IDENTIFICACIÓ DE BIOMARCADORS PROTEÒMICS PER AL DIAGNÒSTIC NO INVASIU DE LA MASLD: METAANÀLISI PROTEÒMICA I APRENENTATGE AUTOMÀTIC

Mario Sánchez Álvarez

TREBALL FINAL DE GRAU DE BIOTECNOLOGIA

Tutor acadèmic: Marta Sanchis Talón

En cooperació amb: Insitut Investigació Sanitària Pere Virgili (IISPV)

Supervisores: Elena Cristinia Rusu Hutu, elena.cristina.rusu.hutu@gmail.com i Mª Teresa Auguet, tauguet.hj23.ics@gencat.cat Departament de Medicina interna Hospital Joan XXIII, IISPV

Índex

1. Dades del centre 4
2. Resum i paraules claus 5
3. Introducció 6
3.1. Malaltia hepàtica esteatòsica associada a disfunció metabòlica 6
3.1.1. Definició de la malaltia 6
3.1.2. Epidemiologia 6
3.1.3. Història natural 7
3.1.4. Fisiopatologia: teoria del "*multiple hit*" 8
3.1.5. Etiologia i factors de risc 10
3.1.6. Diagnòstic 11
3.1.7. Tractament 15
3.2. Metaanàlisi proteòmica 16
3.2.1. Ciències òmiques 16
3.2.2. L'interès en l'estudi de les proteïnes 16
3.2.3. Què és la proteòmica? 17
3.2.3.1. Espectrometria de masses 18
3.2.3.1.1. Preparació de la mostra 18
3.2.3.1.2. Adquisició de dades 19
3.2.3.1.3. Quantificació 20
3.2.4. Metaanàlisis proteòmiques per al diagnosi de MASLD 21
3.3. Aprenentatge automàtic per a la cerca de biomarcadors 22
4. Hipòtesi i objectius 22
5. Metodologia 23
5.1. Cerca bibliogràfica i selecció d'estudis 23
5.2. Selecció de pacients 24
5.2.1. Metadades 24

5.3. Preprocessat de les dades en cru i anàlisi exploratòria 25
5.4. Anàlisi de l'expressió diferencial de les proteïnes 25
5.5. Metaanàlisi per comparativa 26
5.6. Anàlisi d'enriquiment funcional 26
5.6.1. *Reactome pathway enrichment* 26
5.6.2. *GO Terms enrichment* 26
5.7. Xarxa d'interaccions proteïna-proteïna 26
5.8. Aprenentatge automàtic 27
5.8.1. Definició dels conjunts d'entrenament i test 27
5.8.2. Enginyeria de característiques 28
5.8.3. Selecció inicial de característiques 28
5.8.4. Tests d'algoritmes de classificació 29
5.8.5. Ajust d'hiperparàmetres i selecció de característiques final 29
6. Resultats 30
6.1. Estudis seleccionats 30
6.2. Anàlisi exploratori 30
6.3. Proteïnes diferencialment expressades entre SS i MASH 31
6.4. Anàlisi d'enriquiment funcional 32
6.4.1. Vies metabòliques enriquides en la MASLD 32
6.4.2. Termes de la GO enriquits en les DEPs més significatives 32
6.5. Xarxa d'interaccions proteïna-proteïna 33
6.6. Aprenentatge automàtic 34
7. Discussió 38
8. Conclusions 41
9. Bibliografia 42
10. Autoavaluació 49

Data de convocatòria: Juny 2025

Jo, Mario Sánchez Álvarez , amb DNI “48275398P, soc coneixedor de la guia de prevenció del plagi a la URV Prevenció, detecció i tractament del plagi en la docència: guia per estudiants (aprovada el juliol 2017) (http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formaciocompetencies-nuclears/plagi/) i afirmo que aquest TFG no constitueixen cap de les conductes considerades com a plagi per la URV.

Tarragona, 5 de juny de 2025

Mario Sánchez

Digitally signed by Mario Sánchez
Date: 2025.06.05 12:19:43 +02'00'

1. Dades del centre

L'Institut d'Investigació Sanitària Pere Virgili (IISPV) és un institut d'investigació biomèdica situat a la província de Tarragona, sent un centre de referencia tant a nivell nacional com internacional. El seu objectiu és promoure, desenvolupar i difondre la investigació, el coneixement científic i tecnològic, la docència i la formació en l'àmbit de les ciències de la vida i de la salut.

El Grup d'Estudi de Malalties Metabòliques Associades a Insulina-Resistència (GEMMAIR) ha permès la realització d'aquest Treball Final de Grau. Forma part de l'àrea de Nutrició i Metabolisme dins l'IISPV. El seu principal objectiu és l'estudi de la fisiopatologia de l'obesitat, la síndrome metabòlica i la hepatopatia greixosa associada. Concretament, estudien amb detall la malaltia del fetge gras no alcohòlic amb la finalitat de trobar noves dianes terapèutiques mitjançant la investigació de gens que podrien estar implicats i la cerca de biomarcadors de la malaltia.

GEMMAIR és un grup multidisciplinari format per professionals clínics de l'Hopsital Joan XXIII (HJ23) de Tarragona i l'Hospital Sant Joan de Reus (HSJR), i personal tècnic experimental de la Universitat Rovira i Virgili (URV). En l'actualitat, el grup està dirigit per la Dra. Maria Teresa Auguet, cap del servei de Medicina Interna de l'HJ23, investigadora de l'IISPV i professora de la URV.

2. Resum i paraules claus

La malaltia hepàtica esteatòsica associada a disfunció metabòlica (MASLD) presenta una prevalença elevada a escala mundial, afectant aproximadament el 30% de la població adulta, amb una proporció considerable que progressa cap a esteatohepatitis metabòlica (MASH). El diagnòstic actual es basa principalment en tècniques invasives com la biòpsia hepàtica, que presenta limitacions importants, incloent-hi riscos per al pacient i baixa capacitat de generalització.

Davant d'aquesta situació, l'objectiu principal d'aquest estudi és identificar biomarcadors proteòmics que permetin diferenciar, de forma no invasiva, entre els estadis histopatològics d'esteatosi simple (SS) i MASH. Per assolir aquest objectiu, s'ha realitzat una cerca sistemàtica que ha permès identificar tres estudis aptes per a la seva inclusió en una meta-anàlisi proteòmica, amb un total de 169 pacients procedents de cohorts diferents, incrementant així la robustesa i generalització dels resultats obtinguts.

L'anàlisi d'expressió diferencial resultant ha identificat sis proteïnes diferencialment expressades (DEPs) significatives entre els grups de SS i MASH. Les anàlisis funcionals posteriors revelen alteracions destacades en vies metabòliques associades principalment al sistema del complement, així com al transport intracel·lular i al metabolisme.

Finalment, mitjançant tècniques d'aprenentatge automàtic, s'ha desenvolupat un model predictiu basat en 15 associacions proteòmiques, que mostra un rendiment prometedor per diferenciar de manera fiable i no invasiva entre SS i MASH. Tanmateix, calen futures millores en la precisió d'aquest model abans de la seva implementació clínica definitiva.

Paraules clau: MASLD, biomarcadors proteòmics, metaanàlisi, aprenentatge automàtic, diagnòstic no invasiu.

3. Introducció

3.1. Malaltia hepàtica esteatòsica associada a disfunció metabòlica

3.1.1. Definició de la malaltia

La malaltia del fetge gras no alcohòlic (NAFLD, de l'anglès, *non-alcoholic fatty liver disease*) ha esdevingut la malaltia crònica de fetge més comú. Es caracteritza per la presència d'esteatosi (acumulació de greix) en més del 5% dels hepatòcits, juntament amb factors de risc metabòlic, com la diabetis mellitus tipus 2 (T2DM, de l'anglès, *type 2 diabetes mellitus*), el sobrepès o l'obesitat, la dislipèmia (alteració dels nivells de lípids en sang), la hipertensió, i/o alguna altra desregulació metabòlica evident, excloent el consum en excés d'alcohol o altres malalties cròniques del fetge (1).

La NAFLD engloba un ampli rang de condicions, des de la més benigna, l'esteatosi simple o SS (de l'anglès, *simple steatosis*); fins a l'esteatohepatitis o NASH (de l'anglès, *non-alcoholic steatohepatitis*), que és la més severa dins de l'espectre de patologies (2). Des d'aquest estadi, la malaltia pot evolucionar cap a fibrosi i, en els casos més severs, cap a cirrosi i hepatocarcinoma (HCC, de l'anglès, *hepatocellular carcinoma*) que, al seu torn, pot arribar a ocasionar la mort (3).

Al 2023, un consens multisocietat va proposar un nou terme per descriure la malaltia. El motiu principal d'aquest va ser que l'antic terme no englobava correctament a tots els pacients i que els termes *non-alcoholic* i *fatty* resultaven estigmatitzants. El nou terme proposat va ser malaltia hepàtica esteatòsica associada a disfunció o MASLD (de l'anglès, *metabolic dysfunction-associated steatotic liver disease*) (4,5). Aquesta nova nomenclatura permet identificar més pacients amb risc de patir qualsevol dels diferents estadis de la malaltia, incloent aquells que anteriorment eren exclosos per factors com el consum d'alcohol o la presència d'altres malalties hepàtiques (6). Així mateix, els diagnòstics previs de NAFLD es poden reclassificar com a MASLD per reflectir millor els factors metabòlics subjacents a la patologia (7).

3.1.2. Epidemiologia

En les darreres dècades, la prevalença de la MASLD ha augmentat a escala mundial, i s'estima que afecta a, aproximadament, un 30% de la població global (5). El motiu d'aquest augment està estretament lligat amb l'estil de vida sedentari de la població, incloent els hàbits alimentaris pobres i l'activitat física reduïda (8).

La malaltia és més freqüent en homes (67.9%) i acostuma a ser més comú des dels 18 fins als 75 anys (93.8%). En infants, varia entre un 5 i 10%. Cal destacar que la MASLD té una major prevalença entre persones amb obesitat (74.1%), on el 25.9% pateix de T2DM (9).

Un estudi de Younossi *et al*. evidencia importants variacions regionals en la prevalença de la malaltia (10). La MASLD, com s'observa a la Figura 1, té una prevalença més alta a Amèrica del Nord i del Sud, així com a la regió d'Àsia-Pacífic, Austràlia i Nova Zelanda, Orient Mitjà i Europa, però és menys comú a l'Àfrica. Aquest trastorn és freqüent en països amb recursos, la qual cosa s'atribueix a l'àmplia disponibilitat d'aliments obesogènics i altament calòrics, a preus molt assequibles (10).

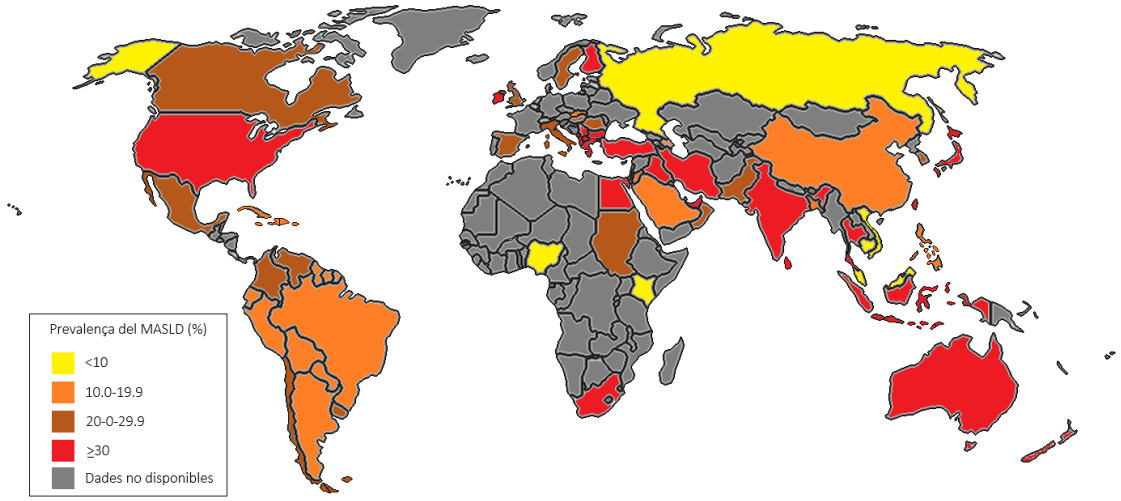

Figura 1. Prevalença mundial de la MASLD. Adaptada de (10). Feta amb *BioRender*.

3.1.3. Història natural

La MASLD inclou principalment dues fases: la SS i l'esteatohepatitis associada a disfunció metabòlica o MASH (de l'anglès, usant la nova terminologia, *metabolic dysfunction-associated steatohepatitis*). La primera es caracteritza per l'acumulació de greix en més del 5% dels hepatòcits, sense evidència d'un dany hepatocel·lular, i es pot considerar un estadi relativament benigne (1). Per contra, quan la malaltia no es tracta, pot avançar cap a la fase de MASH, on la presència d'esteatosi és major i apareix inflamació amb dany hepatocel·lular considerable, presentant balonització cel·lular (11) i podent exhibir fibrosi hepàtica (12). Per una banda, la balonització consisteix en la inflamació dels hepatòcits i l'arrodoniment del seu citoplasma (13); mentre que la fibrosi

és l'acumulació excessiva de teixit connectiu, generalment col·lagen, a la matriu extracel·lular (ME) dels hepatòcits (14). L'estadi de MASH és considerat molt perillós ja que en el 10-15% dels pacients (11) evolucionarà cap a cirrosi i HCC. Estudis publicats en els darrers anys suggereixen que la mortalitat de la MASLD augmenta exponencialment a mida que l'estadi de fibrosi avança (10).

La Figura 2 mostra el desenvolupament de la MASLD i els diferents estadis que pot presentar.

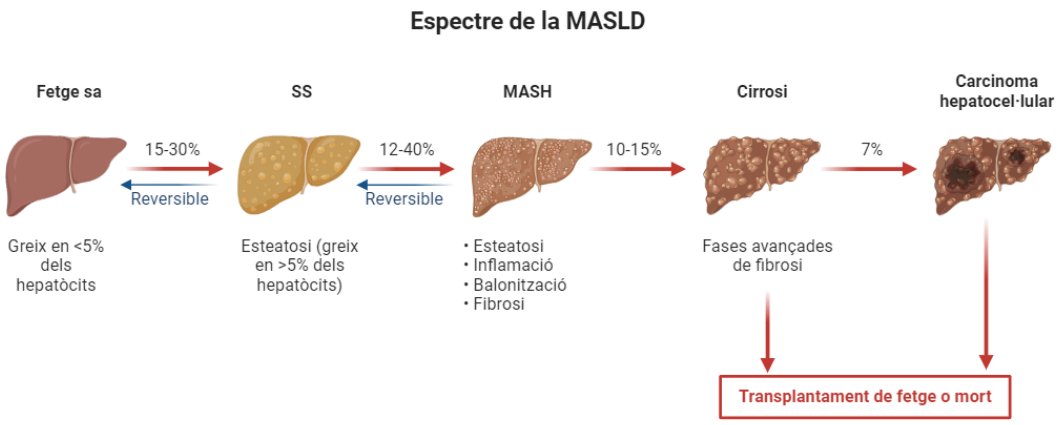

Figura 2. Espectre de la MASLD. Adaptada de (15). Feta amb *BioRender*.

3.1.4. Fisiopatologia: teoria del "*multiple hit*"

El progrés de la malaltia cap a MASH és un procés complex i encara no del tot conegut. En els últims anys, gràcies a la investigació amb models animals, s'ha suggerit que aquest procés està format per dues etapes: la primera consisteix en l'acumulació de greix al fetge, que en última instància provoca resistència a la insulina (RI); mentre que la segona inclou canvis moleculars i cel·lulars que inclouen estrès oxidatiu (EO) i l'oxidació d'àcids grassos (AGs) (2). Això és el que es coneix com la teoria del "*multiple hit*". Aquesta és l'evolució d'una teoria anterior anomenada teoria del "*second hit"*. El primer *hit* es considerava que era l'aparició de RI i d'SS degut a l'acumulació d'AGs al fetge; mentre que el segon era la presència de dany cel·lular, inflamació, fibrosi i altres alteracions patològiques degut a l'EO. La teoria del "*multiple hit*" explica més detalladament el perquè de les conseqüències de la teoria del "*second hit*" (16).

