Development of Innovative Computational Strategies for Molecular Activity Prediction

by

Vishakha Gautam

(PhD20202)

Under the Supervision of Dr. Gaurav Ahuja

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi - 110020

May, 2025

Development of Innovative Computational Strategies for Molecular Activity Prediction

by

Vishakha Gautam

A Thesis

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi - 110020

May, 2025

Certificate

This is to certify that the thesis titled "**Development of Innovative Computational Strategies for Molecular Activity Prediction**" being submitted by **Ms. Vishakha** to the **Indraprastha Institute of Information Technology Delhi,** for the award of the degree of **Doctor of Philosophy**, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May, 2025

Dr. Gaurav Ahuja
Associate Professor
Department of Computational Biology
IIIT Delhi, 110020

Acknowledgment

*First and foremost, I would like to express my gratitude to **GOD Almighty** for granting me the strength and resources to start and complete this Ph.D. journey.*

I am deeply thankful to my Ph.D. supervisor, **Dr. Gaurav Ahuja**, who accepted me as a student in 2020 and provided me with the opportunity to work in his lab. He has consistently offered his unwavering support, guidance, and inspiration. He has allowed me the freedom to conduct my research, engage in stimulating discussions, and always kept me focused on my goals. I am eternally grateful for his invaluable assistance and mentorship.

I also express my sincere thanks to **Prof. Gajendra P. S. Raghava (HOD CB)** and **Prof. Ranjan Bose (the IIIT-D Director)** for admitting me into the PhD program and granting me access to the institute's resources. I am also deeply grateful to **Prof. Debarka Sengupta** and **Dr. Jaspreet Kaur Dhanjal** for serving as my committee members and providing their guidance and expertise. Additionally, I would like to thank **Prof. Gajendra P. S. Raghava, Prof. Debarka Sengupta, Prof. Sriram K, Dr. Vibhor Kumar, Prof. Ganesh Bagler, and Dr. Sanat K Biswas** for their invaluable teachings during my coursework.

I would also like to acknowledge the administrative staff at IIIT-D, specifically **Mrs. Priti Patel, Mrs. Shipra Jain, Mrs. Anshu Dureja**, and **Mr. Raju Biswas,** for always being available to address our queries and resolve our academic issues promptly. My sincere gratitude also goes to **Mr. Imran Khan, Mrs. Sarika, Mr. Mohit**, and **Mr. Kapil Dev Garg** for the timely delivery of my stipend. I would also like to thank the IIITD IT staff, especially **Mr. Adarsh,** for always being there when I needed support. I am also grateful to IIIT-D for providing excellent facilities and infrastructure.

I would like to officially thank the **Department of Science & Technology (DST)** for providing me with the "**Innovation in Science Pursuit for Inspired Research (INSPIRE)**" research fellowship to support my doctoral studies.

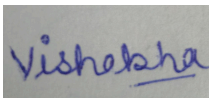
Additionally, I would like to extend my heartfelt thanks to **Mrs. Sangeeta, Mrs. Meenu, Dr. Rupinder Kaur, Dr. Tammanna R. Sahrawat, Dr. Veena Puri,** and **Dr. Ashok Kumar,** who taught me during my graduation and post-graduation, for providing me with the knowledge and direction I needed to embark on this incredible scientific journey.

My heartfelt thanks go to my lab-mates and seniors, both from my own and other labs, for fostering a supportive and enjoyable work environment. I am particularly grateful to my senior,s **Dr. Aayushi Mittal, Dr. Krishan Gupta, Dr. Sumeet Patiyal, Dr. Anjali Dhall, Dr. Neetesh Pandey, Dr. Neelam Sharma, and Dr. Ridam Pal,** for their constant encouragement and mentorship. I would also like to thank my lab mates **Sanjay Kumar Mohanty and Sakshi Arora** for always being there for me, as well as my juniors **Subhadeep, Saveena, Sonam, Arushi, Shiva, Suvendu, and Raidhani,** with whom I have shared countless joys and who have always been encouraging.

