



ACADEMIA ROMÂNĂ

ȘCOALA DE STUDII AVANSATE A ACADEMIEI ROMÂNE

INSTITUTUL DE BIOCHIMIE

**REZUMAT
TEZĂ DE DOCTORAT**

Analiza modificărilor repertoriului imun în timpul îmbătrânirii

CONDUCĂTOR ȘTIINȚIFIC:
CS I Dr. Andrei-J. Petrescu

DOCTORAND:
Eugen Ursu

București, 2025

Cuprins

Cuprins.....	1
I. Introducere.....	1
1. Îmbătrânirea.....	1
2. Biologie computațională și inteligență artificială.....	3
II. Expresia genică la specii diferite dezvăluie noi gene candidate implicate în îmbătrânire și longevitate. 5	5
Prezentare generală.....	5
1. Introducere.....	5
2. Metode.....	5
3. Rezultate și Discuții.....	7
4. Concluzii.....	10
III. Semnătura îmbătrânirii în fibroza pulmonară.....	11
IV. Comunicarea intercelulară este perturbată odată cu îmbătrânirea.....	12
Prezentare generală.....	12
1. Introducere.....	13
2. Metode.....	13
3. Rezultate și Discuții.....	14
4. Concluzii.....	19
V. Compoziția datelor de antrenament negative este critică pentru învățarea repertoriilor imune.....	19
1. Introducere.....	20
2. Metode.....	21
3. Rezultate și Discuții.....	22
VI. Considerații finale.....	29
Bibliografie.....	29

I. Introducere

1. Îmbătrânirea

Îmbătrânirea este un proces biologic complex, cu un impact profund asupra sănătății și a riscului de boală. Cercetările contemporane o consideră un fenomen reglementat, conturat de cele nouă „semne distinctive ale îmbătrânirii” interconectate (instabilitatea genomică, scurtarea telomerilor, modificările epigenetice, pierderea proteostazei, detectarea dereglată a nutrienților, disfuncția mitocondrială, senescența celulară, epuizarea celulelor stem și comunicarea intercelulară alterată)(1). Acest cadru este fundamental, ghidând investigarea mecanismelor celulare și a țintelor de intervenție. Căile cheie, conservate, care influențează îmbătrânirea includ semnalizarea insulină/IGF-1 și căile mTOR(2,3). Modularea experimentală a acestora, alături de intervenții precum restricția calorică (RC) și compuși care imită RC (rapamicina, metformina, terapiile bazate pe NAD+), a extins durata de viață la diverse modele(2,3). Înțelegerea îmbătrânirii are implicații largi pentru sănătatea publică, deoarece vârsta este principalul factor de risc pentru bolile cronice(4). Domeniul gerontologiei caută să vizeze mecanismele îmbătrânirii pentru a întârzia sau a preveni simultan multiple patologii(5).

Îmbătrânirea este strâns legată de dezvoltarea majorității bolilor cronice (tulburări

neurodegenerative, boli cardiovasculare, sindroame metabolice, cancer)(6). Schimbarea demografică către populații mai în vârstă exercită o presiune asupra sistemelor de sănătate și a cheltuielilor, deoarece o mare parte din resurse este alocată gestionării bolilor cronice, legate de vârstă, în special la sfârșitul vieții(4). Sunt necesare strategii proactive, preventive.

Obiectivul cercetării în domeniul îmbătrânirii se mută de la simpla extindere a **duratei de viață (lifespan)** la extinderea **duratei de sănătate (healthspan)** (anii trăiți în stare bună de sănătate)(7). Medicina actuală tratează de obicei bolile cronice izolat, ignorând rădăcina lor comună în procesul fundamental de îmbătrânire(6). O abordare mai holistică implică intervenția în mecanismele biologice ale îmbătrânirii în sine pentru a îmbunătăți rezultatele generale de sănătate și a reduce multimorbiditatea(1). Intervențiile care vizează căi legate de îmbătrânire, cum ar fi rapamicina, senoliticele și amplificatorii de NAD+, sunt studiate pentru capacitatea lor de a îmbunătăți sănătatea pe multiple domenii, acționând asupra semnelor distinctive ale îmbătrânirii (reciclare celulară afectată, inflamație, disfuncție mitocondrială)(8).

Îmbătrânirea este un proces condus de o constelație de procese biologice de deteriorare, distilate în proeminentul model al „semnelor distinctive ale îmbătrânirii” (“hallmarks of aging”) de către López-Otín et al.(1). Acest cadru conectează diverse evenimente celulare, cum ar fi deteriorarea ADN-ului, senescența, întreținerea deficitară a proteinelor și căile metabolice dereglate (semnalizarea insulină/IGF-1, mTOR, sirtuine)(1,9). Din perspectivă evolutivă, îmbătrânirea este explicată de teoriile acumulării mutațiilor, pleiotropiei antagoniste și a somei de unică folosință (“disposable soma”), care se concentrează pe presiunea selectivă și compromisuri (trade-offs)(10). Viziunile mecaniciste, precum teoria radicalilor liberi, se concentrează pe deteriorare, deși înțelegerea actuală este mai nuanțată(11).

Tehnologiile Omice (genomică, transcriptomică, proteomică, metabolomică, epigenomică) oferă o viziune cuprinzătoare, multistratificată a modificărilor moleculare în timpul îmbătrânirii, identificând biomarkeri și urmărind declinul asociat vârstei(12). Biologia computațională și bioinformatica sunt esențiale pentru interpretarea acestor seturi vaste de date, susținând analiza căilor metabolice, modelarea rețelelor și dezvoltarea de predictorii ai vârstei biologice, cum ar fi ceasurile epigenetice(13,14). Inteligența artificială (IA) și învățarea automată (ML) îmbunătățesc în continuare analiza prin descoperirea tiparelor subtile, modelarea traiectoriilor de îmbătrânire, identificarea profilurilor de risc și prioritizarea țăintelor terapeutice, accelerând calea către medicina de precizie pentru îmbătrânire(15). Lucrarea noastră valorifică aceste tehnologii pentru a investiga îmbătrânirea sistemului imunitar și progresia fibrozei.

Biologia comparată oferă perspective prin studierea speciilor cu durate de viață diverse (de exemplu, șobolani-cârțiță golași, balene de Groenlanda), evidențiind mecanisme de longevitate conservate sau adaptate în mod unic, cum ar fi proteostaza îmbunătățită, repararea ADN-ului și răspunsuri robuste la stres(16,17). Studiile inter-specii ajută la distingerea modificărilor legate de vârstă de timpul cronologic, observând modul în care diferite specii gestionează îmbătrânirea pe intervale de timp foarte diferite, arătând adesea că mamiferele cu viață lungă suprimă inflamația și mențin integritatea genomică mai eficient(18). Analiza transcriptomică (RNA-Seq) între specii identifică semnături ale expresiei genice legate de longevitate, cum ar fi reglarea în jos a căilor de creștere și reglarea în sus a genelor de întreținere, oferind indicii despre căile moleculare conservate relevante pentru sănătatea umană(19). **Capitolul II** al

acestei teze utilizează transcriptomica inter-specii pentru a identifica modificări ale expresiei genice, comune și specifice speciei, legate de îmbătrânire, apărarea împotriva stresului și reglarea imunitară.

Fibroza (depunerea excesivă de matrice extracelulară) este o caracteristică esențială a disfuncției tisulare legate de vârstă în organe precum inima, rinichii, ficatul și plămânii(20). Modificările asociate îmbătrânirii (leziuni oxidative, inflamație cronică, activarea fibroblastelor) creează un mediu fibrogenic. Fibroza pulmonară idiopatică (FPI) este un exemplu primordial, asociat vârstei, afectând predominant persoanele de peste 60 de ani(21). Mecanistic, îmbătrânirea promovează fibroza pulmonară prin senescenta celulelor epiteliale, fenotipul secretor asociat senescentei (SASP) pro-fibrotic, disfuncția mitocondrială și semnalizarea inflamatorie persistentă (TGF- β , IL-6)(22,23). **Capitolul III** prezintă o lucrare colaborativă care studiază semnele îmbătrânirii în fibroza pulmonară utilizând un model murin de fibroză pulmonară indusă de bleomicină.

Comunicarea intercelulară perturbată este o marcă sistemică a îmbătrânirii, manifestată ca inflamație cronică de grad scăzut (inflammaging), supraveghere imună afectată și semnalizare endocrină modificată(1). „Inflammaging”(24) este alimentat parțial de celulele senescente care eliberează factori SASP (citokine, chemokine) ce perturbă funcția tisulară și imună, creând o buclă de feedback dăunătoare(25). Imunosenescenta și modificările în semnalizarea endocrină contribuie în continuare la disfuncție(26). Comunicarea alterată este un motor cheie al îmbătrânirii sistemice. În **Capitolul IV**, descriu contribuțiile mele: 1) dezvoltarea unui instrument de bioinformatică, scDiffCom, pentru inferența modificărilor în comunicarea intercelulară din datele scRNA-Seq și 2) utilizarea scDiffCom pentru a construi un atlas al modificărilor legate de vârstă în comunicările intercelulare la șoareci (scAgeCom).

Imunosenescenta – declinul sistemului imunitar legat de vârstă – are ca rezultat răspunsuri diminuate la agenții patogeni, eficacitate redusă a vaccinurilor și boli asociate vârstei crescute(27). O componentă de bază este deteriorarea răspunsului umoral adaptativ, marcată de modificări în compartimentul celulelor B și de o diversitate redusă a repertoriului de anticorpi(28). Secvențierea Repertoriului Receptorilor Imuni Adaptativi (AIRR-Seq) permite o analiză profundă a secvențelor receptorilor celulelor B și T(29). Cu toate acestea, studiile de îmbătrânire se confruntă cu provocări din cauza seturilor de date transversale și a protocoalelor analitice inconsistente(30). Analiza tradițională a repertoriului nu reușește adesea să surprindă modelele complexe, neliniare ale sistemului imunitar în curs de îmbătrânire, ceea ce necesită o tranziție către învățarea automată (ML). **Capitolul V** descrie o lucrare colaborativă în imunologia computațională axată pe impactul datelor de antrenare a repertoriului imunitar asupra modelelor ML, care este crucială pentru dezvoltarea de instrumente pentru o mai bună înțelegere viitoare a dinamicii repertoriului imunitar legate de vârstă.

2. Biologie computațională și inteligență artificială

Tehnologiile omice de înaltă performanță (high-throughput), cum ar fi genomica, transcriptomica, proteomica, metabolomica și epigenomica, au generat seturi extinse de date biologice(13). Amploarea și complexitatea acestor date necesită metode computaționale sofisticate. Biologia computațională și bioinformatica oferă infrastructura esențială pentru

gestionarea datelor, preprocesare și interpretare statistică. Mai recent, învățarea automată (Machine Learning - ML) și inteligența artificială (IA) au devenit proeminente, excelând în detectarea de tipare complexe, adesea neliniare, în datele omice(31). Influența IA se întinde de la domenii precum prezicerea activității de reglare a genelor din secvențele ADN(32) până la analiza populațiilor celulare în transcriptomica unicelulară. Această intersecție facilitează tranziția de la modele descriptive la modele predictive și mecanice, în special în cercetarea îmbătrânirii.

Secvențierea ARN (RNA sequencing - RNA-Seq) este un instrument transformator pentru profilarea globală a expresiei genice, oferind sensibilitate superioară și un interval dinamic mai larg în comparație cu metodele mai vechi(33). Aceasta permite analize ulterioare, cum ar fi expresia diferențială și construcția de rețele, oferind o perspectivă asupra căilor biologice și mecanismelor bolilor(34). Apariția secvențierii ARN unicelulare (scRNA-Seq) a revoluționat și mai mult domeniul, rezolvând expresia genică la nivelul celulei individuale. Spre deosebire de RNA-Seq în masă (bulk), scRNA-Seq dezvăluie diversitatea și eterogenitatea din țesuturile complexe, făcând posibilă identificarea tipurilor de celule rare și a traiectoriilor de origine(35). Cercetarea mea a valorificat în mod extensiv puterea de înaltă rezoluție a acestor tehnologii, iar analizele și descoperirile prezentate în Capitolele II, III și IV s-au bazat pe perspectivele profunde oferite de RNA-Seq și scRNA-Seq.

Învățarea automată (ML) este parte integrantă a științelor vieții moderne pentru detectarea și predicția de tipare, în special în domenii omice și imagistică biomedicală. ML-ul timpuriu s-a concentrat pe acuratețe ridicată, folosind adesea modele de tip „cutie neagră” (black box), cum ar fi rețelele neuronale profunde(36). Din cauza preocupărilor din domenii precum medicina, accentul s-a mutat către învățarea automată interpretabilă. Tehnici precum SHAP și LIME permit cercetătorilor să înțeleagă modul în care modelele iau decizii, legând intrări specifice (de exemplu, gene) de rezultate(37). Această interpretabilitate a fost crucială pentru descoperirea genelor cauzale și decodarea logicii biologice(38). În cercetarea îmbătrânirii, interpretabilitatea este vitală pentru traducerea predicțiilor în semnificație biologică. În propria mea lucrare, interpretabilitatea a fost un pilon central al investigațiilor din Capitolele II, IV și V.

Diversitatea sistemului imunitar adaptiv se bazează pe milioane de receptori unici ai celulelor B (BCR) și anticorpi. Secvențierea Repertoarului de Receptori Imuni Adaptivi (Adaptive Immune Receptor Repertoire Sequencing - AIRR-Seq) permite profilarea în profunzime a acestei diversități, revoluționând studiul imunității umorale într-o varietate de condiții(39). AIRR-Seq surprinde milioane de secvențe BCR în paralel, permițând reconstrucția clonotipurilor și a relațiilor de origine (liniaj) (29). Analiza acestor date vaste și complexe necesită cadre computaționale robuste pentru sarcini precum alinierea secvențelor, gruparea (clustering) și identificarea mutațiilor (30). Cercetarea mea contribuie la acest domeniu prin avansarea tehnicilor de învățare automată pentru AIRR-Seq, concentrându-se pe provocări fundamentale, cum ar fi selectarea datelor de antrenare negative pentru a construi modele mai relevante din punct de vedere biologic. Acest efort fundamental, descris în Capitolul V, urmărește în cele din urmă să reducă lacunele în cartografierea dinamicii repertoriului imun în timpul îmbătrânirii.

II. Expresia genică la specii diferite dezvăluie noi gene candidate implicate în îmbătrânire și longevitate

*Kulaga, A. Y., ***Ursu, E.**, *Toren, D., Tyshchenko, V., Guinea, R., Pushkova, M., Fraifeld, V. E. & Tacutu, R. *Machine Learning Analysis of Longevity-Associated Gene Expression Landscapes in Mammals. Int. J. Mol. Sci.* 22, (2021).

* denota first co-authorship

Declarație de Contribuții

Am contribuit substanțial la designul studiului, la dezvoltarea și implementarea cadrului, la procesarea datelor și la interpretarea rezultatelor. În calitate de co-prim autor, am împărțit responsabilitățile de bază în mod egal cu Anton Kulaga și Dmitri Toren. Împreună, am dezvoltat în colaborare componentele conceptuale și analitice ale studiului, am co-scriș toate secțiunile manuscrisului și am lucrat în comun la deciziile metodologice și interpretative cheie. Studiul a fost realizat sub coordonarea și supravegherea Dr. Robi Tăcutu și Prof. Vadim E. Fraifeld.

Prezentare generală

În domeniul cercetării îmbătrânirii, o provocare centrală este înțelegerea modului în care variațiile în expresia genică influențează longevitatea observată în rândul diferitelor specii. Acest studiu a urmărit să clarifice relațiile complexe dintre diferențele transcriptomice și durata maximă de viață (DMV) prin abordarea limitărilor modelelor liniare convenționale și a impactului factorilor de confuzie, cum ar fi rata metabolică, perioada de gestație și masa corporală.

