UNIVERSITAT ROVIRA i VIRGILI

DESENVOLUPAMENT DE MODELS DE PREDICCIÓ DE COLESTEROL I TRIGLICÈRIDS EN PLASMA A PARTIR D'ESPECTRES DE ^{1}H-RMN

Mireia Gasco Agorreta

TREBALL FINAL DE GRAU – BIOTECNOLOGIA

Tutor acadèmic: Ricardo Cordero Otero

Adreça electrònica: ricardo.cordero@urv.cat

Tutor professional: Sara Samino, Biosfer Teslab - Tarragona

Juny 2023

Jo, Mireia Gasco Agorreta , amb DNI 77793108R, soc coneixedora de la guia de prevenció del plagi a la URV: “Prevenció, detecció i tractament del plagi en la docència: guia per a estudiants” (aprovada el juliol 2017) (http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/) i afirmo que aquet TFG no constitueix cap de les conductes considerades com a plagi per la URV.

Tarragona, juny de 2023

Índex

Agraïments

En primer lloc, vull agrair a tot l'equip de Biosfer Teslab el seu suport i per tractar-me com una més de l'equip, el que m'ha permès aprendre moltíssim i adquirir noves habilitats que no només m'han permès dur a terme aquest treball, sinó que estic segura que també em seran molt útils en un futur. Especialment, vull donar les gràcies a la Sara Samino i al Daniel Rodríguez per ajudar-me i acompanyar-me en tot el desenvolupament del treball i estar sempre disponibles per donar-me un cop de mà quan l'he necessitat.

Finalment, també vull agrair el meu tutor acadèmic, Ricardo Cordero, per les seves recomanacions i indicacions, que han estat molt útils per al desenvolupament d'aquest treball.

Dades del Centre

Biosfer Teslab és una empresa spin-off de la Universitat Rovira i Virgili (URV) i de l'Institut de Recerca Sanitària Pere Virgili (IISPV). Opera en el camp del diagnòstic in vitro amb l'objectiu de prestar serveis analítics per a l'estudi i seguiment de les alteracions del metabolisme. La seva principal finalitat és la d'avançar en el coneixement i millorar la salut de les persones. L'empresa es va fundar al desembre de 2013 i es dedica a aplicar la tecnologia d'anàlisi de ressonància magnètica nuclear (RMN) en mostres de plasma amb l'objectiu de desenvolupar sistemes de diagnòstic in vitro per a buscar nous biomarcadors de malalties. El producte estrella de l'empresa és un test avançat de lipoproteïnes (Liposcale® Test) basat en l'anàlisi del senyal de RMN dels grups metil i metilè, que constitueixen el colesterol i els triglicèrids transportats per les lipoproteïnes, i que permeten caracteritzar exhaustivament el perfil lipoproteic en els seus diferents components. A partir d'aquest test, l'empresa ha aconseguit avançar en l'extracció d'informació dels espectres de RMN, validant l'ús de l'espectroscòpia de RMN per a la caracterització de biofluids i bioteixits en entorns de recerca biomèdica i clínica, podent oferir un perfil global de metabolòmica per RMN que inclou, a més del perfil de lipoproteïnes, el perfil de glicoproteïnes, la caracterització del metaboloma del plasma aquós (LMWM) i la caracterització del lipidoma.

Resum i Paraules Clau

Les malalties cardiovasculars (CVD) tenen un gran impacte en la societat actual, una tendència que ha anat augmentant en els últims anys, pel que és d'especial interès desenvolupar nous mètodes que permetin fer-ne un diagnòstic precoç i acurat. La irrupció de la metabolòmica ha permès que es comencin a utilitzar noves tècniques analítiques, com al ressonància magnètica nuclear (RMN) per a la predicció de lípids, molt relacionats amb el risc cardiovascular. Aquest treball mostra el desenvolupament i avaluació de dos models de basats en el mètode de mínims quadrats parcials (PLS) per a la predicció de la concentració de colesterol i triglicèrids en mostres de plasma prèviament analitzades per RMN.

En primer lloc, es va dur a terme un procés de neteja de les dades de partida, per eliminar possibles "outliers". Les mostres restants es van utilitzar per al calibratge i validació dels dos models de predicció. Es van dur a terme proves per determinar quines condicions generaven els models més precisos, sent els paràmetres a tenir en compte el tipus de divisió de les dades, el percentatge de divisió, el tipus de preprocessat de les dades i les variables latents fetes servir per a la creació dels models. Un cop determinades les millors configuracions en cada cas, es van polir els models manualment per assolir els millors resultats possibles.

Els models finals obtinguts van donar una correlació de r = 0,94017 i r = 0,95394 per al model de colesterol i triglicèrids, respectivament. Es van validar aquests models amb un conjunt extern, diferent dels utilitzats per al calibratge i la validació, per garantir la integritat d'aquests. Els resultats en aquest cas van ser de r = 0,9606 i r = 0,9014, per a colesterol i triglicèrids, respectivament. Aquests resultats demostren que els models entrenats en aquest treball permeten predir acuradament les concentracions de colesterol i triglicèrids en mostres de plasma a partir dels espectres de RMN.

Paraules clau

Malalties cardiovasculars (CVD)

Ressonància magnètica nuclear (RMN)

Mètode dels mínims quadrats parcials (PLS)

Metabolòmica

1. Introducció

1.1. El paper dels lípids en les CVDs

Les malalties cardiovasculars (CVD per les seves sigles en anglès) són totes aquelles patologies amb afectacions en els vasos sanguinis i el cor, englobant malalties com les cardiopaties isquèmiques, els ictus o les cardiopaties hipertenses. En l'àmbit mundial, les CVD constitueixen la primera causa de mortalitat i morbiditat segons els informes de la Organització Mundial de la Salut (OMS), motiu pel qual s'han iniciat nombrosos projectes per reduir-ne l'afectació (Joseph et al., 2017).

Habitualment, aquestes malalties es poden prevenir amb hàbits saludables, encara que hi ha factors ambientals i/o genètics també poden desencadenar la seva aparició. Per això, detectar aquests factors de risc de forma precoç és un dels principals objectius en aquest camp. No només per tal de conèixer en detall el funcionament d'aquestes malalties, sinó per poder detectar els factors de risc de forma precoç en els pacients, augmentant les probabilitats d'un desenllaç favorable.

En aquesta línia, el procediment més habitual és la quantificació del colesterol i, en grau més baix, dels triglicèrids en sang. Aquests dos lípids s'han relacionat directament amb les CVD, sobretot amb malalties com l'aterosclerosi. En els últims anys, però, s'ha observat que el colesterol, que es creia que n'era el causant directe, és només un dels factors que contribueixen al desenvolupament d'aquestes malalties. Així, s'ha vist que altres lípids com els triglicèrids o les mateixes lipoproteïnes que els transporten també tenen un paper fonamental en el desenvolupament de les CVD (Burnett et al., 2020; Farnier et al., 2021; Nordestgaard, 2016). A més a més, estudis recents proposen utilitzar les quantificacions de diversos tipus de lípids a l'hora de determinar el risc cardiovascular d'un pacient asimptomàtic, entre els quals es destaquen els ja esmentats colesterol i triglicèrids, però també altres paràmetres com la lipoproteïna A, la APO B, o el ràtio APO B / APO A. També es destaca la importància del colesterol en lipoproteïnes de baixa densitat (colesterol LDL) i la resta de tipus de colesterol, que poden estar continguts en una gran varietat de lipoproteïnes (Francula-Zaninovic & Nola, 2018).

Aquest tipus de propostes comporten un canvi de paradigma quant a procediments de detecció, que tradicionalment es feien mitjançant l'anàlisi bioquímica de mostres de plasma. L'augment de la quantitat de metabòlits d'interès fa que es plantegin tècniques alternatives a aquestes anàlisis tradicionals, que si bé són molt eficients quan es volen mesurar paràmetres concrets, poden veure's perjudicats per l'augment i la tipologia dels nous lípids a avaluar. D'altra banda, no es pot perdre de vista que és probable que la llista d'aquests paràmetres d'interès continuï augmentant a mesura que es descobreixin més actors en el desenvolupament de les CVD.

1.2. Tècniques d'anàlisi avançades

En aquest context és en el que tècniques d'anàlisi avançades poden ser d'interès. Donada la gran quantitat de metabòlits a analitzar, aquest problema passa a poder-se analitzar des del punt de vista de la metabolòmica i, més concretament, la lipidòmica, que és el camp de la metabolòmica que se centra en l'anàlisi de lípids. Un exemple d'aquestes tècniques és la ressonància magnètica nuclear (RMN). Aquesta tècnica es basa en la detecció del senyal electromagnètic produït en excitar els àtoms d'una mostra mitjançant un camp magnètic.

Aquest procés es produeix en un punt molt proper al punt de ressonància, quan la freqüència d'oscil·lació coincideix amb la freqüència intrínseca del nucli utilitzat (generalment ^{1}H), i depèn de l'entorn químic en el qual es troba l'àtom.(Nagana Gowda & Raftery, 2021).

En les anàlisis clàssiques, l'ús d'aquesta tecnologia no era eficient, per motius com l'elevat cost de l'equipament i la necessitat de fer una preparació prèvia de les mostres molt més complexa. En canvi, quan tenim en compte la dificultat de mesurar certs paràmetres amb els mètodes bioquímics, com el nombre de partícules o la mida de les lipoproteïnes, amb l'augment dels paràmetres que ens interessa mesurar, l'ús d'aquesta tecnologia comença a ser més prometedor.

Així, la RMN, que tradicionalment s'havia utilitzat per a la detecció de compostos i la dilucidació de la seva estructura, pot utilitzar-se també per a la quantificació massiva de molècules en una mostra, mitjançant una sola anàlisi d'aquesta. Aquesta tècnica, encara que és menys sensible que l'espectrometria de masses, té altres característiques que la fan ser més interessant en el camp de la metabolòmica i en l'anàlisi de mostres biològiques. Les característiques més interessants són la seva alta reproductibilitat i la capacitat d'analitzar mostres de fluids biològics sense necessitat d'aplicar-hi mètodes de preparació prèvia que puguin ser destructius. Això permet garantir que tots els components de la mostra es mantinguin correctament per tal que la mesura sigui la més acurada possible (Nagana Gowda & Raftery, 2021).

Els espectres obtinguts de les mesures amb RMN representen tots els metabòlits observables en la mostra, i la seva quantificació és possible mitjançant la mesura de l'àrea del pic, que és proporcional a la concentració de cadascun. En el cas dels lípids, donat que tots presenten grups funcionals molt similars, és complex determinar directament quins pics corresponen a cada grup de lípids, i per això calen eines computacionals avançades que ens permetin establir les relacions entre els pics representats en els espectres i la quantificació dels lípids d'interès.

També cal comentar que aquesta tècnica presenta molts reptes, sobretot en casos on els pics dels espectres es troben superposats, o quan els pics de diferents mostres es troben desplaçats. En aquests casos cal fer un tractament previ dels espectres abans de poder-los utilitzar per a la quantificació dels diferents metabòlits, o bé tenir aquestes casuístiques en compte a l'hora d'elaborar el programari encarregat de fer la quantificació (Nagana Gowda & Raftery, 2023).

Tenint en compte aquests desavantatges, cal explicar per què la RMN segueix sent la tècnica més interessant per a resoldre aquest problema, tenint alternatives com l'espectroscòpia de masses (EM), una tècnica molt utilitzada en l'àmbit de la metabolòmica (Han & Gross, 2022; Heiles, 2021). L'EM presenta una sensitivitat molt més elevada que la RMN i té molta més capacitat detectora, ja que pot arribar a detectar més de 500 metabòlits en una mostra, sempre que s'utilitzin les tècniques adequades. En comparació, la RMN pot arribar a les 200, i sempre segons la resolució de l'espectròmetre (Emwas, 2015).

Tot i aquestes diferències, hi ha tres característiques de la RMN que són clau i la fan destacar per davant de l'EM: la seva reproductibilitat, la mínima preparació de les mostres necessària i el fet que és una tècnica no destructiva. Això permet garantir la fiabilitat dels resultats, així com repetir anàlisis en mostres prèviament analitzades en cas de necessitat. A més, el fet que les mostres no requereixin una preparació prèvia tan complexa com les mostres d'EM pot arribar a compensar el fet que la maquinària de RMN és més complexa i requereix professional més qualificat (Emwas, 2015).

En l'àmbit comercial, existeixen diferents propostes que utilitzen aquesta tecnologia per a l'anàlisi de lípids. Algunes plataformes de metabolòmica ofereixen només la possibilitat d'analitzar mostres de plasma per RMN, però com s'ha explicat prèviament, el més complex és l'anàlisi posterior dels espectres. És per això que l'interès principal recau en eines computacionals que permetin extreure la informació d'interès d'aquests espectres, com LipSpin (Barrilero et al., 2018), i en productes com el test Liposcale (Mallol et al., 2015).

1.3. El test Liposcale

El test Liposcale és un test desenvolupat per Biosfer Teslab que utilitza l'espectroscòpia d' ^{1}H RMN per quantificar no només els triglicèrids i colesterol totals presents en una mostra de plasma, sinó també paràmetres més avançats com el nombre de partícules, la composició de lípids per cadascuna de les subclasses de lipoproteïnes i també la mida d'aquestes partícules. Això és possible ja que, en funció de la mida de la partícula, els grups metil dels lípids continguts en ella ressonen a freqüències lleugerament diferents: com més petita és la partícula, més baixa és la freqüència a la qual ressona el lípid (Mallol et al., 2013a). L'interès de la mida de les lipoproteïnes recau en el fet que s'ha detectat que certs fenotips caracteritzats pel nombre de partícules i per la mida d'aquestes poden ser indicadors de possible risc cardiovascular. Per exemple, els pacients amb diabetis i síndromes metabòliques presenten un fenotip comú amb una elevada concentració de partícules LDL petites que els fa ser més propensos a patir CVDs.

