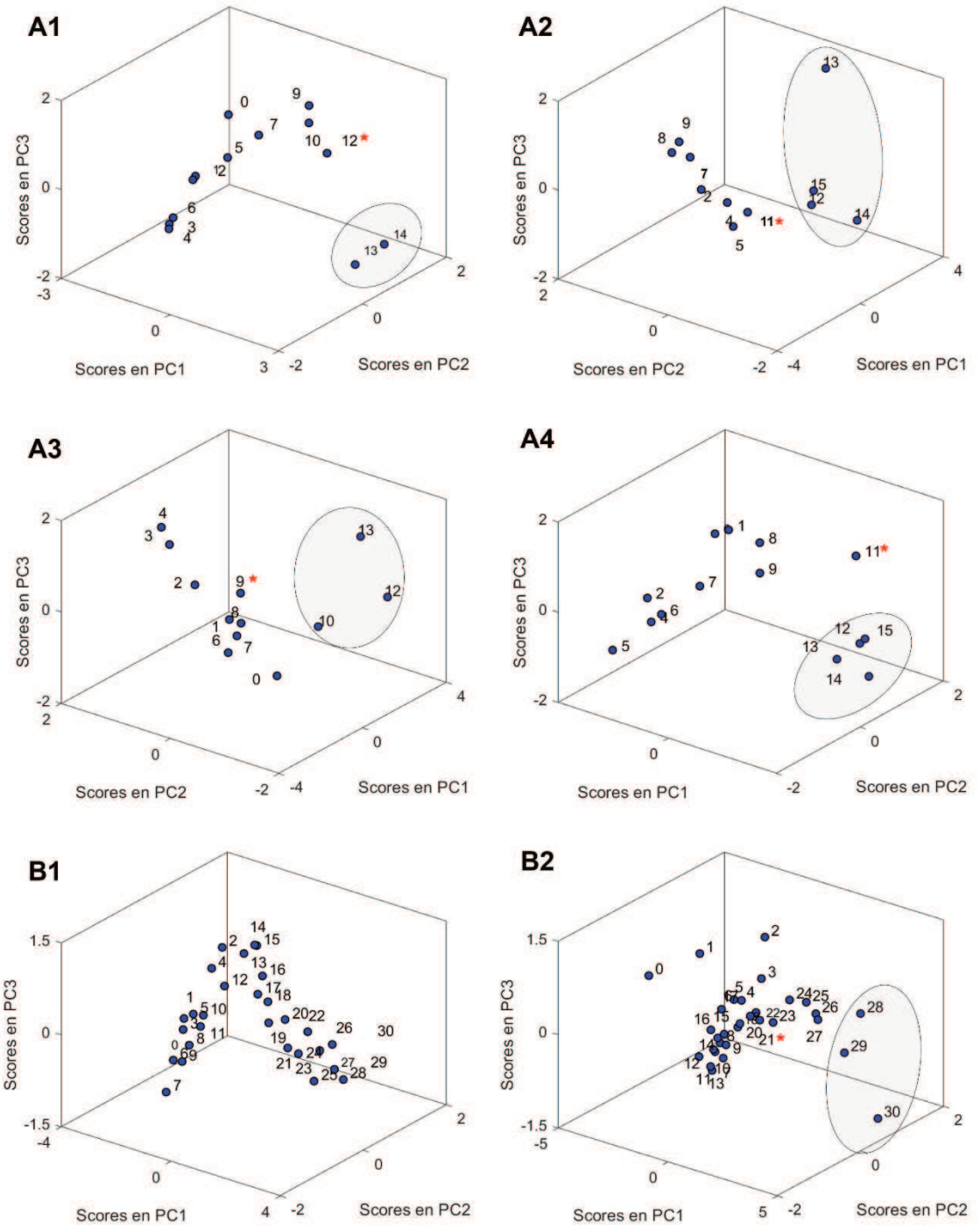


Figura 2.4. Diagramas de *scores* de las muestras de fermentación en las tres primeras PCs para los modelos PCA obtenidos con cada uno de los lotes analizados del Sistema A (A1-A4) y Sistema B (B1-B2). En todos los casos, las tres primeras PCs explican una varianza mayor al 90%. Las etiquetas representan el día de cultivo. Las elipses grises indican el agrupamiento de las muestras de baja viabilidad. El asterisco indica la muestra que corresponde al último día de alta viabilidad según los datos de CPPs.

2.4.3. Análisis exploratorio de los datos de fluorescencia de segundo orden mediante MCR-ALS

En primer lugar, luego de la adquisición de los datos de fluorescencia, se procedió a la importación y ensamblado de las EEMs, para luego efectuar la corrección digital de las señales de *scattering.* En la Figura 2.5, se muestra a modo de ejemplo, la EEM obtenida a partir de la muestra correspondiente al día 1 del lote B1, antes y luego del preprocesamiento (Figura 2.5.A y B, respectivamente).

En la Figura 2.5.B puede observarse que la señal de fluorescencia se distribuye en dos regiones principales, lo cual es consistente con lo descripto en la bibliografía para este tipo de sistemas biológicos. La primera de ellas (referida como R1) se encuentra a longitudes de onda de excitación menores y se origina por los aminoácidos aromáticos (triptófano, tirosina y fenilalanina). Para este sistema en particular, se sabe que estos fluoróforos están presentes en el medio de cultivo como aminoácidos libres y como constituyentes de proteínas. En general en todas las muestras, esta región presenta una mayor intensidad de emisión. Por otra parte, la señal de la otra región (R2) es producida por una serie de fluoróforos biológicos diversos y menos abundantes, tales como vitaminas, coenzimas y cofactores [80]. Debido a la cantidad de especies espectralmente activas en estas muestras y al alto grado de solapamiento de las señales, el uso de algoritmos quimiométricos de segundo orden resultan de gran utilidad para la descomposición de las EEMs y su análisis cualitativo.

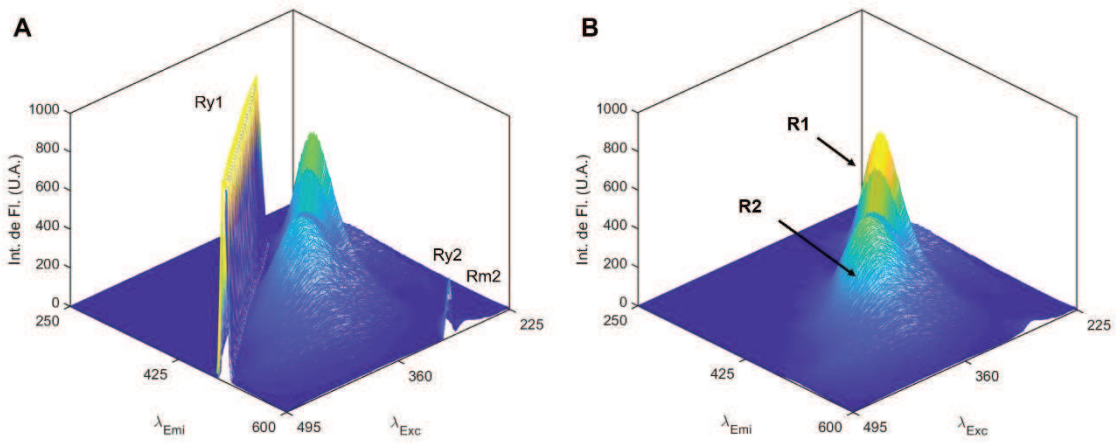


Figura 2.5. EEM de fluorescencia típica, adquirida a partir de la muestra del día 1 del lote B1. **A.** EEM antes del pre-procesamiento, en la que se observan los patrones diagonales característicos de las señales de dispersión de Ry y Rm (primer y segundo orden); **B.** EEM luego de la corrección digital de *scattering.* R1: región espectral que corresponde a la señal generada por aminoácidos aromáticos; R2: región espectral que corresponde a la señal generada por vitaminas y cofactores.

Tal como se mencionó anteriormente, un conjunto de EEMs libre de señales de dispersión permite la obtención de arreglos trilineales mediante el apilamiento de matrices (arreglo de tres vías). En este sentido, una estrategia conveniente para el tratamiento de este tipo de datos son los modelos que se basan en la descomposición trilineal. A diferencia de los métodos bilineales tales como MCR-ALS, los algoritmos trilineales tienen la ventaja de ser menos afectados por fenómenos de ambigüedad, con lo cual suelen presentar unicidad de solución. Entre ellos, el método más frecuentemente reportado en la literatura es el PARAFAC [85]. No obstante, estos métodos resultan comparativamente modelos más rígidos respecto de los supuestos acerca de la estructura de los datos y cualquier leve desvío afecta fuertemente su resolución.

Si bien para los datos adquiridos en este trabajo no se evidenciaron desvíos importantes del supuesto de trilinealidad, la complejidad de las muestras sumada a la considerable diferencia en la cantidad de puntos en una y otra dimensión de cada una de las EEMs (ocasionada por una cuestión de índole instrumental), la capacidad de resolución de PARAFAC se vio fuertemente limitada, con lo cual, no se lo pudo utilizar como método para el análisis de los datos de segundo orden. Por lo tanto, se optó por una estrategia basada en la descomposición bilineal mediante el método MCR-ALS, generando arreglos de EEMs aumentados en columnas, en el sentido de las excitaciones. De esta manera, se fijó a la emisión como modo no aumentado, ya que fue el de mayor resolución (mayor cantidad de puntos).

En todos los casos, se obtuvieron modelos MCR-ALS de cinco componentes con un PVA del 99% y una falta de ajuste del orden del ruido instrumental (estimada a partir del residuo de PCA del arreglo de datos). Como ejemplos representativos de los Sistemas A y B, en la Figura 2.6 se muestran las salidas del modelo MCR-ALS para los lotes A2 y B2. En particular, se incluyen los perfiles espectrales en los modos de excitación (Figura 2.6.A) y emisión (Figura 2.6.B), los cuales son comunes a todos los modelos, y las contribuciones relativas y volumen de perfusión en función del tiempo de proceso de los componentes modelados para el lote A2 (Figura 2.6.C y E) y B2 (Figura 2.6.D y F).

Por un lado, tal como se describió anteriormente, la descomposición mediante MCR-ALS con restricciones permite descomponer la señal instrumental en un conjunto de contribuciones individuales y debido a la interpretabilidad química de este modelo, es posible asociar los perfiles espectrales con espectros reales de compuestos puros que pudieran estar presentes en el medio de cultivo. Si bien no se realizó una comprobación rigurosa de la identidad de cada componente (lo cual requiere de la aplicación de técnicas analíticas adicionales), se procedió a efectuar una identificación

preliminar de posibles compuestos químicos, mediante la comparación directa de los perfiles modelados con información disponible en bases de datos de fluorescencia de analitos puros conocidos. En este sentido se tuvieron en cuenta los máximos de excitación y emisión de cada compuesto y, en caso de encontrarse disponibles, también se tuvo en cuenta la forma de los espectros. Los resultados se resumen en la Tabla 2.2.

El componente 1 (azul) exhibe un máximo de excitación a 290 nm y un máximo de emisión a 367 nm. La posición del máximo y la forma espectral para el perfil de emisión están en concordancia con información reportada para el triptófano [86]. No obstante, el perfil que se observa en el modo excitación se encuentra notablemente distorsionado. Durante el desarrollo del trabajo, esto dio la pauta de que tal deformación podía estar relacionado con diferentes fenómenos fotoquímicos que ocurren debido a las interacciones con la tirosina. En este sentido y de acuerdo a lo descripto en bibliografía, las señales de fluorescencia que se generan a partir de una solución saturada de aminoácidos aromáticos son susceptibles de experimentar diversas interacciones que ocasionan cambios en los espectros y/o apagamiento de señales (tales como *quenching* colisional y filtro interno [86]).

Si bien no se realizó un estudio exhaustivo para caracterizar las interacciones fotoquímicas entre aminoácidos, se pudieron obtener perfiles espectrales altamente compatibles con los de la tirosina y el triptófano e indicativos de su presencia en las muestras de medio de cultivo, mediante una prueba simple de laboratorio. Para este fin, una muestra genérica de fermentación se diluyó 1:50 en agua ultra pura y se adquirió una EEM en las mismas condiciones instrumentales. La misma fue luego analizada mediante MCR-ALS, obteniendo los resultados que se muestran en la Figura 2.7. Como es esperable, la dilución de la muestra causa una pérdida considerable de información espectral, especialmente para los analitos presentes en la región R2 (Figura 2.7.A). Sin embargo, al diluir la muestra, los efectos de interacción entre aminoácidos dejan de ser significativos, razón por la cual se obtiene un patrón de señal diferente en la R1. Además, MCR-ALS logra en esta región resolver dos componentes de manera muy satisfactoria. Dadas sus características espectrales en ambos modos (Figura 2.7.B y C), los perfiles obtenidos se pueden asociar a los aminoácidos triptófano y tirosina, los cuales coinciden notablemente con los reportados en la bibliografía [86]. En el caso de los componentes de R2, debido a que la dilución genera una pérdida muy importante de la intensidad de señal, MCR-ALS sólo es capaz de resolver un único componente, cuyo perfil espectral es el resultante de la combinación de los componentes 2 a 4 mostrados en la Figura 2.6.

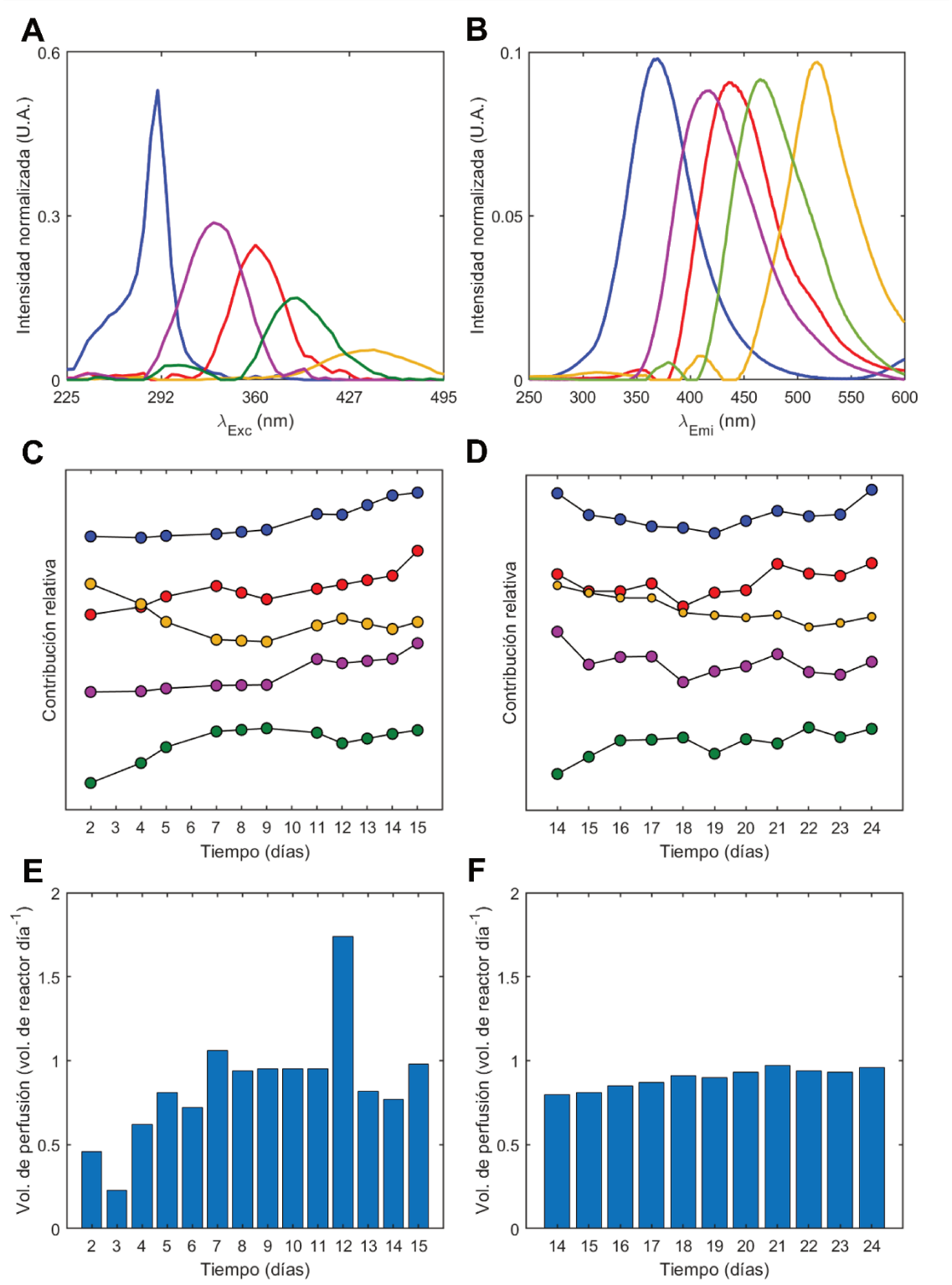


Figura 2.6. Salidas principales de los modelos MCR-ALS con cinco componentes obtenidos con las EEMs de los lotes A2 y B2 para el análisis espectral cualitativo. Componente 1 (azul), componente 2 (rojo), componente 3 (naranja), componente 4 (violeta), componente 5 (verde). En todos los casos, el orden de los componentes se ilustra en función del porcentaje relativo de variabilidad capturada para el modelo (de mayor a menor). **A.** Perfiles espectrales normalizados en el modo excitación (primera submatriz de $\mathbf{C}_{\text{aum}}$ en la Ec. 2.7); **B.** Perfiles espectrales normalizados en el modo emisión (matriz $\mathbf{S}^{\text{T}}$ de la Ec. 2.7); **C.** y **D.** Contribuciones relativas normalizadas de los componentes modelados en función del tiempo de cultivo; **E.** y **F.** Volumen de perfusión en función del tiempo de cultivo.

Tabla 2.2. Caracterización preliminar de posibles fluoróforos presentes en el medio de cultivo de fermentación, a partir de los resultados del modelado mediante MCR-ALS de la información espectral de fluorescencia.

Componente de MCR	Máximo de excitación (nm)	Máximo de emisión (nm)	Posible fluoróforo
C1 (azul)	290	367	Triptófano
C2 (rojo)	360	437	¿NADH?
C3 (naranja)	440	517	¿FAD?
C4 (violeta)	330	416	¿Piridoxina?
C5 (verde)	390	465	¿Tiamina?

Con respecto a los otros aminoácidos aromáticos, se sabe que tanto la fenilalanina como la tirosina también están presentes en el sistema, como aminoácidos libres y como constituyente de proteínas. No obstante, en el caso de las EEMs generadas a partir de muestras sin dilución previa, la tirosina no se resuelve debido a los fenómenos previamente mencionados. Asimismo, la fenilalanina tampoco se resuelve ya que, en virtud de su bajo rendimiento cuántico y teniendo en cuenta que espectralmente está muy solapada con los otros dos aminoácidos, se encuentra muy desfavorecida en señal.

Por otro lado, los componentes 2 y 3 presentan máximos compatibles con los cofactores NADH y FAD, mientras que los compuestos 4 y 5, podrían estar asociados a las vitaminas piridoxina y tiamina, respectivamente [80].

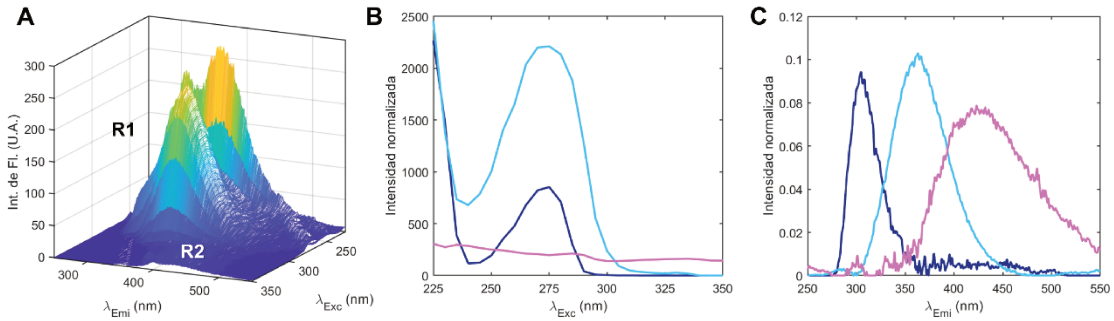


Figura 2.7. Salidas del modelo MCR-ALS con tres componentes obtenido a partir de una EEM adquirida de una muestra genérica del lote A1 diluida 1:50 para la verificación de los componentes presentes en R1. Componente 1 (azul), componente 2 (celeste) y componente 3 (lila). En todos los casos, el orden de los componentes se ilustra en función del porcentaje relativo de variabilidad capturada para el modelo (de mayor a menor). **A.** EEM pre-procesada (se indican las regiones R1 y R2 que equivalen a las mismas regiones espectrales indicadas en la Figura 2.5.B); **B.** Perfiles espectrales normalizados en el modo excitación (matriz $\mathbf{C}$ en la Ec. 2.5); **B.** Perfiles espectrales normalizados en el modo emisión (matriz $\mathbf{S}^T$ de la Ec. 2.5).

Es importante recalcar que la capacidad de resolución del método MCR-ALS se puede ver limitada cuando se trabaja sobre sistemas químicamente complejos y para

los cuales la información *a priori* es limitada [87]. Esto implica, por ejemplo, que MCR-ALS puede arrojar perfiles espectrales que son combinaciones de los espectros de compuestos cuyas señales están altamente solapadas y/o muy desfavorecidas en intensidad y que, por lo tanto, el algoritmo no es capaz de resolver de manera individual. Por estas razones, en algunos casos, la comparación de los perfiles espectrales con la información disponible en la bibliografía no fue suficiente para realizar la caracterización de varios de los componentes modelados (indicados con signo de interrogación en la Tabla 2.2).

Por otra parte, las contribuciones relativas de cada componente en función del tiempo (también denominados de manera coloquial como "*scores* de MCR") mostradas en las Figura 2.6.C y D se obtuvieron a partir del área bajo la curva de los perfiles espectrales contenidos en la matriz $\mathbf{C}_{\mathrm{aum}}$ de la Ec. 2.7. Esta información permite evaluar de manera semicuantitativa cómo varia la cantidad de cada compuesto a lo largo del modo muestral (que, en este caso, corresponde al tiempo de fermentación). Con respecto a esto, las Figura 2.6.C y D revelan, a pesar de las diferencias esperables entre dos procesos diferentes, ciertas particularidades comunes. En este sentido, se observa un importante grado de correlación de los componentes (lo cual es consistente con el régimen de cultivo continuo). Además, se observa en cada caso, que los componentes presentan un cambio en su evolución, lo cual se puede asociar a cambios en el crecimiento de los cultivos y/o al régimen de perfusión. Observando los *scores* de MCR junto con la Figura 2.2 se puede ver que, para el caso del lote A2, los componentes MCR-ALS muestran un tipo de evolución hasta el día 10, y luego experimentan un incremento. En particular en este cultivo, el día 10 marca un cambio importante en el crecimiento de las células. Por su parte, para el lote B2 también se tiene un cambio en la dinámica de los compuestos de MCR-ALS alrededor del día 20 y a partir de ese momento, también tienden a acumularse en el medio. En este caso, los datos de fermentación sugieren que en ese día se produce el fin de una primera fase estacionaria de crecimiento, seguida de una segunda fase exponencial. Dicha etapa se caracteriza por una aceleración en el consumo de glucosa y la acumulación de lactato, junto con un decaimiento cada vez más acentuado de la viabilidad celular.

En relación a las tasas de perfusión, las Figuras 2.6.E y F sugieren que la proporción de volumen de reactor que se intercambia para ambos lotes también podría contribuir a explicar las tendencias observadas en la variación de los componentes de MCR-ALS. En este sentido, se observa que para ambos sistemas hay una estabilización de la cantidad de medio intercambiado alrededor de los días 10 y 20 en cada caso. En particular, en el caso del lote A2, el valor observado en el día 12 corresponde a un error operativo que fue corregido el mismo día. En líneas generales, resulta llamativo que, a

pesar de que en la última etapa de cada cultivo el volumen de perfusión es aproximadamente constante (con lo cual es esperable que las variables se estacionen), la dinámica en que varían los compuestos fluoróforos estaría respondiendo a un complejo balance que involucra tanto al régimen de operación del reactor como al consumo y producción de metabolitos propios del crecimiento celular.

Es evidente que si se quisiera integrar la información de las variables de CPPs con la información espectral obtenida a partir de un enfoque de modelado estadístico, sería necesario incluir herramientas de modelado determinístico. Sin embargo, esta tarea no resultaría sencilla teniendo en cuenta el alto número de variables que podrían considerarse. Asimismo, la falta de información previa para la caracterización completa de los fluoróforos así como el alto grado de correlación observado entre los *scores* de MCR, hacen que esta herramienta *per se* no resulte suficiente para el diseño de una estrategia de monitoreo del proceso. Sin embargo, resulta interesante destacar que la información recabada con esta estrategia ha permitido evidenciar correlaciones entre los componentes modelados y las variables de proceso, que resultan reproducibles entre los diferentes lotes analizados. Esto resultó particularmente atractivo para motivar el diseño de modelos de calibración para el monitoreo de la concentración de etanercept, cuyo desarrollo será tratado en los Capítulos 3 y 4 de este trabajo.

2.4.4. Análisis de agrupamiento mediante PCA a partir de datos espectrales

A partir de las matrices construidas con EEMs desdobladas, se generaron modelos PCA para los diferentes lotes de fermentación. Utilizando los resultados del lote A1 a modo de ejemplo, en la Figura 2.8 se presentan las principales salidas que se obtienen del modelo PCA respecto a su optimización y a la inspección de valores atípicos: PVA y autovalores en función del número de PCs (Figura 2.8.A), RMSECV en función del número de PCs (Figura 2.8.B), diagrama de influencia (Figura 2.8.C), diagramas de *scores* en las PCs significativas (Figura 2.8.D). Para todos los lotes, los resultados fueron equivalentes. Como se observa en la Figura 2.8.A, las primeras dos PCs capturaron más del 95% de variabilidad de los datos. Asimismo, el procedimiento de CV arrojó un mínimo de RMSECV en 2 PC. Por otra parte, de la inspección de los residuos (Figura 2.8.C) y del mapa de *scores* significativos (Figura 2.8.D) no se observaron valores atípicos. En particular, sólo una muestra del lote B2 debió ser descartadas por presentar regiones de saturación de la señal (ver Capítulo 3, Sección 3.4.1).

Una vez validados los modelos PCA, se procedió a realizar el análisis de agrupamiento, para lo cual, se graficaron los *scores* de las muestras en el PC2 vs los *scores* en el PC1. Tal como se realizó con los datos de CPPs, la inspección de grupos

de muestras en los diagramas de *scores* se asoció con las curvas de viabilidad celular. Los resultados se muestran en la Figura 2.9. En particular, en la parte superior de la Figura 2.9 se muestran las mismas curvas de viabilidad presentadas en la Figura 2.2 (a los fines de mejorar la visualización y enfatizar lo que se desea mostrar) para cada lote. En el caso de los procesos del Sistema A, se generaron cuatro modelos PCA individuales. Por otra parte, debido a que los modelos PCA obtenidos tanto con los CPPs como con las EEMs para el caso del lote B1 no mostraron una diferenciación de grupos respecto de la viabilidad, los lotes del Sistema B fueron modelados en conjunto.