1. Introducere

Investigarea diversității în DVS (Durata de Viață Maximă a Speciei) alături de transcripția genică la nivelul speciilor oferă perspective valoroase asupra mecanismelor evolutive ale longevității. Cercetările recente au evidențiat diferențe distincte în profilurile de expresie genică între mamiferele cu viață lungă și cele cu viață scurtă(40–43). În special, speciile cu viață excepțional de lungă prezintă niveluri crescute de gene asociate cu întreținerea și repararea ADN-ului, ubiquitinarea, răspunsurile imune, apoptoza și autofagia(44). Supraexpresia genelor de reparare a ADN-ului este evidențiată în mod constant în studiile de culturi celulare inter-specii(45), iar în câteva organe de mamifere, expresia genelor de răspuns imun se corelează pozitiv cu DVS(42). Adaptările pro-longevitate la nivel transcriptomic sunt clar documentate la specii precum lilieci(46), șobolanii-cârțiță golași(41,47) și balenele(40,44). Aceste dovezi subliniază potențialul studiilor transcriptomice comparative de a descoperi baza genetică a longevității.

Studiile transcriptomice comparative tradiționale s-au bazat adesea pe metode liniare, ignorând procesele biologice non-liniare. Progresele recente oferă, totuși, alte soluții.

2. Metode

Am încadrat problema ca o sarcină de reducere a caracteristicilor, căutând un set minim, extrem

de predictiv de gene asociate cu variația Duratei Maxime de Viață (DMV) în diferite organe. Protocolul nostru integrează date transcriptomice de la specii multiple pentru a detecta gene asociate longevității (GAL) folosind trei metode complementare: modelare liniară pentru tendințe specifice organelor, LightGBM combinat cu SHAP pentru evaluarea influenței și interacțiunilor genelor și modelare cu rețele Bayes-iene pentru a deduce potențiale legături cauzale cu DMV.

Acest studiu a inclus specii de mamifere cu date RNA-Seq disponibile pentru țesuturi sănătoase de ficat, rinichi, plămân, creier sau inimă din Arhiva de Citiri Secvențiale NCBI (NCBI Sequence Read Archive), cu adnotări ale transcriptomului preluate din Baza de Date Ensembl Compara (48). Un total de 408 de probe de la 38 de specii, acoperind cinci organe, au fost procesate folosind un protocol standardizat de cuantificare RNA-Seq. Datele de ortologie și adnotările transcriptomului au fost preluate din Ensembl versiunea 99. În total au rezultat 11831 de gene.

Fastp (versiunea 0.20.1)(49) a gestionat controlul calității și *trimming*-ul. Cuantificarea transcriptelor a fost realizată cu Salmon (versiunea 1.4.0)(50), iar tximport (versiunea 3.12) a agregat expresia la nivel de genă. Numărul brut de citiri a fost normalizat utilizând metoda transcriptelor per milion (TPM) pentru analiza comparativă inter-specii.

Au fost dezvoltate modele liniare specifice organelor pentru a identifica genele asociate cu DMV. Activitatea căilor metabolice a fost evaluată folosind analiza de îmbogățire a seturilor de gene pentru o singură probă (ssGSEA) cu căile KEGG(51).

Pentru a explora tipare non-liniare, a fost utilizată o selecție inversă în doi pași, combinând LightGBM și SHAP. Șase modele de regresie independente au prezis inițial diferite trăsături de istorie a vieții (inclusiv DMV) din expresia genică. Genele identificate în orice model au fost grupate pentru o a doua fază pentru a rafina setul asociat cu DMV. A fost aplicată o validare încrucișată riguroasă pe 5 sub-eșantioane (*5-fold cross-validation*) cu stratificare sortată, repetată de zece ori, cerând valori SHAP non-zero pentru semnificație. O strategie strictă de împărțire a datelor a asigurat că predicțiile s-au bazat pe tipare genice, nu pe identificarea speciei. Primele 15 gene au fost selectate pe baza cotului graficului de importanță a caracteristicilor SHAP (Fig. Suplimentară S4 din Kulaga, Ursu, Toren et al.(52)).

Rețelele Bayes-iene au fost folosite pentru a cartografia independența condiționată și pentru a deduce potențiale conexiuni cauzale cu DMV. Algoritmul SES(53) a fost utilizat pentru selecția caracteristicilor, identificând genele în învelișul Markov al DMV. SES a fost aplicat setului de antrenare cu date imputate, iar semnătura genică rezultată a fost evaluată de un model LightGBM antrenat pe datele neimputate, selectând semnătura cu cea mai mică RMSE.

Cele trei abordări de modelare au generat liste clasificate bazate pe metrici distincte: (1) LightGBM-SHAP: frecvența SHAP mediu absolut non-zero, *Kendall's tau* și SHAP mediu absolut; (2) Regresia Liniară: R² maxim; și (3) Rețele Bayes-iene: frecvența relativă în semnături. O clasificare compozită finală a fost derivată prin însumarea rangurilor a șase metrici cheie: frecvența de selecție, corelația *Kendall's tau*, valoarea SHAP medie absolută, cel mai mare R² liniar, frecvența relativă în rețelele Bayes-iene și indicatorul mențiunii în GenAge(54). Un model liniar multinivel Bayesian cu coeficienți aleatori specifici organelor a fost construit pentru a analiza setul final de gene selectate.

Pe baza clasificării compozite și a unui grafic *elbow* (Fig. Suplimentară S4 din Kulaga, Ursu, Toren et al.(52)), au fost identificate primele 15 gene. O serie de modele LightGBM-SHAP, încorporând subseturi de la primele 5 până la primele 15 gene, au fost dezvoltate. Evaluarea preciziei modelului a indicat faptul că modelul cu primele 6 gene a oferit cel mai bun compromis între simplitate și performanță, arătând o reducere notabilă a pierderii Huber (0.8) în comparație cu modelul cu primele 5 gene.

3. Rezultate și Discuții

Un set de date de expresie genică inter-specii, compilat din date RNA-Seq disponibile public, oferă un cadru comparativ extins. Acest set de date conține date de expresie genică pentru cinci organe majore (ficat, rinichi, plămân, creier și inimă) din 408 de probe, provenind de la 41 de specii de mamifere. Variabilele specifice fiecărei specii, incluzând DMV, masa corporală, temperatura, rata metabolică, perioada de gestație și conținutul GC al ADN-ului mitocondrial, au fost încorporate după normalizare, deoarece reprezintă factori cheie ai longevității(55–57). Pentru a explora legătura dintre expresia genică și DMV, am utilizat regresia liniară, modelarea interpretabilă LightGBM-SHAP și analiza rețelelor bayesiene. Integrarea rezultatelor din aceste modele a identificat gene care se clasează în mod constant ca fiind principalii predictorii ai DMV (vezi Fig. 1).

Modelele liniare au fost utilizate pentru a evalua relația dintre expresia a 11831 de ortologi conservate evolutiv și Longevitatea Maximă a Speciei (LMS) la 33 de specii de mamifere (Fig. 2a). Numărul de gene asociate semnificativ cu LMS ($FDR < 0,05$, $R^2 > 0,3$) a variat în funcție de organ, atingând un vârf în plămân (756) și cel mai scăzut nivel în rinichi (154). Valorile mediane R^2 au fost consistente (0,35–0,38).

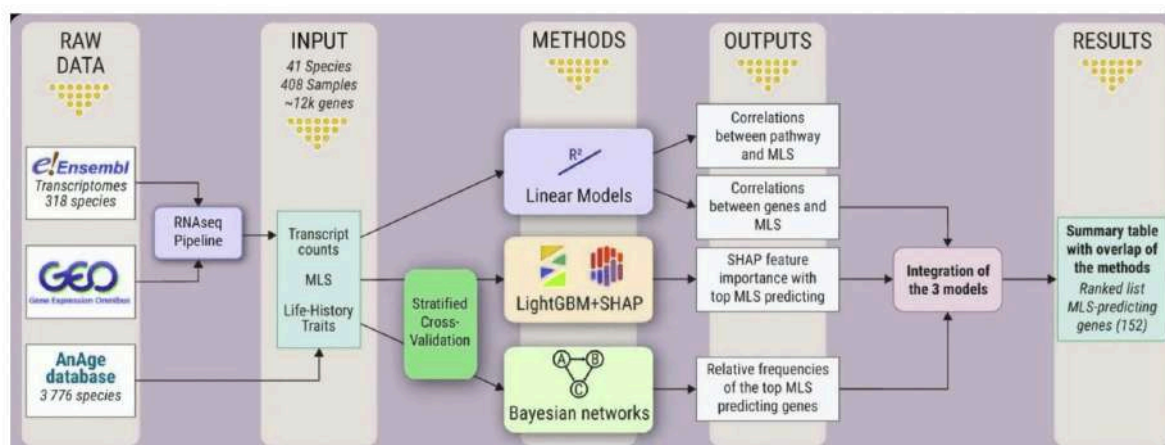


Figura 1. Reprezentarea schematică a fluxului de lucru pentru analiza între specii.

Figură adaptată din Figura 1 din Kulaga, Ursu, Toren. et al. (2021), utilizată sub licența CC BY 4.0.

Doar trei gene (CRYGS, TCFL5 și SPATA20) au fost corelate pozitiv cu DVM în toate cele cinci organe. Concentrându-ne pe creier, ficat și rinichi, numărul de gene asociate cu DVM a crescut la 12, incluzând SPATA20, TCFL5 și CRYGS (Tabel Supl. S2 din Kulaga, Ursu, Toren et al.(52)).

Multe gene asociate cu DVM au fost corelate și cu alte trăsături de istorie a vieții (de exemplu,

masa corporală, rata metabolică), sugerând asocieri indirecte. Doar câteva gene au fost corelate în mod unic cu DVM: una în ficat (CERS4), patru în inimă și 131 în plămâni; niciuna în creier sau rinichi. Totuși, rezultatele pentru inimă și plămâni ar trebui interpretate cu prudență din cauza dimensiunilor mai mici ale eșantioanelor.

Am utilizat abordarea proiecției semnăturii (ssGSEA) pentru a analiza relația dintre activitatea căilor biologice, estimată din expresia genică în organe, și durata de viață maximă (DVM) (Fig. 2b). Analiza noastră s-a concentrat pe căi de îmbătrânire/longevitate deja stabilite, incluzând semnalizarea mTOR și a insulinei, repararea ADN-ului, proteoliza mediata de ubiquitină și adeziunea focală(58–60).

Deși semnalizarea mTOR nu a fost corelată semnificativ cu DVM, multiple căi legate de repararea ADN-ului (repararea erorilor de împerechere – *mismatch*, excizia nucleotidică, excizia bazelor, recombinația omoloagă și joncțiunea capetelor non-omoloage) au prezentat corelații pozitive robuste. Câteva alte căi au fost corelate pozitiv și negativ cu DVM. În mod neașteptat, căi care nu sunt legate în mod tradițional de longevitate, cum ar fi apoptoza, moleculele de adeziune celulară și semnalizarea ErbB, au fost corelate pozitiv cu DVM. Am validat căi de longevitate cunoscute (apoptoza, repararea ADN-ului, răspunsurile imune) și am evidențiat unele mai puțin explorate – incluzând semnalizarea PPAR, metabolismul glutatationului și semnalizarea ErbB – ca fiind domenii promițătoare.

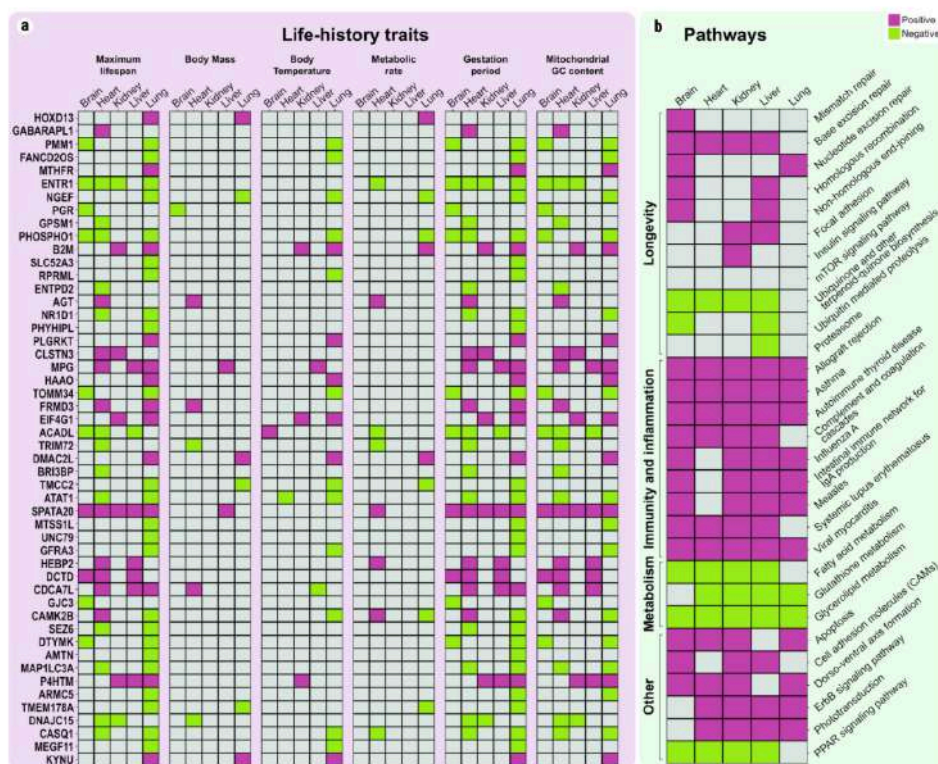


Figura 2. Corelații liniare între expresia genică și căile biologice cu DMV.

(a) Corelații liniare de vârf între expresia genică și trăsăturile speciei. Această hartă termică evidențiază relațiile semnificative statistic ($FDR < 0.05$, $R^2 > 0.3$) între nivelurile de expresie genică și trăsăturile cheie ale speciei. (b) Corelații liniare de vârf între MLS și scorurile de îmbogățire a căii (ES). A doua hartă termică vizualizează asocierile cheie între MLS și scorurile de îmbogățire a căii (ES).

Figură adaptată din Figura 2 Kulaga, Ursu, Toren. et al. (2021), utilizată sub licența CC BY 4.0.

Pentru a identifica relațiile neliniare dintre expresia genică și Durata Maximă de Viață (DMV), am utilizat un cadru de învățare automată interpretativ, folosind LightGBM și SHapley Additive exPlanations (SHAP)(61,37). Un model de referință care utilizează trăsăturile de istorie a vieții speciei (masa corporală, rata metabolică, temperatură, perioada de gestație și conținutul GC mitocondrial) a atins o precizie ridicată ($R^2 = 0,96$). Conținutul GC al ADN-ului mitocondrial și perioada de gestație au fost cei mai influenți predictorii, în concordanță cu descoperirile anterioare(62),(56). În continuare, o selecție descendentă de trăsături în două etape, cu LightGBM-SHAP, a evaluat influența fiecărui gen asupra DMV. Această abordare a rafinat un set inițial la 57 de gene, îmbunătățind substanțial performanța predictivă (Etapa II: R^2 a crescut de la 0,90 la 0,95; MAE a scăzut de la 4,73 la 3,04). Dintre cele 57 de gene, a fost remarcată o suprapunere nesemnificativă, dar consistentă (17 gene) cu genele asociate longevității (GALs) documentate în baza de date GenAge(54), inclusiv GNAS și TERT. Prioritizând genele după valorile absolute medii SHAP, am identificat 57 de gene cu contribuții predictive semnificative ($> 0,1$ ani). Graficul de sinteză SHAP (Fig. 3) ilustrează influența lor relativă.