Taula 1. Lipoproteïnes analitzades pel test Liposcale i les seves característiques principals, segons Feingold et al., 2021

Lipoproteïna	Densitat (g/ml)	Mida (nm)	Lípids principals
VLDL	0,930 – 1,006	30 – 80	Triglicèrids
IDL	1,006 – 1,019	25 – 35	Triglicèrids Colesterol
LDL	1,019 – 1,063	18 – 25	Colesterol
HDL	1,063 – 1,210	5 – 12	Colesterol Fosfolípids

Aquestes lipoproteïnes són complexes formats per un nucli apolar – generalment èsters de colesterol i/o triglicèrids – envoltats per una membrana hidròfila que en facilita el seu transport pel torrent sanguini. Segons la composició i la densitat de les lipoproteïnes, es poden classificar en 6 classes principals (Feingold et al., 2021), de les quals 4 són analitzades pel test Liposcale (Taula 1). L'anàlisi d'aquestes lipoproteïnes és factible gràcies al fet que la 1H

RMN no requereix una preparació de la mostra invasiva. L'estructura i la integritat d'aquestes partícules es podria veure greument malmesa per les centrifugacions requerides en altres tècniques, com pot ser l'EM mencionada amb anterioritat.

Així, el test Liposcale processa els espectres obtinguts de l'anàlisi de les mostres de plasma per RMN i en fa una deconvolució, el que permet quantificar la mida i tipus de les partícules. En concret, s'obtenen la quantitat de partícules grans, petites i mitjanes que hi ha per cada tipus: VLDL, LDL i HDL. Aquests paràmetres són de gran interès per a la investigació, ja que no es poden aconseguir pels mètodes bioquímics tradicionals.

A més d'això, el test també proporciona els valors de colesterol i triglicèrids que hi ha en cadascun d'aquests tipus de partícules – això és, el contingut de colesterol i triglicèrids present en les partícules VLDL, IDL, LDL i HDL. Aquests valors s'obtenen a partir de models de predicció de lípids, basats en el mètode matemàtic de mínims quadrats parcials (PLS). Pel que fa al colesterol i triglicèrids totals, el seu valor es calcula com la suma de les quantificacions individuals de cada tipus de partícula.

Aquestes quantificacions es basen en models de predicció obtinguts a partir de l'anàlisi de les fraccions de VLDL, IDL, LDL i HDL, separades per ultracentrifugació. Durant aquest procés de separació es produeix pèrdua de mostra, el que genera cert error en els models calculats a partir d'aquestes fraccions. A l'hora de calcular el colesterol i triglicèrids totals, aquest error s'acumula i provoca que els resultats es desviïn lleugerament dels aportats per mètodes tradicionals. Aquestes desviacions, però, són proporcionals, de forma que no suposen un greu problema per al test en qüestió.

Tot i així, tenint en compte que l'estàndard actual són les quantificacions per bioquímica, val la pena intentar corregir aquest error per tal d'assolir resultats més propers als dels mètodes tradicionals.

2. Hipòtesi i Objectius

La desviació observada en els resultats del test “Liposcale” es deu a la pèrdua de mostra en el procés d’ultracentrifugació, motiu pel qual s’obtenen resultats proporcionals amb una lleugera desviació. Aquest error es podria evitar quantificant el colesterol i els triglicèrids totals de forma independent, mitjançant models de predicció propis que no presentessin l’error introduït per la ultracentrifugació.

Així, l’objectiu d’aquest treball és desenvolupar i validar dos models de predicció, un per al colesterol total i un altre per als triglicèrids totals, a partir dels espectres RMN de mostres de plasma. Aquests models permetran obtenir valors més precisos d’aquests dos paràmetres, amb els que es calcularan les fraccions contingudes en les diferents partícules. Això permetrà donar resultats tan semblants com sigui possible als assolits pels mètodes bioquímics tradicionals, que són l’estàndard actual en la indústria i el sistema sanitari.

3. Metodologia

3.1. Selecció i anàlisi de les mostres

3.1.1. Obtenció dels grups

Per poder desenvolupar models capaços de predir les concentracions de colesterol i triglicèrids de forma acurada, calien mostres que representessin tot l'espectre de dades possibles, des de valors estàndard fins a casos d'hipercolesterolèmies i hipertrigliceridèmies (Taula 2).

Taula 2. Classificació de la quantitat de colesterol i triglicèrids, Journal of the American College of Cardiology (JACC).

	Bé	**Moderat**	**Alt**	**Molt Alt**
Colesterol*	< 200	200 - 239	> 240	N/A
Triglicèrids*	< 149	150 - 199	200 - 499	> 500

* mg/dl

Tenint això en compte, inicialment es van seleccionar les dades de tres grups (grups 1, 2 i 3) de mostres de plasma provinents de la base de dades de Biosfer Teslab. Aquests tres grups s'havien analitzat tant per mètodes bioquímics com per RMN i permetien cobrir tot l'espectre de valors d'interès, tant per al colesterol com per als triglicèrids, gràcies a les seves característiques diferencials.

3.1.2. Neteja de la base de dades

Un cop seleccionats els grups de dades, es va procedir a fer una neteja de les dades per tal d'eliminar valors atípics i possibles errors. Aquest procés va consistir en creuar les dades obtingudes del les anàlisis bioquímiques de les mostres amb els espectres obtinguts per RMN. Així, es van descartar aquelles mostres que, o bé no presentaven alguna de les dues anàlisis, o bé presentaven incoherències entre les dues. Per exemple, es va detectar que algunes mostres presentaven espectres atípics per als valors de colesterol i/o triglicèrids aconseguits per bioquímica.

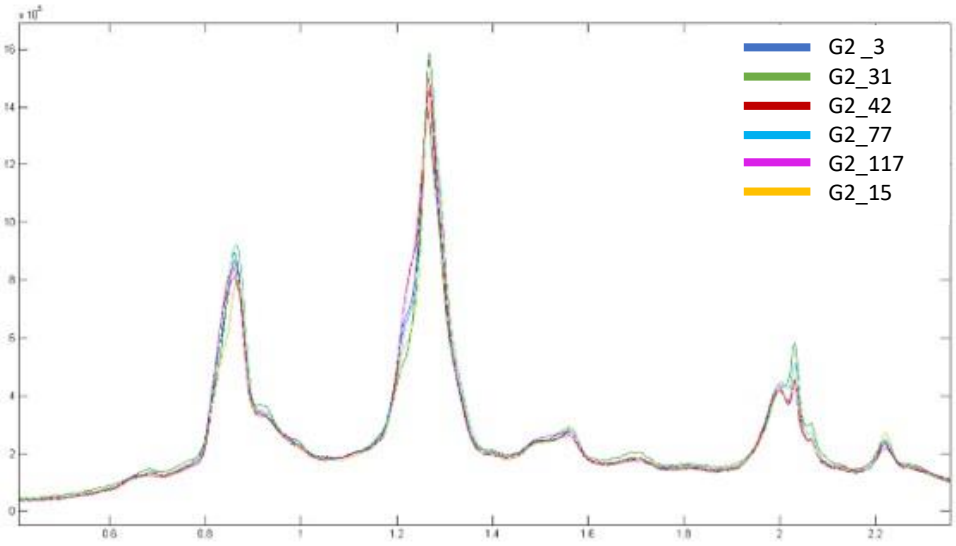

Figura 1. Valor atípic detectat en un dels grups seleccionats. La mostra G2.15 presenta pics similars a mostres amb un valor de colesterol d'entre 150-160 mg/dl, un valor molt allunyat dels 62mg/dl indicats per l'anàlisi bioquímic.

Un cop descartades totes les mostres no vàlides i que podien introduir errors en el model, es va procedir a l'anàlisi estadística de les dades restants per tal de garantir que, efectivament, les mostres seleccionades cobrien tot l'espectre de dades d'interès per al model.

3.1.3. Anàlisi estadístic de les dades

Les dades dels tres grups es van analitzar estadísticament mitjançant RStudio. Es van obtenir taules mostrant els valors estadístics més representatius de cada grup – mitjana, mediana i moda – tant per al colesterol com per als triglicèrids. D'altra banda, també es van generar histogrames i boxplots per tal de poder comprovar el rang de dades cobert pels diferents grups i garantir que disposàvem de mostres en tot l'espectre de dades d'interès.

3.2. Determinació de les regions d'interès de l'espectre de RMN

Un cop seleccionades i analitzades les dades d'interès, es van determinar les regions de l'espectre que eren determinants per a la quantificació de colesterol i triglicèrids, mitjançant l'eina STOCSY.

3.2.1. Determinació de les regions d'interès: STOCSY

L'espectroscòpia de correlació estadística total (STOCSY) és una tècnica d'anàlisi estadístic multivariant que permet obtenir la correlació entre la variable d'interès – en aquest cas, la quantitat de colesterol o triglicèrids – i les diferents coordenades de l'espectre (Barrilero et al., 2015; Cloarec et al., 2005). Aquesta anàlisi permet determinar, en forma de "heatmap", aquelles regions de l'espectre que tenen més correlació amb els paràmetres d'estudi.

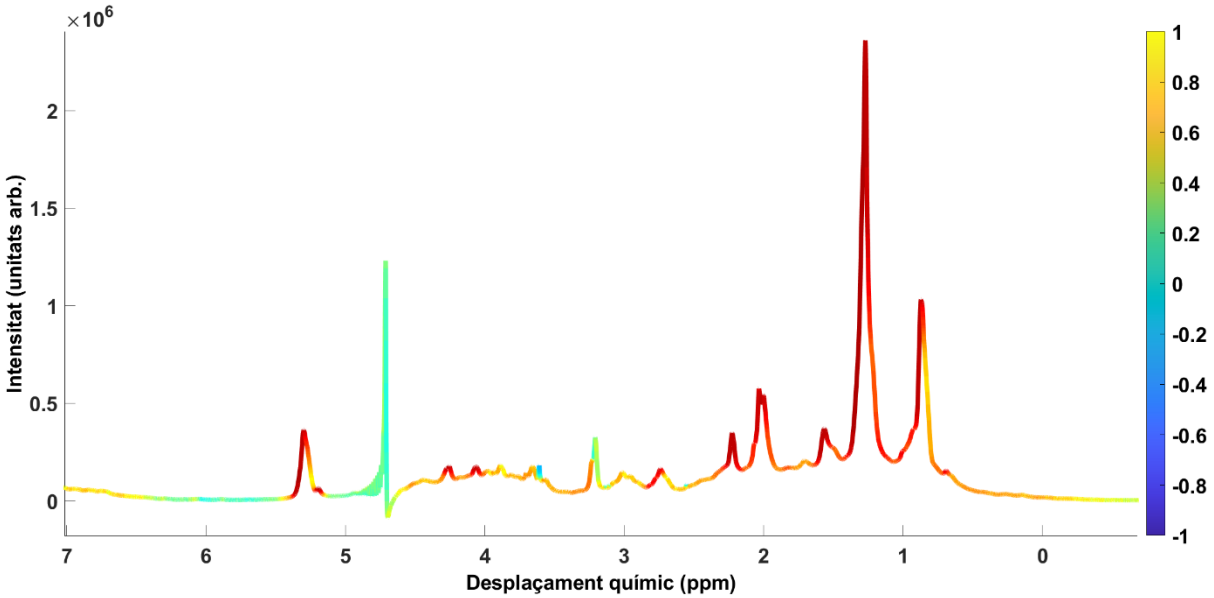

Figura 2. Heatmap d'exemple, obtingut mitjançant l'eina STOCSY.

La Figura 2 mostra un exemple d'un heatmap obtingut per aquesta eina, on es veu l'espectre mitjà de totes les dades analitzades i la correlació obtinguda per a cada zona – com més en vermell es troba una zona, més correlació hi ha amb les variables d'estudi, mentre que com més blava sigui, menys n'hi ha.

3.2.2. Criteris d'inclusió/exclusió de les regions de l'espectre

El principal criteri d'inclusió de les regions va consistir en l'avaluació dels heatmaps obtinguts de l'anàlisi STOCSY, de forma que les regions seleccionades són aquelles que mostren més correlació – les que es troben marcades en vermell més fosc. Així, les zones que presentaven un índex de més de 0,8 eren candidates a ser seleccionades, encara que també es van tenir en compte altres criteris, com l'avaluació dels grups funcionals que podien estar generant els diferents pics.

És per això que es va tenir en compte els grups funcionals específics de cada grup de lípids per tal d'incloure les regions més representatives en cada cas (Figura 2). El principal interès d'aquest pas era poder descartar possibles "falsos positius", és a dir, pics que semblessin tenir correlació en el heatmap, però que no representessin cap grup funcional típic del colesterol ni dels triglicèrids.

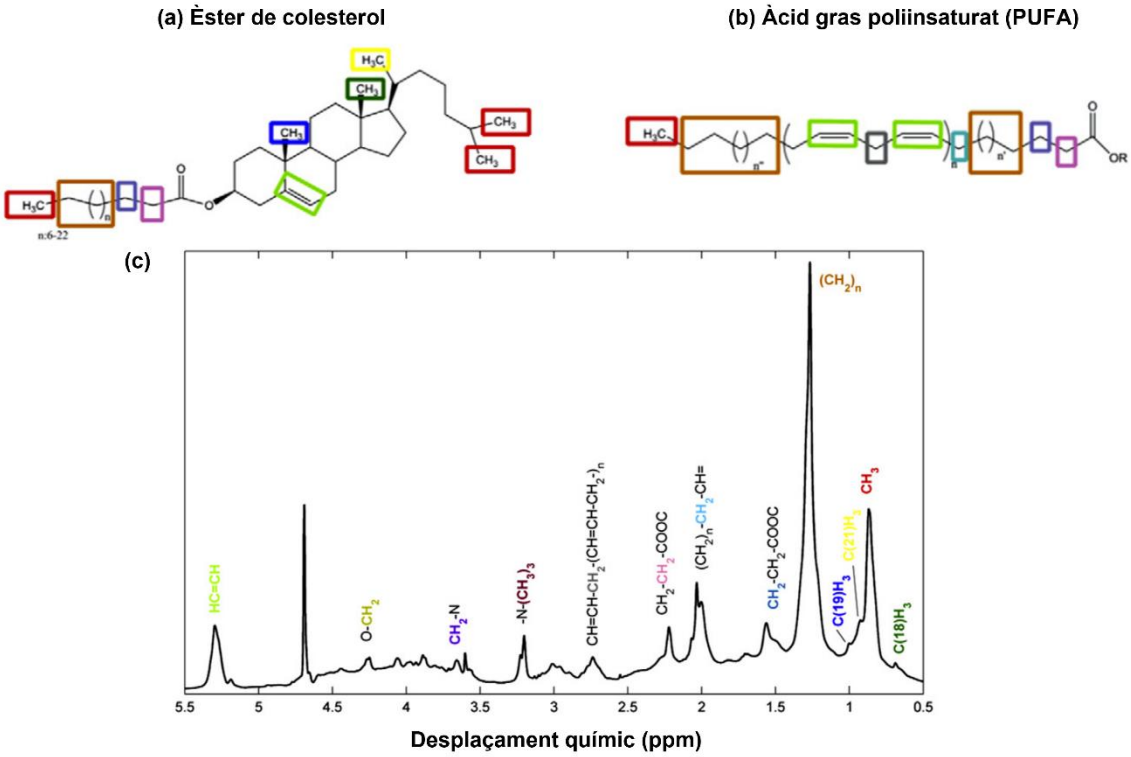

Figura 3. (a) Estructura general dels èsters de colesterol, amb els grups funcionals més representatius indicats en diferents colors. (b) Estructura general dels àcids grassos poliinsaturats, que són els que presenten la major quantitat de grups funcionals diferents. Grups funcionals marcats també en diferents colors. (c) Espectre de RMN indicant els pics on ressonen els grups funcionals dels lípids (a) i (b). Font: (Mallol et al., 2013b)

3.3. Mètodes de separació de les mostres (model i validació)

A l'hora de fer models PLS, cal dividir les dades en dos grups: el primer s'utilitza per a entrenar el model, mentre que el segon s'utilitza per a validar-lo i confirmar que el seu funcionament es correcte. Per fer això, existeixen diferents proporcions en les quals es poden dividir les dades entre els grups del model i els grups de control. Encara que el més habitual és dividir les dades de forma que un 50% constitueixin el model i l'altre 50% el grup control per a la validació, es va voler comprovar les diferències usant diferents proporcions. Així, es van fer proves amb divisions 50%-50%, però també amb 60%-40% i amb 70%-30% per al model i per la validació, respectivament.

