Gemma Sánchez Clavell

Predicció d'albúmina en mostres de sang a partir d'espectres de Ressonància Magnètica Nuclear

Treball Fi de Grau

dirigit pel Dr. Xavier Correig

dirigit pel Sr. Daniel Rodríguez

Grau en Enginyeria Biomèdica

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2022

Índex

1 Introducció

1.1 Albúmina en el cos humà

L'albúmina és una proteïna que es troba en gran proporció als limfòcits, essent la principal proteïna de la sang i la més abundant en els humans. Està sintetitzada al fetge, convertint-la així en un bon indicador sobre la funció sintetitzadora d'aquest, és a dir, sobre l'habilitat del fetge de formar proteïnes com normalment les produeix[1].

La concentració normal d'albúmina en la sang humana va entre 3,5 i 5 g/dl. Els nivells d'albúmina de sang s'utilitzen sovint per estudiar la funció del fetge, com s'ha esmentat anteriorment. No obstant això, una disminució de la seva concentració no sempre indica una funció hepàtica pobra, ja que altres factors com la inflamació, la síndrome nefròtica i la malnutrició també poden jugar un paper important en la seva regulació. Com a normes generals, podem dir:

- Els nivells d'albúmina són normals en condicions de deteriorament del fetge agudes, com l'hepatitis viral aguda o l'hepatotoxicitat induïda per medicaments. La possibilitat d'una malaltia crònica subjacent ha de contemplar-se quan els nivells d'albúmina sèrica s'estenen per sota de 3g/dl.

- Els baixos nivells d'albúmina són més freqüents en malalties hepàtiques cròniques, incloent la cirrosi. En aquests casos, la disminució de la concentració d'albúmina significa un greu dany al fetge. Hi ha algunes excepcions en pacients amb ascites i en pacients que estan rebent grans quantitats de fluids intravenosos en els quals els nivells d'albúmina poden semblar baixos a causa de l'augment del volum de plasma.

Els baixos nivells d'albúmina també poden indicar malalties renals en les quals els ronyons no poden evitar que l'albúmina passi de la sang a l'orina. En aquests casos, es pot sol·licitar la determinació dels nivells d'albúmina o proteïna en l'orina (albuminúria o proteïnúria, respectivament). Els nivells elevats d'albúmina es poden veure en els casos dels pacients amb una disminució de la quantitat de fluid, en dietes de proteïnes molt altes i en algunes variants genètiques. Alguns fàrmacs com ara esteroides anabòlics, hormones de creixement, andrògens i insulina poden augmentar els nivells d'albúmina en la sang[1].

Aquesta proteïna és essencial per al manteniment de la pressió oncòtica, que és necessària per a la correcta distribució dels fluid corporals entre els compartiments intravasculars i extravasculars, localitzats enmig dels teixits. A més, ajuda en la regulació del PH de la sang i pot mediar el metabolisme dels lípids, segregar toxines i resistir l'estrès oxidatiu[2].

L'albúmina forma complexos estables amb molts fàrmacs i xenobiòtics; per tant, te un rol important en la farmacinètica o toxicocinètica de nombrosos compostos tenint, així, activitats antioxidants i pseudoenzimàtiques. La unió de les diferents molècules i el seu transport a través d'aquesta proteïna és una de les seves funcions més importants en molts casos[3].

Una altra aplicació de l'albúmina en sèrum (HSA), segons alguns estudis, és la modulació del comportament dels virus i la seva resistència als fàrmacs. Hi ha patògens que al estar exposats a l'albúmina canvien de forma fenotípica, per tant, canvien els gens expressats si estan amb HSA o no. Això implica canvis en el patogen resistent, fent que pugui perdre la resistència a medicaments múltiples[4]. Actualment, un dels majors problemes de salut és la

resistència de les bactèries als antibiòtics. Els resultats descrits anteriorment poden representar un recurs alternatiu per millorar l'eficàcia dels antibiòtics.

Finalment, referent a la unió de compostos, veurem un últim exemple amb una possible aplicació que pot ésser importat per a la salut de les persones. Les substàncies perfluoroalquíliques (PFAS) són una classe de productes químics industrials, que es produeixen en una varietat d'indústries arreu del món des de fa diverses dècades. El PFAS s'empra en processos de fabricació de productes de la vida quotidiana com envasos d'aliments o teixits a prova d'aigua. Nombrosos PFAS es consideren altament persistent en el medi ambient i es poden detectar en aigua i sòl. El perfluorooctansulfonat (PFOS) i l'àcid perfluorooctanoic (PFOA) són els PFAS més freqüents presents en el medi ambient i en els humans. Les preocupacions de salut humana derivades de l'exposició al PFAS són múltiples. Els resultats adversos, com l'augment dels nivells de colesterol sèric, els efectes immunològics i la reducció del pes de naixement són preocupants. La vinculació de PFAS als components del plasma com l'albúmina o lipoproteïnes encara és una qüestió de discussió, però pot ser el primer pas per a una possible solució[5].

1.2 Tècniques de mesura de l'albúmina en sang

Com bé hem vist, l'albúmina és una proteïna de la sang, per tant, per mesurar-la s'ha de disposar d'una mostra de sang per poder extreure el plasma i, una vegada aïllat, mesurar l'albúmina.

Els mètodes convencionals per a la determinació de l'albúmina es classifiquen aproximadament en dues classes: Els mètodes per determinar l'albúmina fraccionada (precipitació salina i electroforesi), i els mètodes per determinar específicament l'albúmina sola[6].

La precipitació salina és un fenomen físic-químic basat en les interaccions electròlit- no electròlit, en el qual a altes forces iòniques alguns soluts com proteïnes o polímers precipiten degut al augment de les interaccions hidrofòbiques entre ells. Si aquest tècnica s'aplica al sèrum, obtenim un precipitat que conté l'albúmina a mesurar. A partir d'aquí, s'utilitzen mètodes d'electroforesis per fer una estimació de la proteïna d'interès. Després d'aquesta electroforesi, les proteïnes del sèrum es precipiten en el medi de suport i es mantenen amb un tint. Aquest tint és particular per cada proteïna ja que no totes s'uneixen als mateixos compostos, llavors utilitzen l'especificitat lligand-proteïna per assignar el pigment. Finalment, L'albúmina es mesura com un percentatge de la concentració total de proteïnes mitjançant l'estimació de la proporció de tint lligat a ella[7]. Dintre del segon grup trobem, per exemple, la tècnica BCG que s'utilitza per determinar l'augment de l'absorbància a 630 nm causada per l'enllaç entre un reactiu verd a l'albúmina que es dissol en una solució d'amortiment[6].

Actualment, aquest mètode comença a tenir molts inconvenients ja que es una tècnica lenta que no dona resultats a l'instant. És per això que ja s'han desenvolupat altres procediments basats en les reaccions immunològiques de les cèl·lules.

L'immunoassaig és un conjunt de tècniques immunoquímiques analítiques de laboratori que tenen en comú l'ús de complexos immunes, és a dir els resultats de la conjugació de antigen i anticòs, com referències de quantificació d'un analit determinat, que pot ser l'anticòs o l'antigen, usant per a la mesura una molècula com a marcador que forma part de la reacció amb el complex immune de l'assaig químic. La tècnica se basa en la gran especificitat i afinitat dels anticossos pels seus antígens. La seva alta sensibilitat permet la quantificació de compostos presents en líquids orgànics en concentracions reduïdes. Per tant, es pot posar en contacte l'albúmina amb un portador de membrana que conté certa quantitat de grups

d'aldehids. Llavors, es fan diversos tipus d'immuno-reaccions entre el portador i l'albúmina que es conjuga amb un enzim "etiqueta" com la glucosa oxidasa. S'utilitza la mesura de l'activitat enzimàtica per a la determinació de l'albúmina, ja que aquesta activitat esta relacionada amb la quantitat d'antigen que hi ha [6].

1.3 Metabolòmica 1H NMR

La metabolòmica és a tècnica que permet la mesura i anàlisi global dels compostos de baix pes molecular (metabòlits) que es troben en els biofluids i en els teixits dels sers vius i que intervenen en múltiples vies moleculars. Es basa en tecnologies analítiques avançades, com la espectrometria de masses (MS) i espectroscòpia de ressonància magnètica nuclear (NMR), que permeten detectar compostos sense la necessitat de mesures bioquímiques addicionals[8].

L'aplicació d'aquestes tècniques s'estén a diversos camps. Un exemple és l'epidemiologia i la genètica on l'elaboració de perfils metabòlics complets són cada vegada més comuns i permeten el diagnòstic i l'estratificació del risc de les persones en contraure malalties . La majoria d'aquests estudis analitzen compostos en sang que poden ser biomarcador d'una determinada malaltia (com ara el colesterol total o la glucosa). També s'han utilitzat per realitzar perfils metabòlics multivariables de cèl·lules, teixits i fluids biològics per estudiar la variació fisiològica en animals, ja sigui en quan a sexe, edat o fenotips mutants i transgènics. Per últim, la metabolòmica també ha trobat aplicació en estudis de factors com la nutrició i la microflora intestinal. Des de que s'ha demostrat la possibilitat de predir els efectes postdosi de fàrmacs a partir de perfils metabòlics de predosi, aquest enfocament farmacològic s'ha identificat una tecnologia clau en la medicina personalitzada[9][10].

Com s'ha esmentat anteriorment, la NMR té algunes característiques interessants que fan que aquesta tècnica sigui especialment adequada per a l'anàlisi en els estudis de metabolòmica a gran escala; la quantificació ve donada per les àrees dels pics de l'espectre que estan directament relacionades amb la concentració molar d'un nucli específic (generalment 1H). A més, la NMR permet el perfilat metabòlic de biofluids i teixits intactes sense necessitat de costos processos d'extracció o separació de metabòlits.

