ROMANIAN ACADEMY

SCHOOL OF ADVANCED STUDIES OF THE ROMANIAN ACADEMY

INSTITUTE OF BIOCHEMISTRY

**PH.D. THESIS
SUMMARY**

Analysis of the immune repertoire changes during ageing

SCIENTIFIC COORDINATOR:
CS I Dr. Andrei-J. Petrescu

PH.D. STUDENT:
Eugen Ursu

Bucharest, 2025

Table of contents

Table of contents....... 1
I. Introduction....... 1
1. Aging....... 1
2. Computational biology and artificial intelligence....... 3
II. Cross-species gene expression reveals new gene candidates involved in aging and longevity....... 4
Overview....... 5
1. Introduction....... 5
2. Methods....... 5
3. Results and Discussion....... 6
4. Conclusions....... 10
III. Aging signature in lung fibrosis....... 10
Results and Discussion....... 11
Conclusions....... 11
IV. Intercellular communication is disrupted with aging....... 11
Overview....... 12
1. Introduction....... 12
2. Methods....... 12
3. Results and Discussion....... 13
4. Conclusions....... 18
V. Negative training data composition is critical for learning immune repertoires....... 18
1. Introduction....... 18
2. Methods....... 20
3. Results and Discussion....... 20
VI. Final remarks....... 26
Bibliography....... 26

I. Introduction

1. Aging

Aging is a complex biological process with a profound impact on health and disease risk. Contemporary research views it as a regulated phenomenon, outlined by the nine interconnected "hallmarks of aging" (genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intercellular communication)(1). This framework is foundational, guiding the investigation of cellular mechanisms and intervention targets. Key, conserved pathways influencing aging include the insulin/IGF-1 signaling and mTOR pathways(2,3). Experimental modulation of these, alongside interventions like caloric restriction (CR) and CR-mimicking compounds (rapamycin, metformin, NAD+-related therapeutics), has extended lifespan in various models(2,3). Understanding aging has broad implications for public health, as age is the leading risk factor for chronic diseases(4). The field of geroscience seeks to target aging mechanisms to delay or prevent multiple pathologies simultaneously(5).

Aging is intricately linked to the development of most chronic diseases (neurodegenerative disorders, cardiovascular diseases, metabolic syndromes, cancer)(6). The demographic shift towards older populations strains healthcare systems and expenditure, as a large share of resources is allocated to managing chronic, age-related illnesses, especially at the end of life(4). Proactive, preventative strategies are needed.

The goal in aging research is shifting from merely extending **lifespan** to extending **healthspan** (years lived in good health) (7). Current medicine typically treats chronic diseases in isolation, overlooking their common root in the underlying aging process(6). A more holistic approach involves intervening in the biological mechanisms of aging itself to improve overall health outcomes and reduce multimorbidity(1). Interventions targeting aging-related pathways, such as rapamycin, senolytics, and NAD+ enhancers, are being studied for their capacity to improve health across multiple domains by acting on core aging hallmarks (impaired cellular recycling, inflammation, mitochondrial dysfunction)(8).

Aging is driven by a constellation of deteriorating biological processes, distilled into the prominent "hallmarks of aging" model by López-Otín et al.(1). This framework connects diverse cellular events, such as DNA damage, senescence, impaired protein maintenance, and dysregulated metabolic pathways (insulin/IGF-1 signaling, mTOR, sirtuins)(1,9). Evolutionarily, aging is explained by the mutation accumulation, antagonistic pleiotropy, and disposable soma theories, which focus on selective pressure and trade-offs(10). Mechanistic views, like the free radical theory, focus on damage, though current understanding is more nuanced(11).

Omics technologies (genomics, transcriptomics, proteomics, metabolomics, epigenomics) provide a comprehensive, multilayered view of molecular changes during aging, identifying biomarkers and tracing age-associated decline(12). Computational biology and bioinformatics are essential for interpreting these vast datasets, supporting pathway analysis, network modeling, and the development of biological age predictors like epigenetic clocks(13,14). Artificial intelligence (AI) and machine learning (ML) further enhance analysis by uncovering subtle patterns, modeling aging trajectories, identifying risk profiles, and prioritizing therapeutic targets, accelerating the path to precision medicine for aging(15). Our work leverages these technologies to investigate immune system aging and fibrotic progression.

Comparative biology offers insights by studying species with diverse lifespans (e.g., naked mole rats, bowhead whales), highlighting conserved or uniquely adapted longevity mechanisms like enhanced proteostasis, DNA repair, and robust stress responses(16,17). Cross-species studies help distinguish age-related changes from chronological time by observing how different species manage aging over vastly different timeframes, often showing that long-lived mammals suppress inflammation and maintain genomic integrity more effectively(18). Transcriptomic analysis (RNA-Seq) across species identifies longevity-linked gene expression signatures, such as the downregulation of growth pathways and upregulation of maintenance genes, offering clues to conserved molecular pathways relevant to human health(19). **Chapter II** of this thesis uses cross-species transcriptomics to identify shared and species-specific gene expression changes related to aging, stress defense, and immune regulation.

Fibrosis (excessive deposition of extracellular matrix) is a key feature of age-related tissue dysfunction in organs like the heart, kidneys, liver, and lungs(20). Aging-associated changes (oxidative damage, chronic inflammation, fibroblast activation) create a fibrogenic environment. Idiopathic pulmonary fibrosis (IPF) is a prime, age-associated example, predominantly affecting individuals over 60(21). Mechanistically, aging promotes lung fibrosis via epithelial cell senescence, the pro-fibrotic senescence-associated secretory phenotype (SASP), mitochondrial dysfunction, and persistent inflammatory signaling (TGF-β, IL-6)(22,23). **Chapter III** covers collaborative work studying aging signatures in lung fibrosis using a mouse bleomycin-induced pulmonary fibrosis model.

Disrupted intercellular communication is a systemic hallmark of aging, manifesting as chronic low-grade inflammation (inflammaging), impaired immune surveillance, and altered endocrine signaling(1). "Inflammaging"(24) is fueled in part by senescent cells releasing SASP factors (cytokines, chemokines) that disrupt tissue and immune function, creating a deleterious feedback loop(25). Immunosenescence and changes in endocrine signaling further contribute to the dysregulation(26). Altered communication is a key driver of systemic aging. In **Chapter IV**, I describe my contributions: 1) developing a bioinformatics tool, scDiffCom, for inferring changes in intercellular communication from scRNA-Seq data, and 2) using scDiffCom to build an atlas of age-related changes in intercellular communications in mice (scAgeCom).

Immunosenescence—the age-related decline of the immune system—results in diminished pathogen responses, reduced vaccine efficacy, and increased age-associated diseases(27). A core component is the deterioration of the adaptive humoral response, marked by changes in the B cell compartment and reduced antibody repertoire diversity(28). Adaptive Immune Receptor Repertoire Sequencing (AIRR-Seq) allows deep analysis of B and T cell receptor sequences(29). However, aging studies face challenges due to cross-sectional datasets and inconsistent analytical pipelines(30). Traditional repertoire analysis often fails to capture the complex, nonlinear patterns of the aging immune system, prompting a turn toward machine learning (ML). **Chapter V** describes collaborative work in computational immunology focused on the impact of immune repertoire training data on ML models, which is crucial for developing tools for a better future understanding of age-related immune repertoire dynamics.

2. Computational biology and artificial intelligence

High-throughput omics technologies, such as genomics, transcriptomics, proteomics, metabolomics, and epigenomics, have generated extensive biological datasets(13). The magnitude and complexity of this data necessitate sophisticated computational methods. Computational biology and bioinformatics provide the essential infrastructure for data handling, preprocessing, and statistical interpretation. More recently, machine learning (ML) and artificial intelligence (AI) have become prominent, excelling at detecting intricate, often nonlinear, patterns in omics data(31). AI's influence spans domains from predicting gene regulatory activity from DNA sequences(32) to analyzing cell populations in single-cell transcriptomics. This intersection is enabling the transition from descriptive to predictive and mechanistic models, particularly in aging research.

RNA sequencing (RNA-Seq) is a transformative tool for global gene expression profiling, offering superior sensitivity and dynamic range compared to older methods(33). It enables downstream analyses like differential expression and network construction, providing insight into biological pathways and disease mechanisms(34). The emergence of single-cell RNA sequencing (scRNA-Seq) further revolutionized the field by resolving gene expression at the individual cell level. Unlike bulk RNA-Seq, scRNA-Seq reveals the diversity and heterogeneity within complex tissues, making it possible to identify rare cell types and lineage trajectories(35). My research extensively leveraged the high-resolution power of these technologies, and the analyses and discoveries presented in Chapters II, III, and IV relied on the deep insights enabled by RNA-Seq and scRNA-Seq.

Machine learning (ML) is integral to modern life sciences for pattern detection and prediction, especially in omics and biomedical imaging. Early ML focused on high accuracy, often using 'black box' models like deep neural networks(36). Due to concerns in fields like medicine, the focus has shifted to interpretable machine learning. Techniques such as SHAP and LIME allow researchers to understand how models make decisions, linking specific inputs (e.g., genes) to outputs(37). This interpretability has been crucial for uncovering causal genes and decoding biological logic(38). In aging research, interpretability is vital for translating predictions into biological meaning. In my own work, interpretability was a central pillar of the investigations in Chapters II, IV, and V.

The adaptive immune system’s diversity is based on millions of unique B cell receptors (BCRs) and antibodies. Adaptive Immune Receptor Repertoire Sequencing (AIRR-Seq) allows for deep profiling of this diversity, revolutionizing the study of humoral immunity across various conditions(39). AIRR-Seq captures millions of BCR sequences in parallel, enabling the reconstruction of clonotypes and lineage relationships (29). Analyzing this vast, complex data requires robust computational frameworks for tasks like sequence alignment, clustering, and mutation identification (30). My research contributes to this field by advancing machine learning techniques for AIRR-Seq, focusing on foundational challenges such as selecting negative training data to build more biologically relevant models. This foundational effort, described in Chapter V, aims to ultimately bridge gaps in mapping immune repertoire dynamics during aging.

II. Cross-species gene expression reveals new gene candidates involved in aging and longevity

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).*
** denotes first co-authorship*

Statement of Contributions

I contributed substantially to the design of the study, the development and implementation of the framework, data processing, and interpretation of the results. As a co–first author, I shared core responsibilities equally with Anton Kulaga and Dmitri Toren. Together, we collaboratively developed the conceptual and analytical components of the study, co-wrote all sections of the manuscript, and worked jointly on key methodological and interpretative decisions. The study was carried out under the coordination and supervision of Dr. Robi Tăcutu and Prof. Vadim E. Fraifeld.