La Figura 3 mostra un resum d'aquest complex procés.

En primer lloc, l'**acumulació de lípids** ve ocasionada per la síntesi en excés de triglicèrids (TGLs). En aquest escenari, la RI té un paper molt important, ja que, quan es dóna, els IRS-2 (de l'anglès, *insulin receptors substrate* 2), que regulen la SREBP-1c (de l'anglès, *sterol regulatory element-binding protein-1*), es troben desactivats. Això provoca una sobreexpressió de la SREBP-1c , que al seu torn promou la ruta de la lipogènesi *de novo* (DNL, de l'anglès, *de novo lipogenesis*), responsable de la producció d'àcids grassos lliures (FFAs, de l'anglès, *free fatty acids*). Aquesta ruta pot ser també activada per l'excés de glucosa present en la dieta. Per altra banda, la RI també incrementa la lipòlisi del teixit adipós blanc, donant lloc a més FFAs. Aquests són el substrat per a la síntesi de TGLs (16,17).

En segon lloc, l'**EO** s'origina principalment per la producció d'espècies reactives d'oxigen (ROS, de l'anglès, *reactive oxygen species*). Aquestes ROS es generen, sobretot, de dues maneres: D'una banda, quan hi ha un excés de FFAs, aquests poden entrar als mitocondris, on són metabolitzats mitjançant la β-oxidació per produir acetil-CoA. Aquest acetil-CoA és posteriorment utilitzat en el cicle de Krebs per generar energia. Durant aquest procés metabòlic, alguns electrons són transferits a la cadena transportadora d'electrons mitocondrial. Tanmateix, en condicions d'excés AGs, aquests electrons poden reaccionar amb molècules d'oxigen, generant ROS. Aquestes ROS poden oxidar els lípids cel·lulars, formant peròxids lipídics que activen una resposta inflamatòria. D'altra banda, les ROS també poden ser generades per la lipòlisi dels TGLs. En aquest procés intervenen microsomes i peroxisomes (16).

En tercer lloc, l'**estrès del reticle endoplasmàtic** també és ocasionat per l'excés de FFAs. Aquest estrès provoca que es sintetitzin proteïnes mal plegades, fet que desencadena una resposta contra aquestes (UPR, de l'anglès, *unfolded protein response*) (16). Els factors que indueixen aquesta resposta inclouen hiperglucèmia, dany mitocondrial, augment de colesterol i EO. L'UPR també activa la SREBP-1c (17).

Per últim, la **lipotoxicitat** és el resultat de la generació de lípids lipotòxics a partir dels FFAs. Aquests agreugen les condicions dins els mitocondris, l'EO i la inflamació. Aquesta inflamació indueix a les cèl·lules de Kupffer a secretar citocines proinflamatòries, per una banda; i a les cèl·lules estelades hepàtiques a produir fibrosi hepàtica, per l'altra. Aquestes últimes són la principal causa de la progressió de MASH (16).

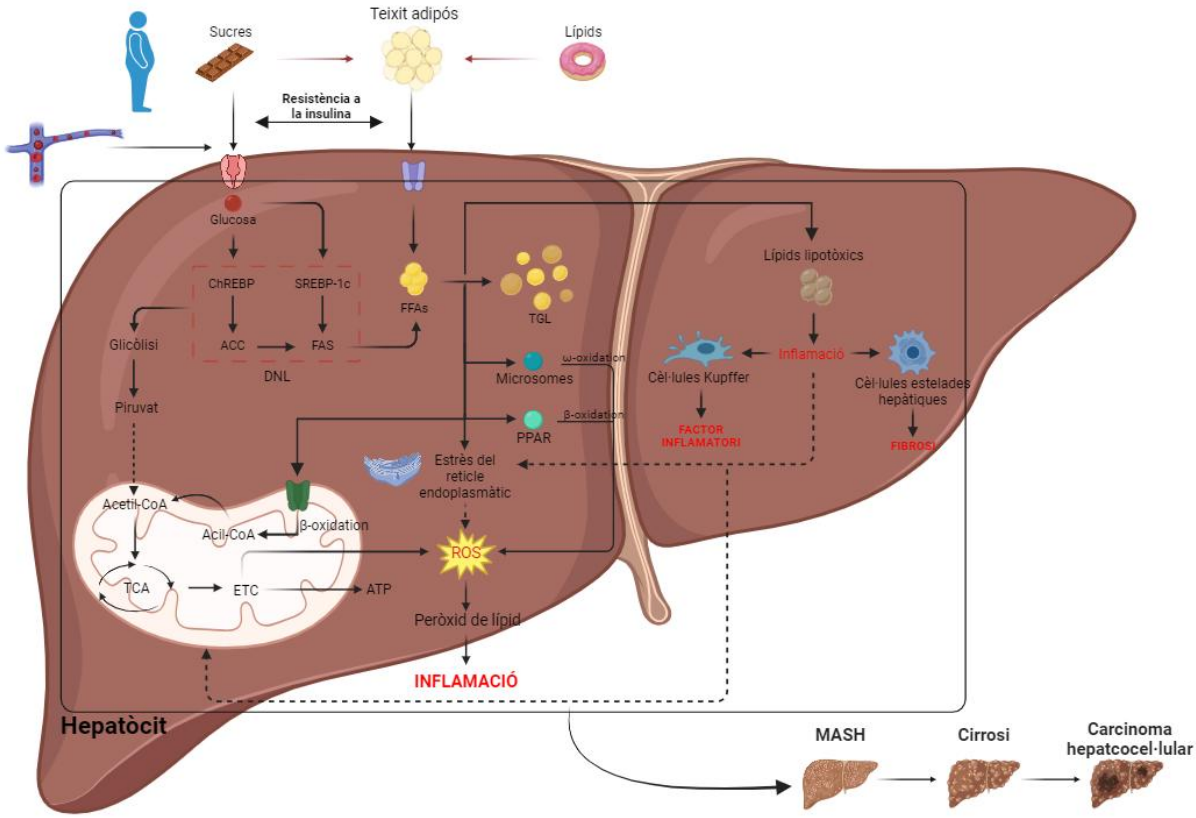

Figura 3. Esquema de la fisiopatologia de la MASLD. Adaptada de (15). Feta amb *BioRender*. ChREBP: *carbohydrate response element-binding protein,* SREBP-1c: *sterol regulatory element-binding protein-1c*, ACC: *acetyl-CoA carboxylase*, FAS: *fatty acid synthase*, DNL: *de novo lipogenesis*, FFAs: *free fatty acids*, TCA: *tricarboxylic acid cycle*, ETC: *electron transport chain*, PPAR: *proliferator-activated receptor*.

3.1.5. Etiologia i factors de risc

La MASLD és una malaltia multifactorial ja que el seu desenvolupament i progressió són causats per la combinació de factors genètics, ambientals i estil de vida (18). Les persones que pateixen la malaltia acostumen a presentar característiques de la síndrome metabòlica (SM), com són l'obesitat, la T2DM, la RI o la dislipèmia (19), les quals són considerades importants factors de risc. A més, estudis han demostrat l'increment de la prevalença de malalties cardiovasculars en pacients que pateixen la MASLD (2).

Estudis genòmics han identificat polimorfismes de nucleòtid únic en alguns gens dins de diferents cohorts amb MASLD. Per una banda, el gen *PNPLA3* codifica per la proteïna adiponectina, la qual participa en la hidròlisi de TGLs. Quan presenta el polimorfisme rs738409 C>G, el procés d'hidròlisi s'inhibeix provocant l'acumulació lipídica hepàtica promovent la MASLD (10,20). Per altra banda, el polimorfisme rs58542926 C>T del gen *TM6SF2* està associat a estadis avançats de fibrosi i cirrosi (10).

Els factors ambientals tenen un paper clau en el desenvolupament de la malaltia, especialment els hàbits alimentaris, l'activitat física i els factors socioeconòmics. La majoria de pacients tendeixen a consumir aliments poc nutritius i calòrics, , i a portar una vida sedentària (10), fet que contribueix a l'aparició de RI, component clau de la SM (21), i alteracions en la microbiota intestinal (disbiosi intestinal). Aquesta última pot augmentar la permeabilitat intestinal, facilitant la translocació sistèmica de bacteris i la inflamació intestinal i hepàtica (22). A més, factors com la cultura, l'educació i els ingressos també estan fortament associats al risc de patir MASLD (10).

Altres factors determinants inclouen els epigenètics, que poden provocar la metilació de gens associats a la fibrosi, induint així aquest procés (10); i l'exposició a contaminants i microplàstics, que afavoreixen l'augment de TGLs en sang i l'acumulació de greix al fetge (23,24).

3.1.6. Diagnòstic

Amb la nova terminologia, el diagnòstic de MASLD, tal com es mostra en la Figura 4, recau principalment en la detecció d'esteatosi juntament amb almenys 1 dels 5 criteris cardiometabòlics: inicis de desregulació dels nivells de glucosa, T2DM, sobrepès/obesitat, hipertensió o dislipèmia (7). Tanmateix, distingir amb precisió en quin estadi de la malaltia es troba un pacient és clau per tal de poder tractar-lo correcta i eficaçment. De moment, la biòpsia de fetge és la tècnica que ho permet fer amb major exactitud. Per contra, té limitacions inherents pel fet de ser una tècnica invasiva que comporta riscos per a la salut del pacient (3,25). És per això que també s'utilitzen altres tècniques no invasives, no tan sensibles, com són l'estudi d'imatges abdominals o l'anàlisi de biomarcadors en sèrum (3).

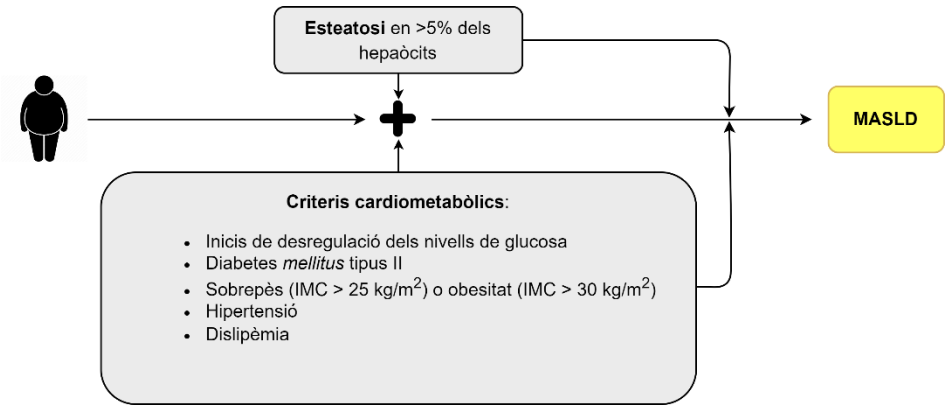

Figura 4. Criteri per diagnosticar la MASLD.

El diagnòstic a través d'una **ecografia** abdominal és una tècnica no invasiva, econòmica, lliure de radiació i àmpliament disponible. Per contra, únicament pot detectar la presència d'esteatosi hepàtica quan aquesta supera un 30%. És per això que només permet diagnosticar MASLD, és a dir, no permet distingir si el pacient es troba en l'estadi de SS o MASH. Com ja s'ha comentat anteriorment, aquesta distinció és determinant ja que pot portar a complicacions més severes com estadis avançats de fibrosis o, fins i tot, HCC (3,26).

Existeixen altres tècniques, com la ressonància magnètica per imatge o per espectroscòpia, que són molt més sensibles i específiques en relació a la quantificació de l'esteatosi, però que es veuen limitades per la seva disponibilitat i cost (3). D'altres, com la elastografia (*FibroScan*), poden mesurar la rigidesa del fetge i detectar estadis de fibrosi (26). Aquesta última té limitacions en pacients amb un índex de massa corporal (IMC) superior a 35 kg/m^2 o que pateixen de T2DM (25). Tanmateix, aquestes tècniques continuen sense ser capaces de distingir entre SS i MASH (3).

Taula 1. Comparació de les diferents tècniques de diagnosi de la MASLD.

Tècnica	Detecció	Distinció	Desavantatges
Ecografia	Imatge	Grau SS	Pacients amb >30% esteatosi hepàtica i només detecta SS
Elastografia	Rigidesa del fetge	Presència de fibrosi	Només detecta fibrosi
Ressonància magnètica	Imatge	Grau SS	Costosa i només detecta SS
Biòpsia	Histologia	Qualsevol estadi de la malaltia	Invasiva, riscs

El diagnòstic a través d'una **biòpsia** de fetge és considerat l'estàndard d'or ja que, de moment, és l'única tècnica que permet distingir clarament els estadis de SS i MASH, a més a més de detectar la presència de fibrosi. Malauradament, és una tècnica amb limitacions, ja que és molt invasiva, amb risc de complicacions que poden, fins i tot, comportar mortalitat. Adicionalment, el diagnòstic a partir d'una biòpsia està estretament relacionat amb l'experiència i coneixement de l'hepatòleg, i la tècnica està subjecta a la variabilitat en l'obtenció de la mostra, ja que aquesta no representa la malaltia de manera global perquè la inflamació i fibrosi poden no estar distribuïdes uniformement (2,26).

El NAS (de l'anglès, *NAFLD Activity Score*) és un sistema de puntuació àmpliament utilitzat pel pronòstic de la MASLD a partir d'una biòpsia mitjançant l'avaluació d'un seguit de criteris: esteatosi (0-3), inflamació lobular (0-3) i balonització hepatocel·lular (0-2). La suma de les puntuacions obtingudes segons cada criteri dóna coma resultat el diagnòstic, que pot ser: SS o sense diagnòstic de MASH (0-2), en el llindar SS-MASH (3,4) o MASH (5-8) (2). Una modificació d'aquest sistema és el que utilitza Kleiner, que afegeix un nou criteri en relació a la fibrosi (0-4), mostrat en la Taula 2 (27). Existeixen altres sistemes també basats en el NAS com el de Bedossa (28), Brunt (27) o la puntuació conjunta d'esteatosi, activitat i fibrosi (SAF) (29).

Taula 2. Puntuacions Kleiner pel diagnòstic de MASLD a partir d'una biòpsia. Adaptada de (27).

Criteri	Grau del criteri observat	Puntuació
Esteatosi	<5%	0
	5-33%	1
	>33-66%	2
	>66%	3
Inflamació lobular	Nul·la	0
	<2 zones	1
	2-4 zones	2
	>4 zones	3
Balonització hepatocel·lular	Nul·la	0
	Alguns hepatòcits arrodonits	1
	Molts hepatòcits amb balonització prominent	2
Fibrosi	Nul·la	0
	Fibrosi pericel·lular o periportal	1
	Fibrosi pericel·lular i periportal	2
	Ponts de fibrosi	3
	Cirrosi	4

L'**anàlisi de biomarcadors** en sang és una de les tècniques que està guanyant més rellevància recentment, gràcies a la seva capacitat per detectar els diferents estadis de la MASLD, inclòs el grau de fibrosi (30). És una tècnica no invasiva que facilita el procés d'anàlisi, i permet donar una visió global de la condició del pacient, a diferència de la biòpsia, que únicament dona una visió reduïda d'una fracció del fetge (25). Per contra, cal destacar que pot no sempre fer un diagnòstic precís. Tanmateix, és una tècnica en

constant desenvolupament i s'espera que algun dia pugui arribar a substituir les biòpsies (31).

De la mateixa manera que en les biòpsies, existeixen sistemes de puntuació basats en biomarcadors per tal de determinar en quin estadi de la malaltia es troba un pacient. La Taula 3 mostra un resum dels diferents biomarcadors que permeten distingir els diferents estadis de la MASLD.

Taula 3 Puntuacions pel diagnòstic dels diferents estadis de MASLD a partir de biomarcadors. Adaptada de (30).