1.3.1 Fonaments del NMR

Quan una mostra s'introdueix a l'espectròmetre, alguns espins de la mostra s'alineen amb el camp magnètic constant constituint un estat baix d'energia, la resta es troben en un estat d'alta energia. La distribució dels espins entre aquests dos estats pot ser alterada aplicant un pols de radiofreqüència (RF) d'una freqüència específica coneguda com a freqüència de Larmor. Una vegada el pols RF s'apaga, es registra la pèrdua d'energia de qualsevol espín excitat per recuperar el seu estat d'equilibri. El senyal obtingut es coneix com la desintegració lliure i conté la suma de la pèrdua d'energia de tots els espins excitats.[11]

Els espins en una molècula experimenten un entorn magnètic lleugerament diferent depenent dels nuclis circumdants. Els diferents ambients magnètics fan que una molècula mostri un conjunt de senyals dispersats al llarg de l'eix de freqüència d'un espectre de NMR, que estan relacionats amb els seus grups funcionals (per exemple, metil, metilè, aromàtics, etc.). Aquest signatura espectral caracteritza les diferents espècies moleculars i revela la seva composició estructural[11].

1.3.2 Aplicació de perfils 1H-NMR de biofluids

Els biofluids s'utilitzen habitualment en estudis metabolòmics perquè contenen centenars de metabòlits i les mostres es poden obtenir d'una forma no invasiva o mínimament invasiva.

En metabolòmica, el nucli més utilitzat per ser analitzat amb NMR és el protó (1H-NMR) a causa de la seva alta sensibilitat, relaxació ràpida, abundància i la seva presència en els compostos orgànics[11]. L'espectre 1H-NMR dels biofluids consisteix en un conjunt de senyals superposats d'una gran nombre de compostos a concentracions molt diferents, fent que la identificació i quantificació fiables sigui una tasca difícil. A més, la complexitat espectral s'amplifica a causa del pH, la força iònica, la composició d'ions metàl·lics i altres processos d'intercanvi químic perquè afecten grups específics de metabòlits, causant variacions de desplaçaments químics entre mostres[12]. Això també afecta a la quantificació ja que augmenta o disminueix el senyal en funció d'aquest desplaçament[13]. Per tant, podem dir que l'elaboració de perfils 1H-NMR és un procés laboriós que requereix una manipulació intensiva de dades (processament espectral, identificació i quantificació).

1.3.3 Aplicació de perfils 1H-NMR de sèrum/plasma: Les tres finestres moleculars

La sang conté molècules de diferents tipus com ara, proteïnes, lípids, colesterols, metabòlits i ions de baix pes molecular, i les seves concentracions van des de nM a mM. El sèrum procedent de la sang sembla ser el derivat preferit per l'elaboració de perfils H-NMR, ja que l'espectre sèric manca de senyals d'interferència assignats a la coagulació de proteïnes i additius anticoagulants. [11]

Els pics dels metabòlits de baix pes molecular (LMWM) es superposen amb els senyals amplis de macromolècules (principalment lipoproteïnes i albúmina) en l'espectre 1 H-NMR del sèrum. Aquesta complexitat evita l'ús d'un únic experiment H-NMR per la caracterització completa de la bioquímica de la sang. Ala-Korpela i els col·laboradors van proposar la implementació d'un model de tres finestres moleculars que permet la quantificació exhaustiva de les classes de lipoproteïnes i els lípids constituents, l'albúmina, i una ran varietat de metabòlits de baix pes molecular, incloent aminoàcids, metabòlits relacionats amb la glicòlisi i cossos de cetona[9].

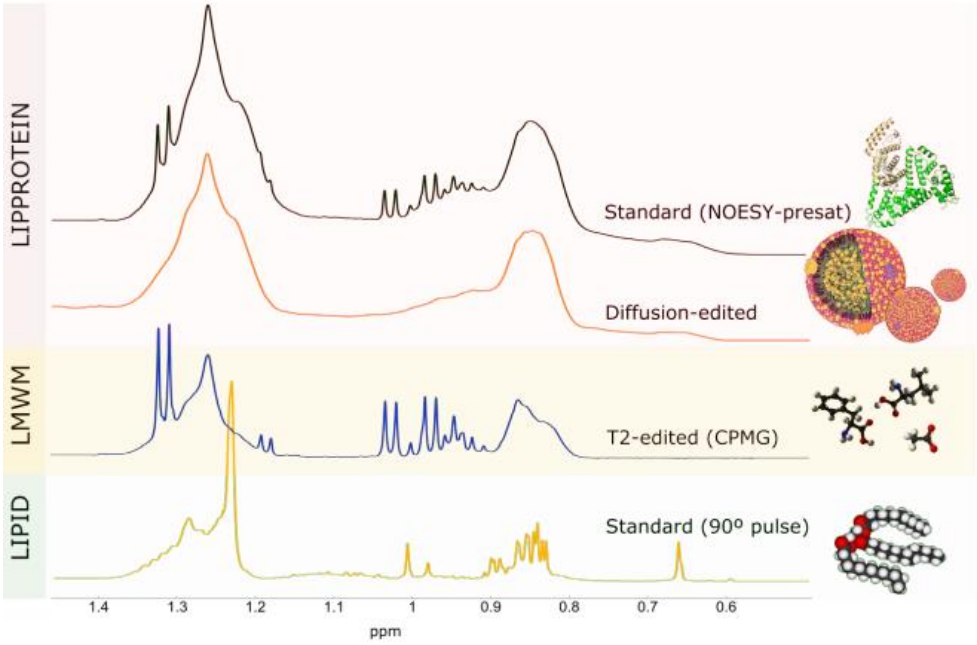


Figura 1. Regions de metil i metilè utilitzant el model de tres finestres moleculars i exemples de compostos que es poden mesurar en cada finestra. [R. Ba. Regadera, "Development of ^{1}H-NMR Serum Pro fi ling Methods for High-Throughput Metabolomics," 2017.]

L'elaboració de perfils H-NMR del sèrum ens ha portat a una comprensió més profunda sobre la patogènesis i les malalties, així com a la identificació de biomarcador metabòlics, com ja he esmentat, per al diagnòstic de malalties o el seguiment dels tractaments emprats. Per tant, els perfils H-NMR podrien millorar el diagnòstic clínic de diversos tipus de malalties com el càncer, la malaltia d'Alzheimer, la diabetis tipus 2 i malalties cardiovasculars, entre altres.

1.4 Mesura d'albúmina amb espectroscòpia NMR

La quantificació de metabòlits de baix pes molecular (LMWM) per 1H NMR s'utilitza rutinàriament en metabolòmica sèrica ja que ofereix un alt rendiment. La unió de les proteïnes modifica les propietats de moviment del LMWM i el seu senyal s'atenua parcialment amb un filtre, juntament amb el fons de la proteïna. En conseqüència, els senyals LMWM quantificats no reflecteixen la concentració total en el sèrum, sinó la part no vinculant. Les mostres biològiques complexes, com en el cas del sèrum i el plasma, impliquen múltiples fonts d'intercanvi químic, com la quilació d'ions metàl·lics, la protonació, l'intercanvi de protons amb aigua i la unió molecular, el que porta a modificacions dels senyal, com ara canvis espectrals, disminució d'amplitud i eixamplament de la línia. En aquest sentit, la capacitat de les proteïnes plasmàtiques per unir petites molècules pot afectar l'anàlisi quantitatiu de la NMR de LMWM per dos processos d'intercanvi simultanis. En els estudis on s'ha realitzat espectroscòpia NMR i albúmina, el que s'ha fet és analitzar l'afinitat d'unió dels metabòlits a HSA en mostres de sèrum i promoure l'alliberament de la seva fracció unida . Per fer-ho, s'utilitza una combinació de CMPG que és un filtre de NMR, i l'anàlisi de resolució de corbes multivariants , permetent la separació de senyals LMWM i proteïnes[14].

Per tant, només hi ha constància d'estudis realitzats amb NMR per trobar metabòlits que tinguin un lligand amb l'albúmina. D'aquesta manera, es pot intuir en quines rutes metabòliques te un paper important aquesta proteïna i com pot influir com a biomarcador. No hi ha cap publicació que mesuri la concentració d'albúmina amb NMR.

1.5 Objectius TFG

Degut a que no hi ha cap registre sobre la mesura de la quantitat d'albúmina en espectroscòpia de ressonància magnètica nuclear, el principal objectiu d'aquest projecte és la correcta determinació d'albúmina en mostres de sèrum utilitzant la tecnologia de NMR. Per aconseguir-ho s'elaboraran models de predicció mitjançant espectres LED de NMR, a partir de mesures d'albúmina determinada per mètodes tradicionals.

2 Metodologia

2.1 Conjunt de dades

2.1.1 Conjunt 1

En aquest conjunt trobem mostres de plasma de 340 pacients amb diabetis de tipus 2, obesitat i síndrome metabòlica. Aquestes mostres van ser recollides per la Unitat de Lípids i arteriosclerosi de l'Hospital Universitari Sant Joan de Reus amb l'objectiu de mesurar el nombre de partícules, les diferents mides que hi ha en elles (petites, mitjanes o grans) i els perfils de glicoproteïna Glyc-A i Glyc-B en 9 subfraccions de lipoproteïnes [15]. Dintre de la base de dades, no es podia distingir les diferents malalties a partir de les seves ID. Totes les persones participants en l'estudi varen donar el seu consentiment informat per tal de poder utilitzar les mostres de sang en estudis de metabolòmica.

Inicialment aquest data set contenia 340 mostres que, després de passar els diferents filtres de qualitat dels espectres i els criteris per l'eliminació de outlayers, s'ha reduït a 317 mostres en total.