D'altra banda, també es van tenir en compte les diferències entre els grups que conformen la base de dades utilitzada. Donat que cada grup aporta dades en un rang diferent, dividir les dades de forma aleatòria, sense tenir en compte els grups, podia causar una manca de dades en un cert rang en un dels dos grups – model i validació. És per això que es van fer diferents subdivisions de les dades, unes sense tenir en compte els grups i les altres tenint-los en compte (Taula 7).

Taula 3. Divisió de les dades en els diferents datasets utilitzats per a les proves dels models de colesterol i triglicèrids.

DATASETS PER LA CREACIÓ DE MODELS

DATASET		Nº DADES MODEL				Nº DADES VALIDACIÓ			
		Grup1	Grup2	Grup3	Total	Grup1	Grup2	Grup3	Total
COLESTEROL (N = 940)	DS01 (50 – 50)	-	-	-	470	-	-	-	470
	DS02 (60 – 40)	-	-	-	564	-	-	-	376
	DS03 (70 – 30)	-	-	-	658	-	-	-	282
	DS04 (50 – 50)	230	154	86	470	230	154	86	470
	DS05 (60 – 40)	276	185	103	564	184	123	69	376
	DS06 (70 – 30)	322	216	120	658	138	92	52	282
TRIGLICÈRIDS (N = 480)	DS07 (50 - 50)	-	-	-	240	-	-	-	240
	DS08 (60 – 40)	-	-	-	288	-	-	-	192
	DS09 (70 – 30)	-	-	-	336	-	-	-	144
	DS10 (50 – 50)	-	154	86	240	-	154	86	240
	DS11 (60 – 40)	-	185	103	288	-	123	69	192
	DS12 (70 – 30)	-	216	120	336	-	92	52	144

La construcció dels datasets es va fer de forma aleatòria, seleccionant a l'atzar la quantitat de dades requerida d'entre totes les que conformen la base de dades. En el cas dels datasets DS04, DS05, DS06, DS10, DS11 i DS12, com que es volia tenir en compte la distribució no uniforme de les dades en els diferents grups, es va fixar la proporció de dades de cada grup que s'havia d'incloure en el dataset.

3.4. Metodologia per a la generació dels models

3.4.1. El mètode PLS i l'efecte de les variables latents (VLs)

La regressió de mínims quadrats parcials és un mètode estadístic multivariant que analitza la relació entre variables per trobar un subespai de variables latents que sintetitza les variables de predicció o independents (X) amb l'objectiu d'entendre la dispersió de les variables dependents o observades (Y) de forma lineal. Aquesta tècnica descompon les variables d'estudi en una sèrie de components latents (o variables latents, VLs) que capturen la variabilitat conjunta més important en ambdós conjunts de variables, cosa que permet construir un model predictiu. En altres, paraules, les variables latents representen l'estructura subjacent de les dades i permeten modelar-ne la relació.

Així, la quantitat de VLs que s'utilitzen per a fer un model PLS ha de ser suficient per a poder capturar correctament la relació entre els dos grups de dades, però alhora s'ha de controlar la seva quantitat. Fer un model PLS amb excessives VLs podria comportar problemes d'inestabilitat – excessiva sensibilitat a les variacions en les dades d'entrenament – i de sobreajustament – el model podria capturar soroll o variabilitat aleatòria en comptes de la veritable estructura de les dades. Per aquest motiu és recomanable utilitzar mètodes de validació creuada, que permeten determinar quin és el nombre de VLs més adequat estadísticament per a un model en concret. Per a la generació dels models, per coherència amb els procediments previs de l'empresa Biosfer Teslab, es va utilitzar el mètode anomenat "Venetian Blinds".

3.4.2. Metodologia utilitzada per a l'elaboració i selecció dels models

Per cada dataset (Taula 3), es van dur a terme 100 proves amb diferents seleccions de dades, totes de forma aleatòria, per tal de determinar quina configuració era millor generalment. Per a cada prova, es va avaluar el resultat d'utilitzar dos mètodes de preprocessat diferents – autoscale i mean centering – i l'efecte d'incloure entre 4 i 8 variables latents. D'aquesta forma, per a cada dataset es van obtenir 1000 possibles models, resultants de les combinacions de totes les variables tingudes en compte. Per a determinar la qualitat dels models obtinguts, es va fer servir el coeficient de correlació de Pearson (r), segons la correlació existent entre la bioquímica del grup de validació i la proporcionada pel model.

A més, els diferents datasets es van avaluar estadísticament per tal de determinar quina combinació de variables – proporció en les mostres del model i la validació, tipus de divisió de les dades, variables latents i tipus de preprocessat – proporcionava els resultats més estables al llarg de totes les proves. A partir d'aquí, es va seleccionar el millor model dels que complien les condicions idònies i es va procedir a fer-ne un refinament manual, ajustant les zones de l'espectre seleccionades amb anterioritat. Això es va fer tant per al model de colesterol com per al de triglicèrids, de forma independent per a cadascun.

El tractament inicial de les dades i la generació dels models es van dur a terme mitjançant MATLAB (versió R2023a), juntament amb la PLS Toolbox d'Eigenvector (versió 9.2.1). L'anàlisi estadística dels resultats obtinguts, així com l'anàlisi prèvia dels grups de dades inicials, es van dur a terme amb RStudio (versió 2023.03.01).

3.4.3. Metodologia utilitzada per al càlcul de les concentracions parcials

Un cop seleccionats els models, per tal de poder-los incorporar en futures versions del test Liposcale® calia comprovar que es podien obtenir valors individuals de les concentracions de cada tipus de partícula – VLDL, IDL, LDL i HDL – sense perdre precisió. Per fer això, es van utilitzar els resultats dels models preexistents en el test per obtenir la proporció de colesterol i triglicèrids totals que representava cada tipus (Taula 4).

Taula 4. Exemple de la metodologia utilitzada per obtenir els valors individuals de la concentració de VLDL, IDL, LDL i HDL amb dades d'una mostra de colesterol. La primera fila mostra els valors obtinguts amb el test Liposcale®. La segona, el tant per 1 que representa cada tipus de partícula en la concentració total de colesterol. La tercera fila és el càlcul de les concentracions de cada tipus de partícula a partir de la concentració total obtinguda del model i els tants per 1 calculats en la segona fila

	Total	**VLDL**	**IDL**	**LDL**	**HDL**
Valors per Liposcale®	198,43 mg/dl	21,05 mg/dl	14,43 mg/dl	120,76 mg/dl	42,18 mg/dl
Tant per 1	1	0,11	0,07	0,61	0,21
Valors amb el model	203,87 mg/dl	22,42 mg/dl	14,27 mg/dl	124,36 mg/dl	42,81 mg/dl

Aquesta metodologia permet ajustar els resultats obtinguts amb els models existents, alleugerint els errors introduïts a causa del procés d'ultracentrifugació que es va utilitzar per separar les diferents fraccions. Donat que s'espera que el valor total aconseguit amb el model sigui més precís que el de Liposcale®, les concentracions parcials obtingudes a partir d'ell també haurien de ser-ho.

4. Resultats i Discussió

4.1. Anàlisi de les dades

En fer l'anàlisi de les dades, es va observar que el grup 1 (n = 460) contenia mostres de colesterol de pacients generalment sans, amb valors que entre els 100 i els 200 en la majoria de casos. El grup 2 estava format dades tant de colesterol (n = 323) com de triglicèrids (n = 312) de pacients sans, però amb valors lleugerament per sobre dels recomanats. El grup 3, igual que el grup 2, contenia dades de colesterol (n = 189) i triglicèrids (n = 172) de pacients amb colesterol alt i amb hipertrigliceridèmies (Taules 5 i 6).

Encara que inicialment es va plantejar excloure el grup 1 per la manca de dades de triglicèrids, es va considerar que eren mostres prou rellevants pel model de colesterol com per a mantenir-les. Aquesta decisió es va prendre pel fet que els altres dos grups gairebé no presentaven mostres en el rang baix (fins a 200 mg/dl) el que podia ser perjudicial per al model.

Així, finalment es van seleccionar tres grups per al model de colesterol i dos grups per al model de triglicèrids, ja que en aquest cas les mostres aportades sí que cobrien per complet el rang de valors d'interès, com veurem en més detall en l'anàlisi estadística.

Cal destacar que la divisió entre els diferents grups es basa en la procedència de les mostres, motiu pel qual un mateix grup pot contenir un nombre diferent de mostres de colesterol que de triglicèrids. Aquesta diferència es deu a la neteja de valors erronis en les fases inicials del tractament de dades explicada en apartats anteriors, on es van detectar mostres que només tenien mesures bioquímiques d'un dels dos paràmetres.

Taula 5. Estadística descriptiva de les mostres de colesterol dels diferents grups. Dades en mg/dl.

Colesterol	n	Mitjana	Mediana	Moda
G1	460	143	140	125
G2	308	217	210	163
G3	172	244	242	267
TOTAL	940	223	218	163

Com s'ha comentat amb anterioritat, s'observa que el grup 1 aporta mostres en un rang completament diferent que els altres dos grups, per la qual cosa val la pena incloure'l encara que no presenti valors de triglicèrids (Figura 4 i 5).

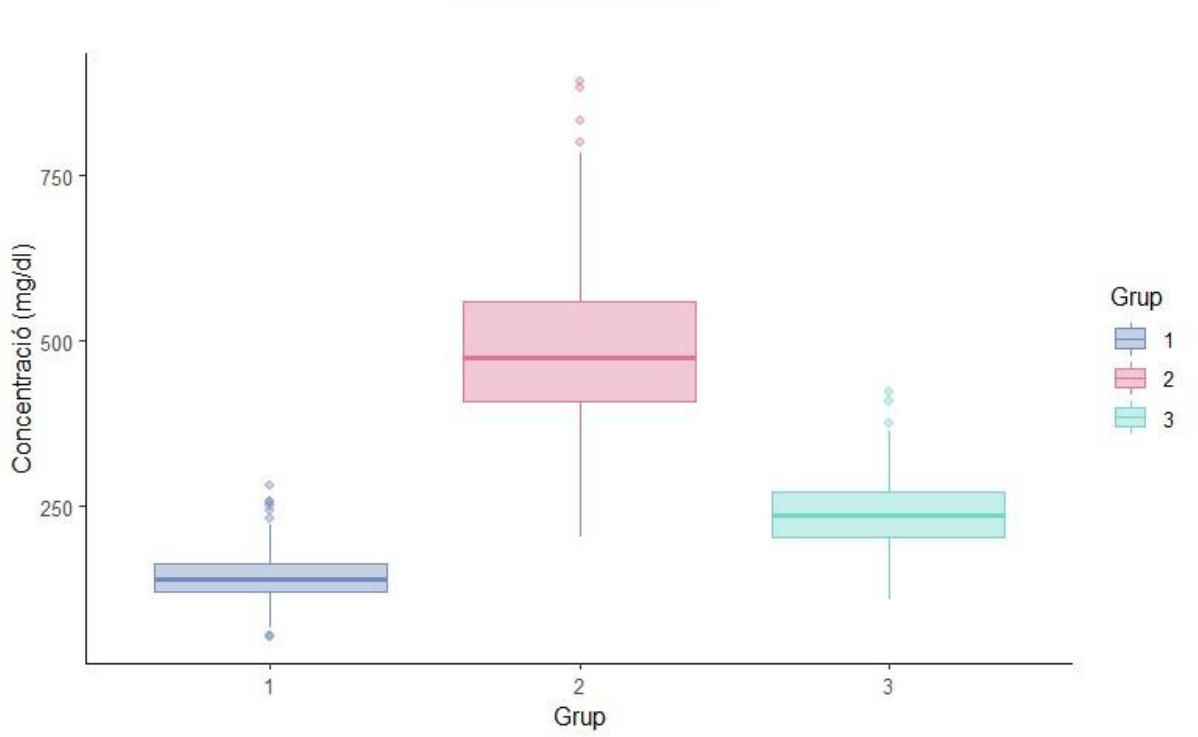

Figura 4. Boxplot de les mostres de colesterol dels tres grups. S'observa com entre els tres grups s'aconsegueix cobrir adequadament tot el rang de valors d'interès.