Puntuació	Biomarcadors	Altres variables
Detecció de MASLD		
FLI (de l'anglès, *Fatty Liver Index*)	GGT, TGL	IMC, circumferència de cintura (cm)
MASLD *Screening score*	ALT/AST *ratio*, TGL, glucosa	Edat, IMC
Detecció de SS		
HSI (de l'anglès, *Hepatic Steatosis Index*)	ALT/AST *ratio*	Sexe, T2DM, IMC
SteatoTest	*apha-2 macroglobulin*, *apolipoprotein A1*, haptoglobina, bilirubina, GGT, ALT, TGL, colesterol	Edat, IMC, sexe
Detecció de MASH		
IL-6	IL-6	-
VCAM-1	VCAM-1	-
Citoqueratina 18	Citoqueratina 18	-
Detecció de fibrosi		
ELF (de l'anglès, *Enhanced Liver Fibrosis*)	TIMP1, HA, pèptid aminoterminal del procol·lagen III	-
APRI	APRI	-
FIB-4 (de l'anglès, *fibrosis-4*)	ALT, AST, nombre de plaquetes	Edat
BARD *score* (de l'anglès, *ALT/AST ratio, presence of diabetes and BMI*)	AST, ALT	IMC, diabetis

GGT: *gamma-glutamyl transpeptidase,* ALT: *alanine aminotransferase*, AST: *aspartate transaminase*, IL-6: interleucina-6, VCAM-1: *vascular cell adhesion protein 1*, TIMP1: *tissue inhibitor of matrix metalloproteinase 1*, HA: *hyaluronic acid*, APRI: *AST to platelet ratio index*, BMI: IMC

Aquests biomarcadors s'acostumen a utilitzar com a diagnòstic precoç de la malaltia i com a complement de la biòpsia. Tot i ser una tècnica emergent, encara no s'ha

trobat un panell de biomarcadors el suficientment precís que sigui capaç de distingir amb exactitud els diferents estadis de la MASLD (31).

3.1.7. Tractament

El tractament de la MASLD, juntament amb les malalties derivades d'aquesta i els factors de risc relacionats amb la SM, molt sovint consisteix en intervenir en diferents aspectes de la vida del pacient, com per exemple modificant l'estil de vida o subministrant medicació (2,3). El més comú acostuma a ser modificacions en la dieta i en l'exercici físic. Estudis mostren que la dieta mediterrània, baixa en hidrats de carboni i rica en greixos monosaturats, omega-3, fibra i proteïna, redueix el greix al fetge i ajuda a tractar la MASLD. La font de proteïna és clau, ja que la carn vermella pot incrementar el risc de malalties cardiovasculars. Altres elements com la vitamina E i la cafeïna també contribueixen al tractament, actuant com a antioxidants i disminuint la inflamació hepàtica. A més, la cafeïna redueix el risc de fibrosi (2,3).Per altra banda, fer exercici físic, especialment aeròbic, provoca un augment en el consum de glucosa que permet reduir la RI i evitar l'acumulació de greix al fetge, oferint així una protecció contra la MASLD (3).

Conjuntament amb la dieta i exercici físic, nombrosos estudis estan investigant amb diferents fàrmacs. Aquests acostumen a ser antioxidants, sensibilitzadors d'insulina i hipolipemiants (agents reductors de greix). Es classifiquen en funció de la seva diana terapèutica (3,32). La Taula 4 mostra un resum dels fàrmacs més comuns avaluats en la fase III de diferents assajos clínics.

Taula 4. Fàrmacs pel tractament de la MASLD. Adaptada de (3,32).

Fàrmac	Diana terapèutica	Efecte
Metabolisme lipídic		
Olitpraz	Inhibidor de LXR-α	Reducció de la síntesi d'AGs però augment de l'oxidació de lípids en pacients amb MASLD.
Resmetirom	Agonista selectiu del receptor THR-β	Increment del metabolisme lipídic hepàtic i reducció de la lipotoxicitat en pacients amb MASH.
Metabolisme de la glucosa		
Pioglitazone	Agonista de PPARγ	Sensibilitzador d'insulina que permet tractar la T2DM.
Metabolisme d'àcids biliars		

Àcid obeticòl·lic	Agonista de FXR	Reducció del greix del fetge i la fibrosi en pacients amb MASH.
Estrès oxidatiu, inflamació i fibrosi		
N-acetilcisteïna	Equilibri oxidant-antioxidant	Protector del fetge.
Vitamina E	Estrès oxidatiu	Protecció contra el dany oxidatiu causat pels radicals lliures i la toxicitat mitocondrial.

LXR-α: *liver X receptor- α*, PPARγ: *proliferator-activated receptor γ*, THR-β: *thyroid hormone receptor-β*, FXR: *farnesoid X nuclear receptor*

Tots els fàrmacs anteriors encara estan sota estudi, en les darreres fases d'assajos clínics (32). Tanmateix, recentment, al març de 2024, la FDA (de l'anglès, *Food and Drug Administration*) va aprovar l'ús del Resmetirom als Estats Units (33,34).

Quan les estratègies anteriors no funcionen i el pacient presenta un IMC elevat i es troba en estadis avançats de la malaltia, pot ser necessària la cirurgia bariàtrica. Aquesta és capaç de reduir dràsticament el greix del fetge, fet que ajuda a revertir el desenvolupament de la MASLD (35). Principalment existeixen dos tipus de cirurgia: la restrictiva, que disminueix la capacitat gàstrica; i la malabsortiva, que redueix l'absorció intestinal (35). El *bypass* gàstric és la tècnica més comuna que combina la restrictiva amb la malabsortiva (3,35).

En les fases més severes de la malaltia, on hi ha presència de cirrosi i, fins i tot, HCC, i els tractaments anteriors no han sigut efectius, és necessari fer un transplantament de fetge ja que els pacients tenen un gran risc de morir degut a la disfunció hepàtica (36).

3.2. Metaanàlisi proteòmica

3.2.1. Ciències òmiques

El terme "òmica" descriu la caracterització i quantificació col·lectives de grans conjunts de dades que inclouen el genoma, el transcriptoma, el proteoma, el microbioma i l'epigenoma que influeixen en l'estructura, la funció i la dinàmica d'un procés biològic.

3.2.2. L'interès en l'estudi de les proteïnes

Les proteïnes són macromolècules implicades en funcions biològiques importants, com la replicació de la informació genòmica, la regulació de la transcripció, la senyalització, l'estructura, la catàlisi de reaccions i el transport de molècules. Sovint,

moltes proteïnes es troben desregulades i expressades diferencialment, fet que les converteix en possibles dianes de fàrmacs o biomarcadors clínics (37). A més, s'estima que existeixen al voltant de 300.000 proteïnes al cos humà, considerant les modificacions post-traduccionals (PTMs, de l'anglès, *post-translational modifications*) i l'empalmament, mentre que el genoma humà només conté 20.000 gens codificants. Això fa que l'estudi de les proteïnes sigui un camp molt complex (37,38).

La proteòmica és l'estudi a gran escala de les proteïnes i en els darrers anys ha contribuït significativament a l'ampliació del coneixement que es tenia sobre malalties complexes i multifactorials com l'Alzheimer o el Parkinson. També ha permès la identificació de biomarcadors utilitzables com a empremtes moleculars en el diagnosi de diferents tipus de càncer, entre d'altres (38).

Una gran quantitat de proteïnes regulen processos cel·lulars que són clau per al correcte funcionament de la cèl·lula. A més, el *target* principal de molts fàrmacs acostumen a ser proteïnes. És per això que la proteòmica és essencial per entendre mecanismes cel·lulars i moleculars, i per poder diagnosticar i tractar malalties eficaçment (38).

3.2.3. Què és la proteòmica?

La **proteòmica** és la ciència òmica que estudia el **proteoma**, és a dir, el conjunt de proteïnes expressades en un organisme o sistema biològic. Aquesta disciplina permet analitzar amb detall les funcions, estructures i interaccions de les proteïnes, així com els seus canvis sota diferents condicions fisiològiques, com ara els diversos estadis d'una malaltia (38).

Dins de la proteòmica, la **proteòmica d'expressió** se centra en l'anàlisi dels canvis qualitatius i quantitatius en els nivells de proteïnes produïdes en un sistema biològic. Encara que tècnicament les proteïnes són sintetitzades (traduïdes) a partir de l'ARN missatger, el terme 'expressió proteica' s'utilitza habitualment per referir-se a la quantitat de proteïna present en un moment determinat (39). Aquesta branca és especialment útil per identificar diferències proteiques en situacions concretes, com el desenvolupament de patologies o la resposta a tractaments (40).

Una de les grans fortaleses de la proteòmica, gràcies a la seva naturalesa **d'anàlisi a gran escala**, és que constitueix una **tècnica clau per a la identificació de biomarcadors**. Aquesta capacitat d'analitzar simultàniament un gran nombre de

proteïnes la converteix en una eina imprescindible per avançar en el diagnòstic precoç, la monitorització i el desenvolupament de teràpies personalitzades (41,42).

3.2.3.1. Espectrometria de masses

L'espectrometria de masses (EM) és l'eina més popular a l'hora d'identificar, caracteritzar i quantificar proteïnes i les seves PTMs amb un alt rendiment i a gran escala. La Figura 5 mostra un esquema del flux de treball de la proteòmica usant l'EM.

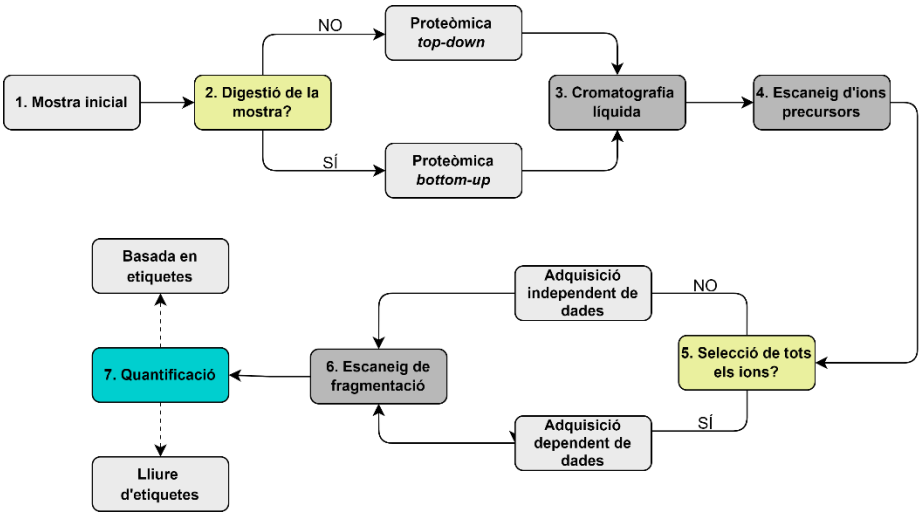

Figura 5. Flux de treball de la proteòmica.

3.2.3.1.1. Preparació de la mostra

La separació prèvia de proteïnes és essencial per a l'anàlisi proteòmica, ja que permet detectar proteïnes de baixa abundància i diferenciar aquelles que no es poden identificar només amb espectrometria de masses (EM). Tradicionalment, aquesta separació es feia amb tècniques basades en gel com el 2D PAGE (de l'anglès, *two-dimensional polyacrylamide gel electrophoresis*), però són lentes i laborioses. La cromatografia líquida (CL) és preferida per la seva major eficiència i velocitat, i perquè es pot acoblar directament a l'EM (CL-MS/MS) (43).

En general, existeixen 2 aproximacions a l'hora de preparar la mostra i analitzar-la: la proteòmica ***bottom-up*** o ***shotgun***, que digereix les proteïnes en pèptids abans de l'anàlisi (44), i la ***top-down***, que permet analitzar les proteïnes intactes ja que no les digereix (45). La primera és la més usada ja que els pèptids són més fàcils de separar i analitzar que les proteïnes (44), però pot no distingit fàcilment les proteïnes formades per pèptids comuns(46). La segona tècnica permet detectar formes actives de proteïnes,

incloent PTMs, però té moltes més dificultats tècniques (45). La Taula 5 mostra una comparació més detallada d'ambdues tècniques.

Taula 5 Comparació proteòmica *bottom-up* i *top-down*. Adaptada de *(44–46)*.

	Bottom-up	***Top-down***
Què s'analitza?	Pèptids	Proteïnes intactes
Preparació mostra	Proteòlisi	Cap (difícil mantenir proteïnes intactes en l'extracció)
Rendiment i escalabilitat	Permet l'anàlisi de diferents mostres amb mescles complexes	Menys rendiment, es necessiten mostres petites i mescles poc complexes
Especificitat	Pot trobar pèptids comuns en diferents proteïnes	PTMs i estructura sencera
Instrumentació i cost	Espectròmetres econòmics i àmpliament disponibles	Espectròmetres d'alta resolució, cars i menys comuns

3.2.3.1.2. Adquisició de dades

Quan les mostres ja estan preparades, cal seleccionar una estratègia d'adquisició de dades. Quan la mostra s'analitza a través d'un espectròmetre, el primer que es fa és ionitzar les mostres, donant lloc a ions precursors. Un escaneig de precursors (EM1) permet detectar-los i seleccionar-los. Els seleccionats es fragmenten mitjançant col·lisió i els fragments resultants s'analitzen (EM2). Els espectres obtinguts es comparen amb bases de dades per identificar pèptids i proteïnes (47).

Existeixen 2 estratègies: l'adquisició dependent de dades (**DDA**, de l'anglès, *data dependent acquisition*), que fragmenta els ions més abundants, i l'adquisició independent (**DIA**, de l'anglès, *data independent acquisition*), que fragmenta tots els ions, independentment de la seva abundància. És important destacar que en la DIA, l'espectre d'EM1 es fracciona en finestres i es duu a terme l'EM2 per cadascuna d'elles (47). La Figura 6 mostra una comparació visual entre aquestes dues estratègies.

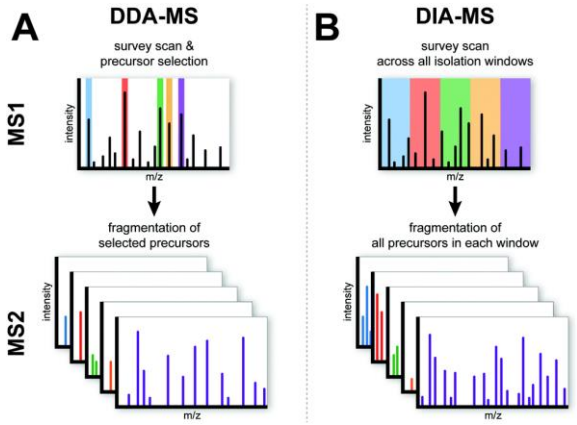

Figura 6. Comparació DDA vs DIA. Font: *(48)*. MS1: EM1, MS2: EM2.

3.2.3.1.3. Quantificació

Un dels principals atractius de l'EM és la possibilitat de poder quantificar les proteïnes tant si són molt abundants com si presenten nivells molt baixos d'expressió. La quantificació es pot dur a terme mitjançant 2 tècniques (43):

Per una banda, la **quantificació lliure d'etiquetes** (LFQ, de l'anglès, *label-free quantification*). No utilitza cap marcador isotòpic o químic per quantificar les proteïnes. La quantificació es pot basar en el recompte d'espectres de fragmentació (EM2) o en la mesura de la intensitat dels pics obtinguts després de l'EM1. Aquesta última opció és la més utilitzada.(49). La Taula 6 mostra una comparativa entre ambdues aproximacions.

Taula 6. Comparació de les 2 aproximacions de LFQ. Adaptada de *(49)*.

	Recompte d'espectres	**Mesura d'intensitat**
Precisió	Baixa, sobretot per proteïnes de baixa abundància	Precís i alta sensibilitat
Complexitat	Fàcil implementació	Anàlisi complex
Aplicabilitat	Proteïnes d'alta abundància i anàlisis qualitatius	Proteïnes de baixa abundància i anàlisis quantitatius precisos

Per l'altra, els **mètodes basats en etiqueta**. La quantificació té lloc gràcies a l'ús d'isòtops o etiquetes isobàriques. Aquest marcatge es pot fer a l'inici de la preparació de la mostra, de manera que es marquen les proteïnes; o bé després de la proteòlisi, aconseguint etiquetar els pèptids. En tots dos casos, aquesta tècnica permet la multiplexació de varies mostres dins d'un mateix experiment (43). Un exemple és la tècnica iTRAQ (de l'anglès, *isobaric tags for relative and absolute quantification*), que

permet la multiplexació de mostres mitjançant etiquetes isobàriques, amb la quantificació basada en l'alliberació d' ions *reporter* (50,51). Existeixen altres tècniques que usen etiquetes com TMT (de l'anglès, *tandem mass tags*), i altres que usen isòtops com SILAC (de l'anglès, *stable-isotope labeling with amino acids in cell culture*) (43). La Figura 7 mostra un esquema del funcionament dels ions *reporter* d'iTRAQ.