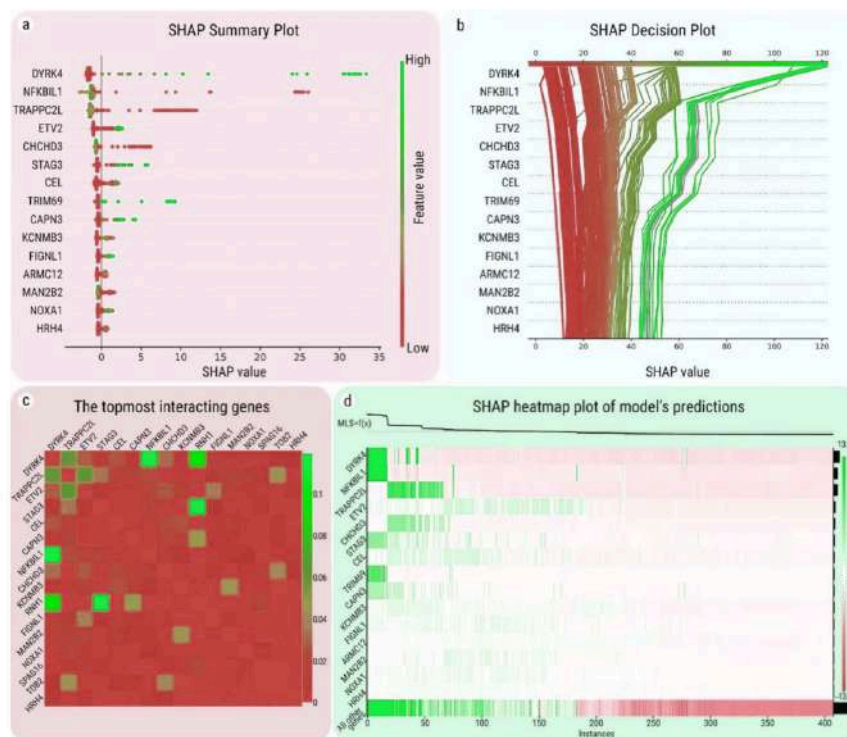


Figura 3. Explicații SHAP ale Expresiei Genice și Predicțiilor MLS

(a) Diagrama Sumarizatoare SHAP (SHAP Summary Plot). (b) Diagrama Decizională SHAP pentru Predicții Individuale (SHAP Decision Plot). (c) Harta Termică de Interacțiune SHAP pentru Perechi de Gene (SHAP Interaction Heatmap). (d) Harta Termică bazată pe SHAP a Contribuțiilor Genice per Eșantion (SHAP-based Heatmap of Gene Contributions per Sample).

Figura adaptată din Figura 3 Kulaga, Ursu, Toren. et al. (2021), utilizată sub licența CC BY 4.0.

Cinci dintre primele 15 gene (DYRK4, NFKBIL1, TRAPPC2L, ETV2 și CHCHD3) au influențat substanțial predicțiile duratei maxime de viață la mamifere (MLS), fiecare modificând predicțiile cu peste un an. Asocierea genelor TRAPPC2L și ETV2 cu îmbătrânirea este nedocumentată, sugerând că acestea sunt noi candidați promițători pentru cercetarea longevității. Cuantificarea direcției și forței asocierii utilizând coeficientul tau-b al lui Kendall între expresia genică și contribuțiile SHAP a identificat gene puternic pro-longevitate ($\tau \geq 0.6$) precum NEIL1,

CALCOCO2 și LRR1, și gene puternic anti-longevitate ($\tau \leq -0.6$), incluzând C6orf89 și PPP1CA. CALCOCO2 ($\tau = 0.72$) a apărut în mod constant ca un predictor pozitiv puternic al MLS în multiple organe (de exemplu, plămân $R^2 = 0.56$, creier $R^2 = 0.54$, ficat $R^2 = 0.53$, inimă $R^2 = 0.39$). În schimb, cea mai puternică genă anti-longevitate, C6orf89 ($\tau = -0.79$), a fost un predictor negativ în inimă ($R^2 = 0.61$) și ficat ($R^2 = 0.40$).

Genele care influențează durata de viață la mamifere (MLS) interacționează adesea complex, ducând la efecte combinate diferite de impacturile individuale. Am analizat aceste interacțiuni folosind valorile de interacțiune SHAP, care cuantifică relațiile de cooperare sau antagoniste. Fig. 3c rezumă forțele de interacțiune între genele de top asociate cu MLS.

Apoi am aplicat modelarea cu rețele bayesiene pentru a explora potențialele asocieri cauzale între expresia genică și MLS, identificând relații robuste, independente de redundanță sau corelații indirecte(63). Folosind conceptul de pătură Markov (Markov blanket) și algoritmul SES(53), am efectuat 50 de iterații pentru a genera semnături genice potențial cauzale (Fig. Supl. S3 din Kulaga, Ursu, Toren et al.(52)). Genele cu frecvență ridicată, indicând robustețe, au inclus NOXA1 (1.00), C6orf89 (0.94), NEU2 (0.94), NDUFA6 (0.90), RBM46 (0.82), KCNMB3 (0.72) și CEL (0.60). Din punct de vedere biologic, aceste rezultate coroborează descoperirile LightGBM-SHAP (de exemplu, NOXA1, C6orf89, CEL). Analiza de îmbogățire funcțională a implicat puternic mecanismele mitocondriale ca fiind centrale pentru longevitate.

Am găsit o consistență substanțială între modelele liniare și cele LightGBM-SHAP. Integrarea rețelelor bayesiene a întărit rolul genelor NOXA1, C6orf89 și CEL, indicând mitocondriile, repararea ADN-ului și reglarea metabolică ca procese cheie ale longevității. Integrând rezultatele bayesiene și LightGBM-SHAP, am constatat că genele robuste bayesiene (NOXA1, C6orf89, NEU2, NDUFA6, RBM46, KCNMB3, CEL) nu au avut toate cele mai mari valori SHAP. Totuși, CEL și KCNMB3 s-au clasat printre primele 10 gene cu cel mai mare impact conform SHAP. Două gene, NOXA1 și KCNMB3, au fost asociate robust cu MLS prin toate cele trei metode. Aceste gene fiabile, validate încrucișat, sunt candidați puternici pentru investigații funcționale viitoare asupra longevității.

Clasamentul Compozit și Determinarea Semnăturii Genice de Bază. Am stabilit un clasament compozit prin agregarea metricilor de performanță din toate cele trei modele. Acest clasament a fost utilizat cu modele liniare bayesiene multinivel pentru a identifica o semnătură genică de bază optimă și concisă. Evaluând subgrupuri (de la primele 3 la primele 13 gene) pe baza devianței penalizate, primele 11 gene clasate au fost selectate ca semnătură de bază, echilibrând simplitatea modelului și precizia predicției. Acest set final (detaliat în Tabelul 1 din teza integrală și în publicație) oferă candidați genici robuști pentru cercetarea longevității la mamifere.

4. Concluzii

Acest studiu a analizat expresia genică la 41 de specii de mamifere și cinci organe, utilizând modele liniare, neliniare și rețele bayesiene, pentru a identifica determinanții genetici ai Longevității Maxime (MLS). Peste 1800 de gene au arătat corelații semnificative cu MLS, adesea specifice organului. Analiza căilor de semnalizare a confirmat căile de longevitate deja

stabilite (de exemplu, repararea ADN-ului) și a sugerat altele noi (de exemplu, semnalizarea PPAR). Învățarea automată interpretabilă (LightGBM-SHAP) și rețelele bayesiene au identificat gene candidate suplimentare, cu o suprapunere limitată cu genele asociate experimental cu longevitatea (LAGs), reflectând diferențele dintre variația naturală și intervenția genetică. În special, CEL, NOXA1, CALCOCO2 și KCNMB3 au fost identificate în mod constant ca fiind candidați robuști pentru longevitate. Cercetarea evidențiază valoarea integrării modelării complementare pentru a diseca durata de viață a mamiferelor, oferind candidați promițători și un cadru metodologic solid.

III. Semnătura îmbătrânirii în fibroza pulmonară

Toren, D., Yanai, H., Abu Taha, R., Bunu, G., **Ursu, E.**, Ziesche, R., Tacutu, R. & Fraifeld, V. E. *Systems biology analysis of lung fibrosis-related genes in the bleomycin mouse model. Sci. Rep. 11, 19269 (2021).*

Declarație de Contribuții

În calitate de autor secundar, contribuția mea principală a fost o analiză de modelare liniară inter-specii, corelând expresia genelor pro- și anti-fibrotice cu durata maximă de viață (MLS) în organele mamiferelor. De asemenea, am interpretat rezultatele și am asistat la pregătirea manuscrisului. Proiectul a fost conceput și coordonat de echipele Prof. Vadim E. Fraifeld și Dr. Robi Tăcutu.

Rezultate și Discuții

Acest studiu a identificat 216 Gene Unice Asociate cu Fibroza Pulmonară (GUAFP) utilizând un model murin de fibroză pulmonară indusă de bleomicină. Intervențiile genetice au fost metoda principală. Aproximativ 43,5% din aceste gene au arătat activitate anti-fibrotică, 50% au fost pro-fibrotice, iar 6,5% au avut rezultate inconsistente.

Asocieri între GUAFP și Longevitate. Corelarea GUAFP cu Genele Asociate cu Longevitatea (GAL) din baza de date GenAge ([58.64](#)) a dezvăluit o legătură direcțională puternică: 11 din 12 gene pro-longevitate au fost anti-fibrotice, și 5 din 6 gene anti-longevitate au fost pro-fibrotice (testul exact al lui Fisher, $p = 0,001$).

Analiza ulterioară a datelor de expresie pulmonară inter-specii a arătat că 34 de GUAFP s-au corelat semnificativ cu Durata Maximă de Viață (MLS) a mamiferelor, o frecvență de 2,34 ori mai mare decât cea așteptată (testul exact al lui Fisher, $p = 6,4E-05$). Acest lucru susține ipoteza că genele legate de fibroză joacă un rol conservat în reglarea longevității, subliniind fibroza ca un factor critic al patologiei legate de vârstă. Concluzii

Genele pro-longevitate (GAL) sunt, în general, anti-fibrotice, iar genele anti-longevitate sunt în mare parte pro-fibrotice, dezvăluind o legătură genetică puternică, comună, între fibroza pulmonară și îmbătrânire. Analiza funcțională întărește acest lucru, cu grupuri anti-fibrotice bogate în GAL și grupuri pro-fibrotice predominant anti-GAL.

Acest lucru se aliniază cu faptul că îmbătrânirea este legată de căile care guvernează repararea țesuturilor, inflamația și metabolismul, care sunt critice pentru ambele condiții. Dovezile actuale sugerează că fibroza și îmbătrânirea apar din mecanisme genetice/moleculare comune, deși cauzalitatea necesită mai multe studii.

IV. Comunicarea intercelulară este perturbată odată cu îmbătrânirea

*Lagger, C., ***Ursu, E.**, Equey, A., Avelar, R. A., Pisco, A. O., Tacutu, R. & de Magalhães, J. P. *scDiffCom: a tool for differential analysis of cell-cell interactions provides a mouse atlas of aging changes in intercellular communication*. *Nat. Aging* 3, 1446–1461 (2023).

* denota first co-authorship

Declarația de Contribuții

În calitate de co-prim autor al acestui studiu, am contribuit în mod egal la proiectarea, implementarea și interpretarea proiectului alături de Dr. Cyril Lagger. În mod specific, am co-dezvoltat instrumentele scDiffCom, scAgeCom și scAgeComShiny; am curat și analizat baza de date privind interacțiunile ligand-receptor; și am efectuat analize pentru a susține constatările cheie. Am fost, de asemenea, implicat activ în interpretarea rezultatelor și în co-redactarea manuscrisului. Cadrul metodologic de bază și fluxul de lucru analitic au fost dezvoltate în colaborare între mine și C.L. și reprezintă obiectivul principal al acestui capitol. Studiul a fost supervizat în comun de Dr. Robi Tăcutu și Prof. João Pedro de Magalhães.

Prezentare generală

Dereglementarea comunicării intercelulare (CIC) este un semn distinctiv fundamental al îmbătrânirii, contribuind la diverse procese fiziologice și patologice. Pentru a investiga sistematic aceste modificări, introducem scDiffCom și scAgeCom, două instrumente complementare pentru analiza alterărilor comunicării celulă-celulă legate de vârstă.

scDiffCom este un pachet R pentru analiza diferențială a CIC utilizând date de transcriptomică unicelulară, bazându-se pe o bază de date curată de aproximativ 5.000 de interacțiuni ligand-receptor (ILR) pentru a compara rețelele de comunicare între diferite condiții.

Construit pe baza scDiffCom, scAgeCom este un atlas cuprinzător al modificărilor CIC legate de vârstă, integrând date din 23 de țesuturi de șoarece și 58 de seturi de date scRNA-seq din Tabula Muris Senis și Calico Murine Aging Cell Atlas. Această resursă dezvăluie schimbări sistemice legate de vârstă în semnalizarea intercelulară, incluzând:

- Activitate imună și inflamație crescută
- Semnalizare de dezvoltare redusă
- Angiogeneză și remodelare a matricei extracelulare afectate
- Metabolism lipidic dereglat

scAgeCom identifică liganzi, receptori și tipuri de celule specifice care conduc aceste procese și este disponibil public la <https://scagecom.org>.

1. Introducere

Îmbătrânirea este un proces biologic complex, marcat de o dereglare semnificativă a ICC, recunoscută ca o caracteristică definitorie a îmbătrânirii(1,65). Cercetările existente evidențiază mai multe modificări ale ICC odată cu îmbătrânirea, cum ar fi *inflammaging* (inflamația asociată vârstei)(66), supravegherea imună afectată(67) și secreția crescută de SASP(68). Deși ICC este dificil de măsurat direct, progresele în analiza expresiei genice la nivel de celulă unică permit inferența acesteia(69,70). Studiile existente privind îmbătrânirea la nivel de celulă unică se concentrează adesea pe detectarea rețelelor ICC separat în probele tinere și cele în vârstă, ceea ce ignoră schimbările în intensitatea interacțiunii și nu dispune de un cadru statistic pentru cuantificarea modificărilor. Pentru a aborda aceste limitări, am dezvoltat scDiffCom, un pachet R pentru analiza diferențială a ICC. Am aplicat scDiffCom la seturi de date scRNA-seq privind îmbătrânirea din Tabula Muris Senis(71) și Atlasul Celular Murin al Îmbătrânirii Calico(72) pentru a crea scAgeCom, un atlas la scară largă care cartografiază schimbările ICC asociate vârstei în 23 de țesuturi de șoarece. Această analiză confirmă dereglarea sistemică, cu o reglare pozitivă globală a activității sistemului imunitar și a inflamației și un declin al proceselor precum organizarea matricei extracelulare și creșterea tisulară.

2. Metode

Recuperarea și Procesarea Interacțiunilor Ligand–Receptor (ILR) și adnotarea cu termeni GO, căi KEGG și resurse despre îmbătrânire. Am compilat seturi de date de înaltă calitate, curate, privind interacțiunile ligand–receptor (ILR) din șapte baze de date publice (de ex., CellChat, NicheNet, CellPhoneDB), excluzând interacțiunile prezise computațional. ILR-urile au fost adnotate cu termeni Gene Ontology (GO) utilizând o metodă personalizată bazată pe intersecția bazată pe grafic a termenilor ligandului și receptorului din Ensembl. Căile KEGG au fost atribuite numai dacă atât ligandul, cât și receptorul se aflau în aceeași cale. Pentru a lega ILR-urile de îmbătrânire, am integrat date din baze de date legate de îmbătrânire (GenAge, LongevityMap, CellAge, HAGR) și am cuantificat articolele PubMed care fac referire la fiecare genă ILR (sau omologul său uman) în contextul îmbătrânirii.

Scorarea IIC (Media Geometrică), Detectia, Analiza Diferențială și Clasificarea. scDiffCom calculează un scor de interacțiune inter-celulară (IIC) ca media geometrică a expresiei ligandului și receptorului pentru a reflecta natura multiplicativă a interacțiunilor, făcând ca logaritmul schimbării de pliuri ($\log FC$) al acestuia să fie media aritmetică a $\log FC$ -urilor genelor. Metoda utilizează trei teste de permutare simultane pentru a evalua specificitatea IIC în două condiții (A, B) și expresia diferențială între ele, testând doar IIC-urile exprimate. Distribuțiile nule sunt generate prin amestecarea etichetelor, iar p-valorile specificității unilaterale și p-valorile expresiei diferențiale bilaterale sunt calculate și ajustate Benjamini-Hochberg. Un IIC este „detectat” dacă este exprimat, specific ($p\text{-adj.} < 0,05$) și clasat în top (80%). Este „exprimat diferențial” dacă p-valoarea sa diferențială ajustată cu FDR este $< 0,05$. IIC-urile detectate sunt clasificate (UP, DOWN, FLAT, NSC) pe baza DE ajustate și a $\log FC$, prioritizând testul diferențial. Compararea cu analiza standard a expresiei diferențiale a genelor evidențiază necesitatea integrării rezultatelor la nivel de IIC și la nivel de genă pentru o captură și clasificare fiabilă a interacțiunilor. Analiza de Supra-Reprezentare (ORA) este utilizată pentru a detecta trăsături supra-reprezentate (de ex., termeni GO, căi) în Interacțiunile Inter-Celulare (IIC) reglate diferențial.