Overview

In the field of aging research, a central challenge is to understand how variations in gene expression influence the longevity observed among different species. This study seeked to clarify the complex relationships between transcriptomic differences and maximum lifespan (MLS) by addressing the limitations of conventional linear models and the impact of confounding factors like metabolic rate, gestation period, and body mass.

1. Introduction

Investigating the diversity in MLS alongside gene transcription across species provides valuable insights into the evolutionary mechanisms of longevity. Recent research has revealed distinct differences in gene expression profiles between long- and short-lived mammals(40–43). Notably, exceptionally long-lived species show elevated levels of genes associated with DNA maintenance and repair, ubiquitination, immune responses, apoptosis, and autophagy(44). The overexpression of DNA repair genes is consistently highlighted in cross-species cell culture studies(45), and in several mammalian organs, immune response gene expression positively correlates with MLS(42). Pro-longevity adaptations in the transcriptome are clearly documented in species like bats(46), naked mole rats(41,47), and whales(40,44). This evidence underscores the potential of comparative transcriptomic studies to uncover the genetic basis of longevity.

Traditional comparative transcriptomic studies often relied on linear methods, overlooking non-linear biological processes. Recent advances, however, offer other solutions.

2. Methods

We framed the problem as a feature reduction task, seeking a minimal, highly predictive set of genes associated with Maximum Life Span (MLS) variation across different organs. Our pipeline integrates transcriptomic data from multiple species to detect longevity-associated genes (LAGs) using three complementary methods: linear modeling for organ-specific trends, LightGBM combined with SHAP for evaluating gene influence and interactions, and Bayesian network modeling to infer potential causal links with MLS.

This study included mammalian species with available RNA-Seq data for healthy liver, kidney, lung, brain, or heart tissues from the NCBI Sequence Read Archive, with transcriptome annotations sourced from Ensembl Compara Database(48). A total of 408 samples from 38 species spanning five organs were processed using a standardized RNA-Seq quantification pipeline. Orthology data and transcriptome annotations were sourced from Ensembl release 99. Overall 11,831 genes yielded.

Fastp (version 0.20.1)(49) handled quality control and trimming. Transcript quantification was performed with Salmon (version 1.4.0)(50), and tximport (version 3.12) aggregated expression at the gene level. Raw read counts were normalized using the transcripts per million (TPM) method for cross-species comparative analysis.

Organ-specific linear models were developed to identify genes associated with MLS. Pathway activity was assessed using single-sample gene set enrichment analysis (ssGSEA) with KEGG pathways(51).

To explore non-linear patterns, a two-step backward selection was used, combining LightGBM and SHAP. Six independent regression models first predicted different life-history traits (including MLS) from gene expression. Genes identified in any model were pooled for a second phase to refine the MLS-associated set. A rigorous 5-fold cross-validation with sorted stratification, repeated ten times, was applied, requiring non-zero SHAP values for significance. A strict data-splitting strategy ensured predictions relied on gene patterns, not species identification. The top 15 genes were selected based on the elbow of the SHAP feature importance graph (Supp. Fig. S4 from Kulaga, Ursu, Toren et al.(52).).

Bayesian networks were employed to map conditional independence and infer potential causal connections with MLS. The SES algorithm(53) was used for feature selection, identifying genes in the Markov blanket of MLS. SES was applied to the imputed training set, and the resulting gene signature was evaluated by a LightGBM model trained on the non-imputed data, selecting the signature with the lowest RMSE.

The three modeling approaches yielded ranked lists based on distinct metrics: (1) LightGBM-SHAP: frequency of non-zero mean absolute SHAP, Kendall’s tau, and mean absolute SHAP; (2) Linear Regression: maximum R2; and (3) Bayesian Networks: relative frequency in signatures. A final composite ranking was derived by summing the ranks of six key metrics: frequency of selection, Kendall’s tau correlation, mean absolute SHAP value, highest linear R2, relative frequency in Bayesian networks, and GenAge mention indicator(54). A Bayesian multilevel linear model with organ-specific random coefficients was constructed to analyze the final set of selected genes.

Based on the composite ranking and an elbow plot (Supp. Fig. S4 from Kulaga, Ursu, Toren et al.(52)), the top 15 genes were identified. A series of LightGBM-SHAP models, incorporating subsets from the top 5 to the top 15 genes, were developed. Model accuracy evaluation indicated that the top-6 gene model offered the best trade-off between simplicity and performance, showing a notable reduction in Huber loss (0.8) compared to the top-5 model.

3. Results and Discussion

A cross-species gene expression dataset, compiled from publicly available RNA-Seq data, offers a broad comparative framework. This dataset contains gene expression data across five major organs (liver, kidney, lung, brain, and heart) from 408 samples across 41 mammalian species. Species-specific variables, including MLS, body mass, temperature, metabolic rate, gestation period, and mitochondrial DNA GC content, were incorporated after normalization, as they are key longevity factors(55–57). To explore the link between gene expression and MLS, we used linear regression, interpretable LightGBM-SHAP modeling, and Bayesian network analysis. Integrating results from these models identified genes consistently ranking as top MLS predictors (see Fig. 1).

Linear models were used to assess the relationship between the expression of 11,831 evolutionarily conserved orthologs and Maximum Lifespan (MLS) in 33 mammalian species (Fig. 2a). The number of genes significantly associated with MLS (FDR < 0.05, R^2 > 0.3) varied by organ, peaking in the lung (756) and bottoming in the kidney (154). Median R2 values were consistent (0.35–0.38).

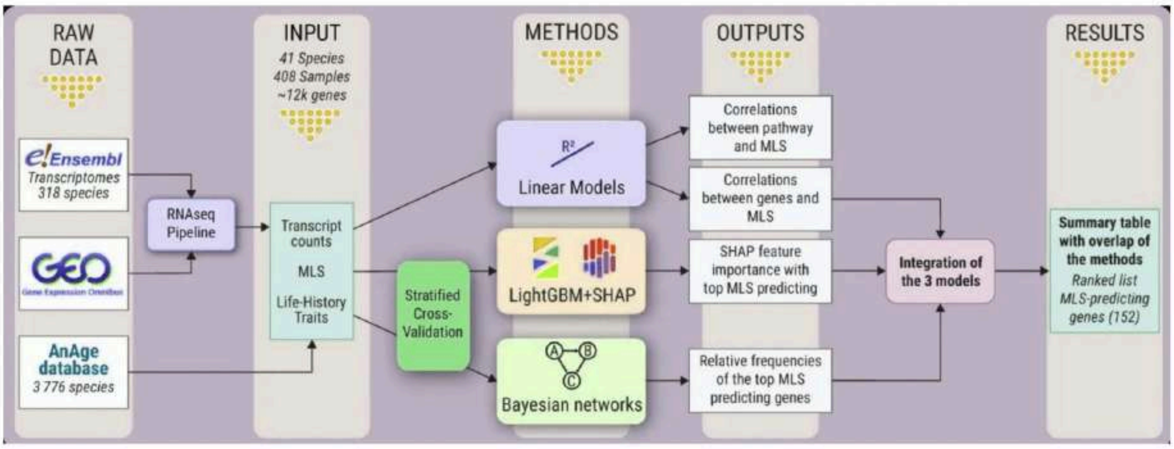

RAW DATA
INPUT
41 Species
408 Samples
~12k genes
METHODS
OUTPUTS
RESULTS
Ensembl
Transcriptomes
318 species
GEO
Gene Expression Omnibus
AnAge database
3 776 species
RNAseq Pipeline
Transcript counts
MLS
Life-History Traits
Stratified Cross-Validation
R²
Linear Models
LightGBM+SHAP
A
B
C
Bayesian networks
Correlations between pathway and MLS
Correlations between genes and MLS
SHAP feature importance with top MLS predicting
Relative frequencies of the top MLS predicting genes
Integration of the 3 models
Summary table with overlap of the methods
Ranked list MLS-predicting genes (152)

Figure 1. Schematic representation of the cross-species analysis workflow.
Figure adapted from Figure 1 from Kulaga, Ursu, Toren. et al. (2021), used under license CC BY 4.0.

Only three genes (CRYGS, TCFL5, and SPATA20) positively correlated with MLS across all five organs. Focusing on the brain, liver, and kidney increased MLS-associated genes to 12, including SPATA20, TCFL5, and CRYGS (Supp. Table S2 from Kulaga, Ursu, Toren et al.(52)).

Many MLS-associated genes also correlated with other life-history traits (e.g., body mass, metabolic rate), suggesting indirect associations. Only a few genes correlated uniquely with MLS: one in the liver (CERS4), four in the heart, and 131 in the lung; none in the brain or kidney. However, heart and lung results should be interpreted cautiously due to smaller sample sizes.

We used the signature projection approach (ssGSEA) to analyze the relationship between biological pathway activity, estimated from gene expression across organs, and maximum lifespan (MLS) (Fig. 2b). Our analysis focused on established aging/longevity pathways, including mTOR and insulin signaling, DNA repair, ubiquitin-mediated proteolysis, and focal adhesion(58–60).

While mTOR signaling did not significantly correlate with MLS, multiple DNA repair-related pathways (mismatch, nucleotide excision, base excision, homologous recombination, and non-homologous end-joining) showed robust positive correlations. Several other pathways were positively and negatively correlated with MLS. Unexpectedly, pathways not traditionally linked to longevity, such as apoptosis, cell adhesion molecules, and ErbB signaling, positively correlated with MLS. We validated known longevity pathways (apoptosis, DNA repair, immune responses) and highlighted underexplored ones—including PPAR signaling, glutathione metabolism, and ErbB signaling—as promising areas.