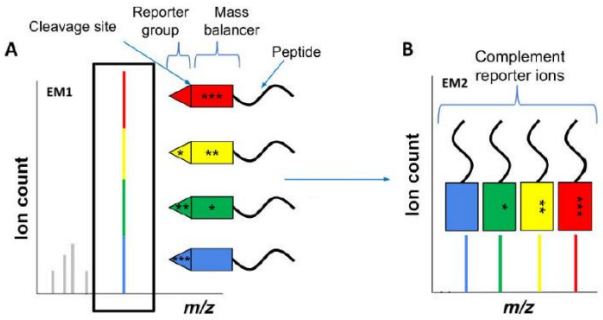

Figura 7. Esquema iTRAQ. Adaptada de *(50)*.

3.2.4. Metaanàlisis proteòmiques per al diagnosi de MASLD

Una metaanàlisi és essencialment una "anàlisi d'anàlisis"; és a dir, utilitzar mètodes estadístics per combinar els resultats de múltiples estudis sobre un mateix tema. Aquesta combinació permet a una metaanàlisi proporcionar un resum objectiu i quantitatiu resultats de múltiples estudis individuals, millorant el poder i la precisió estadístics (52). A més, mentre que els estudis individuals realitzats pel mateix grup de recerca poden estar influenciats per biaixos específics del laboratori, les metaanàlisis poden proporcionar un nivell d'evidència més robust i fiable (53). De fet, una metaanàlisi es considera una evidència altament fiable, situant-se a la part superior de la jerarquia d'evidències en investigació mèdica (52,53).

Tot i que en els darrers anys s'han realitzat diverses metaanàlisis centrades en l'estudi del transcriptoma hepàtic per caracteritzar la progressió de la MASLD, actualment no s'ha identificat cap metaanàlisi prèvia centrada específicament en la patologia de la MASH des d'una perspectiva proteòmica. Aquest tipus d'anàlisi té l'avantatge de permetre la identificació de biomarcadors generalitzables en diferents poblacions, fet que en reforça la validesa i en facilita l'aplicació en múltiples contextos clínics (52).

3.3. Aprenentatge automàtic per a la cerca de biomarcadors

L'anàlisi de grans conjunts de dades òmiques presenta dificultats per als mètodes tradicionals, ja que la seva alta dimensionalitat i complexitat fan que siguin poc pràctics i difícils de visualitzar. En aquest context, els mètodes d'aprenentatge automàtic (ML, de l'anglès, *machine learning*) esdevenen una eina clau, capaços de gestionar grans volums de dades amb relacions no lineals i diverses distribucions. Aquests algoritmes permeten identificar patrons, predir resultats i classificar grups a partir de dades com els perfils proteòmics, contribuint a una millor comprensió biològica i possibilitant tractaments mèdics personalitzats basats en el perfil biomolecular específic de cada pacient (54).

4. Hipòtesi i objectius

El descobriment de biomarcadors per a la MASLD és cada cop més rellevant, donat l'increment continu de la prevalença d'aquesta malaltia i les limitacions associades a la biòpsia hepàtica, una tècnica invasiva, costosa i amb riscos potencials per al pacient. En aquest context, la identificació de biomarcadors precisos i no invasius pot facilitar un diagnòstic més accessible i fiable, especialment pel que fa a la diferenciació entre SS i MASH,una distinció fonamental per establir un pronòstic i una estratègia terapèutica adequada.

La hipòtesi central d'aquest treball planteja que una metaanàlisi de dades proteòmiques pot permetre la identificació de biomarcadors perifèrics específics capaços de discriminar amb precisió entre els estadis d'SS i MASH.

L'objectiu principal és identificar un perfil de biomarcadors perifèrics que diferenciï de manera fiable entre els estadis histopatològics d'SS i MASH.

Els objectius específics que es deriven són:

- Realitzar una metaanàlisi de dades proteòmiques per identificar biomarcadors diferencials associats a cada estadi de SS i MASH.
- Desenvolupar un model predictiu basat en nivells proteics circulants mitjançant tècniques de ML, amb l'objectiu amb l'objectiu d'identificar pacients amb MASH de forma no invasiva i amb alta fiabilitat.

5. Metodologia

5.1. Cerca bibliogràfica i selecció d'estudis

Durant pràcticament un mes (3 de juny de 2024 – 28 de juny de 2024), es va dur a terme la cerca bibliogràfica i l'obtenció de les dades necessàries per poder realitzar la metaanàlisi. La base de dades principalment utilitzada va ser *PubMed*. El flux de treball va consistir en buscar tots els articles que parlessin de la MASLD per posteriorment filtrar-los i obtenir-ne aquells que complien els criteris d'inclusió explicats a continuació. Després de revisar cada article i seguir el procediment que mostra la Figura 8, es va escollir un total de 3 estudis segons els criteris d'inclusió següents:

- Els diferents estadis de la malaltia han d'estar diagnosticats a través d'una biòpsia.
- La proteòmica ha d'estar feta en plasma o sèrum.
- La cohort d'estudis ha de ser una població adulta amb obesitat.
- La proteòmica ha de ser no dirigida (és a dir, no enfocada a cap grup de proteïnes en particular).

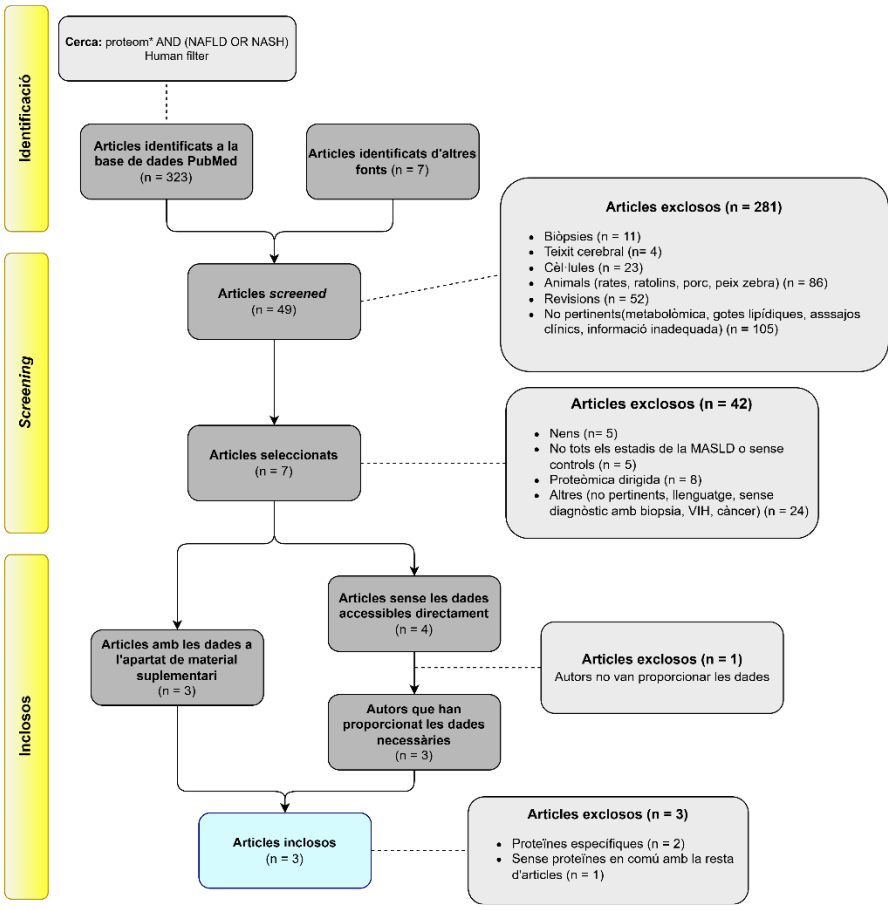

Figura 8. Diagrama PRISMA de la cerca bibliogràfica i selecció d'estudis.

5.2. Selecció de pacients

Després de seleccionar els articles i donat que alguns incloïen pacients amb estadis histopatològics no relacionats amb els nostres objectius (per exemple, pacients amb cirrosi), es va revisar quins pacients triar per a la metaanàlisi.

El primer estudi (55) incloïa 19 pacients, classificats com a control i NASH. Tot i ser anomenats com a controls en l'estudi inicial, els pacients d'aquest no tenien el fetge sa (sinó que es correspondrien amb MASLD sense MASH, és a dir, SS). Tots van ser seleccionats per a la metaanàlisi.

El segon estudi (25) utilitzava dues cohorts de pacients procedents de 2 estudis diferents. Un d'ells disposava de pacients control, és a dir, sense MALSD diagnosticada, i cirròtics, mentre que l'altre de pacients amb MASLD (SS i NASH). Aquest estudi (56), fet sobre 20 pacients, va ser el seleccionat. En l'article no quedava clar com havien fet la classificació dels pacients, així que es va contactar amb els autors per tal d'obtenir el NAS *score* de les biòpsies hepàtiques i poder fer la classificació adequada per a la metaanàlisi. Es van acabar seleccionant 19 pacients, ja que un no presentava NAS *score* en les dades proporcionades pels autors.

El darrer estudi (57) disposava de pacients control, SS i MASH. Tanmateix, seguint el criteri anterior, únicament es van seleccionar els pacients amb SS i MASH, aportant un total de 131 pacients. Entre tots els estudis, es van seleccionar 169 pacients.

5.2.1. Metadades

Les metadades, com l'edat, el sexe i l'IMC, són elements clau en les metaanàlisis, ja que permeten detectar i controlar com aquests factors influeixen en els nivells de certes proteïnes circulants. Per exemple, estudis han demostrat que l'adipositat està associada amb una varietat de biomarcadors proteics, i que aquestes associacions poden variar segons el sexe (58). A més, s'han identificat diferències significatives en biomarcadors circulants entre homes i dones, reflectint vies biològiques distintes implicades en malalties cardiovasculars (59). Per tant, la inclusió d'aquestes metadades en les anàlisis augmenta la precisió dels resultats i facilita la seva aplicació en diferents contextos clínics. Es van seleccionar l'IMC i la presència de T2DM com a covariables de la metaanàlisi, ja que eren les úniques metadades comunes en tots 3 estudis.

5.3. Preprocessat de les dades en cru i anàlisi exploratòria

Havent seleccionat els estudis i pacients, va ser necessari preprocessar les dades per tal de permetre la comparabilitat de les dades en l'anàlisi estadística. Es van seguir els passos documentats en *MSstats*, un paquet de R especialitzat en proteòmica (60). El preprocessat es va fer creant un *script* en R que seguia els passos següents:

- Eliminació de totes aquelles proteïnes on el percentatge de valors no disponibles (NA, de l'anglès, *not available*) fos superior al 80%.
- Transformació de les dades mitjançant logaritme. Es va utilitzar *log(dada+1)* per evitar valors negatius i l'error que donaria el càlcul de *log(0)*.
- Normalització mitjançant equalització de medianes, segons la Equació 1.

$$Dada_{x,y} = Dada_{x,y} \times \frac{Mediana_{global}}{Mediana_y}$$

Equació 1. Normalització.

- Imputació de les dades no disponibles (proteïnes amb un percentatge de dades NA <80%) mitjançant l'algoritme d'intel·ligència artificial kNN (de l'anglès, *k-nearest neighbours*). Breument, busca els "k" punts més propers al NA en el conjunt de dades i fa una estimació basada en els valors dels altres punts (61). Es va utilitzar el mètode inclòs en la llibreria *VIM*, versió 6.2.2 (62).

Per cada estudi, es va fer una anàlisi exploratòria de les dades. Es va inspeccionar la distribució de les dades a cada pas del preprocessat per monitoritzar l'efecte que tenien les diferents transformacions. Es va utilitzar la llibreria *ggplot2*, versió 3.5.1 (63).

5.4. Anàlisi de l'expressió diferencial de les proteïnes

L'anàlisi d'expressió diferencial individual de cada estudi es va fer mitjançant la llibreria *limma*, versió 3.60.3 (64). Es va usar la tècnica *empirical bayes*, implementada com una funció dins de *limma*, que ajusta l'error estàndard dels coeficients estimats per tal de millorar les variàncies estimades de les proteïnes individuals. S'utilitza quan el nombre de mostres per cada condició és petit per tal de poder fer tests estadístics més fiables i potents (65).

El resultat d'aquest pas és la taula de proteïnes diferencialment expressades o **DEPs** (de l'anglès, *differentially expressed proteins*), incloent el *fold change*, *p*-valor i intervals de confiança.

5.5. Metaanàlisi per comparativa

És important destacar que la metaanàlisi només es pot dur a terme amb aquelles proteïnes que siguin comunes en els 3 estudis. Un cop trobades, es va fer servir la funció *rem_mv* (66) de la llibreria *MetaVolcanoR*, versió 1.14.0 (67), per a realitzar la metaanàlisi combinant els valors de *fold change* dels diferents estudis tenint en compte els intervals de confiança i utilitzant un model d'efectes aleatoris.

5.6. Anàlisi d'enriquiment funcional

5.6.1. *Reactome pathway enrichment*

Amb el resultat obtingut de la metaanàlisi es va dur a terme una anàlisi d'enriquiment de conjunt de gens o GSEA (de l'anglès, *Gene Set Enrichment Analysis*). Per a poder-la fer, primer es van mapejar els identificadors KEGG (de l'anglès, *Kyoto Encyclopedia of Genes and Genomes*) dels gens que codifiquen per les proteïnes detectades mitjançant la llibreria d'R *clusterProfiler*, versió 4.8.2 (68), la qual usava la llibreria *org.Hs.eg.db*, versió 3.17.0 (69) per connectar-se a bases de dades d'anotació del genoma humà. Tot seguit, es van ordenar les d'acord al seu valor de *log2FoldChange* i es va realitzar l' anàlisi d'enriquiment de gens de vies biològiques a partir de la informació de la base de dades *Reactome*. Aquesta anàlisi es va dur a terme amb la llibreria d'R *ReactomePA*, versió 1.44.0 (70), mitjançant la funció *gsePathway* (71).

5.6.2. *GO Terms enrichment*

Es va fer una altra anàlisi d'enriquiment funcional per identificar els termes de la *Gene Ontology* (GO), com són processos biològics, funcions moleculars i components cel·lulars, sobre-representats dins de les DEPs més significatives (p-valor ≤ 0.1). Aquesta és un tipus d'anàlisi de sobre-representació o ORA (*over-representation analysis*). Per dur-la a terme, aquestes DEPs es van importar al programa *Cytoscape*, versió 3.10.2 (72), i es van analitzar a través del *plug-in ClueGO*, versió 2.5.10 (73).

5.7. Xarxa d'interaccions proteïna-proteïna

Es van seleccionar les DEPs més significatives (p-valor ≤ 0.1) i es van importar a *Cytoscape*. A través del *plug-in StringApp*, versió 2.1.1 (74), es van analitzar les interaccions entre les proteïnes codificades per aquests.

5.8. Aprenentatge automàtic

Un dels objectius principals és descobrir una empremta de proteïnes circulants capaç de distingir pacients amb MASH dels que no en tenen. Per aconseguir-ho, es va usar la llibreria *scikit-learn*, versió 1.5.1 (75), de Python. La Figura 9 mostra el flux de treball per arribar a aconseguir el model definitiu d'IA. També mostra una representació d'un procés anomenat validació creuada de tres subconjunts (3FCV, de l'anglès, *3-fold cross-validation*), posteriorment explicat.