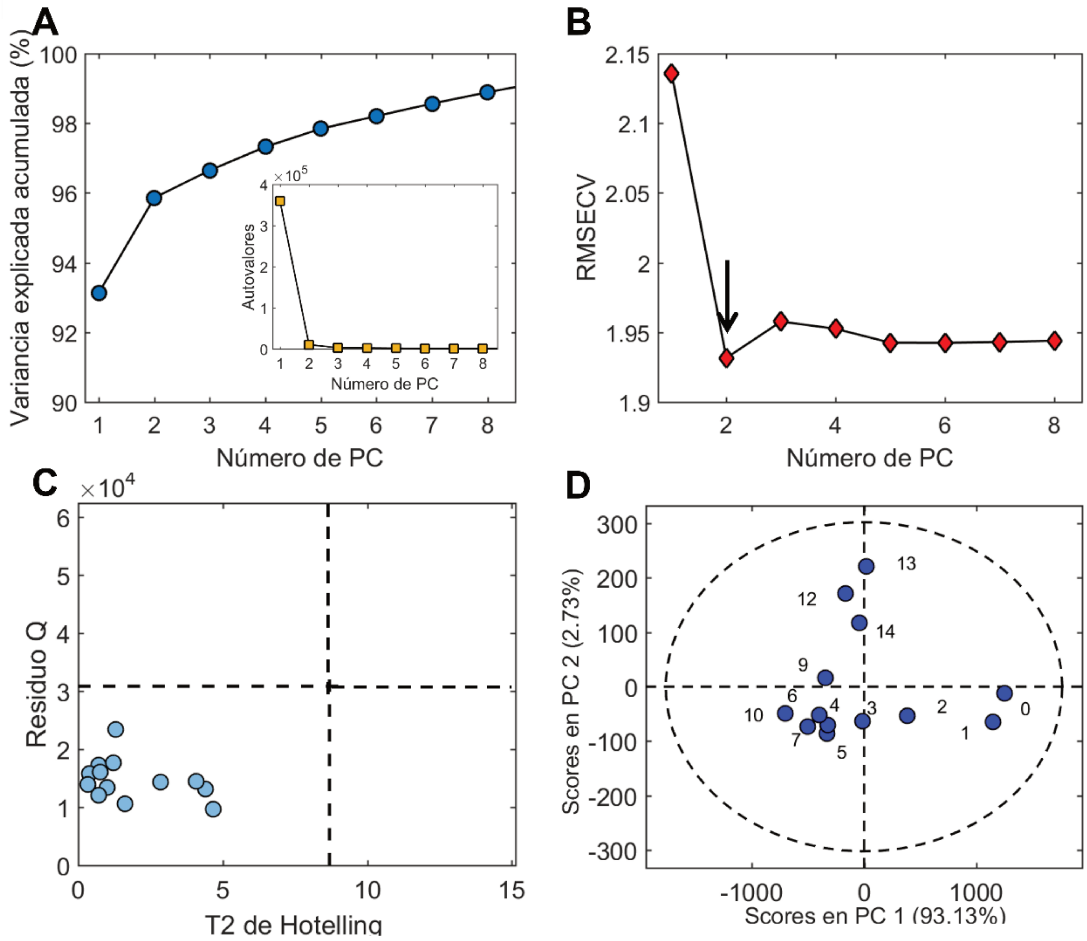


Figura 2.8. Salidas principales del modelo PCA para los datos espectrales (EEMs desdobladas) del lote A1: **A.** PVA y autovalores (figura incrustada) en función del número de PC; **B.** RMSECV en función del número de PC (la flecha indica el mínimo global); **C.** diagrama de influencia (las líneas discontinuas representan los límites de confianza estimados con un 0,05 de nivel de significancia); **D.** *scores* en PC2 vs PC1 (la línea discontinua representa la elipse de confianza calculada con un nivel de significancia de 0,05).

Como se mencionó anteriormente, una particularidad que presenta este bioproceso es la caída impredecible de la viabilidad celular. Si sólo se tienen en cuenta las curvas de viabilidad en función del tiempo, es evidente que dicho parámetro cae de manera considerable luego de los días 12, 11, 9, 11 y 24 para los lotes A1, A2, A3, A4

y B2, respectivamente, mientras que el lote B1 este efecto es menos importante. Por otra parte, una inspección detenida de los diagramas de *scores* de la Figura 2.9 revela que las muestras tienden a diferenciarse conforme avanza el tiempo, en la dirección del PC2. Sorpresivamente, y a diferencia de lo que se observa en los modelos PCA que sólo tienen en cuenta variables de proceso (Figura 2.4), al modelar los datos espectrales, se observa en todos los casos, que la muestra que corresponde al último día en que la viabilidad se mantiene alta según la medida univariada de viabilidad, en los diagramas de *scores* se encuentra más cerca de las muestras de baja viabilidad (ver muestras 12, 11, 9, 11 y 21 para los lotes A1, A2, A3, A4 y B2, respectivamente, en la Figura 2.9). Sin embargo, la Figura 2.4 revelaba que, si sólo se consideran datos de CPPs, la última muestra que corresponde a una medida de viabilidad alta, se agrupa con las muestras en las que dicho parámetro también es alto (estas muestras están indicadas con asteriscos en la Figura 2.4). En consecuencia, esta tendencia representa un criterio alternativo para el monitoreo de la viabilidad en el proceso a partir de la información espectral.

Desde el punto de vista quimiométrico, es interesante observar que, a diferencia de lo que ocurre con los datos de CPPs, cuando se modelan datos espectrales, las muestras tienden a agruparse en el sentido del PC2, a pesar de que dicha PC captura una variabilidad del modelo muy pequeña en comparación a la PC1. Este hecho refleja que en un modelo PCA no necesariamente la máxima variabilidad está asociada con la dirección de máxima separabilidad de las muestras.

Por otra parte, es interesante interpretar los resultados del modelado exploratorio de datos desde el punto de vista biotecnológico. Debido al hecho de que un bioproceso se basa en el uso de células vivas, es esperable que un evento negativo del proceso esté relacionado con cambios no observables (por ejemplo, cambios metabólicos) que comienzan antes de que el impacto sobre los CPPs resulte evidente u observable. Asimismo, este *delay* temporal puede ser del orden de días en el caso de los cultivos de células animales, debido a su baja tasa de duplicación. De esta manera, para la eficiencia del proceso, resulta ventajoso contar con estrategias adicionales que permitan anticipar el momento en que la viabilidad decae. En este caso, se muestra cómo la huella espectral de fluorescencia registrada durante la fermentación podría alertar sobre cambios en las condiciones del cultivo, que derivan luego en una caída de la viabilidad. Este efecto anticipado ofrece la posibilidad de efectuar una inferencia de tipo prospectiva sobre el proceso y, por lo tanto, permite tomar una decisión sobre la fermentación de manera anticipada. Por ejemplo, implementar acciones correctivas, o bien, finalizar el proceso, ahorrando tiempo y recursos.

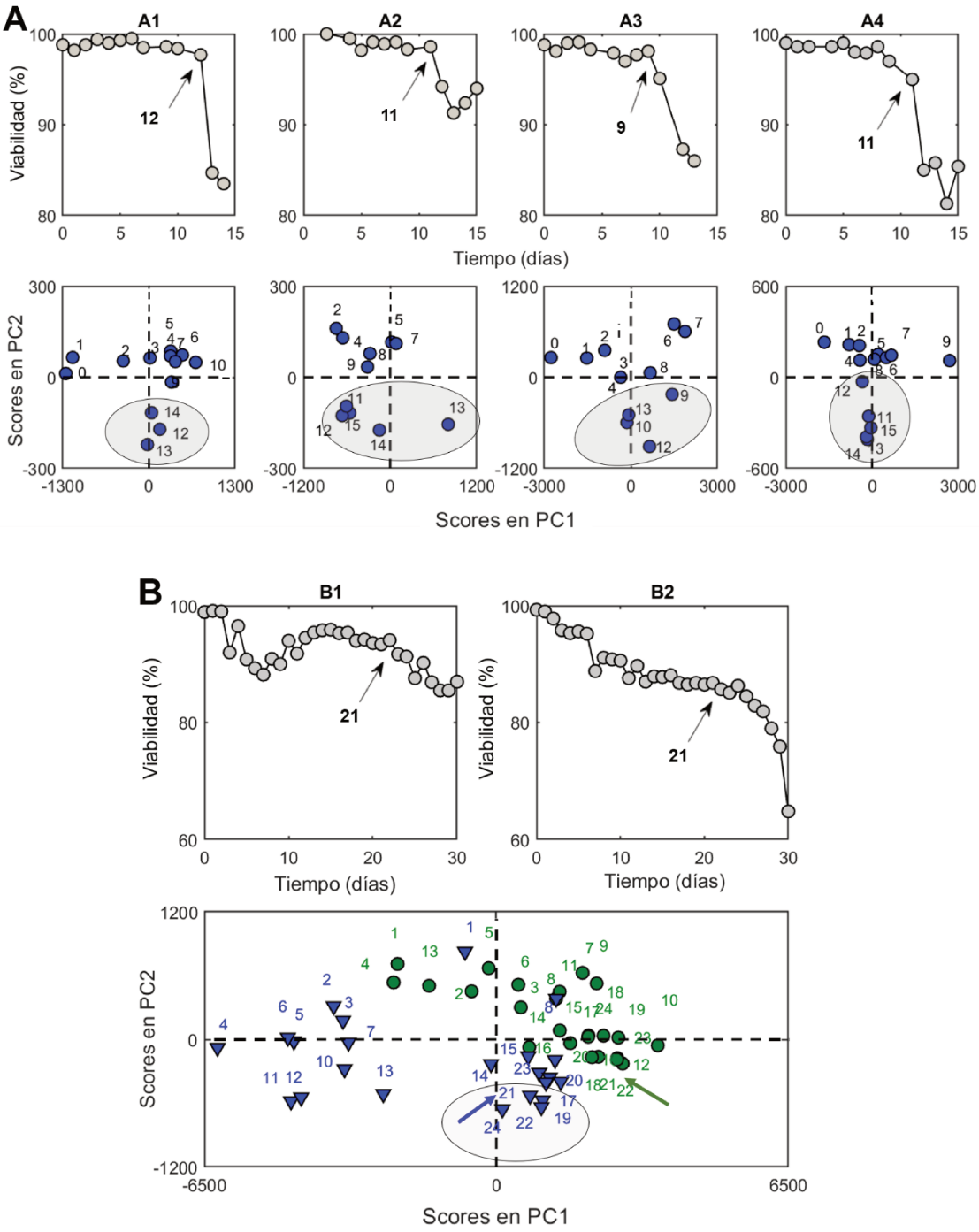


Figura 2.9. Comparación entre las mediciones *off-line* de viabilidad celular y el modelado PCA de datos espectrales multivariados para análisis de agrupamiento. **A.** lotes del Sistema A. **B.** lotes del Sistema B. En ambos casos se muestra la viabilidad celular en función del tiempo de proceso (arriba) y los *scores* obtenidos de modelos PCA con 2 PCs para análisis de agrupamiento (abajo): se graficaron los *scores* de las muestras en PC2 (aproximadamente 5% de varianza explicada) vs PC1 (aproximadamente 90% de varianza explicada). Para los lotes del Sistema A (A1 a A4) se generaron modelos PCA individuales. En el caso del Sistema B, los lotes B1 (círculos verdes) y B2 (triángulos azules) se modelaron de manera conjunta. Las elipses grises indican los *clusters* de muestras consideradas de baja viabilidad. En todas las gráficas, las etiquetas de los puntos representan el tiempo de cultivo en días. Las flechas negras señalan el día a partir del cual la viabilidad comienza a decaer, lo que sucede unos pocos días antes de la interrupción del proceso. Las flechas de colores señalan los grupos de muestras de los lotes B1 y B2 alrededor del día 21 para enfatizar las diferencias.

Con respecto al Sistema B, resulta importante aclarar algunas cuestiones respecto de la ventana de tiempo seleccionada para el análisis espectral. Como se mencionó anteriormente, con las muestras de fermentación de los días 25 a 30 se lograron medir los CPPs pero no fue posible la adquisición de las EEMs. Sin embargo, los resultados que se muestran en la Figura 2.9.B sugieren una cierta similitud con lo que ocurre en el Sistema A y, en este sentido, el período analizado para el Sistema B parece ser representativo del efecto que se desea mostrar. En particular, en la Figura 2.9.B, si se comparan las curvas de viabilidad en relación a los *scores* de las muestras para el modelo PCA del Sistema B, puede verse claramente que entre los días 21-24, el lote B1 aún mantiene una viabilidad por encima del 90%, mientras que para el lote B2, ya se observa una mayor disminución en este parámetro. De esta manera, los *scores* del período 21-24 del lote B2 (flecha azul) tienden a diferenciarse del resto de las muestras en el sentido negativo del PC2, como ocurre para los lotes del Sistema A. Contrariamente, las muestras del lote B1 del mismo período no presentan un agrupamiento marcado respecto del resto de las muestras (flecha verde). Por lo tanto, es razonable suponer que, para el Sistema B, aquellas muestras consideradas de alta viabilidad desde el punto de vista de los CPPs se encuentren más cercanas en términos espectrales a las muestras de baja viabilidad. En síntesis, esto indica que podría establecerse un criterio equivalente para el monitoreo de la viabilidad celular en los lotes de escala industrial.

Debido a la cantidad insuficiente de datos espectrales, los resultados obtenidos para el Sistema B se consideran preliminares y su análisis finalizó en la etapa exploratoria. No obstante, los resultados obtenidos para el Sistema A motivaron el desarrollo de un modelo de clasificación para fines predictivos. Este se presenta en la siguiente sección. En virtud de las observaciones realizadas con los datos espectrales, el criterio de clasificación de las muestras en alta y baja viabilidad de aquí en adelante, responde al análisis de agrupamiento discutido en esta sección y no a los resultados obtenidos mediante PCA con datos de CPPs.

Finalmente, es importante señalar que la estrategia de utilizar EEMs desdobladas (en lugar de espectros adquiridos a una única longitud de onda de excitación) se fundamenta en el hecho de que los datos de segundo orden permiten barrer una cantidad mucho mayor de información del sistema que los de primer orden, aumentando la robustez de los modelos.

2.4.5. Desarrollo de un modelo de clasificación para el monitoreo prospectivo de la viabilidad celular basado en el método PLS-DA con EEMs desdobladas

El PCA permitió realizar el análisis de agrupamiento a partir de los datos generados para procesos finalizados. Teniendo en cuenta una potencial aplicación tecnológica y con el objetivo de generar un modelo predictivo para la inferencia prospectiva de la viabilidad celular en lotes futuros del mismo bioproceso, se diseñó un modelo multivariado de clasificación basado en el uso de las EEMs desdobladas, mediante el método PLS-DA.

Se desarrolló un modelo PLS-DA con las muestras correspondientes a los lotes del Sistema A. En este sentido, las muestras de los lotes A1-A3 se utilizaron para el ajuste del modelo y CV, mientras que el lote A4 sirvió como conjunto de validación externa. Para obtener los vectores lógicos de clases para el método PLS, las muestras fueron asignadas a la clase 1 (alta viabilidad) o a la clase 2 (baja viabilidad), de acuerdo a lo observado en los diagramas de *scores* de la Figura 2.9. De esta manera, se ajustó un modelo PLS-DA con un número óptimo de 3 LVs y con un porcentaje de varianza explicada acumulada superior al 90%. Los resultados de clasificación para los conjuntos de entrenamiento y validación se muestran gráficamente en la Figura 2.10. Asimismo, las matrices de confusión y los índices de clasificación calculados para las predicciones del conjunto de entrenamiento, CV y validación externa se compilan en la Tabla 2.3.

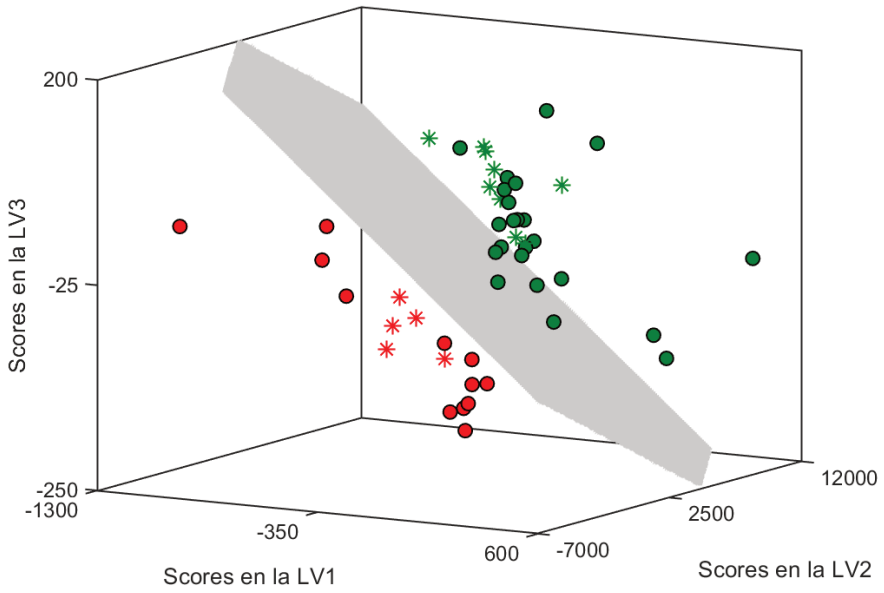


Figura 2.10. Modelo PLS-DA de 3 LVs obtenido para el Sistema A: los círculos y los asteriscos representan los *scores* de las muestras en los conjuntos de entrenamiento y validación externa, respectivamente. Las muestras de la clase 1 (alta viabilidad) y la clase 2 (baja viabilidad) están representadas en verde y rojo, respectivamente. El plano gris representa el plano discriminante calculados a partir de los *scores* de PLS.

Tabla 2.3.A. Matrices de confusión obtenidas a partir de las predicciones del modelo PLS-DA para los conjuntos de entrenamiento, CV y validación externa.

	Nominal / Predicha	**Clase 1**	**Clase 2**
Entrenamiento	Clase 1	25	0
	Clase 2	0	12
CV	Clase 1	25	0
	Clase 2	0	12
Validación externa	Clase 1	9	0
	Clase 2	0	5

Tabla 2.3.B. Resumen de los índices de clasificación asociados al modelo PLS-DA desarrollado para el Sistema A.

	Muestras	**Índice de clasificación**	**Clase 1**	**Clase 2**
Entrenamiento	Lotes A1-A3	Especificidad	1.00	1.00
		Sensibilidad	1.00	1.00
		Precisión	1.00	1.00
		NER	1.00	
		Exactitud	1.00	
CV	Lotes A1-A3 (ciegos venecianos con 10 particiones)	Especificidad	1.00	1.00
		Sensibilidad	1.00	1.00
		Precisión	1.00	1.00
		NER	1.00	
		Exactitud	1.00	
Validación externa	Lote A4	Especificidad	1.00	1.00
		Sensibilidad	1.00	1.00
		Precisión	1.00	1.00
		NER	1.00	
		Exactitud	1.00	

Como se puede notar en la Figura 2.10, el modelo PLS-DA permitió el ajuste satisfactorio de los datos, estableciendo una clara discriminación entre las clases. En la Figura 2.10 se incluye también una representación del plano discriminante del modelo, calculado a partir de los *scores* de PLS. Además, los resultados arrojados en las matrices de confusión (Tabla 2.3.A) así como los índices de clasificación que se calculan a partir de ellas son satisfactorios (Tabla 2.3.B). En particular, se obtuvo una NER porcentual del 100% para las predicciones de los conjuntos de entrenamiento, CV y validación. Especialmente, este parámetro calculado para la CV y la validación externa resultan muy importantes para descartar un sobreajuste del modelo. Por lo tanto, es posible afirmar que se obtuvo un modelo parsimonioso y con una buena capacidad predictiva.

Finalmente, es importante remarcar que el modelo predictivo propuesto no tiene como objetivo proveer una estimación cuantitativa directa de la viabilidad celular, es decir que no sustituye a la técnica univariada de referencia de monitoreo. Por el contrario, la estrategia desarrollada proporciona información cualitativa que estaría asociada con el estadio fisiológico del cultivo. Esto representa información valiosa sobre el proceso ya que, como fue demostrado, puede ayudar a anticipar el momento en que la viabilidad celular comienza a declinar, permitiendo así una mayor eficiencia en la toma de decisiones.

2.5. Conclusiones del capítulo

El modelado integral de datos espectrales a través de una combinación de diversas herramientas quimiométricas permite extraer una gran cantidad de información a partir de los datos de variables de proceso y espectrales.

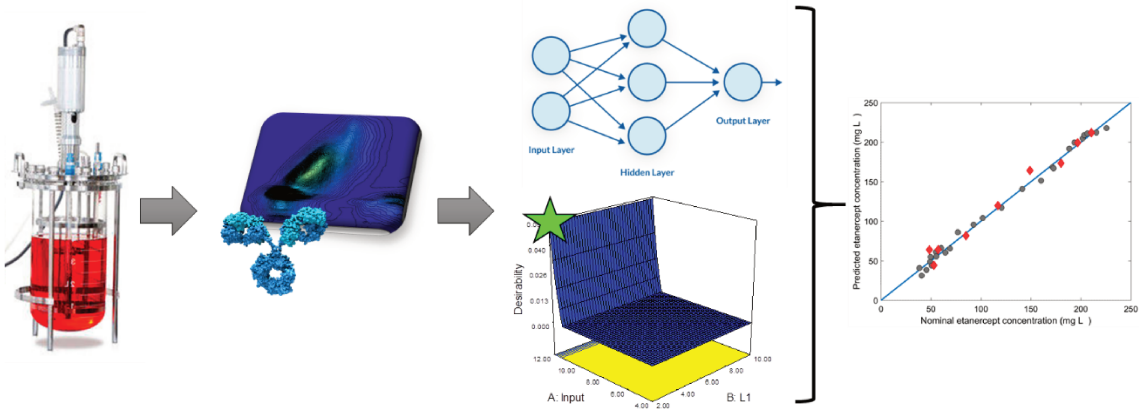
En particular, los resultados de MCR-ALS permitieron una identificación preliminar de posibles fluoróforos presentes en el medio de fermentación y se observaron correlaciones entre *scores* de MCR y variables de proceso que resultan interesantes para la elaboración de modelos predictivos cuantitativos.

Por otra parte, los análisis de agrupamiento con datos espectrales revelaron un consistente patrón diferencial que permitió establecer un nuevo criterio para el seguimiento de la viabilidad de los cultivos, generando una herramienta para la inferencia prospectiva de este CPP y brindando la capacidad de aumentar la eficiencia en la toma de decisiones respecto al monitoreo del bioproceso. En este sentido, los hallazgos efectuados a partir del análisis de agrupamiento motivaron el desarrollo de un modelo predictivo basado en el método PLS-DA de clasificación multivariada, cuyo desempeño en términos de ajuste y capacidad predictiva resultó óptima (tasa de no error porcentual del 100%). La aplicación del modelo obtenido permitiría efectuar una inferencia prospectiva acerca de la evolución de la viabilidad en nuevos lotes del bioproceso en escala piloto.

La metodología desarrollada está en concordancia con los principios de la PAT ya que se basa en el uso de una técnica instrumental robusta y económica, y con la rapidez suficiente para la implementación *at-line* del método de monitoreo. Asimismo, la misma estrategia podría aplicarse de manera *in-* u *on-line*, en virtud de que la técnica analítica tiene la capacidad de ser automatizada.

CAPÍTULO 3

Desarrollo de una estrategia de PAT cuantitativa para el monitoreo de la proteína recombinante



La imagen del biorreactor corresponde a una vista parcial de una fotografía tomada de https://www.labwrench.com/equipment/21988/sartorius-group-biostat-a

3.1. Introducción

En el Capítulo 2 se comentaron los fundamentos básicos de la regresión PLS. Tal como fue descripto, cuando en el vector $\mathbf{y}$ de la Ec. 2.9 se incluye una variable cuantitativa continua, entonces PLS se trata de un modelo de regresión y, desde el punto de vista quimiométrico, se utiliza como un método de calibración de primer orden. El algoritmo PLS es una poderosa técnica predictiva que ha sido ampliamente utilizada para el desarrollo de métodos analíticos basados en datos espectrales para la determinación multianalito en muestras de diversa naturaleza [88-90]. En el contexto de la PAT aplicada a cultivos celulares, existen numerosos desarrollos reportados basados en PLS [33,34]. No obstante, una de las limitaciones del método en su versión clásica es que el modelo asume linealidad entre la variable respuesta y las predictoras. En el contexto de la calibración analítica, esto implica una relación lineal entre la señal instrumental y la concentración del analito de interés. Los desvíos de la linealidad en las técnicas espectroscópicas pueden deberse a diferentes circunstancias. En particular, en el contexto de los métodos de PAT para el monitoreo de bioprocesos, suele ocurrir que la señal espectral es generada directamente a partir de muestras químicamente complejas y saturadas, con lo cual la ley de Beer-Lambert deja de ser válida no sólo como resultado de la saturación, sino porque el o los analitos de interés suelen estar afectados por los efectos de la matriz [27].

Si bien existen variantes no lineales del modelo PLS clásico, tales como el PLS cuadrático [91,92] o el Kernel-PLS [93], estos métodos paramétricos pueden presentar la limitación de asumir como ciertos algunos supuestos sobre los datos que, en ocasiones, no son evidentes o no son fáciles de probar. Por ejemplo, en el caso del PLS cuadrático, se asume una relación de tipo polinomial entre la señal y la concentración. Asimismo, al día de hoy, el modelo PLS sigue siendo un objeto de investigación activa y continúan siendo reportadas nuevas variantes del método, que incluyen el tratamiento de problemas no lineales [94,95].

3.1.1. Redes neuronales artificiales (ANNs) en calibración multivariada de primer orden: el perceptrón multicapa (MLP)

Una alternativa interesante para lidiar con los problemas de no linealidad en calibración multivariada ha sido, desde hace algunas décadas, el uso de algoritmos basados en ANNs [96].

Desde el punto de vista de la ingeniería computacional, las ANNs constituyen un tipo de algoritmo de inteligencia artificial, que se encuadra dentro del subgrupo de métodos denominados como aprendizaje maquinal (*machine learning*). En líneas

generales, las técnicas de *machine learning* comparten la característica de ser algoritmos que tienen la capacidad de aprender a partir de un conjunto de datos, sin ser expresamente programados para ello, es decir, de manera automática. Al igual que los métodos estadísticos clásicos, los algoritmos de *machine learning* pueden ser de aprendizaje supervisado (clasificación y regresión) o no supervisado (agrupamiento). Dada su versatilidad y flexibilidad, estos algoritmos no paramétricos resultan convenientes para el tratamiento de problemas no lineales, especialmente, cuando los datos carecen de ciertas propiedades o estructura. En particular, las ANNs deben su nombre a la manera en que fueron originalmente diseñadas, basándose en la forma en que se organizan y aprenden las neuronas biológicas (algoritmos bioinspirados). Si bien los fundamentos de las ANNs han sido propuestos hace ya algunas décadas, su desarrollo y aplicaciones se han disparado en los últimos años debido fundamentalmente a la capacidad cada vez más accesible para generar, almacenar y procesar grandes bases de datos. En este sentido, vale la pena hacer una distinción entre aquellas ANNs que desde un punto de vista histórico se pueden denominar como "redes clásicas", ya que tienen un cierto tiempo de existencia (desde mediados del siglo XX) y las ANNs más recientes, que reciben el nombre de "redes de aprendizaje profundo" o *deep learning* (de estas últimas, los primeros desarrollos datan de la década del 2000). Este tipo de redes representan en la actualidad un área muy caliente de investigación y desarrollo tecnológico. En este sentido, el *deep learning* se ha constituido en un área en sí misma que atraviesa transversalmente buena parte de las disciplinas que refieren al análisis masivo de datos.