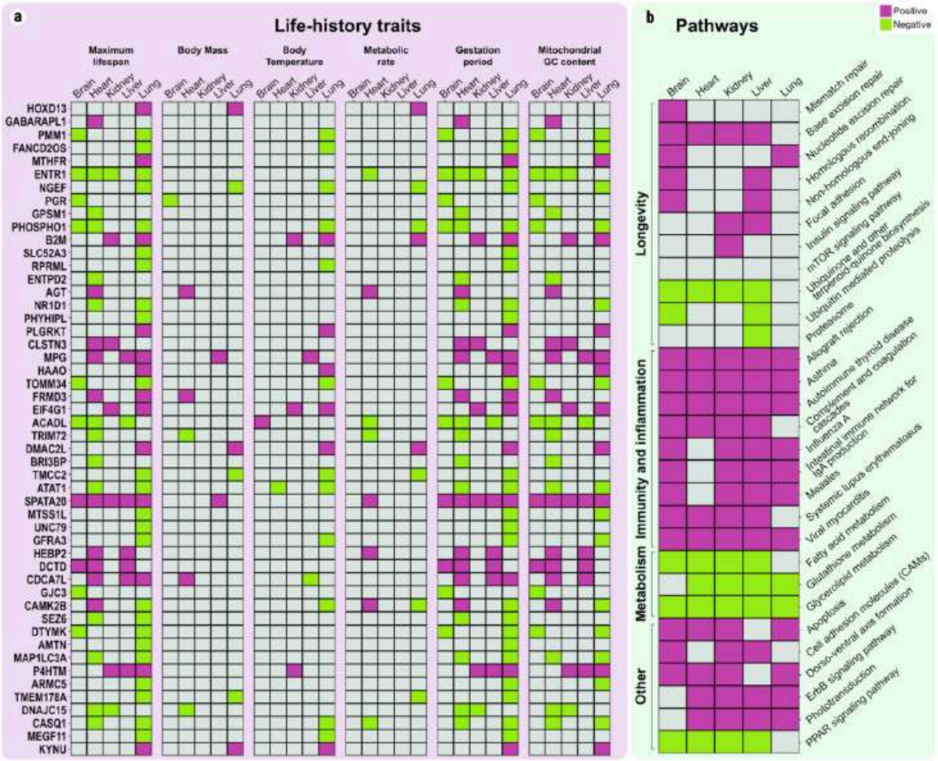


Figure 2. Linear correlations between gene expression and biological pathways with MLS.
(a) Top Linear Correlations Between Gene Expression and Species Traits. This heatmap highlights statistically significant relationships (FDR < 0.05, R² > 0.3) between gene expression levels and key species traits. (b) Top Linear Correlations Between MLS and Pathway Enrichment Scores The second heatmap visualizes key associations between MLS and pathway enrichment scores (ES).
Figure adapted from Figure 2 Kulaga, Ursu, Toren. et al. (2021), used under license CC BY 4.0.

To identify non-linear relationships between gene expression and Maximal Lifespan (MLS), we employed an interpretable machine learning framework using LightGBM and SHapley Additive exPlanations (SHAP)(61),(37). A baseline model using species life-history traits (body mass, metabolic rate, temperature, gestation period, and mitochondrial GC content) achieved high accuracy (R2 = 0.96). Mitochondrial DNA GC content and gestation period were the most influential predictors, consistent with previous findings(62),(56). Next, a two-stage backward feature selection with LightGBM-SHAP assessed individual gene influence on MLS. This approach refined an initial set to 57 genes, substantially improving predictive performance (Stage II: R2 increased from 0.90 to 0.95; MAE decreased from 4.73 to 3.04). Among the 57 genes, a non-significant but consistent overlap (17 genes) was noted with documented longevity-associated genes (LAGs) in the GenAge database(54), including GNAS and TERT. Prioritizing genes by mean absolute SHAP values, we found 57 genes with significant predictive contributions (> 0.1 years). The SHAP summary plot (Fig. 3) illustrates their relative influence.

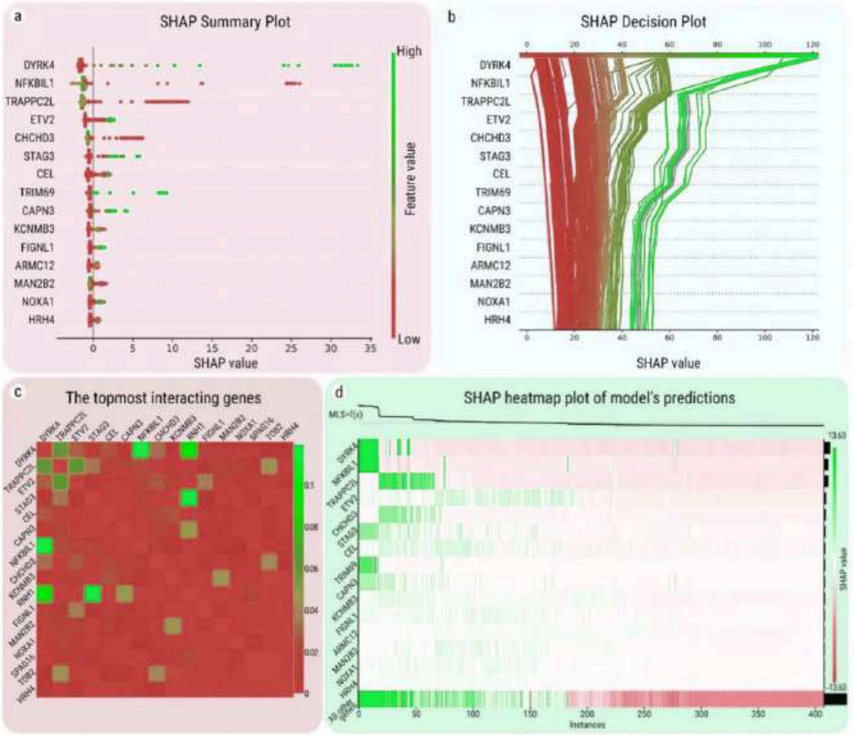


Figure 3. SHAP Explanations of Gene Expression and MLS Predictions
(a) SHAP Summary Plot. (b) SHAP Decision Plot for Individual Predictions. (c) SHAP Interaction Heatmap for Gene Pairs. (d) SHAP-based Heatmap of Gene Contributions per Sample.Figure adapted from Figure 3 Kulaga, Ursu, Toren. et al. (2021), used under license CC BY 4.0.

Five of the top 15 genes (DYRK4, NFKBIL1, TRAPPC2L, ETV2, and CHCHD3) substantially influenced MLS predictions, each shifting predictions by over one year. The association of TRAPPC2L and ETV2 with aging is undocumented, suggesting they are promising new candidates for longevity research. Quantifying the direction and strength of association using Kendall's tau-b between gene expression and SHAP contributions identified strongly pro-longevity genes (tau ≥ 0.6) like NEIL1, CALCOCO2, and LRR1, and strongly anti-longevity genes (tau ≤ −0.6), including C6orf89 and PPP1CA. CALCOCO2 (tau = 0.72) consistently emerged as a strong positive predictor of MLS across multiple organs (e.g., lung $R^2 = 0.56$, brain $R^2 = 0.54$, liver $R^2 = 0.53$, heart $R^2 = 0.39$). Conversely, the strongest anti-longevity gene, C6orf89 (tau = −0.79), was negatively predictive in the heart ($R^2 = 0.61$) and liver ($R^2 = 0.40$).

Genes influencing mammalian lifespan (MLS) often interact complexly, leading to combined effects different from individual impacts. We analyzed these using SHAP interaction values, which quantify cooperative or antagonistic relationships. Fig. 3c summarizes interaction strengths among top MLS-associated genes.

We then applied Bayesian network modeling to explore potential causal associations between gene expression and MLS, identifying robust relationships independent of redundancy or indirect correlations(63). Using the Markov blanket concept and the SES algorithm(53), we conducted 50 iterations to generate potentially causal gene signatures (Supp. Fig. S3 from Kulaga, Ursu, Toren et al.(52)). Highly frequent genes, indicating robustness, included NOXA1 (1.00), C6orf89 (0.94), NEU2 (0.94), NDUFA6 (0.90), RBM46 (0.82), KCNMB3 (0.72), and CEL (0.60). Biologically, these results corroborate LightGBM-SHAP findings (e.g., NOXA1, C6orf89, CEL). Functional enrichment strongly implicated mitochondrial mechanisms as central to longevity.

We found substantial consistency between linear and LightGBM-SHAP models. Bayesian network integration reinforced NOXA1, C6orf89, and CEL, pointing to mitochondria, DNA repair, and metabolic regulation as key longevity processes. Integrating Bayesian and LightGBM-SHAP results, we found robust Bayesian genes (NOXA1, C6orf89, NEU2, NDUFA6, RBM46, KCNMB3, CEL) did not all have the highest SHAP values. However, CEL and KCNMB3 ranked among the top 10 most impactful genes by SHAP. Two genes, NOXA1 and KCNMB3, were robustly associated with MLS across all three methods. These reliable, cross-validated genes are strong candidates for future functional longevity investigation.

Composite Ranking and Core Gene Signature Determination. We established a composite ranking by aggregating performance metrics from all three models. This rank was used with Bayesian multilevel linear models to identify an optimal, concise core gene signature. Evaluating subsets (top 3 to top 13 genes) based on penalized deviance, the top 11 ranked genes were selected as the core signature, balancing model simplicity and prediction accuracy. This final set (detailed in Table 1 in the full thesis and in the publication) provides robust candidate genes for mammalian longevity research.

4. Conclusions

This study analyzed gene expression across 41 mammalian species and five organs, using linear, non-linear, and Bayesian network models, to identify genetic determinants of Maximum Lifespan (MLS). Over 1800 genes showed significant, often organ-specific, MLS correlations. Pathway analysis confirmed established longevity pathways (e.g., DNA repair) and suggested novel ones (e.g., PPAR signaling). Interpretable machine learning (LightGBM-SHAP) and Bayesian networks identified additional candidate genes, with limited overlap with experimental longevity-associated genes (LAGs), reflecting differences between natural variation and genetic intervention. Notably, CEL, NOXA1, CALCOCO2, and KCNMB3 were consistently identified as robust longevity candidates. The research highlights the value of integrating complementary modeling to dissect mammalian lifespan, offering promising candidates and a solid methodological framework.

III. Aging signature in lung fibrosis

*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).*

Statement of Contributions

As second author, my main contribution was a cross-species linear modeling analysis, correlating pro- and anti-fibrotic gene expression with maximum lifespan (MLS) in mammalian organs. I also interpreted results and assisted with manuscript preparation. The project was conceived and coordinated by Prof. Vadim E. Fraifeld and Dr. Robi Tăcutu's teams.

Results and Discussion

This study identified 216 unique Pulmonary Fibrosis-Related Genes (PFRGs) using a bleomycin-induced lung fibrosis mouse model. Genetic interventions were the primary method. Approximately 43.5% of these genes showed anti-fibrotic activity, 50% were pro-fibrotic, and 6.5% had inconsistent outcomes.

Associations between PFRGs and Longevity. Cross-referencing PFRGs with Longevity-Associated Genes (LAGs) from the GenAge database(58,64) revealed a strong directional link: 11 out of 12 pro-longevity genes were anti-fibrotic, and 5 out of 6 anti-longevity genes were pro-fibrotic (Fisher's exact test, p = 0.001).

Further analysis of cross-species lung expression data showed that 34 PFRGs significantly correlated with mammalian Maximum Lifespan (MLS), a frequency 2.34 times higher than expected (Fisher's exact test, p = 6.4E−05). This supports the hypothesis that fibrosis-related genes play a conserved role in longevity regulation, underscoring fibrosis as a critical driver of age-related pathology.

Conclusions

Pro-longevity genes (LAGs) are generally anti-fibrotic, and anti-longevity genes are largely pro-fibrotic, revealing a strong, shared genetic link between pulmonary fibrosis and aging. Functional analysis reinforces this, with anti-fibrotic clusters rich in LAGs and pro-fibrotic clusters primarily anti-LAGs.