Després d'analitzar la distribució d'albúmina en les mostres, es va separar el conjunt amb 3 grans grups diferents per treballar en ells per separat perquè eren molt diversos entre ells per poder-los ajuntar a priori. La separació d'aquests grups es va realitzar gràcies a les diferents identificacions de les mostres, ja que es veia clarament l'existència dels tres grups.

2.1.2 Conjunt 2

Aquest és un estudi amb pacients de cirrosi. En aquest conjunt no es va eliminar cap mostra del total de mostres que se'ns proporcionà inicialment. Comptem amb 65 mostres que tenen hipoalbuminèmia, és a dir, els nivells d'albúmina en plasma estan per sota dels nivells mínims en una mostra no patològica. Totes aquestes persones o els seus representants legals i el comitè d'ètica de cada hospital del consorci EASL Clif implicats en l'estudi va donar el consentiment informat per investigacions en ciències òmiques.

Dintre d'aquestes mostres hi ha quatre grups: (1) Controls (HS), (2) Cirrosi Compensada (CC), (3) Descompensació aguda (AD) i (4) Fallo renal agut crònic (ACLF). A més, cinc mostres sense haver patit un cicle de congelació[16]. No obstant això, en aquest treball, s'han tractat les mostres de la mateixa manera sense distinció grupal.

2.1.3 Conjunt 3

En aquest conjunt teníem 280 mostres en un principi i, en aplicar els mateixos filtres que ens altres dos conjunts, se'ns reduí el data set a 165 mostres en total d'un únic grup. Totes les persones participants en l'estudi varen donar el seu consentiment informat per tal de poder utilitzar les mostres de sang en estudis de metabolòmica.

Aquestes mostres també són de població general i no ens consta que hi hagi cap patologia a tenir en compte a l'hora de treballar i utilitzar-les.

2.1.4 Preparació i adquisició d'espectres

Tots els conjunts són mostres de plasma sense cap patologia a priori i tractades de la mateixa manera. Els reactius utilitzats per a la preparació de les mostres van ser aigua deuterada de Euroisotop, les sals NaH2PO4 i Na2HPO4 de Labkem per a preparar el tampó fosfat salí (PBS, per les seves sigles en anglès) i NaOH de POCH per a ajustar el pH del PBS.

Per a la preparació, les mostres es van descongelar i homogeneïtzar abans de ser preparades de forma automàtica utilitzant el robot manipulador de líquids Gilson. Breument, es van diluir 200 µl de mostra plasmàtica amb 300 µl de PBS amb una concentració de 50 mM (pH = 7.4), es van afegir 50 µl d'aigua deuterada (D2O) i es van homogeneïtzar directament al tub de RMN en condicions controlades de temperatura 4ºC.

Les mostres diluïdes es van analitzar per espectroscòpia de ressonància magnètica nuclear de protó (1H-NMR) utilitzant un espectròmetre Ultrashield Plus 600 de Bruker, operant a una freqüència de 600 MHz (14.1 T) a 305,95 K. Els espectres es van adquirir utilitzant el pols longitudinal eddy corrent delay (LED).

2.2 Anàlisi estadístic i multivariat

2.2.1 STOCSY

La espectroscòpia de correlació total estadística (STOCSY) és un mètode ben establert que aprofita la multicolinealitat de les variables en un conjunt d'espectres (en este cas, espectres 1H-NMR) per a generar un espectre NMR pseudodimensional que mostra les correlacions entre els diferents pics dels espectres que ha agafat i el metabòlit d'interès[17].

En particular, els trets de la mateixa molècula mostraran correlacions fortes d'intensitat positives i aquesta informació pot ser emprada en les assignacions espectrals/estructurals.[18] Mentre que les correlacions altes entre pics solen indicar ressonància estructuralment relacionades, és a dir, pics que sorgeixen de la mateixa molècula, les correlacions més baixes poden al·ludir a relacions biològiques, per exemple, entre molècules de la mateixa via bioquímica o intermediaris de reaccions[19].

STOCSY es basa en les propietats de la matriu de correlació C, calculada a partir d'un conjunt d'espectres de mostres segons l'equació següent

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}_1{}^{t}\mathbf{X}_2$$

On X1 i X2 són les matrius experimentals autoescalades de n x v1 i n x v2, respectivament; n és el número d'espectres i v1 i v2 són el número de variables en els espectres de cada matriu. C és, per tant, una matriu de v1 x v2 on cada valor és un coeficient de correlació entre dos variables de les matrius X1 i X2. [17]

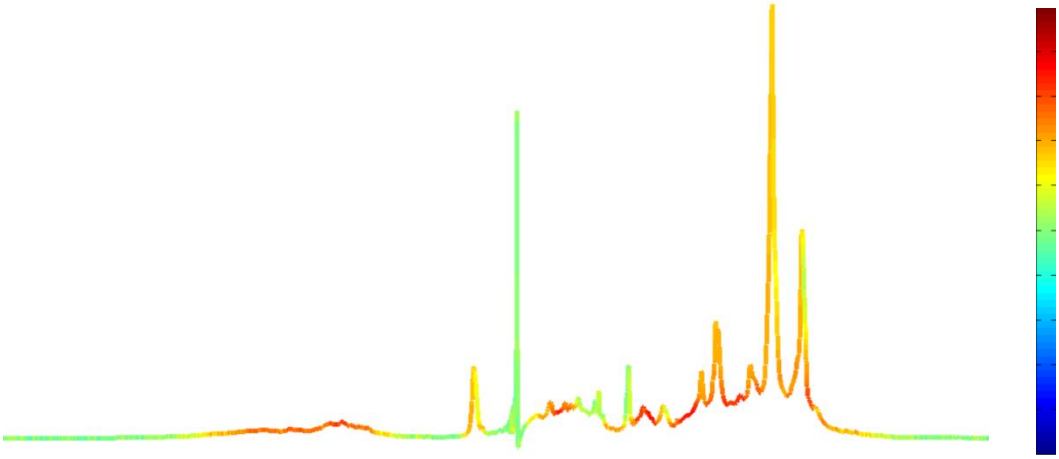


Figura 2: Exemple d'un espectre STOCSY on les correlacions positives són els colors càlids i les negatives els colors freds.

Per tant, es va realitzar aquesta tècnica per determinar quines regions de l'espectre van mostrar l'associació més forta amb les variables a predir, en el meu cas, l'albúmina.

2.2.2 T-test

El t-test és un tipus d'estadística diferencial utilitzada per determinar si hi ha una diferència significativa entre les mitjanes de dos grups.

En el meu cas es va utilitzar per veure si la predicció de l'albúmina es veia afectada per les diferències que hi havia entre els dos subgrups del grup 2 del conjunt 1. Com ja he esmentat, el conjunt 1 es va desglossar en 3 grups diferents. El segon agrupament es va dividir en casos i controls, per tant hi havia diferències considerables entre ells en els resultats. És per això que a partir d'aquets moment, el grup 2 només consta d'un dels subconjunts, amb un total de 72 mostres.

2.2.3 Detecció Outlayers

Els outlayers són mostres que el seu espectre 1H-NMR i/o els valors de les variables clíniques/bioquímiques es troben molt allunyades de la resta, de mostres, que considerem normals. A partir d'aquí, per detectar aquests outlayers s'apliquen una sèrie de criteris en funció de l'estudi que es realitza.

En aquest treball, concretament, s'han fet ús de tres comprovacions necessàries. Primerament, totes les mostres van passar per un control de qualitat del software de Liposcale. Aquest control és automàtic i informa de quines són les possibles mostres outlayers. Una vegada fet, també es comprovà que les mostres tinguessin un nivell de triglicèrids totals menor a 600 perquè si la concentració és més elevada, el software perd precisió a l'hora de obtenir els resultats i l'error comès és superior al permès. Per últim, es va mirar les correlacions que donaven els diferents grups de cadascun dels conjunts en STOCSY per veure si hi havia algun problema afegit en algun grup en particular.

A partir d'aquesta detecció es comproven totes les mostres seleccionades i, una per una, es decideix si forma part del conjunt total o s'elimina. Això es fa buscant una justificació amb les característiques de l'espectre, ja sigui amb intensitat, soroll que pugui haver a la mostra o

altres diferències que puguin semblar a priori extranyes. A més, cal veure quina és la lectura de l'albúmina que s'ha obtingut d'aquests outlayers perqué també es pot veure una certa influència d'aquest metabòlit en l'espectre.

2.2.4 Normalització i escalat

Abans de fer un model PLS, s'ha de triar quin tipus de normalització i escalat s'ha d'utilitzar per a que els resultats siguin el millor possible. La normalització i l'escalat s'utilitzen per a que totes les mostres, per diferents que siguin, es troben dins d'un mateix interval.

En el nostre cas, s'han realitzat diferents models amb diferents pre-processats per veure quin d'ells permetia predir millor l'albúmina.

S'han utilitzat 3 tipus de pre-processat: Z-score, auto scale i *mean centering*. El Z-score es va realitzar calculant la desviació estàndard i la mitjana aritmètica per columnes de cadascun dels punts dels espectres, una vegada obtenim aquests valors s'aplica la fórmula següent $z = (x - \mu)/\sigma$ on z és el nou valor del punt de l'espectre, x és el punt original, μ la mitjana i, finalment, σ la desviació estàndard. Auto scale es similar al Z-score però, en aquest cas, en comptes de la mitjana, s'utilitza el valor mínim d'aquela variable essent la fórmula d'aquesta manera $z = (x - xmin)/\sigma$. Finalment, *mean centering* el que fa és substreure a cada variable (columna) el seu valor mitjà i restar-lo a cadascuna de les mostres. Això es realitza en bucle fins que la mitjana de la variable sigui 0. Cal afegir que MATLAB proporciona els dos darrers pre-processats de forma automàtica quan afegeixes la matrius de espectres a la toolbox dels PLS, mentre que el Z-score s'ha fet de forma manual, també utilitzant MATLAB, amb el càlcul previ de les mitjanes i les desviacions estàndards. D'aquesta manera es pot observar millor la influència del pre-processat i dels errors que es poden tenir a l'hora de calcular-los.