En el contexto de la química analítica, una revisión bibliográfica indica que, en general, la mayoría de los problemas analíticos relacionados con la calibración de primer orden en sistemas no lineales puede ser abordado con el uso de redes clásicas [97]. En este sentido, uno de los tipos de ANN que más ha sido reportado en la literatura es el algoritmo conocido como perceptrón multicapa (MLP). Debido a su flexibilidad y capacidad para lidiar con problemas no lineales, el uso de MLP (y de ANNs en general) ha demostrado ser adecuado para el desarrollo de métodos de calibración aplicado a bioprocesos [98-100].

El MLP constituye un método supervisado que permite resolver un problema de regresión no lineal, proyectando un conjunto de variables de entrada (predictoras) en una función no lineal y ajustando sus parámetros libres de manera automática mediante un proceso iterativo de minimización del error de predicción. Para entender el fundamento del MLP, es necesario realizar una breve descripción de su antecesor, el perceptrón simple (SP), que es el nombre que se le dio al primer algoritmo de ANN

descripto y que constituye el modelo matemático de neurona más simple [101]. Una representación esquemática de este algoritmo se muestra en la Figura 3.1.

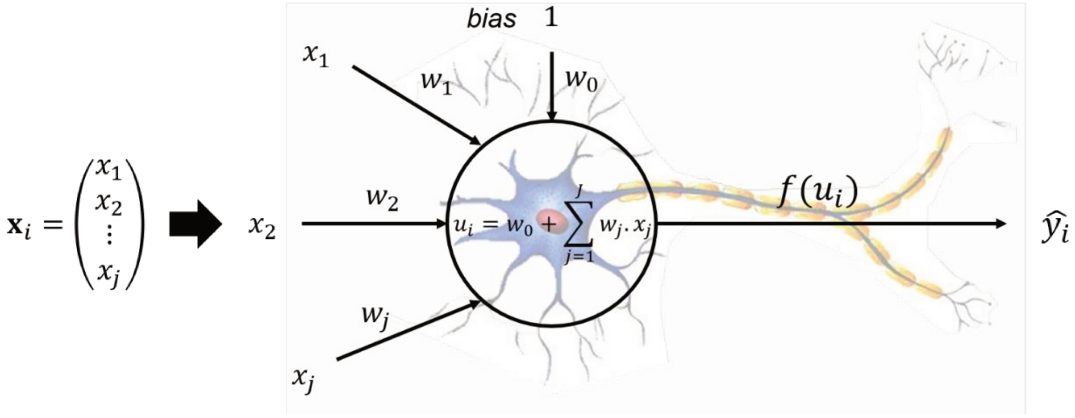


Figura 3.1. Modelo matemático de neurona (perceptrón simple). El vector $\mathbf{x}_i$ representa las señales correspondientes al i-ésimo patrón de calibración (cuyas variables predictoras son los escalares x_1, x_2 hasta x_j) que es el *input* del perceptrón. En el cuerpo neurona se calcula una suma ponderada u_i, en la que intervienen las variables de entrada, el *bias* y los pesos sinápticos w_0 y w_1, w_2 hasta w_j. La cantidad u_i sirve como argumento de la función de transferencia f. La imagen de la función de transferencia constituye la predicción de la variable respuesta del modelo $\hat{y}_i$ (salida de la neurona).

El *input* del SP es un arreglo matricial $\mathbf{X}$ de I muestras de entrenamiento $\times J$ variables predictoras (bloque X). A su vez, a cada muestra de entrenamiento le corresponde un valor nominal (conocido) de la variable respuesta contenido en un vector $\mathbf{y}$ (bloque Y). Como se observa, dada una muestra i del conjunto de entrenamiento cuya señal constituye un dato de tipo vectorial $\mathbf{x}_i$, el SP se alimenta del conjunto de J escalares x_j que definen al vector $\mathbf{x}_i$. Las entradas se conectan al cuerpo neuronal a través de los denominados "pesos sinápticos", que están representados por los escalares w_j. Adicionalmente, el SP hace uso de una variable de entrada que es independiente de las variables predictoras y que se conoce como *bias*. El *bias* suele tomar el valor 1 por defecto y está conectado al cuerpo neuronal por un peso simbolizado como w_0. Posteriormente, se efectúan dos operaciones matemáticas: en primer lugar, se calcula una suma ponderada u_i (también llamada "salida lineal") de acuerdo a:

$$u_i = w_0 + \sum_{j=1}^{J} w_j . x_j \tag{3.1}$$

donde los pesos sinápticos w_0 y w_j constituyen los parámetros libres del perceptrón que permiten lograr el ajuste de los datos durante su entrenamiento. Como se intuye de la Ec. 3.1, el *bias* hace las veces de parámetro de "ordenada al origen" del modelo y permite efectuar traslaciones de la variable u_i para brindar mayor flexibilidad durante la

optimización de su capacidad predictiva. La combinación lineal de entradas y pesos u_i sirve en una segunda instancia como argumento para una función f que se denomina "función de activación neuronal" (también denominada "función de transferencia" o "salida no lineal"). En este sentido, la función de transferencia más simple que puede emplearse en un modelo de SP es la función signo[18], aunque el algoritmo admite en sus diferentes aplicaciones una variedad de funciones de diferente complejidad (logística, logarítmica, tangente hiperbólica, gaussiana, entre otras). En cualquiera de los casos, la imagen de u_i dada por la función f constituye la salida de la neurona, que representa la predicción $\hat{y}_i$ de la variable respuesta en la i-ésima muestra de entrenamiento[19].

El SP constituye la unidad estructural de un MLP, el cual se conforma de un conjunto de perceptrones interconectados (también llamados neuronas o nodos) dispuestos en una serie de capas (*layers*), tal como se esquematiza en la Figura 3.2. El número de capas y la cantidad de neuronas por capa que caracteriza un MLP es lo que se conoce como "arquitectura" o "topología" de la red. En relación al nombre de las capas, existe cierta divergencia en la literatura. En este trabajo se adopta el siguiente criterio de nomenclatura. Se considera como capa de entrada (*input layer*) al conjunto de variables predictoras más el *bias* de entrada (cuyas neuronas se representan en color verde en la Figura 3.2). Por otra parte, la o las neuronas que efectúan las operaciones finales para dar la predicción del modelo conforman la denominada capa de salida (*output layer*). Debido a que todos los modelos MLP desarrollados en este trabajo se aplican a problemas de calibración para la predicción de la concentración de un único analito de interés, en todos los casos, la capa de salida está formada por una única neurona (representada en color rojo en la Figura 3.2). Finalmente, todas las neuronas que forman parte de la capa intermedia constituyen la denominada capa oculta (*hidden layer*) del MLP (representada en azul en la Figura 3.2). Es importante mencionar en este punto que un MLP puede tener una o varias capas ocultas. A medida que aumenta tanto el número de capas como la cantidad de neuronas en cada capa de un MLP, su arquitectura se vuelve más compleja y esto conlleva un aumento considerable del número de parámetros libres del modelo. La arquitectura de una red MLP suele representarse con una secuencia de números separados por guiones, que indican el número de nodos en cada capa.

Por otra parte, tal como se observa en la Figura 3.2, en el modelo de MLP considerado en este trabajo, todas las neuronas de una misma capa se relacionan con

[18] La función signo o sgn es una función discontinua que, dado un número real, determina su signo. De esta manera, los únicos valores posibles del conjunto imagen de esta función son -1, 0 y 1.
[19] Dependiendo del tipo de función de transferencia empleada, las variables predictoras pueden requerir algún escalado adicional, por lo que en esos casos, la predicción del SP resulta en realidad, la imagen de la función reescalada.

las neuronas de la o las capas adyacentes y no existen conexiones intracapa. Además, a excepción de la capa de entrada, en cada una de las neuronas se efectúan las dos operaciones descriptas más arriba, es decir, la suma ponderada y la proyección de la misma en una función de activación. En el modelo MLP utilizado aquí, las neuronas de la capa de entrada efectúan directamente la proyección de las variables en la función de activación, sin realizar la suma pesada. En particular, el tipo de función de transferencia mayormente empleado para los algoritmos de MLP (y que es la de elección en este trabajo) es la función sigmoidea, del tipo $1/(1+e^{-x})$.

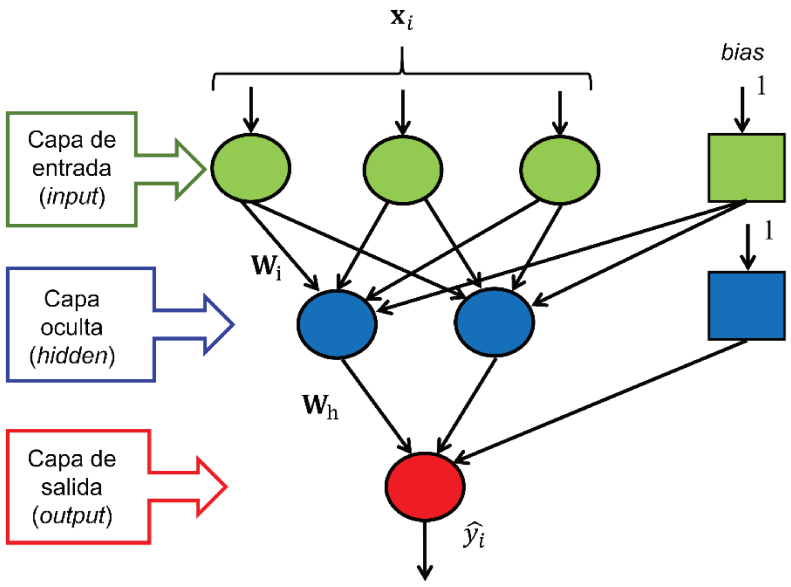


Figura 3.2. Representación esquemática de un MLP: $\mathbf{x}_i$ representa el i-ésimo patrón de calibración; $\mathbf{W}_\mathrm{i}$ y $\mathbf{W}_\mathrm{h}$ son las matrices de pesos sinápticos cuyos elementos representan respectivamente, las conexiones entre las neuronas de la capa de entrada y oculta, y entre las neuronas de la capa oculta y la de salida; $\hat{y}_i$ simboliza la predicción de la variable respuesta para la i-ésima muestra de entrenamiento.

Por otra parte, el algoritmo matemático que permite el ajuste automático de los pesos sinápticos durante el entrenamiento del modelo es lo que se denomina como "algoritmo de aprendizaje". Si bien existe una variedad importante de algoritmos de aprendizaje descriptos y que continúan siendo desarrollados sobre todo para ANNs de mayor complejidad, la técnica matemática por excelencia que es normalmente utilizada en los modelos MLP consiste en un método iterativo conocido como retropropagación del error mediante el vector gradiente (*error back-propagation*). Es por eso que el MLP (así como otras ANNs que utilizan este algoritmo de aprendizaje) suele denominarse también como red *feed-forward* o de *back-propagation.* En este sentido, es común emplear en la jerga de las redes neuronales las expresiones "propagación hacia adelante" para referirse al paso de una muestra de entrenamiento a través del algoritmo,

y "propagación hacia atrás" para denominar a la etapa de ajuste de pesos que opera desde la última capa hacia la primera. En el algoritmo de aprendizaje interviene el bloque Y de valores nominales de la variable respuesta de interés, razón por la que el MLP se considera un algoritmo de *machine learning* supervisado.

El fundamento básico del método de retropropagación consiste en establecer un criterio de error de predicción que debe ser minimizado. En cada iteración, dicho criterio es *sensado* a la salida de la red y utilizado como una medida de ajuste hacia atrás de los pesos sinápticos hasta que se alcanza la convergencia. Una iteración del algoritmo de retropropagación (también conocida como "época") se completa cuando todas las muestras del conjunto de entrenamiento son propagadas hacia adelante una vez, y se produce un único ajuste de todos los pesos de la red.

Matemáticamente, el método de aprendizaje *back-propagation* se puede resumir de la siguiente manera. Sea $e(w_0, w_1, w_2, \dots, w_j)$ la función escalar de $\mathbb{R}^j$ en $\mathbb{R}$ asociada al criterio de error de la predicción de una red MLP y que depende de j pesos sinápticos totales, se define como gradiente de e (∇e) al campo vectorial dado por:

$$\nabla e = \left(\frac{\partial e}{\partial w_0}, \frac{\partial e}{\partial w_1}, \frac{\partial e}{\partial w_2}, \dots, \frac{\partial e}{\partial w_j}\right) \tag{3.2}$$

Tal como se conoce del cálculo vectorial, una de sus propiedades fundamentales es que, evaluado en cada punto $(w_0, w_1, w_2, \dots, w_j)$ del espacio $\mathbb{R}^j$, el gradiente apunta en la dirección de máximo incremento de la función e. De esta manera, la minimización del criterio de error se logra ajustando los pesos sinápticos en la dirección opuesta a la de ∇e. Para el j-ésimo peso de la p-ésima capa en la n-ésima iteración, la proporción de ajuste $\Delta w_i^p(n)$ está dada por:

$$\Delta w_i^p(n) = -\mu \nabla e \tag{3.3}$$

donde μ es una constante conocida como "velocidad de aprendizaje". Este parámetro permite modular la proporción en que se modifican los pesos, evitando los "saltos bruscos" sobre la función error que pudieran entorpecer la convergencia. Normalmente, la primera época se inicializa con pesos asignados al azar, cuyos valores son luego actualizados en cada nueva iteración hasta que se alcanza algún criterio de convergencia. La expresión matemática explícita y el costo computacional de cálculo de ∇e dependen esencialmente de la naturaleza de la función de transferencia y de la complejidad de la red. Es interesante mencionar que, a lo largo del tiempo, se han propuesto numerosas variantes para esta metodología, aunque esta filosofía de aprendizaje continúa teniendo gran vigencia en buena parte de los algoritmos de *machine learning.*

Uno de los cuellos de botella que frecuentemente se produce cuando se desea desarrollar un método de calibración basado en el MLP tiene que ver con la optimización de su arquitectura. En este sentido, a pesar de que los valores de los pesos sinápticos son automáticamente ajustados por el algoritmo, el número de capas y el número de neuronas en cada capa no son parámetros que se ajustan solos y su determinación *a priori* puede no resultar intuitiva, especialmente en problemas no lineales de alta dimensionalidad. A diferencia de los métodos paramétricos y debido a su flexibilidad, las redes neuronales suelen ser bastante más sensibles al sobreajuste. Esto implica directamente que, si la topología de una red está sobredimensionada, es probable que el algoritmo aprenda muy bien los datos de entrenamiento, pero tenga una limitada capacidad predictiva (también denominada capacidad de generalización). Por esta razón es que resulta crítica la determinación de una arquitectura óptima del MLP en el desarrollo de un método de calibración.

En la literatura es posible encontrar una variedad importante de estrategias y criterios para lograr este objetivo. Sin embargo, a diferencia de los métodos paramétricos, las redes mantienen cierta naturaleza de "caja negra". Por esta razón, muchos de los criterios de optimización están basados en técnicas de tipo "prueba y error", por lo que pueden resultar tediosas o insumir mucho tiempo y/o costo computacional. En este contexto, una alternativa interesante que surge desde la quimiometría es la utilización de herramientas de diseño y optimización experimental, en virtud de que la optimización de un MLP requiere de la evaluación de la *performance* del modelo para diferentes combinaciones de sus parámetros topológicos.

3.1.2. Elementos de diseño y optimización experimental

En el contexto del desarrollo de métodos analíticos, el diseño y optimización de experimentos constituye un conjunto de metodologías cuyo objetivo principal es determinar condiciones experimentales que permitan obtener el mejor rendimiento analítico, minimizando el número de experimentos (lo que permite, a su vez, la optimización de recursos) [17]. En líneas generales, la filosofía subyacente en una técnica de diseño y optimización de experimentos consiste en obtener un modelo estadístico para estudiar la relación entre los factores (variables independientes) y una o más respuestas experimentales (variables dependientes). Esta estrategia suele llevarse a cabo en dos etapas: una fase de preselección o cribado, más conocida con el término en inglés *screening*, en la que se analiza la influencia de un conjunto de factores sobre la respuesta experimental (descartando aquellos que no resultan estadísticamente significativos); y una fase de optimización propiamente dicha, en la que se busca modelar la relación respuesta-factores con el objetivo de efectuar

predicciones y hallar combinaciones óptimas de factores que maximicen o minimicen una o varias respuestas de interés. La etapa de optimización desde el enfoque multivariado ofrece numerosas ventajas tales como permitir evaluar posibles interacciones entre los factores, ampliar el entorno experimental y minimizar el número de ensayos. Esta etapa también contempla la posibilidad de optimizar varias respuestas de manera simultánea. La técnica por excelencia que se utiliza con estos objetivos es la denominada RSM [102].

Luego de la fase de *screening* de factores, la implementación de la RSM habitualmente se lleva a cabo de la siguiente manera:

1. Elaboración de una matriz de diseño de experimentos de optimización: en esta primera instancia, se seleccionan los factores y las respuestas experimentales que se desean estudiar y se determina el entorno experimental (es decir, el rango de variación de los factores). Asimismo, se debe seleccionar un diseño estadístico que permita estudiar los efectos de los factores sobre la respuesta y sus interacciones de manera simultánea, estableciendo experimentos con combinaciones de diferentes niveles de factores y minimizando el número de ensayos. Además, un diseño estadístico debe gozar de algunas bondades matemáticas que servirán durante la etapa de modelado (tales como ortogonalidad y rotabilidad, entre otras). En este contexto, los diseños de optimización más frecuentemente utilizados son el diseño factorial completo a tres niveles (FFD), el diseño central compuesto (CCD) y el diseño de Box-Behnken (BBD).
2. Ejecución de los experimentos y construcción de los modelos de superficie de respuesta: en esta etapa, se registran las respuestas experimentales para cada uno de los experimentos sugeridos por el diseño. Posteriormente, se utilizan herramientas estadísticas para modelar la relación de cada respuesta con el conjunto de factores. Esta tarea puede llevarse a cabo mediante técnicas clásicas de ajuste por cuadrados mínimos (modelos lineales de regresión) o bien, mediante el uso de algoritmos basados en ANNs.
3. Determinación del óptimo experimental: si se desea optimizar una única respuesta experimental, entonces la localización del óptimo puede efectuarse de manera relativamente simple mediante la inspección visual de una representación gráfica del modelo obtenido en la etapa anterior. Normalmente, el óptimo experimental suele corresponder a un punto mínimo o máximo de la superficie de respuesta. Por otra parte, si se desean optimizar más de una respuesta experimental de manera simultánea, entonces, se debe integrar la información obtenida en la etapa anterior, para generar una solución de compromiso. Existen diferentes estrategias

para lograr este objetivo, aunque la técnica por excelencia en el ámbito de la química analítica es la que se basa en la función Deseabilidad D de Derringer y Suich [103].

La aplicación de la RSM para la optimización de la arquitectura del modelo de calibración MLP desarrollado en esta etapa del trabajo no implicó trabajo experimental de mesada, sino que cada "experimento" propuesto por el diseño estadístico consistió en una ejecución del algoritmo de ANN, bajo diferentes configuraciones de los parámetros topológicos. Por esta razón es que no fue necesaria la utilización de un diseño de *screening* de factores y directamente se procedió a implementar un diseño de optimización. En este sentido, es importante destacar que si el número de factores es manejable, no siempre es necesario realizar la etapa de *screening,* ya que en un diseño de optimización es posible incluir todos los factores que se deseen, permitiendo luego durante la etapa de modelado, discriminar aquellos factores que no resultan significativos para la respuesta.

3.2. Objetivos específicos del capítulo

En este capítulo se plantearon los siguientes objetivos específicos:

- desarrollar una estrategia de PAT cuantitativa para el monitoreo *at-line* de etanercept en un proceso fermentativo, basado en datos multivariados de fluorescencia;
- evaluar el tipo de correlación entre la señal multivariada de fluorescencia y la concentración de etanercept mediante el algoritmo PLS;
- implementar estrategias de diseño y optimización experimental para llevar a cabo la optimización de un modelo de calibración basado en MLP-ANN.

3.3. Materiales y métodos

3.3.1. Datos

En esta etapa del trabajo, se desarrolló un método de calibración a partir de los datos de fluorescencia y de concentración de etanercept generados para los lotes del proceso a escala industrial (Sistema B), de acuerdo a las técnicas descriptas en la Sección 2.3 del Capítulo 2. Si bien este desarrollo se realizó únicamente para los lotes de la escala industrial, la metodología empleada se puede aplicar de manera análoga para los lotes del Sistema A.

A partir de las EEMs generadas con las 48 muestras adquiridas de los lotes B1 y B2, se descartaron las muestras correspondientes a los días para los cuales no se

tenían datos de concentración de etanercept[20]. Cada una de las EEM fue vectorizada y el total de los datos se dispuso en un arreglo matricial, generando un único *pool* de datos (bloque X). Además, para la construcción del bloque Y se utilizaron los resultados de la cuantificación de etanercept mediante el método cromatográfico de referencia (HPLC) como valores nominales. Este conjunto inicial contó con un total de 36 muestras.

En particular, el bloque X se modeló mediante PCA para la exploración de *outliers* espectrales, de acuerdo a la metodología descripta en la Sección 2.3.5. Luego de la inspección de valores atípicos, el conjunto de datos se dividió en los subconjuntos de calibración (entrenamiento) y validación externa. Para garantizar que todo el rango de concentración estuviera bien representado en ambos grupos, los valores de concentraciones del bloque Y se ordenaron de manera ascendente y luego se graficaron dichos valores en función del número de muestra. De esta manera, se evidenciaron claramente tres grupos o niveles de concentración (baja, media y alta), seleccionando de manera aleatoria tres muestras de cada nivel para el conjunto de validación. Así, el número de muestras de los subconjuntos de calibración y validación fue, respectivamente, 75% y 25% de la cantidad total de mediciones.

3.3.2. Modelado cuantitativo de EEMs desdobladas mediante PLS. Estudio de la correlación entre la señal espectral y la concentración de proteína recombinante

El primer paso en el desarrollo del método de calibración fue explorar la correlación entre la señal multivariada de fluorescencia y la concentración de proteína recombinante. Con este objetivo, se utilizó el conjunto de calibración para construir modelos PLS. El número óptimo de LVs fue estimado mediante una estrategia de LOO-CV, junto con el criterio estadístico propuesto por Haaland y Thomas en 1988 [74]. Se ha descripto en la bibliografía especializada que no necesariamente el número óptimo de LVs para un modelo PLS es aquel para el cual el PRESS alcanza un mínimo global. Por el contrario, y atendiendo el principio de parsimonia, el criterio propuesto por Haaland y Thomas permite determinar un óptimo de LVs asociado a un mínimo local de PRESS, efectuando una prueba estadística con los resultados de la CV. En este sentido, la prueba se basa en calcular cocientes entre los valores de PRESS para cada valor de LV y el PRESS mínimo global. Cada uno de estos cocientes se asemeja al estadístico F (cociente de varianzas) y por lo tanto pueden utilizarse en una prueba estadística de comparación. Calculando la probabilidad asociada a cada valor F con un número de

[20] En este proceso en particular y para optimizar recursos del laboratorio de control de calidad, no se suele determinar la concentración de producto en todas las muestras obtenidas durante los primeros 10 días de fermentación, puesto que se sabe que la cantidad de proteína recombinante se mantiene en un valor basal aproximadamente constante.

grados de libertad igual a la cantidad de muestras de calibración, se ha propuesto en base a resultados empíricos, que un criterio conveniente para elegir el número de LVs consiste en seleccionar aquel para el cual la probabilidad asociada al cociente de PRESS cae por debajo de 0,75. Este fue el criterio empleado para la optimización de los modelos PLS.

Con el objetivo de explorar aquellas regiones espectrales de mayor contribución a la correlación con la concentración de etanercept, se efectuó una selección de variables a través del método PLS por intervalos (iPLS) [104], y los resultados se asociaron a regiones espectrales específicas de las EEMs de fluorescencia. Para cada uno de los modelos generados, se computaron los valores de los estadísticos raíz del error cuadrático medio de la calibración ($RMSEC_{PLS}$) y de la CV ($RMSECV_{PLS}$), y el coeficiente de determinación entre los valores predichos y nominales de concentración para la CV (R^2-CV_{PLS}). Entre los modelos obtenidos, se seleccionó como modelo óptimo para continuar con el estudio aquel que presentara el menor valor de $RMSECV_{PLS}$, ya que constituye un indicador confiable de sobreajuste y capacidad predictiva. Además, los resultados de la CV permitieron analizar el tipo de correlación (lineal/no lineal) entre la señal espectral y la concentración de etanercept. Esta tarea se efectuó mediante inspección visual de las salidas de PLS y a través de la prueba estadística residuos parciales aumentados de Mallows (APaRP) [105], la cual se basó en probar la hipótesis nula de que los datos de concentración predicha vs nominales en la CV son lineales (con un nivel de significancia de 0,05).

Por otra parte, el modelo PLS optimizado se utilizó para efectuar predicciones con las muestras del conjunto de validación. De esta manera, se calcularon los estadísticos raíz del error cuadrático medio de predicción ($RMSEP_{PLS}$), error relativo porcentual de predicción ($REP\%_{PLS}$) y coeficiente de determinación para las concentraciones predichas vs nominales en el conjunto de validación (R^2-$pred_{PLS}$). En particular, el error relativo porcentual de predicción se calcula como:

$$\mathrm{REP\%} = \frac{\mathrm{RMSEP}}{\overline{y_{\mathrm{cal}}}}\, 100 \tag{3.4}$$

donde $\overline{y_{\mathrm{cal}}}$ es la concentración promedio del analito en el conjunto de calibración.

3.3.3. Desarrollo de un método de calibración basado en MLP. Optimización del modelo mediante la RSM

Se propuso un modelo de calibración MLP con una topología general $A - N1 - N2 - O$. La capa de salida se compuso sólo por una única neurona ya que el modelo está pensado para la predicción de una única respuesta (concentración de etanercept). El resto de la arquitectura del modelo fue optimizado mediante la RSM, partiendo de un

diseño estadístico de tipo BBD con cinco puntos centrales [17]. Los factores seleccionados para el diseño fueron el número de nodos en A (Factor A), el número de nodos en $N1$ (Factor B) y el número de nodos en $N2$ (Factor C). Los niveles de cada factor correspondientes a los valores codificados $+1$ y -1 del diseño se variaron entre 4 y 12 para A, 2 y 10 para $N1$ y 0 y 8 para $N2$. El hecho de haber incluido como factor del diseño al número de nodos en A (número de variables predictoras para alimentar al modelo) se fundamenta en el concepto de que los datos espectrales de fluorescencia son altamente redundantes. Por esta razón, para priorizar el principio de parsimonia y minimizar el costo computacional, la estrategia de elección en este contexto suele ser el uso de *scores* de PCA de las muestras (compresión del bloque X) para alimentar al MLP. Sin embargo, la literatura especializada revela que los criterios para determinar el número óptimo de PCs necesarios para comprimir un bloque espectral que luego será utilizado en un modelo de regresión no lineal, no necesariamente son los mismos que se adoptan para modelos lineales. En este sentido, en virtud de que PCA es un modelo de descomposición lineal, puede ocurrir que la información acerca de la estructura no lineal de los datos sea capturada por PCs cuya contribución relativa al modelo PCA sea poco significativa [97]. De esta manera, el número de componentes a utilizar para el entrenamiento del MLP también debe ser optimizado.