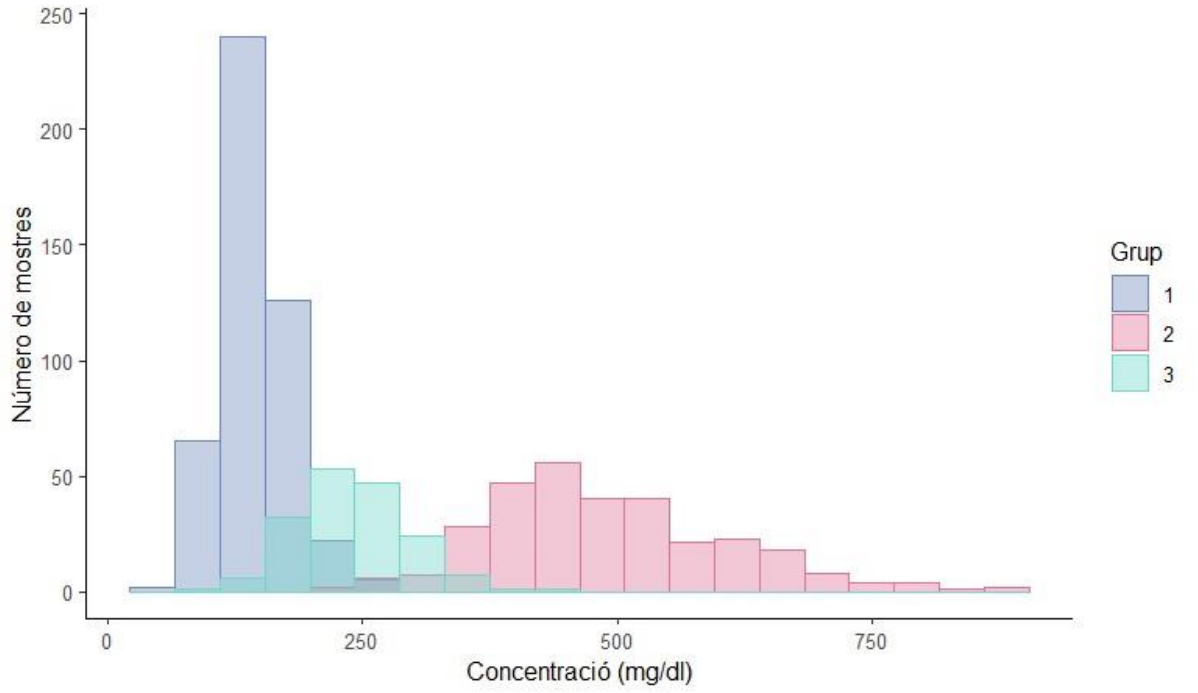

Figura 5. Boxplot de les mostres de colesterol dels tres grups. S'observa com entre els tres grups s'aconsegueix cobrir adequadament tot el rang de valors d'interès.

Taula 6. Estadística descriptiva de les mostres de triglicèrids dels diferents grups. Dades en mg/dl.

Triglicèrids	n	Mitjana	Mediana	Moda
G2	308	175	142	50
G3	172	322	283	201
TOTAL	480	227	210	50

En el cas de les mostres de triglicèrids, els grups 2 i 3 contenien prou mostres com per a cobrir tot el rang de valors d'interès, des de valors molt baixos fins a valors de fins a 700 mg/dl. En mostres d'hipertrigliceridèmies molt severes, per sobre dels 700 mg/dl, les mostres es dilueixen abans de fer-ne l'anàlisi per RMN per tal de poder garantir que la mesura obtinguda és correcte. És per aquest motiu que les mostres utilitzades no superen els 700 mg/dl (Figura 6 i 7), ja que el model no haurà de treballar amb mostres amb valors tan elevats, per la qual cosa ens podem centrar en valors més baixos per tal d'aconseguir més precisió.

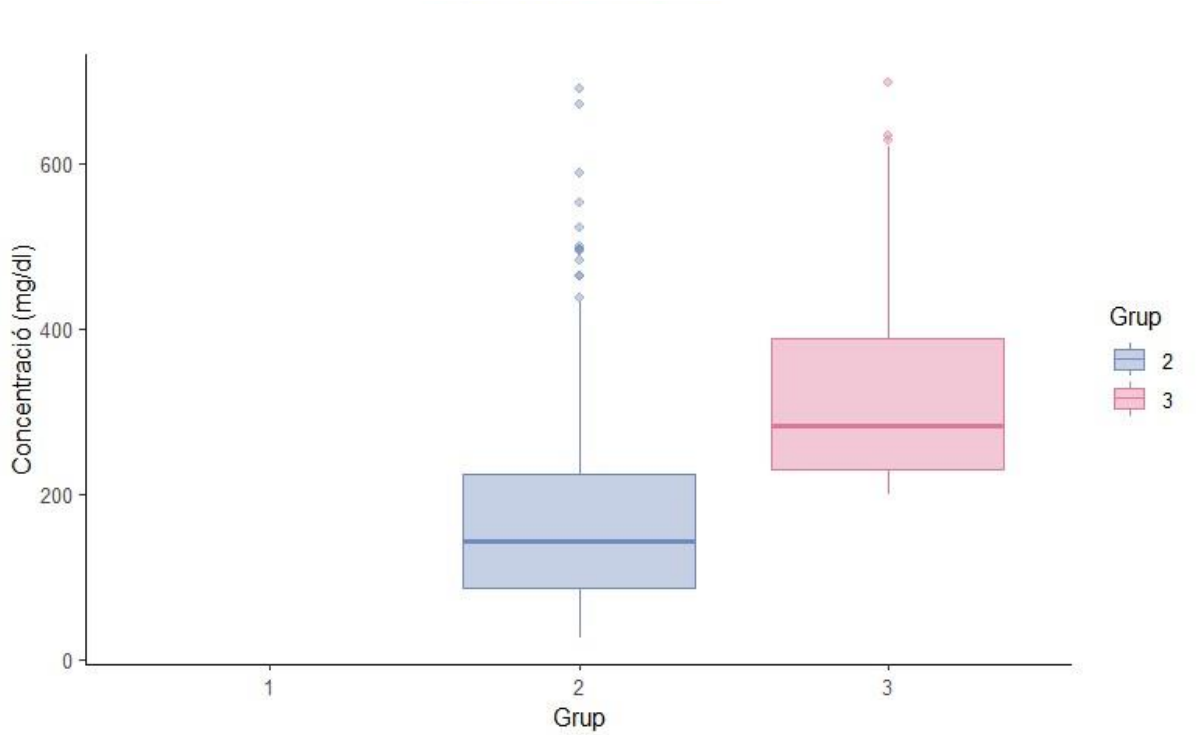

Figura 6. Boxplot de les mostres de triglicèrids dels dos grups. S'observa com entre els dos grups s'aconsegueix cobrir adequadament tot el rang de valors d'interès.

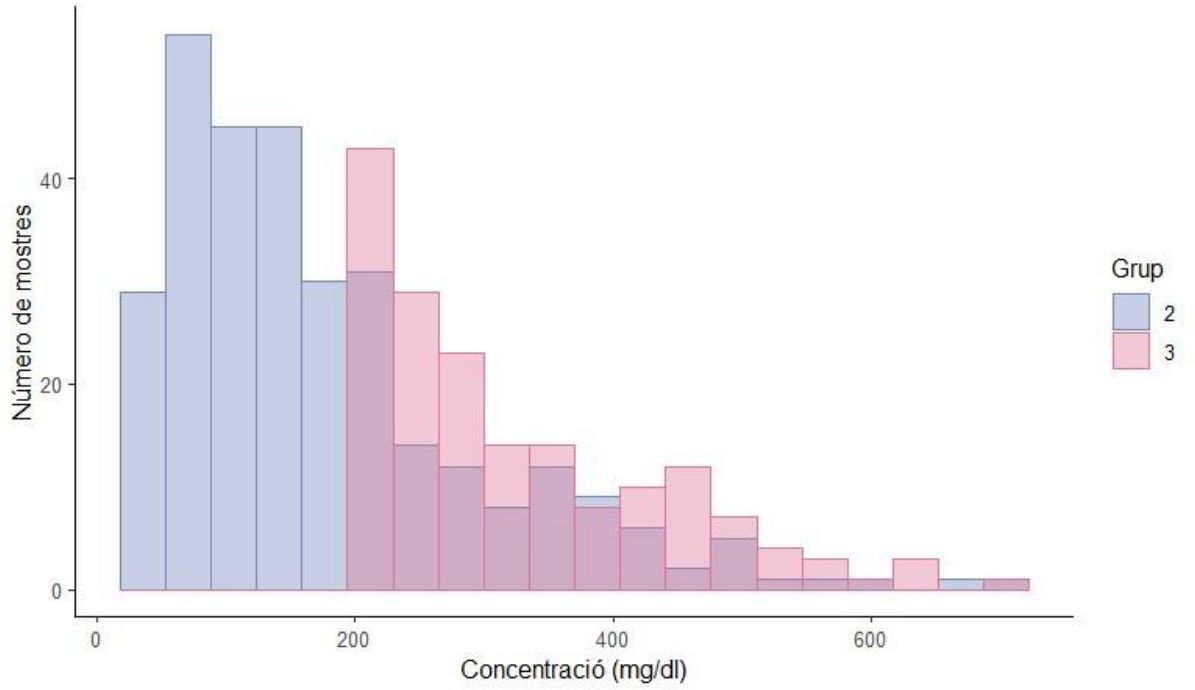

Figura 7. Histograma de les mostres de triglicèrids dels dos grups. S'observa com entre els dos grups s'aconsegueix cobrir adequadament tot el rang de valors d'interès.

4.2. Resultats del STOCSY

Els resultats obtinguts dels dos STOCSY van mostrar que les zones amb més correlació, aquelles marcades en vermell en els heatmaps, eren molt semblants en ambdós casos, a causa de la similitud dels dos lípids a avaluar. Tot i això, s'observen diferències subtils en les regions dels pics que tenen més correlació: el colesterol correla més en la zona dreta dels pics, mentre que els triglicèrids ho fan més en la part esquerra (Figura 8).

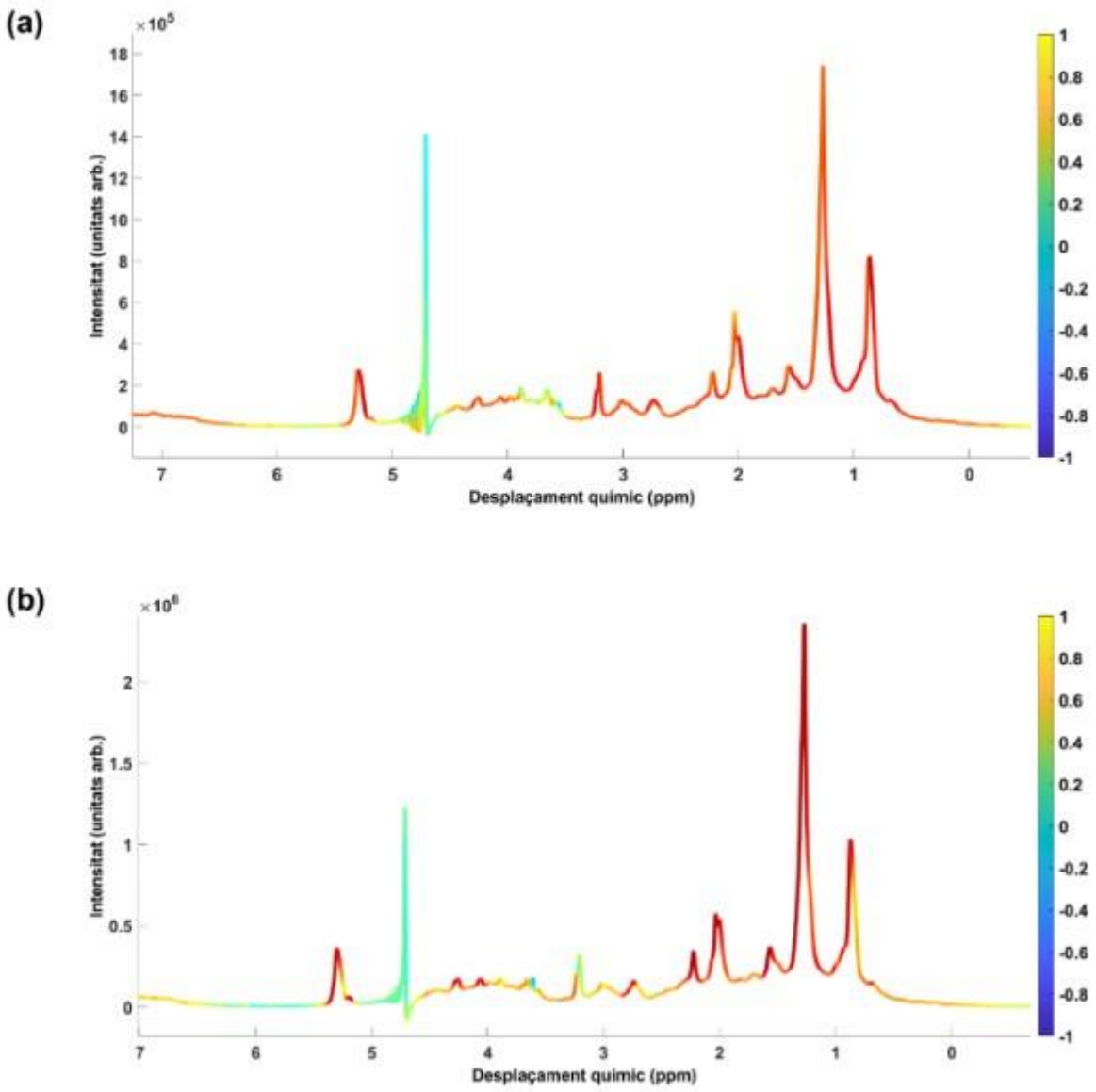

Figura 8. Heatmaps obtinguts del STOCSY per al colesterol (a) i els triglicèrids (b). S'observen les zones de l'espectre que correlen més amb la quantificació dels dos paràmetres (zones en vermell fosc) i les que correlen menys (zones en groc i verd).

Això es deu al fet que les diferents lipoproteïnes ressonen en diferents regions de l'espectre, segons la seva mida i la seva composició. Com hem vist, les partícules més riques en triglicèrids són les VLDL, que degut a les seves propietats ressonen en la zona esquerra dels pics. D'altra banda, el colesterol és més comú en partícules com les HDL i les LDL, que ressonen més a la dreta (Figura 7). Aquesta diferència va permetre seleccionar regions lleugerament diferents a l'hora de fer els models, per tal d'obtenir més precisió en la quantificació.