2.2.5 PLS

2.2.5.1 Definició

La regressió parcial per mínims quadrats (PLS) és un mètode d'estimació. PLS ha estat cada vegada més utilitzat entre els químics i altres científics com una tècnica per tractar dades altament col·lineals. S'aplica gairebé de manera rutinària en espectroscòpia, on un objectiu es predir la concentració d'un compost química d'un espectre[20].

El PLS és un procediment utilitzat comunament per al calibratge multivariat, la mineria de dades i altres anàlisis de dades. Modela una matriu de resposta, Y, com una combinació lineal d'un conjunt de variables, recollides en la matriu X. PLS assumeix que X i Y són manifestacions del mateix conjunt de variables latents (LV), és a dir, les variables X i Y estan relacionades entre si a través d'aquestes variables LV. X és la matriu opcionalment escalada i centrada de variables predictores, per exemple, espectres digitalitzats, i y és el vector de la variable de resposta única, per exemple, la concentració d'albúmina de les mostres, també opcionalment escalat i centrat[21].

Per veure si un model és robust o no, MATLAB dona informació sobre aquest una vegada ha estat creat. En primer lloc, es fa la validació del model i a partir dels resultats predits, es correlen amb els valors reals i s'obté el coeficient de Pearson (R de la regressió). Per últim, es mostra la R^2 del model que indica l'estabilitat de l'esmentat i que no està relacionat amb el coeficient de Pearson que nosaltres calculem en la regressió.

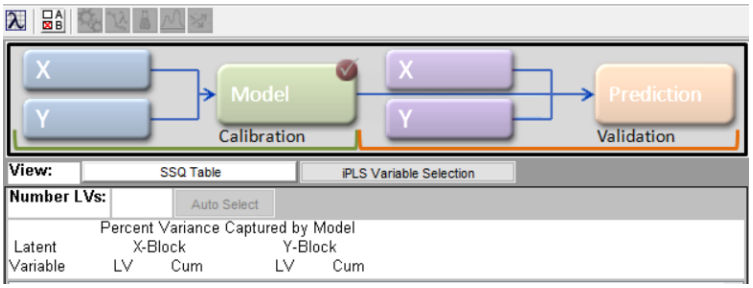


Figura 3: Toolbox de matlab utilitzada per fer els PLS.

2.2.5.2 Aplicació en el treball

En el meu cas particular, per fer els diferents models PLS s'han utilitzat per a la matriu X, un % dels espectres de les mostres situades a cadascun dels grups que hi ha dintre del conjunt 1, és a dir, un % del grup 1, un % del grup 2 i un altre % del grup 3. D'aquesta manera aconseguim que hi hagi mostres te tots els tipus per a fer l'entrenament del model. Com a vector y, hem utilitzat la quantitat d'albúmina que hi ha en les mateixes mostres emprades per a la matriu X. Cal esmentar que, en la matriu X no agafàvem tots els punts dels diferents espectres, sinó que agafàvem diferents regions segons el que sortia al STOCSY. Això a permès poder fer més d'un model utilitzant els mateixos espectres, ja que en cadascun d'ells hi havia diferents punts. Finalment, una vegada fet el model PLS, es van utilitzar el % restant de la resta de mostres i els conjunts 2 i 3 (tots tres per separat), per poder validar-lo. Per tant, la matriu de resposta Y és la predicció d'albúmina dels conjunts 2 i 3 i d'un % del conjunt 1 que dóna el model creat. Per poder fer la predicció de les mostres, s'ha d'introduir els espectres d'aquestes ja que el que fa el model és predir la concentració d'albúmina a partir de l'espectre NMR de la mostra. Tot això, es pot veure representat a la taula 1.

A més, tant la matriu X, com el vector y han estat pre-processats de la mateixa manera, utilitzant els diferents pre-processats per veure quin dona millor resultats.

MODEL	TRAINING SET	TEST SET
MODEL 1	70% mostres del grup 1 70% mostres del grup 2 70% mostres del grup 3 175 mostres en total del conjunt 1	30% restant del grup 1 30% restants del grup 2 30% restant del grup 3 83 mostres restants del conjunt 1
MODEL 2	60% mostres del grup 1 60% mostres del grup 2 60% mostres del grup 3 151 mostres totals del conjunt 1	40% restant del grup 1 40% restant del grup 2 40% restant del grup 3 107 mostres restants del conjunt 1 Els conjunts sencers 2 i 3
MODEL 3	50% mostres del grup 1 50% mostres del grup 2 50% del mostres grup 3	50% mostres del grup 1 50% mostres del grup 2 50% mostres del grup 3

	129 mostres totals del conjunt 1	129 mostres restants del conjunt 1
MODEL 4	75% mostres del grup 1	25% restant del grup 1
	75% mostres del grup 2	25% restants del grup 2
	75% mostres del grup 3	25% restant del grup 3
	185 mostres en total del conjunt 1	73 mostres restants del conjunt 1

Taula 1: Mostres utilitzades en el training set i en el test set durant la creació de models PLS

2.3 Eina software

2.3.1 MATLAB

MATLAB és un entorn de computació numèrica i un llenguatge de programació que permet manipular fàcilment matrius, dibuixar funcions i dades, implementar noves funcions i algorismes i fer funcions específiques segons les toolbox que hi hagi instal·lada.

En el meu cas particular, Matlab ha estat l'eina més utilitzada durant aquest treball perquè era el programa òptim que incorporava les toolbox necessàries per poder realitzar els STOCSY corresponents, els diferents PLS i per poder tractar les dades i analitzar-les de la millor manera possible.

2.3.2 Python i R

Python és un llenguatge avançat de programació i per tenir-lo vaig instal·lar-me el programa anaconda que és un distribuïdor lliure i de codi obert dels llenguatges de programació python i R per al processament de dades de gran escala, analítica predictiva i computació científica, que tracta de simplificar la gestió i desplegaments de paquets. Personalment, s'ha utilitzat per poder fer els t-test necessaris per l'estudi.

2.3.3 Liposcale

Liposcale® és una prova de lipoproteïna avançada basada en l'espectroscòpia 1H NMR (DOSY). Aquest nou enfocament permet mesurar els coeficients de difusió, que estan associats amb les diferents subclasses de lipoproteïnes, i calcular directament les seves mides a través de l'equació de Stokes-Einstein. A més, aquesta tècnica no només estima el nombre de partícules de les classes principals (VLDL, IDL, LDL, i HDL) sinó que també quantifica la concentració de colesterol i triglicèrids d'aquest últim. Per obtenir el nombre de partícules de cada subfracció de lipoproteïnes, el volum espacial de molècules de lípids totals s'ha de dividir per la mida mitjana de les partícules de lipoproteïna[22].

En aquest treball s'ha utilitzat el següent software per la obtenció dels diferents espectres dels conjunts, per al control de qualitat i, per últim, per establir els paràmetres que decideixen els outlayers que hi ha.

3 Resultats

3.1 Desenvolupament de la base de dades i anàlisi dels outlayers

Per desenvolupar el treball s'han utilitzat 3 conjunts diferents de mostres, cadascun amb diferents característiques per poder avaluar la funcionalitat dels diferents models. No obstant això, els PLS només s'han realitzat amb el conjunt 1 , mentre que el 2 i el 3 s'han utilitzat per validar-los quan ja donaven bons resultats. Per tant, en aquest apartat ens centrarem més amb el conjunt 1 que amb la resta.

En primer lloc es va realitzar una comprovació de tot el conjunt per assegurar que totes les mostres comptaven amb nivells d'albúmina i amb el seu respectiu espectre i ens vam trobar que, en alguns casos, hi havia mostres que apareixien amb valors buits d'albúmina o que no tenien espectre. Aquestes mostres van ser eliminades ja que no es podien utilitzar per a l'estudi perquè només es disposava de la meitat de la informació necessària. Després, es comprovà els criteris que es van estipular per detectar els outlayers i poder eliminar totes aquelles mostres que els complien. Amb això, es van eliminar 23 mostres, deixant al conjunt 1 amb 317 mostres en comptes de 340.

Un com eliminades les mostres, es va fer un primer STOCSY per veure les correlacions amb l'albúmina.

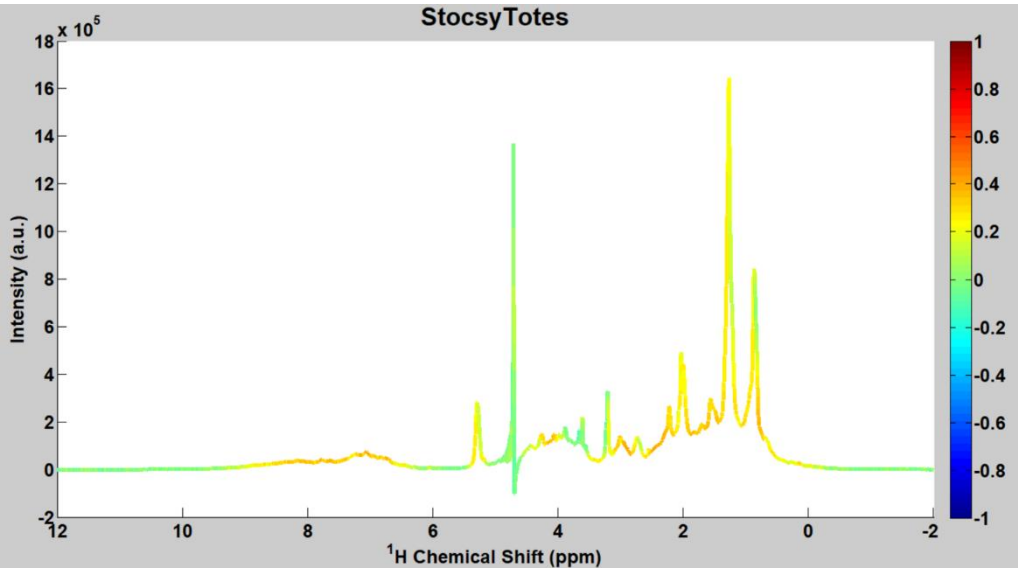


Figura 4: STOCSY del conjunt 1.