Por otra parte, si bien existen otros parámetros ajustables *a priori* del entrenamiento para un modelo MLP (tales como los parámetros del algoritmo de aprendizaje), estos se mantuvieron constantes y en valores estándares de acuerdo a lo descripto en bibliografía y no fueron tenidos en cuenta en el diseño experimental.

En cada experimento sugerido por el BBD se entrenó un modelo MLP con el juego de calibración, realizando además una CV mediante la técnica de ciegos venecianos con 3 particiones (para disminuir el costo computacional y por lo tanto el tiempo de ejecución). Asimismo, para controlar el entrenamiento, se seleccionó al azar un conjunto de muestras de tamaño igual al 20% del juego de entrenamiento. Esta partición del conjunto de calibrado es lo que se conoce habitualmente como conjunto de monitoreo (*monitoring*) y el error asociado a las predicciones que se realizan sobre dicho conjunto permite controlar el algoritmo de aprendizaje para evitar el sobreajuste de la red [97]. En la Figura 3.3 se representa de manera esquemática la variación del error de predicción asociada a los conjuntos de entrenamiento y monitoreo en función del número de épocas. Se ha descripto en la bibliografía que, en general, el error de entrenamiento disminuye de manera monótona con el número de iteraciones, mientras que el error de monitoreo alcanza un mínimo y luego se estabiliza o bien, crece nuevamente. Habiendo fijado el número de épocas en un número arbitrariamente alto, un criterio apropiado para detener el algoritmo de *back-propagation* consiste en registrar

el error de monitoreo y guardar los pesos obtenidos en la época para la cual dicho error alcanza un mínimo.

El entrenamiento de cada red permitió luego el cálculo de las respuestas experimentales $RMSEC_{BBD}$, $RMSECV_{BBD}$ y R^2-CV_{BBD} para cada experimento del diseño. Posteriormente, se aplicó la metodología de RSM para determinar una arquitectura óptima para el modelo MLP de calibración. Una vez establecida la topología más apropiada, se procedió a efectuar una LOO-CV y predecir las muestras del juego de validación externa para evaluar la autoconsistencia y capacidad predictiva del modelo MLP optimizado. Con los resultados de todas las predicciones se calcularon los estadísticos $RMSEC_{MLP}$, $RMSECV_{MLP}$, R^2-CV_{MLP}, $RMSEP_{MLP}$, $REP\%_{MLP}$ y R^2-$pred_{MLP}$ del modelo final. Por último, la exactitud y precisión del modelo obtenido se evaluó mediante la técnica estadística bivariada de región de confianza elíptica (EJCR) [106].

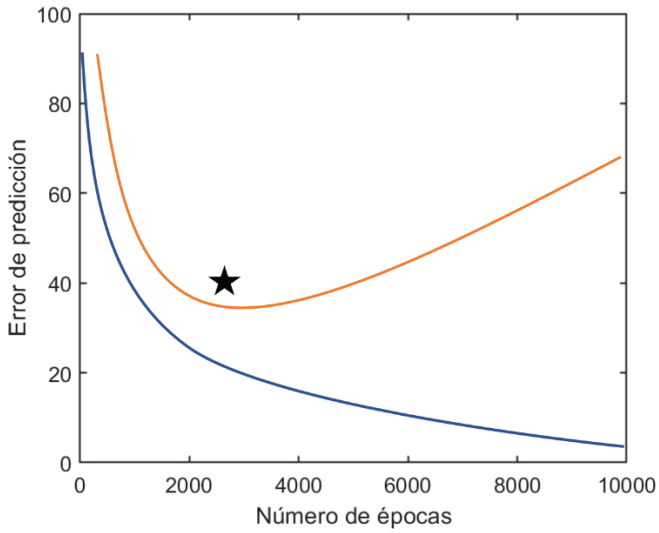


Figura 3.3. Variación del error de predicción del MLP en función del número de épocas para los subconjuntos de entrenamiento (azul) y monitoreo (naranja). La estrella indica la época (iteración) en que se detiene el entrenamiento.

3.3.4. Software.

La manipulación de datos e implementación de pruebas estadísticas se realizó en MATLAB R2017b. Los algoritmos PLS e iPLS fueron llevados a cabo en la interfaz libre de MATLAB MVC1 [107]. Las metodologías de diseño y optimización de experimentos fueron ejecutadas en el software Stat-Ease Design-Expert 8.0.0 (Stat-Ease, Inc., Minneapolis, USA). Los algoritmos PCA y MLP fueron implementados en la interfaz PLS Toolbox 8.7.1 (2019), en su versión de prueba gratuita disponible a través del enlace http://www.eigenvector.com (Eigenvector Research, Inc., Manson, WA USA 98831).

3.4. Resultados y discusión

3.4.1. Inspección de valores atípicos (outliers)

El conjunto de partida de datos espectrales (EEMs desdobladas) estuvo compuesto por un total de 36 muestras, las cuales fueron dispuestas en un único arreglo matricial de tamaño 36×38.500 para luego efectuar la inspección de *outliers* mediante PCA. La implementación del algoritmo se realizó de acuerdo a las especificaciones detalladas en la Sección 2.3.5 del Capítulo 2. Se obtuvo de esta manera un modelo PCA con 2 PCs cuyo PVA fue del 96,5%, lo cual demuestra la gran redundancia de los datos espectrales. Asimismo, la inspección de *outliers* se basó en el diagrama de influencia (residuo Q vs T^2 de Hotelling) y en el diagrama de *scores* en las PCs significativas (PC2 vs PC1). Estos resultados se resumen gráficamente en la Figura 3.4. Como puede observarse, sólo una muestra del lote B2 (día 9) expuso un comportamiento atípico, por lo que fue descartada del conjunto. Un análisis detallado de dicha muestra reveló que esta presentaba regiones de saturación de la señal de fluorescencia que no habían sido advertidas durante la generación del dato. Esta técnica de evaluación de valores atípicos mediante PCA constituye una manera de poner de manifiesto la ventaja de primer orden, en virtud de que los algoritmos son capaces de etiquetar muestras que presentan un comportamiento atípico con respecto al resto del conjunto. Esto tiene particular interés durante la etapa de predicción de muestras incógnitas reales, ya que el modelo tiene la capacidad de identificar muestras atípicas, previniendo así posibles resultados analíticos incorrectos [20].

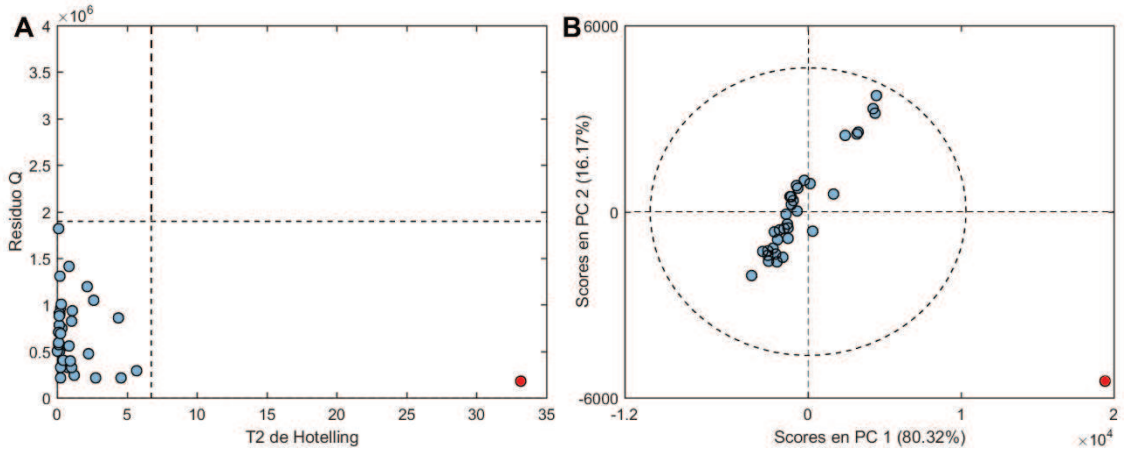


Figura 3.4. Inspección de *outliers* espectrales mediante PCA del bloque X. **A.** Diagrama de influencia (Residuo Q vs T^2 de Hotelling). Las líneas discontinuas representan los límites de confianza estimados con un 0,05 de nivel de significancia; **B.** Scores en PC2 vs scores en PC1. La línea discontinua elíptica representa el límite de confianza bivariado calculado con un nivel de significancia de 0,05. En rojo se representa el valor atípico identificado.

3.4.2. Modelado PLS

Luego de la inspección y remoción de valores atípicos, el conjunto de datos de trabajo estuvo compuesto por un total de 35 muestras. Este se dividió en los subconjuntos de calibración (entrenamiento) y validación, asignando 26 y 9 muestras a cada grupo, respectivamente, de acuerdo a lo detallado en la Sección 3.2.1.

El subconjunto de calibración se utilizó para construir modelos PLS y evaluar la relación entre la señal de fluorescencia y la concentración de etanercept. En este sentido, se empleó el método de selección de variables iPLS para explorar las regiones espectrales de las EEMs de mayor covarianza con la variable respuesta de interés.

El método iPLS consiste en dividir el rango espectral en un número de intervalos predeterminado y ajustar un modelo PLS para cada uno de ellos. De esta manera se puede evaluar cómo varía el RMSECV de PLS en diferentes regiones del dato. Se espera que aquel o aquellos intervalos que arrojen menores valores de RMSECV serán los que más correlación guardan con el bloque Y. Los resultados del iPLS se muestran en la Figura 3.5.A. En la misma, se muestra el RMSECV relativo porcentual para cada uno de los intervalos (barras azules), junto con un espectro normalizado de la matriz de calibrado (verde). Se debe recordar que el "espectro" en realidad corresponde a una EEM desdoblada, por lo que la cantidad total de sensores (puntos) equivale al producto de las dimensiones de cada EEM, es decir $55 \times 700 = 38500$. Los números de sensores del iPLS permiten identificar las regiones espectrales de las EEMs que arrojan los menores valores de RMSECV (rectángulos rojos). En este sentido, con la información de los sensores en el espectro vectorizado, se procedió a identificar en una EEM original, las regiones espectrales a las que corresponden los dos conjuntos de intervalos de menor RMSECV, lo cual se representa en la Figura 3.5.B mediante los rectángulos rojos sobre un diagrama de contorno de una EEM genérica de muestra de fermentación.

Una vez identificadas estas dos regiones, se procedió a calibrar cuatro modelos PLS: modelo 1, con la región seleccionada 1 (RS1); modelo 2, con la región seleccionada 2 (RS2); modelo 3, con las regiones RS1 y RS2 juntas; y modelo 4, con el dato completo. En cada caso, se computaron los RMSECV relativos porcentuales[21] obteniendo, respectivamente, los valores 27,3; 20,6; 20,7 y 19,5%. Como se puede ver, la selección de regiones espectrales no mejora sustancialmente el resultado que se obtiene modelando los datos completos. De esta manera, y teniendo en cuenta que la adquisición de datos en regiones espectrales puntuales requiere un mayor grado de sofisticación instrumental (sobre todo pensando en una implementación automatizada del método), se decidió continuar trabajando con las EEMs completas. En este sentido,

[21] Su definición es equivalente al REP%, sólo que con los datos de la CV.

se seleccionó el modelo 4 para realizar una caracterización más detallada de los resultados que se obtienen mediante el modelado de los datos con PLS.

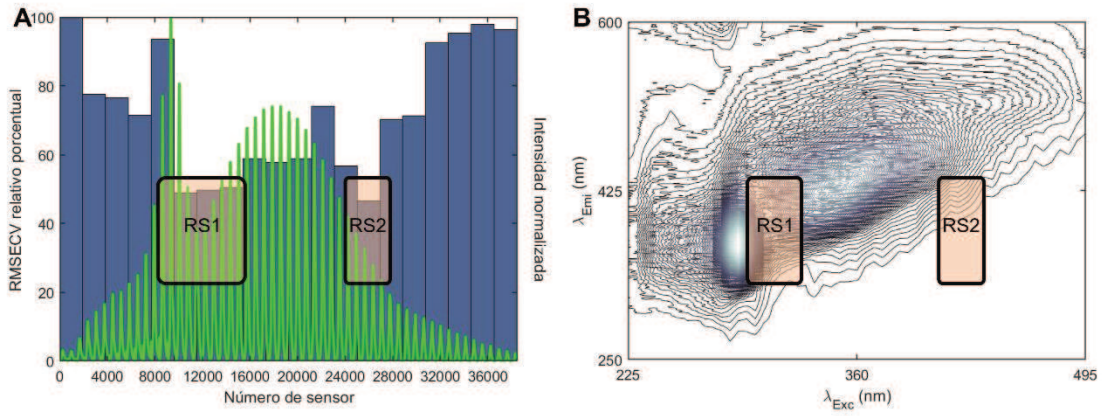


Figura 3.5. Estudio de selección de variables para el modelado PLS. **A.** RMSECV relativo porcentual (barras azules) en función del número de sensores obtenido mediante el método iPLS y espectro (EEM vectorizada) normalizado (línea verde). Los rectángulos rojos indican los sensores para los cuales se obtienen los menores valores de error de predicción. **B.** Mapa de contorno de una EEM genérica de muestra de fermentación en la que se indican las regiones espectrales en 2D a las que corresponden los sensores de mínimo RMSECV relativo porcentual arrojados por el método iPLS.

El criterio de Haaland y Thomas para el modelo PLS final obtenido arrojó un valor óptimo de LVs igual a 5. Las salidas del modelo respecto de su capacidad predictiva se presentan en la Figura 3.6. En la esta se incluyen los resultados de concentración predicha vs nominal para el procedimiento de LOO-CV (Figura 3.6.A) y para el conjunto de validación externa (Figura 3.6.B). Asimismo, los valores de los estadísticos $RMSEC_{PLS}$, $RMSECV_{PLS}$ y R^2-CV_{PLS} asociados a la calibración fueron, respectivamente, 7,5 mg L^{-1}, 23,0 mg L^{-1} y 0,878, mientras que los estadísticos $RMSEP_{PLS}$, $REP\%_{PLS}$ y R^2-$pred_{PLS}$ asociados a la validación externa fueron 15,4 mg L^{-1}, 12,6% y 0,956, respectivamente.

Como se observa claramente en la Figura 3.6, existe una relación no lineal entre la señal de fluorescencia y la concentración de etanercept, especialmente a valores de concentración bajos. Además de la inspección visual, la prueba estadística APaRP arrojó como resultado un p-valor de 0,01, indicando el rechazo de la hipótesis nula formulada y, en consecuencia, confirmando que la relación entre señal espectral y concentración es efectivamente no lineal.

Los resultados obtenidos indican que el modelado PLS no resulta una estrategia conveniente para el desarrollo de un modelo de calibración para este sistema. Teniendo en cuenta las condiciones experimentales de generación de los datos espectrales y la discusión realizada en el Capítulo 2, es evidente que existe un fuerte efecto matriz sobre

el analito, especialmente a concentraciones menores a los 100 mg L^{-1}. Es interesante destacar además que, si bien en el diagrama de *scores* de PCA (Figura 3.4.B) se observa una relación lineal en la variación de la información espectral en función al número de muestra (tiempo de cultivo), debido a que el método de descomposición es no supervisado, este no permite evidenciar la no linealidad con la concentración de proteína, como sí lo hace PLS. En este sentido, la no linealidad del sistema también puede interpretarse en virtud de la manera en que evoluciona la concentración de etanercept en función del tiempo de cultivo. En la Figura 2.2 del Capítulo 2 se observa claramente que esta variable no evoluciona de manera lineal durante el proceso.

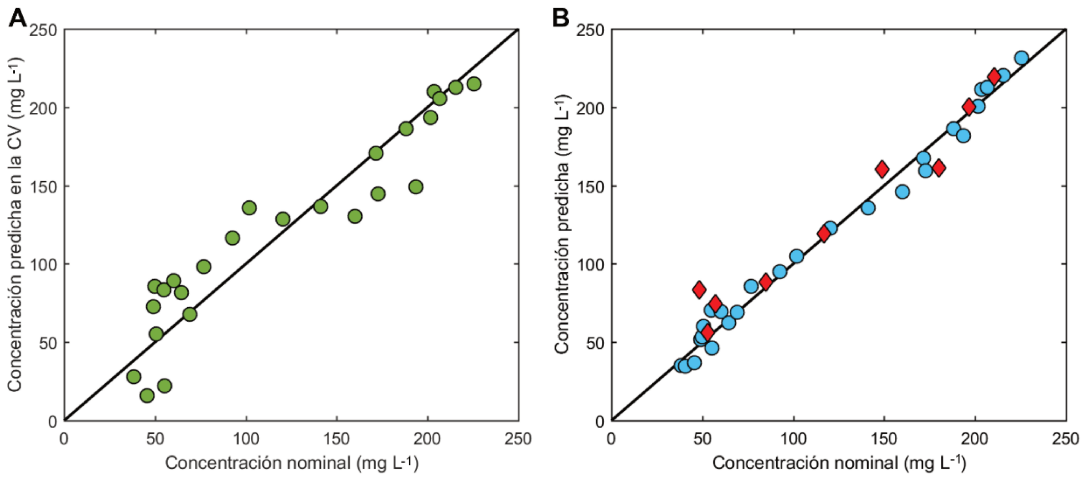


Figura 3.6. Predicciones del modelo PLS (5 LVs). **A.** Concentración de etanercept predicha en la CV vs concentración nominal (conjunto de calibrado). **B.** Concentración de etanercept predicha vs concentración nominal para el conjunto de calibración (círculos celestes) y validación externa (rombos rojos). En todos los casos, la línea negra representa la curva de ajuste ideal 1:1.

3.4.3. Optimización, entrenamiento y validación del método de calibración basado en MLP-ANN

Los resultados obtenidos mediante la estrategia PLS justifican la necesidad de un enfoque de modelado no lineal. En este sentido, se optó por desarrollar un método de calibración basado en la red neuronal MLP. La optimización de su arquitectura se efectuó mediante la RSM, de acuerdo a lo descripto en la Sección 3.2.3. Los 17 experimentos sugeridos por el BBD se muestran en la Tabla 3.1.

Los datos compilados en la Tabla 3.1 se utilizaron para modelar estadísticamente la relación entre la arquitectura de la red y su capacidad predictiva, utilizando una estrategia de regresión múltiple. A fin de encontrar el modelo de regresión múltiple que mejor ajustara el conjunto de datos para cada respuesta, se evaluó la significancia de cada factor mediante el test ANOVA. En todos los casos, el modelo lineal resultó significativo. En particular, la respuesta $RMSEC_{BBD}$ fue transformada para mejorar el

ajuste. Más detalles sobre la transformación de respuestas puede encontrarse en la literatura especializada (por ejemplo, [17]). En la Tabla 3.2 se resumen los resultados del ANOVA para el modelado de cada respuesta, indicando los factores que resultaron significativos, el tipo de transformación de la respuesta, los p-valores para evaluar la significancia del modelo y su falta de ajuste y los coeficientes de determinación (R^2) y de determinación ajustado (R^2-adj). Estos últimos dos parámetros representan el porcentaje de varianza capturada por cada modelo.

Tabla 3.1. Experimentos del BBD para la optimización del modelo de calibración MLP.

Orden estándar[a]	**Orden de ensayo**[b]	**Factor A** A	**Factor B** $N1$	**Factor C** $N2$	**Resp. 1** **$RMSEC_{BBD}$**	**Resp. 2** **$RMSECV_{BBD}$**	**Resp. 3** **R^2-CV_{BBD}**
1	2	4	2	4	10,82	18,38	0,922
2	5	12	2	4	4,68	14,07	0,955
3	16	4	10	4	7,97	17,65	0,929
4	17	12	10	4	5,27	13,13	0,961
5	4	4	6	0	10,03	17,55	0,929
6	14	12	6	0	5,96	12,89	0,963
7	10	4	6	8	9,65	16,04	0,941
8	6	12	6	8	5,15	13,33	0,960
9	8	8	2	0	6,02	16,67	0,936
10	9	8	10	0	7,53	17,53	0,93
11	11	8	2	8	6,15	14,84	0,949
12	13	8	10	8	7,64	18,52	0,921
13	3	8	6	4	6,99	17,22	0,932
14	12	8	6	4	6,76	19,70	0,911
15	15	8	6	4	6,86	23,26	0,875
16	7	8	6	4	7,15	16,84	0,935
17	1	8	6	4	6,87	16,91	0,934

[a]Se refiere al orden estándar del diseño.
[b]Se refiere al orden de ejecución de los experimentos para garantizar la aleatoriedad.

Tabla 3.2. Modelado de las respuestas experimentales mediante RSM, a partir de los resultados obtenidos de los experimentos del BBD.

Respuesta	Formulación del modelo en los predictores	Factor/es significativos[a,b]	Transformación de la respuesta[c]	p-valor del ANOVA[d]		R^2-adj	R^2
				Modelo	Falta de ajuste		
RMSEC	Lineal	A	Raíz cuadrada inversa	<0,0001	0,0034	0,84	0,83
RMSECV	Lineal	A	Ninguna	0,0020	0,4655	0,50	0,47
R^2-CV	Lineal	A	Ninguna	0,0029	0,5538	0,48	0,44

[a]A: número de nodos en A; B: número de nodos en $N1$; C: número de nodos en $N2$.
[b]Términos con p-valores menores que 0,05.
[c]Ver Ref. [17].
[d]Se considera significativo cuando el p-valor es menor que 0,05.

Los resultados mostrados en la Tabla 3.2 indican que para las tres respuestas se pudieron obtener modelos significativos, a pesar de que algunos de los estadísticos asociados no fueron tan satisfactorios. En este contexto, resulta notable que el único factor significativo para las tres respuestas sea el número de nodos en la capa de entrada (Factor A). En particular, en el caso del $RMSEC_{BBD}$, los valores de R^2 y R^2-adj son relativamente buenos, aunque el modelo presenta una falta de ajuste significativa. Por otra parte, las respuestas $RMSECV_{BBD}$ y R^2-CV_{BBD} presentan valores poco satisfactorios de los coeficientes de determinación. No obstante, con el objetivo de encontrar una topología óptima para el modelo MLP, se procedió a implementar la función deseabilidad para la optimización simultánea de las tres respuestas. Para este fin, los criterios de optimización de la función deseabilidad fueron la minimización de los estadísticos $RMSEC_{BBD}$ y $RMSECV_{BBD}$, y la maximización del R^2-CV_{BBD}, otorgando mayor importancia relativa al $RMSECV_{BBD}$, ya que es el que mejor representa la capacidad de generalización de la red. Para la mejor solución de deseabilidad obtenida (con un valor de 0,053), en la Figura 3.7.A se muestra la superficie de respuesta D como función del número de neuronas en la capa A y de la $N1$, para un valor fijo de $N2$. Como puede verse, el número de neuronas de entrada debe ser alto, mientras que los valores para $N1$ y $N2$ no son críticos. Teniendo esto en consideración y atendiendo al principio de parsimonia, se optó por el modelo más simple, es decir, aquel con una arquitectura $12-2-0-1$ (eligiendo los valores mínimos del entorno experimental para las capas ocultas). Asimismo, es interesante mencionar que el óptimo de PCs para la compresión del bloque X que predice el criterio de Haaland y Thomas es de apenas 5 componentes (bastante menor al óptimo de 12 determinado mediante la RSM).

Para la arquitectura final elegida, los valores predichos para $RMSEC_{BBD}$, $RMSECV_{BBD}$ y R^2-CV_{BBD} por el modelo de regresión múltiple RSM fueron 5,2 ± 0,3 mg L^{-1}, 14,3 ± 1,4 mg L^{-1} y 0,950 ± 0,010, respectivamente. El modelo MLP optimizado arrojó los valores 5,4 mg L^{-1}, 13,9 mg L^{-1} y 0,956 para los estadísticos $RMSEC_{MLP}$, $RMSECV_{MLP}$ and R^2-CV_{MLP}, respectivamente. Asimismo, tras efectuar predicciones con el conjunto de validación externa, los valores de $RMSEP_{MLP}$, $REP\%_{MLP}$ y R^2-$pred_{MLP}$ fueron 8,6 mg L^{-1}, 7,1% y 0,982, respectivamente. En la Figura 3.7 también se presentan los resultados de las predicciones efectuadas con el modelo MLP optimizado, durante la LOO-CV (Figura 3.7.B) y con el conjunto de validación externa (Figura 3.7.C).

Es importante destacar que, a pesar de que los resultados del ajuste de los modelos de RSM por mínimos cuadrados (LS) no fueron tan buenos, el hecho de haber obtenido valores experimentales de las tres respuestas que caen dentro de los límites predichos por los modelos de RSM para la mejor solución de D, confirma que la estrategia de RSM-LS fue adecuada para la optimización del MLP. Por otro lado, como

criterio adicional para validar estos resultados, se procedió a modelar la misma matriz de factores y respuestas (Tabla 3.1) mediante una técnica no paramétrica, basada en el uso de redes neuronales con funciones de base radial (RBF-ANN) [108], arrojando resultados muy similares y confirmando, de esta manera, la potencia del método paramétrico de ajuste de las respuestas experimentales.