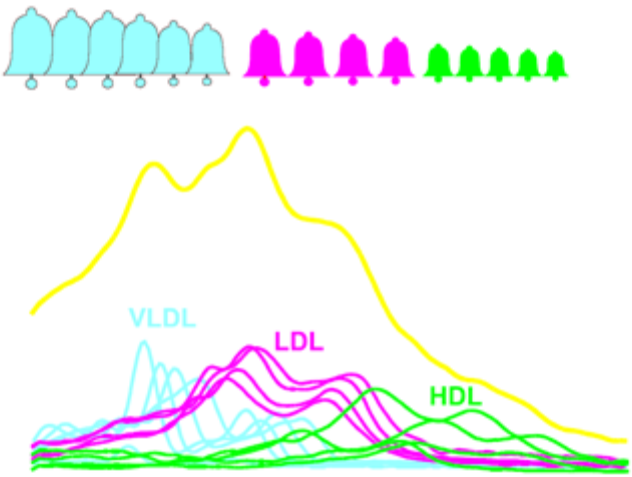

Figura 9. Representació de les funcions utilitzades per la quantificació de diferents lipoproteïnes en un estudi del 2006 publicat a la revista "Clinics in laboratory medicine". Es veu com les VLDL (en blau) ressonen més a l'esquerra, les HDL (en verd) més a la dreta, i les LDL (en lila) ressonen al centre. Les il·lustracions de campanes fan referència a les diferents mides de partícules avaluades, 6 per a les VLDL, 4 per les LDL i 5 per les HDL (Jeyarajah et al., 2006).

A partir dels heatmaps obtinguts dels STOCSYs, es van seleccionar els pics d'interès, tant per al colesterol com per als triglicèrids. En cada cas, les regions seleccionades van ser les utilitzades com a punt de partida per a dur a terme els models. Per al colesterol es van seleccionar 8 regions (Taula 7), mentre que per als triglicèrids es van seleccionar 9 regions (Taula 8) al llarg de tot l'espectre, que eren les que incloïen les zones que havíem mostrat més correlació.

Taula 7. Regions inicials seleccionades per al model de colesterol. Es mostren les 8 regions escollides inicialment a partir dels resultats observats en el heatmap del STOCSY.

Regió	Inici	Final
Reg1	5,309	5,219
Reg2	3,282	3,054
Reg3	2,812	2,675
Reg4	2,217	2,078
Reg5	2,011	1,750
Reg6	1,558	1,382
Reg7	1,247	0,901
Reg8	0,866	0,602

Taula 8. Regions inicials seleccionades per al model de triglicèrids. Es mostren les 9 regions escollides inicialment a partir dels resultats observats en el heatmap del STOCSY.

Regió	Inici	Final
Reg1	5,396	5,268
Reg2	5,221	5,161
Reg3	4,295	4,195
Reg4	4,120	4,001
Reg5	2,312	2,141
Reg6	2,014	1,939
Reg7	1,645	1,495
Reg8	1,424	0,853
Reg9	0,918	0,853

Es pot observar que els pics seleccionats en base al heatmap del STOCSY concorden amb els grups funcionals presents en el colesterol i els diferents tipus de triglicèrids (recordar Figura 3) . A més, tenir en compte aquesta informació va permetre descartar pics que en el STOCSY semblava que tenien bona correlació, però que no coincidien amb cap grup funcional d'interès. Va ser el cas del pic del grup O-CH_3, entre 4,2 i 4,3 ppm, que genera un pic que semblava rellevant segons el STOCSY, però que no és del nostre interès perquè no està present en els lípids que volem avaluar.

4.3. Models finals

4.3.1. Resultats per al model de colesterol

Un cop realitzades les proves dels 6 datasets de les mostres de colesterol es van analitzar els resultats obtinguts per a cadascun dels 1000 models generats en cada dataset. El primer que es va avaluar va ser l'impacte de les diferents condicions d'estudi:

(a) **Tipus de divisió**: es van agrupar els datasets segons si s'havien tingut en compte els grups o no (DS01, DS02 i DS03 contra DS04, DS05 i DS06) (Figura 10).
(b) **Percentatge de divisió**: es van agrupar els datasets segons el percentatge de divisió de les dades entre el model i la validació (DS01 amb DS04, DS02 amb DS05 i DS03 amb DS06) (Figura 11).
(c) **Tipus de preprocessat**: es van agrupar les dades dels diferents datasets segons el preprocessat que se'ls havia aplicat (Figura 12).
(d) **Nombre de VLs**: es van agrupar les dades segons el nombre de VLs que tenia cadascun dels models (Figura 13).

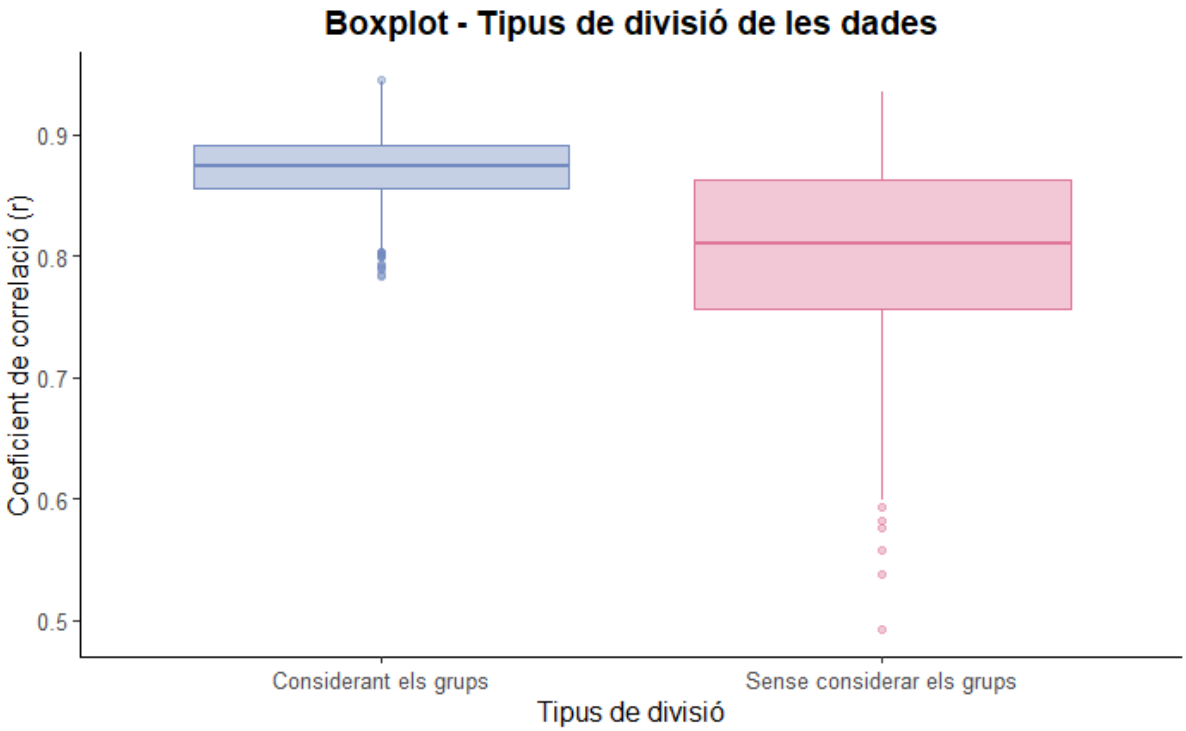

Figura 10. Boxplot dels resultats obtinguts amb les proves dels models de colesterol, segons el tipus de divisió de les dades utilitzat (a).

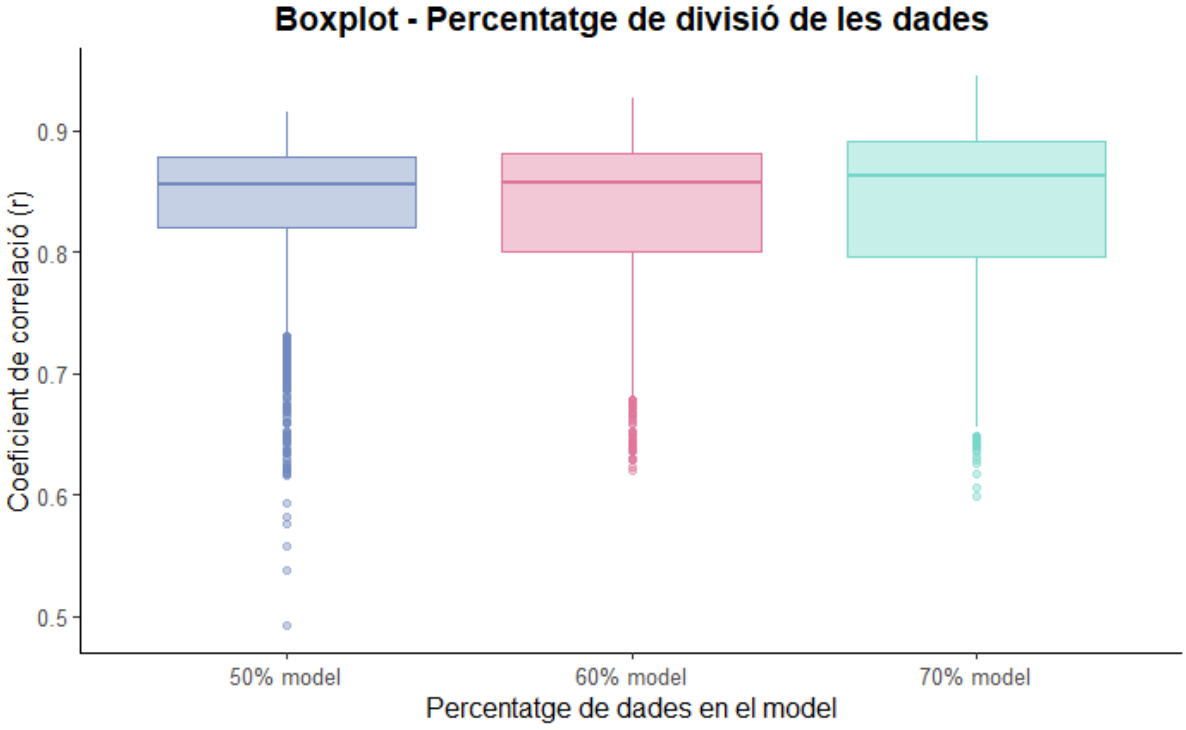

Figura 11. Boxplot dels resultats obtinguts amb les proves dels models de colesterol, segons el percentatge de divisió de les dades utilitzat (b).

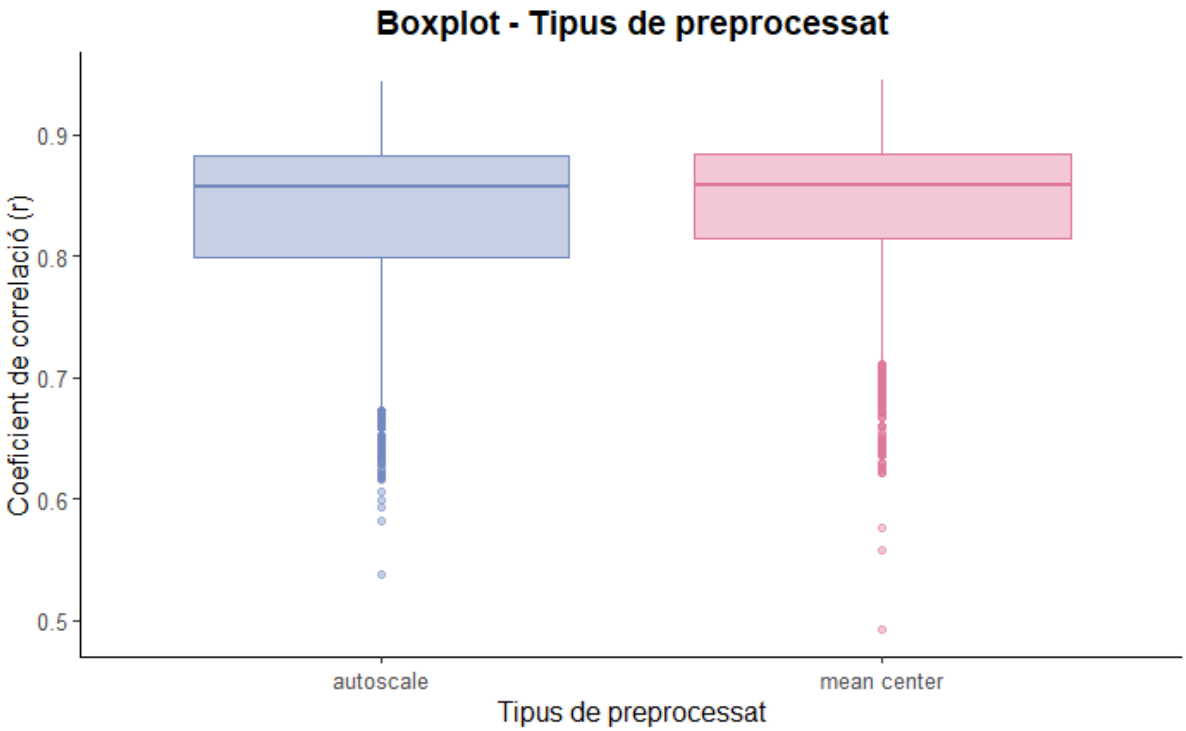

Figura 12. Boxplot dels resultats obtinguts amb les proves dels models de colesterol, segons el tipus de preprocessat de les dades utilitzat (c).