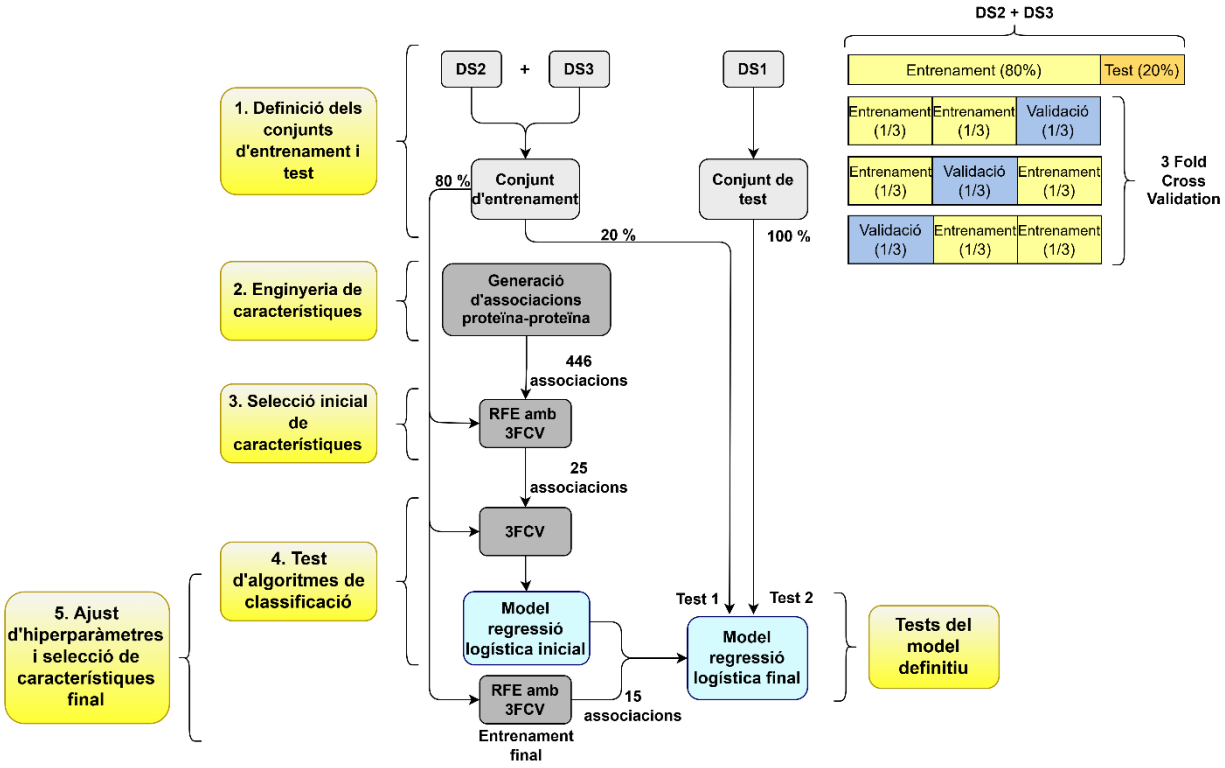

Figura 9. Flux de treball de la generació del model d'IA, amb la representació de la 3FCV.

5.8.1. Definició dels conjunts d'entrenament i test

Per entrenar i validar els models, es van combinar les dades del segon i tercer estudis (DS2 i DS3). Inicialment, el 80% d'aquestes dades combinades es van fer servir per a la selecció inicial de característiques i les proves preliminars dels algoritmes. Les característiques es defineixen com les dades d'entrada per entrenar els models, i corresponen a les ràtios entre proteïnes calculades individualment per cada pacient. Posteriorment, aquest mateix 80% es va fer servir per ajustar els hiperparàmetres i seleccionar les característiques definitives. El 20% restant va servir com a primer conjunt de validació (Test 1). El primer estudi (DS1) es va reservar exclusivament com a conjunt final de test per avaluar el rendiment del model sobre dades totalment noves (Test 2).

5.8.2. Enginyeria de característiques

Inicialment es van usar els nivells individuals de proteïnes com a dades d'entrada per entrenar els models. Després de tot el procés detallat a continuació, el model va obtenir una precisió en el Test 2 de 0.44. D'aquesta manera, com que utilitzar nivells individuals de proteïnes no va donar resultats satisfactoris, es van explorar ràtios entre proteïnes (característiques) associades a les mateixes vies KEGG, sota la hipòtesi que els canvis relatius entre proteïnes funcionalment relacionades poden aportar característiques més robustes per al model. Aquest enfocament va generar un total de de 446 associacions úniques (característiques), distribuïdes en 45 vies KEGG diferents, que es van utilitzar en les següents etapes de selecció i optimització de característiques. Les vies KEGG es van obtenir a través de la llibreria *KEGGREST*, versió 1.44.1 (76).

5.8.3. Selecció inicial de característiques

Posteriorment, es va aplicar una selecció de característiques inicial mitjançant l'eliminació recursiva de característiques (RFE, de l'anglès, *recursive feature elimination*) amb quatre algoritmes classificadors: *stochastic gradient descent* (SGD), *support vector machines* (SVM) amb nucli lineal, *AdaBoost* i *Random Forest*. Aquesta reducció de característiques és necessària tant per evitar el sobreajustament com per reduir la complexitat del model, ja que un nombre elevat de variables pot dificultar l'entrenament, empitjorar la interpretabilitat i disminuir la capacitat de generalització. La selecció es va realitzar mitjançant 3FCV a través de la funció *GridSearchCV*, dins de la llibreria *scikit-learn*. Aquesta funció maximitza la mètrica de l'F1 *score*, que es calcula com mostra l'Equació 2. En aquest procés, els pacients (cadascú amb les mateixes 446 associacions) es divideixen en 3 parts, alternant-ne una per validació i les altres dues per entrenament, repetint aquest procediment fins que totes les parts actuen com a conjunt de validació. A més, a cada iteració, es busca la millor combinació d'hiperparàmetres dels models per optimitzar els resultats. Finalment, es van seleccionar totes aquelles característiques que es trobaven dins del top 100 de més importància en almenys tres dels quatre models analitzats.

$$F1\ score = 2 \times \frac{Precisió\ \times Sensibilitat}{Precisió + Sensibilitat}$$

Equació 2. Càlcul F1 *score*.

5.8.4. Tests d'algoritmes de classificació

Amb aquestes característiques inicialment seleccionades, concretament 25, es va realitzar una nova 3FCV per comparar 15 algoritmes classificadors diferents. Per avaluar el rendiment de cada algoritme, es van avaluar diferents mètriques: exactitud, precisió, sensibilitat, especificitat i F1 *score*. Això va permetre identificar el model més robust i estable per a la classificació.

5.8.5. Ajust d'hiperparàmetres i selecció de característiques final

Havent seleccionat el model més òptim, en aquest cas el de regressió logística, es va realitzar una darrera RFE amb 3FCV per ajustar definitivament els hiperparàmetres i seleccionar les característiques finals del model. La selecció es va basar en la rellevància de cada característica, entesa com la seva contribució al procés de classificació segons els coeficients del model. Concretament, es van escollir les 15 característiques amb major pes en la predicció, és a dir, aquelles que tenien una influència més gran en la discriminació entre SS i MASH.

6. Resultats

6.1. Estudis seleccionats

La Taula 7 presenta una breu descripció dels estudis que es van seleccionar per la metaanàlisi.

Taula 7. Descripció dels estudis seleccionats per la metaanàlisi.

Estudi	Any de publicació	País	Tècnica	Teixit	Classificació	N total	N per grup	% dones
10.1042/BSR20190395	2020	Regne Unit	LC-EM/EM (mode DIA)	Plasma	Kleiner	19	10 SS / 9 MASH	53%
10.15252/msb.20188793	2019	Dinamarca	LC-EM/EM (mode DDA)	Plasma	NAS score	19	13 SS / 6 MASH	47%
10.1016/j.jprot.2024.105317	2024	Espanya	LC-EM/EM amb TMT (mode DDA)	Sèrum	Kleiner	131	66 SS / 65 MASH	100%

6.2. Anàlisi exploratori

La Figura 10 mostra l'evolució de la distribució de les dades en cada pas del preprocessat. Cada fila correspon a cada estudi i cada columna a cada pas del preprocessat. Les gràfiques il·lustren clarament com els valors bruts (a l'esquerra) presenten distribucions molt asimètriques i amb valors extrems, mentre que després de la transformació, normalització i imputació (cap a la dreta), les dades mostren distribucions molt més simètriques i tractables per a anàlisis estadístiques.

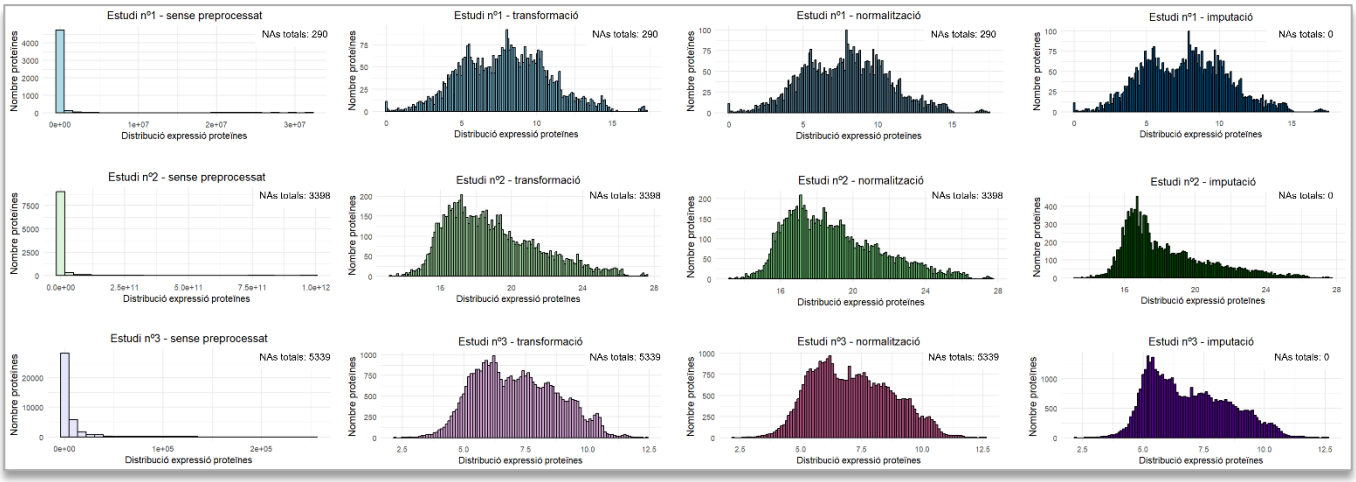

Figura 10. Evolució de les dades en les diferents fases del preprocessat.

6.3. Proteïnes diferencialment expressades entre SS i MASH

En l'anàlisi de l'expressió diferencial de les proteïnes entre les condicions SS i MASH, es va identificar un total de **119 proteïnes** comunes en els tres estudis analitzats. D'aquestes, només **6** van resultar ser significants, amb un *p*-valor ≤ 0,05. Per visualitzar aquests resultats, es va generar un *Volcano Plot* (Figura 11), on cada punt representa una proteïna. En aquest gràfic, les proteïnes amb un valor de *fold change* positiu (marcades en vermell) estan **sobreexpressades** en la condició MASH en comparació amb SS. Per contra, les proteïnes amb un *fold change* negatiu (marcades en blau) es troben **infraexpressades** en MASH respecte a SS. Finalment, les proteïnes situades al punt 0 indiquen l'absència de canvis en l'expressió entre ambdues condicions.

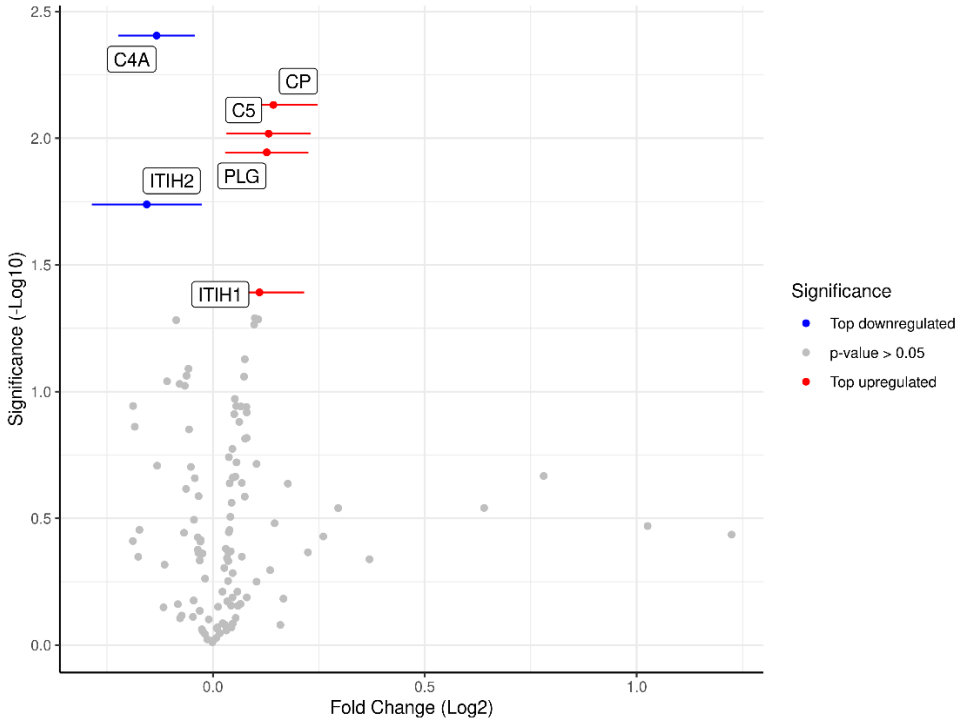

Figura 11. *Volcano plot* de l'expressió diferencial de les proteïnes en la comparativa entre SS i MASH. Cada punt identifica una DEP. La seva posició ve determinada pel *p*-valor (eix Y), representat en escala logarítmica negativa, i pel *fold change* (eix X), representant la variació en l'expressió proteica entre ambdues condicions. Les DEPs etiquetades i en color blau i vermell representen les DEPs més significants (*p*-valor ≤ 0.05). En color vermell es troben representades aquelles que estan sobreexpresades, en blau els que estan infraexpressades. Les barres d'error mostren com varia el valor de *fold change* entre els diferents conjunts de dades.

6.4. Anàlisi d'enriquiment funcional

6.4.1. Vies metabòliques enriquides en la MASLD

L'anàlisi GSEA va identificar vies metabòliques alterades tant positivament com negativament, tal com es mostra a la Figura 12. Les tres primeres vies del gràfic estan relacionades amb el sistema del complement (SC), un component fonamental de la immunitat innata (77), destacant-ne l'enriquiment positiu. D'altra banda, les vies metabòliques enriquides negativament es troben associades al transport intracel·lular, el metabolisme i les PTMs.

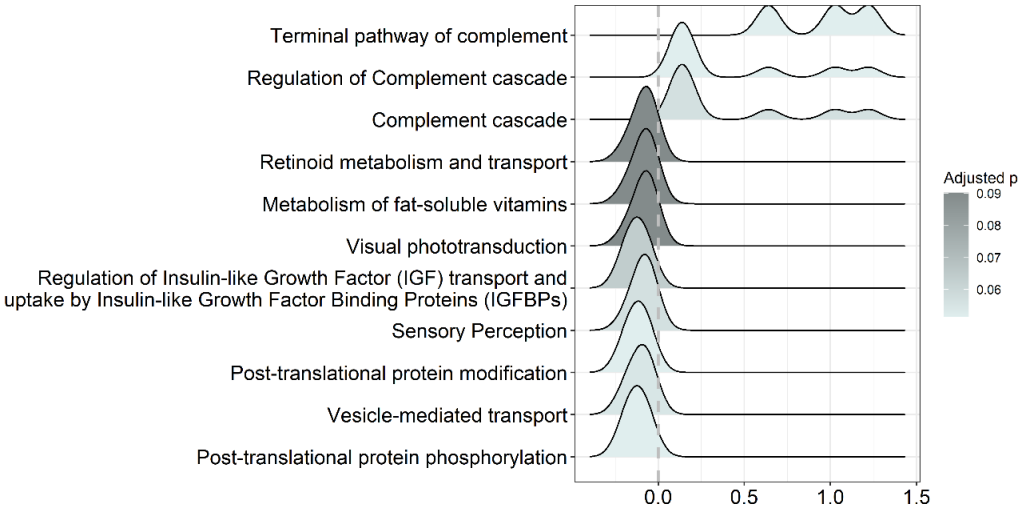

Figura 12. Vies més enriquides en la MASLD obtingudes a partir d'una anàlisi GSEA. Cada pic representa un conjunt de proteïnes, on l'àrea sota la corba indica la magnitud del senyal d'enriquiment. Les vies amb pics a la dreta del punt 0.0 (eix X) estan positivament enriquides en processos metabòlics, mentre que les situades a l'esquerra indiquen enriquiment negatiu. Els conjunts de proteïnes properes al punt 0.0 no mostren enriquiment destacat. El gradient de color reflecteix el *p*-valor: els tons més clars indiquen *p*-valors més petits (més significatius, mínim 0.05), i els tons més foscos representen *p*-valors més grans (màxim 0.09).