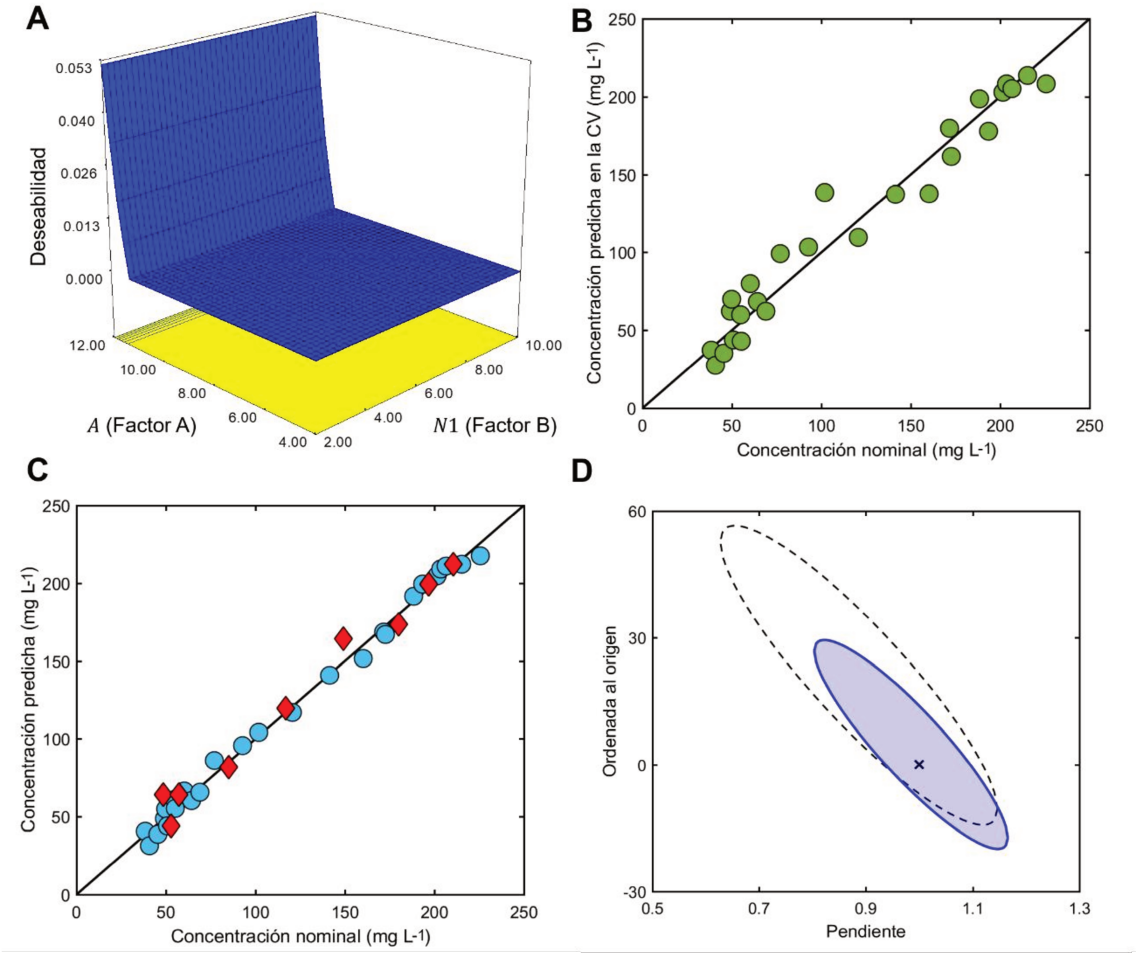


Figura 3.7. A. Resultados de la optimización del modelo MLP mediante la RSM. Deseabilidad como función del número de neuronas en la capa de entrada y capa oculta 1, para un valor de 0 neuronas (fijo) en la segunda capa oculta. **B.** Concentración de etanercept predicha en la CV vs concentración nominal. **C.** Concentración de etanercept predicha vs nominal para el conjunto de calibración (círculos celestes) y validación externa (rombos rojos). En ambos casos, la línea negra representa la curva de ajuste ideal 1:1. **D.** Elipse de confianza del test EJCR para la evaluación de la exactitud y precisión del método MLP (celeste). El asterisco corresponde al punto cuyas coordenadas representan los valores teóricos de pendiente y ordenada al origen de la recta de regresión ideal 1:1 (1 y 0, respectivamente). En líneas negras discontinuas, se muestra la elipse de confianza que se obtendría con las predicciones de PLS (a modo ilustrativo).

El modelo MLP obtenido muestra un rendimiento analítico notablemente superior a PLS dada la naturaleza no lineal del problema. En particular, como se mencionó

anteriormente, los resultados sugieren que la respuesta espectral del analito es fuertemente afectada por la matriz de la muestra, especialmente a valores bajos de concentración. Esto es consistente con el hecho que el modelo PLS arroja valores de predicción relativamente buenos para concentraciones de etanercept mayores a los 100 mg L^{-1} (Figura 3.6.B). No obstante, el método MLP permite no sólo una mejor precisión, sino la cuantificación de la proteína recombinante en un rango de concentración más amplio.

Finalmente, el test EJCR se utilizó para evaluar la exactitud y precisión del modelo de calibración MLP (Figura 3.7.D). En este sentido, es posible afirmar que el método es exacto, en virtud de que la elipse de confianza contiene al punto cuyas coordenadas representan los valores teóricos de pendiente y ordenada al origen de la curva de calibrado ideal 1:1 (1 y 0, respectivamente). Asimismo, la escala de los ejes de la figura permite inferir que el área de la elipse de confianza es razonablemente pequeña, lo cual está directamente ligado a la precisión del método. Esto también puede confirmarse observando la elipse de confianza que se obtendría con las predicciones de PLS (Figura 3.7.D, en líneas negras discontinuas).

3.5. Conclusiones del capítulo

La metodología de calibración desarrollada representa una estrategia multivariada novedosa para la cuantificación (monitoreo) *at-line* de etanercept en un proceso fermentativo, la cual se basa en datos de fluorescencia de segundo orden que son directamente adquiridos a partir de muestras del bioproceso.

El MLP demostró ser un algoritmo adecuado para el desarrollo del método de calibración, cuyo rendimiento analítico demostró ser superior al de PLS, en virtud de la no linealidad de los datos evidenciada en el sistema bajo estudio. Esto último, además, resulta interesante en el sentido del desarrollo de estrategias de PAT ya que la no linealidad en la relación señal-concentración resulta frecuente cuando se desean adquirir señales analíticas directamente desde el seno del bioproceso.

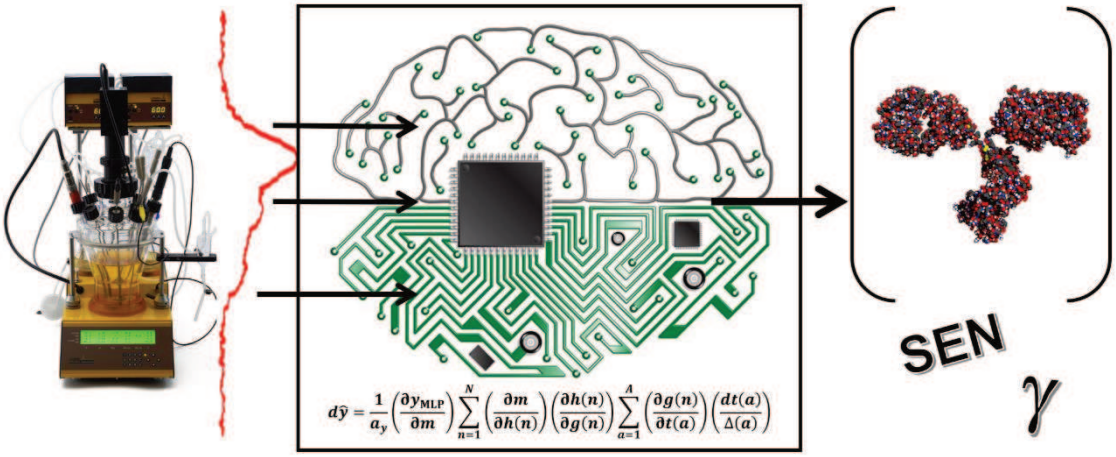
El método analítico desarrollado se encuentra en consonancia con los principios de la PAT, debido a que se trata de un método basado en una técnica instrumental que no requiere pretratamiento alguno de la muestra y que, además, es robusta, rápida, automatizable y relativamente económica.

Una limitación importante en la actualidad que concierne a los modelos basados en redes neuronales es que, a diferencia de sus contrapartes paramétricas, no están tan bien caracterizados desde el punto de vista estadístico. Esto impacta directamente en la posibilidad de calcular cifras analíticas de mérito (AFOMs) para una caracterización

más rigurosa, comparabilidad y transferencia/validación en la industria de este tipo de métodos de calibración. En este sentido, el estudio de la propagación de errores en modelos no paramétricos es un tema de gran interés en investigación estadística y quimiométrica. De esta manera, el próximo y último capítulo de esta tesis está dedicado al estudio de algunas AFOMs en modelos de calibración de tipo MLP.

CAPÍTULO 4

Cifras analíticas de mérito y calibración con redes neuronales. Estimación de la sensibilidad para el caso del perceptrón multicapa



La imagen del biorreactor corresponde a una fotografía tomada de https://www.lambda-instruments.com/es/fermentador-y-biorreactor/

4.1. Introducción

Tal como se describió en la Introducción de la presente tesis, una etapa fundamental durante el desarrollo de un método de calibración consiste en la caracterización de su rendimiento analítico, lo que se realiza mediante el cálculo de las denominadas cifras analíticas de mérito (AFOMs).

Las AFOMs constituyen un conjunto de parámetros numéricos que sirven para caracterizar la *performance* de una metodología de análisis, esencialmente en términos de su habilidad predictiva y capacidad de detección. En este sentido, por ejemplo, las AFOMs se utilizan durante el desarrollo de un método como criterio para la optimización de las condiciones experimentales que permitan obtener la mejor *performance* analítica. Por otro lado, las AFOMs constituyen una herramienta clave en la validación de metodologías, no sólo porque brindan indicadores numéricos simples que dan cuenta de sus bondades analíticas, sino porque, además, permiten establecer comparaciones objetivas entre diferentes métodos analíticos. Este último aspecto cobra una importancia fundamental en lo que respecta a la validación de metodologías ante autoridades regulatorias para su transferencia a la industria, ya que buena parte de los estándares internacionales para la validación de un método fijan determinados valores de AFOMs que este debe cumplir para ser aplicado como parte de las herramientas de control de calidad de un proceso industrial [28].

Una revisión bibliográfica permite evidenciar en las últimas dos décadas, un interés creciente en la industria en relación a la aplicación de técnicas espectrales para el monitoreo de procesos [109-111] y, en particular, en el uso de ANNs para el tratamiento de datos [112-117]. Esta situación se basa en el hecho de que la generación de señales analíticas mediante detección espectral en línea se suele realizar a partir de muestras complejas y/o saturadas, razón por la cual la presencia de no linealidades en la relación señal-concentración es más frecuente. En particular, el uso de redes de tipo MLP se ha extendido notablemente en química analítica, debido a la mayor flexibilidad que presentan estos algoritmos en comparación con los modelos clásicos, tales como PCR y PLS [118-120]. No obstante, a diferencia de los métodos paramétricos, el MLP ha sido menos caracterizado desde el punto de vista estadístico. Esto implica esencialmente que, si bien estos métodos han demostrado ser muy útiles desde el punto de vista predictivo en contextos no lineales, los elementos teóricos que permiten realizar inferencias sobre estos modelos aún no han sido completamente desarrollados. En este sentido, la posibilidad de estimar la incertidumbre asociada a la predicción de un MLP impacta de manera directa sobre la capacidad de calcular las AFOMs de un modelo de calibración. Esto trae aparejadas consecuencias de índole tecnológicas ya que la

incapacidad para el cálculo de AFOMs representa un cuello de botella para la validación e implementación de protocolos basados en ANNs en la industria.

Si bien es cierto que para el caso del MLP (y otras ANNs en general), la propagación de errores y la estimación de intervalos de predicción ha sido explorada en la literatura [121-123], tales contribuciones permanecen aún por fuera del área de la química analítica y no es posible todavía efectuar el cálculo de AFOMs en modelos de calibración MLP de manera rigurosa. En este sentido, Allegrini y Olivieri en 2016 desarrollaron las ecuaciones para el cálculo de AFOMs en redes basadas en funciones de base radial (RBF), las cuales constituyen otra clase de ANN comúnmente empleada para calibración [124]. Por otro lado, la propagación de errores también ha sido empleada para discutir cuestiones relacionadas con la sensibilidad[22], y su aplicación tanto para la optimización de modelos como para la selección de variables [125-129].

En base a lo expuesto es que en esta última instancia del trabajo de tesis se han utilizado las contribuciones previas como marco teórico y metodológico para desarrollar ecuaciones referidas al cálculo de AFOMs en modelos MLP de calibración de primer orden. En particular, en este capítulo se reportan, por primera vez, las ecuaciones para la estimación de la sensibilidad (SEN) y la sensibilidad analítica (γ) en calibración MLP. Las ecuaciones desarrolladas han sido validadas a partir de dos enfoques: simulación Montecarlo y *bootstrap*.

Finalmente, las contribuciones de este capítulo se aplicaron para la caracterización de la sensibilidad del modelo de calibración PAT desarrollado en el Capítulo 3 para el monitoreo de etanercept, en forma comparativa con el método univariado de referencia (HPLC).

4.1.1. Estimación de SEN *y* γ *en calibración de primer orden. Marco teórico y metodológico*

Una de las AFOMs que reviste una importancia fundamental al momento de caracterizar un modelo de calibración es la SEN. La IUPAC establece diferentes definiciones de este parámetro, según el escenario de calibración del que se trate (univariado o multivarido) [130]. Sin embargo, existe una definición intuitiva general que establece que, en términos cualitativos, la SEN es la respuesta que tiene un modelo (cambio en señal analítica) frente a un estímulo (cambio en la concentración de analito).

[22] El término "sensibilidad" puede generar cierta confusión, ya que adquiere diferente significado según el contexto. En general, desde el punto de vista matemático, la sensibilidad de un modelo que depende de un conjunto de parámetros (ya sea determinístico o estocástico) se puede entender como una medida que refleja cuánto se modifica la salida o predicción del modelo dado un valor fijo de la variable de entrada, para diferentes valores de los parámetros que lo definen. En cambio, en química analítica, la sensibilidad de un modelo de calibración a parámetros fijos es una razón de cambio entre señal y concentración, es decir, entre las variables de entrada y salida, respectivamente.

Esto implica que la SEN de un modelo se relaciona directamente con su capacidad para distinguir entre cambios pequeños de concentración del analito. De este hecho se deriva la idea intuitiva de que un método será más "sensible" cuanto mayor sea su capacidad para detectar entre muestras con variaciones pequeñas en la concentración del analito. En consecuencia, la importancia en la estimación de SEN radica en que dicho parámetro está directamente asociado con otras AFOMs importantes:

- γ, constituye una medida relativizada de la SEN, que es independiente del tipo de señal instrumental y que posibilita la comparación entre métodos desarrollados con diferentes técnicas analíticas;
- selectividad, que permite evaluar la posibilidad de cuantificar un analito en presencia de interferentes;
- incertidumbre de predicción y límites de detección (LOD) y de cuantificación (LOQ), parámetros clave para caracterizar la precisión y la capacidad de detección de un método analítico.

Es importante mencionar que, si bien existen numerosas AFOMs y de diferente jerarquía, todas las cifras anteriormente mencionadas se encuentran directa o indirectamente relacionadas con lo que Valcárcel y Ríos (1999) [131] definen como las propiedades supremas de un método de calibración analítica: exactitud y representatividad.

Según establece la IUPAC para la calibración univariada clásica, la SEN se define como la pendiente de la curva de calibración [132,133], lo que puede interpretarse como una variación en señal (x) por variación unitaria en concentración (y)[23]. Este resultado se deriva directamente de la ecuación de predicción para un modelo de calibración de orden cero. Sea el modelo poblacional asociado a una calibración univariada

$$x = \beta_0 + \beta_1 y + \varepsilon \tag{4.1}$$

donde β_0 y β_1 son los parámetros a estimar y ε, el error del modelo, la estimación de la concentración $\hat{y}$ a partir de la señal medida para un muestra incógnita x se calcula como:

$$\hat{y} = \frac{x - \widehat{\beta_0}}{\widehat{\beta_1}} \tag{4.2}$$

donde $\widehat{\beta_0}$ y $\widehat{\beta_1}$ representan la ordenada al origen y la pendiente de la recta de regresión, respectivamente (parámetros estimados). De la Ec. (4.2), se sigue que, bajo el supuesto de calibración infinitamente precisa, la SEN es igual a:

$$\mathrm{SEN} = \frac{dx}{dy} = \widehat{\beta_1} \tag{4.3}$$

[23] Si bien en calibración univariada se suele simbolizar a la señal analítica como y, aquí se prefirió utilizar una notación análoga a la calibración multivariada para evitar conflictos.

Por otra parte, en calibración de primer orden (por ejemplo PLS), la estimación $\hat{y}$ a partir de una señal multivariada de una muestra test $\mathbf{x}$, se define como:

$$\hat{y} = \mathbf{b}_{\mathrm{PLS}}^{\mathrm{T}}\mathbf{x} \tag{4.4}$$

donde $\mathbf{b}_{\mathrm{PLS}}$ es el vector de coeficientes de regresión del modelo. Sin embargo, en este escenario de calibración, no es posible extender de manera directa la definición de SEN univariada, ya que la señal espectral representada por $\mathbf{x}$ contiene información tanto del analito como de los interferentes. Una de las soluciones que se propusieron originalmente para solucionar este problema fue la introducción del concepto de señal neta de analito (NAS) [134]. Sin embargo, debido a las limitaciones prácticas para el cálculo de la NAS, la definición moderna general de SEN para calibración multivariada, se basa en la teoría de propagación de errores [130]. Bajo los supuestos de ruido independiente e idénticamente distribuido (*iid*) y calibración infinitamente precisa ($\mathbf{b}_{\mathrm{PLS}}$ no es afectado por la incertidumbre de la calibración), es posible probar, mediante la teoría de propagación de errores, que la incertidumbre en la predicción del modelo PLS (σ_y) responde a la expresión [126,135]:

$$\sigma_y = \|\mathbf{b}_{\mathrm{PLS}}\|\sigma_x \tag{4.5}$$

donde σ_x es la incertidumbre o desvío estándar asociado a la señal instrumental (ruido). De esta manera, se define la SEN como la inversa de la norma euclídea del vector de coeficientes de regresión, es decir [136]:

$$\mathrm{SEN} = \frac{1}{\|\mathbf{b}_{\mathrm{PLS}}\|} \tag{4.6}$$

o bien, también se deduce de la Ec. 4.5 que:

$$\mathrm{SEN} = \frac{\sigma_x}{\sigma_y} \tag{4.7}$$

Las igualdades que establecen las Ecs. 4.6 y 4.7 representan, respectivamente, las definiciones teórica y operacional de la SEN. Por otro lado, la Ec. 4.7 muestra una analogía directa con la definición de SEN para calibración univariada y también constituye la base metodológica para la estimación de SEN mediante simulación Montecarlo, que se describe en la siguiente sección. Este enfoque ha sido ampliamente utilizado para la derivación y validación de ecuaciones de SEN en diversos escenarios de calibración multivariada [28,137].

En calibración MLP no existe una expresión análoga simple de la Ec. 4.6. Sin embargo, la estrategia para su deducción en este trabajo se basó en encontrar una expresión equivalente para $\mathbf{b}_{\mathrm{PLS}}$, derivando la ecuación de predicción del modelo MLP, de manera tal de poder escribir:

$$\mathrm{SEN} = \frac{1}{\|\mathbf{b}_{\mathrm{MLP}}\|} \tag{4.8}$$

Resulta importante mencionar que la elección de la notación empleada es para establecer una analogía con la calibración PLS, pero nada tiene que ver con la idea de que en un modelo de red neuronal se estiman coeficientes de regresión. Por otro lado, tal como se mostrará más adelante y a diferencia de lo que ocurre en PLS, el vector $\mathbf{b}_{\mathrm{MLP}}$ resulta dependiente de la concentración de analito, lo que implica que cada muestra puede ser caracterizada por un valor de sensibilidad. Este hecho es esperable debido a que, en calibración no lineal, la pendiente de la función que relaciona la señal con la concentración no es constante.

Una vez que se obtiene el parámetro de SEN, es posible entonces calcular γ como [138]:

$$\gamma = \frac{\mathrm{SEN}}{\sigma_x} \tag{4.9}$$

Este parámetro también resulta muestra-dependiente. Tal como se mencionó anteriormente, la γ resulta un parámetro muy útil para la comparación de métodos analíticos que han sido desarrollados a partir de fuentes instrumentales diferentes.

4.2. Objetivos específicos del capítulo

En este capítulo se plantearon los siguientes objetivos específicos:

- derivar las expresiones matemáticas que permitan estimar de manera teórica la SEN y la γ de un modelo de calibración basado en MLP-ANN;
- validar las ecuaciones desarrolladas mediante datos simulados y experimentales;
- caracterizar la sensibilidad del método PAT desarrollado en el Capítulo 3 y compararla con la del método HPLC de referencia.

4.3. Materiales y métodos

4.3.1. Validación de las ecuaciones desarrolladas a partir de datos simulados mediante método Montecarlo

Uno de los enfoques empleados para validar las ecuaciones propuestas para el cálculo de SEN consistió en llevar a cabo una simulación de tipo Montecarlo. La simulación Montecarlo constituye una técnica estadística numérica que permite aproximar una solución mediante simulaciones aleatorias de un gran número de datos [139]. Su aplicación en este trabajo consistió en obtener una estimación empírica de la SEN asociada a un modelo de calibración, a partir de lo que establece la Ec. 4.7. En este sentido, asumiendo una calibración fija libre de ruido, es posible explorar la propagación

de la incertidumbre desde la señal a la concentración predicha, mediante el agregado de una pequeña cantidad de ruido *iid* a un conjunto de señales multivariadas de muestras test, y luego evaluar la estadística asociada a las predicciones.

Para llevar adelante esta estrategia, se generaron cuatro sistemas multianalito de datos simulados, cada uno caracterizado por un tipo de no linealidad diferente. Cada sistema estuvo formado por un analito en presencia de tres interferencias. Los espectros a concentración unitaria de los cuatro componentes, utilizados para generar la totalidad de los datos simulados, se muestran en la Figura 4.1.A. Se puede observar que todos están parcialmente solapados, sobre una región espectral de 200 longitudes de onda (escala arbitraria). En todos los casos, el analito es el componente 1 (negro), mientras que los componentes 2-4 (tonos celestes) representan los interferentes. En cada sistema, se consideró un tipo de no linealidad particular para generar la relación entre señal y concentración. A saber:

a. sistema A (cuadrático):

$$x(j) = \sum_{n=1}^{4} s_n(j)(y_n + y_n^2) \tag{4.10}$$

b. sistema B (irracional):

$$x(j) = \sum_{n=1}^{4} s_n(j) y_n^{0.3} \tag{4.11}$$

c. sistema C (sigmoideo):

$$x(j) = \sum_{n=1}^{4} s_n(j) \frac{1}{1 + e^{-(y_n - 1)}} \tag{4.12}$$

d. sistema D (logarítmico):

$$x(j) = \sum_{n=1}^{4} s_n(j) \ln y_n \tag{4.13}$$

En las Ecs. 4.10-4.13, $x(j)$ es la señal de la muestra a la longitud de onda j, n es el número total de componentes, $s_n(j)$ e y_n representan, respectivamente, la señal a la misma longitud de onda y la concentración del n-ésimo componente puro. En la Figura 4.1.B se muestra cómo varía la señal total (medida como el valor de la norma euclídea de $\mathbf{x}$) en función del componente 1 (analito) puro.

La generación de los datos y el procedimiento empleado para las etapas de calibración y predicción fue el mismo para los cuatro sistemas. En cada caso, las muestras del conjunto de calibración estuvieron formadas por los cuatro componentes, cuyas concentraciones se establecieron en el rango de 1-5 en unidades arbitrarias, de acuerdo a una matriz de diseño factorial completo de tipo 5^4 (en total, 625 muestras de

calibración). Para los conjuntos de calibrado no se incorporó ningún nivel de ruido, ni en señales ni en concentraciones. Esto se debe al hecho de que en la definición operacional de SEN (Ec. 4.7) se requiere del agregado de ruido *iid* únicamente en las muestras de predicción, para evaluar cómo este se propaga a la concentración estimada, sin tener en cuenta el efecto de la incertidumbre que aporta calibración. Como ejemplo representativo, en la Figura 4.1.C se muestran 30 muestras arbitrarias del conjunto de calibrado simulado para el sistema A.

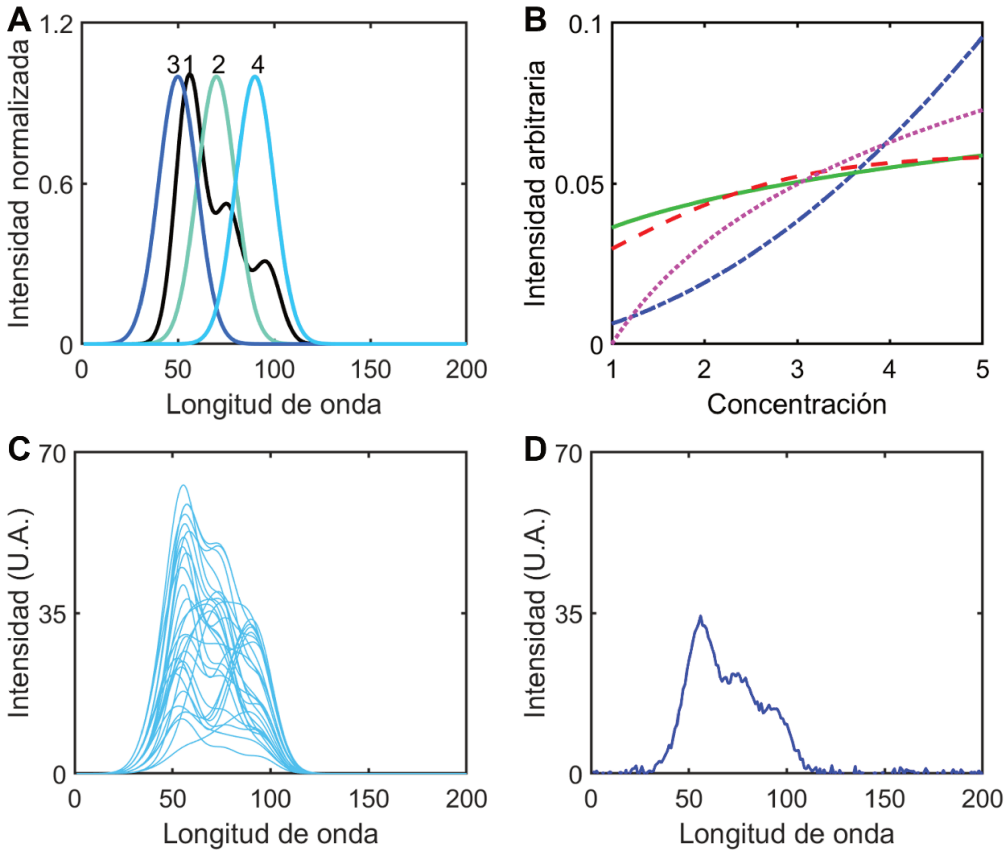


Figura 4.1.A. Espectros a concentración unitaria de los cuatro componentes utilizados para la generación de todos los sistemas simulados. En cada modelo de calibración, el componente 1 (negro) es el analito, mientras que los componentes 2-4 son las interferencias. **B.** Señal total vs concentración del analito puro para los sistemas A-D: cuadrático (líneas discontinuas azules), irracional (línea verde continua), sigmoideo (línea discontinua roja) y logarítmico (líneas discontinuas magenta). **C.** Espectros de 30 muestras arbitrarias del conjunto de calibrado generado para el Sistema A. **D.** Una muestra test genérica del Sistema A, simulada con 4,5 unidades de concentración y 1% de nivel de ruido.

Por otra parte, para cada sistema no lineal, se generaron 11 conjuntos de muestras test de 500 réplicas cada uno, conteniendo un nivel de concentración de analito comprendido en el rango de 1-5 unidades de concentración (11 puntos equidistantes, diferentes del calibrado) y el resto de los componentes a un valor fijo. Además, por cada nivel de concentración del analito, se adicionó a las señales

espectrales una cantidad creciente de ruido *iid*. Se evaluaron cinco niveles de ruido: 0,01; 0,05; 0,10; 0,50 y 1,00% del máximo de señal en el conjunto de calibración. Para cada sistema, el número total de conjuntos test individuales generados fue de 55, con un total de muestras de 27.500 (11 niveles de concentración, 5 niveles de ruido, 500 réplicas). En la Figura 4.1.D se muestra una muestra test genérica simulada con un nivel de concentración de 4,5 unidades y un nivel de ruido del 1,00%, a modo de ejemplo.

Con el juego de calibración se procedió a entrenar un MLP de tres capas, utilizando para ello un algoritmo *ad hoc* escrito en MATLAB. Algunas de sus características son las siguientes:

- tipo de función de transferencia: sigmoidea;
- inicialización de pesos al azar;
- método de entrenamiento: algoritmo de *error back-propagation*;
- criterio de finalización del entrenamiento: estabilización de la raíz cuadrada del error cuadrático medio en el conjunto de monitoreo (RMSEM). Para ello, se partió al azar el conjunto de calibrado en los subconjuntos de entrenamiento y monitoreo, utilizando 60% y 40% del número total de muestras de calibración, respectivamente;
- se incluyen por defecto dos neuronas *bias*, una en la capa de entrada y otra, en la capa oculta, con sus respectivos pesos.