Dezvoltarea scAgeComShiny. Aplicația web interactivă Shiny, scAgeComShiny, a fost construită cu golem (73). Utilizează plotly (74) pentru diagrame de dispersie și rrvgo (75)/REVIGO(76) pentru vizualizarea termenilor GO și reducerea redundanței. Aplicația este containerizată cu Docker și implementată prin ShinyProxy.

Statistici și Reproducibilitate. Dimensiunea eșantionului a fost limitată de disponibilitatea publică. Nu a fost posibilă randomizarea sau orbirea. Au fost utilizate teste non-parametrice (DE și ORA), cu corecția Benjamini–Hochberg aplicată în mod constant pentru testarea multiplă.

3. Rezultate și Discuții

Interacțiuni Ligand-Receptor (LRI) din baze de date existent

Pentru a analiza comunicarea intercelulară (ICC) din datele scRNA-seq, abordarea noastră a necesitat mai întâi o bază de date extinsă de interacțiuni Ligand-Receptor (LRI). Pentru a maximiza acoperirea diferitelor tipuri de interacțiuni, am compilat LRI din șapte resurse disponibile public, incluzând CellChat (79), CellPhoneDB (80), CellTalkDB (81), NATMI/connectomeDB2020 (82), ICELLNET (83), NicheNet (84), și SingleCellSignalR (85). Abordarea noastră a dus la generarea a două baze de date curate de interacțiuni ligand–receptor (LRI). Baza de date LRI pentru șoarece conține 4.582 de interacțiuni, dintre care 3.479 sunt simple și 1.103 sunt complexe. În mod similar, baza de date LRI pentru om include 4.785 de interacțiuni, cu 3.648 de interacțiuni simple și 1.137 de interacțiuni complexe. Date detaliate pot fi găsite în Fig. 6a și Tabele Suplimentare 1 și 2 din Lager, Ursu et al. (86).

Adnotarea funcțională a LRI-urilor. Am implementat un cadru standardizat pentru a adnota toate interacțiunile ligand–receptor (LRI), asigurând relevanța biologică și facilitând analizele ulterioare. Pentru a îmbogăți LRI-urile cu adnotări relevante din punct de vedere biologic, am atribuit termeni Gene Ontology (GO) (87) interacțiunilor într-un mod semnificativ, prioritizând termenii relevanți pentru interacțiunea în sine, mai degrabă decât pentru genele individuale. LRI-urile au fost legate de căile KEGG (88) numai dacă toate genele dintr-o anumită interacțiune erau prezente în aceeași cale. În anticiparea analizelor legate de îmbătrânire, am mapat genele LRI de șoarece la articole PubMed care le leagă de îmbătrânire și boli legate de vârstă (excluzând cancerul). În plus, am făcut o referire încrucișată a genelor cu bazele de date cheie legate de îmbătrânire, inclusiv GenAge (54), CellAge (89,90), LongevityMap (91), baza de date Gene Expression (92) a Resurselor Genomice de Îmbătrânire Umană (HAGR) (54).

Analiza comunicării diferențiale celulă-celulă cu scDiffCom

Pachetul R scDiffCom detectează schimbări semnificative în comunicarea celulă-tip (CCI) între două condiții în seturile de date scRNA-seq (Fig. 4), lucrând cu obiecte R Seurat care conțin etichete de tip celular și condiție(93–95). Acesta atribuie un scor CCI bazat pe expresia medie a ligandului și receptorului, robust la părtinirea numărului total de celule(96). CCI-urile sunt validate pe baza a trei criterii: expresie suficientă, specificitate printr-un test de permutare de tip CellPhoneDB(80,97) și un scor relativ ridicat. CCI-urile sunt apoi clasificate ca reglate în sus (UP), reglate în jos (DOWN), stabile (FLAT) sau fără schimbare semnificativă (NSC). scDiffCom

utilizează Analiza Supra-Reprezentării (ORA) pentru a analiza mii de Interacțiuni Celulă-Celulă (CCI) clasificate — Reglate în sus (UP), Reglate în jos (DOWN) sau Stabile (FLAT) — pe un fundal al celorlalte. Această abordare evită părtinirea tradițională a îmbogățirii bazate pe gene.

Îmbătrânirea dereglează mai multe aspecte ale comunicării celulă-celulă

Atlasul scAgeCom, construit cu scDiffCom pe 58 de seturi de date scRNA-seq murine din TMS(71) și Calico(72), detaliază modificările comunicării intercelulare (ICC) legate de vârstă în 23 de organe, abordând dimorfismul sexual (Fig. 5 și Text Suplimentar 1, Fig. Suplimentară 1 din Lager, Ursu et al.(86)). Această resursă este disponibilă online (<https://scagecom.org/>) (Fig. 6). Dintre cele 393.035 de interacțiuni celulă-celulă (CCI) detectate, 18% au fost reglate diferențial în funcție de vârstă (5% reglate în sus, 13% reglate în jos). 1.135 de interacțiuni ligand-receptor au fost filtrate pentru a minimiza descoperirile false. Benchmarking-ul a confirmat că scorul CCI al scDiffCom este superior comparațiilor la nivel de genă (Ext. Data Fig. 2 din Lager, Ursu et al.(86)), metoda sa ORA evită părtinirea termenilor GO (Fig. Suplimentare 2 și 3 din Lager, Ursu et al.(86)) și bazele de date LRI sunt cruciale (Fig. Suplimentară 4 din Lager, Ursu et al.(86)). Aceste rezultate stabilesc scDiffCom ca un instrument robust pentru detectarea CCI-urilor mediate de proteine, relevante din punct de vedere biologic, din datele scRNA-seq, făcând din atlasul scAgeCom o resursă valoroasă pentru cercetarea îmbătrânirii.

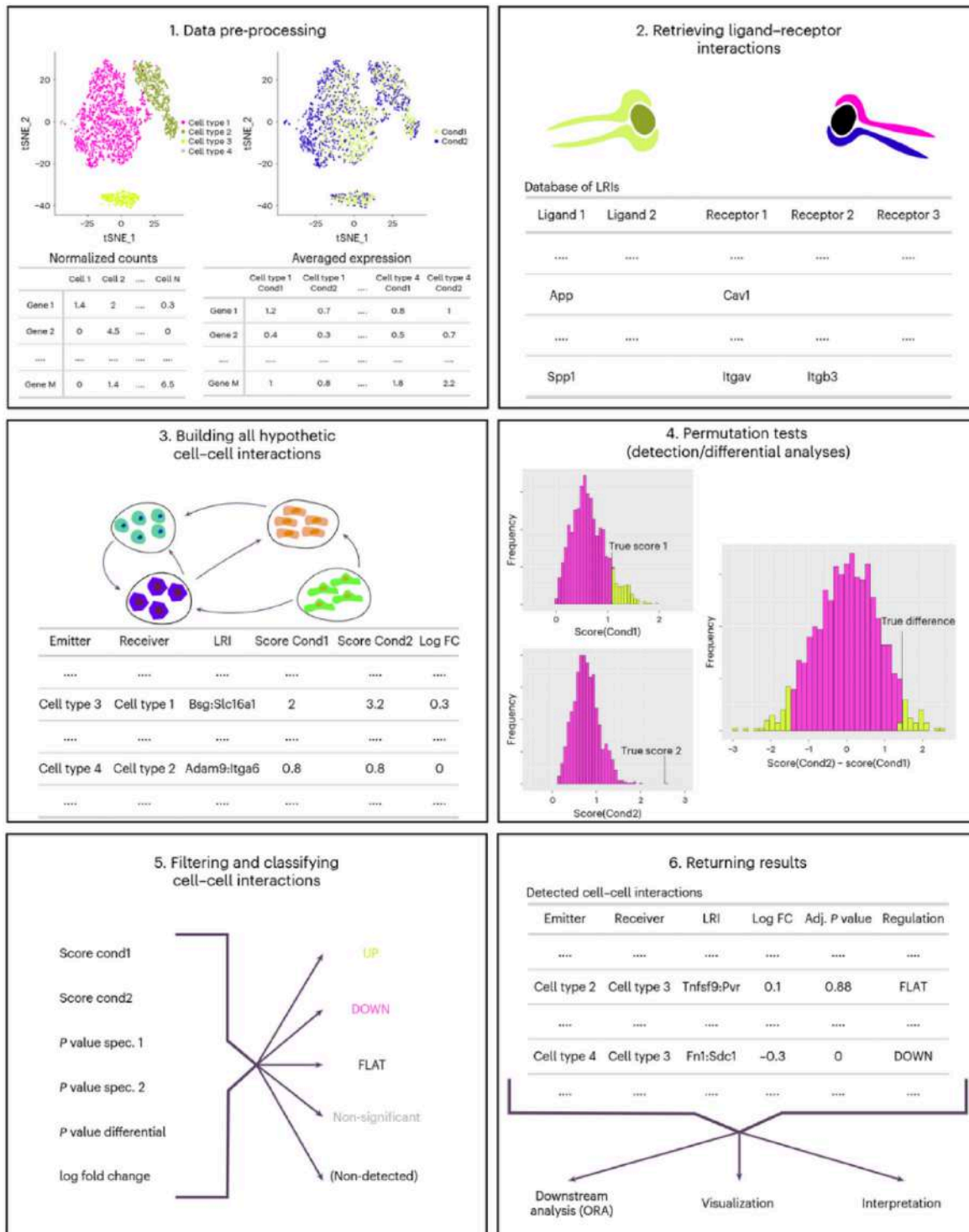


Figura 4. Rezumatul Fluxului de Lucru al scDiffCom.

Citirile de secvențiere ARN monocelulară (scRNA-seq) sau identificatorii moleculari unici (UMIs) sunt mai întâi agregați pe tipuri de celule și condiții experimentale (1). Datele de expresie genică sunt apoi mapate pe baza de date curatoriată de interacțiuni ligand-receptor (LRIs) (2) pentru a deduce toate interacțiunile celulă-celulă (CCIs) posibile între tipurile de celule (3). Sunt efectuate teste statistice de permutare pentru a evalua relevanța biologică a fiecărei CCI și pentru a detecta expresia diferențială între condiții (4). CCIs sunt clasificate ulterior pe baza metricilor calculate, incluzând scorurile de interacțiune, valorile P și log-schimbarea de pli (5). Rezultatele sunt compilate într-un format structurat adecvat pentru analiza și interpretarea ulterioară (6). FC, schimbare de pli (fold change); tSNE, încorporare stocastică de vecinătate distribuită t (t-distributed stochastic neighbor embedding).

Figura adaptată din Figura 2 din Lager, Ursu. et al. (2023), utilizată sub licența CC BY 4.0

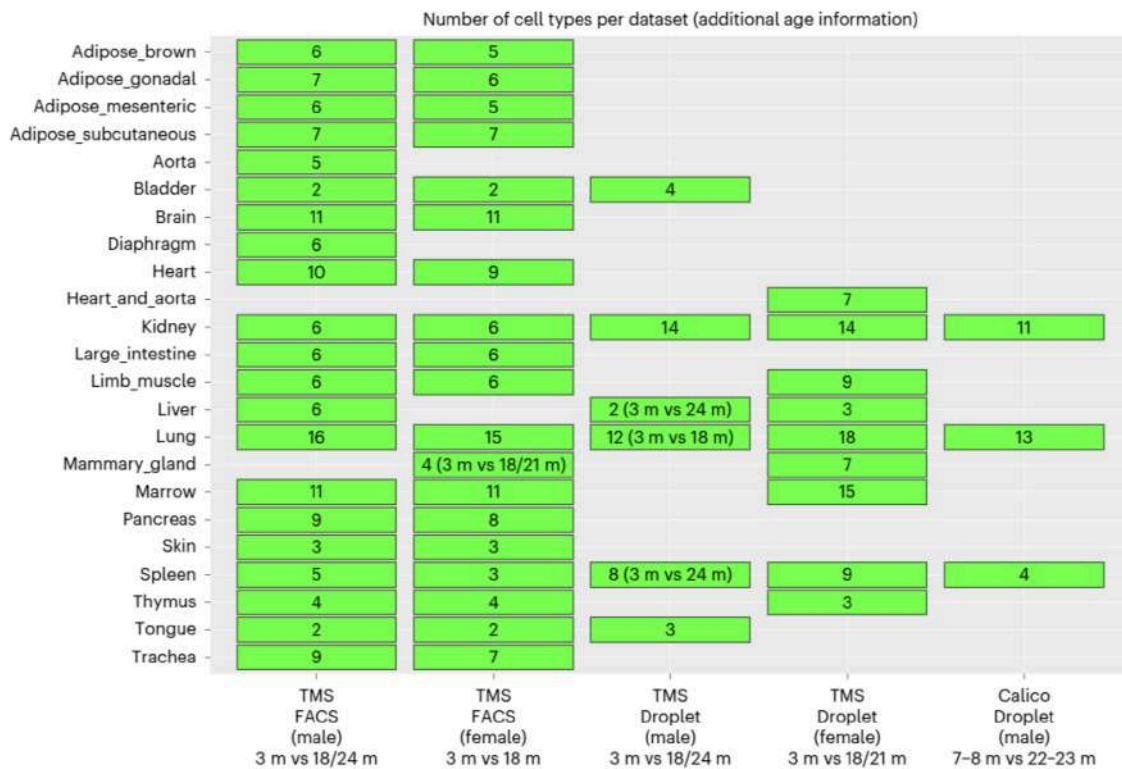


Figura 5. Cele 58 de Seturi de Date de Îmbătrânire scRNA-seq din scAgeCom.

Figură adaptată din Figura 3 din Lager, Ursu. et al. (2023), utilizată sub licență CC BY 4.0

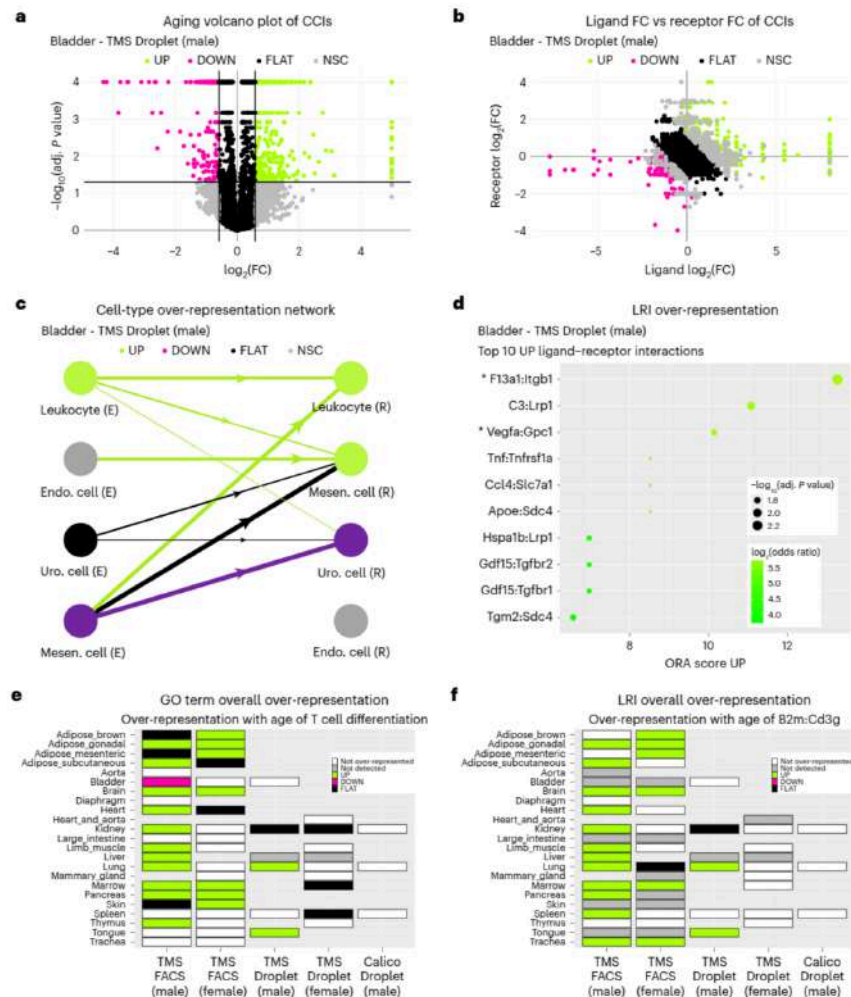


Figura 6. Prezentare generală a rezultatelor vizuale în scAgeCom.

scAgeCom oferă o varietate de rezultate grafice, accesibile la <https://scagecom.org/>, pentru a facilita explorarea schimbărilor legate de îmbătrânire în comunicarea intercelulară.