This aligns with aging being tied to pathways governing tissue repair, inflammation, and metabolism, which are critical to both conditions. Current evidence suggests that fibrosis and aging arise from shared genetic/molecular mechanisms, though causality requires more study.

IV. Intercellular communication is disrupted with aging

Lagger, C., **<u>*Ursu, E.</u>, 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).*

** denotes first co-authorship*

Statement of Contributions

As co–first author of this study, I contributed equally to the design, implementation, and interpretation of the project alongside Dr. Cyril Lagger. Specifically, I co-developed the scDiffCom, scAgeCom, and scAgeComShiny tools; curated and analyzed the ligand–receptor interaction database; and performed analyses to support key findings. I was also actively involved in interpreting the results and co-writing the manuscript. The core methodological framework and analytic pipeline were developed collaboratively between myself and C.L., and represent the primary focus of this chapter. The study was jointly supervised by Dr. Robi Tăcutu and Prof. João Pedro de Magalhães.

Overview

Intercellular communication (ICC) dysregulation is a fundamental hallmark of aging, contributing to various physiological and pathological processes. To investigate these changes systematically, we introduce scDiffCom and scAgeCom, two complementary tools for analyzing age-related alterations in cell–cell communication.

scDiffCom is an R package for differential ICC analysis using single-cell transcriptomics data, leveraging a curated database of approximately 5,000 ligand–receptor interactions (LRIs) to compare communication networks across different conditions.

Built upon scDiffCom, scAgeCom is a comprehensive atlas of age-related ICC changes, integrating data from 23 mouse tissues and 58 scRNA-seq datasets from Tabula Muris Senis and the Calico Murine Aging Cell Atlas. This resource reveals systemic age-related shifts in intercellular signaling, including:

- Increased immune activity and inflammation
- Reduced developmental signaling
- Impaired angiogenesis and extracellular matrix remodeling
- Dysregulated lipid metabolism

scAgeCom identifies specific ligands, receptors, and cell types driving these processes and is publicly available at https://scagecom.org.

1. Introduction

Aging is a complex biological process marked by significant dysregulation of ICC, recognized as a hallmark of aging(1,65). Existing research highlights several ICC alterations with aging, such as inflammaging(66), impaired immune surveillance(67), and increased SASP secretion(68). While ICC is challenging to measure directly, advancements in single-cell gene expression analysis allow for its inference(69,70). Existing single-cell aging studies often focus on detecting ICC networks separately in young and old samples, which overlooks shifts in interaction strength and lacks a statistical framework for quantifying changes. To address these limitations, we developed scDiffCom, an R package for differential ICC analysis. We applied scDiffCom to aging scRNA-seq datasets from Tabula Muris Senis(71) and the Calico Murine Aging Cell Atlas(72) to create scAgeCom, a large-scale atlas mapping age-associated ICC changes across 23 mouse tissues. This analysis confirms systemic dysregulation, with a global upregulation of immune system activity and inflammation, and a decline in processes like extracellular matrix organization and tissue growth.

2. Methods

Retrieving and Processing Ligand–Receptor Interactions (LRIs) and annotation with GO terms, KEGG pathways, and aging resources. We compiled curated, high-quality ligand–receptor interaction (LRI) datasets from seven public databases (e.g., CellChat, NicheNet, CellPhoneDB), excluding computationally predicted interactions. LRIs were annotated with Gene Ontology (GO) terms using a custom method based on graph-based intersection of ligand and receptor terms from Ensembl. KEGG pathways were assigned only if both the ligand and receptor were in the same pathway. To link LRIs to aging, we integrated data from aging-related databases (GenAge, LongevityMap, CellAge, HAGR) and quantified PubMed articles referencing each LRI gene (or its human homolog) in the context of aging.

CCI Scoring (Geometric Mean), Detection, Differential Analysis, and Classification. scDiffCom calculates a cell-cell interaction (CCI) score as the geometric mean of ligand and receptor expression to reflect the multiplicative nature of interactions, making its log fold change (logFC) the arithmetic mean of the gene logFCs. The method uses three simultaneous permutation tests to assess CCI specificity in two conditions (A, B) and differential expression between them, testing only expressed CCIs. Null distributions are generated by shuffling labels, and one-sided specificity and two-sided differential expression p-values are computed and Benjamini-Hochberg adjusted. A CCI is "detected" if expressed, specific (adj. $p < 0.05$), and top-ranked (80%). It is "differentially expressed" if its FDR-adjusted differential p-value is < 0.05. Detected CCIs are classified (UP, DOWN, FLAT, NSC) based on adjusted DE and logFC, prioritizing the differential test. Benchmarking against standard differential gene expression highlights the necessity of integrating CCI-level and gene-level results for reliable interaction capture and classification. Over-Representation Analysis (ORA) is used to detect over-represented features (e.g., GO terms, pathways) in differentially regulated Cell-Cell Interactions (CCIs).

scAgeComShiny Development. The interactive Shiny web application, scAgeComShiny, was built with golem (73). It uses plotly (74) for scatter plots and rrvgo (75)/REVIGO(76) for GO term visualization and redundancy reduction. The application is containerized with Docker and deployed via ShinyProxy.

Statistics and Reproducibility. Sample size was limited by public availability. No randomization or blinding was possible. Non-parametric tests (DE and ORA) were used, with Benjamini–Hochberg correction consistently applied for multiple testing.

3. Results and Discussion

LRIs from existing databases

To analyze intercellular communication (ICC) from scRNA-seq data, our approach first required an extensive LRI database. To maximize coverage of different interaction types, we compiled LRIs from seven publicly available resources, including CellChat (79), CellPhoneDB (80), CellTalkDB (81), NATMI/connectomeDB2020 (82), ICELLNET (83), NicheNet (84), and SingleCellSignalR (85). Our approach resulted in the generation of two curated ligand–receptor interaction (LRI) databases. The mouse LRI database contains 4,582 interactions, of which 3,479 are simple and 1,103 are complex. Similarly, the human LRI database includes 4,785 interactions, with 3,648 simple and 1,137 complex interactions. Detailed data can be found in Fig. 6a and Supp. Tables 1 and 2 from Lagger, Ursu et al. (86).

Functional annotation of LRIs. We implemented a standardized framework to annotate all ligand–receptor interactions (LRIs), ensuring biological relevance and facilitating downstream analyses. To enrich LRIs with biologically meaningful annotations, we assigned Gene Ontology

(GO) terms (87) in a meaningful way to interactions, prioritizing terms relevant to the interaction itself rather than the individual genes. LRIs were linked to KEGG pathways (88) only if all genes in a given interaction were present in the same pathway. In anticipation of aging-related analyses, we mapped mouse LRI genes to PubMed articles linking them to aging and age-related diseases (excluding cancers). Additionally, we cross-referenced genes with key aging-related databases, including GenAge (54), CellAge (89,90), LongevityMap (91), Gene Expression (92) database of the Human Ageing Genomic Resources (HAGR) (54).

Differential cell–cell communication analysis with scDiffCom

The R package scDiffCom detects significant changes in cell-type communication (CCI) between two conditions in scRNA-seq datasets (Fig. 4), working with R Seurat objects containing cell-type and condition labels(93–95). It assigns a CCI score based on average ligand and receptor expression, robust to total cell number bias(96). CCIs are validated based on three criteria: sufficient expression, specificity via a CellPhoneDB-like permutation test(80,97), and a high relative score. CCIs are then classified as upregulated (UP), downregulated (DOWN), stable (FLAT), or non-significant change (NSC). scDiffCom employs Over-Representation Analysis (ORA) to analyze thousands of classified Cell-Cell Interactions (CCIs)—Upregulated (UP), Downregulated (DOWN), or Stable (FLAT)—against a background of the rest. This approach avoids traditional gene-based enrichment bias.

Aging dysregulates several aspects of cell–cell communication

The scAgeCom atlas, constructed with scDiffCom across 58 murine scRNA-seq datasets from TMS(71) and Calico(72), details age-related intercellular communication (ICC) changes across 23 organs, addressing sex dimorphism (Fig. 5 and Supp. Text 1, Supp. Fig. 1 from Lagger, Ursu et al.(86)). This resource is online (https://scagecom.org/) (Fig. 6). Of 393,035 detected cell-cell interactions (CCIs), 18% were differentially regulated with age (5% upregulated, 13% downregulated). 1,135 ligand-receptor interactions were filtered out to minimize false discoveries. Benchmarking confirmed scDiffCom's CCI score is superior to gene-level comparisons (Ext. Data Fig. 2 from Lagger, Ursu et al.(86)), its ORA method avoids GO term bias (Supp. Figs. 2 and 3 from Lagger, Ursu et al.(86)), and LRI databases are crucial (Supp. Fig. 4 from Lagger, Ursu et al.(86)). These results establish scDiffCom as a robust tool for detecting biologically relevant, protein-mediated CCIs from scRNA-seq data, making the scAgeCom atlas a valuable resource for aging research.