6.4.2. Termes de la GO enriquits en les DEPs més significatives

L'anàlisi ORA va identificar processos biològics significativament sobre-representats dins de les DEPs, amb un *p*-valor ≤ 0.1. Els resultats, presentats a la Figura 13, mostren una clara divisió entre processos relacionats amb la regulació negativa de la coagulació sanguínia i la fibrinòlisi. D'entre tots, destaca la regulació negativa de la fibrinòlisi. A l'eix Y del gràfic es mostren els processos més enriquits, mentre que l'eix X indica el percentatge de proteïnes associades a cada terme.

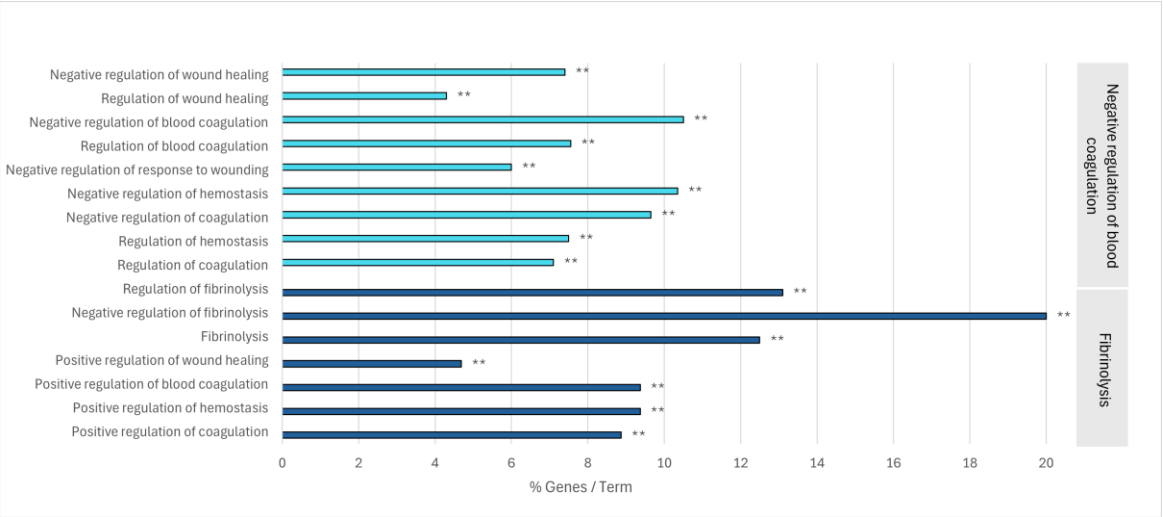

Figura 13. Termes de la GO més enriquits dins del conjunt de proteïnes significatives (p-valor ≤ 0.1) obtinguts a partir d'una anàlisi ORA. A l'eix Y es troben els processos més enriquits, l'eix X representa el percentatge de gens per terme. ** indica un p-valor ≤ 0.05.

6.5. Xarxa d'interaccions proteïna-proteïna

L'anàlisi de les interaccions entre proteïnes o PPI (de l'anglès, *protein-protein interaction*) va generar la xarxa mostrada a la Figura 14, que destaca les interaccions més significatives entre DEPs (p-valor $\leq$ 0.1). Principalment estan implicades en el metabolisme lipídic (78), mecanismes de coagulació sanguínia (79), ME (80) i SC (77).

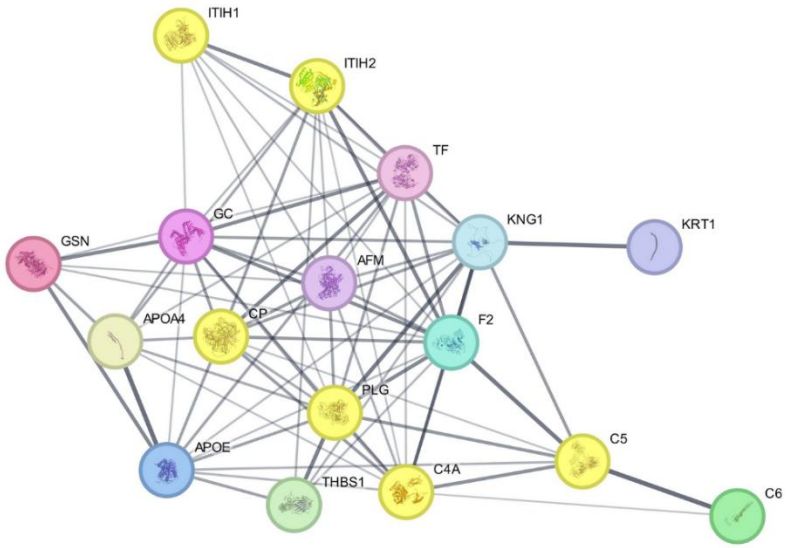

Figura 14. Xarxa PPI.

6.6. Aprenentatge automàtic

Per seleccionar els conjunts d'entrenament i test, es va aplicar una anàlisi de components principals (PCA, de l'anglès, *principal component analysis*) sobre les dades combinades dels tres estudis. La Figura 15 mostra la projecció en els dos primers components principals (PC1 i PC2), on PC1 captura el 95,9% de la variabilitat. Es va observar una clara separació al llarg de PC1, especialment entre el segon estudi i els altres dos. Aquesta diferenciació indica que l'estudi 2 conté dades amb característiques diferents, fet que pot contribuir a millorar la capacitat de generalització del model. Per aquest motiu, es van escollir els estudis 2 i 3 com a conjunt d'entrenament, amb l'objectiu de garantir diversitat en les dades, i es va reservar l'estudi 1 com a conjunt de test.

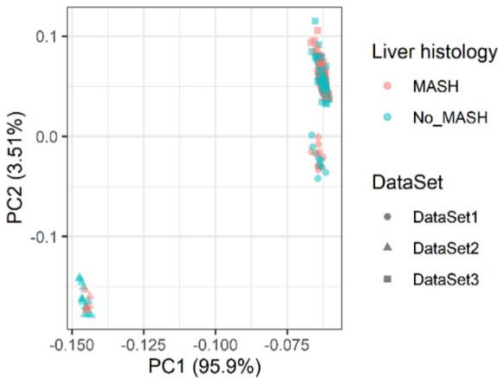

Figura 15. PCA de 2 components dels estudis seleccionats. L'eix X representa la component principal 1, l'eix 2 representa la component principal 2. Els cercles representen els pacients del primer estudi, els triangles els del segon i els quadrats els del tercer.

Després de la selecció inicial de característiques a partir de les 446 associacions úniques entre proteïnes, se'n van identificar 25 que es trobaven el en el top 100 més important en al menys 3 dels 4 algorismes testats. Aquestes van servir per testejar 15 algoritmes de classificació, els resultats dels quals es mostren a la Figura 16. Entre tots els models analitzats, el que utilitza l'algoritme de **regressió logística** (marcat en vermell a la figura) va obtenir el millor rendiment segons les mètriques utilitzades (exactitud, precisió, etc).

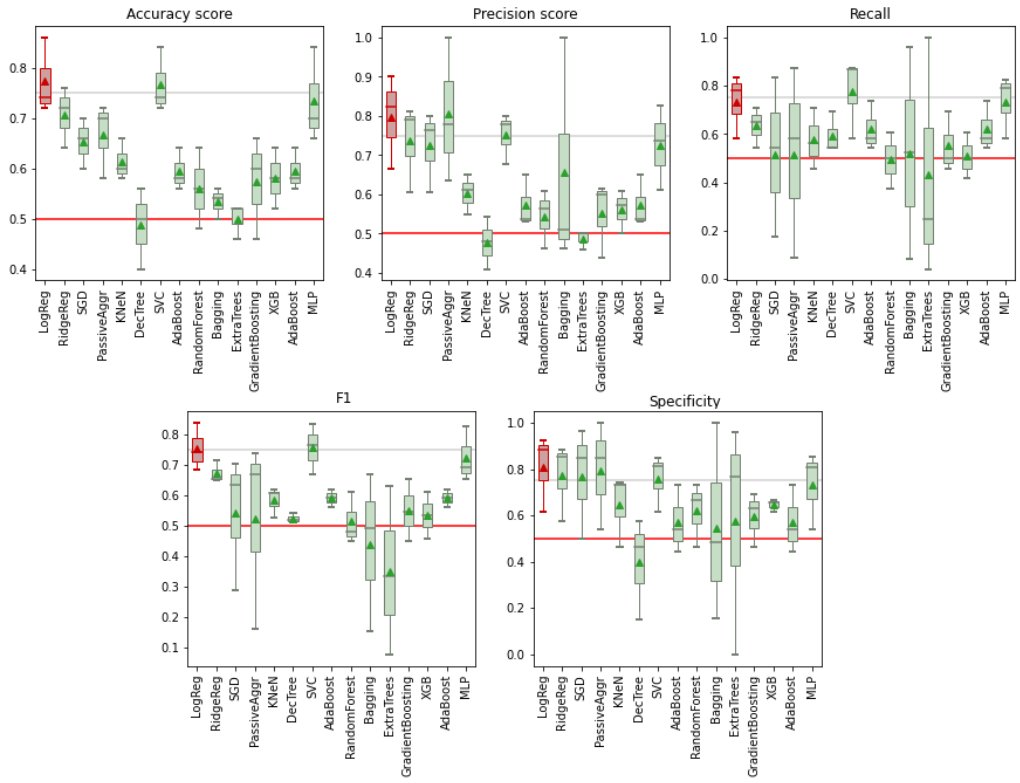

Figura 16. Box plots comparant les mètriques de classificació per als diferents algoritmes amb línia base (0.5) i model òptim destacat en vermell.

Aquest model es va tornar a entrenar i, després de realitzar la segona selecció de característiques mitjançant l'estratègia RFE amb validació creuada, es va reduir el nombre final de característiques a 15. Els resultats d'aquesta última selecció de característiques es poden observar a la Figura 17. Encara que utilitzar fins a 23 relacions podria semblar lleugerament més favorable, no aporta millores significatives en l'F1 *score*, sinó que augmenta innecessàriament la complexitat del model.

Per altra banda, les 15 característiques seleccionades, agrupades segons les vies KEGG en què participen, són les mostrades a la Figura 18a.

Finalment, el rendiment del model es pot resumir de dues formes:

1. A través de les matrius de confusió obtingudes durant l'entrenament i els dos tests, que es poden veure a la Figura 18b. Cada matriu compara les prediccions inferides pel model amb l'etiqueta real del perfil proteòmic.
2. Mitjançant la Taula 8, que recull de manera resumida les mètriques de rendiment aconseguides a cadascuna de les fases mencionades anteriorment. També mostra els valors d'AUC (de l'anglès, *area under the curve*). Aquesta àrea representa la relació entre les prediccions realment vertaderes i les falses

positives. Com més proper a 1 sigui el seu valor, millor distingeix el model entre SS i MASH.

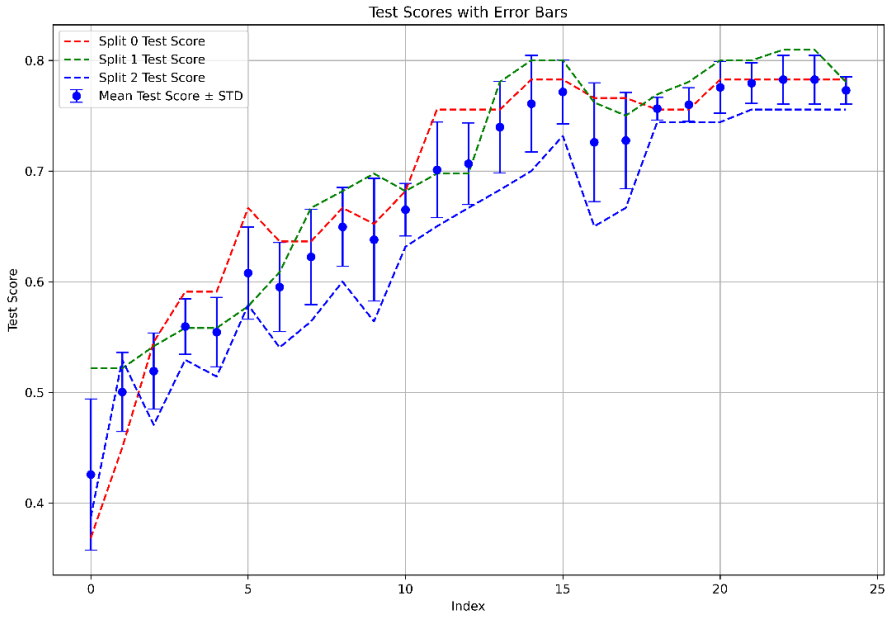

Figura 17. Rendiment (F1 *score*) del model de regressió logística amb diferents nombres de característiques (eix X), calculat mitjançant RFE amb validació creuada.

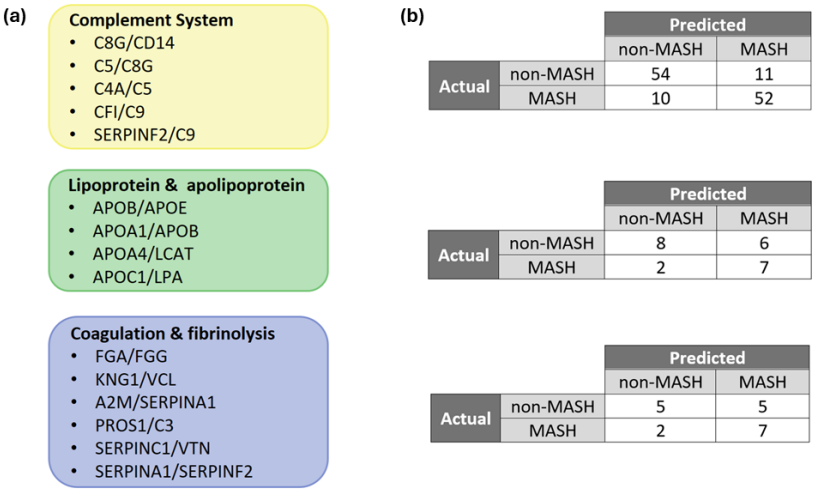

Figura 18. *(a)* Característiques seleccionades del model de regressió logística agrupades en les vies KEGG on participen. (b) *Rendiment del model* de regressió logística *en els conjunts d'entrenament i test. De dalt a baix, matriu de confusió per al 80% dels conjunts 2 i 3 (conjunt d'entrenament), matriu de confusió per al 20% dels conjunts 2 i 3 (test 1) i per al conjunt 1 (test 2).*

Taula 8. Mètriques de rendiment per a totes les particions de dades (dades d'entrenament, test 1 i test 2) realitzades durant l'anàlisi de ML.

Data	Descripció	Exactitud	Precisió	Sensibilitat	F1 *score*	Especificitat	AUC
Data d'entrenament	80% de DS2 i DS3	0.835	0.825	0.839	0.832	0.831	0.956
Test 1	20% de DS2 i DS3	0.652	0.539	0.778	0.636	0.571	0.746
Test 2	DS1	0.631	0.538	0.778	0.667	0.500	0.622

7. Discussió

La identificació de biomarcadors específics que permetin distingir clarament entre els estadis d'SS i MASH representa una necessitat clínica rellevant, especialment tenint en compte l'increment constant en la prevalença de la malaltia i les limitacions inherents a la biòpsia hepàtica. En aquest context, la metaanàlisi de dades proteòmiques es presenta com una estratègia prometedora per aconseguir un diagnòstic fiable, precís i no invasiu.