I would also like to express my gratitude to my friends **Biswa, Mohit, Mansi, Asra, Sakshi, Sanjay, Sukriti, Gayatri, Sadiyah, Shyam, Tharma, Sagnik, Atish, Madhu, Asrar, Akshit, Pradeep, Akshaya, Abhishek, Harsh, Yatish, Deepak, and Tejas** whom I met during my Ph.D. and who have shared my challenges and supported me during difficult times. I would also like to thank my colleagues **Nishant, Shubham, Manas, Alok and Ritu** who have always been there whenever I needed them. Additionally, I would like to thank all my old friends, especially **Pooja, Kiran, Ambir, Sherry, Amit, Swapnil, Vivek, Arun, and Jitender**, and all my cousins, especially **Yashu, Puju di, Vishu di, Kittu Massi, Honie Jiju, Himanshu Jiju, and Shanu Mosa ji**, who have been unwavering pillars of support in this journey.

I would also like to thank the IIIT-D administrative staff for providing hostel amenities. Special thanks to **Rajeev Ji and Malti Ma'am (hostel warden), as well as Shanti Ma'am,** for always being available and attending to tasks promptly. I am also grateful to the **mess vendor Trilok Uncle**, the canteen staff, especially **Akshay Bhaiya**, and the "Ravi Tea Stall" and its proprietor **Rajesh Ji** for providing me with nourishment and energy during my time away from home.

Finally, I would like to express my deepest gratitude to my beloved **grandparents** for encouraging and supporting my pursuit of higher education. I would also like to thank my wonderful parents, **Mrs. Champa Sharma** and **Mr. Mahipal Sharma,** for their unwavering support and inner strength during the most challenging phases of my life. I would also like to thank my brother **Satyam** for his love, care, and support throughout this journey. And to all the people I have met during this journey who have shaped my path and inspired me in countless ways.



Vishakha

Abstract

Abstract

Understanding and predicting complex biological processes, from the interactions of molecules to the aging of cells, relies heavily on the ability to extract useful information from different types of biological data. Biological features are the measurable characteristics extracted from this data and can exist in many forms. They are the molecule shapes displayed as graphs, the levels of gene expression in single cells, and the visible features seen in cell images: shape, texture, and motion. Using these different forms of features successfully is highly important to generating good models and learning more about biology. However, getting these various features in front of a computer and extracting the most relevant ones is very challenging and often requires special techniques specific to the type of data and biological questions. This thesis presents three computational tools to address these challenges by focusing on robust feature engineering across multiple biological scales and data types.

First, deepGraphh utilizes graph-based structural traits to predict molecular activity. It avoids the need for conventional, pre-computed descriptors by using a suite of Graph Neural Networks (GNNs), such as Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), Directed Acyclic Graph (DAG) networks, and Attentive FP, to learn representations straight from molecular graphs. This web service streamlines model development, parameter tuning, and validation, offering performance comparable to traditional descriptor-based approaches. Quantitative Structure-Activity Relationship (QSAR) modeling is made easier with deepGraphh, an open-source web service that is accessible and performs similarly to descriptor-based techniques. deepGraphh was used to predict the permeability of human and microbiome-generated metabolites across the blood-brain barrier.

Second, EcTracker analyzes single-cell RNA sequencing data to identify ectopically expressed genes and characterize cell types. This R/Shiny-based web server compares gene expression to physiological norms, enabling the identification of cell identities and regulatory networks through regulon analysis. By reanalyzing a CRISPRi dataset, EcTracker revealed previously ambiguous identities in SMAD2 knockout cells, highlighting its ability to uncover critical regulatory insights.

Finally, scCamAge leverages single-cell microscopy images and associated bioactivity measurements to predict cellular states, particularly aging. This multimodal deep learning engine, packaged in Docker, analyzes single-cell microscopy images (trained initially on approximately one million yeast cells) to predict cellular age and functional bioactivities by capturing complex spatiotemporal and morphological features. The conservation of visual aging indicators and their promise for high-throughput screening are highlighted by its validation utilizing genetic and chemical perturbations, as well as their exceptional

capacity to predict senescence in human fibroblasts based on yeast training data, highlighting the conserved morphometric features related to senescence in human cells.

Collectively, deepGraphh, EcTracker, and scCamAge provide a comprehensive suite of tools for robust feature engineering and predictive modeling across diverse biological data, facilitating a deeper understanding of complex biological processes.