Con respecto a la arquitectura, en todos los casos, la capa de entrada constó de 4 neuronas, ya que las entradas para los MLP fueron *scores* obtenidos de la compresión de la matriz de calibrado mediante PCA. Por otra parte, todas las redes constaron de una única neurona en la capa de salida (predicción de una única variable) y, para determinar la cantidad de neuronas ocultas, se seleccionó como valor óptimo el menor número de neuronas que minimizara el RMSEM.

Una vez entrenados los MLP de los cuatro sistemas, se procedió a realizar las predicciones en cada uno de los 55 conjunto de datos test correspondientes. Cada ciclo de predicción se basó en realizar 500 estimaciones de la concentración del analito a un dado nivel de concentración y un determinado nivel de ruido. Como prueba de concepto para validar las ecuaciones desarrolladas para el cálculo teórico de SEN (Ec. 4.8) se llevó a cabo la siguiente estrategia. Con los resultados de las predicciones, en primer lugar, se computó en cada ciclo el valor de $\mathbf{b}_{\mathrm{MLP}}$ promedio que permitió efectuar el cálculo teórico de SEN a cada nivel de concentración de analito, de acuerdo a la Ec. 4.8. Por otra parte, la Ec. 4.7 permite establecer que:

$$\sigma_x = \mathrm{SEN}\ \sigma_y \tag{4.14}$$

En la simulación Montecarlo, cada ciclo de predicción de 500 réplicas permitió obtener un valor de desviación estándar en concentración, el cual resulta representativo del parámetro σ_y en la Ec. 4.14, mientras que los valores de σ_x corresponden al nivel de ruido resultante en las muestras test tras el agregado de distintos niveles de ruido *iid* durante la simulación. De esta manera, fue posible obtener una estimación empírica de SEN ($\mathrm{SEN_{MC}}$) como la pendiente de la recta de regresión calculada a partir de los valores σ_x vs σ_y. Finalmente, los valores de SEN obtenidos de manera teórica y mediante método Montecarlo fueron comparados estadísticamente con el test EJCR [106].

4.3.2. Validación de las ecuaciones desarrolladas a partir de datos experimentales mediante técnica de bootstrap

Para poder validar las ecuaciones desarrolladas a partir de un sistema experimental, se utilizó el mismo juego de datos de fluorescencia del Capítulo 3 (ver Sección 3.3.1). En primer lugar, se procedió a estimar un valor teórico de SEN para cada muestra del conjunto, computando el valor de $\mathbf{b}_{\mathrm{MLP}}$ mediante una estrategia de entrenamiento-predicción tipo *leave-one-out*, empleando para ello la red MLP desarrollada en el Capítulo 3.

Por otra parte, la prueba de concepto para validar el cálculo teórico de SEN a partir de datos experimentales se llevó a cabo mediante la técnica de *bootstrap*. Este método constituye una valiosa técnica estadística de remuestreo que permite resolver problemas de estimación. Fue formalizada por Bradley Efron en 1979 [140] y su gran versatilidad se debe a que posee una formulación intuitiva que resulta más flexible que los enfoques de estimación clásicos. El método permite estimar una medida de dispersión de un estadístico cuya distribución muestral es desconocida, a partir de una única muestra aleatoria. Para conseguir este objetivo, el método calcula estadísticos muestrales sobre un número grande de muestras *bootstrap* obtenidas a partir del remuestreo con reemplazo del conjunto original de datos [141,142].

El enfoque de *bootstrap* permitió obtener una estimación empírica de la incertidumbre de predicción asociada a cada muestra del conjunto de datos experimentales de fluorescencia. Para implementar esta metodología, en primer lugar, un 25% del total de 35 muestras se reservó como juego de validación fijo. A partir del 75% restante (conjunto de calibrado), se generaron conjuntos *bootstrap* de entrenamiento de tamaño muestral igual a 35, realizando el remuestreo aleatorio con reposición de dicho conjunto. Este procedimiento se repitió de manera iterativa 1000 veces. En cada ciclo *bootstrap*, se entrenó una red MLP de arquitectura fija (previamente optimizada, según lo desarrollado en el capítulo anterior) y, empleando el juego de

validación fijo, se efectuaron las predicciones correspondientes de cada una las muestras de validación. Todo este proceso se repitió para cuatro particiones diferentes del juego original (es decir, cuatro conjuntos de validación fijos diferentes), de manera de garantizar que todas las muestras del conjunto de datos de partida sean predichas por la red. Esto permitió obtener 1000 predicciones para cada muestra del conjunto original de datos, cuyo desvío estándar representa una estimación del parámetro σ_y en la Ec. 4.7. Asimismo, el valor de incertidumbre en la señal de fluorescencia (ruido instrumental o σ_x) se estimó a partir del residuo obtenido de la descomposición PCA del bloque espectral con un número óptimo de PCs. En virtud de la Ec. 4.7, se logró estimar un valor de SEN empírico ($\mathrm{SEN}_{\mathrm{BS}}$) para cada muestra del conjunto de datos experimentales, a partir del cociente entre los parámetros σ_x y σ_y (este último estimado por *bootstrap*).

4.3.3. Software

La manipulación de datos, simulaciones e implementación de pruebas estadísticas y método *bootstrap* se realizó en MATLAB R2017b. Los algoritmos PLS y MLP fueron llevados a cabo en la interfaz libre de MATLAB MVC1 [107].

4.4. Resultados y discusión

4.4.1. Deducción de las ecuaciones para el cálculo de SEN *en MLP-ANN*

En este trabajo se derivó de manera teórica la expresión para el cálculo de la SEN (Ec. 4.8) en un modelo de calibración basado en una red tipo MLP de tres capas y función de transferencia sigmoidea. Es importante mencionar que el enfoque matemático aquí presentado es general y se puede extender, en principio, a otras ANNs con diferentes funciones de transferencia. Por otro lado, tal como se describió anteriormente, estimar la SEN implica conocer la varianza en la concentración predicha como función de la incertidumbre en la señal instrumental. Esta última se puede medir convenientemente mediante la matriz de covarianza del error y, en el caso del ruido *iid*, a través de la varianza en la señal instrumental.

Para la derivación de una fórmula específica para $\mathbf{b}_{\mathrm{MLP}}$ se consideró una red de tres capas como la que se muestra en la Figura 4.2, formada por una capa de entrada, una capa oculta y una capa de salida. La capa de entrada se compone de A neuronas (el número de PCs utilizados para modelar la varianza espectral en la matriz de datos de entrenamiento) más el *bias* de entrada.

Para una dada muestra de entrenamiento i, los datos de entrada de la red son A *scores* de la muestra, contenidos en un vector $\mathbf{t}$ de $A \times 1$, que constituye la proyección del espectro de la muestra $\mathbf{x}$ de tamaño $J \times 1$ en el espacio de los *loadings* $\mathbf{P}$ de PCA, es decir:

$$\mathbf{t} = \mathbf{P}^{\mathrm{T}}\mathbf{x} \tag{4.16}$$

Debido a que el conjunto imagen de la función de transferencia empleado varía de -1 a $+1$, cada elemento $t(a)$ del vector de *scores* $\mathbf{t}$ debe ser escalado a $t_{\mathrm{MLP}}(a)$, de acuerdo a:

$$t_{\mathrm{MLP}}(a) = \frac{2t(i,a) - [\max(\mathbf{t}_a) + \min(\mathbf{t}_a)]}{\max(\mathbf{t}_a) - \min(\mathbf{t}_a)} \tag{4.17}$$

donde $\max(\mathbf{t}_a)$ y $\min(\mathbf{t}_a)$ son el máximo y el mínimo valor de la a-ésima columna de la matriz de *scores* de calibración $\mathbf{T}$.

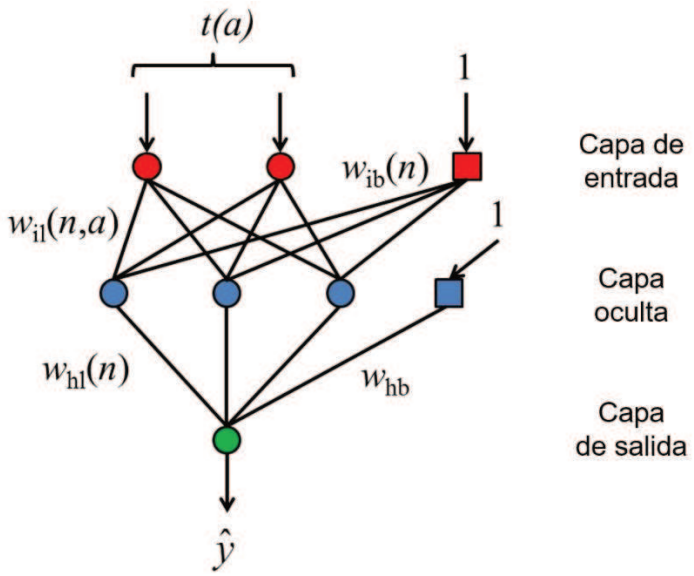


Figura 4.2. Representación esquemática de una MLP típica con dos neuronas en la capa de entrada, en rojo, más el *bias* (en general, A neuronas más el *bias*), tres neuronas en la capa oculta, en azul (en general, N neuronas) y una única neurona en la capa de salida, en verde (en general, O neuronas). Las neuronas se representan con círculos, mientras que los dos *bias* se representan con cuadrados y sus valores por defecto son iguales a 1. Los pesos asociados a las conexiones intercapa son los siguientes: $w_{\mathrm{il}}(n,a)$ representa el peso sináptico entre la a-ésima neurona de entrada y la n-ésima neurona oculta; $w_{\mathrm{ib}}(n)$ es el peso entre el *bias* de entrada y la n-ésima neurona oculta; $w_{\mathrm{hl}}(n)$ es la conexión entre la n-ésima neurona oculta y la neurona de salida, mientras que w_{hb}, entre el *bias* de la capa oculta y la neurona de salida.

Una vez escalado, el a-ésimo elemento del vector $\mathbf{t}_{\mathrm{MLP}}$ constituye el *input* (entrada) de cada una de las A neuronas de la capa de entrada, que son proyectados sobre la función no lineal de transferencia, de manera que la salida de las cada neurona de entrada está dada por:

$$\frac{1}{1+\exp\left[-t_{\mathrm{MLP}}(a)\right]} \tag{4.18}$$

En todas las neuronas de la red, la función de transferencia es de tipo sigmoidea (logística), como lo expresa la Ec. 4.18. Sin embargo, con el objetivo de simplificar la notación y el álgebra en las ecuaciones sucesivas, las expresiones exponenciales se escriben en términos de la función tangente hiperbólica $(\tanh)$, de acuerdo a la siguiente equivalencia con la función logística:

$$\frac{1}{1+\exp\left[-t_{\mathrm{MLP}}(a)\right]} = \frac{1+\tanh\left[t_{\mathrm{MLP}}(a)/2\right]}{2} \tag{4.19}$$

Las salidas de las neuronas en la capa de entrada se multiplican por los pesos asociados a las conexiones entre la a-ésima neurona de entrada y la n-ésima neurona oculta ($w_{\mathrm{il}}(n,a)$), y se agrega el *bias*, obteniendo el valor $g(n)$ dado por:

$$g(n) = w_{\mathrm{ib}}(n) + \sum_{a=1}^{A} w_{\mathrm{il}}(n,a)\frac{1+\tanh\left[t_{\mathrm{MLP}}(a)/2\right]}{2} \tag{4.20}$$

donde $w_{\mathrm{ib}}(n)$ es el peso asociado a la conexión entre el *bias* de entrada y la n-ésima neurona oculta. En cada neurona oculta, la suma ponderada dada por $g(n)$ representa el argumento de la función de activación, que efectúa una nueva proyección no lineal para obtener el valor $h(n)$ como:

$$h(n) = \frac{1+\tanh\left[g(n)/2\right]}{2} \tag{4.21}$$

Entre la capa oculta y la de salida, se efectúa un procedimiento similar. Los valores $h(n)$ son pesados por las conexiones $w_{\mathrm{hl}}(n)$ entre la n-ésima neurona oculta y la neurona de salida, y junto con el *bias* de la capa oculta, se obtiene el valor m dado por:

$$m = w_{\mathrm{hb}} + \sum_{n=1}^{N} w_{\mathrm{hl}}(n)h(n) \tag{4.22}$$

donde w_{hb} es la conexión del *bias* en la capa oculta y la capa de salida. Por último, la salida de la capa oculta que representa la salida de la red y_{MLP}, estará dada por la proyección de m sobre la función de transferencia, es decir:

$$y_{\mathrm{MLP}} = \frac{1+\tanh\left(m/2\right)}{2} \tag{4.23}$$

Para obtener la estimación de la variable respuesta predicha por la red $\hat{y}$, se debe reescalar el valor y_{MLP} para recuperar su rango original, de acuerdo a:

$$\hat{y} = \frac{y_{\mathrm{MLP}} - b_y}{a_y} \tag{4.24}$$

donde a_y y b_y son los factores de escalado.

Tal como puede observarse en las ecuaciones previas, la predicción de la red $\hat{y}$ está dada por una serie de funciones multivariables anidadas. Por lo tanto, para conocer

cómo varía $\hat{y}$ respecto de las entradas de la red, es necesario derivar parcialmente la expresión completa para $\hat{y}$ utilizando la regla de la cadena. En efecto, la derivada de $\hat{y}$ viene dada por:

$$d\hat{y} = \frac{1}{a_y}\left(\frac{\partial y_{\mathrm{MLP}}}{\partial m}\right)\sum_{n=1}^{N}\left(\frac{\partial m}{\partial h(n)}\right)\left(\frac{\partial h(n)}{\partial g(n)}\right)\sum_{a=1}^{A}\left(\frac{\partial g(n)}{\partial t(a)}\right)\left(\frac{dt(a)}{\Delta(a)}\right) \tag{4.25}$$

que puede ser fácilmente calculada a partir de las Ecs. 4.18-24 como:

$$d\hat{y} = \frac{1}{4a_y\cosh^2(m/2)}\sum_{j=1}^{J}\frac{w_{\mathrm{hl}}(j)}{4\cosh^2[g(j)/2]}\sum_{a=1}^{A}\frac{w_{\mathrm{il}}(j,a)dt(a)}{2\Delta(a)\cosh^2[t_{\mathrm{MLP}}(a)/2]} \tag{4.26}$$

En las Ecs. 4.25-26, el parámetro $\Delta(a)$ está dado por $\max(\mathbf{t}_a) - \min(\mathbf{t}_a)$.

De la Ec. 4.26 se puede establecer una relación entre el diferencial del a-ésimo *score* y el diferencial del espectro de la muestra $\mathbf{x}$ como:

$$dt(a) = \mathbf{p}(a)^T d\mathbf{x} = \sum_{j=1}^{J} p(j,a)dx(j) \tag{4.27}$$

De esta manera, es posible ahora definir un vector $\mathbf{b}_{\mathrm{MLP}}$ de tamaño $J \times 1$, cuyo j-ésimo elemento estará dado por:

$$b_{\mathrm{MLP}}(j) = \frac{1}{4a_y\cosh^2(m/2)}\sum_{n=1}^{N}\frac{w_{\mathrm{hl}}(n)}{4\cosh^2[g(n)/2]}\sum_{a=1}^{A}\frac{w_{\mathrm{il}}(n,a)p(j,a)}{2\Delta(a)\cosh^2[t_{\mathrm{MLP}}(a)/2]} \tag{4.28}$$

Así, el $d\hat{y}$ puede reescribirse como el producto escalar

$$d\hat{y} = \mathbf{b}_{\mathrm{MLP}}^{\mathrm{T}} d\mathbf{x} \tag{4.29}$$

Debe notarse que, a diferencia de lo que ocurre en PLS, el vector $\mathbf{b}_{\mathrm{MLP}}$ es dependiente de la muestra, debido la presencia del factor $\mathbf{t}_{\mathrm{MLP}}$ y funciones dependientes de él en la Ec. 4.28, dado que $\mathbf{t}_{\mathrm{MLP}}$ representa el vector de *scores* escalados de una muestra específica.

Finalmente, la varianza en concentración viene dada por la esperanza de $d\hat{y}^2$, la cual se deriva de:

$$E(d\hat{y}^2) = \sigma_{\hat{y}}^2 = E\left(\mathbf{b}_{\mathrm{MLP}}^{\mathrm{T}} d\mathbf{x} d\mathbf{x}^{\mathrm{T}} \mathbf{b}_{\mathrm{MLP}}\right) = \mathbf{b}_{\mathrm{MLP}}^{\mathrm{T}} \sum_{\mathbf{x}} \mathbf{b}_{\mathrm{MLP}} \tag{4.30}$$

donde $\sum_{\mathrm{x}}$ es la matriz de covarianza del error para los datos espectrales, es decir, la esperanza de la matriz $d\mathbf{x}d\mathbf{x}^{\mathrm{T}}$. En caso de estructura de ruido tipo *iid*, la matriz $\sum_x$ es diagonal y sus elementos no nulos representan la varianza espectral σ_x^2, por lo que:

$$\sigma_{\hat{y}}^2 = \sigma_x^2 \|\mathbf{b}_{\mathrm{MLP}}\|^2 \tag{4.31}$$

Reescribiendo esta última ecuación, es posible entonces obtener el valor de SEN como el cociente entre las incertidumbres en x e $\hat{y}$, tal como expresa la Ec. 4.7, es decir:

$$\mathrm{SEN} = \sqrt{\frac{\sigma_x^2}{\sigma_y^2}} = \frac{\sigma_x}{\sigma_y} = \frac{1}{\|\mathbf{b}_{\mathrm{MLP}}\|} \tag{4.32}$$

Finalmente, debe notarse que si en la neurona de salida no se incluye una función de transferencia no lineal, sólo se debe remover el factor $\frac{1}{4\cosh^2(m/2)}$ de la Ec. 4.28.

4.4.2. Validación mediante simulación Montecarlo

Se generaron cuatro sistemas no lineales de datos simulados. Con cada conjunto de calibración se entrenó una red MLP que, en todos los casos, presentó una arquitectura óptima de $4-2-1$. Con las redes entrenadas, se procedió a efectuar predicciones en los diferentes conjuntos de datos de prueba.

Para cada ciclo de predicción, se computaron los valores de SEN de la Ec. 4.8, calculados en forma teórica de acuerdo a lo establecido en la Sección 4.4.1. En cada ronda de 500 predicciones a una dada concentración de analito y un nivel determinado de ruido, se computó el valor de $\mathbf{b}_{\mathrm{MLP}}$ promedio para efectuar el cálculo de SEN. Paralelamente, se calculó el desvío estándar σ_y asociado a las 500 predicciones para cada conjunto de muestras test. Para cada nivel de concentración de analito en los conjuntos de test (11 niveles en total) se obtuvo un valor de σ_y para cada nivel de ruido σ_x (4 niveles). Esto permitió obtener las dispersiones σ_x vs σ_y, para luego proceder a estimar la $\mathrm{SEN}_{\mathrm{MC}}$ a partir de la pendiente de la recta de regresión calculada en cada caso. Al igual que para los cálculos teóricos, se obtuvo de esta forma un valor de $\mathrm{SEN}_{\mathrm{MC}}$ para cada nivel de concentración de analito. A modo de ejemplo, en la Figura 4.3 se muestran los gráficos de dispersión σ_x vs σ_y y las rectas de regresión ajustadas, con los respectivos valores de $\mathrm{SEN}_{\mathrm{MC}}$, obtenidas para el sistema cuadrático. Para todos los niveles de analito, puede verse que la relación entre σ_x y σ_y es perfectamente lineal. Los valores calculados de $\mathrm{SEN}_{\mathrm{MC}}$ carecen de unidades porque para los datos simulados se trabajó en una escala arbitraria de concentraciones.

Los valores de SEN calculados en forma teórica y mediante simulación Montecarlo se compararon de manera gráfica y estadística. En primer lugar, en la Figura 4.4 se presentan los resultados de la comparación gráfica para los cuatro sistemas simulados. En este sentido, en las Figura 4.4.A-D se muestra la variación de SEN calculada en forma teórica (cruces rojas) y mediante Montecarlo (círculos azules) como función de la concentración de analito, para los sistemas cuadrático, irracional, sigmoideo y logarítmico, respectivamente. Se puede observar que los valores de sensibilidad estimados mediante los dos métodos son muy similares entre sí para todos los sistemas. Asimismo, se puede evidenciar que cuando los valores de SEN se grafican en función

de la concentración de analito, en todos los casos, la sensibilidad se comporta cualitativamente igual que la primera derivada de la función no lineal que gobierna la relación entre señal y concentración. En este caso, dado que los datos son simulados, dichas funciones se conocen de manera exacta y están dadas por las Ecs. 4.10 a 4.13. De esta manera, en la Figura 4.4 se representan en líneas negras discontinuas, las derivadas normalizadas de $x(j)$ con respecto a y_n para cada una de las funciones no lineales empleadas en la simulación de los datos. Este resultado permite además ratificar la estrategia propuesta para validar las ecuaciones desarrolladas para el cálculo teórico de SEN, ya que existe una correspondencia directa entre la Ecs. 4.8 y el concepto general de SEN dado por la Ec. 4.7.

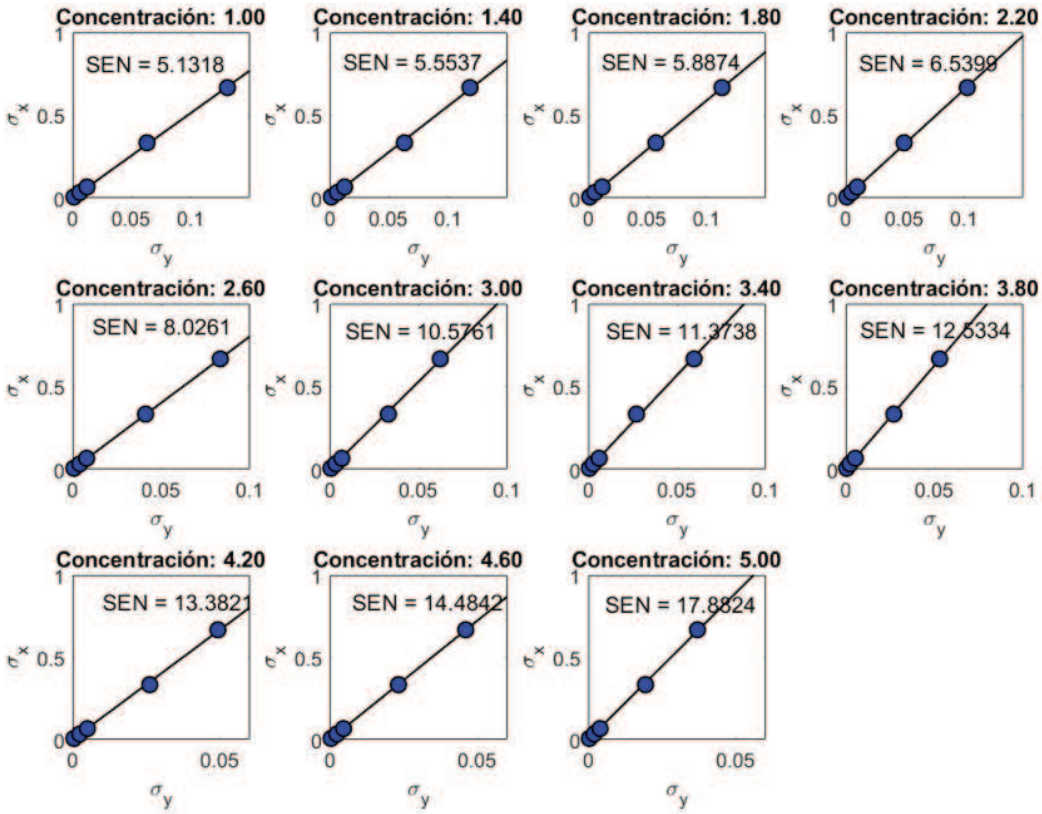


Figura 4.3. Cálculo de $\mathrm{SEN}_{\mathrm{MC}}$ mediante simulación Montecarlo para los 11 conjuntos de muestras test generados para el sistema no lineal cuadrático (sistema A). En cada panel se muestra la dispersión σ_x vs σ_y, con su respectiva recta de regresión ajustada y el valor de SEN (obtenido de la pendiente de la recta de regresión).

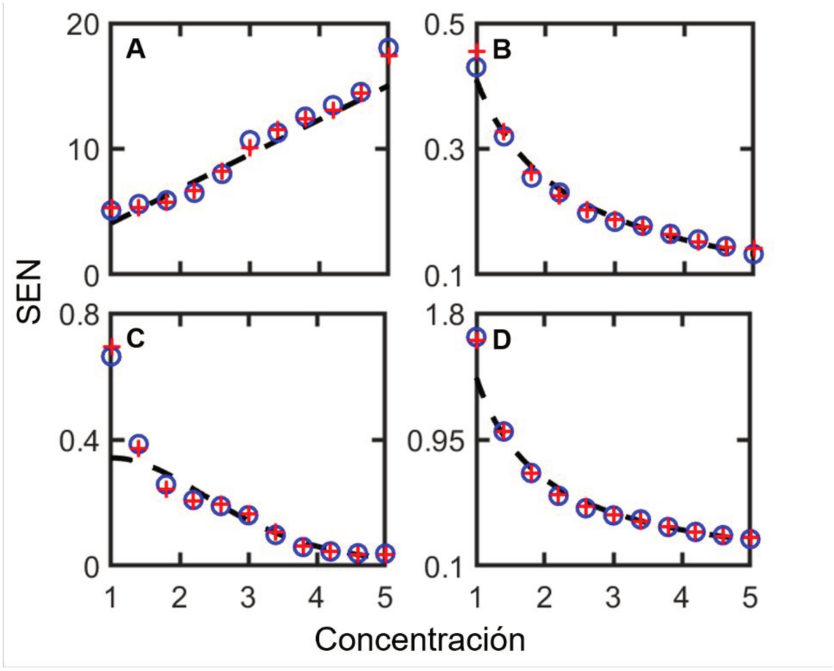


Figura 4.4. Cálculo de SEN teórica de la Ec. 5.8 (cruces rojas) y $\mathrm{SEN}_{\mathrm{MC}}$ (círculos azules) como función de la concentración de analito en las muestras test de los sistemas simulados **A.** cuadrático, **B.** irracional, **C.** sigmoideo y **D.** logarítmico. En A-D la curva negra discontinua representa la primera derivada normalizada de las Ecs. 4.10-13, respectivamente (derivada de $x(j)$ con respecto a y_n).

Por otra parte, la comparación entre los valores de SEN estimados mediante las dos metodologías también se llevó a cabo de manera estadística, utilizando la prueba EJCR, cuya salida gráfica se muestra en la Figura 4.5. Para realizar el test estadístico, se compendiaron los pares de resultados de SEN obtenidos para los cuatro sistemas, de manera que se efectuó una prueba de comparación única. De esta forma, puede verse en la Figura 4.5 (figura principal) que los valores de SEN teóricos y Montecarlo se alinean de manera muy satisfactoria a una recta de pendiente 1 y ordenada 0 (recta ideal). En este sentido, la prueba estadística EJCR consiste justamente en testear la hipótesis nula de que los parámetros estimados en la recta de regresión que surge de la dispersión SEN teórica vs SEN Montecarlo incluyen en sus intervalos de confianza a los valores 1 y 0 para la pendiente y la ordenada, respectivamente. Este resultado se muestra gráficamente en la figura insertada de la Figura 4.5. Debido a que la región elíptica de confianza que surge de la prueba estadística bivariada incluye al punto $(1, 0)$ entonces es posible, con un nivel de confianza del 95%, aceptar la hipótesis nula.