Com es pot veure en la Figura 4, no s'aprecia cap tipus de correlació amb l'albúmina. Degut a aquesta situació, es va decidir fer un segon anàlisi del conjunt 1 per poder donar una explicació i per intentar millorar el STOCSY i poder aconseguir l'objectiu principal. Per les identificacions de cada mostra, es va poder detectar que hi havia 3 grups diferents dintre del mateix conjunt i es va realitzar un histograma d'aquets grups per poder veure com l'albúmina es distribuïa, ja que, si la distribució era molt diferent s'havien de tractar els grups per separat. Aquest va ser el resultat:

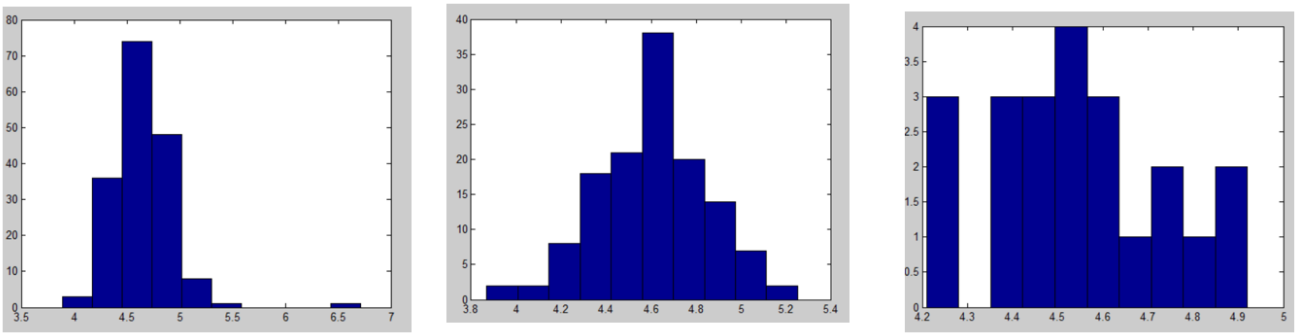


Figura 5: Histogrames dels diferents grups dintre del conjunt 1 que mostres la distribució d'albúmina en ells. D'esquerra a dreta els histogrames corresponen al grup 1, grup 2 i grup 3, respectivament.

Es pot apreciar que, les distribucions en cadascun dels grups són molt diferents i que, per tant, s'han de separar les mostres en aquests agrupaments segons la ID. El primer grup compta amb 163 mostres, el segon amb 131 i, finalment, el tercer grup amb 24 mostres.

Una vegada separats, es va tornar a comprovar STOCSY per cadascun dels grups que es van formar. El grup 1 i 3 ja correlaven de forma correcta i ja es podia apreciar zones d'interès on ressonava l'albúmina (Figura 6 i 7). En el grup 2, es seguia veient una imatge semblant a la Figura 4. En conseqüència, es va separar amb dos subgrups diferents que també estaven marcats per les seves ID, denominant-los, casos i controls. Fent un tercer STOCSY a aquest grup separant els cassos i controls, els resultats ja van ser optimistes perquè havíem aconseguit que el subgrup cassos correles d'igual manera que el grup 1 i 3 (Figura 8). El subgrup control seguia donant un aspecte semblant a la Figura 4.

En aquest punt, es va decidir fer un t-test per veure la influència, segons les variables de Liposcale, de l'albúmina. Com hipòtesis nul·la es considera que no hi ha influència en els valors de Liposcale per a l'albúmina, com a alternativa, sí hi ha influència i per tant el grup control pot portar inconvenients a l'hora de realitzar l'estudi. El resultat d'aquest t-test va esser un p-value igual a 0.00389. Amb aquest p-value s'ha de donar per bona la hipòtesi alternativa i per tant s'ha d'eliminar el grup control de la base de dades. Per tant, s'eliminen 58 mostres més del conjunt 1, deixant un total de 259 mostres on 163 pertanyen al grup 1, 72 al grup 2 i 24 al grup 3.

Pel que fa al conjunt 2 i 3, es va aplicar els criteris de detecció de outlayers i nomes es van trobar 3 mostres sospitoses del conjunt 2 que van ser eliminades com a outlayers. Per tant, no va haver gaires canvis en aquets conjunts ja que des d'un principi les bases de dades van estar prou netes.

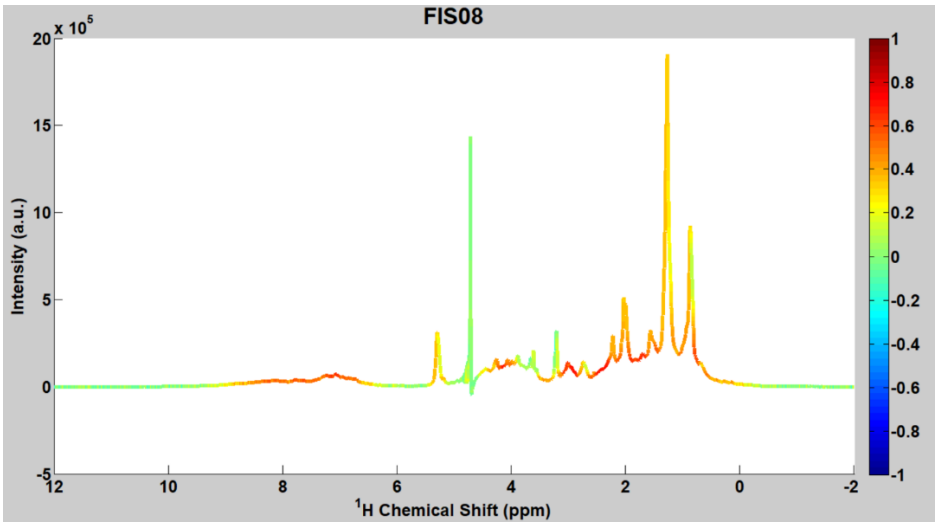


Figura 6: STOCSY del grup 1 del conjunt 1.

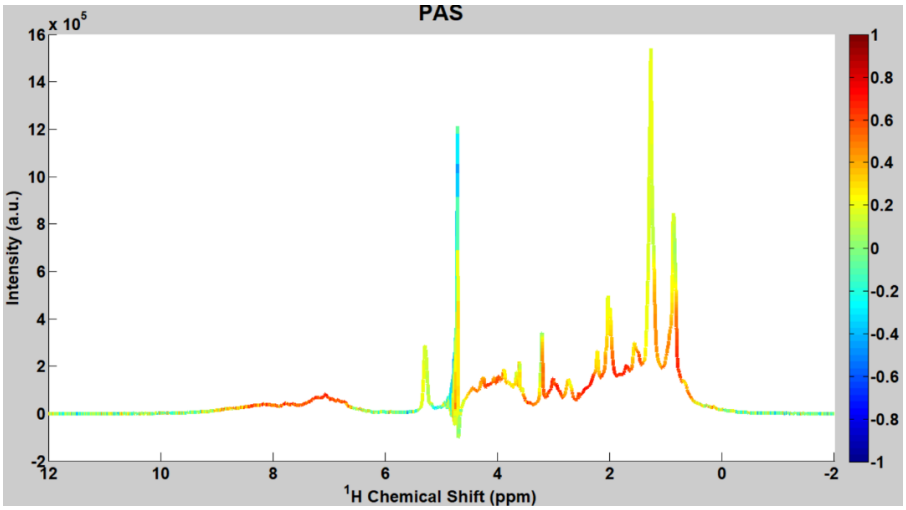


Figura 7: STOCSY del grup 3 del conjunt 1.

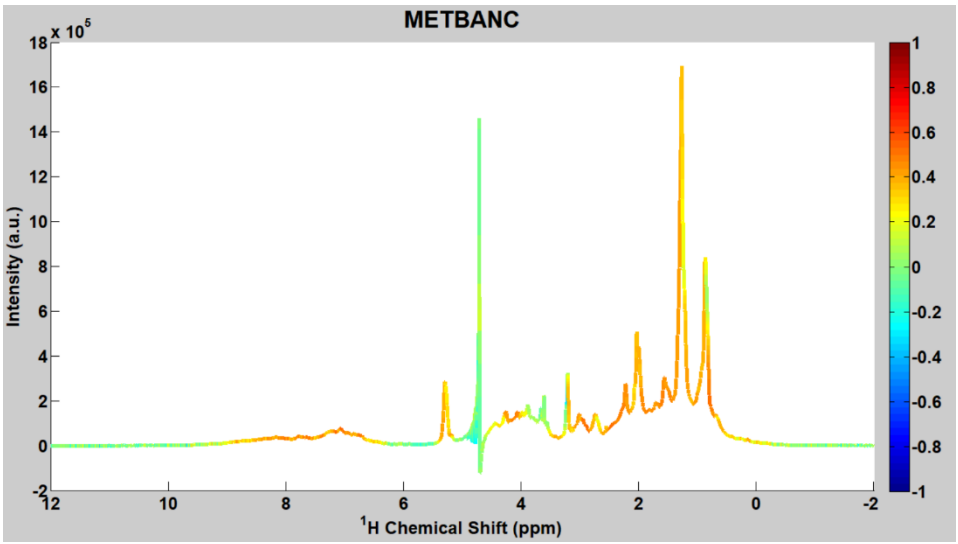


Figura 8: STOCSY del subgrup cassos del grup 2 del conjunt 1.