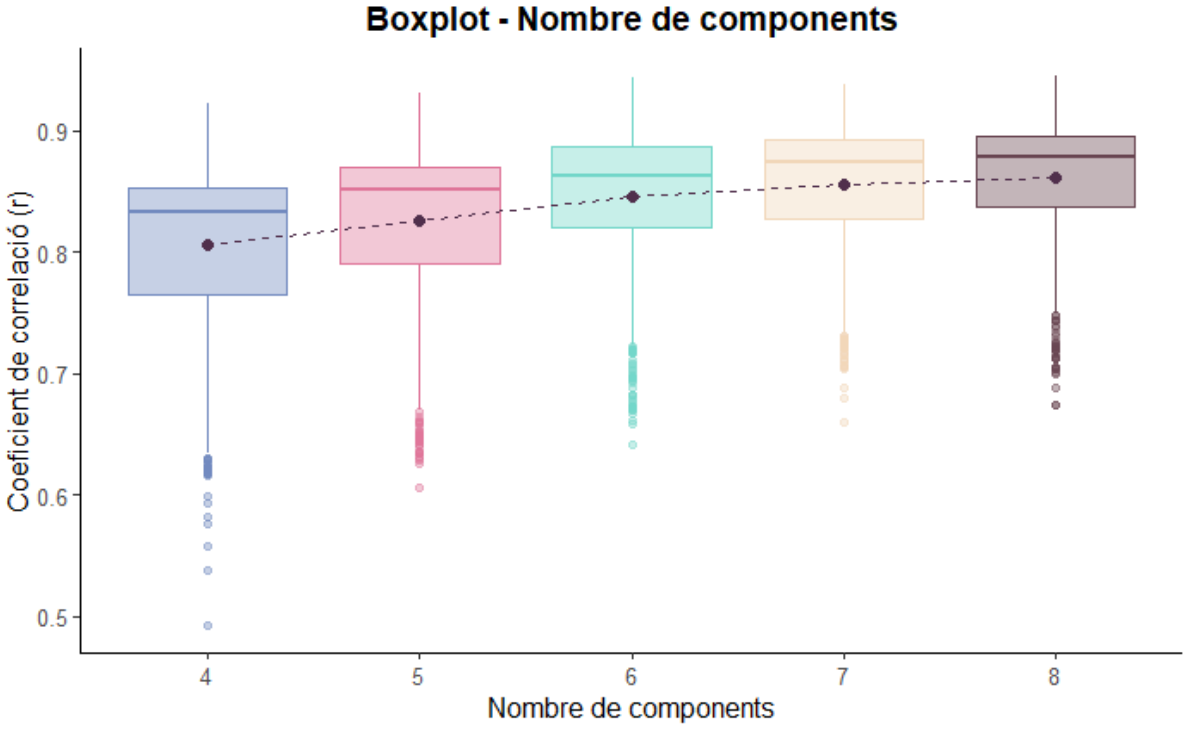

Figura 13. Boxplot dels resultats obtinguts amb les proves dels models de colesterol, segons el nombre de VLs (components) utilitzat (d)

A partir de les Figures 10, 11, 12 i 13 es va determinar que les úniques variables que tenien un impacte representatiu eren el tipus de divisió de les dades i el nombre de components tinguts en compte a l'hora de fer l'anàlisi PLS. Així, es va determinar que el model final havia de tenir en compte els diferents grups a l'hora de dividir les dades per fer els grups de model/control i que el nombre de VLs havia d'estar entre 6 i 8.

A banda de l'elaboració dels gràfics anteriors, es va fer una anàlisi ANOVA per tal de garantir que les diferents condicions d'estudi tenien un impacte significant en els resultats (Taula 9). Els resultats van demostrar que totes les condicions d'estudi eren significants, és a dir, que tenien un efecte rellevant sobre el valor del coeficient de Pearson.

*Taula 9. Anàlisi ANOVA de les dades obtingudes de les proves per als models de colesterol. Els * indiquen la significança dels diferents factors (un asterisc indica significança al nivell de 0,05, dos asteriscs indiquen significança al nivell de 0,01 i tres asteriscs indiquen significança al nivell de 0,001)*

FACTOR	GRAUS DE LLIBERTAT	SUMA DELS QUADRATS	MITJANA DELS QUADRATS	F-VALOR	P-VALOR
Tipus de Divisió	1	7,017	7,017	3011,588	< 2e-16***
Percentatge de Divisió	2	0,020	0,010	4,304	0,0136*
Preprocessat	1	0,0079	0,079	33,728	6,66e-9***
VLs	1	2,337	2,337	1002,957	< 2e-16***

De tots els models generats que complien aquests requisits, es van seleccionar els més prometedors, aquells que obtenien coeficients de correlació més elevats, per a cadascuna de les combinacions possibles dels paràmetres restants. Així, es van seleccionar 6 models, els millors per a cada combinació de tipus de preprocessat i tipus de divisió de les dades (Taula 10).

Taula 10. Models per al colesterol seleccionats inicialment. Tots els models seleccionats provenen dels datasets 4 a 6, que són els que generaven els grups de dades del model i la validació tenint en compte la proporció entre els diferents grups.

ID MODEL	PERCENTATGE DE DIVISIÓ	TIPUS DE PREPROCESSAT	NOMBRE DE COMPONENTS	VALOR R
COL01	50% model 50% validació	Autoscale	6	0,9158
COL02	50% model 50% validació	Mean Center	7	0,9146
COL03	60% model 40% validació	Autoscale	6	0,9200
COL04	60% model 40% validació	Mean Center	8	0,9228
COL05	70% model 30% validació	Autoscale	6	0,9357
COL06	70% model 30% validació	Mean Center	7	0,9383

Els 6 models seleccionats es van treballar de forma manual, cadascun per separat, per intentar millorar-ne els resultats. Les principals modificacions que es van fer van consistir en ajustar lleugerament les regions de l'espectre seleccionades amb l'eina STOCSY, per tal d'intentar ajustar encara més les regions d'interès i poder obtenir els millors resultats possibles.

Un cop fetes totes aquestes proves finals, el model que va donar millors resultats va ser el COL05, del qual es va aconseguir millorar el coeficient de correlació fins a 0,94017 (Taula 11). Els resultats d'aquest model es van comparar amb els resultats aconseguits amb la metodologia del test Liposcale® i es va comprovar que la predicció del colesterol total era millor amb aquest nou model que amb el test original (Figura 14).

Taula 11. Configuració del model final de colesterol, desenvolupat a partir del model COL05.

MODEL	PERCENTATGE DE DIVISIÓ	TIPUS DE PREPROCESSAT	NOMBRE DE COMPONENTS	VALOR R
COLESTEROL	70% model 30% validació	Autoscale	5	0,94017

D'altra banda, es van calcular les concentracions individuals del colesterol VLDL, IDL, LDL i HDL, utilitzant els resultats del test Liposcale® per extreure'n la proporció del total en cadascun dels casos. Les concentracions obtingudes es van comparar amb les obtingudes pel test Liposcale®, per tal d'avaluar les diferències obtingudes amb cadascuna de les dues metodologies. Es va observar que, mentre que els resultats de VLDL, IDL i LDL eren molt similars, existia certa diferència en els resultats per al colesterol HDL (Figura 15). Degut a la

manca de referències en aquests valors, ja que no disposàvem de les mesures per bioquímica d'aquests paràmetres, no es va poder determinar si aquesta diferència en el colesterol HDL corresponia també a una diferència amb les quantificacions per mètodes tradicionals.

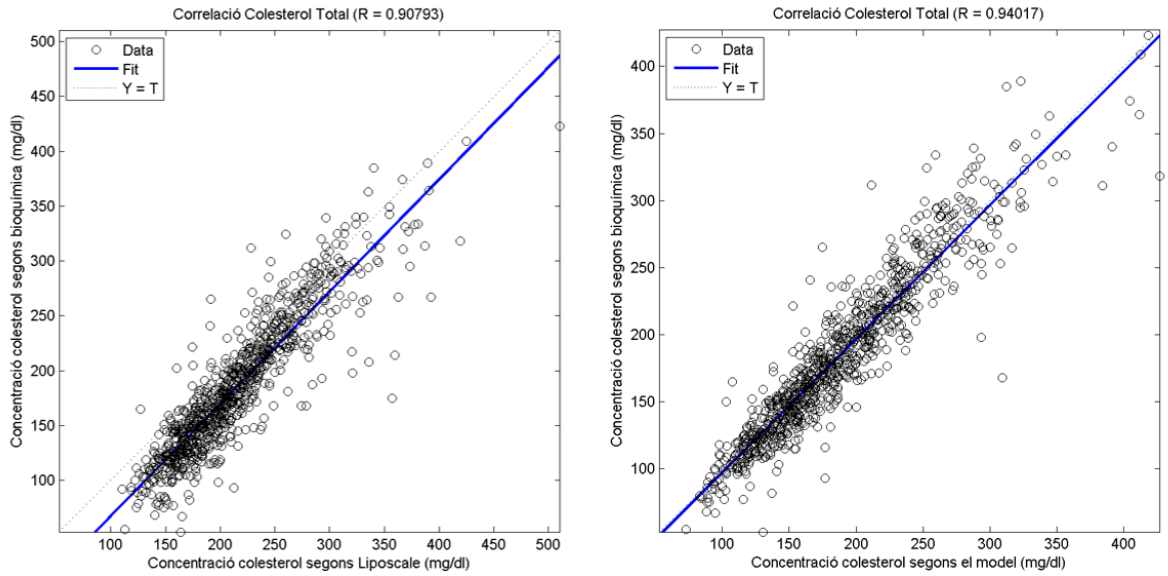

Figura 14. A la dreta, correlació de la predicció del model de colesterol amb els valors de bioquímica. A l'esquerra, correlació de la predicció del test Liposcale® amb la bioquímica. Les dades utilitzades són les mateixes en ambdós casos.

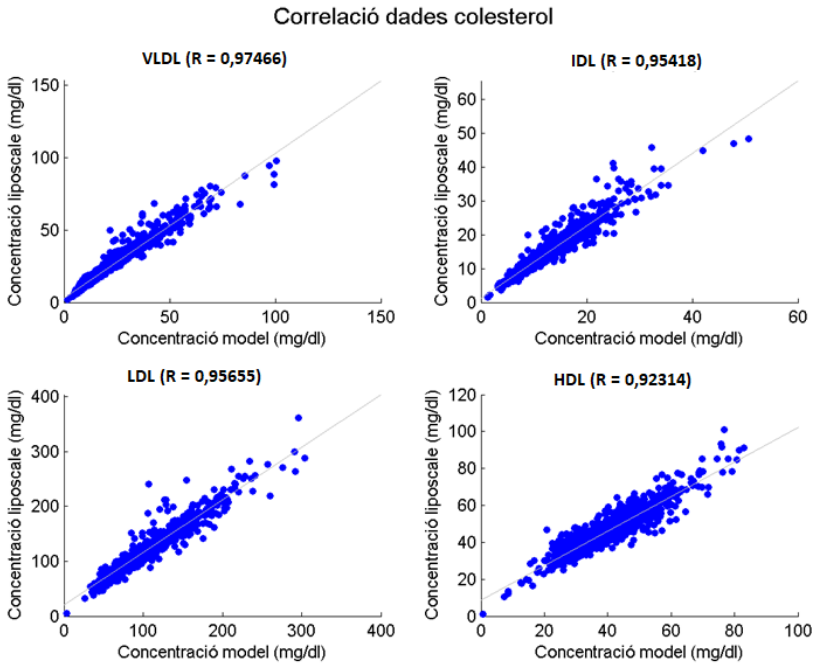

Figura 15. Correlació entre els resultats obtinguts amb el nou model i els valors obtinguts amb el test Liposcale®

4.3.2. Resultats per al model de triglicèrids

El procediment va ser el mateix que per al model de colesterol. En aquest cas, en canvi, els gràfics que en el cas del colesterol havien estat prou per determinar quins eren els paràmetres de més interès, no donaven gairebé informació (Figura 16, 17, 18 i 19). L'anàlisi ANOVA, però, va confirmar que tres de les quatre condicions d'estudi eren significatives (Taula 12).

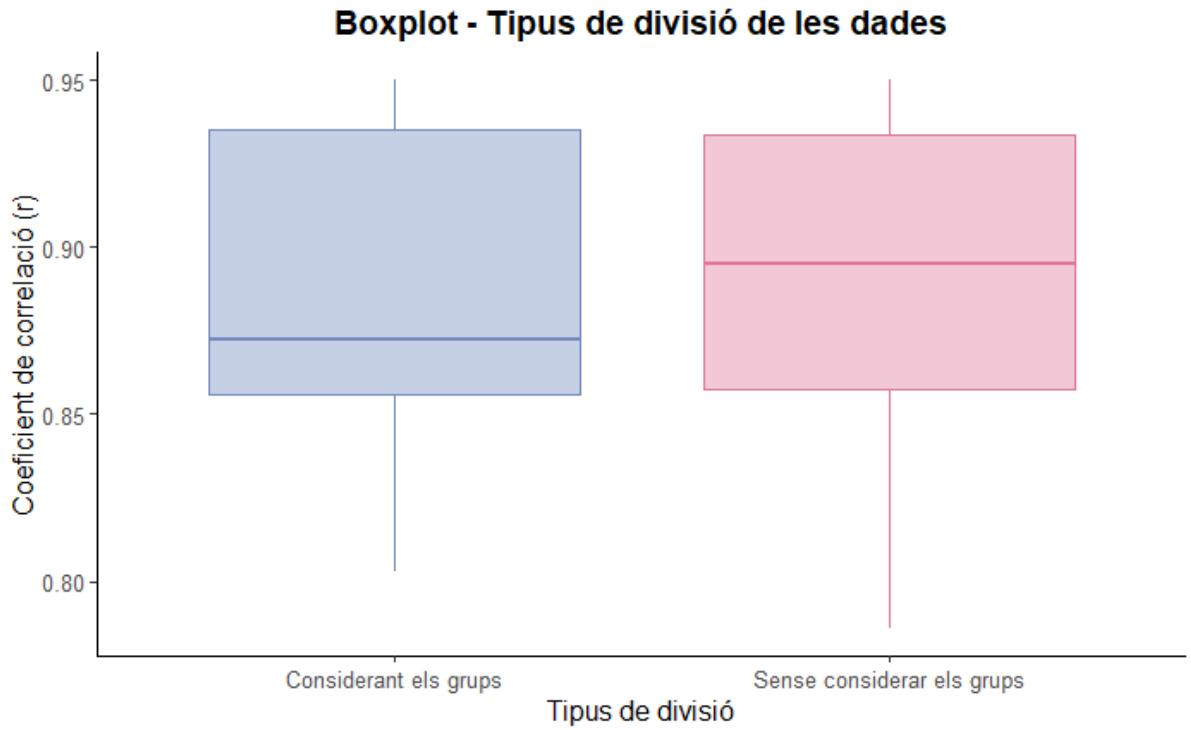

Figura 16. Boxplot dels resultats obtinguts amb les proves dels models de triglicèrids, segons el tipus de divisió de les dades utilitzat (a).