Figură adaptată după Figura 4 din Lager, Ursu. et al. (2023), utilizată sub licența CC BY 4.0

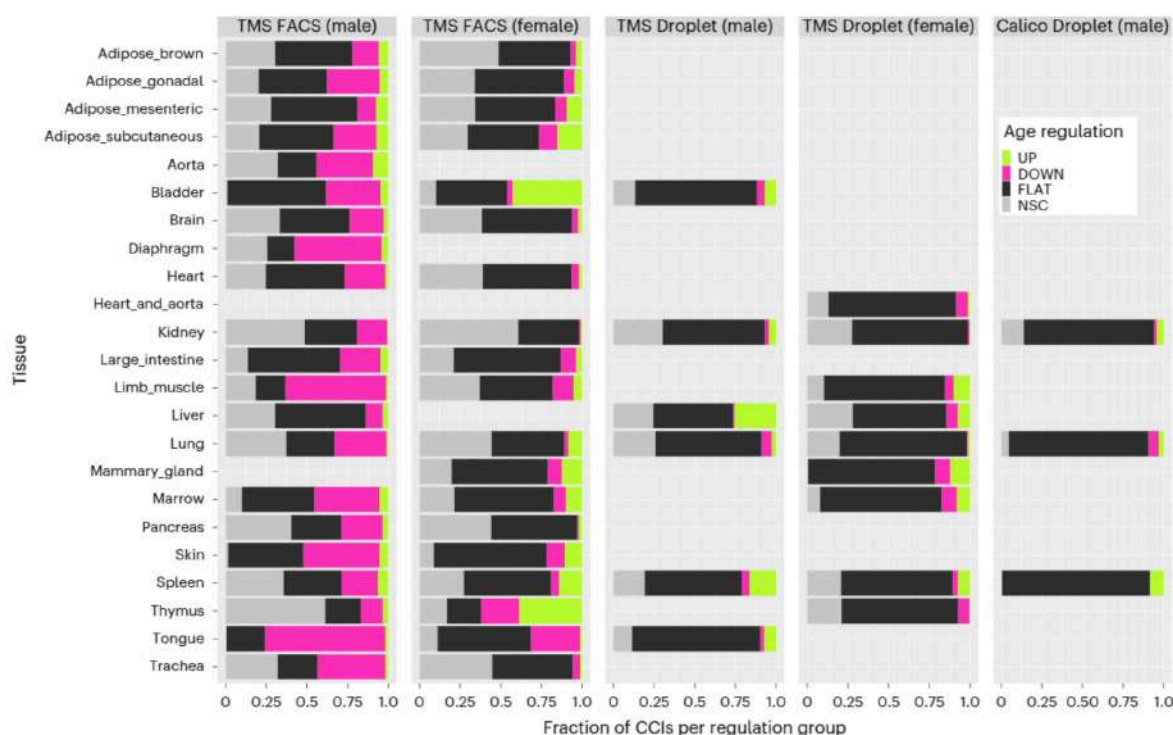


Figura 7. Variația tiparelor de reglare în funcție de vârstă ale CCI-urilor în seturile de date scAgeCom.

Proporțiile de CCI-uri clasificate ca fiind supreglate (UP), subreglate (DOWN), stabile (FLAT) sau fără modificări semnificative (NSC) în funcție de vârstă diferă între seturile de date, în funcție de tipul de țesut, sex și metodologia experimentală. În general, seturile de date TMS FACS prezintă un zgomot mai mare, reflectat de o proporție mai mare de CCI-uri clasificate ca NSC, comparativ cu seturile de date bazate pe Droplet. În plus, seturile de date TMS FACS (masculin) afișează o fracțiune notabil mai mare de CCI-uri subreglate în funcție de vârstă, în raport cu celelalte condiții.

Figură adaptată din Figura 6 din Lager, Ursu. et al. (2023), utilizată sub licența CC BY 4.0

Reglarea interacțiunilor celulă-celulă (ICC) odată cu vârsta este foarte variabilă, seturile de date TMS FACS (masculi) arătând multe ICC-uri reglate negativ. Seturile de date FACS sunt mai zgomotoase decât seturile de date Droplet. Am prioritizat schimbările care apar în mai multe țesuturi, implicând gene noi de îmbătrânire, în secretomică sau dependente de sex.

A fost confirmată o reglare sistemică în sens ascendent a proceselor inflamatorii, imune și virale, incluzând interacțiunile B2m: Cd3g, Tnfrsf12: Tnfrsf12a și Ccl5 (Ext. Data Fig. 3b din Lager, Ursu et al. (86)). B2M este prezent în cinci secretome. Slpi: Plscr1, deși reglat ascendent în opt țesuturi, este în mare măsură neexplorat în contextul îmbătrânirii.

Metabolismul lipidelor este dereglat (Ext. Data Fig. 4 din Lager, Ursu et al. (86)) cu tipare specifice sexului: ICC-urile legate de Apoe sunt reglate ascendent la masculi, dar reglate descendent la femele, și invers pentru ICC-urile legate de App. Aceste proteine asociate bolii Alzheimer (AD) au probabil roluri sistemice în îmbătrânire, susținute de detecția în secretomă.

O observație frapantă este reglarea descendentă a organizării matricei extracelulare (MEC) și a

aderenței celulare (Ext. Data Fig. 5 din Lager, Ursu et al.(86)). Acest declin, determinat de colagene, cadherine și metalopeptidaze cu integrine, este cel mai puternic în țesutul conjunctiv, celulele epiteliale și endoteliale.

Creșterea, dezvoltarea, supraviețuirea, diferențierea și angiogeneza, de asemenea, scad (Ext. Data Fig. 6 din Lager, Ursu et al.(86)), sugerând o regenerare afectată. Comunicarea redusă între celulele stem și către celulele endoteliale leagă îmbătrânirea de capacitatea regenerativă diminuată (Ext. Data Fig. 5c din Lager, Ursu et al.(86)).

scAgeCom dezvăluie tipare dimorfice sexuale (Fig. 7); de exemplu, în setul de date TMS FACS Plămân, 13% din ICC-uri (incluzând App, Pecam1 și Itgb1) au arătat o expresie mai puternică la masculii tineri, care a scăzut odată cu vârsta la masculi, dar a crescut la femele, evidențiind necesitatea abordărilor personalizate.

4. Concluzii

În ciuda limitărilor semnificative, scAgeCom oferă un atlas extins al îmbătrânirii comunicării intercelulare la șoareci, oferind perspective noi asupra modificărilor de comunicare specifice țesutului și specifice sexului. Contribuțiile cheie includ:

- Cartografierea comunicării intercelulare cuprinzătoare în 23 de țesuturi
- Identificarea potențialelor ținte terapeutice (de ex., B2m, Mif, Angpt1, Apoe)
- Noi ipoteze privind mecanismele de îmbătrânire, inclusiv modificările metabolismului lipidic și mecanismele declinului vascular
- Potențial de integrare cu seturi de date de senescentă și proteomică pentru validare ulterioară

În viitor, analiza încrucișată cu alte atlase de îmbătrânire, cum ar fi atlasul SASP, va fi esențială pentru a ne rafina înțelegerea modului în care comunicarea intercelulară influențează procesul de îmbătrânire.

V. Compoziția datelor de antrenament negative este critică pentru învățarea repertoriilor imune

**Ursu, E., *Minnegalieva, A., Rawat, P., Chernigovskaya, M., Tacutu, R., Sandve, G. K., Robert, P. A. & Greiff, V. Training data composition determines machine learning generalization and biological rule discovery. Nature Machine Intelligence (2025). doi:10.1101/2024.06.17.599333*

** denota first co-authorship*

Declarație privind Contribuțiile

În calitate de co-prim autor al acestui studiu, am contribuit în mod egal cu Aygul Minnegalieva la proiectarea, execuția și redactarea lucrării. Împreună cu Aygul Minnegalieva și Prof. Victor Greiff, am conceput analizele și vizualizările centrale studiului. De asemenea, am efectuat analize cheie și generarea de figuri în colaborare cu Aygul Minnegalieva și Dr. Philippe A. Robert. În plus, am co-redactat prima versiune a manuscrisului alături de Aygul Minnegalieva și

Prof. Victor Greiff și am participat la revizuirea versiunii finale. Natura colaborativă a acestui studiu este reflectată în contribuțiile noastre egale. Acest capitol se axează pe munca pe care am dezvoltat-o în comun și pe interpretarea rezultatelor în contextul mai larg al modelării imunologice sub îndrumarea Prof. Victor Greiff.

1. Introducere

Modelele de învățare automată (ML) supravegheată depind în mod critic de compoziția seturilor de date etichetate, în special de definirea eşantioanelor negative (care reprezintă absența clasei țintă) în clasificarea binară(98–101,102,103). Acest factor este insuficient studiat în ceea ce privește influența sa asupra generalizării modelului și extracției regulilor biologice. Contextul predicției legării anticorp-antigen, cu strategiile sale variate de construire a datelor negative, oferă un cadru pentru a studia aceste efecte. Lucrările anterioare arată că alegerea exemplilor negative afectează precizia predictivă și generalizarea în modelele de interacțiune anticorp-antigen și TCR-antigen(104–106,107–111); cu toate acestea, impactul asupra interpretabilității (mecanismele de legare deduse) este neexaminat. Pentru a investiga, am utilizat „Absolut!”(106) pentru a genera date sintetice anticorp-antigen cu diverse strategii de seturi de date negative (Fig. 8), concentrându-ne pe regiunile CDRH3. Rețelele neuronale simple și DeepLIFT(112–114) au fost utilizate pentru a evalua regulile biologice învățate (Fig. 8). Modelele antrenate cu date negative mai similare cu clasa pozitivă au generalizat mai bine la datele neobservate și au dezvăluit reguli de legare diferite în comparație cu cele antrenate cu exemple negative disimilare. Aceste tendințe s-au menținut și pe date experimentale (Fig. 8,9)(115). Proiectarea și selecția exemplilor negative este o componentă critică, adesea trecută cu vederea, pentru construirea de modele ML robuste și interpretabile în aplicațiile biologice.

Acest capitol final oferă o contribuție metodologică la modelele de învățare automată a repertoriilor receptorilor imuni, abordând provocările legate de calitatea datelor de antrenare și de atribuirea specificității antigenice. Acesta îmbunătățește limitările actuale de modelare AIRR-Seq pentru a spori interpretabilitatea și relevanța biologică, stabilind o bază computațională pentru studiile viitoare privind îmbătrânirea imună și schimbările repertoriului imun adaptiv.

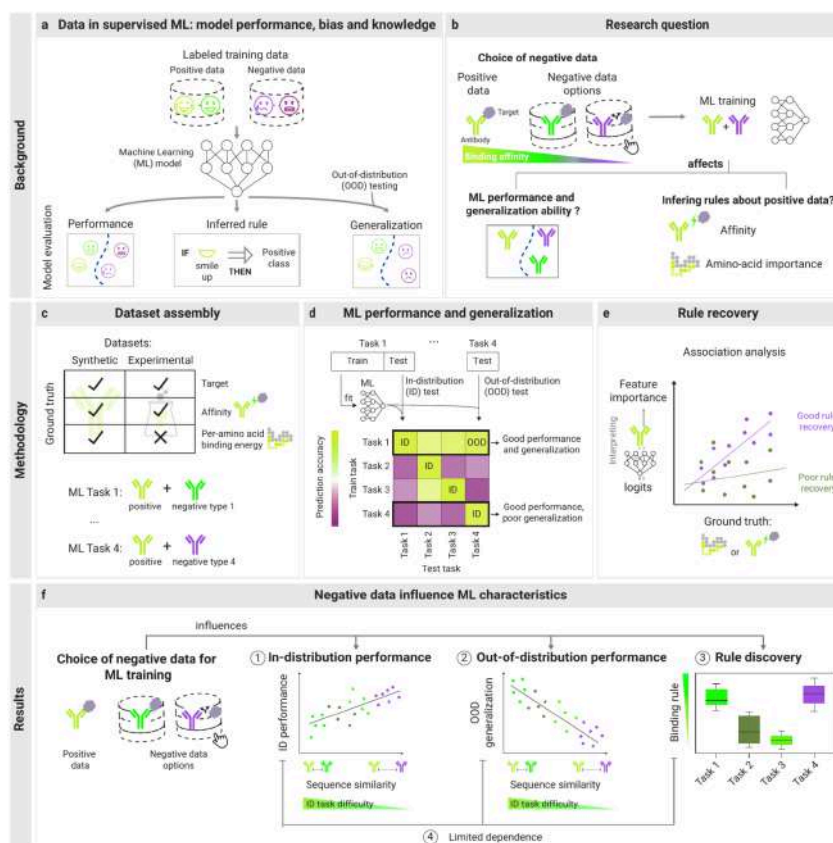


Figura 8. Compoziția setului de date de antrenament modelează generalizarea și descoperirea de reguli a ML.

Figură adaptată după Figura 1 din Ursu, E. et al. (2025), utilizată sub licența CC BY 4.0

2. Metode

Un set de date de legare anticorp-antigen semi-sintetic a fost generat utilizând cadrul de simulare Absolut! (Robert et al., Akbar et al.(106,116)). Acest set de date combină secvențe autentice de CDRH3 de șoarece(117) cu structuri de antigen PDB formate în grilă 3D. Secvențele CDRH3 au fost andocate pe antigene rigide pentru a calcula energiile de legare utilizând potențialul statistic Miyazawa–Jernigan într-un model de grilă 3D (cadrul Absolut!, Robert et al.(106)). Secvențele au fost etichetate în funcție de percentila de afinitate de legare: afinitate mare (top 1%), slabă (1–5%) și ne-liganți (restul de 95%). Au fost utilizate date de la zece antigene.

Pentru sarcinile de predicție, seturi de date echilibrate (30k antrenare, 10k testare) au fost create pentru fiecare antigen. Probele pozitive (liganți din top 1%) au fost comparate cu patru definiții de clasă negativă: Vs 1 (un singur antigen diferit), Vs 9 (nouă alte antigene agregate), Vs Slab (liganți slabi, 1–5%) și Vs Ne-ligant (>5%). Reproducibilitatea a fost asigurată cu șase împărțiri antrenare-testare și patru semințe aleatoare per împărțire.

Modelul de bază de Învățare Automată a fost SN10(106), o rețea neuronală feedforward superficială antrenată pe CDRH3-uri codificate one-hot (intrare cu 220 de neuroni, strat ascuns ReLU cu 10 unități, ieșire sigmoidă). Acesta a fost comparat cu un Transformer mai profund și variante SN10 bazate pe PLM, utilizând încorporări pre-antrenate ESM2b (1280 dim.) și

AntiBERTa2 (1024 dim.)[\(118,119,120\)](#), generate prin EmbedAIRR. Interpretarea a fost explorată utilizând DeepLIFT[\(114\)](#) pentru a cuantifica contribuția aminoacizilor. În plus, Analiza Specifică Epitopului a implicat construirea de seturi de date restrânse la secvențele care se leagă de un singur epitop dominant (de exemplu, 1H0DE1).