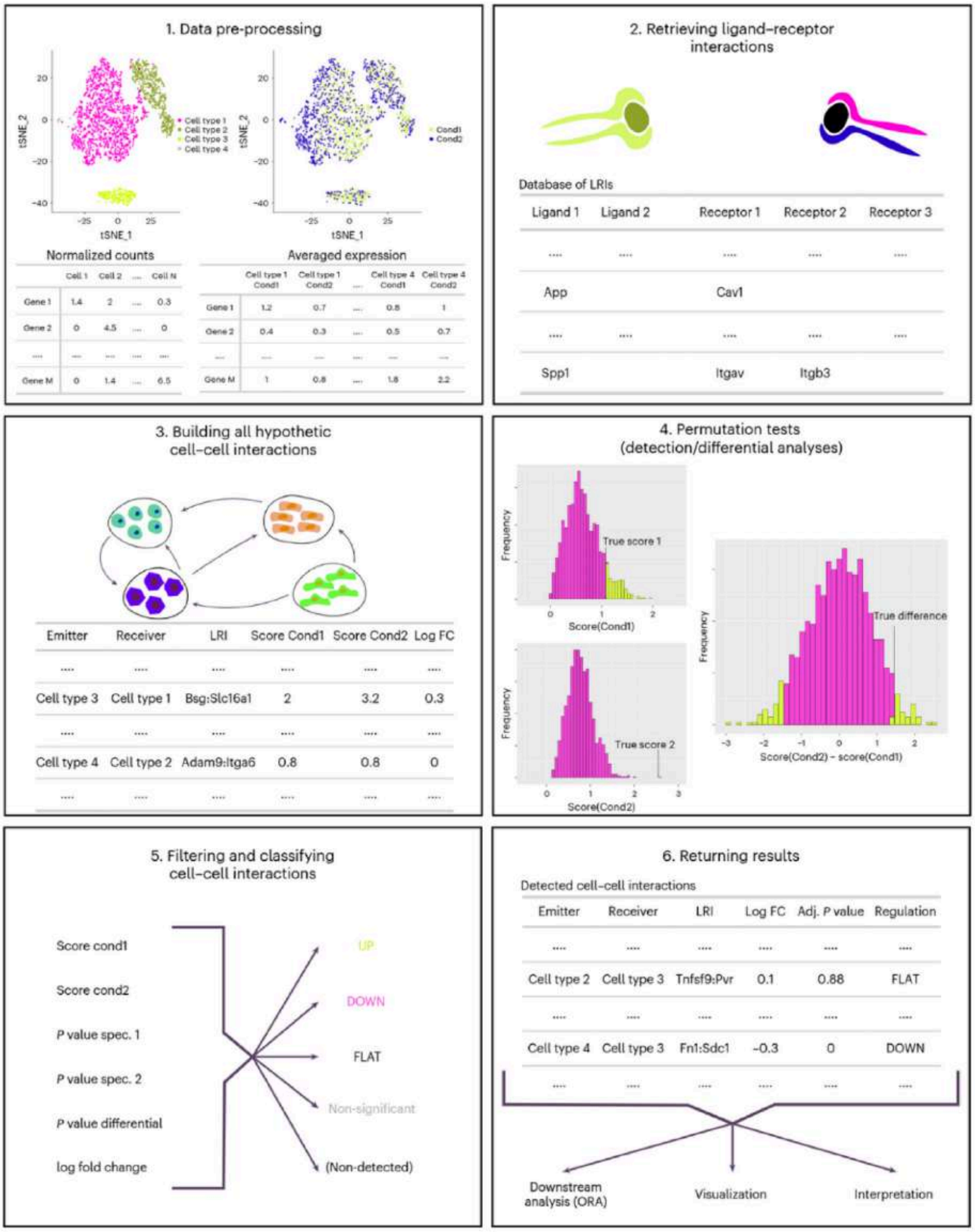
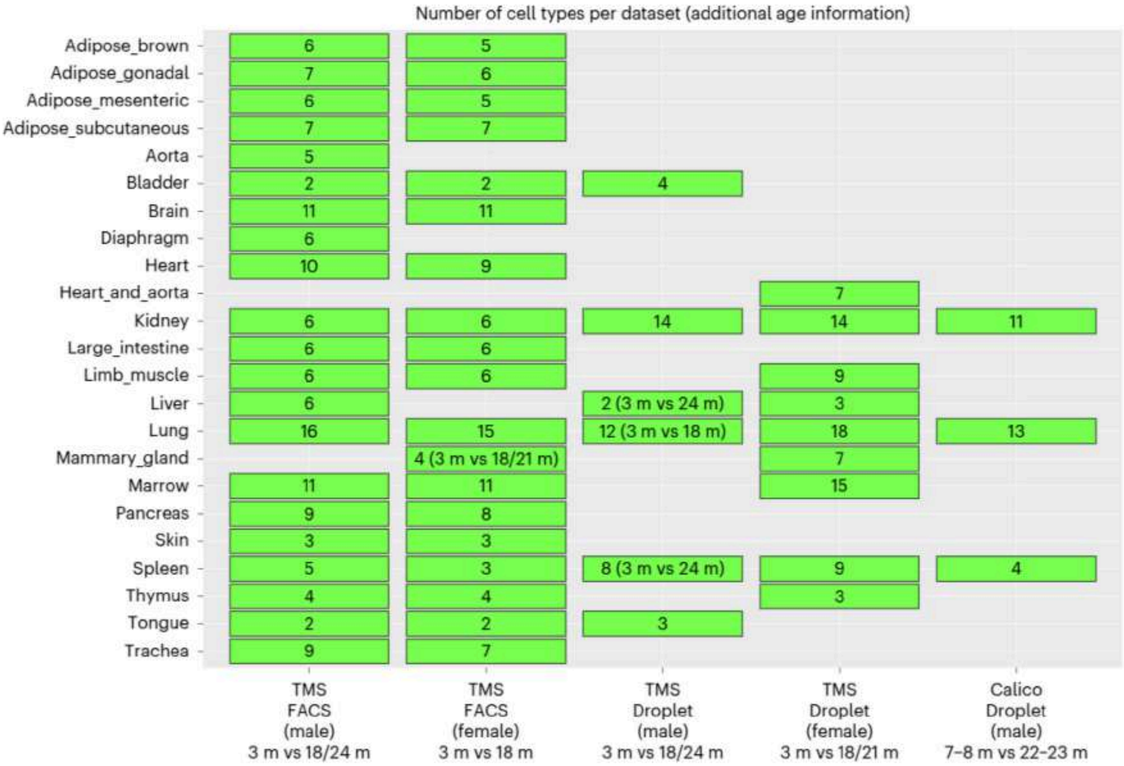


Figure 4. Workflow Summary of scDiffCom.

Single-cell RNA-seq read counts or unique molecular identifiers (UMIs) are first aggregated by cell types and experimental conditions (1). Gene expression data are then mapped onto the curated database of ligand–receptor interactions (LRIs) (2) to infer all possible cell–cell interactions (CCIs) between cell types (3). Statistical permutation tests are performed to assess the biological relevance of each CCI and to detect differential expression between conditions (4). CCIs are subsequently classified based on computed metrics, including interaction scores, P values, and log fold change (5). The results are compiled into a structured format suitable for downstream analysis and interpretation (6). FC, fold change; tSNE, t-distributed stochastic neighbor embedding.

Figure adapted from Figure 2 from Lagger, Ursu. et al. (2023), used under license CC BY 4.0



Number of cell types per dataset (additional age information)

	TMS FACS (male) 3 m vs 18/24 m	TMS FACS (female) 3 m vs 18 m	TMS Droplet (male) 3 m vs 18/24 m	TMS Droplet (female) 3 m vs 18/21 m	Calico Droplet (male) 7–8 m vs 22–23 m
Adipose_brown	6	5			
Adipose_gonadal	7	6			
Adipose_mesenteric	6	5			
Adipose_subcutaneous	7	7			
Aorta	5				
Bladder	2	2	4		
Brain	11	11			
Diaphragm	6				
Heart	10	9			
Heart_and_aorta				7	
Kidney	6	6	14	14	11
Large_intestine	6	6			
Limb_muscle	6	6		9	
Liver	6		2 (3 m vs 24 m)	3	
Lung	16	15	12 (3 m vs 18 m)	18	13
Mammary_gland		4 (3 m vs 18/21 m)		7	
Marrow	11	11		15	
Pancreas	9	8			
Skin	3	3			
Spleen	5	3	8 (3 m vs 24 m)	9	4
Thymus	4	4		3	
Tongue	2	2	3		
Trachea	9	7			

Figure 5. The 58 scRNA-seq Aging Datasets Used to Build scAgeCom.

Figure adapted from Figure 3 from Lagger, Ursu. et al. (2023), used under license CC BY 4.0

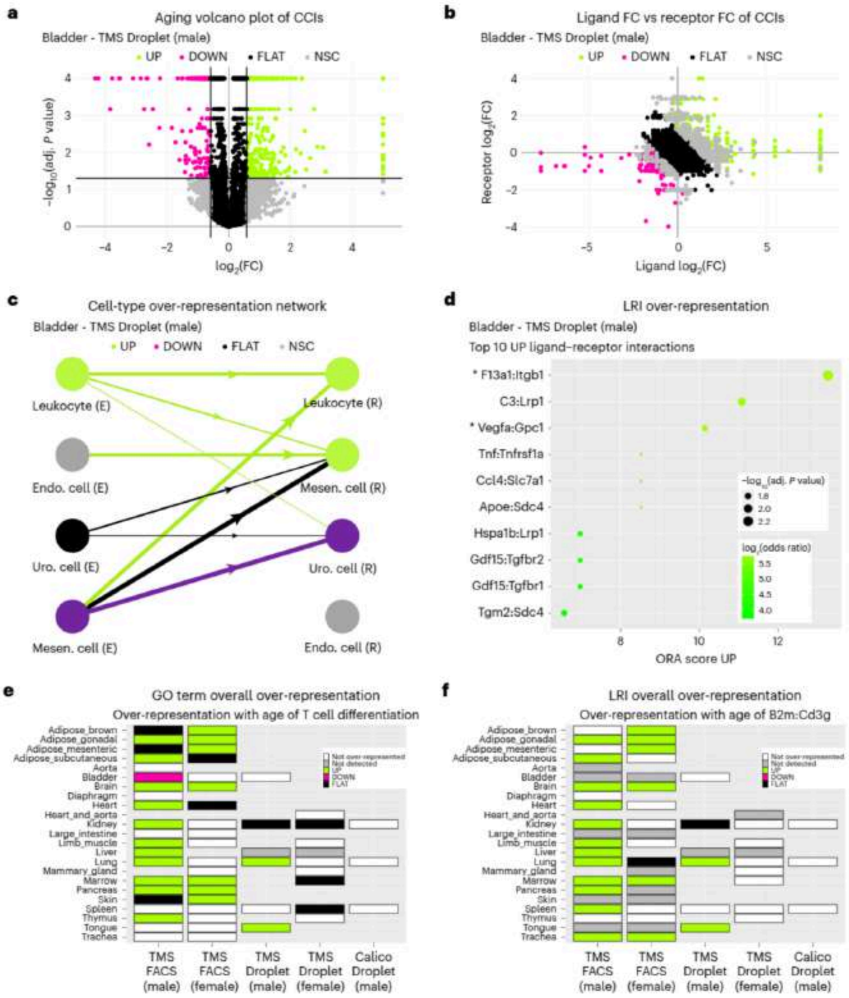


Figure 6. Overview of Visual Outputs in scAgeCom.
scAgeCom offers a variety of graphical outputs, accessible at https://scagecom.org/, to facilitate exploration of aging-related changes in intercellular communication.
Figure adapted from Figure 4 from Lagger, Ursu. et al. (2023), used under license CC BY 4.0