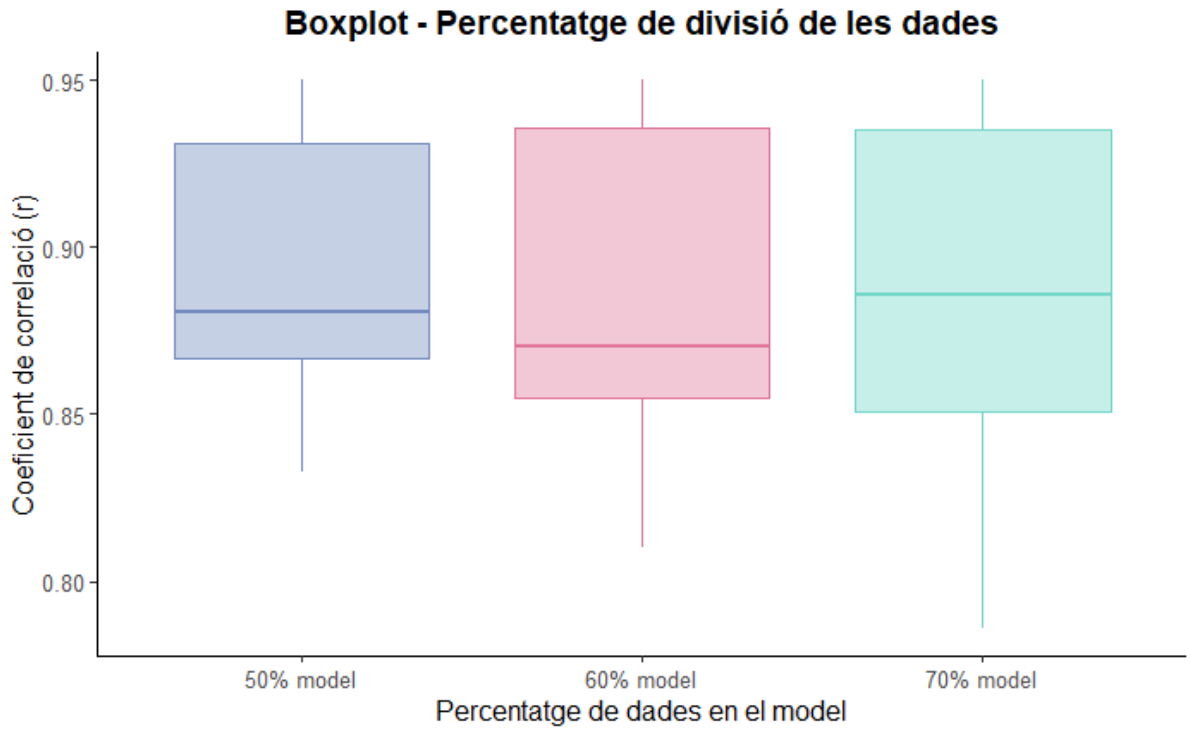

Figura 17. Boxplot dels resultats obtinguts amb les proves dels models de triglicèrids, segons el percentatge de divisió de les dades utilitzat (b).

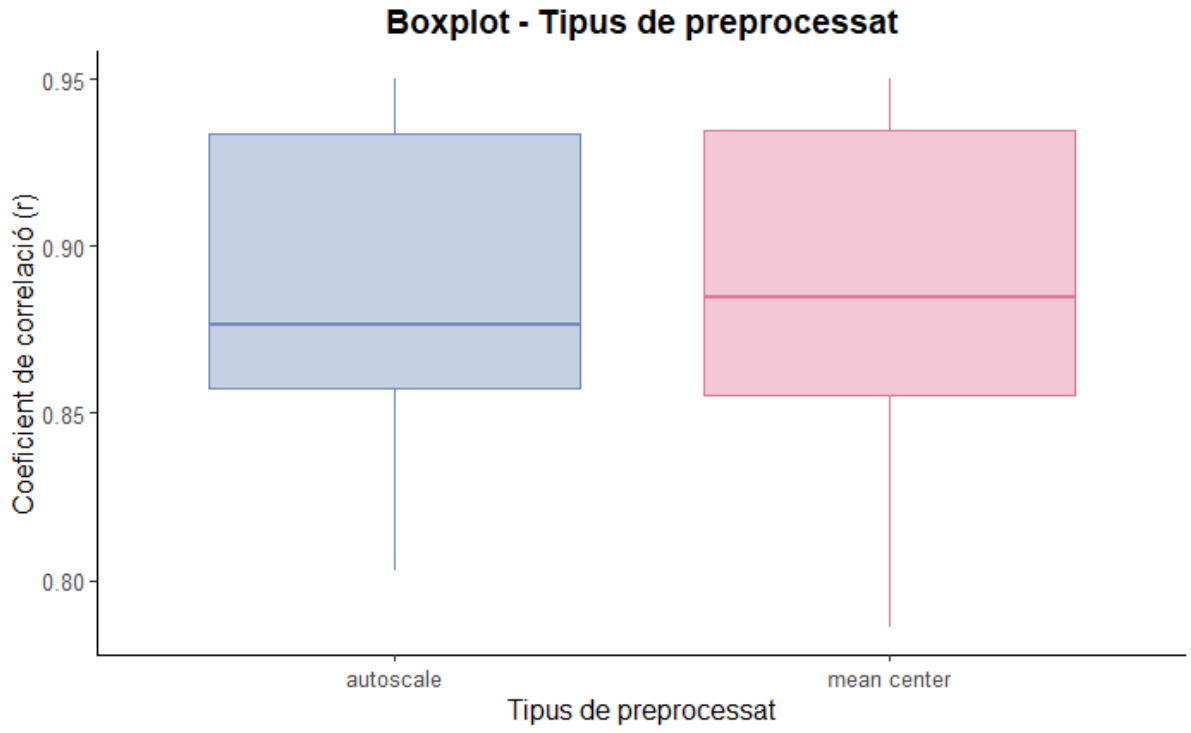

Figura 18. Boxplot dels resultats obtinguts amb les proves dels models de triglicèrids, segons el tipus de preprocessat de les dades utilitzat (c).

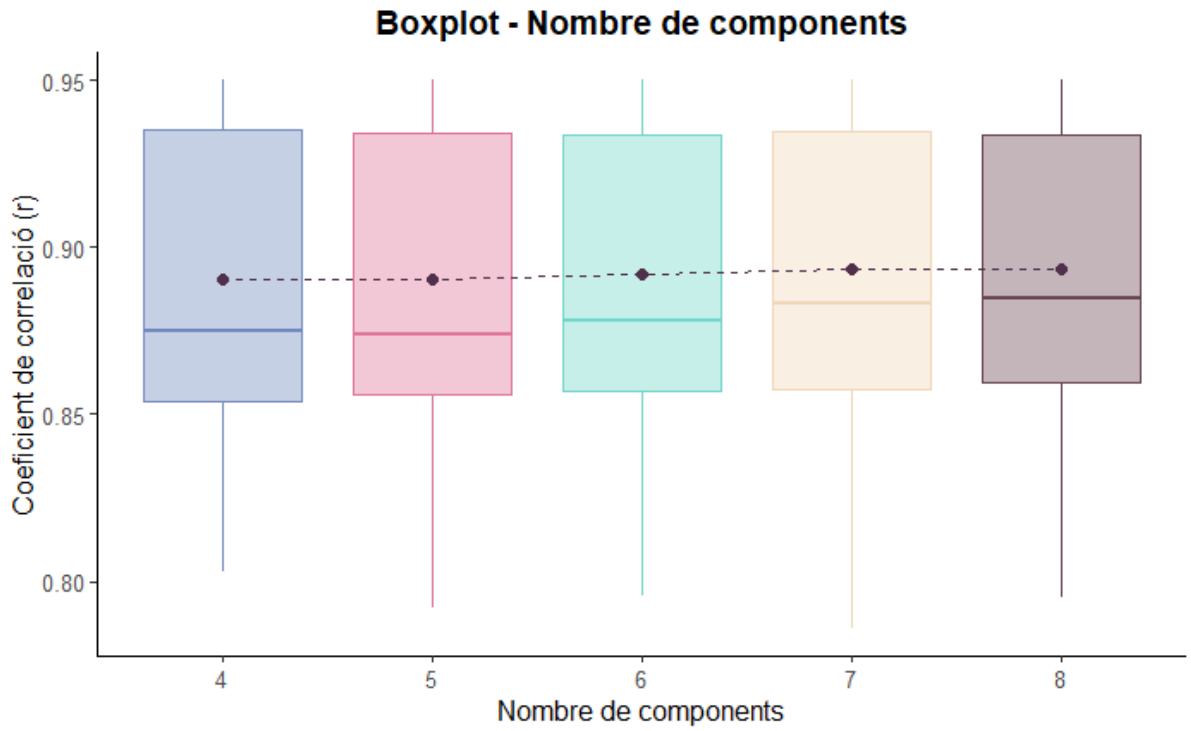

Figura 19. Boxplot dels resultats obtinguts amb les proves dels models de triglicèrids, segons nombre de components (VLs) utilitzat (d).

*Taula 12. Anàlisi ANOVA de les dades obtingudes de les proves per als models de triglicèrids. Els * indiquen la significança dels diferents factors (un asterisc indica significança al nivell de 0,05, dos asteriscs indiquen significança al nivell de 0,01 i tres asteriscs indiquen significança al nivell de 0,001)*

FACTOR	GRAUS DE LLIBERTAT	SUMA DELS QUADRATS	MITJANA DELS QUADRATS	F-VALOR	P-VALOR
Tipus de Divisió	1	0,0296535	0,0296535	18,520014	0.0000171***
Percentatge de Divisió	2	0,0205065	0,0102532	6,403632	0,0016669**
Preprocessat	1	0,0027947	0,0027947	1,745436	0,1865019
VLs	1	0,0099798	0,0099798	6,232863	0,0125667*

Amb els resultats del test ANOVA, un cop s'havia comprovat que els paràmetres avaluats sí que afectaven el coeficient de correlació – a excepció del tipus de preprocessat, que en aquest cas no era prou determinant en la qualitat del model – es van seleccionar 6 models amb diferents combinacions dels paràmetres, pel mateix procediment que en el cas dels models de colesterol (Taula 13).

Taula 13. Models per als triglicèrids seleccionats inicialment. Es van seleccionar els millors models de cadascuna de les combinacions de tipus de divisió i percentatge de divisió.

ID MODEL	TIPUS DE DIVISIÓ	PERCENTATGE DE DIVISIÓ	TIPUS DE PREPROCESSAT	NOMBRE DE COMPONENTS	VALOR R
TRI01	No proporcional	50% model 50% validació	Autoscale	5	0,94996
TRI02	No proporcional	60% model 40% validació	Autoscale	4	0,94981
TRI03	No proporcional	70% model 30% validació	Autoscale	4	0,94991
TRI04	Proporcional	50% model 50% validació	Autoscale	8	0,94998
TRI05	Proporcional	60% model 40% validació	Autoscale	6	0,94990
TRI06	Proporcional	70% model 30% validació	Autoscale	7	0,94988

Igual que en el cas dels models de colesterol, aquests sis models es van refinar manualment per tal d'assolir la precisió més gran possible en la predicció. El model que va donar millors resultats va ser el TRI06, que es va aconseguir refinar fins arribar a un valor de correlació de 0,9539 (Taula 14). En comparar aquest resultat amb la correlació assolida pel test Liposcale® es va veure que, encara que el valor de r assolit era molt similar, el nou model era més precís en valors de triglicèrids alts, ja que obtenia un pendent més proper a 1 (Figura 20).

Taula 14. Configuració del model final de triglicèrids, desenvolupat a partir del model TRI06.

MODEL	PERCENTATGE DE DIVISIÓ	TIPUS DE PREPROCESSAT	NOMBRE DE COMPONENTS	VALOR R
TRIGLICÈRIDS	70% model 30% validació	Autoscale	3	0,95394

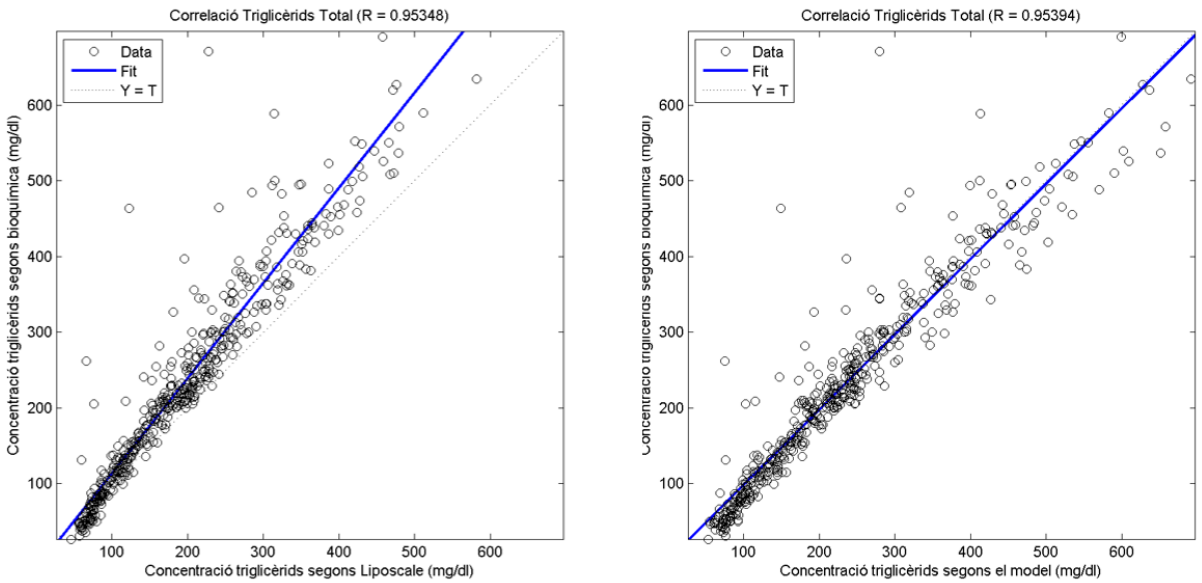

Figura 20. A la dreta, correlació de la predicció del model de triglicèrids amb els valors de bioquímica. A l'esquerra, correlació de la predicció del test Liposcale® amb la bioquímica. Les dades utilitzades són les mateixes en ambdós casos.

També es van calcular els valors de VLDL, IDL, LDL i HDL a partir del nou model, i es van comparar amb els que donava el test Liposcale® per a les mateixes dades. Els resultats van mostrar que hi havia un alt grau de correlació en tots els casos, per la qual cosa la pèrdua de precisió en aquest cas era mínima (Figura 21).