3. Rezultate și Discuții

1. Compoziția secvențelor din setul de date de antrenare influențează performanța predicției în sarcinile de clasificare binară in-distribution (ID) și out-of-distribution (OOD)

1.1 Configurația învățării automate pe date sintetice și experimentale

Am început prin a investiga modul în care diferitele definiții ale clasei negative afectează performanța și generalizabilitatea modelelor de învățare automată supravegheată pentru predicția legării anticorp–antigen. Fiecare dintre cele patru sarcini a utilizat un set identic de probe pozitive — secvențe CDRH3 cu afinitate ridicată — dar a diferit prin modul în care a fost definită clasa negativă.

Am utilizat date sintetice de secvențe CDRH3, adnotate cu energii de legare pentru zece antigene. Pentru fiecare antigen, clasa pozitivă a constatat în secvențe care se încadrează în percentila de afinitate de top 1%. Am definit patru tipuri de clase negative: **vs Nelegare** (*Non-binder*): secvențe CDRH3 din cea mai mică percentilă de 95% a energiei de legare la același antigen; **vs Slabă** (*Weak*): liganzi slabi în percentila 1–5% pentru același antigen, formând un set disjunct față de vs Nelegare; **vs 1**: liganzi cu afinitate ridicată (top 1%) la un singur antigen distinct, excluzând cele care se leagă și de antigenul clasei pozitive; **vs 9**: o extindere a vs 1, cuprinzând liganzi cu afinitate ridicată la fiecare dintre celelalte nouă antigene, reprezentate în mod egal.

Am antrenat modele SN10. Această arhitectură a fost aleasă pentru interpretabilitatea sa și performanța anterioară de *benchmarking* și a arătat rezultate comparabile cu modelele mai profunde. Pentru a valida concluziile noastre derivate din seturile de date sintetice, am replicat aceeași configurație experimentală utilizând setul de date care vizează HER2 publicat de Porebski et al. [\(115\)](#).

1.2 Acuratețea Predicției în Distribuție (ID) Depinde de Compoziția Setului de Date de Antrenare

Pentru a evalua capacitatea modelelor de a învăța reguli generalizabile, am măsurat mai întâi acuratețea în distribuție (ID) — performanța pe datele de testare cu aceleași definiții ale claselor pozitive și negative utilizate în timpul antrenării (Fig. 9a). La toate cele patru tipuri de sarcini, modelele au obținut o acuratețe ID ridicată, cu valori mediane care depășesc 0.85. A apărut un clasament clar între sarcini: modelele au funcționat cel mai bine pe „vs Nelegare” (interval: 0.97–1.00, mediană: 0.99), urmate de „vs 1” (interval: 0.94–1.00, mediană: 0.98), „vs 9” (interval: 0.91–0.98, mediană: 0.94) și, în final, „vs Slabă” (interval: 0.85–0.98, mediană: 0.92), care s-a dovedit a fi cea mai dificilă (Fig. 9b). Mai mult, modelul SN10 a depășit regresia logistică (LR), în special în sarcina „vs Slabă”. Acest lucru sugerează că SN10 poate valorifica dependențele inter-poziționale în secvențe pe care LR nu le poate.

1.3 Disimilaritatea Secvenței Explică Variabilitatea în Acuratețea ID

Pentru a înțelege de ce performanța ID a diferit între sarcini, am examinat relația dintre acuratețea modelului și disimilaritatea secvenței între seturile de date pozitive și negative. Folosind matricile de greutate pozițională (PWM), am calculat distanțele Jensen–Shannon (JSD) pentru a cuantifica divergența distribuțională. Am constatat că JSD a crescut în ordinea: „vs Slabă” < „vs 9” < „vs 1” < „vs Nelegare”, reflectând o divergență mai mare a secvenței între clasele pozitive și negative (Fig. 9c). Acest gradient a oglindit îndeaproape tendința de acuratețe ID, iar JSD a fost corelat semnificativ cu performanța modelului pentru sarcinile „vs 9” ($r = 0.94$), „vs 1” ($r = 0.73$) și „vs Nelegare” ($r = 0.77$) (toate $p < 0.05$), dar nu și pentru „vs Slabă” ($r = 0.30$, $p \geq 0.05$) sau pentru oricare dintre controalele amestecate (Supp. Table 1, Supp. Fig. 3a din Ursu, Minnegalieva et al. [\(121\)](#)).

1.4 Acuratețea în Afara Distribuției (OOD) Este De Asemenea Modelată de Compoziția Setului de Date

Am evaluat generalizarea în afara distribuției (OOD) (Fig. 9d). Acuratețea OOD a fost mai mică decât performanța în distribuție (ID), așa cum era de așteptat.

Modelele antrenate pe „vs Nelegare” au arătat cea mai mare scădere (mediana ID 0.99 la OOD 0.72–0.82) (Supp. Fig. 3d din Ursu, Minnegalieva et al. [\(121\)](#)). Modelele antrenate pe „vs 9” s-au generalizat cel mai bine la „vs 1” (0.94), în timp ce modelele „vs 1” au performat moderat (0.78 la „vs 9,” 0.72 la „vs Nelegare”). Alegerea antigenului negativ „vs 1” a afectat rezultatele (Supp. Fig. 5 din Ursu, Minnegalieva et al. [\(121\)](#)).

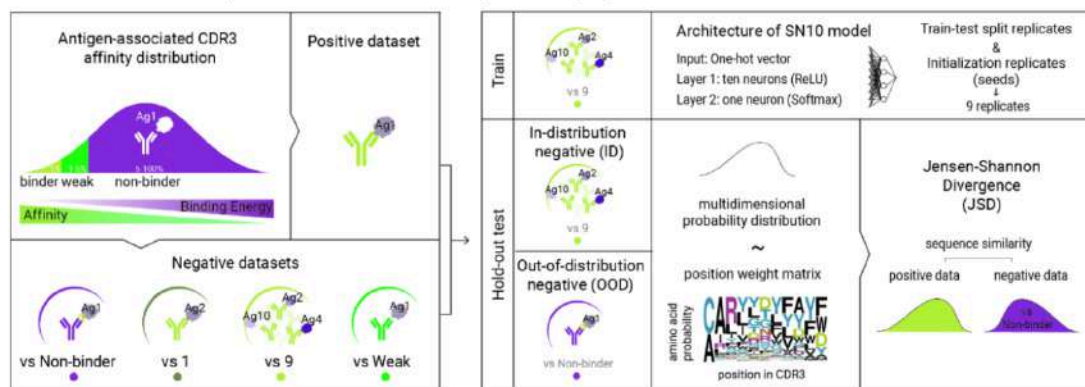
„Vs Slabă” a fost cel mai dificil test OOD (acuratețe 0.58–0.71). În mod surprinzător, modelele antrenate pe „vs Slabă” s-au generalizat cel mai bine în ansamblu (0.90–0.96) (Fig. 9d).

Acuratețea ID ridicată nu garantează generalizarea. De exemplu, „vs Nelegare” a produs o performanță ID excelentă, dar slabă OOD, în timp ce „vs Slabă” a fost cea mai dificilă sarcină ID, dar cea mai robustă în generalizare. Acest lucru subliniază rolul critic al designului clasei negative pentru modelele de învățare automată a legării anticorp–antigen.

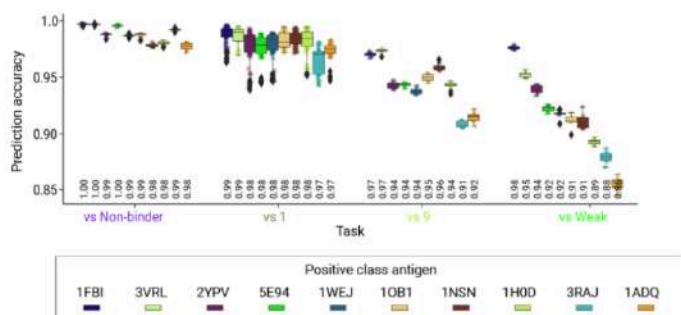
1.5 Datele Experimentale Confirmă Constatările din Seturile de Date Sintetice

Pentru a valida aceste constatări în date din lumea reală, am analizat un set de date experimentale de legare anticorp–antigen publicat recent de Porebski et al. [\(115\)](#), care, la fel ca datele sintetice „Absolut!”, include secvențe CDRH3 adnotate cu afinitate împotriva HER2 (a se vedea Metodele).

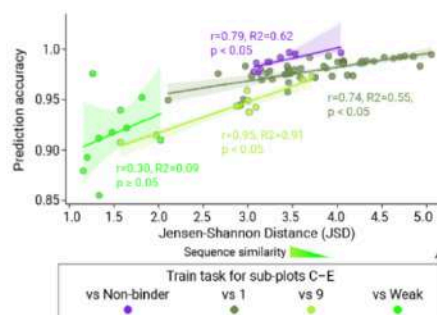
a Classification dataset composition, ML architecture, training, and testing regimes



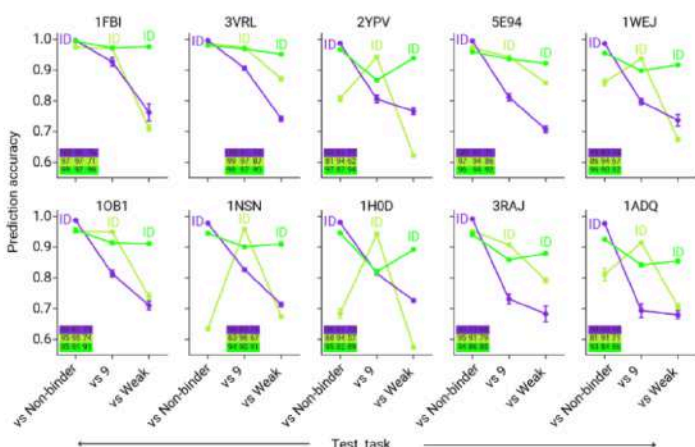
b In-distribution prediction accuracy across tasks and positive class antigens



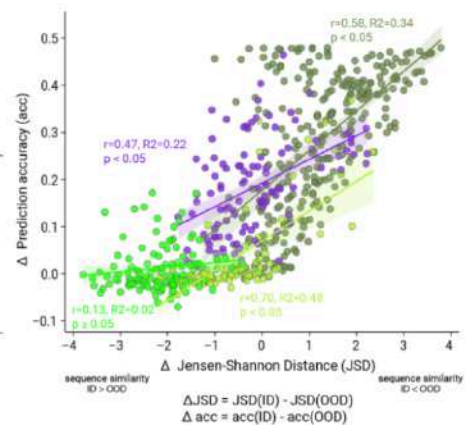
c In-distribution prediction accuracy and sequence similarity correlation



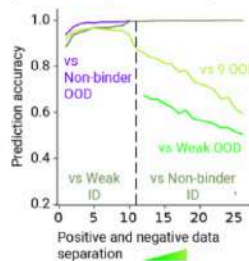
d Out-of-distribution prediction accuracy across tasks



e Out-of-distribution prediction accuracy and sequence similarity correlation



f Performance vs. affinity-wise positive-negative data separation



g ID and OOD performance on independent experimental dataset

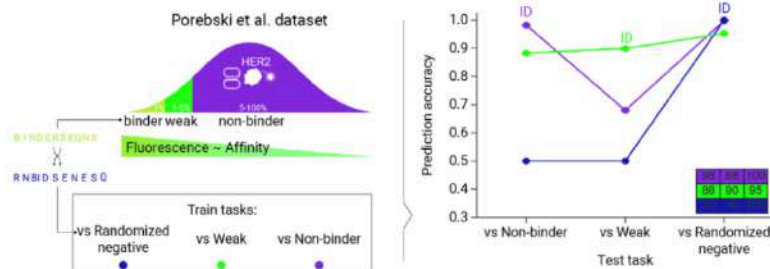


Figura 9. Acuratețea clasificării variază în funcție de antigeni, sarcinile de predicție a legării și similaritatea secvențelor pozitiv-negative.

Figură adaptată după Figura 2 din Ursu, E. et al. (2025), utilizată sub licența CC BY 4.0

Modelele SN10 au fost antrenate pe fiecare sarcină și evaluate pe ambele seturi de testare, atât ID, cât și OOD (Fig. 9f). Modelul „vs Non-binder” a întâmpinat dificultăți la sarcina „vs Weak” (precizie OOD de 68%), în timp ce modelul „vs Weak” a generalizat bine (precizie OOD de 88%). Așa cum s-a întâmplat în experimentele sintetice, antrenarea „vs Weak” a dus la limite de decizie mai strânse și mai generalizabile, care au echilibrat performanța pe toate sarcinile. Prin contrast, antrenarea „vs Non-binder” a condus la limite de decizie mai permissive, care au oferit o rată de rechemare (recall) ridicată în contextele ID, dar au dus la rezultate fals pozitive în scenariile OOD.

2. Compoziția setului de date de antrenament determină acuratețea recuperării regulilor biologice

După ce am stabilit că modul de compoziție a datelor de antrenament negative influențează acuratețea predicției și generalizarea în sarcinile de învățare automată (IA) supervizată binară bazată pe secvențe, am investigat în continuare dacă acesta afectează și capacitatea modelului de a învăța reguli de legare semnificative din punct de vedere biologic.

2.1 Compoziția Setului de Date de Antrenament Impactează Învățarea Energiei de Legare a Anticorpilor

Am întrebat mai întâi dacă modelele antrenate au capturat peisajul energetic al secvențelor de anticorpi. Deși unele studii au explorat predicția simultană a statusului de legare și a afinității ([122,123](#)), clasificarea binară (legare vs. non-legare) este adesea mai ușor de realizat experimental. Pentru a evalua dacă modelul nostru, SN10, învață implicit relațiile secvență-energie de legare, am calculat corelația între logiții de ieșire prezisă (adică, activările brute pre-sigmoid) și energiile de legare reale (ground-truth) pentru fiecare secvență. Acești logiți reflectă încrederea modelului în prezicerea apartenenței la clasa pozitivă (Fig. 10a). O distribuție 2D reprezentativă pentru antigenul 3VRL în sarcina „vs Weak” este prezentată în Fig. 10b (insert).

În general, modelele antrenate pe seturile de date „vs 1” și „vs 9” nu au reușit să învețe regulile energetice per-secvență, cu corelații slabe de $-0,05$ și, respectiv, $-0,19$ (Fig. 10b). Excepții notabile au fost antigenele 3VRL și 5E94, unde modelele au învățat energiile de legare mai eficiente, indiferent de tipul setului de date negativ. În contrast, modelele antrenate cu negative „vs Weak” sau „vs Non-binder” au învățat în mod constant asocieri energetice semnificative pentru toți antigenii, atingând corelații Pearson mediane de $-0,62$ (interval: $-0,33$ la $-0,90$) și, respectiv, $-0,51$ (interval: $-0,29$ la $-0,85$).

Aceste rezultate indică faptul că capacitatea de a învăța reguli bazate pe energie depinde puternic de tipul de date de antrenament negative. Acest lucru este susținut de o varianță semnificativă a valorilor de corelație între sarcini (ANOVA unidirecțională, $p = 8,7e-58$) și între antigeni (ANOVA unidirecțională, $p = 4,6e-25$).

Pentru a evalua dacă învățarea regulilor energiei de legare afectează performanța modelului, am corelat asocierile logit-energie cu acuratețea predicției atât în ID (in-distribution), cât și în OOD (out-of-distribution). Asocieri semnificative au fost observate în setările ID pentru sarcinile „vs Weak” ($r = -0,77$) și „vs Non-binder” ($r = -0,83$) (Fig. 10c). Pentru sarcinile OOD, acuratețea a fost asociată semnificativ cu învățarea regulilor bazate pe logit în toate cazurile (Fig. supl. 4b

din Ursu, Minnegalieva et al. (121)).

2.2 Compoziția Setului de Date de Antrenament Impactează Învățarea Contribuției Poziționale la Energia de Legare

După ce am arătat că SN10 poate capta reguli energetice per-secvență, am investigat în continuare dacă acesta ar putea învăța și contribuțiile specifice poziției ale reziduurilor de aminoacizi la legare — un determinant esențial al recunoașterii imune. Ipoteza noastră a fost că regulile învățate corect ar avea ca rezultat corelații negative între valorile de atribuire (calculate prin DeepLIFT (114)) și energia de legare per-reziduu: reziduurile cu legare mai puternică (energie mai mică) ar trebui să primească o atribuire mai mare în predicțiile pozitive.