3.2 Selecció de les regions d'interès

STOCSY es va utilitzar com a part del procediment de selecció de variable abans de la construcció dels models de predicció. Com es pot apreciar a les diferents figures anteriors, les regions amb una major correlació coincideixen tot i ser grups diferents. Per tant, en tots els grups utilitzarem els mateixos punts per poder fer un bon model.

Inicialment, comptàvem amb 10 regions amb ressonància d'albúmina alta. Deu intervals són molts de punts de l'espectre i per tant aquesta quantitat s'havia de reduir d'alguna forma. Mitjançant prova i error a l'hora de realitzar els PLS es va veure un comportament satisfactori al agafar certs intervals en concret i deixar-ne altres fora. També es va veure que alguns dels punts eren essencials per poder fer una bona predicció ja que si no estaven la regressió que hi havia entre la predicció i les mesures reials era mot poc precisa. Finalment, es va aconseguir reduir aquestes zones a 6 regions d'interès.

Quan es parla de intervals, ens referim de quin punt a quin punt de l'espectre del STOCSY agafem segons la correlació d'aquest. Entenent això, els 10 intervals (en ppm) de l'espectre que s'han escollit en un principi són: [7.8 - 6.6], [5.4 – 5.1], [4.3 – 4], [3.1 – 2.85], [2.8 – 2.65], [2.5 – 2.17], [2.1 – 1.9], [1.6 – 1.45], [1.4 – 1] i [1 – 0.7].

Finalment, les 6 regions (en ppm) importants que s'han realitzat per als PLS són: [7.8 - 6.6], [4.3 – 4], [3.1 – 2.85], [2.1 – 1.9], [1.4 – 1] i [1 – 0.7].

3.3 Estimació de la concentració d'albúmina amb els models PLS

Durant aquest treball s'han realitzat aproximadament 50 models PLS diferents, la majoria han estat realitzats per comparar diferents característiques, com per exemple el pre-processat, per poder fer una selecció adient de cadascun dels components que formen un model PLS robust. A la taula següent es mostren els models més rellevants per aquesta investigació.

Com podem observar en la Taula 2, el paràmetre que indica un model PLS robust és la R de la seva regressió en comparar la predicció d'aquest amb el valor real d'albúmina. Aquesta R ha d'estar entre 0 i 1. També s'ha de mirar les variables latents (VL) de cadascun dels models i veure si amb l'acumulació de variància podem afegir o llevar alguna VL sense produir gaires canvis als models i, sobretot, sense produir *overfitting*.

Hi ha models PLS que són prou robustos per a fer una bona predicció. És per això, que també es va intentar fer un model per als espectres NOESY. En aquests espectres no hi ha cap tipus de filtre i per tant hi ha molt de soroll però si es pogués realitzar seria un avanç important perquè estalviaria temps a l'hora d'adquirir els espectres ja que no caldria filtrar-los com en el LED.

MODEL	TRAINING SET	TEST SET	PRE-PROCESSAT	VL	RANGS	R
MODEL 1	175 mostres conjunt1	83 mostres conjunt1	Mean centering	5	10 regions	0.75
			Autoscale	8	10 regions	0.80
				5	10 regions	0.81

MODEL 2	151 mostres conjunt1	107 mostres conjunt1	Autoscale	5	10 regions	0.83
				5	6 regions	0.85
				3	6 regions	0.78
				4	6 regions	0.82
			Mean centering	5	6 regions	0.75
		Conjunt2	Autoscale	5	6 regions	0.81
		Conjunt3	Autoscale	5	6 regions	0.78
MODEL 3	129 mostres conjunt1	129 mostres conjunt1	Autoscale	5	6 regions	0.82
MODEL 4	185 mostres conjunt 1	73 mostres conjunt1	Z-score	5	6 regions	0.82
				4	6 regions	0.80
MODEL NOESY	151 mostres conjunt1	107 mostres conjunt1	Autoscale	5	6 regions	0.83

Taula 2: resultats més rellevants dels diferents models que s'han realitzat durant el transcurs d'aquest treball.

El model amb millor predicció és el model 2 amb 5 Variables latents i amb les 6 regions més importants de l'espectre. Aquest model és el que s'ha validat amb el conjunt 2 i el 3, donant una predicció del 80% aproximadament en els dos casos. A conseqüència d'això, es va provar de fer un model amb els 3 tipus de conjunts però no millorava el model actual i per tant es va descartar.

Tal i com es mostra a la taula 2, el model Noesy té una bona predicció per als mateixos espectres del conjunt 1 utilitzant el pols Noesy. No obstant això, en intentar validar el model amb els espectres Noesy dels altres conjunts, la predicció més alta que es va obtindré va ser del 50% i ja no es va entrar en més detall per a poder millorar aquest resultat.

Les VL s'han escollit a partir de la gràfica de l'acumulació de variància capturada que es troba a sota (figura 9) on cada punt representa el nombre de variables latents, el primer és 1VL, el segon equival a dues variables latents, etc.

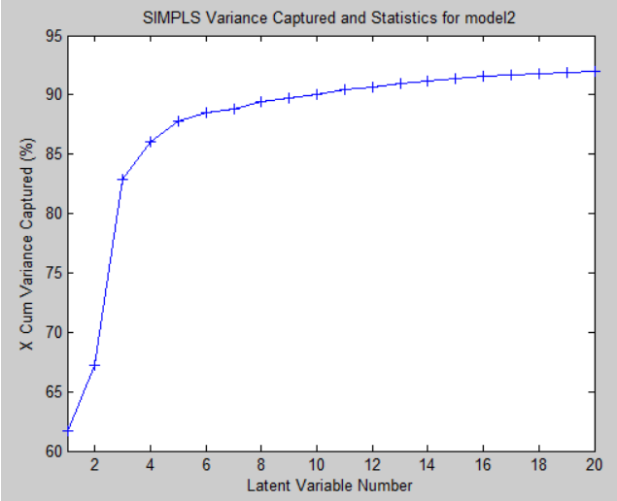
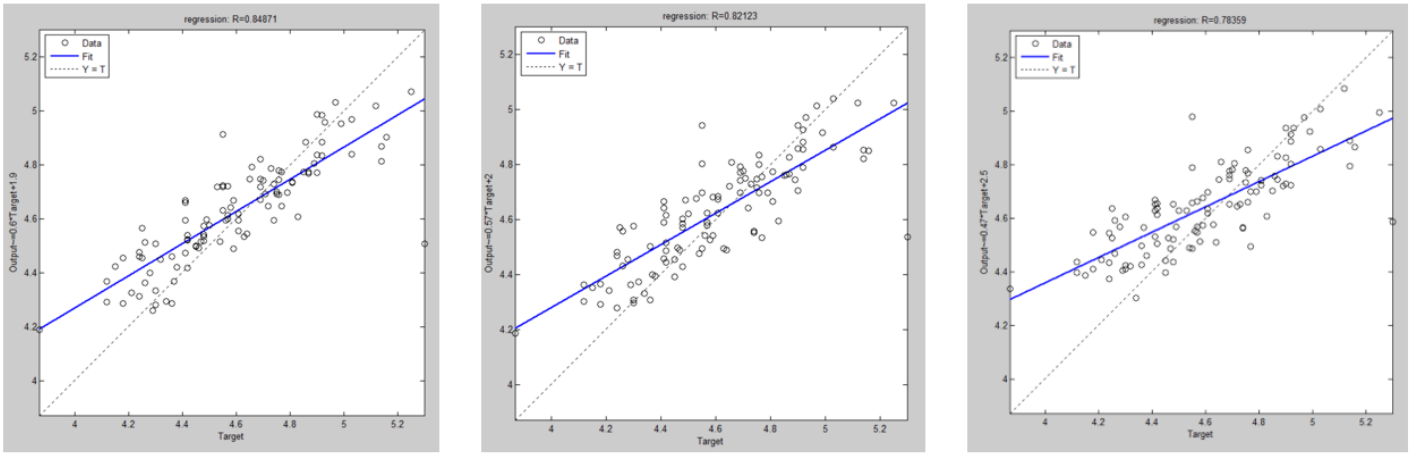


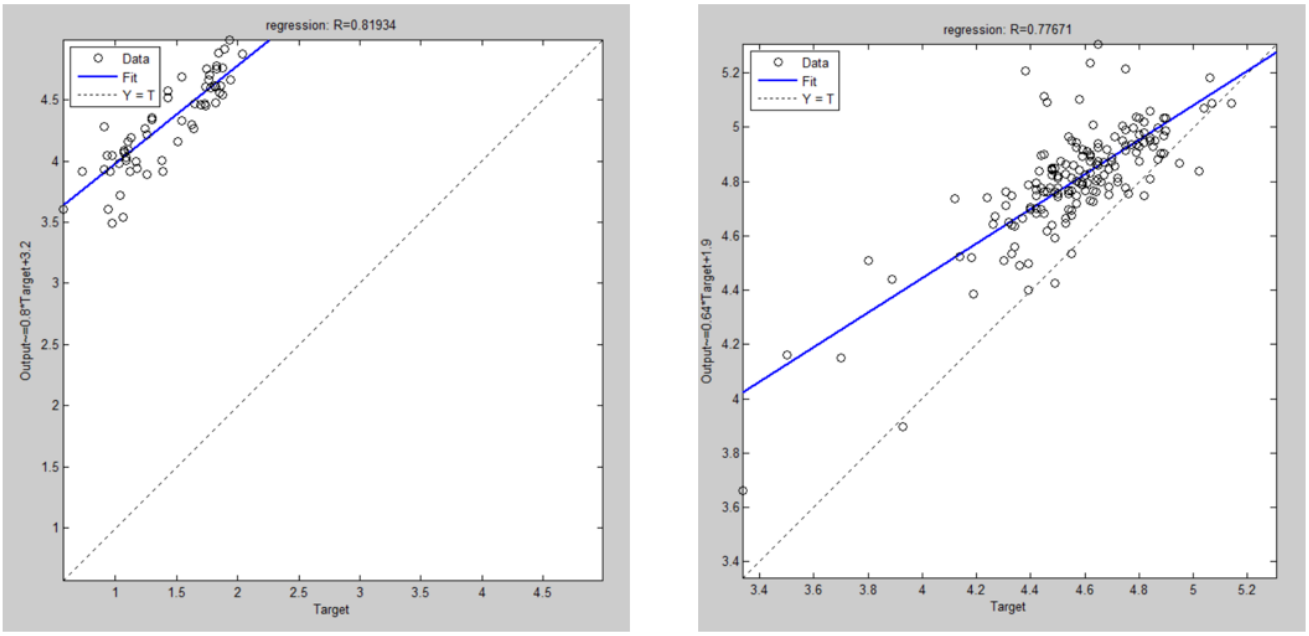
Figura 9: Gràfica de l'acumulació de variància capturada del Model 2 on es pot veure que a partir de la quinta variable latent, el model comença a estancar-se.