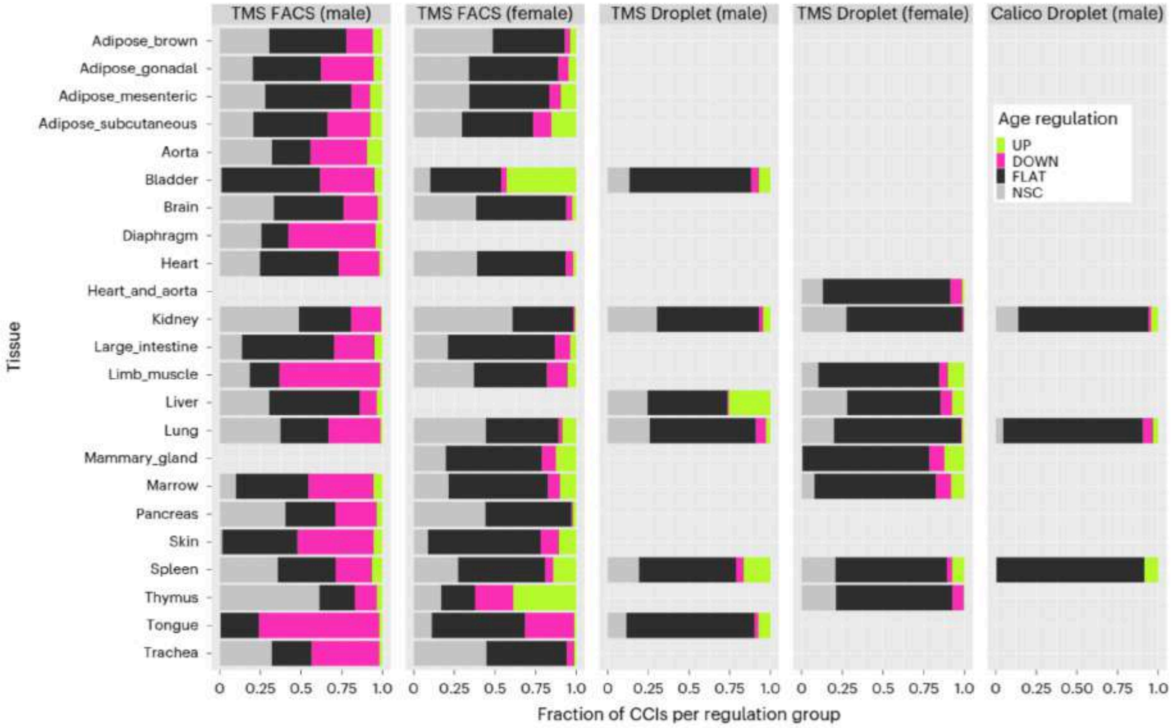


Figure 7. Variation in CCI age regulation patterns across scAgeCom datasets.
The proportions of CCIs classified as upregulated (UP), downregulated (DOWN), stable (FLAT), or non-significantly changing (NSC) with age differ across datasets, depending on tissue type, sex, and experimental methodology. Overall, TMS FACS datasets exhibit greater noise, reflected by a higher proportion of NSC-classified CCIs, compared to Droplet-based datasets. Additionally, TMS FACS (male) datasets display a notably higher fraction of age-downregulated CCIs relative to other conditions.
Figure adapted from Figure 6 from Lagger, Ursu. et al. (2023), used under license CC BY 4.0

The regulation of cell-cell interactions (CCIs) with age is highly variable, with TMS FACS (male) datasets showing many downregulated CCIs. FACS datasets are noisier than Droplet datasets. We prioritized changes across multiple tissues, involving novel aging genes, in secretomics, or sex-dependent.

A systemic upregulation of inflammatory, immune, and viral processes was confirmed, including B2m:Cd3g, Tnfsf12:Tnfrsf12a, and Ccl5 interactions (Ext. Data Fig. 3b from Lagger, Ursu et al.(86)). B2M is present in five secretomes. Slpi:Plscr1, despite upregulation in eight tissues, is largely unstudied in aging.

Lipid metabolism is dysregulated (Ext. Data Fig. 4 from Lagger, Ursu et al.(86)) with sex-specific patterns: Apoe-related CCIs are upregulated in males but downregulated in females, and vice versa for App-related CCIs. These AD-related proteins likely have systemic aging roles, supported by secretome detection.

A striking observation is the downregulation of extracellular matrix (ECM) organization and cell adhesion (Ext. Data Fig. 5 from Lagger, Ursu et al.(86)). This decline, driven by collagens, cadherins, and metallopeptidases with integrins, is strongest in connective tissue, epithelial, and endothelial cells.

Growth, development, survival, differentiation, and angiogenesis also decline (Ext. Data Fig. 6 from Lagger, Ursu et al.(86)), suggesting impaired regeneration. Decreased communication among stem cells and to endothelial cells links aging to reduced regenerative capacity (Ext. Data Fig. 5c from Lagger, Ursu et al.(86)).
scAgeCom reveals sex-dimorphic patterns (Fig. 7); e.g., in the TMS FACS Lung dataset, 13% of CCIs (including App, Pecam1, and Itgb1) showed stronger expression in young males, decreased with age in males, but increased in females, highlighting the need for personalized approaches.

4. Conclusions

Despite significant limitations, scAgeCom offers an extensive atlas of ICC aging in mice, providing novel insights into tissue-specific and sex-specific communication changes. Key contributions include:

- Comprehensive ICC mapping across 23 tissues
- Identification of potential therapeutic targets (e.g., B2m, Mif, Angpt1, Apoe)
- New hypotheses on aging mechanisms, including lipid metabolism shifts and mechanisms vascular decline
- Potential for integration with senescence and proteomics datasets for further validation

Moving forward, cross-analysis with other aging atlases, such as the SASP atlas, will be instrumental in refining our understanding of how ICC influences the aging process.

V. Negative training data composition is critical for learning immune repertoires

<u>Ursu, E.</u>, **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*
** denotes first co-authorship*

Statement of Contributions
As co–first author of this study, I contributed equally with Aygul Minnegalieva to the design, execution, and writing of the work. Together with Aygul Minnegalieva and Prof. Victor Greiff, I designed the analyses and visualizations central to the study. I also performed key analyses and figure generation in collaboration with Aygul Minnegalieva and Dr. Philippe A. Robert. In addition, I co-authored the first draft of the manuscript alongside Aygul Minnegalieva and Prof. Victor Greiff, and participated in revising the final version. The collaborative nature of this study is reflected in our equal contributions. This chapter focuses on the work we jointly developed and interpreted the results within the broader context of immunological modeling under the guidance of Prof. Victor Greiff.

1. Introduction

Supervised machine learning (ML) models rely critically on the composition of labeled datasets, particularly the definition of negative samples (representing the absence of the target class) in binary classification(98–101,102,103). This factor is understudied concerning its influence on model generalization and biological rule extraction. The context of antibody–antigen binding prediction, with its varied negative data construction strategies, provides a framework to study these effects. Prior work shows negative example choice impacts predictive accuracy and generalization in antibody–antigen and TCR–antigen interaction models(104–106,107–111); however, the impact on interpretability (binding mechanisms inferred) is unexamined. To investigate, we used "Absolut!"(106) to generate synthetic antibody–antigen data with various negative dataset strategies (Fig. 8), focusing on CDRH3 regions. Simple neural networks and DeepLIFT(112–114) were used to assess learned biological rules (Fig. 8). Models trained with negative data more similar to the positive class generalized better to unseen data and revealed different binding rules compared to those trained with dissimilar negatives. These trends held on experimental data (Fig. 8,9)(115). The design and selection of negative examples is a critical, often overlooked, component for building robust and interpretable ML models in biological applications.

This final chapter offers a methodological contribution to machine learning models of immune receptor repertoires, addressing challenges in training data quality and antigen specificity assignment. It improves on current AIRR-Seq modeling limitations to increase interpretability and biological relevance, establishing computational groundwork for future studies on immune aging and adaptive immune repertoire changes.

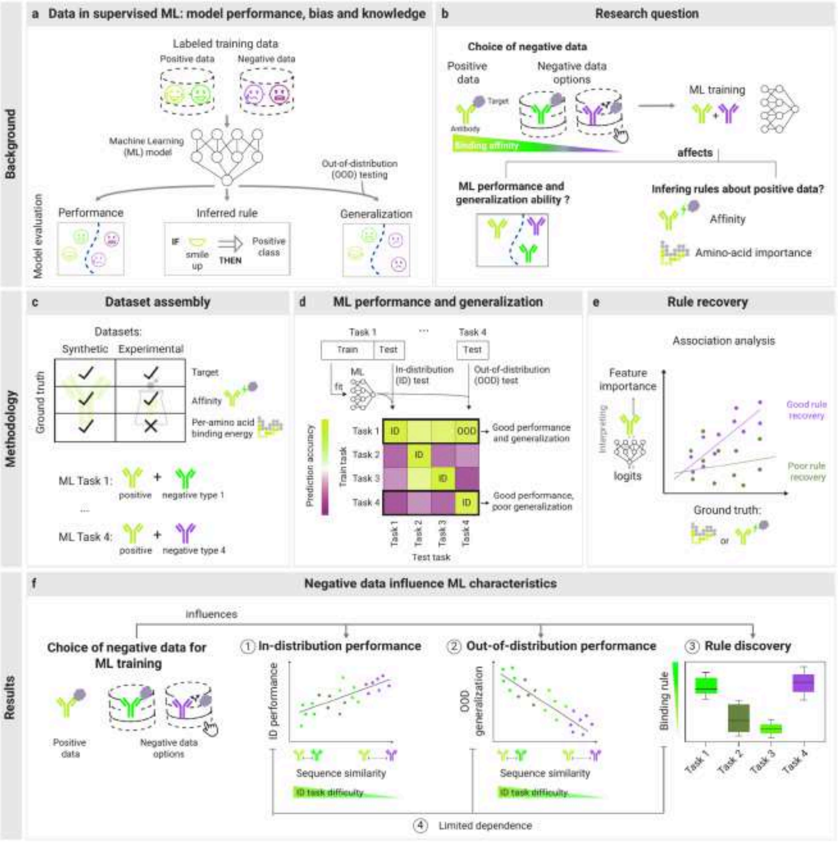


Figure 8. Training dataset composition shapes ML generalization and rule discovery.
Figure adapted from Figure 1 from Ursu, E. et al. (2025), used under license CC BY 4.0

2. Methods

A semi-synthetic antibody–antigen binding dataset was generated using the Absolut! simulation framework (Robert et al., Akbar et al.(106,116)). This dataset combines authentic mouse CDRH3 sequences(117) with 3D-lattice-formatted PDB antigen structures. CDRH3s were docked onto rigid antigens to calculate binding energies using the Miyazawa–Jernigan statistical potential in a 3D-lattice model (Absolut! framework, Robert et al.(106)). Sequences were labeled by binding affinity percentile: high-affinity (top 1%), weak (1–5%), and non-binders (remaining 95%). Data from ten antigens were used.

For prediction tasks, balanced datasets (30k train, 10k test) were created per antigen. Positive samples (top 1% binders) were compared against four negative class definitions: Vs 1 (different single antigen), Vs 9 (pooled nine other antigens), Vs Weak (weak binders, 1–5%), and Vs Non-binder (>5%). Reproducibility was ensured with six train–test splits and four random seeds per split.