Mitjançant la combinació dels tres conjunts de dades disponibles, la metaanàlisi ha permès identificar sis DEPs. El perfil global d'aquestes proteïnes suggereix un enriquiment positiu de les vies relacionades amb el SC, un component essencial de la immunitat innata (77). D'entre les sis DEPs identificades, destaquen les proteïnes C4A i C5, totes dues integrants d'aquest sistema (77). El SC ha estat àmpliament estudiat en relació amb la progressió de la MASLD. L'estudi realitzat per Guo *et al.* (81) indica que nivells elevats de C3 i els seus derivats, així com de C5, s'associen estretament amb la RI, el percentatge de greix al fetge i la inflamació hepàtica, indicadors clau de la gravetat de la malaltia. Tot i això, en aquesta metaanàlisi no s'ha observat una associació consistent entre el C3 i la presència de MASH en les cohorts analitzades. De manera similar, no s'ha detectat una relació positiva amb la proteïna C4A. Aquesta absència podria deure's a una activació preferencial de la via alternativa o de la via de les lectines, les quals no impliquen directament la proteïna C4 i, per tant, poden generar menys C4A (82). Aquesta activació selectiva podria estar modulada per factors com l'EO, la presència de dipòsits de greix en el fetge o alteracions metabòliques associades a la SM (83). Tot plegat posa de manifest la complexitat del paper del SC en la patogènesi de la MASLD i la necessitat de continuar aprofundint en aquest àmbit.

Un altre resultat rellevant és l'increment dels nivells de ceruloplasmina (CP), malgrat que seva relació amb la MASLD és encara controvertida. Un estudi recent realitzat per Jiang *et al.* (84) suggereix que aquesta proteïna podria constituir una potencial diana terapèutica ja que la seva eliminació s'ha associat amb una reducció de l'acumulació de lípids i inflamació, i de la fibrosi, contribuint així a disminuir el dany hepàtic. Aquests resultats vinculen la CP amb estadis avançats de la malaltia. Contràriament, altres estudis han identificat nivells baixos de CP en infants amb MASH (85) o en pacients no diabètics (86), mentre que alguns estudis no han trobat cap relació clara amb la patologia (87).

El plasminogen (PLG) és una altra de les proteïnes associades positivament amb la progressió de la MASLD. Aquest és un precursor de la plasmina, un enzim que degrada la fibrina, procés que s'anomena fibrinòlisi (88). A més, s'ha observat que la fibrina, juntament amb les plaquetes, són les principals encarregades de la formació de coàguls (89). D'aquesta manera, el PLG, gràcies a la degradació de fibrina, indirectament també regula negativament la coagulació sanguínia. Tanmateix, l'estudi publicat per Nawaz *et al.* (90) ha evidenciat que l'augment de PLG en pacients amb MASH comporta també un increment en l'expressió de l'inhibidor de l'activador de plasminogen o PAI-1 (de l'anglès, *plasminogen activator inhibitor-1*), el principal regulador negatiu de la fibrinòlisi. Aquest fet pot afavorir la formació de coàguls (91).

Pel que fa als inhibidors inter-α-tripsina amb cadenes pesants, aquests tenen un paper crucial en la fibrosi ja que presenten una elevada afinitat per l'àcid hialurònic, un component principal de la ME dels hepatòcits, contribuint així a la seva estabilització (92). No obstant això, un estudi recent (93) posa de manifest que l'expressió de l'inhibidor amb cadena pesant 3 disminueix a mesura que la malaltia progressa. Aquest comportament diferencial suggereix la necessitat de dur a terme futurs estudis que permetin caracteritzar amb més profunditat els mecanismes de regulació i funció d'aquests inhibidors en el context de la MASLD.

Un altre dels objectius principals d'aquest estudi era explorar la viabilitat dels biomarcadors proteòmics com a eina diagnòstica fiable per a la MASH. Tot i que les mesures directes dels nivells de proteïnes han resultat poc consistents a l'hora d'extrapolar els resultats a noves cohorts, l'ús de ràtios entre proteïnes que participen en les mateixes vies metabòliques ha demostrat ser una estratègia més robusta. Aquest enfocament ha estat especialment efectiu en vies associades al SC, les lipoproteïnes, les apolipoproteïnes i els mecanismes de coagulació. Aquestes vies semblen jugar un paper clau en la progressió de la malaltia afavorint la infiltració de neutròfils en el fetge i promovent una desregulació metabòlica i una inflamació sistèmica que augmenten el risc d'afeccions cardiovasculars associades (94). L'estratègia també ha resultat útil en l'estudi de les vies implicades en la coagulació i fibrinòlisi. De fet, s'ha observat que els pacients amb MASH presenten nivells elevats de marcadors de coagulació al plasma, cosa que els situa en un risc més alt de patir episodis de trombosi venosa (95).

Gràcies a l'aplicació d'algorismes de ML, s'han pogut identificar 15 relacions entre proteïnes candidates a actuar com a biomarcadors diagnòstics de MASH, obtenint

un F1 *score* de 0.67 en mostres completament independents. Tot i els desafiaments inherents, les ràtios proteiques identificades representen una aproximació prometedora per a la detecció no invasiva de MASH. Els resultats obtinguts en els conjunts de test (precisions de 0,539 i 0,583, i especificitats de 0,571 i 0,500) evidencien una certa variabilitat entre cohorts, et que subratlla el caràcter exploratori del model desenvolupat. L'objectiu principal ha estat demostrar la viabilitat d'utilitzar tècniques de ML per a la identificació de biomarcadors circulants, amb la consciència que refinaments posteriors podrien millorar-ne significativament el rendiment classificatori. En comparació amb marcadors tradicionals com ALT/AST, FIB-4 o fins i tot la biòpsia hepàtica, aquest enfocament ofereix una alternativa potencialment més específica i menys invasiva.

Tanmateix, aquest estudi presenta algunes limitacions importants. D'una banda, la comprensió de la patogènesi i l'etiologia multifactorial de la MASLD, així com la identificació d'un tractament efectiu, continua sent limitada per la manca d'estudis proteòmics disponibles. A diferència de la transcriptòmica, les bases de dades de proteòmica estan encara poc desenvoulpades, cosa que ha restringit considerablement l'abast de la metaanàlisi realitzada. A més, les dades es trobaven sovint disperses en materials suplementaris o en repositoris especialitzats, dificultant i alentint el procés de recopilació, que finalment ha donat lloc a la inclusió de només tres conjunts de dades proteòmiques. Aquest fet posa de manifest la necessitat urgent de centralitzar i millorar l'accessibilitat de les dades proteòmiques en repositoris públics per facilitar estudis de major escala en un futur pròxim.

Un altre obstacle important ha estat la manca d'uniformitat en l'anotació de les dades entre estudis, tant pel que fa als mètodes analítics emprats com a les metadades clíniques disponibles. L'absència d'informació detallada en alguns d'ells —com l'IMC o paràmetres histològics clau com el grau de fibrosi o inflamació— ha limitat la possibilitat d'ajustar els models per covariables rellevants, fet que probablement hauria contribuït a augmentar la robustesa i la fiabilitat de les conclusions.

A més, la representativitat de les cohorts utilitzades és limitada, ja que la majoria dels participants eren individus de raça caucàsica, i en el cas d'un dels estudis (DS3), exclusivament dones. Aquesta manca de diversitat planteja interrogants sobre la generalitzabilitat dels resultats a altres grups ètnics o poblacionals. També cal destacar que la mida reduïda de les mostres en els estudis proteòmics analitzats afegeix incertesa addicional respecte a la solidesa del model desenvolupat.

Pel que fa al ML, malgrat que el model va mostrar resultats prometedors en els conjunts d'entrenament, va experimentar una davallada de rendiment en l'aplicació a conjunts de dades completament independents. Aquest fet posa de manifest que la capacitat de generalització del model és encara limitada i que seran necessaris refinaments metodològics addicionals per millorar-ne l'eficàcia diagnòstica.

Malgrat aquestes limitacions, aquest estudi introdueix un enfocament innovador que vincula els resultats obtinguts a través d'una metaanàlisi de proteòmica circulant no dirigida en pacients amb MASH, una aproximació que, fins on s'ha pogut constatar, és pionera en aquest àmbit. La integració de tècniques de ML per identificar una signatura proteòmica consistent entre diverses cohorts representa un avenç significatiu. En aquest sentit, s'ha prioritzat la generalització dels resultats mitjançant la reserva d'una cohort completa com a conjunt de prova independent, fet que ha permès avaluar amb més realisme la capacitat real del model per extrapolar a nous conjunts de dades.

8. Conclusions

Aquest treball aporta avenços significatius en la identificació de biomarcadors per al diagnòstic no invasiu de la MASLD. Mitjançant una metaanàlisi proteòmica combinada amb tècniques de ML, s'han identificat alteracions moleculars associades a MASH que es mantenen consistents en diferents cohorts de pacients. Els resultats han permès destacar sis proteïnes diferencialment expressades, així com diversos biomarcadors potencials vinculats amb la cascada del complement, les apolipoproteïnes, les lipoproteïnes i les proteïnes implicades en processos de coagulació.

Tot i aquests avenços, la variabilitat observada entre cohorts subratlla la necessitat de dur a terme nous estudis amb mostres més àmplies i heterogènies per tal de validar aquestes troballes i identificar biomarcadors més robustos i generalitzables. En conjunt, aquest estudi representa una contribució rellevant per a la comprensió dels mecanismes moleculars implicats en la MASLD, amb el potencial de millorar el diagnòstic, seguiment clínic i la gestió global dels pacients afectats per aquesta patologia hepàtica.

9. Bibliografia

1. Chalasani N, Younossi Z, Lavine JE, Diehl AM, Brunt EM, Cusi K, et al. The diagnosis and management of non-alcoholic fatty liver disease: practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association. Hepatology. junio de 2012;55(6):2005-23.

2. Pouwels S, Sakran N, Graham Y, Leal A, Pintar T, Yang W, et al. Non-alcoholic fatty liver disease (NAFLD): a review of pathophysiology, clinical management and effects of weight loss. BMC Endocr Disord. 14 de marzo de 2022;22(1):63.

3. Ahmed A, Wong RJ, Harrison SA. Nonalcoholic Fatty Liver Disease Review: Diagnosis, Treatment, and Outcomes. Clinical Gastroenterology and Hepatology. noviembre de 2015;13(12):2062-70.

4. Rinella ME, Lazarus JV, Ratziu V, Francque SM, Sanyal AJ, Kanwal F, et al. A multisociety Delphi consensus statement on new fatty liver disease nomenclature. Journal of Hepatology. diciembre de 2023;79(6):1542-56.

5. Rinella ME, Sookoian S. From NAFLD to MASLD: updated naming and diagnosis criteria for fatty liver disease. Journal of Lipid Research. enero de 2024;65(1):100485.

6. Huang J, Ou W, Wang M, Singh M, Liu Y, Liu S, et al. MAFLD Criteria Guide the Subtyping of Patients with Fatty Liver Disease. RMHP. febrero de 2021;Volume 14:491-501.

7. Hagström H, Vessby J, Ekstedt M, Shang Y. 99% of patients with NAFLD meet MASLD criteria and natural history is therefore identical. Journal of Hepatology. febrero de 2024;80(2):e76-7.

8. Rivera-Esteban J, Jiménez-Masip A, Muñoz-Martínez S, Augustin S, Guerrero RA, Gabriel-Medina P, et al. Prevalence and Risk Factors of MASLD and Liver Fibrosis amongst the Penitentiary Population in Catalonia: The PRISONAFLD Study. JCM. 24 de noviembre de 2023;12(23):7276.

9. Wang T, Xi Y, Raji A, Crutchlow M, Fernandes G, Engel SS, et al. Overall and subgroup prevalence of non-alcoholic fatty liver disease and prevalence of advanced fibrosis in the United States: An updated national estimate in National Health and Nutrition Examination Survey (NHANES) 2011-2018. Annals of Hepatology. enero de 2024;29(1):101154.

10. Younossi Z, Anstee QM, Marietti M, Hardy T, Henry L, Eslam M, et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol. enero de 2018;15(1):11-20.

11. Enomoto H. Liver fibrosis markers of nonalcoholic steatohepatitis. WJG. 2015;21(24):7427.

12. Kosaka K, Kubota Y, Adachi N, Akita S, Sasahara Y, Kira T, et al. Human adipocytes from the subcutaneous superficial layer have greater adipogenic potential

and lower PPAR-γ DNA methylation levels than deep layer adipocytes. American Journal of Physiology-Cell Physiology. 1 de agosto de 2016;311(2):C322-9.

13. Lackner C. Hepatocellular ballooning in nonalcoholic steatohepatitis: the pathologist's perspective. Expert Review of Gastroenterology & Hepatology. abril de 2011;5(2):223-31.

14. Antar SA, Ashour NA, Marawan ME, Al-Karmalawy AA. Fibrosis: Types, Effects, Markers, Mechanisms for Disease Progression, and Its Relation with Oxidative Stress, Immunity, and Inflammation. IJMS. 16 de febrero de 2023;24(4):4004.

15. Chen YH, Wu WK, Wu MS. Microbiota-Associated Therapy for Non-Alcoholic Steatohepatitis-Induced Liver Cancer: A Review. IJMS. 20 de agosto de 2020;21(17):5999.

16. Guo X, Yin X, Liu Z, Wang J. Non-Alcoholic Fatty Liver Disease (NAFLD) Pathogenesis and Natural Products for Prevention and Treatment. IJMS. 7 de diciembre de 2022;23(24):15489.

17. Buzzetti E, Pinzani M, Tsochatzis EA. The multiple-hit pathogenesis of non-alcoholic fatty liver disease (NAFLD). Metabolism. agosto de 2016;65(8):1038-48.

18. Vidal-Cevallos P, Sorroza-Martínez AP, Chávez-Tapia NC, Uribe M, Montalvo-Javé EE, Nuño-Lámbarri N. The Relationship between Pathogenesis and Possible Treatments for the MASLD-Cirrhosis Spectrum. IJMS. 16 de abril de 2024;25(8):4397.

19. Huang PL. A comprehensive definition for metabolic syndrome. Disease Models & Mechanisms. 30 de abril de 2009;2(5-6):231-7.

20. Tilson SG, Morell CM, Lenaerts A, Park SB, Hu Z, Jenkins B, et al. Modeling PNPLA3-Associated NAFLD Using Human-Induced Pluripotent Stem Cells. Hepatology. diciembre de 2021;74(6):2998-3017.

21. Hill MA, Yang Y, Zhang L, Sun Z, Jia G, Parrish AR, et al. Insulin resistance, cardiovascular stiffening and cardiovascular disease. Metabolism. junio de 2021;119:154766.

22. Bashiardes S, Shapiro H, Rozin S, Shibolet O, Elinav E. Non-alcoholic fatty liver and the gut microbiota. Molecular Metabolism. septiembre de 2016;5(9):782-94.

23. Arciello M, Gori M, Maggio R, Barbaro B, Tarocchi M, Galli A, et al. Environmental Pollution: A Tangible Risk for NAFLD Pathogenesis. IJMS. 7 de noviembre de 2013;14(11):22052-66.

24. Kannan K, Vimalkumar K. A Review of Human Exposure to Microplastics and Insights Into Microplastics as Obesogens. Front Endocrinol. 18 de agosto de 2021;12:724989.

25. Niu L, Geyer PE, Wewer Albrechtsen NJ, Gluud LL, Santos A, Doll S, et al. Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. Molecular Systems Biology. marzo de 2019;15(3):e8793.

26. Manilgama SR, Hettiarachchi NM. Diagnosis of nonalcoholic fatty liver disease. J of Cey Coll of Phy. 24 de junio de 2022;53(1):35-42.

27. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology. junio de 2005;41(6):1313-21.