Table of Contents

List of Abbreviations

1D	1 Dimensional
2D	2 Dimensional
3D	3 Ddimensional
AFP	Attentive FP
BBB	blood-brain barrier
CLS	chronological lifespan
CNN	convolutional neural network
CT	Computed Tomography
DAG	Directed Acyclic Graphs
DGE	differential gene expression
EHR	electronic health record
ET	Extra Tree Classifier
fastMNN	Fast mutual nearest neighbors correction
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph neural networks
GUI	graphical user interface
hESCs	human embryonic stem cells
HMDB	Human Metabolome Database
k-NN	k-Nearest Neighbors
KEGG	Kyoto Encyclopedia of Genes and Genomes
LASSO	Least Absolute Shrinkage and Selection Operator
MCC	Matthews correlation coefficient

MCS	Maximum Common Subgraph
MD	Molecular Dynamics
MLP	multilayer perceptron
MRI	Magnetic resonance imaging
mRNA	messenger RNA
NGS	next-generation sequencing
OCHEM	Online chemical modeling environment
PDD	phenotypic drug discovery
PET	positron emission tomography
QC	quality control
QSAR	Quantitative Structure-Activity Relationship
RF	Random Forest
RNA-Seq	RNA Sequencing
ROC	receiver operating characteristic
rRNA	ribosomal RNA
SBEA	Stouffer’s score-based enrichment analysis
scRNA-seq	Single-cell ribonucleic acid sequencing
SGD	Stochastic Gradient Descent
SQL	Structured Query Language
SVM	Support Vector Machines
SVR	Support Vector Regression
TDD	traditional drug discovery
tRNA	transfer RNA
WT	Wild-type

List of genes and their description

Gene	Description
SMAD2	SMAD family member 2
ID1	DNA-Binding Protein Inhibitor ID-1
ID3	Inhibitor Of DNA Binding 3
HLA-B	Major Histocompatibility Complex, Class I, B
HLA-C	Major Histocompatibility Complex, Class I, C
RARRES2	Retinoic Acid Receptor Responder 2
CCND1	Cyclin D1
TEAD1	TEA Domain Transcription Factor 1
MYBL2	MYB Proto-Oncogene Like 2
E2F4	E2F Transcription Factor 4
MYC	MYC Proto-Oncogene, BHLH Transcription Factor
CALD1	Caldesmon 1
pdx3	Pyridoxine (pyridoxamine) phosphate oxidase
pho89	Plasma membrane Na+/Pi cotransporter
prs3	5-phospho-ribosyl-1(alpha)-pyrophosphate synthetase
met2	L-homoserine-O-acetyltransferase
met17	O-acetyl homoserine-O-acetyl serine sulfhydrylase

ipt1	Inositolphosphotransferase
gre3	Aldose reductase; involved in methylglyoxal
cys4	Cystathionine beta-synthase
adh2	Glucose-repressible alcohol dehydrogenase II
kgd1	Subunit of the mitochondrial alpha-ketoglutarate dehydrogenase complex
sod2	Mitochondrial manganese superoxide dismutase
trx1	Cytoplasmic thioredoxin isoenzyme
gsy2	Glycogen synthase
alt2	Catalytically inactive alanine transaminase
msw1	Mitochondrial tryptophanyl-tRNA synthetase

List of Figures

List of Tables

List of Publications (link)

Thesis Related Publications (3)

- **Gautam V**, Duari S, Solanki S, Gupta M, Mittal A, Arora S, Aggarwal A, Sharma AK, Tyagi S, Pankajbhai RK, Sharma A, Chauhan S, Satija S, Kumar S, Mohanty SK, Tayal J, Dixit NK, Sengupta D, Mehta A, Ahuja G. scCamAge: A context-aware prediction engine for cellular age, aging-associated bioactivities, and morphometrics. **Cell Rep. 2025** Feb 6;44(2):115270. doi: 10.1016/j.celrep.2025.115270. Epub ahead of print. PMID: 39918957. ***(first author; co-corresponding author)***

- **Gautam V**, Gupta R, Gupta D, Ruhela A, Mittal A, Mohanty SK, Arora S, Gupta R, Saini C, Sengupta D, Murugan NA, Ahuja G. deepGraphh: AI-driven web service for graph-based quantitative structure-activity relationship analysis. **Brief Bioinform.** 2022 Sep 20;23(5):bbac288. doi: 10.1093/bib/bbac288. PMID: 35868454. ***(first author)***