The core Machine Learning model was SN10(106), a shallow feedforward neural network trained on one-hot encoded CDRH3s (220-neuron input, 10-unit ReLU hidden layer, sigmoid output). This was benchmarked against a deeper Transformer and PLM-based SN10 variants using pre-trained ESM2b (1280 dim.) and AntiBERTa2 (1024 dim.) embeddings(118,119,120), generated via EmbedAIRR. Interpretability was explored using DeepLIFT(114) to quantify amino acid contribution. Additionally, Epitope-Specific Analysis involved constructing datasets restricted to sequences binding to the single dominant epitope (e.g., 1H0DE1).

3. Results and Discussion

1. Training dataset sequence composition influences prediction performance in ID and OOD binary classification tasks

1.1 Machine learning setup on synthetic and experimental data

We began by investigating how different definitions of the negative class affect the performance and generalizability of supervised machine learning models for antibody–antigen binding prediction. Each task of the four tasks used an identical set of positive samples—high-affinity CDRH3 sequences—but differed in how the negative class was defined.

We used synthetic CDRH3 sequence data annotated with binding energies for ten antigens. For each antigen, the positive class consisted of sequences falling in the top 1% affinity percentile. We defined four types of negative classes: **vs Non-binder**: CDRH3 sequences from the lowest 95% binding energy percentile to the same antigen; **vs Weak**: Weak binders within the 1–5% percentile for the same antigen, forming a set disjoint from vs Non-binder; **vs 1**: High-affinity binders (top 1%) to a single, distinct antigen, excluding those that also bind to the positive-class antigen; **vs 9**: An expansion of vs 1, comprising high-affinity binders to each of the other nine antigens, equally represented.

We trained SN10 models. This architecture was chosen for its interpretability and previous benchmarking performance and showed comparable results to deeper models. To validate our conclusions derived from synthetic datasets, we replicated the same experimental setup using the HER2-targeting dataset published by Porebski et al. (115).

1.2 In-distribution (ID) Prediction Accuracy Depends on Training Dataset Composition

To evaluate the capacity of models to learn generalizable rules, we first measured in-distribution (ID) accuracy—performance on test data with the same positive and negative class definitions used during training (Fig. 9a). Across all four task types, models achieved high ID accuracy, with median values exceeding 0.85. A clear ranking emerged across tasks: models performed best on “vs Non-binder” (range: 0.97–1.00, median: 0.99), followed by “vs 1” (range: 0.94–1.00, median: 0.98), “vs 9” (range: 0.91–0.98, median: 0.94), and finally “vs Weak” (range: 0.85–0.98, median: 0.92), which proved most challenging (Fig. 9b). Moreover, the SN10 model outperformed logistic regression (LR), particularly in the “vs Weak” task. This suggests that SN10 can leverage inter-positional dependencies in the sequences that LR cannot.

1.3 Sequence Dissimilarity Explains Variability in ID Accuracy

To understand why ID performance differed across tasks, we examined the relationship between model accuracy and sequence dissimilarity between positive and negative datasets. Using position weight matrices (PWMs), we calculated Jensen–Shannon distances (JSD) to quantify distributional divergence. We found that JSD increased in the order: “vs Weak” < “vs 9” < “vs 1” < “vs Non-binder”, reflecting greater sequence divergence between the positive and negative classes (Fig. 9c). This gradient closely mirrored the ID accuracy trend, and JSD was significantly correlated with model performance for “vs 9” ($r = 0.94$), “vs 1” ($r = 0.73$), and “vs Non-binder” ($r = 0.77$) tasks (all $p < 0.05$), but not for “vs Weak” ($r = 0.30$, $p \geq 0.05$) or any of the shuffled controls (Supp. Table 1, Supp. Fig. 3a from Ursu, Minnegalieva et al. (121)).

1.4 Out-of-distribution (OOD) Accuracy Is Also Shaped by Dataset Composition

We assessed out-of-distribution (OOD) generalization (Fig. 9d). OOD accuracy was lower than in-distribution (ID) performance, as expected.

Models trained on "vs Non-binder" showed the largest drop (median ID 0.99 to OOD 0.72-0.82) (Supp. Fig. 3d from Ursu, Minnegalieva et al.(121)). "Vs 9"-trained models generalized best to "vs 1" (0.94), while "vs 1" models performed moderately (0.78 to "vs 9," 0.72 to "vs Non-binder"). The "vs 1" negative antigen choice affected results (Supp. Fig. 5 from Ursu, Minnegalieva et al.(121)).

"Vs Weak" was the hardest OOD test (accuracy 0.58–0.71). Surprisingly, "vs Weak"-trained models generalized best overall (0.90–0.96) (Fig. 9d).

High ID accuracy does not guarantee generalization. E.g., "vs Non-binder" yielded excellent ID but poor OOD performance, while "vs Weak" was the hardest ID task but the most robust in generalization. This highlights the critical role of negative class design for antibody–antigen binding ML models.

1.5 Experimental Data Confirm Findings from Synthetic Datasets

To validate these findings in real-world data, we analyzed an experimental antibody–antigen binding dataset recently published by Porebski et al. (115), which, like the synthetic "Absolut!" data, includes affinity-annotated CDRH3 sequences against HER2 (see Methods).

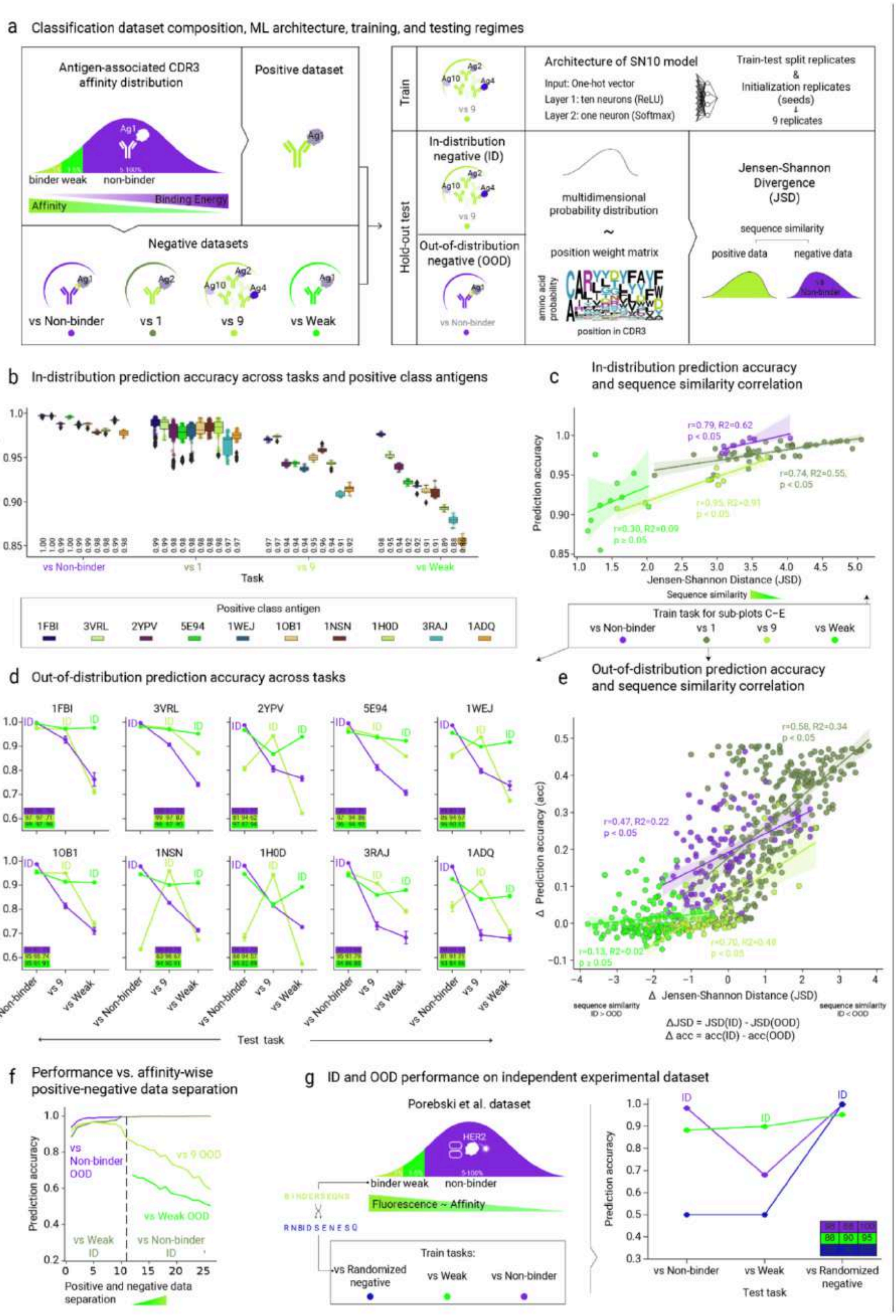


__Figure 9. Classification accuracy varies across antigens, binding prediction tasks, and positive-negative sequence similarity.__ Figure adapted from Figure 2 from Ursu, E. et al. (2025), used under license CC BY 4.0

SN10 models were trained on each task and evaluated on both ID and OOD test sets (Fig. 9f). The “vs Non-binder” model struggled on the “vs Weak” task (68% OOD accuracy), while the “vs

Weak” model generalized well (88% OOD accuracy). Just as in the synthetic experiments, “vs Weak” training led to tighter, more generalizable decision boundaries that balanced performance across tasks. By contrast, “vs Non-binder” training led to more permissive decision boundaries that offered high recall within ID contexts but resulted in false positives under OOD scenarios.

2. Training dataset composition determines the accuracy of biological rule recovery

Having established that the composition of negative training data influences prediction accuracy and generalization in supervised binary sequence-based machine learning (ML) tasks, we next investigated whether it also affects the model's ability to learn biologically meaningful binding rules.

2.1 Training Dataset Composition Impacts the Learning of Antibody Binding Energy

We first asked whether trained models captured the energy landscape of antibody sequences. Although some studies have explored simultaneous prediction of binding status and affinity (122,123), binary classification (binding vs. non-binding) is often more experimentally tractable. To assess whether our model, SN10, implicitly learns sequence-to-binding energy relationships, we computed the correlation between predicted output logits (i.e., raw pre-sigmoid activations) and ground-truth binding energies on a per-sequence basis. These logits reflect the model's confidence in predicting positive class membership (Fig. 10a). A representative 2D distribution for antigen 3VRL in the "vs Weak" task is shown in Fig. 10b (inset).

Overall, models trained on “vs 1” and “vs 9” datasets failed to learn per-sequence energy rules, with weak correlations of –0.05 and –0.19, respectively (Fig. 10b). Notable exceptions were antigens 3VRL and 5E94, where models learned binding energies more effectively regardless of negative dataset type. By contrast, models trained with “vs Weak” or “vs Non-binder” negatives consistently learned meaningful energy associations across all antigens, achieving median Pearson correlations of –0.62 (range: –0.33 to –0.90) and –0.51 (range: –0.29 to –0.85), respectively.