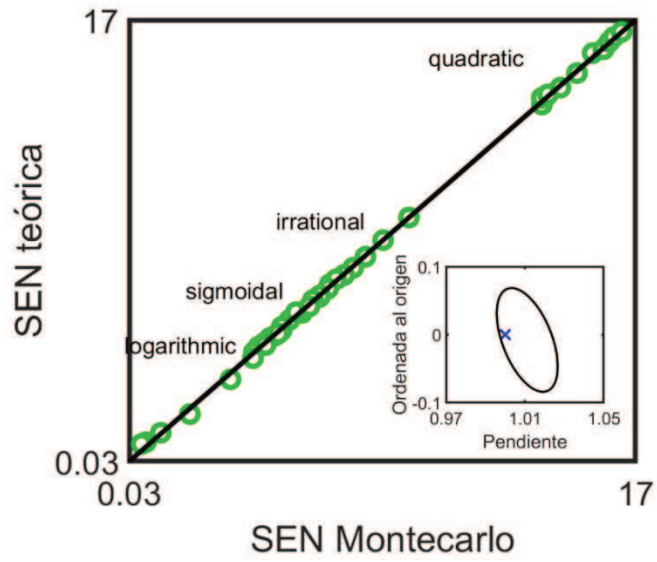


Figura 4.5. Comparación estadística de los valores de SEN teóricos y calculados mediante Montecarlo para los resultados obtenidos con los cuatro sistemas no lineales simulados. Se realizó una única regresión (figura principal) y prueba EJCR (figura insertada) con todos los datos de SEN. La recta negra representa la curva identidad ideal SEN teórica vs SEN Montecarlo.

4.4.3. Validación mediante técnica de bootstrap y caracterización del método de calibración desarrollado para la cuantificación at-line de etanercept. Comparación con el método de referencia

El segundo enfoque propuesto para validar la fidelidad de las ecuaciones teóricas desarrolladas para la estimación de la sensibilidad en modelos de calibración MLP se llevó a cabo con los datos de fluorescencia empleados en el Capítulo 3, mediante la técnica de *bootstrap.* A partir del conjunto original, se generaron 4 particiones diferentes de entrenamiento y validación, de manera de garantizar que todas las muestras del conjunto sean utilizadas una vez como muestras de predicción. Para cada una de ellas, los ciclos de *bootstrap* generaron 1000 predicciones de concentración de etanercept, que permitió calcular los valores de σ_y para cada muestra. Asimismo, con el valor de σ_x estimado a partir del residuo de un modelo PCA óptimo para la compresión del bloque espectral completo, se calcularon los valores de $\mathrm{SEN}_{\mathrm{BS}}$ de acuerdo a la Ec. 4.7. En la Figura 4.6 se muestran, a modo de ejemplo, los histogramas correspondientes a 1000 predicciones para las 9 muestras de una de las particiones del conjunto de datos.

Por otro lado, se procedió a comparar los valores de $\mathrm{SEN}_{\mathrm{BS}}$ con los de SEN estimados de manera teórica. Para ello, se empleó una estrategia tipo *leave-one-out* en ciclos de entrenamiento-predicción de una red MLP a arquitectura fija de manera de obtener una estimación de SEN teórica asociada a cada muestra del conjunto de datos.

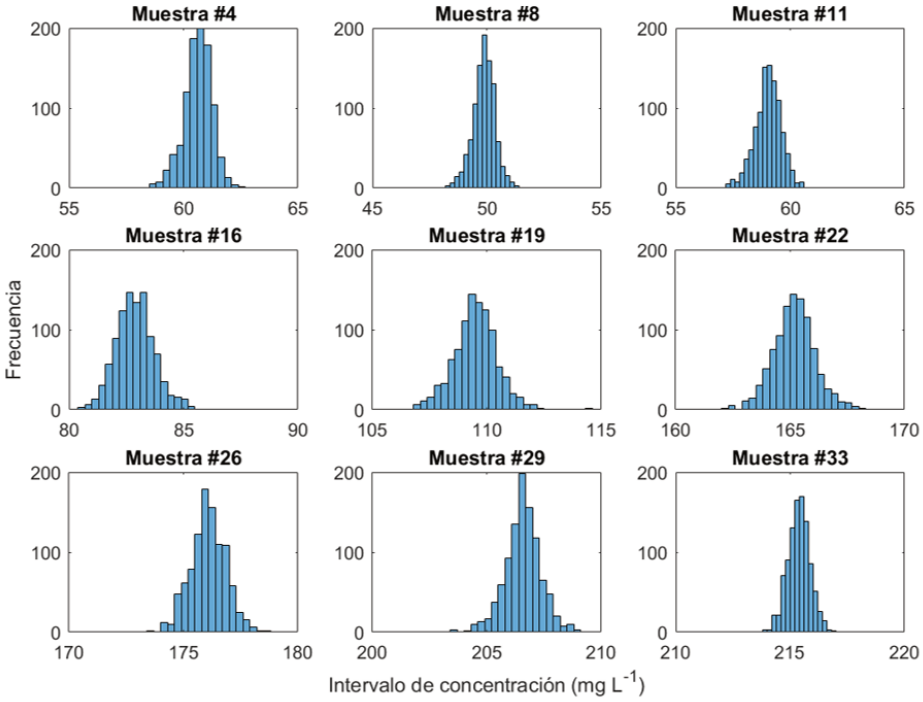


Figura 4.6. Histogramas para 1000 predicciones *bootstrap* de nueve muestras del conjunto experimental de fluorescencia para la estimación de σ_y y posterior cálculo de $\mathrm{SEN}_{\mathrm{BS}}$.

Debido a que, en el caso de la estrategia de validación basada en datos experimentales, la metodología implicó una cantidad muy grande de cálculos complejos que se ven fuertemente influenciados por los errores intrínsecos de los datos reales, era dable esperar que el grado de similitud entre los valores estimados mediante una y otra técnica sea menor que en el caso de los datos simulados. Por esta razón, la comparación se realizó únicamente de manera gráfica/cualitativa. Los resultados se muestran en la Figura 4.7.A, en la cual se representan los valores de SEN teóricos (círculos azules) y $\mathrm{SEN}_{\mathrm{BS}}$ (círculos verdes) en función de la concentración nominal de etanercept de cada muestra del conjunto de datos. Se observa un grado de consistencia razonable entre las sensibilidades calculadas por ambas técnicas, lo cual también permite validar las ecuaciones teóricas desarrolladas.

Asimismo, tal como ocurrió en la estrategia Montecarlo, se comprobó que la variación cualitativa de la sensibilidad en función de la concentración se asemeja a la primera derivada de la función no lineal que relaciona la señal espectral con la concentración. En este caso, la expresión analítica de dicha función es desconocida. No obstante, para poder poner de manifiesto este resultado, se procedió a aproximar la función no lineal involucrada a partir de los resultados de CV obtenidas mediante calibración PLS (Capítulo 3). En este sentido, en la Figura 4.7.B se muestran los valores de concentración predicha en la CV de PLS en función de la concentración nominal de

etanercept (misma gráfica que la mostrada en la Figura 3.6.A del capítulo anterior). A partir las predicciones de CV se estimó la función no lineal de manera gráfica, realizando un suavizado e interpolación de los datos (Figura 4.7.B, línea negra discontinua). Con la curva obtenida se calculó su primera derivada de manera numérica, mediante el método de Savitzky-Golay. El resultado obtenido se normalizó de manera conveniente para poder representar la función derivada junto con las estimaciones de sensibilidad en la Figura 4.7.A (línea negra discontinua). Nuevamente se observa una clara consistencia entre la estimación teórica de sensibilidad y su definición general representada por la Ec. 4.7.

Finalmente, con el objetivo de caracterizar al modelo MLP desarrollado en el Capítulo 3, que permitiría la implementación de un método de monitoreo *at-line* de etanercept en el contexto de la PAT, se procedió a calcular la sensibilidad promedio del conjunto de calibración, cuyo valor fue igual a 4,2 mg^{-1} L. Para poder establecer una comparación con el método de referencia (HPLC univariado), se calcularon los valores correspondientes de γ de acuerdo a la Ec. 4.9. Los valores obtenidos fueron 0,05 y 1,70 mg^{-1} L para el método de referencia y el basado en MLP, respectivamente. Este último resultado constituye una prueba de cómo la calibración multivariada favorece a incrementar la sensibilidad de una técnica analítica, lo cual ha sido ampliamente demostrado en la bibliografía [143]. Naturalmente, este incremento de la sensibilidad, a su vez, tiene un impacto directo sobre otras AFOMs de interés, tales como LOD y LOQ.

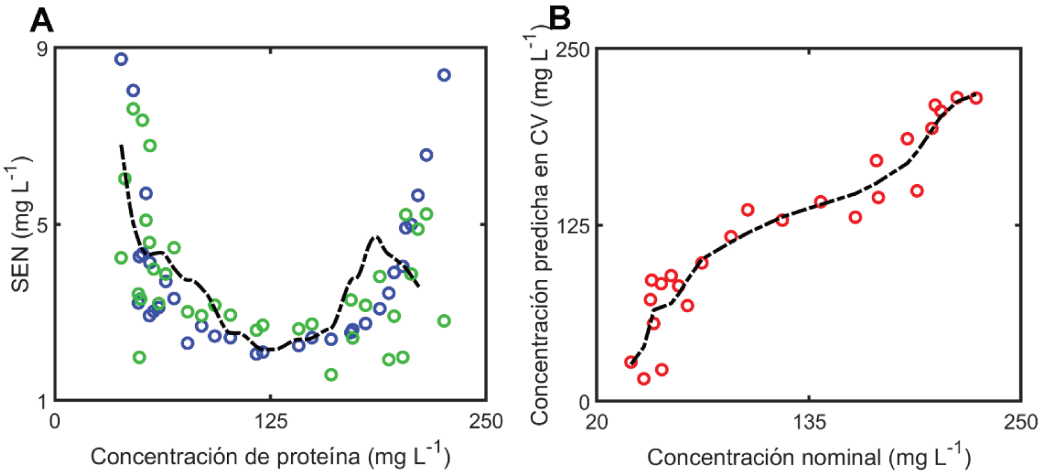


Figura 4.7. A. SEN teórica y SEN_{BS} en función de la concentración de proteína recombinante (etanercept). La línea negra discontinua representa la primera derivada normalizada de la función no lineal que relaciona la señal de fluorescencia con la concentración del analito (calculada en forma numérica a partir de una aproximación de la función no lineal con los resultados de predicciones de CV en PLS). **B.** Valores predichos vs nominales de concentración de etanercept mediante CV en el modelo PLS reportado en el Capítulo 3. La línea negra discontinua representa una aproximación de la función no lineal que relaciona la señal de fluorescencia con la concentración del analito (calculada mediante suavizado e interpolación de los resultados de CV).

4.4.4. MVC1-GUI: actualización de una interfaz gráfica de usuario para calibración de primer orden que incluye modelos no lineales y cálculo de cifras de mérito

MVC1 es una interfaz gráfica desarrollada en MATLAB y que fue publicada por primera vez en el año 2004 [107]. El programa permite la implementación de varios de los algoritmos más frecuentemente empleados para el desarrollo de modelos de calibración de primer orden, junto con algunas otras utilidades para el manejo y visualización de los datos. Durante el desarrollo de esta tesis, las nuevas contribuciones efectuadas en lo que respecta a la implementación de algoritmos de calibración basados en redes neuronales y cálculo de AFOMs, en particular de tipo MLP, fueron compiladas para dar lugar a una versión actualizada de la GUI. Asimismo, se mejoraron diferentes aspectos prácticos y estéticos del programa, generando una herramienta más versátil e intuitiva.

El diseño de la ventana principal de la nueva versión de MVC1 se muestra en la Figura 4.8. El paquete para ejecutar la GUI dentro del entorno de MATLAB, junto con el manual de usuario y datos ejemplos se pueden descargar libremente a través del enlace https://fbcb.web1.unl.edu.ar/laboratorios/ladaq/download/.

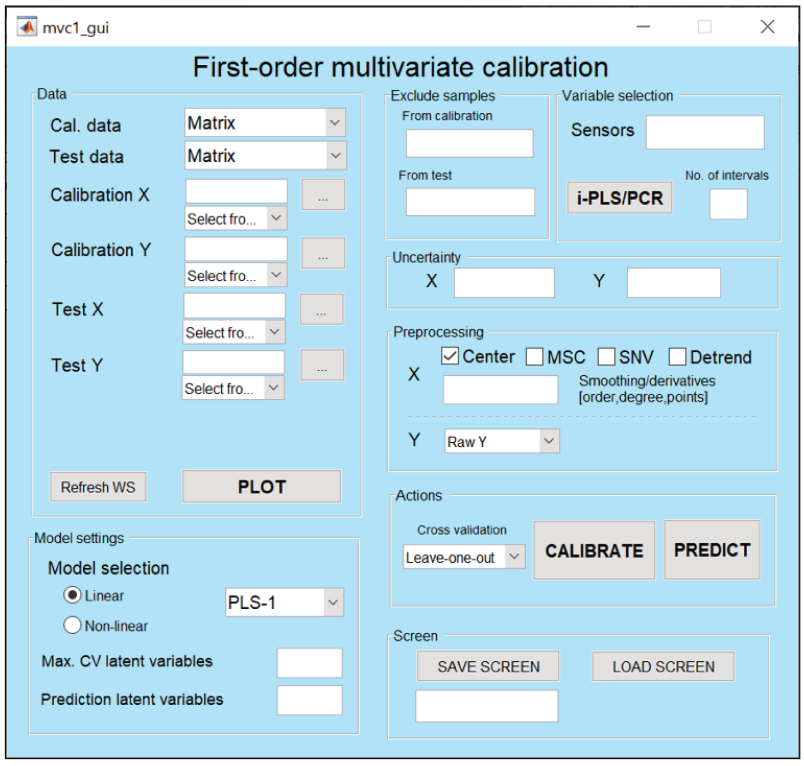


Figura 4.8. Ventana principal de la GUI MVC1 (versión actualizada) para la implementación amigable de algoritmos de calibración lineales (PCR, PLS, entre otros) y no lineales (MLP y RBF). Los paneles principales incluyen carga y visualización de datos (*Data*), selección y parámetros del modelo (*Model settings*), herramientas para manejo de muestras (*Exclude samples*) y selección de variables (*Variable selection*), preprocesamiento (*Preprocessing*), entrenamiento/calibración de modelos y predicción de muestras test (*Actions*).

4.4.5. Perspectivas futuras

Tal como se describió en la introducción de este capítulo, la sensibilidad constituye una de las AFOMs de mayor relevancia en calibración, ya que de ella se derivan otras cifras de gran importancia. En este contexto, por ejemplo, es bien conocido que, bajo el supuesto de ruido *iid*, la incertidumbre asociada a la predicción $\hat{y}$ de una dada muestra incógnita en cualquier modelo de calibración, puede conocerse a partir del cálculo de la varianza $\sigma_{\hat{y}}^2$, de acuerdo a la siguiente expresión de tres términos [135]:

$$\sigma_{\hat{y}}^2 = \sigma_x^2 \mathrm{SEN}^{-2} + h\sigma_x^2 \mathrm{SEN}^{-2} + h\sigma_{\mathrm{ycal}}^2 \tag{4.33}$$

donde σ_{ycal}^2 es la varianza asociada a las concentraciones de las muestras de calibración y h se conoce como leva de la muestra. La Ec. 4.33 da cuenta de las fuentes de error que influyen en la incertidumbre de predicción de un modelo. En este sentido, el primer término se relaciona con la incertidumbre de la señal instrumental de la muestra problema, el segundo, con la incertidumbre asociada a las señales del calibrado y el tercero, con la incertidumbre asociada a las concentraciones del conjunto de calibración [28]. Se puede ver que los términos que se relacionan con la incertidumbre proveniente del calibrado se encuentran escalados por el parámetro h. La estimación de los parámetros SEN y h también son necesarios para el cálculo de otras AFOMs de gran interés analítico y tecnológico, tales como el LOD y el LOQ.

Conceptualmente, la leva de la muestra se define como un parámetro adimensional que mide la posición relativa de una muestra al centro del espacio de calibración. Esta se debe poder estimar tanto para muestras de calibración como para muestras incógnita y puede ser expresada en términos de concentración, en términos de variables espectrales o en términos de variables latentes. Si bien el cálculo de la leva es relativamente simple en calibración univariada, su estimación no es trivial en escenarios de calibración multivariada. En la literatura se han hecho aportes muy importantes en lo que respecta al cálculo de este parámetro en modelos de calibración de primer orden y superior, tanto para métodos lineales como no lineales [144]. Sin embargo, no existe una expresión general que pueda ser directamente extrapolada para el caso del MLP y la manera de estimarla aún no ha sido descripta. Dada la enorme flexibilidad que caracteriza a este modelo no paramétrico, el cálculo del parámetro h no resulta una tarea sencilla. Por estas razones, el estudio de cifras tales como $\sigma_{\hat{y}}$, LOD y LOQ no ha sido abordado en el desarrollo de esta investigación, ya que todas dependen, en primera medida, de poder estimar los valores de leva en el MLP.

4.5. Conclusiones del capítulo

Se desarrolló un método sistemático para la estimación en forma teórica de la sensibilidad y la sensibilidad analítica de un modelo de calibración de primer orden basado en redes de tipo MLP.

Los resultados muestran que las ecuaciones desarrolladas pueden ser validadas mediante los enfoques de simulación Montecarlo y *bootstrap* con datos reales.

En base a las contribuciones realizadas en este capítulo, fue posible por primera vez calcular la sensibilidad en un modelo de calibración de primer orden basado en redes de tipo MLP. Esto permitió calcular y caracterizar la sensibilidad del método de calibración PAT desarrollado en el Capítulo 3.

Desde el punto de vista de la investigación básica, este trabajo ha contribuido a la caracterización estadística de modelos de calibración basados en un tipo específico de red neuronal artificial. En este sentido, los resultados del análisis teórico y las ecuaciones desarrolladas constituyen una estrategia robusta y simple para la estimación de la sensibilidad en el contexto de la calibración analítica, sin costos computacionales excesivos.

La estimación de la sensibilidad resulta de fundamental importancia para el cálculo de otras AFOMs que contribuyan a la validación de métodos analíticos. A su vez, esto presenta un impacto tecnológico relevante, ya que la caracterización completa de las AFOMs de un método de calibración basado en MLP brindaría a las industrias la posibilidad de describir y validar nuevas metodologías de análisis al mismo nivel que aquellos basados en modelos multivariados lineales. Como consecuencia, esta situación permitiría a su vez, incrementar la flexibilidad de cambios posteriores a la aprobación ante autoridades regulatorias respecto de la validación de nuevas herramientas de monitoreo de procesos, facilitando de esta manera la implementación de iniciativas acordes con los principios de la PAT.

Conclusiones

Este trabajo de tesis ha puesto de manifiesto uno de los atributos que caracterizan al quehacer científico contemporáneo. Es evidente que las problemáticas científico-tecnológicas actuales revisten un grado de complejidad tal que implican necesariamente de abordajes interdisciplinarios de investigación y desarrollo. En este sentido, en esta tesis han confluido eficaz y provechosamente diversas disciplinas relacionadas con la química, la biología, la matemática, la estadística y las ciencias computacionales. En suma, todas estas áreas atraviesan a la quimiometría y sus aplicaciones.

En línea con el título y los objetivos de la tesis, es posible afirmar que en este trabajo de investigación se realizaron aportes relevantes tanto para la quimiometría básica como aplicada.

En relación a la quimiometría básica, las contribuciones sustanciales se pueden sintetizar de la siguiente manera:

- el estudio exhaustivo de metodologías para el preprocesamiento de matrices de excitación-emisión de fluorescencia permite evidenciar que no existe un método general que resulte óptimo para efectuar la corrección de las señales de dispersión de manera computacional, cualquiera sea el escenario. Por el contrario, la estrategia ideal sería la combinación conveniente de todas ellas, guiada por el buen criterio del analista y los objetivos posteriores de análisis de los datos. En este sentido, el estudio realizado permite vislumbrar la utilidad que reviste poder contar con diversas herramientas que muestren diferentes ventajas, de manera tal de combinarlas para explotar al máximo las potencialidades de cada una de ellas (Capítulo 1);
- el estudio sistemático y comparativo motivó el desarrollo de una metodología novedosa para la corrección de *scattering* basada en el principio de la conservación de la señal (Capítulo 1);
- el uso de herramientas de diseño y optimización experimental (RSM) demostró ser una estrategia interesante en relación a la optimización de modelos de calibración basados en redes neuronales artificiales del tipo perceptrón multicapa. En particular, la RSM permitió, minimizando el número de ensayos computacionales, analizar de manera global la relación entre la capacidad predictiva del algoritmo de red neuronal y los diferentes parámetros topológicos que lo definen. En este sentido, esta estrategia quimiométrica demostró ser útil como herramienta para la construcción de un modelo de calibración no paramétrico en donde *a priori* su arquitectura óptima resulta desconocida (Capítulo 3);
- se desarrollaron y validaron ecuaciones para la estimación de la sensibilidad y la sensibilidad analítica en modelos de calibración basados en el

perceptrón multicapa. Este hecho representa un primer e importante paso hacia la caracterización completa en términos estadísticos y analíticos de los modelos de calibración basados en este tipo de red neuronal artificial (Capítulo 4).

- se desarrollaron y actualizaron algoritmos y programas computacionales (GUI) de libre acceso, con el objetivo de brindar a la comunidad científica herramientas que integran técnicas de preprocesamiento (EEM_corr) y de calibración de primer orden (MVC1), que faciliten la implementación de metodologías quimiométricas de una manera más intuitiva y amigable (Capítulo 1 y Capítulo 4).

Por otra parte, en relación a las aplicaciones de la quimiometría para el estudio y monitoreo del bioproceso de etanercept en el contexto de la PAT, las principales contribuciones fueron las siguientes:

- la utilización de manera integrada de diversas metodologías quimiométricas de primer y segundo orden demostró ser una estrategia sumamente útil para modelar y explorar desde un enfoque multivariado la gran cantidad de información contenida tanto en los datos de variables de proceso como en los datos de fluorescencia. En este sentido, el uso de algoritmos con fines exploratorios permitió, no sólo efectuar una caracterización preliminar de la composición del medio de cultivo, sino también establecer nuevos criterios respecto al monitoreo de la viabilidad celular. En relación a este último aspecto, se probó cómo el modelado adecuado de los datos multidimensionales de fluorescencia permite extraer huellas espectrales que reflejan información clave relacionada con la evolución de la viabilidad de los cultivos. Asimismo, se obtuvo un modelo cualitativo predictivo que permitiría la implementación de esta metodología como una estrategia de PAT para el monitoreo prospectivo de la viabilidad celular en futuros lotes del mismo bioproceso (Capítulo 3);
- el uso de redes neuronales artificiales con fines cuantitativos junto con herramientas de RSM, permitió obtener una metodología confiable y robusta para la cuantificación de etanercept directamente a partir de muestras de medio de cultivo, sin necesidad de efectuar ningún pretratamiento ni recurrir a técnicas separativas. En este sentido, el método analítico desarrollado se encuadra perfectamente en los principios de la PAT, ya que se basa en una técnica instrumental fácilmente automatizable y de relativo bajo costo. Además, el método permite cuantificar la proteína recombinante de manera *at-line*, brindando una herramienta de monitoreo mucho más rápida, sencilla

y eficiente en términos de proceso que el método univariado actualmente empleado. Vale destacar también que, como una consecuencia directa de la filosofía de la PAT y de la calibración multivariada, se propone un método limpio desde el punto de vista ambiental (Capítulo 3);

- la capacidad de poder calcular de manera confiable la sensibilidad y la sensibilidad analítica de métodos de calibración basados en un tipo específico de red neuronal artificial que hasta el momento no ha sido completamente caracterizado, constituye un paso crucial en el camino hacia la validación y capacidad de implementación en la industria de nuevas metodologías analíticas en el contexto de la PAT (Capítulo 4);
- es importante destacar que tanto el método PAT cualitativo (Capítulo 3) como el cuantitativo (Capítulo 4) se pueden aplicar a partir de la generación de un único tipo de dato, recolectado de un mismo conjunto de muestras. Esto implica, desde el punto de vista práctico, un ahorro considerable de tiempo en relación a la generación de la señal analítica. Asimismo, ambas metodologías serían fácilmente adaptables a su implementación de manera *on-* o *in-line* ya que la espectroscopía de fluorescencia (al igual que muchas otras técnicas espectrales) brinda la posibilidad de generar datos directamente desde el seno del bioproceso, mediante la incorporación al reactor de sensores de fibra óptica autoclavables.

Finalmente, a modo de conclusión global y de reflexión personal, este trabajo de tesis pone de manifiesto que ciencia básica y aplicada claramente conforman y deben seguir siendo un matrimonio indisoluble, cuya retroalimentación constante permitirá garantizar el sólido avance científico-tecnológico. En este sentido, se ha vislumbrado cómo el planteo de una problemática tecnológica real en un contexto alejado de la quimiometría, ha disparado numerosas contribuciones que enriquecen no sólo a la disciplina en sí misma, sino que, naturalmente, su retroalimentación hacia los problemas tecnológicos generará soluciones prácticas que, seguramente, podrán ser extensivas a diversos contextos aplicados.

Bibliografía

[1] SM Mercier, B Diepenbroek, RH Wijffels, M Streefland (2014) Multivariate PAT solutions for biopharmaceutical cultivation: current progress and limitations. *Trends Biotechnol* 32 (6):329-336.

[2] CF Mandenius, K Graumann, TW Schultz, A Premstaller, IM Olsson, E Petiot, C Clemens, M Welin (2009) Quality-by-design for biotechnology-related pharmaceuticals. *Biotechnol J* 4 (5):600-609.

[3] AS Rathore, H Winkle (2009) Quality by design for biopharmaceuticals. *Nat Biotechnol* 27 (1):26-34.

[4] FDA (2004) *Guidance for Industry PAT – A Framework for Innovative Pharmaceutical Manufacturing and Quality Assurance.* http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070305.pdf.

[5] SM Mercier, B Diepenbroek, MCF Dalm, RH Wijffels, M Streefland (2013) Multivariate data analysis as a PAT tool for early bioprocess development data. *J Biotechnol* 167 (3):262-270.

[6] J Glassey, KV Gernaey, C Clemens, TW Schulz, R Oliveira, G Striedner, C-F Mandenius (2011) Process analytical technology (PAT) for biopharmaceuticals. *Biotechnol J* 6 (4):369-377.

[7] W Sommeregger, B Sissolak, K Kandra, M von Stosch, M Mayer, G Striedner (2017) Quality by control: Towards model predictive control of mammalian cell culture bioprocesses. *Biotechnol J* 12 (7):1600546.

[8] AS Rathore, R Bhambure, V Ghare (2010) Process analytical technology (PAT) for biopharmaceutical products. *Anal Bioanal Chem* 398 (1):137-154.