A continuació, veurem les diferents regressions resultants del Model 2 i el Model Noesy, descrits en la Taula 2, per tenir una visió més gràfica dels resultats.



a) b) c)

Figura 10: Regressió Mo utilitzant el 2 entre la predicció i el valor real de la concentració d'albúmina segons el nombre de variables latents utilitzades, a) 5VL, b) 4VL i c) 3VL.



a) b)

Figura 11: Regressió del Model 2 amb 6 regions i 5 VL entre la predicció d'albúmina del conjunt 2 i 3 i el valor real d'aquests, on la gràfica de l'esquerra és la comparació en el conjunt 2 i la gràfica de la dreta és la del conjunt 3.

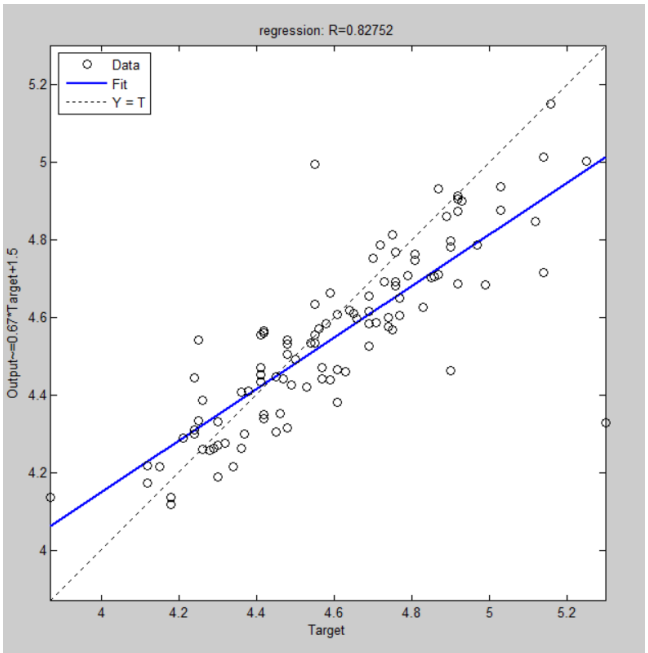


Figura 12: Regressió del Model Noesy entre la predicció d'aquest i el valor real d'albúmina

4 Discussió

Com ja s'ha parlat anteriorment, l'albúmina és una proteïna molt important per al cos humà amb funcions úniques i imprescindibles per a la salut. És per això que aquest treball té tanta importància ja que, amb un simple espectre de RMN, s'ha aconseguit calcular la concentració d'albúmina de la mostra.

Els estudis anteriors ja han demostrat que els mètodes estadístics multivariables poden ser molt eficaços per a la predicció de lípids i proteïnes en matrius biològiques com el sèrum o el plasma. Els resultats reportats confirmen el potencial de l'anàlisi dels models PLS per desenvolupar models robustos per se aplicats.

Entrant més en detall amb els resultats, podem observar com els diferents pre-processats canvien de forma radical les prediccions. El primer model va servir per fer la tria sobre el pre-processat i, es veu clarament que l'*auto scale* dóna millors prediccions que el *mean centering* és per això que en aquest pre-processat només hi ha un parell de proves. En el model 2 es va acabar de confirmar aquestes sospites i ja es va decidir utilitzar *auto scale* en la resta. Abans de concloure definitivament el tipus de pre-processat que s'utilitzaria, es va fer una prova amb el *z-score* (Model 4) per veure si els resultats que ja teníem milloraven notablement. Finalment, es va veure que no hi havia molta millora dels resultats amb *auto scale* comparats amb els de *z-score,* per tant, es va decidir realitzar els models amb el pre-processat *auto scale* perquè donaven millors resultats i es feia automàticament amb MATLAB.

També cal afegir que les zones que agafes segons els STOCSY són importants per la realització d'un bon model PLS. Segons els STOCSY dels resultats, podem dir que els triglicèrids totals de les mostres i les variables de Liposcale, influeixen a l'hora d'obtenir correlacions d'albúmina en l'espectre. Així ha quedat demostrat amb els STOCSY del conjunt 1 on el grup casos i controls, tot i pertànyer a la mateixa base de dades i provenir del mateix grup, un correlava igual que la resta de grups del conjunt i l'altre s'ha eliminat perquè no s'obtenia la correlació adequada per a l'albúmina i no es podia extreure informació.

Les VL són una combinació lineal de les variables originals, en aquest cas, els punts que componen l'espectre NMR, de manera que ens interessa un resultat òptim amb un nombre reduït de VL, per evitar un sobreentrenament del model . És per això que en el model 2 hi ha diferències en quan a rangs utilitzats i el nombre de variables latents que s'utilitzen. És una evidència que els intervals de l'espectre influeixen amb la predicció i així es demostra amb els resultats. Hi ha una regressió més bona amb el model 2 quan s'utilitzen només els 6 rangs importants detectats amb prova i error, que no utilitzant tots els rangs. En conseqüència, el criteri per seleccionar els diferents rangs és que tinguin informació complementària entre ells i, dintre d'aquests 10 rangs, hi ha 4 que no el compleixen. En quan a les variables latents, s'ha provat el mateix model amb diferents VL per veure si reduint les VL es podia aconseguir un resultat semblant que amb 5.

A partir de la gràfica de la Figura 9, es pot veure com en el Model 2, al passar de 3 variables latents a 5, només augmenta un 5% més de variància. Això indica que, entre la 3a i la 5a variable latent només s'explica un 5% de la variabilitat total de les mostres. Per tant, es va decidir implementar el mateix model amb diferents VL per veure si els resultats no variaven en excés a l'hora de reduir aquestes variables, ja que l'objectiu és tenir la millor predicció amb el menor nombre de VL possible, tal i com ja s'ha esmentat anteriorment. Es pot veure a la taula 2 com el canvi de 3 a 5 VL és considerable (increment de 0,78 a 0,84 en la regressió), però, de 4 a 5 només varia d'un 82% de predicció a un 85%, aproximadament. Després de tot, es va decidir que 5 variables no eren gaires i per tant, el model 2 amb aquest nombre de VL és el més òptim que s'ha obtingut.

Seguint amb el mateix model, podem veure en la Figura 10 les regressions amb el diferent nombre de VL. En aquestes gràfiques s'observen certes mostres amb uns nivells d'albúmina no comparables a la resta, ja sigui per que són les mostres amb major nivells d'albúmina o pel cas contrari. El model 2 s'ha entrenat amb un conjunt específic, i tot i escollir les mostres per al training set de forma aleatòria, es sap que no s'han agafat els mínims i els màxims de la concentració d'albúmina. És per això, que la predicció d'aquestes mostres (que són poques) no és acurada. Com a possible millora sobre aquest model seria utilitzar, per al training set, mostres amb uns nivells d'albúmina majors i menors als que hi ha actualment. Si això no és possible, s'ha de contemplar la possibilitat de haver de fer un model específic per a aquestes mostres amb uns nivells d'albúmina més extrems o mirar la possibilitat de contra restar aquest problema dintre del mateix model.

No obstant, el model 2 segueix essent un model robust ja que en provar-lo amb mostres que no ha vist mai com el conjunt 2 i 3, no dona uns resultats dolents. De fet, la validació amb tots dos conjunts dóna una regressió aproximadament de R=0.8. Veient aquests resultats, es va intentar realitzar un model amb mostres dels 3 conjunts però no va resultar factible perquè empitjorava el model actual. També es va intentar treure el conjunt 2 i fer un model amb les mostres del conjunt 1 i del conjunt 3 que són més semblants, però seguia sense donar uns resultats millors que els que ja es tenien.

Veient els bons resultats que s'han obtingut amb els espectres LED, s'ha intentat veure què passa si en comptes de tenir espectres LED, es tenen espectres CMPG o Noesy. En fer els respectius STOCSY, en els CMPG es va veure com només ressonaven les zones dels lípids que, precisament és el que es veu amb els LED i, per tant, es va deixar córrer. Amb els Noesy, la cosa canvia perquè en realitzar el seu STOCSY, les zones d'interès eren semblants a les zones que obteníem amb el LED. D'aquesta manera es va crear un model Noesy d'igual forma que s'havia fet amb el model 2. Tal i com es pot veure en la taula 2 dels resultats, el model Noesy va donar un resultat prou bo com per discutir si val la pena o no elaborar un estudi per a aquests tipus d'espectres i intentar millorar el model Noesy que havia sortit. Finalment, en validar aquest model amb mostres Noesy del conjunt 3, es va veure que no era un model robust tot i la seva predicció inicial amb mostres del conjunt 1. De totes formes, podria ser un inici per intentar veure la possibilitat de predir l'albúmina o altres metabòlits amb aquests tipus d'espectre.