These results indicate that the capacity to learn energy-based rules is strongly dependent on the type of negative training data. This is supported by significant variance in correlation values across tasks (one-way ANOVA, p = 8.7e–58) and across antigens (one-way ANOVA, p = 4.6e–25).

To evaluate whether learning of binding energy rules affects model performance, we correlated logit–energy associations with both ID and OOD prediction accuracy. Significant associations were observed in ID settings for “vs Weak” (r = –0.77) and “vs Non-binder” (r = –0.83) tasks (Fig. 10c). For OOD tasks, accuracy was significantly associated with logit-based rule learning across all cases (Supp. Fig. 4b from Ursu, Minnegalieva et al. (121)).

2.2 Training Dataset Composition Impacts Learning of Position-Wise Contribution to Binding Energy

Having shown that SN10 can capture per-sequence energy rules, we next investigated whether it could also learn position-specific contributions of amino acid residues to binding—an essential determinant of immune recognition. Our hypothesis was that correctly learned rules would result in negative correlations between attribution values (computed via DeepLIFT (114)) and per-residue binding energy: residues with stronger binding (lower energy) should receive higher attribution in positive predictions.

Consistent with per-sequence results, the strongest per-residue rule learning occurred in models trained on “vs Weak” (median r = –0.69) and “vs Non-binder” (median r = –0.71) datasets. In contrast, models trained on “vs 9” and “vs 1” tasks showed weaker or even reversed associations (–0.36 and –0.01, respectively). Notably, for antigens such as 3VRL, 5E94, and 3RAJ, correlations remained high across all negative datasets—suggesting that in some cases, positive data alone may suffice to learn position-level rules.
Task type significantly influenced rule learning (one-way ANOVA, p = 3.5e–43). No significant correlation was found between ID prediction accuracy and per-residue rule learning (Fig. 10e), underscoring the importance of evaluating explainability metrics independently. However, OOD accuracy did correlate with attribution-energy agreement in most “vs 1” and “vs 9” settings (Supp. Fig. 4b from Ursu, Minnegalieva et al. (121)). In contrast, such associations were not observed for models trained on “vs Weak” or “vs Non-binder” datasets.

2.3 Investigating the Extent of Additivity in Learned Rules

Although “vs Weak” and “vs Non-binder” tasks yielded highly similar attribution profiles, their differing out-of-distribution (OOD) performance prompted us to explore whether these models relied more on additive or non-additive decision rules. Specifically, we aimed to understand whether the differences in generalization could be explained by the extent of feature interactions learned during training. To assess this, we trained logistic regression models. These models lack the capacity for feature interaction. As such, they serve as a benchmark for purely additive rule learning. The findings imply that “vs Non-binder” tasks can be solved primarily using additive features, while “vs Weak” demands more complex, interaction-driven representations. This aligns with the superior performance of SN10 over logistic regression in harder tasks like “vs Weak” (Supp. Fig. 3c from Ursu, Minnegalieva et al. (121)), supporting the notion that SN10 benefits from its ability to model feature interactions.

Discussion

The definition of positive and negative data fundamentally influences machine learning model behavior, yet remains an underexplored area in immunological ML applications. In this study, we demonstrated that both in-distribution (ID) and out-of-distribution (OOD) performance, as well as the interpretability of learned rules, are strongly dependent on the composition of the training dataset in the context of antibody-antigen binding prediction.
More broadly, our results underscore that training data curation is not merely a preparatory step but a central design consideration in developing robust and generalizable ML models. Careful planning and justification of dataset composition are critical for achieving valid predictive outcomes and biologically meaningful rule discovery, particularly when models are applied beyond the immediate scope of their training data.
Our work demonstrates that the definition of negative training data has a profound impact on ML model behavior, including prediction accuracy, generalization (ID vs. OOD), and the discovery of biologically meaningful binding rules in antibody-antigen interactions. Despite the importance of this topic, it remains underexplored in the antibody domain compared to its treatment in TCR-epitope prediction studies (107–111).

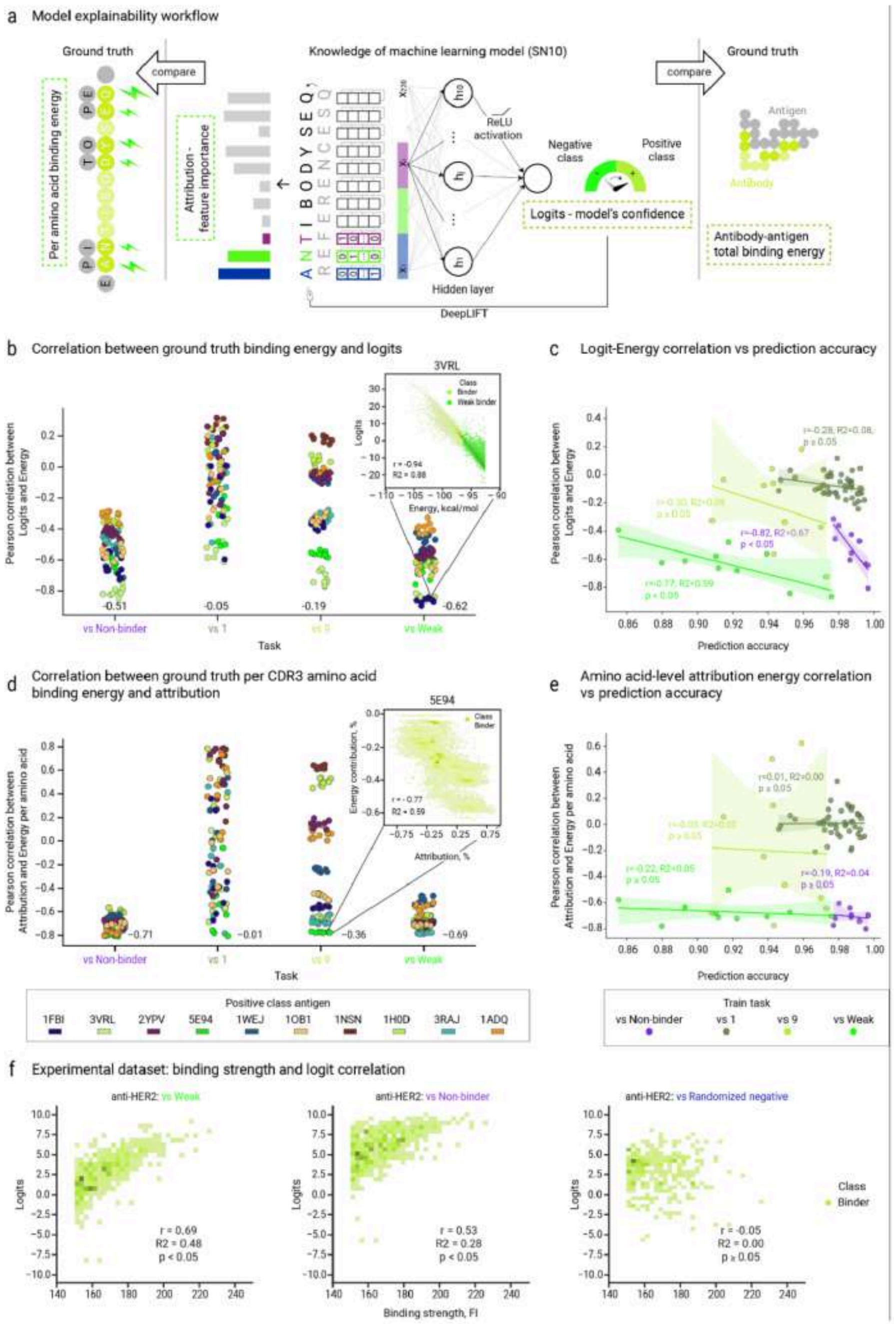


Figure 10. The composition of the negative dataset shapes the model's ability to learn binding rules for positive sequences.

Figure adapted from Figure 3 from Ursu, E. et al. (2025), used under license CC BY 4.0

Our findings emphasize that: Careful definition of negative datasets is critical for interpretable and generalizable ML models in immunology; OOD performance should be an explicit metric in evaluating rule discovery and model interpretability and current attribution methods have limitations, and novel techniques are needed to capture feature interactions, such as epistasis, that likely underlie complex biological phenomena (124–126).

In summary, our findings demonstrate that: Models trained on datasets with closely matched positives and negatives generalize better and learn more robust biological rules; Negative dataset design is not merely a technical step but a central determinant of model behavior and interpretability; Future work should systematically explore strategies for selecting hard negatives

(127), investigate near-miss sequences, and develop attribution methods that capture complex binding determinants.

VI. Final remarks

The work presented in this thesis was driven by a central question: how can we better understand the biological complexity of aging through the lenses of modern omics, computational biology, and machine learning? To address this, I integrated multi-dimensional datasets, developed novel computational tools, and collaborated across disciplines to explore key aging mechanisms from a systems-level perspective.

Across five chapters, I approached aging through distinct yet interconnected angles. In Chapter II, I explored transcriptomic signatures across 41 mammalian species, uncovering conserved and organ-specific gene expression patterns associated with lifespan. Chapter III expanded on the intersection between aging and fibrotic remodeling in the lungs, highlighting both conserved molecular patterns and unique aging signatures in a well-characterized fibrosis model. In Chapter IV, I focused on intercellular communication—a hallmark of aging that remains understudied—and co-developed scDiffCom and scAgeCom, providing a scalable generalizable framework to map and quantify age-associated signaling disruptions. Finally, in Chapter V, I turned to the adaptive immune system, combining machine learning with AIRR-Seq data to examine how model performance and biological insight are shaped by the composition of training datasets—a foundational contribution for future studies of immune repertoire aging.

Throughout this thesis, I have sought not only to generate results, but to contribute tools, frameworks, and methodologies that will empower others in the aging research community and in the life sciences. This includes interpretable machine learning strategies, cross-species analytical pipelines, and robust datasets made publicly available for broader reuse. My research reflects the belief that deep biological insight often emerges at the intersection of data richness, methodological rigor, and cross-disciplinary collaboration, but also fundamental biological understanding.

While many questions remain, I hope this thesis provides a meaningful step toward understanding the molecular determinants of aging. Future work may expand on the tools developed here, applying them to longitudinal human datasets, interventional studies, or therapeutic screens. The ultimate goal remains the same: to untangle the complexity of aging in ways that can inform strategies for healthier and longer lives.

Bibliography

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. Nikolich-Žugich 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* PackageMXM : 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, Grandclaudon 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. Lagger 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 ΔΔG 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;