În concordanță cu rezultatele per-secvență, cea mai puternică învățare a regulilor per-reziduu a avut loc în modelele antrenate pe seturile de date „vs Weak” (mediană $r = -0,69$) și „vs Non-binder” (mediană $r = -0,71$). În contrast, modelele antrenate pe sarcinile „vs 9” și „vs 1” au arătat asocieri mai slabe sau chiar inversate ($-0,36$ și, respectiv, $-0,01$). În special, pentru antigene precum 3VRL, 5E94 și 3RAJ, corelațiile au rămas ridicate în toate seturile de date negative — sugerând că, în unele cazuri, datele pozitive singure pot fi suficiente pentru a învăța reguli la nivel de poziție.

Tipul sarcinii a influențat semnificativ învățarea regulilor (ANOVA unidirecțională, $p = 3,5e-43$). Nu a fost găsită nicio corelație semnificativă între acuratețea predicției ID și învățarea regulilor per-reziduu (Fig. 10e), subliniind importanța evaluării metricei de explicabilitate în mod independent. Cu toate acestea, acuratețea OOD a corelat cu acordul atribuire-energie în majoritatea setărilor „vs 1” și „vs 9” (Fig. supl. 4b din Ursu, Minnegalieva et al. (121)). În contrast, astfel de asocieri nu au fost observate pentru modelele antrenate pe seturile de date „vs Weak” sau „vs Non-binder”.

2.3 Investigarea Amploarei Aditivității în Regulele Învățate

Deși sarcinile „vs Weak” și „vs Non-binder” au produs profiluri de atribuire foarte similare, performanța lor diferită în afara distribuției (OOD) ne-a determinat să explorăm dacă aceste modele se bazează mai mult pe reguli de decizie aditive sau non-aditive. Mai exact, am urmărit să înțelegem dacă diferențele de generalizare ar putea fi explicate prin amploarea interacțiunilor de caracteristici învățate în timpul antrenamentului. Pentru a evalua acest lucru, am antrenat modele de regresie logistică. Aceste modele nu au capacitatea de interacțiune a caracteristicilor. Ca atare, ele servesc ca etalon pentru învățarea pur aditivă a regulilor. Constatările implică faptul că sarcinile „vs Non-binder” pot fi rezolvate în principal folosind caracteristici aditive, în timp ce „vs Weak” necesită reprezentări mai complexe, bazate pe interacțiune. Acest lucru se aliniază cu performanța superioară a SN10 față de regresia logistică în sarcini mai dificile, cum ar fi „vs Weak” (Fig. supl. 3c din Ursu, Minnegalieva et al. (121)), susținând ideea că SN10 beneficiază de capacitatea sa de a modela interacțiunile caracteristicilor.

Discuție

Definirea datelor pozitive și negative influențează fundamental comportamentul modelelor de învățare automată, dar rămâne o zonă insuficient explorată în aplicațiile imunologice de *machine learning* (ML). În acest studiu, am demonstrat că atât performanța în distribuție (ID) și

în afara distribuției (OOD), cât și interpretabilitatea regulilor învățate, depind puternic de compoziția setului de date de antrenament în contextul predicției legării anticorp-antigen.

Mai general, rezultatele noastre subliniază faptul că organizarea datelor de antrenament nu este doar un pas pregătitor, ci o considerație de design centrală în dezvoltarea de modele ML robuste și generalizabile. Planificarea și justificarea atentă a compoziției setului de date sunt critice pentru obținerea de rezultate predictive valide și pentru descoperirea de reguli cu semnificație biologică, în special atunci când modelele sunt aplicate dincolo de scopul imediat al datelor lor de antrenament.

Lucrarea noastră demonstrează că definirea datelor negative de antrenament are un impact profund asupra comportamentului modelului ML, incluzând precizia predicției, generalizarea (ID vs. OOD) și descoperirea de reguli de legare cu semnificație biologică în interacțiunile anticorp-antigen. În ciuda importanței sale, acest subiect rămâne insuficient explorat în domeniul anticorpilor în comparație cu modul în care este tratat în studiile de predicție a epitopilor TCR [\(107–111\)](#).

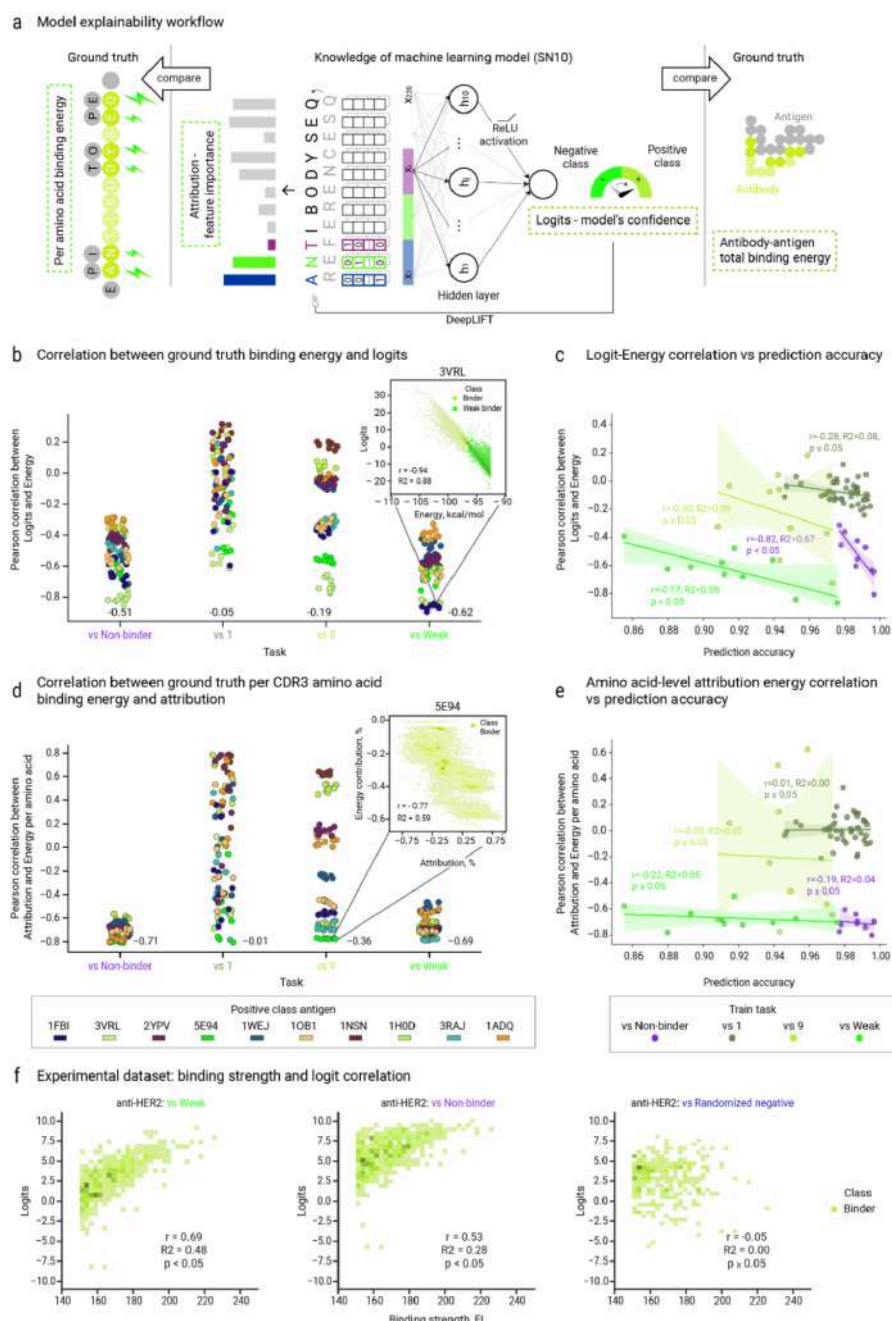


Figura 10. Compoziția setului de date negativ modelează capacitatea modelului de a învăța regulile de legare pentru secvențele pozitive.

Figură adaptată după Figura 3 din Ursu, E. et al. (2025), utilizată sub licența CC BY 4.0

Descoperirile noastre subliniază că: Definirea atentă a seturilor de date negative este crucială pentru modelele ML interpretabile și generalizabile în imunologie; Performanța OOD ar trebui să fie o metrică explicită în evaluarea descoperirii de reguli și a interpretabilității modelului, iar metodele actuale de atribuire au limitări, fiind necesare noi tehnici pentru a surprinde interacțiunile de caracteristici, cum ar fi epistazia, care stau probabil la baza fenomenelor biologice complexe (124–126).

În concluzie, descoperirile noastre demonstrează că: Modelele antrenate pe seturi de date cu eşantioane pozitive și negative bine echilibrate generalizează mai bine și învață reguli biologice mai robuste; Designul setului de date negative nu este doar un pas tehnic, ci un determinant

central al comportamentului și interpretabilității modelului; Lucrările viitoare ar trebui să exploreze în mod sistematic strategii pentru selectarea cazurilor negative dificile (hard negatives) (127), să investigheze secvențele de tip "aproape-ratate" (near-miss sequences) și să dezvolte metode de atribuire care să surprindă determinanții complecși ai legării.

VI. Considerații finale

Lucrarea prezentată în această teză a fost motivată de o întrebare centrală: cum putem înțelege mai bine complexitatea biologică a îmbătrânirii prin prisma omicii moderne, a biologiei computaționale și a învățării automate (machine learning)? Pentru a aborda acest subiect, am integrat seturi de date multidimensionale, am dezvoltat instrumente computaționale inovatoare și am colaborat interdisciplinar pentru a explora mecanismele cheie ale îmbătrânirii dintr-o perspectivă de sistem.

De-a lungul celor cinci capitole, am abordat îmbătrânirea din unghiuri distincte, dar interconectate. În Capitolul II, am explorat semnăturile transcriptomice la 41 de specii de mamifere, descoperind tipare de expresie genică conservate și specifice organelor, asociate cu durata de viață. Capitolul III a extins intersecția dintre îmbătrânire și remodelarea fibrotică în plămâni, evidențiind atât tipare moleculare conservate, cât și semnături unice de îmbătrânire într-un model de fibroză bine caracterizat. În Capitolul IV, m-am concentrat pe comunicarea intercelulară — o trăsătură distinctivă a îmbătrânirii care rămâne insuficient studiată — și am co-dezvoltat scDiffCom și scAgeCom, oferind un cadru scalabil și generalizabil pentru a mapa și cuantifica perturbările de semnalizare asociate vârstei. În cele din urmă, în Capitolul V, m-am îndreptat către sistemul imunitar adaptiv, combinând învățarea automată cu date AIRR-Seq pentru a examina modul în care performanța modelului și înțelegerea biologică sunt influențate de compoziția seturilor de date de antrenament — o contribuție fundamentală pentru studiile viitoare privind îmbătrânirea repertoriului imun.

Pe tot parcursul acestei teze, am căutat nu doar să generez rezultate, ci să contribui cu instrumente, cadre și metodologii care să sprijine comunitatea de cercetare a îmbătrânirii și în științele vieții. Aceasta include strategii de învățare automată interpretabile, protocoale analitice inter-specifice și seturi de date robuste, puse la dispoziție public pentru o reutilizare mai largă. Cercetarea mea reflectă convingerea că o înțelegere biologică profundă apare adesea la intersecția dintre bogăția datelor, rigoarea metodologică și colaborarea interdisciplinară, dar și înțelegerea biologică fundamentală.

Deși multe întrebări rămân deschise, sper că această teză oferă un pas semnificativ spre înțelegerea determinanților moleculari ai îmbătrânirii. Lucrările viitoare pot extinde instrumentele dezvoltate aici, aplicându-le la seturi de date umane longitudinale, studii intervenționale sau screening-uri terapeutice. Obiectivul final rămâne același: de a descurca complexitatea îmbătrânirii în moduri care pot informa strategii pentru vieți mai sănătoase și mai lungi.

Bibliografie

1. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging.

Cell. 2013 Jun 6;153(6):1194–217.

2. Kenyon CJ. The genetics of ageing. *Nature*. 2010 Mar 25;464(7288):504–12.
3. Johnson SC, Rabinovitch PS, Kaeberlein M. mTOR is a key modulator of ageing and age-related disease. *Nature*. 2013 Jan 17;493(7432):338–45.
4. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020 Oct 17;396(10258):1204–22.
5. Kaeberlein M, Rabinovitch PS, Martin GM. Healthy aging: The ultimate preventative medicine. *Science*. 2015 Dec 4;350(6265):1191–3.
6. Kennedy BK, Berger SL, Brunet A, Campisi J, Cuervo AM, Epel ES, et al. Geroscience: linking aging to chronic disease. *Cell*. 2014 Nov 6;159(4):709–13.
7. Seals DR, Justice JN, LaRocca TJ. Physiological geroscience: targeting function to increase healthspan and achieve optimal longevity. *J Physiol (Lond)*. 2016 Apr 15;594(8):2001–24.
8. Mannick JB, Del Giudice G, Lattanzi M, Valiante NM, Praestgaard J, Huang B, et al. mTOR inhibition improves immune function in the elderly. *Sci Transl Med*. 2014 Dec 24;6(268):268ra179.
9. Partridge L, Deelen J, Slagboom PE. Facing up to the global challenges of ageing. *Nature*. 2018 Sep 5;561(7721):45–56.
10. Williams GC. Pleiotropy, natural selection, and the evolution of senescence. *Evolution*. 1957 Dec;11(4):398–411.
11. Harman D. Aging: a theory based on free radical and radiation chemistry. *J Gerontol*. 1956 Jul;11(3):298–300.
12. López-Otín C, Galluzzi L, Freije JMP, Madeo F, Kroemer G. Metabolic control of longevity. *Cell*. 2016 Aug 11;166(4):802–21.
13. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. 2017 May 5;18(1):10.1186/s13059-017-1215–1.
14. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol*. 2013;14(10):R115.
15. Zhavoronkov A, Mamoshina P, Vanhaelen Q, Scheibye-Knudsen M, Moskalev A, Aliper A. Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing Res Rev*. 2019 Jan;49:49–66.
16. Austad SN. Comparative biology of aging. *J Gerontol A Biol Sci Med Sci*. 2009 Feb 1;64(2):199–201.
17. de Magalhães JP, Budovsky A, Lehmann G, Costa J, Li Y, Fraifeld V, et al. The Human Ageing Genomic Resources: online databases and tools for biogerontologists. *Aging Cell*. 2009 Feb;8(1):65–72.
18. Ma S, Sun S, Geng L, Song M, Wang W, Ye Y, et al. Caloric Restriction Reprograms the Single-Cell Transcriptional Landscape of *Rattus Norvegicus* Aging. *Cell*. 2020 Mar 5;180(5):984-1001.e22.

19. Fang X, Nevo E, Han L, Levanon EY, Zhao J, Avivi A, et al. Genome-wide adaptive complexes to underground stresses in blind mole rats *Spalax*. *Nat Commun*. 2014 Jun 3;5:3966.
20. Wynn TA, Ramalingam TR. Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nat Med*. 2012 Jul 6;18(7):1028–40.
21. Selman M, Pardo A. Revealing the pathogenic and aging-related mechanisms of the enigmatic idiopathic pulmonary fibrosis: an integral model. *Am J Respir Crit Care Med*. 2014 May 15;189(10):1161–72.
22. Schafer MJ, White TA, Iijima K, Haak AJ, Ligresti G, Atkinson EJ, et al. Cellular senescence mediates fibrotic pulmonary disease. *Nat Commun*. 2017 Feb 23;8:14532.
23. Rojas M, Mora AL, Kapetanaki M, Weathington N, Gladwin M, Eickelberg O. Aging and lung disease: clinical impact and cellular and molecular pathways. *Ann Am Thorac Soc*. 2015 Dec;12(12):S222–7.
24. Franceschi C, Campisi J. Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. *J Gerontol A Biol Sci Med Sci*. 2014 Jun;69 Suppl 1:S4–9.
25. Campisi J. Aging, cellular senescence, and cancer. *Annu Rev Physiol*. 2013;75:685–705.
26. Fülöp T, Larbi A, Pawelec G. Human T cell aging and the impact of persistent viral infections. *Front Immunol*. 2013 Sep 13;4:271.
27. Nikolic-Zugich J. The twilight of immunity: emerging concepts in aging of the immune system. *Nat Immunol*. 2018 Jan;19(1):10–9.
28. Frasca D, Blomberg BB. Aging induces B cell defects and decreased antibody responses to influenza infection and vaccination. *Immun Ageing*. 2020 Nov 19;17(1):37.
29. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med*. 2015 Nov 20;7:121.
30. Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol*. 2015 Nov;36(11):738–49.
31. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015 Jun;16(6):321–32.
32. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet*. 2019 Jul;20(7):389–403.
33. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009 Jan;10(1):57–63.
34. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016 Jan 26;17(1):13.
35. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*. 2009 May;6(5):377–82.
36. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018 Apr;15(141).

37. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. *NeurIPS Proceedings* [Internet]. 2017 [cited 2020 Dec 11]; Available from: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
38. Azodi CB, Tang J, Shiu S-H. Opening the black box: interpretable machine learning for geneticists. *Trends Genet.* 2020 Jun;36(6):442–55.
39. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol.* 2014 Feb;32(2):158–68.
40. Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, et al. Insights into the evolution of longevity from the bowhead whale genome. *Cell Rep.* 2015 Jan 6;10(1):112–22.
41. Gorbunova V, Seluanov A, Zhang Z, Gladyshev VN, Vijg J. Comparative genetics of longevity and cancer: insights from long-lived rodents. *Nat Rev Genet.* 2014 Aug;15(8):531–40.
42. Fushan AA, Turanov AA, Lee S-G, Kim EB, Lobanov AV, Yim SH, et al. Gene expression defines natural changes in mammalian lifespan. *Aging Cell.* 2015 Jun;14(3):352–65.
43. Ma S, Gladyshev VN. Molecular signatures of longevity: Insights from cross-species comparative studies. *Semin Cell Dev Biol.* 2017 Oct;70:190–203.
44. Toren D, Kulaga A, Jethva M, Rubin E, Snezhkina AV, Kudryavtseva AV, et al. Gray whale transcriptome reveals longevity adaptations associated with DNA repair and ubiquitination. *Aging Cell.* 2020 Jul;19(7):e13158.
45. Ma S, Upneja A, Galecki A, Tsai Y-M, Burant CF, Raskind S, et al. Cell culture-based profiling across mammals reveals DNA repair and metabolism as determinants of species longevity. *eLife.* 2016 Nov 22;5.
46. Huang Z, Whelan CV, Foley NM, Jebb D, Touzalin F, Petit EJ, et al. Longitudinal comparative transcriptomics reveals unique mechanisms underlying extended healthspan in bats. *Nat Ecol Evol.* 2019 Jul;3(7):1110–20.
47. Hilton HG, Rubinstein ND, Janki P, Ireland AT, Bernstein N, Fong NL, et al. Single-cell transcriptomics of the naked mole-rat reveals unexpected features of mammalian immunity. *PLoS Biol.* 2019 Nov 21;17(11):e3000528.
48. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, et al. Ensembl comparative genomics resources. *Database (Oxford).* 2016 Feb 20;2016.
49. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* 2018 Sep 1;34(17):i884–90.
50. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods.* 2017 Apr;14(4):417–9.
51. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 2012 Jan;40(Database issue):D109–14.
52. Kulaga AY, Ursu E, Toren D, Tyshchenko V, Guinea R, Pushkova M, et al. Machine

Learning Analysis of Longevity-Associated Gene Expression Landscapes in Mammals. *Int J Mol Sci.* 2021 Jan 22;22(3).

53. Lagani V, Athineou G, Farcomeni A, Tsagris M, Tsamardinos I. Feature Selection with the *R* Package *MXM*: Discovering Statistically Equivalent Feature Subsets. *J Stat Softw.* 2017;80(7).
54. Tacutu R, Thornton D, Johnson E, Budovsky A, Barardo D, Craig T, et al. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res.* 2018 Jan 4;46(D1):D1083–90.
55. Muradian KK, Utko NA, Mozzhukhina TG, Litoshenko AY, Pishel IN, Bezrukov VV, et al. Pair-wise linear and 3D nonlinear relationships between the liver antioxidant enzyme activities and the rate of body oxygen consumption in mice. *Free Radic Biol Med.* 2002 Dec 15;33(12):1736–9.
56. Lehmann G, Segal E, Muradian KK, Fraifeld VE. Do mitochondrial DNA and metabolic rate complement each other in determination of the mammalian maximum longevity? *Rejuvenation Res.* 2008 Apr;11(2):409–17.
57. Lehmann G, Muradian KK, Fraifeld VE. Telomere length and body temperature-independent determinants of mammalian longevity? *Front Genet.* 2013 Jun 13;4:111.
58. Yanai H, Budovsky A, Barzilay T, Tacutu R, Fraifeld VE. Wide-scale comparative analysis of longevity genes and interventions. *Aging Cell.* 2017 Dec;16(6):1267–75.
59. Tacutu R, Budovsky A, Yanai H, Fraifeld VE. Molecular links between cellular senescence, longevity and age-related diseases - a systems biology perspective. *Aging (Albany NY).* 2011 Dec;3(12):1178–91.
60. Wolfson M, Budovsky A, Tacutu R, Fraifeld V. The signaling hubs at the crossroad of longevity and age-related disease networks. *Int J Biochem Cell Biol.* 2009 Mar;41(3):516–20.
61. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017;
62. Ricklefs RE. Life-history connections to rates of aging in terrestrial vertebrates. *Proc Natl Acad Sci USA.* 2010 Jun 1;107(22):10314–9.
63. Li J, Liu L, Le TD. Practical approaches to causal relationship exploration. Cham: Springer International Publishing; 2015.
64. Tacutu R, Thornton D, Johnson E, Budovsky A, Barardo D, Craig T, et al. Human ageing genomic resources: New and updated databases. *Nucl Acids Res.* 46:1083– 1090.
65. Miller HA, Dean ES, Pletcher SD, Leiser SF. Cell non-autonomous regulation of health and longevity. *eLife.* 2020 Dec 10;9.
66. Franceschi C, Garagnani P, Parini P, Giuliani C, Santoro A. Inflammaging: a new immune-metabolic viewpoint for age-related diseases. *Nat Rev Endocrinol.* 2018 Oct;14(10):576–90.
67. Ovadya Y, Landsberger T, Leins H, Vadai E, Gal H, Biran A, et al. Impaired immune surveillance accelerates accumulation of senescent cells and aging. *Nat Commun.* 2018

Dec 21;9(1):5435.

68. Fafián-Labora JA, O’Loghlen A. Classical and nonclassical intercellular communication in senescence and ageing. *Trends Cell Biol.* 2020 Aug;30(8):628–39.
69. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell-cell interactions and communication from gene expression. *Nat Rev Genet.* 2021;22(2):71–88.
70. Shao X, Lu X, Liao J, Chen H, Fan X. New avenues for systematically inferring cell-cell communication: through single-cell transcriptomics data. *Protein Cell.* 2020 Dec;11(12):866–80.
71. Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature.* 2020 Jul 15;583(7817):590–5.
72. Kimmel JC, Penland L, Rubinstein ND, Hendrickson DG, Kelley DR, Rosenthal AZ. Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Res.* 2019 Dec;29(12):2088–103.
73. A Framework for Robust Shiny Applications [R package golem version 0.3.1] [Internet]. 2021 [cited 2021 Jun 11]. Available from: <https://cran.r-project.org/web/packages/golem/index.html>
74. Plotly Technologies Inc. Collaborative data science Publisher: Plotly Technologies Inc. Montréal, QC; 2015.
75. Sayols S. rrvgo: a Bioconductor package to reduce and visualize Gene Ontology term . Bioconductor; 2020.
76. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE.* 2011 Jul 18;6(7):e21800.
77. Meissner F, Scheltema RA, Mollenkopf H-J, Mann M. Direct proteomic quantification of the secretome of activated immune cells. *Science.* 2013 Apr 26;340(6131):475–8.
78. Tüshaus J, Müller SA, Kataka ES, Zaucha J, Sebastian Monasor L, Su M, et al. An optimized quantitative proteomics method establishes the cell type-resolved mouse brain secretome. *EMBO J.* 2020 Oct 15;39(20):e105693.
79. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan C-H, et al. Inference and analysis of cell-cell communication using CellChat. *Nat Commun.* 2021 Feb 17;12(1):1088.
80. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc.* 2020 Apr;15(4):1484–506.
81. Shao X, Liao J, Li C, Lu X, Cheng J, Fan X. CellTalkDB: a manually curated database of ligand-receptor interactions in humans and mice. *Brief Bioinformatics.* 2021 Jul 20;22(4).
82. Hou R, Denisenko E, Ong HT, Ramilowski JA, Forrest ARR. Predicting cell-to-cell communication networks using NATMI. *Nat Commun.* 2020 Oct 6;11(1):5011.
83. Noël F, Massenet-Regad L, Carmi-Levy I, Cappuccio A, Grandclaude M, Trichot C, et al. Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat Commun.* 2021 Feb 17;12(1):1089.

84. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods*. 2020 Feb;17(2):159–62.
85. Cabello-Aguilar S, Alame M, Kon-Sun-Tack F, Fau C, Lacroix M, Colinge J. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res*. 2020 Jun 4;48(10):e55.
86. Lager C, Ursu E, Equey A, Avelar RA, Pisco AO, Tacutu R, et al. scDiffCom: a tool for differential analysis of cell-cell interactions provides a mouse atlas of aging changes in intercellular communication. *Nat Aging*. 2023 Nov 2;3(11):1446–61.
87. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: Tool for the unification of biology. *Nat Genet*. 2000 May;25(1):25–9.
88. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017 Jan 4;45(D1):D353–61.
89. Avelar RA, Ortega JG, Tacutu R, Tyler EJ, Bennett D, Binetti P, et al. A multidimensional systems biology analysis of cellular senescence in aging and disease. *Genome Biol*. 2020 Apr 7;21(1):91.
90. Tejada-Martinez D, Avelar RA, Lopes I, Zhang B, Novoa G, de Magalhães JP, et al. Positive selection and enhancer evolution shaped lifespan and body mass in great apes. *Mol Biol Evol*. 2022 Feb 3;39(2).
91. Budovsky A, Craig T, Wang J, Tacutu R, Csordas A, Lourenço J, et al. LongevityMap: a database of human genetic variants associated with longevity. *Trends Genet*. 2013 Oct;29(10):559–60.
92. Unable to find information for 44699.
93. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. 2015 May;33(5):495–502.
94. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018 Jun;36(5):411–20.
95. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019 Jun 13;177(7):1888-1902.e21.
96. Büttner M, Ostner J, Müller CL, Theis FJ, Schubert B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat Commun*. 2021 Nov 25;12(1):6876.
97. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature*. 2018 Nov 14;563(7731):347–53.
98. Hummer AM, Schneider C, Chinery L, Deane CM. Investigating the Volume and Diversity of Data Needed for Generalizable Antibody-Antigen $\Delta\Delta G$ Prediction. *BioRxiv*. 2023 May 19;
99. Yang R, Mao J, Chaudhari P. Does the data induce capacity control in deep learning?. *International Conference on Machine Learning*. 2022;25166.
100. Alvarez-Melis D, Fusi N. Geometric dataset distances via optimal transport. *Advances in*

Neural Information Processing Systems. 2020;33:21428–39.

101. Wang T, Isola P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *International conference on machine learning*. 2020;9929.
102. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, editors. *Dataset shift in machine learning*. The MIT Press; 2008.
103. Li XL, Liu B, Ng SK. Negative training data can be harmful to text classification. *Proceedings of the 2010 conference on empirical methods in natural language processing*. 2010;218.
104. Schneider C, Buchanan A, Taddese B, Deane CM. DLAB: deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics*. 2022 Jan 3;38(2):377–83.
105. Krützfeldt L-M, Schubach M, Kircher M. The impact of different negative training data on regulatory sequence predictions. *PLoS ONE*. 2020 Dec 1;15(12):e0237412.
106. Robert PA, Akbar R, Frank R, Pavlović M, Widrich M, Snapkov I, et al. Unconstrained generation of synthetic antibody-antigen structures to guide machine learning methodology for antibody specificity prediction. *Nat Comput Sci*. 2022 Dec 19;2(12):845–65.
107. Montemurro A, Jessen LE, Nielsen M. NetTCR-2.1: Lessons and guidance on how to develop models for TCR specificity predictions. *Front Immunol*. 2022 Dec 6;13:1055151.
108. Grazioli F, Mösch A, Machart P, Li K, Alqassem I, O'Donnell TJ, et al. On TCR binding predictors failing to generalize to unseen peptides. *Front Immunol*. 2022 Oct 21;13:1014256.
109. Deng L, Ly C, Abdollahi S, Zhao Y, Prinz I, Bonn S. Performance comparison of TCR-pMHC prediction tools reveals a strong data dependency. *Front Immunol*. 2023 Apr 18;14:1128326.
110. Dens C, Laukens K, Bittremieux W, Meysman P. The pitfalls of negative data bias for the T-cell epitope specificity challenge. *Nat Mach Intell*. 2023 Oct 5;
111. Gao Y, Gao Y, Dong K, Wu S, Liu Q. Reply to: The pitfalls of negative data bias for the T-cell epitope specificity challenge. *Nat Mach Intell*. 2023 Oct 5;5(10):1063–5.
112. Sundararajan M, Taly A, Yan Q. [1611.02639] Gradients of Counterfactuals. *arXiv*. 2016 Nov 8;
113. Karim MR, Islam T, Shajalal M, Beyan O, Lange C, Cochez M, et al. Explainable AI for bioinformatics: methods, tools and applications. *Brief Bioinformatics*. 2023 Sep 20;24(5).
114. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. *International conference on machine learning*. 2017;3145.
115. Porebski BT, Balmforth M, Browne G, Riley A, Jamali K, Fürst MJLJ, et al. Rapid discovery of high-affinity antibodies via massively parallel sequencing, ribosome display and affinity screening. *Nat Biomed Eng*. 2024 Mar;8(3):214–32.
116. Akbar R, Robert PA, Weber CR, Widrich M, Frank R, Pavlović M, et al. In silico proof of principle of machine learning-based antibody design at unconstrained scale. *MAbs*. 2022;14(1):2031482.

117. Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook S, et al. Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Rep.* 2017 May 16;19(7):1467–78.
118. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science.* 2023 Mar 17;379(6637):1123–30.
119. Leem J, Mitchell LS, Farmery JHR, Barton J, Galson JD. Deciphering the language of antibodies using self-supervised learning. *Patterns (N Y).* 2022 Jul 8;3(7):100513.
120. Barton J, Galson JD, Leem J. Enhancing Antibody Language Models with Structural Information. *BioRxiv.* 2024 Jan 4;
121. Ursu E, Minnegalieva A, Rawat P, Chernigovskaya M, Tacutu R, Sandve GK, et al. Training data composition determines machine learning generalization and biological rule discovery. *BioRxiv.* 2024 Jun 19;
122. Hu F, Jiang J, Wang D, Zhu M, Yin P. Multi-PLI: interpretable multi-task deep learning model for unifying protein-ligand interaction datasets. *J Cheminform.* 2021 Apr 15;13(1):30.
123. Pei Q, Wu L, Zhu J, Xia Y, Xie S, Qin T, et al. Breaking the barriers of data scarcity in drug-target affinity prediction. *Brief Bioinformatics.* 2023 Sep 22;24(6).
124. Adams RM, Kinney JB, Walczak AM, Mora T. Epistasis in a Fitness Landscape Defined by Antibody-Antigen Binding Free Energy. *Cell Syst.* 2019 Jan 23;8(1):86-93.e3.
125. Starr TN, Thornton JW. Epistasis in protein evolution. *Protein Sci.* 2016 Jul;25(7):1204–18.
126. Cocco S, Posani L, Monasson R. Minimal epistatic networks from integrated sequence and mutational protein data. *BioRxiv.* 2023 Sep 25;
127. Xu L, Lian J, Zhao WX, Gong M, Shou L, Jiang D, et al. Negative Sampling for Contrastive Representation Learning: A Review. *arXiv.* 2022;