28. Bedossa P. Histological Assessment of NAFLD. Dig Dis Sci. mayo de 2016;61(5):1348-55.

29. El Ray A, Paradis V, Montasser A, Elghannam M, Shemis M, Nessim I, et al. Usefulness of the SAF score to characterize NAFLD/NASH in non-cirrhotic HCV patients. Egypt Liver Journal. 20 de agosto de 2022;12(1):47.

30. Kupčová V, Fedelešová M, Bulas J, Kozmonová P, Turecký L. Overview of the Pathogenesis, Genetic, and Non-Invasive Clinical, Biochemical, and Scoring Methods in the Assessment of NAFLD. IJERPH. 24 de septiembre de 2019;16(19):3570.

31. Nassir F. NAFLD: Mechanisms, Treatments, and Biomarkers. Biomolecules. 13 de junio de 2022;12(6):824.

32. Alonso-Peña M, Del Barrio M, Peleteiro-Vigil A, Jimenez-Gonzalez C, Santos-Laso A, Arias-Loste MT, et al. Innovative Therapeutic Approaches in Non-Alcoholic Fatty Liver Disease: When Knowing Your Patient Is Key. IJMS. 27 de junio de 2023;24(13):10718.

33. Keam SJ. Resmetirom: First Approval. Drugs. junio de 2024;84(6):729-35.

34. Sookoian S, Pirola CJ. Resmetirom for treatment of MASH. Cell. junio de 2024;187(12):2897-2897.e1.

35. Karin Papapietro V. Cirugía para la obesidad: efectos generales, beneficios y riesgos. Revista Médica Clínica Las Condes. marzo de 2012;23(2):189-95.

36. Thuluvath PJ, Hanish S, Savva Y. Waiting List Mortality and Transplant Rates for NASH Cirrhosis When Compared With Cryptogenic, Alcoholic, or AIH Cirrhosis. Transplantation. enero de 2019;103(1):113-21.

37. Cohen L, Walt DR. Highly Sensitive and Multiplexed Protein Measurements. Chem Rev. 9 de enero de 2019;119(1):293-321.

38. Pierce JD, Fakhari M, Works KV, Pierce JT, Clancy RL. Understanding proteomics. Nursing & Health Sciences. marzo de 2007;9(1):54-60.

39. Bradley BP, Kalampanayil B, O'Neill MC. Protein Expression Profiling. En: Tyther R, Sheehan D, editores. Two-Dimensional Electrophoresis Protocols [Internet]. Totowa, NJ: Humana Press; 2009 [citado 26 de abril de 2025]. p. 455-68. (Methods in

Molecular Biology; vol. 519). Disponible en: http://link.springer.com/10.1007/978-1-59745-281-6_30

40. Tamang S. Proteomics: Types, Methods, Steps, Applications [Internet]. 2023 [citado 22 de julio de 2024]. Disponible en: https://microbenotes.com/proteomics/

41. He Q, Chiu J. Proteomics in biomarker discovery and drug development. J of Cellular Biochemistry. agosto de 2003;89(5):868-86.

42. Rajcevic U. Proteomics strategies for target identification and biomarker discovery in cancer. Front Biosci. 2009;Volume(14):3292.

43. Zhang Z, Wu S, Stenoien DL, Paša-Tolić L. High-Throughput Proteomics. Annual Rev Anal Chem. 12 de junio de 2014;7(1):427-54.

44. Duong VA, Lee H. Bottom-Up Proteomics: Advancements in Sample Preparation. IJMS. 10 de marzo de 2023;24(6):5350.

45. Cupp-Sutton KA, Wu S. High-throughput quantitative top-down proteomics. Mol Omics. 2020;16(2):91-9.

46. Gregorich ZR, Chang YH, Ge Y. Proteomics in heart failure: top-down or bottom-up? Pflugers Arch - Eur J Physiol. junio de 2014;466(6):1199-209.

47. B4B: Module 8 - Data acquisition overview [Internet]. 2020 [citado 22 de julio de 2024]. Disponible en: https://www.youtube.com/watch?v=PYE8osIlnM4

48. Elhamraoui Z. What are DIA and DDA and how does the data looks like ? [Internet]. Medium. 2024 [citado 22 de julio de 2024]. Disponible en: https://medium.com/@zahraelhamraoui1997/what-are-dia-and-dda-and-how-does-the-data-looks-like-e18a3a532142

49. Megger DA, Bracht T, Meyer HE, Sitek B. Label-free quantification in clinical proteomics. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. agosto de 2013;1834(8):1581-90.

50. PROTEÓMICA CUANTITATIVA: CUANTIFICACIÓN BASADA EN MS2 [Internet]. SERVICIOS DE APOYO A LA INVESTIGACIÓN. 2022 [citado 23 de julio de 2024]. Disponible en: https://saishnp.com/2022/03/21/proteomica-cuantitativa-cuantificacion-basada-en-ms2/

51. Aggarwal K. Shotgun proteomics using the iTRAQ isobaric tags. Briefings in Functional Genomics and Proteomics. 10 de mayo de 2006;5(2):112-20.

52. Lee YH. An overview of meta-analysis for clinicians. Korean J Intern Med. 1 de marzo de 2018;33(2):277-83.

53. Adamowicz K, Arend L, Maier A, Schmidt JR, Kuster B, Tsoy O, et al. Proteomic meta-study harmonization, mechanotyping and drug repurposing candidate prediction with ProHarMeD. npj Syst Biol Appl. 10 de octubre de 2023;9(1):49.

54. Ng S, Masarone S, Watson D, Barnes MR. The benefits and pitfalls of machine learning for biomarker discovery. Cell Tissue Res. octubre de 2023;394(1):17-31.

55. Hou W, Janech MG, Sobolesky PM, Bland AM, Samsuddin S, Alazawi W, et al. Proteomic screening of plasma identifies potential noninvasive biomarkers associated with significant/advanced fibrosis in patients with nonalcoholic fatty liver disease. Bioscience Reports. 31 de enero de 2020;40(1):BSR20190395.

56. Junker AE, Gluud L, Holst JJ, Knop FK, Vilsbøll T. Diabetic and nondiabetic patients with nonalcoholic fatty liver disease have an impaired incretin effect and fasting hyperglucagonaemia. J Intern Med. mayo de 2016;279(5):485-93.

57. Bertran L, Rusu EC, Guirro M, Aguilar C, Auguet T, Richart C. Circulating proteomic profiles in women with morbid obesity compared to normal-weight women. Journal of Proteomics. enero de 2025;310:105317.

58. Pang Y, Kartsonaki C, Lv J, Fairhurst-Hunter Z, Millwood IY, Yu C, et al. Associations of Adiposity, Circulating Protein Biomarkers, and Risk of Major Vascular Diseases. JAMA Cardiol. 1 de marzo de 2021;6(3):276.

59. Lau ES, Paniagua SM, Guseh JS, Bhambhani V, Zanni MV, Courchesne P, et al. Sex Differences in Circulating Biomarkers of Cardiovascular Disease. Journal of the American College of Cardiology. septiembre de 2019;74(12):1543-53.

60. Föll M, Fahrner M. Galaxy Training Network. Galaxy Training Network; [citado 25 de julio de 2024]. MaxQuant and MSstats for the analysis of label-free data. Disponible en: https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/maxquant-msstats-dda-lfq/tutorial.html

61. Kong W, Hui HWH, Peng H, Goh WWB. Dealing with missing values in proteomics data. Proteomics. diciembre de 2022;22(23-24):2200092.

62. VIM documentation [Internet]. [citado 30 de julio de 2024]. Disponible en: https://cran.r-project.org/web/packages/VIM/index.html

63. ggplot2 documentation [Internet]. [citado 30 de julio de 2024]. Disponible en: https://rdrr.io/cran/ggplot2/man/

64. Gordon Smyth [Cre A. limma [Internet]. Bioconductor; 2017 [citado 25 de julio de 2024]. Disponible en: https://bioconductor.org/packages/limma

65. ebayes function - RDocumentation [Internet]. [citado 26 de julio de 2024]. Disponible en: https://www.rdocumentation.org/packages/limma/versions/3.28.14/topics/ebayes

66. rem_mv: A function to perform the Random Effect Model (REM)... in MetaVolcanoR: Gene Expression Meta-analysis Visualization Tool [Internet]. [citado 29 de julio de 2024]. Disponible en: https://rdrr.io/bioc/MetaVolcanoR/man/rem_mv.html

67. Bioconductor [Internet]. [citado 30 de julio de 2024]. MetaVolcanoR. Disponible en: http://bioconductor.org/packages/MetaVolcanoR/

68. Bioconductor [Internet]. [citado 29 de julio de 2024]. clusterProfiler. Disponible en: http://bioconductor.org/packages/clusterProfiler/

69. Bioconductor [Internet]. [citado 29 de julio de 2024]. org.Hs.eg.db. Disponible en: http://bioconductor.org/packages/org.Hs.eg.db/

70. Bioconductor [Internet]. [citado 29 de julio de 2024]. ReactomePA. Disponible en: http://bioconductor.org/packages/ReactomePA/

71. gsePathway function - RDocumentation [Internet]. [citado 29 de julio de 2024]. Disponible en: https://www.rdocumentation.org/packages/ReactomePA/versions/1.16.2/topics/gsePathway

72. What is Cytoscape? [Internet]. [citado 30 de julio de 2024]. Disponible en: https://cytoscape.org/what_is_cytoscape.html

73. Cytoscape App Store - ClueGO [Internet]. [citado 30 de julio de 2024]. Disponible en: https://apps.cytoscape.org/apps/cluego

74. Cytoscape App Store - stringApp [Internet]. [citado 30 de julio de 2024]. Disponible en: https://apps.cytoscape.org/apps/stringapp

75. scikit-learn [Internet]. [citado 30 de julio de 2024]. Disponible en: https://www.researchgate.net/publication/51969319_Scikit-learn_Machine_Learning_in_Python

76. Tenenbaum D. KEGGREST [Internet]. Bioconductor; 2017 [citado 30 de diciembre de 2024]. Disponible en: https://bioconductor.org/packages/KEGGREST

77. Rus H, Cudrici C, Niculescu F. The Role of the Complement System in Innate Immunity. IR. 2005;33(2):103-12.

78. Wang F, Kohan AB, Lo CM, Liu M, Howles P, Tso P. Apolipoprotein A-IV: a protein intimately involved in metabolism. Journal of Lipid Research. agosto de 2015;56(8):1403-18.

79. Mast AE, Wolberg AS, Gailani D, Garvin MR, Alvarez C, Miller JI, et al. SARS-CoV-2 suppresses anticoagulant and fibrinolytic gene expression in the lung. eLife. 8 de marzo de 2021;10:e64330.

80. Hamm A, Veeck J, Bektas N, Wild PJ, Hartmann A, Heindrichs U, et al. Frequent expression loss of Inter-alpha-trypsin inhibitor heavy chain (ITIH) genes in multiple human solid tumors: A systematic expression analysis. BMC Cancer. diciembre de 2008;8(1):25.

81. Guo Z, Fan X, Yao J, Tomlinson S, Yuan G, He S. The role of complement in nonalcoholic fatty liver disease. Front Immunol. 29 de septiembre de 2022;13:1017467.

82. Yaseen S, Demopulos G, Dudler T, Yabuki M, Wood CL, Cummings WJ, et al. Lectin pathway effector enzyme mannan-binding lectin-associated serine protease-2

can activate native complement C3 in absence of C4 and/or C2. FASEB j. mayo de 2017;31(5):2210-9.

83. Efectes de la reperfusió després de la isquèmia cerebral: proteïnes d'estrès cel·lular, estrès oxidatiu i activació del complement. En [citado 30 de julio de 2024]. Disponible en: http://hdl.handle.net/10803/145439

84. Jiang Q, Wang N, Lu S, Xiong J, Yuan Y, Liu J, et al. Targeting hepatic ceruloplasmin mitigates nonalcoholic steatohepatitis by modulating bile acid metabolism. Liu F, editor. Journal of Molecular Cell Biology. 4 de abril de 2024;15(9):mjad060.

85. Nobili V, Siotto M, Bedogni G, Ravà L, Pietrobattista A, Panera N, et al. Levels of Serum Ceruloplasmin Associate With Pediatric Nonalcoholic Fatty Liver Disease. J pediatr gastroenterol nutr. abril de 2013;56(4):370-5.

86. Wang Q, Zhou D, Wang M, Zhu M, Chen P, Li H, et al. A Novel Non-Invasive Approach Based on Serum Ceruloplasmin for Identifying Non-Alcoholic Steatohepatitis Patients in the Non-Diabetic Population. Front Med. 20 de junio de 2022;9:900794.

87. Arefhosseini S, Pouretedal Z, Tutunchi H, Ebrahimi-Mameghani M. Serum copper, ceruloplasmin, and their relations to metabolic factors in nonalcoholic fatty liver disease: a cross-sectional study. European Journal of Gastroenterology & Hepatology. abril de 2022;34(4):443-8.

88. Pezzino S, Luca T, Castorina M, Puleo S, Latteri S, Castorina S. Role of Perturbated Hemostasis in MASLD and Its Correlation with Adipokines. Life. 7 de enero de 2024;14(1):93.

89. Coagulación, genética y reestenosis postangioplastia [Internet]. [citado 30 de julio de 2024]. Disponible en: https://www.revespcardiol.org/es-coagulacion-genetica-y-reestenosis-posta-articulo-X030089329700405X

90. Nawaz SS, Siddiqui K. Plasminogen activator inhibitor-1 mediate downregulation of adiponectin in type 2 diabetes patients with metabolic syndrome. Cytokine: X. marzo de 2022;4(1):100064.

91. Baker SK, Strickland S. A critical role for plasminogen in inflammation. Journal of Experimental Medicine. 6 de abril de 2020;217(4):e20191865.

92. Liao Z, Zhang H, Liu F, Wang W, Liu Y, Su C, et al. m^6 A-Dependent ITIH1 Regulated by TGF-β Acts as a Target for Hepatocellular Carcinoma Progression. Advanced Science. noviembre de 2024;11(42):2401013.

93. Talari NK, Mattam U, Kaminska D, Sotomayor-Rodriguez I, Rahman AP, Péterfy M, et al. Hepatokine ITIH3 protects against hepatic steatosis by downregulating mitochondrial bioenergetics and de novo lipogenesis. iScience. mayo de 2024;27(5):109709.

94. Carli F, Della Pepa G, Sabatini S, Vidal Puig A, Gastaldelli A. Lipid metabolism in MASLD and MASH: From mechanism to the clinic. JHEP Reports. diciembre de 2024;6(12):101185.

95. Pandey N, Anand SK, Kaur H, Richard KSE, Chandaluri L, Butler ME, et al. Enhanced venous thrombosis and hypercoagulability in murine and human metabolic dysfunction-associated steatohepatitis. Journal of Thrombosis and Haemostasis. diciembre de 2024;22(12):3572-80.

10. Autoavaluació

Aquest Treball de Fi de Grau ha suposat un repte tant des del punt de vista tècnic com metodològic, ja que ha requerit integrar coneixements de biologia molecular, bioinformàtica, estadística i intel·ligència artificial. Considero que el treball ha assolit satisfactòriament els objectius plantejats, especialment pel que fa a la identificació de biomarcadors diferencials entre SS i MASH i el desenvolupament d'un model predictiu mitjançant AA.

Destaco positivament l'estructura metodològica rigorosa seguida: des d'una cerca bibliogràfica sistemàtica i una metaanàlisi robusta, fins al disseny i validació d'un model de classificació amb dades reals. A més, s'han aplicat pràctiques habituals en la recerca científica, com l'estandardització de dades, la imputació de valors perduts i l'ús de validació creuada.

Un dels principals punts forts ha estat la capacitat d'adaptació davant limitacions reals, com ara la manca de dades completes o l'alta variabilitat entre cohorts. L'ús de ràtios proteiques funcionals com a característiques ha estat una decisió fonamentada i innovadora, que ha permès millorar el rendiment del model.

Com a aspectes millorables, reconec que el model predictiu requereix una validació addicional amb cohorts externes abans de ser traslladat a l'àmbit clínic.

En conjunt, considero que aquest treball reflecteix un alt nivell d'aprenentatge i maduresa investigadora, i m'ha permès desenvolupar habilitats transversals essencials per a la meva futura carrera científica.