[9] J Zhu (2012) Mammalian cell protein expression for biopharmaceutical production. *Biotechnol Adv* 30 (5):1158-1170.

[10] B Nitish, A Rathore (2011) Use of multivariate data analysis (MVDA) for generating process understanding from manufacturing data of biotech processes. *Proc Indian Natl Sci Acad* 77:133-142.

[11] ND Lourenço, JA Lopes, CF Almeida, MC Sarraguça, HM Pinheiro (2012) Bioreactor monitoring with spectroscopy and chemometrics: a review. *Anal Bioanal Chem* 404 (4):1211-1237.

[12] AS Rathore, N Bhushan, S Hadpe (2011) Chemometrics applications in biotech processes: a review. *Biotechnol Prog* 27 (2):307-315.

[13] S Challa, R Potumarthi (2012) Chemometrics-Based Process Analytical Technology (PAT) Tools: Applications and Adaptation in Pharmaceutical and Biopharmaceutical Industries. *Appl Biochem Biotechnol* 169:66–76.

[14] J Menezes, A Ferreira, L Rodrigues, L Brás, T Alves (2009), 4.10 - Chemometrics Role within the PAT Context: Examples from Primary Pharmaceutical Manufacturing, En: *Comprehensive chemometrics*, (Ed: S Brown, R Tauler, B Walczak), Elsevier, p. 313-355.

[15] S Wold (1972) Splin-funktioner-ett nytt verktyg i data analysen. *Kem Tidskr* 84:34-37.

[16] SD Brown (1988) Chemometrics: A textbook. D. L. Massart. B. G. M. Vandeginste, S. N. Deming, Y. Michotte, and L. Kaufman, Elsevier, Amsterdam, 1988. ISBN 0-444-42660-4. Price Dfl 175.00. *J Chemom* 2 (4):298-299.

[17] L Vera Candioti, MM De Zan, MS Cámara, HC Goicoechea (2014) Experimental design and multiple response optimization. Using the desirability function in analytical methods development. *Talanta* 124:123-138.

[18] P Oliveri, C Malegori, E Mustorgi, M Casale (2021) Qualitative pattern recognition in chemistry: Theoretical background and practical guidelines. *Michrochem J* 162:105725.

[19] DL Massart, BGM Vandeginste, LMC Buydens, S De Jong, PJ Lewi, J Smeyers-Verbeke (1998), Chapter 8 Straight line regression and calibration, En: *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier, p. 171-230.

[20] AC Olivieri, GM Escandar (2014), Chapter 1 - Calibration Scenarios, En: *Practical Three-Way Calibration*, (Ed: AC Olivieri, GM Escandar), Elsevier, Boston, p. 1-9.

[21] GM Escandar, HC Goicoechea, A Muñoz de la Peña, AC Olivieri (2014) Second- and higher-order data generation and calibration: A tutorial. *Anal Chim Acta* 806:8-26.

[22] M Montemurro, GG Siano, MR Alcaráz, HC Goicoechea (2017) Third order chromatographic-excitation–emission fluorescence data: Advances, challenges and prospects in analytical applications. *TrAC, Trends Anal Chem* 93:119-133.

[23] GM Escandar, AC Olivieri, NM Faber, HC Goicoechea, A Muñoz de la Peña, RJ Poppi (2007) Second- and third-order multivariate calibration: data, algorithms and applications. *TrAC, Trends Anal Chem* 26 (7):752-765.

[24] AC Olivieri, GM Escandar (2019) Analytical chemistry assisted by multi-way calibration: A contribution to green chemistry. *Talanta* 204:700-712.

[25] AC Olivieri, GM Escandar (2014), Chapter 3 - Experimental Three-way/Second-order Data, En: *Practical Three-Way Calibration*, (Ed: AC Olivieri, GM Escandar), Elsevier, Boston, p. 27-45.

[26] AP Teixeira, R Oliveira, PM Alves, MJT Carrondo (2009) Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative. *Biotechnol Adv* 27 (6):726-732.

[27] D Harvey (2000), Chapter 3 - The language of analytical chemistry, En: *Modern Analytical Chemistry*, (Ed: D Harvey), McGrow-Hill, Boston, p. 35-52.

[28] AC Olivieri (2014) Analytical Figures of Merit: From Univariate to Multiway Calibration. *Chem Rev* 114 (10):5358-5378.

[29] J Vessman, RI Stefan, JFv Staden, K Danzer, W Lindner, DT Burns, A Fajgelj, H Müller (2001) Selectivity in analytical chemistry (IUPAC Recommendations 2001). *Pure Appl Chem* 73 (8):1381-1386.

[30] M Streefland, DE Martens, EC Beuvery, RH Wijffels (2013) Process analytical technology (PAT) tools for the cultivation step in biopharmaceutical production. *Eng Life Sci* 13 (3):212-223.

[31] AO Kirdar, KD Green, AS Rathore (2008) Application of multivariate data analysis for identification and successful resolution of a root cause for a bioprocessing application. *Biotechnol Prog* 24 (3):720-726.

[32] N Bhushan, S Hadpe, AS Rathore (2012) Chemometrics applications in biotech processes: assessing process comparability. *Biotechnol Prog* 28 (1):121-128.

[33] A Teixeira, C A.M. Portugal, N Carinhas, J Dias, J Crespo, P Alves, MJT Carrondo, R Oliveira (2009) In Situ 2D Fluorometry and Chemometric Monitoring of Mammalian Cell Cultures. *Biotechnol Bioeng* 102:1098-1106.

[34] B Li, M Shanahan, A Calvet, KJ Leister, AG Ryder (2014) Comprehensive, quantitative bioprocess productivity monitoring using fluorescence EEM spectroscopy and chemometrics. *Analyst* 139 (7):1661-1671.

[35] DAM Pais, RMC Portela, MJT Carrondo, IA Isidro, PM Alves (2019) Enabling PAT in insect cell bioprocesses: In situ monitoring of recombinant adeno-associated virus production by fluorescence spectroscopy. *Biotechnol Bioeng* 116 (11):2803-2814.

[36] C Rafferty, K Johnson, J O'Mahony, B Burgoyne, R Rea, K Balss (2020) Analysis of chemometric models applied to Raman spectroscopy for monitoring key metabolites of cell culture. *Biotechnol Progr* 36:e2977.

[37] AO Kirdar, G Chen, J Weidner, AS Rathore (2010) Application of near-infrared (NIR) spectroscopy for screening of raw materials used in the cell culture medium for the production of a recombinant therapeutic protein. *Biotechnol Prog* 26 (2):527-531.

[38] B Li, PW Ryan, M Shanahan, KJ Leister, AG Ryder (2011) Fluorescence excitation-emission matrix (EEM) spectroscopy for rapid identification and quality evaluation of cell culture media components. *Appl Spectrosc* 65 (11):1240-1249.

[39] Y Jin, SJ Qin, Q Huang, V Saucedo, Z Li, A Meier, S Kundu, B Lehr, S Charaniya (2019) Classification and Diagnosis of Bioprocess Cell Growth Productions Using Early-Stage Data. *Ind Eng Chem Res* 58 (30):13469-13480.

[40] R Bro (2003) Multivariate calibration: What is in chemometrics for the analytical chemist? *Anal Chim Acta* 500 (1):185-194.

[41] B Hassett, E Singh, E Mahgoub, J O'Brien, SM Vicik, B Fitzpatrick (2018) Manufacturing history of etanercept (Enbrel(®)): Consistency of product quality through major process revisions. *MAbs* 10 (1):159-165.

[42] J Trygg, J Gabrielsson, T Lundstedt (2009), 2.01 - Background Estimation, Denoising, and Preprocessing, En: *Comprehensive Chemometrics*, (Ed: S Brown, R Tauler, B Walczak), Elsevier, p. 1-8.

[43] K Kumar, M Tarai, AK Mishra (2017) Unconventional steady-state fluorescence spectroscopy as an analytical technique for analyses of complex-multifluorophoric mixtures. *TrAC, Trends Anal Chem* 97:216-243.

[44] Å Rinnan, CM Andersen (2005) Handling of first-order Rayleigh scatter in PARAFAC modelling of fluorescence excitation–emission data. *Chemom Intell Lab Syst* 76 (1):91-99.

[45] M Bahram, R Bro, C Stedmon, A Afkhami (2006) Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. *J Chemom* 20 (3-4):99-105.

[46] JR Lakowicz (2006), 2 Chapter 2 - Instrumentation for Fluorescence Spectroscopy, En: *Principles of Fluorescence Spectroscopy*, (Ed: JR Lakowicz), Springer US, Boston, MA, p. 27-61.

[47] BE Wilson, BR Kowalski (1989) Quantitative analysis in the presence of spectral interferents using second-order nonbilinear data. *Anal Chem* 61 (20):2277-2284.

[48] S Elcoroaristizabal, R Bro, J García, L Alonso (2015) PARAFAC models of fluorescence data with scattering: A comparative study. *Chemom Intell Lab Syst* 142:124-130.

[49] Å Rinnan, KS Booksh, R Bro (2005) First order Rayleigh scatter as a separate component in the decomposition of fluorescence landscapes. *Anal Chim Acta* 537 (1):349-358.

[50] K Kumar, AK Mishra (2013) Analysis of dilute aqueous multifluorophoric mixtures using excitation–emission matrix fluorescence (EEMF) and total synchronous fluorescence (TSF) spectroscopy: A comparative evaluation. *Talanta* 117:209-220.

[51] M McKnight Diane, W Boyer Elizabeth, K Westerhoff Paul, T Doran Peter, T Kulbe, T Andersen Dale (2001) Spectrofluorometric characterization of dissolved organic

matter for indication of precursor organic material and aromaticity. *Limnology and Oceanography* 46 (1):38-48.

[52] LG Thygesen, Å Rinnan, S Barsberg, JKS Møller (2004) Stabilizing the PARAFAC decomposition of fluorescence spectra by insertion of zeros outside the data area. *Chemom Intell Lab Syst* 71 (2):97-106.

[53] CM Andersen, R Bro (2003) Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. *J Chemom* 17 (4):200-215.

[54] RG Zepp, WM Sheldon, MA Moran (2004) Dissolved organic fluorophores in southeastern US coastal waters: correction method for eliminating Rayleigh and Raman scattering peaks in excitation–emission matrices. *Marine Chemistry* 89 (1):15-36.

[55] PHC Eilers, PM Kroonenberg (2014) Modeling and correction of Raman and Rayleigh scatter in fluorescence landscapes. *Chemom Intell Lab Syst* 130:1-5.

[56] MJ Rodríguez-Cuesta, R Boqué, FX Rius, D Picón Zamora, M Martínez Galera, A Garrido Frenich (2003) Determination of carbendazim, fuberidazole and thiabendazole by three-dimensional excitation–emission matrix fluorescence and parallel factor analysis. *Anal Chim Acta* 491 (1):47-56.

[57] S Elcoroaristizabal, R Callejón, J Amigo, J Ocaña, ML Morales, C Ubeda (2016) Fluorescence Excitation-Emission Matrix Spectroscopy as a Tool for Determining Quality of Sparkling Wines. *Food Chem* 206:284-290.

[58] R Bro (1999) Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemom Intell Lab Syst* 46 (2):133-147.

[59] J Kim, H-M Cho, G Kim (2018) Significant production of humic fluorescent dissolved organic matter in the continental shelf waters of the northwestern Pacific Ocean. *Sci Rep* 8:4887.

[60] W Mendoza, E Weiss, B Schieber, B Greg Mitchell (2017) Controls on the distribution of fluorescent dissolved organic matter during an under-ice algal bloom in the western Arctic Ocean: Distribution of FDOM in the Arctic Ocean. *Global Biogeochem Cycles* 31 (7):1118-1140.

[61] U Wünsch, K Murphy, C Stedmon (2015) Fluorescence Quantum Yields of Natural Organic Matter and Organic Compounds: Implications for the Fluorescence-based Interpretation of Organic Matter Composition. *Front Mar Sci* 2:1-15.

[62] PHC Eilers (2004) Parametric Time Warping. *Anal Chem* 76 (2):404-411.

[63] R Bro, AK Smilde (2014) Principal component analysis. *Anal Methods* 6 (9):2812-2831.

[64] D Ballabio, V Consonni (2013) Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Anal Methods* 5:3790-3798.

[65] R Tauler, A de Juan (2015), Chapter 5 - Multivariate Curve Resolution for Quantitative Analysis, En: *Data Handling in Science and Technology*, (Ed: AM de la Peña, HC Goicoechea, GM Escandar, AC Olivieri), Elsevier, Boston, p. 247-292.

[66] N Ceaglio, M Bollati-Fogolín, M Oggero, M Etcheverrigaray, R Kratje (2014), 6.2 - High cell density cultivation process, En: *Animal Cell Biotechnology*, (Ed: H Hauser, R Wagner), De Gruyter, Berlin, p. 427-454.

[67] K Esbensen, P Geladi (2009), 2.13 - Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History, Practice, En: *Comprehensive chemometrics*, (Ed: S Brown, R Tauler, B Walczak, p. 211-226.

[68] A Olivieri (2018), Chapter 4 - Principal Component Analysis, En: *Introduction to Multivariate Calibration*, (Ed: A Olivieri), Springer, p. 57-71.

[69] AC Olivieri, GM Escandar (2014), Chapter 9 - Partial Least-Squares with Residual Bilinearization, En: *Practical Three-Way Calibration*, (Ed: AC Olivieri, GM Escandar), Elsevier, Boston, p. 157-195.

[70] AC Olivieri, GM Escandar (2014), Chapter 2 - Data Properties, En: *Practical Three-Way Calibration*, (Ed: AC Olivieri, GM Escandar), Elsevier, Boston, p. 11-26.

[71] A de Juan, R Tauler (2016), Chapter 2 - Multivariate Curve Resolution-Alternating Least Squares for Spectroscopic Data, En: *Data Handling in Science and Technology*, (Ed: C Ruckebusch), Elsevier, p. 5-51.

[72] W Windig, J Guilment (1991) Interactive self-modeling mixture analysis. *Anal Chem* 63 (14):1425-1432.

[73] AC Olivieri (2020) Second-order multivariate calibration with the extended bilinear model: Effect of initialization, constraints, and composition of the calibration set on the extent of rotational ambiguity. *J Chemom* 34 (3):e3130.

[74] DM Haaland, EV Thomas (1988) Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal Chem* 60 (11):1193-1202.

[75] A Olivieri (2018), Chapter 7 - Partial Least-Squares Model, En: *Introduction to Multivariate Calibration*, (Ed: A Olivieri), Springer, p. 103-121.

[76] M Cocchi, A Biancolillo, F Marini (2018), Chapter Ten - Chemometric Methods for Classification and Feature Selection, En: *Comprehensive Analytical Chemistry*, (Ed: J Jaumot, C Bedia, R Tauler), Elsevier, p. 265-299.

[77] D Ballabio, F Grisoni, R Todeschini (2018) Multivariate comparison of classification performance measures. *Chemom Intell Lab Syst* 174:33-44.

[78] EMEA (2006) *Guideline on the environmental risk assessment of medicinal products for humans use CHMP/SWP/4447/00.*

https://www.ema.europa.eu/en/environmental-risk-assessment-medical-products-human-use.

[79] ICH (2005) *International conference on harmonization of technical requirements for registration of pharmaceuticals for human use. Validation of analytical procedures: Text and methodology Q2 (R1).* https://www.ema.europa.eu/en/documents/scientific-guideline/ich-q-2-r1-validation-analytical-procedures-text-methodology-step-5_en.pdf.

[80] SM Faassen, B Hitzmann (2015) Fluorescence spectroscopy and chemometric modeling for bioprocess monitoring. *Sensors* 15 (5):10271-10291.

[81] T-T Wong (2015) Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit* 48 (9):2839-2846.

[82] D Ballabio (2015) A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. *Chemom Intell Lab Syst* 149:1-9.

[83] AC Olivieri, H-L Wu, R-Q Yu (2009) MVC2: A MATLAB graphical interface toolbox for second-order multivariate calibration. *Chemom Intell Lab Syst* 96 (2):246-251.

[84] CD Agarabi, BK Chavez, SC Lute, EK Read, S Rogstad, D Awotwe-Otoo, MR Brown, MT Boyne II, KA Brorson (2017) Exploring the linkage between cell culture process parameters and downstream processing utilizing a plackett-burman design for a model monoclonal antibody. *Biotechnol Progr* 33 (1):163-170.

[85] AC Olivieri, GM Escandar (2014), Chapter 5 - Parallel Factor Analysis: Trilinear Data, En: *Practical Three-Way Calibration*, (Ed: AC Olivieri, GM Escandar), Elsevier, Boston, p. 65-92.

[86] JR Lakowicz (2006), Chapter 8 - Quenching of fluorescence, En: *Principles of Fluorescence Spectroscopy*, (Ed: JR Lakowicz), Springer US, Boston, MA, p. 277-330.

[87] A de Juan, SC Rutan, R Tauler (2020), 2.10 - Two-way Data Analysis: Multivariate Curve Resolution, Iterative Methods, En: *Comprehensive Chemometrics (Second Edition)*, (Ed: S Brown, R Tauler, B Walczak), Elsevier, Oxford, p. 153-171.

[88] S Wold, M Sjöström, L Eriksson (2001) PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58 (2):109-130.

[89] HC Goicoechea, AC Olivieri (1999) Simultaneous determination of rifampicin, isoniazid and pyrazinamide in tablet preparations by multivariate spectrophotometric calibration. *J Pharm Biomed Anal* 20 (4):681-686.

[90] I Eide, G Neverdal, B Thorvaldsen, R Arneberg, B Grung, OM Kvalheim (2004) Toxicological evaluation of complex mixtures: fingerprinting and multivariate analysis. *Environ Toxicol Pharmacol* 18 (2):127-133.

[91] S Wold, N Kettaneh-Wold, B Skagerberg (1989) Nonlinear PLS modeling. *Chemom Intell Lab Syst* 7 (1):53-65.

[92] A Höskuldsson (1992) Quadratic PLS regression. *J Chemom* 6 (6):307-334.

[93] R Rosipal, L Trejo (2001) Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. *J Mach Learn Res* 2:97-123.

[94] RD Cook, L Forzani (2020) Envelopes: A new chapter in partial least squares regression. *J Chemom* 34 (10):e3287.

[95] RD Cook, L Forzani (2021) PLS regression algorithms in the presence of nonlinearity. *Chemom Intell Lab Syst* 213:104307.

[96] J Zupan (1995) Neural Networks in Chemistry. *Angew Chem Int Ed* 32:469-470.

[97] F Despagne, D Luc Massart (1998) Neural Networks in Multivariate Calibration. *The Analyst* 123:157R-178R.

[98] JA Lopes, JC Menezes (2004) Multivariate monitoring of fermentation processes with non-linear modelling methods. *Anal Chim Acta* 515 (1):101-108.

[99] K-I Lee, Y-S Yim, S-W Chung, J Wei, JI Rhee (2005) Application of artificial neural networks to the analysis of two-dimensional fluorescence spectra in recombinant E coli fermentation processes. *J Chem Technol Biotechnol* 80 (9):1036-1045.

[100] O Paquet-Durand, S Assawarajuwan, B Hitzmann (2017) Artificial neural network for bioprocess monitoring based on fluorescence measurements: Training without offline measurements. *Eng Life Sci* 17 (8):874-880.

[101] J Lopez, E Caicedo Bravo (2009), Capítulo 2 - Redes neuronales perceptrón y adaline, En: *Una aproximación práctica a las Redes Neuronales Artificiales*, Universidad del Valle, p. 37-73.

[102] RH Myers, DC Montgomery, C Anderson-Cook (2016), 1. Introduction, En: *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, Wiley.

[103] G Derringer, R Suich (1980) Simultaneous optimization of several response variables. *J Qual Technol* 12 (4):214-219.

[104] L Nørgaard, A Saudland, J Wagner, JP Nielsen, L Munck, S Engelsen (2000) Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Appl Spectrosc* 54:413-419.

[105] CL Mallows (1986) Augmented Partial Residuals. *Technometrics* 28 (4):313-319.

[106] J Riu, FX Rius (1997) Method comparison using regression with uncertainties in both axes. *TrAC, Trends Anal Chem* 16 (4):211-216.

[107] AC Olivieri, HC Goicoechea, FA Iñón (2004) MVC1: an integrated MatLab toolbox for first-order multivariate calibration. *Chemom Intell Lab Syst* 73 (2):189-197.

[108] PC Giordano, HC Goicoechea, AC Olivieri (2017) SRO_ANN: An integrated MatLab toolbox for multiple surface response optimization using radial basis functions. *Chemom Intell Lab Syst* 171:198-206.

[109] H Chung, H Lee, C-H Jun (2001) Determination of Research Octane Number using NIR Spectral Data and Ridge Regression. *Bull Korean Chem Soc* 22:37-42.

[110] C Musmann, K Joeris, S Markert, D Solle, T Scheper (2016) A review of spectroscopic methods and their applicability for high-throughput characterization of mammalian cell cultures in automated cell culture systems. *Eng Life Sci* 16:405-416.

[111] SP Gurden, JA Westerhuis, AK Smilde (2002) Monitoring of batch processes using spectroscopy. *AIChE Journal* 48 (10):2283-2297.

[112] G Rounaghi, R Mohammad Zadeh Kakhki, T Heidari (2011) Artificial Neural Networks Applied for Simultaneous Analysis of Mixtures of Nitrophenols by Conductometric Acid–Base Titration. *Ind Eng Chem Res* 50 (19):11375-11381.

[113] H-C Yao, M Sun, X-F Yang, Z-Z Zhang, H Li (2011) Simultaneous determination of captopril and hydrochlorothiazide by time-resolved chemiluminescence with artificial neural network calibration. *J Pharm Anal* 1 (1):32-38.

[114] B Wang, G Liu, Y Dou, L Liang, H Zhang, Y Ren (2009) Quantitative analysis of diclofenac sodium powder via near-infrared spectroscopy combined with artificial neural network. *J Pharm Biomed Anal* 50 (2):158-163.

[115] C Bessant, S Saini (1999) Simultaneous Determination of Ethanol, Fructose, and Glucose at an Unmodified Platinum Electrode Using Artificial Neural Networks. *Anal Chem* 71 (14):2806-2813.

[116] F Marini (2009) Artificial neural networks in foodstuff analyses: Trends and perspectives A review. *Anal Chim Acta* 635 (2):121-131.

[117] D Stratiev, I Marinov, R Dinkov, I Shishkova, I Velkov, I Sharafutdinov, S Nenov, T Tsvetkov, S Sotirov, M Mitkova, N Rudnev (2015) Opportunity to Improve Diesel-Fuel Cetane-Number Prediction from Easily Available Physical Properties and Application of the Least-Squares Method and Artificial Neural Networks. *Energy & Fuels* 29 (3):1520-1533.

[118] G Hanrahan (2010) Computational Neural Networks Driving Complex Analytical Problem Solving. *Anal Chem* 82 (11):4307-4313.

[119] M Jalali-Heravi (2009) Neural Networks in Analytical Chemistry. *Methods Mol Biol* 458:78-118.

[120] ZB Alfassi, Z Boger, Y Ronen (2009), Chapter 13 - Artificial Neural Networks in Analytical Chemistry, En: *Statistical Treatment of Analytical Data*, Wiley-Blackwell, Oxford.

[121] JTG Hwang, AA Ding (1997) Prediction Intervals for Artificial Neural Networks. *J Am Stat Assoc* 92 (438):748-757.

[122] L Zhang, P Luh, K Kasiviswanathan (2003) Energy clearing price prediction and confidence interval estimation with cascaded neural networks. *IEEE Trans Power Syst* 18:99-105.

[123] E Mazloumi, G Rose, G Currie, S Moridpour (2011) Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction. *Eng Appl Artif Intell* 24:534-542.

[124] F Allegrini, AC Olivieri (2016) Sensitivity, Prediction Uncertainty, and Detection Limit for Artificial Neural Network Calibrations. *Anal Chem* 88 (15):7807-7812.

[125] EPPA Derks, MSS Pastor, LMC Buydens (1995) Robustness analysis of radial base function and multi-layered feed-forward neural network models. *Chemom Intell Lab Syst* 28 (1):49-60.

[126] K Faber, BR Kowalski (1997) Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. *J Chemom* 11 (3):181-238.

[127] EPPA Derks, MS Sánchez Pastor, LMC Buydens (1996) Response to "Comment on a recent sensitivity analysis of radial base function and multi-layer feed-forward neural network models". *Chemom Intell Lab Syst* 34 (2):299-301.

[128] PB Harrington, A Urbas, C Wan (2000) Evaluation of neural network models with generalized sensitivity analysis. *Anal Chem* 72 (20):5004-5013.

[129] X Zeng, DS Yeung (2001) Sensitivity analysis of multilayer perceptron to input and weight perturbations. *IEEE transactions on neural networks* 12 (6):1358-1366.

[130] F Allegrini, AC Olivieri (2020), 2.20 - Figures of Merit, En: *Comprehensive Chemometrics (Second Edition)*, (Ed: S Brown, R Tauler, B Walczak), Elsevier, Oxford, p. 441-463.

[131] M Valcárcel, Á Ríos (1999) A metrological hierarchy for analytical chemistry. *TrAC, Trends Anal Chem* 18:68-75.

[132] D Klaus, LA Currie (1998) Guidelines for calibration in analytical chemistry. Part I. Fundamentals and single component calibration (IUPAC Recommendations 1998). *Pure Appl Chem* 70 (4):993-1014.

[133] LA Currie (1999) Nomenclature in Evaluation of Analytical Methods Including Detection and Quantification Capabilities IUPAC Recommendations 1995, in Validation of Analytical Methods. *Anal Chim Acta* 391 (2):105-126.

[134] A Lorber, K Faber, BR Kowalski (1997) Net Analyte Signal Calculation in Multivariate Calibration. *Anal Chem* 69 (8):1620-1626.

[135] AC Olivieri (2018), Chapter 10 - Analytical Figures of Merit, En: *Introduction to Multivariate Calibration*, Springer.

[136] AC Olivieri, NM Faber, J Ferré, R Boqué, JH Kalivas, H Mark (2006) Uncertainty estimation and figures of merit for multivariate calibration (IUPAC Technical Report). *Pure Appl Chem* 78 (3):633-661.

[137] F Allegrini, P D. Wentzell, A Olivieri (2015) Generalized error-dependent prediction uncertainty in multivariate calibration. *Anal Chim Acta* 903:51-60.

[138] RK Skogerboe, CL Grant (1970) Comments OH the Definitions of the Terms Sensitivity and Detection Limit. *Spectrosc Lett* 3 (8-9):215-220.

[139] DP Kroese, T Brereton, T Taimre, ZI Botev (2014) Why the Monte Carlo method is so important today. *WIREs Comput Stat* 6 (6):386-392.

[140] B Efron (1979) Bootstrap Methods: Another Look at the Jackknife. *Ann Stat* 7 (1):1-26, 26.

[141] M Dathe, M Otto (1996) Confidence intervals for calibration with neural networks. *Fresenius J Anal Chem* 356 (1):17-20.

[142] Y Xu, R Goodacre (2018) On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J Anal Test* 2 (3):249-262.

[143] MR Alcaraz, MJ Culzoni, HC Goicoechea (2016) Enhanced fluorescence sensitivity by coupling yttrium-analyte complexes and three-way fast high-performance liquid chromatography data modeling. *Anal Chim Acta* 902:50-58.

[144] F Allegrini, AC Olivieri (2017) Recent advances in analytical figures of merit: heteroscedasticity strikes back. *Anal Methods* 9 (5):739-743.