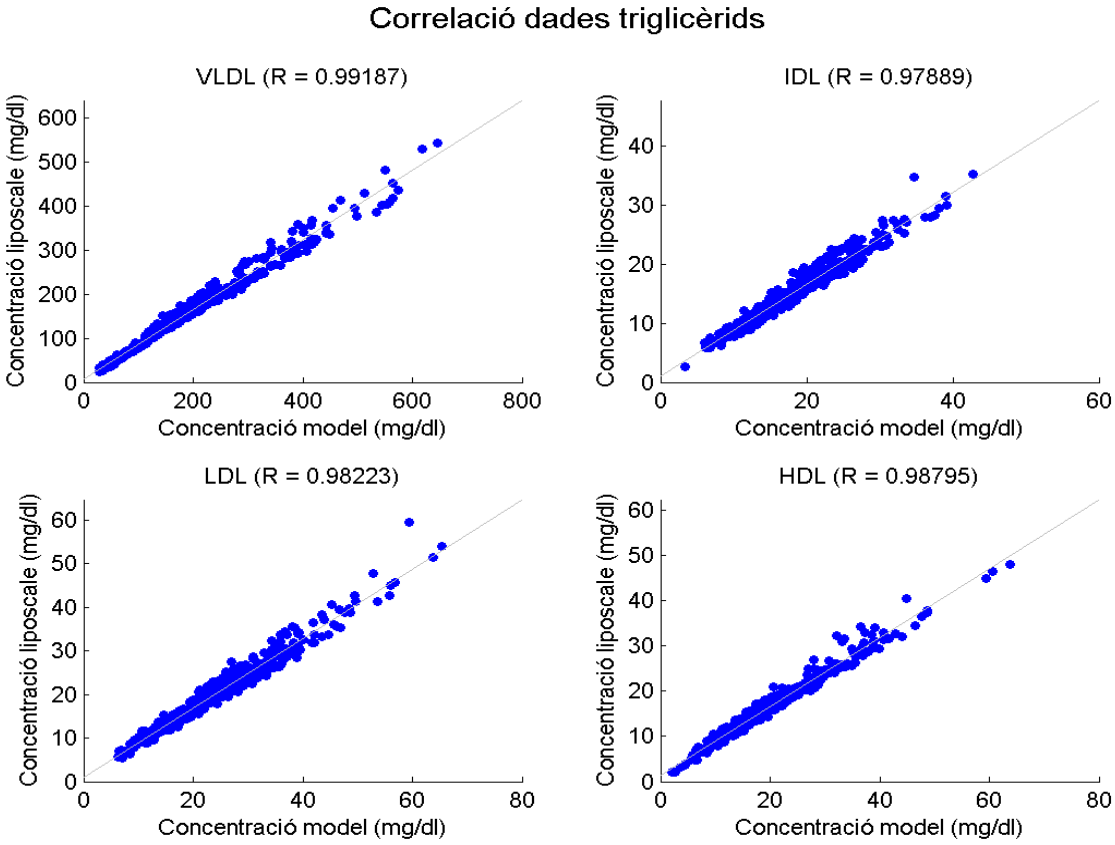

Figura 21. Correlació entre els resultats obtinguts amb el nou model i els valors obtinguts amb el test Liposcale®

4.4. Validació amb un conjunt extern

Finalment, es va utilitzar un conjunt de dades extern – això és, que no s'havien utilitzat per a la generació del model ni per a la seva validació – per tal de garantir la integritat dels dos models desenvolupats. Aquesta comprovació final va servir per posar a prova els models i determinar si aportaven resultats competitius fora de les dades usades per a la seva generació.

En els dos casos, es va observar que les correlacions obtingudes eren superiors a R = 0,9 (Figura 22 i 23). En el cas del model de colesterol, la correlació supera l'obtinguda amb les dades de validació del model – en aquest cas R = 0,96, mentre que amb les dades de validació obteníem R = 0,94. En el cas del model de triglicèrids, la correlació baixa de R = 0,96 a R = 0,90, a causa de dues mostres en les quals la predicció és més imprecisa.

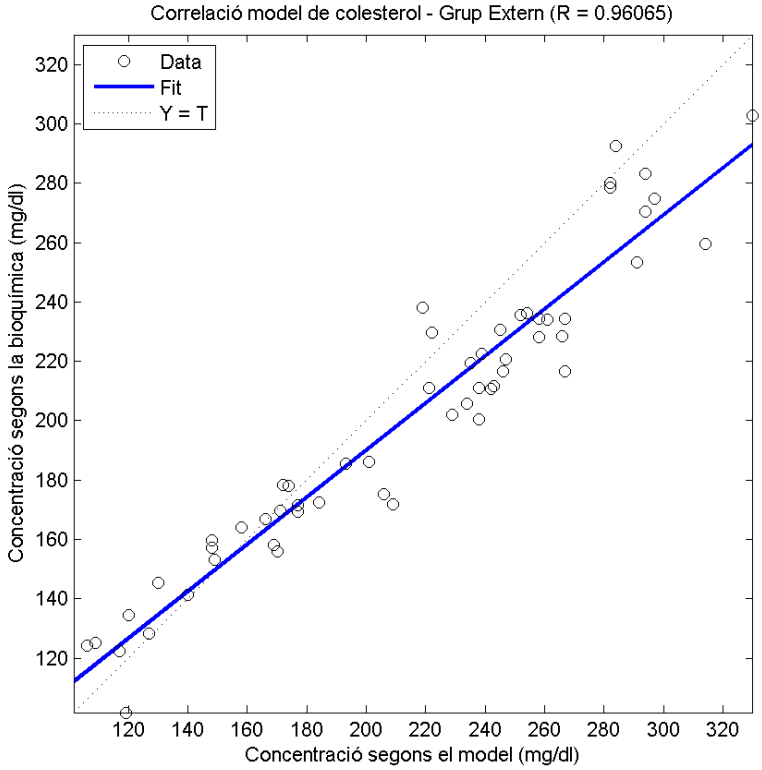

Figura 22. Correlació del model de colesterol amb les dades de bioquímica del grup de validació extern.

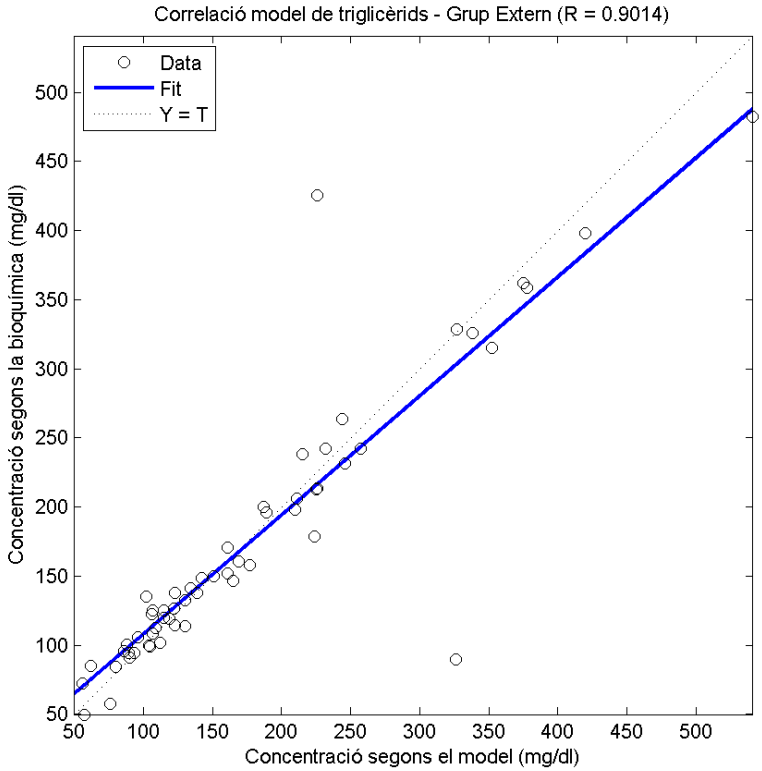

Figura 23. Correlació del model de triglicèrids amb les dades de bioquímica del grup de validació extern.

5. Conclusions

L'objectiu principal del treball consistia a desenvolupar i validar dos models de predicció, un per al colesterol i un altre per als triglicèrids, a partir d'espectres de RMN de mostres de plasma. Seguint la metodologia descrita anteriorment, s'ha aconseguit desenvolupar aquests dos models de forma satisfactòria, obtenint correlacions molt altes entre els valors de bioquímica i els valors predits pels models. A més a més, també s'ha comprovat que els resultats obtinguts amb aquests models són més precisos que els que s'obtenien amb el test Liposcale®.

Per anar més enllà, s'ha comprovat la correlació entre els valors individuals de colesterol i triglicèrids VLDL, IDL, LDL i HDL segons Liposcale® amb els calculats a partir dels nous models. En tots els casos, aquesta correlació ha estat molt elevada, i per això es pot afirmar que no hi ha pèrdua de precisió en aquests valors si s'utilitzen els models desenvolupats en aquest treball. És per això que en versions futures del test Liposcale® s'incorporaran aquests dos models per tal de millorar encara més els resultats que estava donant fins ara, amb l'objectiu d'apropar-se el més possible al "gold standard" actual, que són les anàlisis per bioquímica.

6. Bibliografia

Barrilero, R., Gil, M., Amigó, N., Dias, C. B., Wood, L. G., Garg, M. L., Ribalta, J., Heras, M., Vinaixa, M., & Correig, X. (2018). LipSpin: A New Bioinformatics Tool for Quantitative 1H NMR Lipid Profiling. *Analytical Chemistry*, *90*(3), 2031–2040. https://doi.org/10.1021/acs.analchem.7b04148

Barrilero, R., Llobet, E., Mallol, R., Brezmes, J., Masana, L., Zulet, M. Á., Martínez, J. A., Ribalta, J., Bulló, M., & Correig, X. (2015). Design and evaluation of standard lipid prediction models based on 1H-NMR spectroscopy of human serum/plasma samples. *Metabolomics*, *11*(5), 1394–1404. https://doi.org/10.1007/s11306-015-0796-5

Burnett, J. R., Hooper, A. J., & Hegele, R. A. (2020). Remnant Cholesterol and Atherosclerotic Cardiovascular Disease Risk. In *Journal of the American College of Cardiology* (Vol. 76, Issue 23, pp. 2736–2739). Elsevier Inc. https://doi.org/10.1016/j.jacc.2020.10.029

Cloarec, O., Dumas, M. E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J. C., Holmes, E., & Nicholson, J. (2005). Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Analytical Chemistry*, *77*(5), 1282–1289. https://doi.org/10.1021/ac048630x

Emwas, A. H. M. (2015). The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods in Molecular Biology*, *1277*, 161–193. https://doi.org/10.1007/978-1-4939-2377-9_13

Farnier, M., Zeller, M., Masson, D., & Cottin, Y. (2021). Triglycerides and risk of atherosclerotic cardiovascular disease: An update. *Archives of Cardiovascular Diseases*, *114*(2), 132–139. https://doi.org/10.1016/j.acvd.2020.11.006

Feingold, K. R., Anawalt, B., Blackman, M. R., Boyce, A., Chrousos, G., & Corpas, E. (2021). *Introduction to Lipids and Lipoproteins*. Endotext. https://www.ncbi.nlm.nih.gov/books/NBK305896/#lipid_intro.INTRODUCTION

Francula-Zaninovic, S., & Nola, I. A. (2018). Management of Measurable Variable Cardiovascular Disease' Risk Factors. *Current Cardiology Reviews*, *14*(3), 153–163. https://doi.org/10.2174/1573403x14666180222102312

Han, X., & Gross, R. W. (2022). The foundations and development of lipidomics. In *Journal of Lipid Research* (Vol. 63, Issue 2). American Society for Biochemistry and Molecular Biology Inc. https://doi.org/10.1016/j.jlr.2021.100164

Heiles, S. (2021). Advanced tandem mass spectrometry in metabolomics and lipidomics-methods and applications. *Analytical and Bioanalytical Chemistry*. https://doi.org/10.1007/s00216-021-03425-1/Published

Jeyarajah, E. J., Cromwell, W. C., & Otvos, J. D. (2006). Lipoprotein Particle Analysis by Nuclear Magnetic Resonance Spectroscopy. In *Clinics in Laboratory Medicine* (Vol. 26, Issue 4, pp. 847–870). https://doi.org/10.1016/j.cll.2006.07.006

Joseph, P., Leong, D., McKee, M., Anand, S. S., Schwalm, J. D., Teo, K., Mente, A., & Yusuf, S. (2017). Reducing the global burden of cardiovascular disease, part 1: The epidemiology and risk factors. In *Circulation Research* (Vol. 121, Issue 6, pp. 677–694). Lippincott Williams and Wilkins. https://doi.org/10.1161/CIRCRESAHA.117.308903

Mallol, R., Amigó, N., Rodríguez, M. A., Heras, M., Vinaixa, M., Plana, N., Rock, E., Ribalta, J., Yanes, O., Masana, L., & Correig, X. (2015). Liposcale: A novel advanced lipoprotein test based on 2D diffusion-ordered 1H NMR spectroscopy. *Journal of Lipid Research*, *56*(3), 737–746. https://doi.org/10.1194/jlr.D050120

Mallol, R., Rodriguez, M. A., Brezmes, J., Masana, L., & Correig, X. (2013a). Human serum/plasma lipoprotein analysis by NMR: Application to the study of diabetic dyslipidemia. In *Progress in Nuclear Magnetic Resonance Spectroscopy* (Vol. 70, pp. 1–24). Elsevier B.V. https://doi.org/10.1016/j.pnmrs.2012.09.001

Mallol, R., Rodriguez, M. A., Brezmes, J., Masana, L., & Correig, X. (2013b). Human serum/plasma lipoprotein analysis by NMR: Application to the study of diabetic dyslipidemia. In *Progress in Nuclear Magnetic Resonance Spectroscopy* (Vol. 70, pp. 1–24). Elsevier B.V. https://doi.org/10.1016/j.pnmrs.2012.09.001

Nagana Gowda, G. A., & Raftery, D. (2021). NMR-Based Metabolomics. In *Advances in Experimental Medicine and Biology* (Vol. 1280, pp. 19–37). Springer. https://doi.org/10.1007/978-3-030-51652-9_2

Nagana Gowda, G. A., & Raftery, D. (2023). NMR Metabolomics Methods for Investigating Disease. In *Analytical Chemistry* (Vol. 95, Issue 1, pp. 83–99). American Chemical Society. https://doi.org/10.1021/acs.analchem.2c04606

Nordestgaard, B. G. (2016). Triglyceride-Rich Lipoproteins and Atherosclerotic Cardiovascular Disease: New Insights from Epidemiology, Genetics, and Biology. *Circulation Research*, *118*(4), 547–563. https://doi.org/10.1161/CIRCRESAHA.115.306249