Per últim, cal afegir que per comprovar que en el model 2 que s'ha agafat com a òptim no hi ha overfitting, per part del pre-processat o per l'elecció de nombre de variables latents, es va realitzar un model amb la meitat de mostres al training set i l'altra meitat al test set. Aquest model és el model 3 de la taula 2 dels resultats i s'ha utilitzat el mateix nombre de VL i l'*auto scale* com al model 2 per veure si podíem fiar-nos dels resultats obtinguts. Aquest model 3 va aconseguir una predicció del 82%, únicament amb la meitat de les mostres. Això demostra que no hi ha overfitting possible i que, una vegada més, el model 2 amb 5 variables latents i 6 regions d'interès és robust i funciona bé.

5 Conclusions

En aquest estudi s'ha avaluat el rendiments dels models de predicció PLS basats en l'espectroscòpia 1H-NMR. Es conclou que la ressonància magnètica nuclear permet una predicció robusta de l'albúmina, sent una tècnica que, tot i seguir en estudi, ofereix avantatges en front a altres tècniques de quantificació, com ja s'ha comentat. En vista de l'anterior, és de gran importància tenir una predicció dels nivells d'albúmina sense haver de fer-ho mitjançant processos bioquímics, ja que s'estalvia temps i recursos. A més, és interessant saber la concentració d'albúmina d'una mostra per les característiques i funcions que te aquesta proteïna, permeten una millora en la prevenció de malalties que es puguin tenir.

En resum, els models de regressió dels PLS podrien ser utilitzats com una eina general per quantificar l'albúmina en estudis metabolòmics en matrius de plasma. Per altra banda, també seria de gran importància anar més enllà en els models de predicció perquè el PLS és el més senzill de tots i, d'aquesta manera, potser s'aconseguiria una millora de la predicció d'albúmina que, fins ara, està en un 85% aproximadament.

En conclusió, es pot confirma de forma objectiva que s'ha aconseguit un model de predicció PLS robust per als espectres LED de 1H-NMR que, donades les variables de Liposcale, obté un resultat sobre la concentració d'albúmina de la mostra amb una alta correlació amb les mesures bioquímiques, assolint així l'objectiu principal del treball.

6 Referències

[1] S. Sugio, A. Kashima, S. Mochizuki, M. Noda, and K. Kobayashi, "Crystal structure of human serum albumin at 2.5 Å resolution," *Protein Eng.*, vol. 12, no. 6, pp. 439–446, 1999, doi: 10.1093/protein/12.6.439.

[2] S. Li, Y. Cao, and F. Geng, "Genome-Wide Identification and Comparative Analysis of Albumin Family in Vertebrates," *Evol. Bioinforma.*, vol. 13, 2017, doi: 10.1177/1176934317716089.

[3] M. Poór *et al.*, "Interaction of citrinin with human serum albumin," *Toxins (Basel).*, vol. 7, no. 12, pp. 5155–5166, 2015, doi: 10.3390/toxins7124871.

[4] C. Pimentel *et al.*, "Human pleural fluid and human serum albumin modulate the behavior of a hypervirulent and multidrug-resistant (MDR) acinetobacter Baumannii representative strain," *Pathogens*, vol. 10, no. 4, pp. 1–13, 2021, doi: 10.3390/pathogens10040471.

[5] M. Forsthuber *et al.*, "Albumin is the major carrier protein for PFOS, PFOA, PFHxS, PFNA and PFDA in human plasma," *Environ. Int.*, vol. 137, no. September 2019, p. 105324, 2020, doi: 10.1016/j.envint.2019.105324.

[6] U. K. P. Application, A. Gb, G. B. Ag, H. Ch, and B. Wade, "UK Patent Application „9,GB," vol. 1986, no. 8608216, 1986.

[7] L. Slater, P. M. Carter, and J. R. Hobbs, "Technical Bulletin No. 34. Measurement of albumin in the sera of patients.," *Ann. Clin. Biochem.*, vol. 12, no. 1, pp. 33–40, 1975, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11592272

[8] R. Barrilero *et al.*, "Design and evaluation of standard lipid prediction models based on 1H-NMR spectroscopy of human serum/plasma samples," *Metabolomics*, vol. 11, no. 5, pp. 1394–1404, 2015, doi: 10.1007/s11306-015-0796-5.

[9] P. Soininen, A. J. Kangas, P. Würtz, T. Suna, and M. Ala-Korpela, "Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics," *Circ. Cardiovasc. Genet.*, vol. 8, no. 1, pp. 192–206, 2015, doi: 10.1161/CIRCGENETICS.114.000216.

[10] O. Beckonert *et al.*, "Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts.," *Nat. Protoc.*, vol. 2, no. 11, pp. 2692–2703, 2007, doi: 10.1038/nprot.2007.376.

[11] R. B. Regadera, "Development of ^{1}H-NMR Serum Pro fi ling Methods for High-Throughput Metabolomics," 2017.

[12] J. D. Bell, J. C. C. Brown, G. Kubal, and P. J. Sadler, "NMR-invisible lactate in blood plasma," *FEBS Lett.*, vol. 235, no. 1–2, pp. 81–86, 1988, doi: 10.1016/0014-5793(88)81238-9.

[13] W. J. Bligh, E.G. and Dyer, "Canadian Journal of Biochemistry and Physiology," *Can. J. Biochem. Physiol.*, vol. 37, no. 8, 1959.

[14] R. Barrilero *et al.*, "Unravelling and Quantifying the 'nMR-Invisible' Metabolites Interacting with Human Serum Albumin by Binding Competition and T2 Relaxation-Based Decomposition Analysis," *J. Proteome Res.*, vol. 16, no. 5, pp. 1847–1856, 2017, doi: 10.1021/acs.jproteome.6b00814.

[15] J. Moreno-Vedia *et al.*, "Triglyceride-Rich Lipoproteins and Glycoprotein A and B Assessed by 1H-NMR in Metabolic-Associated Fatty Liver Disease," *Front. Endocrinol. (Lausanne).*, vol. 12, no. January, pp. 1–11, 2022, doi: 10.3389/fendo.2021.775677.

[16] "H-NMR METABOLOMICS RESULTS REPORT".

[17] O. Cloarec *et al.*, "Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic 1H NMR data sets," *Anal. Chem.*, vol. 77, no. 5, pp. 1282–1289, 2005, doi: 10.1021/ac048630x.

[18] H. C. Keun *et al.*, "Heteronuclear 19F-1H statistical total correlation spectroscopy as a tool in drug metabolism: Study of flucloxacillin biotransformation," *Anal. Chem.*, vol. 80, no. 4, pp. 1073–1079, 2008, doi: 10.1021/ac702040d.

[19] C. J. Sands, M. Coen, T. M. D. Ebbels, E. Holmes, J. C. Lindon, and J. K. Nicholson, "Data-driven approach for metabolite relationship recovery in biological 1H NMR data sets using iterative statistical total correlation spectroscopy," *Anal. Chem.*, vol. 83, no. 6, pp. 2075–2082, 2011, doi: 10.1021/ac102870u.

[20] C. Goutis, "Partial least squares algorithm yields shrinkage estimators," *Ann. Stat.*, vol. 24, no. 2, pp. 816–824, 1996, doi: 10.1214/aos/1032894467.

[21] S. Wold *et al.*, "The PLS model space revisited," *J. Chemom.*, vol. 23, no. 2, pp. 67–68, 2009, doi: 10.1002/cem.1171.

[22] R. Mallol *et al.*, "Liposcale: A novel advanced lipoprotein test based on 2D diffusion-ordered 1H NMR spectroscopy," *J. Lipid Res.*, vol. 56, no. 3, pp. 737–746, 2015, doi: 10.1194/jlr.D050120.

7 Annex

7.1 T-test amb python

```
#T-test:
import numpy as np
import scipy.stats as stats
control= df[df['Grup']==0].Albumina
casos= df[df['Grup']==1].Albumina
tvalue, pvalue = stats.ttest_ind(casos,control,alternative='two-sided')
print(pvalue)
```

```
0.00389318421955788
```

Figura 13: Codi utilitzat per fer el t-test amb python i el resultat del p-value

7.2 Codi Matlab

```
%Create new data set:
uiopen('C:\Users\gemma\Escritorio\NMR_Metabolomic_Results_J001.xls',1)%load the .csv or .xls file
Espectres = dataset(transpose(data(:,[2:end]))); %create the new data set
Espectres.label{1} = textdata(1,[2:end]); %Add the name of the samples (row)
Espectres.axisscale{2} = data(:,1); %Add the axisscale to the columns

%Plot diferents samples with diferents colors:
plot(Espectres.axisscale{2},Espectres.data([1:19 21:29 31:60 62 63 65],:), 'b');
hold on
plot(Espectres.axisscale{2},Espectres.data([30 61 64],:), 'r');
hold off

%Plot regression:
plotregression(Albumina.data([1:19 21:end],:),pls_pred.pred{2}([1:19 21:end]),'regression');

%Stocsy:
stocsy(Espectres.data(:,:), Albumina.data(:,:), Espectres.axisscale{2}, 'title');

%Get Rang
rang1 = getRang(Espectres.axisscale,1.4,1);

%Select the points that are in the ranges of interest
model.include{2} = [rang1 rang3 rang4 rang7 rang9 rang10];

%Create a new data set from other data sets
model2 = [model;predict];

%Create a new data set from samples of other data set
Espectress = dataset(Espectres.data([4:6 8 9 253:end],:)); %select the samples and do the data set
Espectress.label{1} = Espectres.label{1}([4:6 8 9 253:end],:); %Add the name of the samples
Espectress.axisscale{2} = Espectres.axisscale{2};%Add the axisscale to the columns

%to create a PLS model, you only have to write in the consol the command
%"pls" and MATLAB will open a new window with the necessary tool to do it.
```