- **Gautam V**, Mittal A, Kalra S, Mohanty SK, Gupta K, Rani K, Naidu S, Mishra T, Sengupta D, Ahuja G. EcTracker: Tracking and elucidating ectopic expression leveraging large-scale scRNA-seq studies. **Brief Bioinform. 2021** Nov 5;22(6):bbab237. doi: 10.1093/bib/bbab237. PMID: 34184038. ***(first author)***

Other Publications

- Duari S, **Gautam V**, Ahuja G. Protocol for cellular age prediction in yeast and human single cells using transfer learning. STAR Protoc. 2025 Aug 11;6(3):104023. doi: 10.1016/j.xpro.2025.104023. Epub ahead of print. PMID: 40794494; PMCID: PMC12357082. ***(co-first author)***

- Arora S, Mittal A, Duari S, Chauhan S, Dixit NK, Mohanty SK, Sharma A, Solanki S, Sharma AK, **Gautam V**, Gahlot PS, Satija S, Nanshi J, Kapoor N, Cb L, Sengupta D, Mehrotra P, Ghosh TS, Ahuja G. Discovering geroprotectors through the explainable artificial intelligence-based

platform AgeXtend. **Nat Aging.** 2024 Dec 3. doi: 10.1038/s43587-024-00763-4. Epub ahead of print. PMID: 39627462. ***(contributing author)***

- Mohanty SK, Maryam S, **Gautam V**, Mittal A, Gupta K, Arora R, Bhadra W, Mishra T, Sengupta D, Ahuja G. Transcriptional advantage influence odorant receptor gene choice. **Brief Funct Genomics.** 2023 May 18;22(3):281-290. doi: 10.1093/bfgp/elac052. PMID: 36542133. ***(contributing author)***

- Saproo S, Sarkar SS, **Gautam V**, Konyak CW, Dass G, Karmakar A, Sharma M, Ahuja G, Gupta A, Tayal J, Mehta A, Naidu S. Salivary protein kinase C alpha and novel microRNAs as diagnostic and therapeutic resistance markers for oral squamous cell carcinoma in Indian cohorts. **Front Mol Biosci.** 2023 Jan 10;9:1106963. doi: 10.3389/fmolb.2022.1106963. PMID: 36703917; PMCID: PMC9871261. ***(contributing author)***

- Mittal A, Mohanty SK, **Gautam V**, Arora S, Saproo S, Gupta R, Sivakumar R, Garg P, Aggarwal A, Raghavachary P, Dixit NK, Singh VP, Mehta A, Tayal J, Naidu S, Sengupta D, Ahuja G. Artificial intelligence uncovers carcinogenic human metabolites. **Nat Chem Biol**. 2022 Nov;18(11):1204-1213. doi: 10.1038/s41589-022-01110-7. Epub 2022 Aug 11. PMID: 35953549. ***(co-first author)***

- Gupta R, Mittal A, Agrawal V, Gupta S, Gupta K, Jain RR, Garg P, Mohanty SK, Sogani R, Chhabra HS, **Gautam V**, Mishra T, Sengupta D, Ahuja G. OdoriFy: A conglomerate of artificial intelligence-driven prediction engines for olfactory decoding. **J Biol Chem.** 2021 Aug;297(2):100956. doi: 10.1016/j.jbc.2021.100956. Epub 2021 Jul 12. PMID: 34265305; PMCID: PMC8342790. ***(contributing author)***

Introduction

Chapter 1: Introduction

The precise prediction of molecular function and cellular attributes is a key component in advancing biomedical science and the pace of drug development. The ability to computationally evaluate the interactions of molecules and biological targets and how they affect cellular behavior is valuable for prioritizing potential drug candidates, understanding disease mechanisms, and ultimately developing more effective therapies. The value of molecular activity prediction is underscored by its key role in enhancing the drug discovery process, mainly when experimental data used to train predictive models are often limited and variable (Zadorozhny et al. 2025). Machine learning algorithms, grounded explicitly in deep learning architectures, have become the primary methodologies in this field, promising to improve the accuracy and efficiency of drug molecule activity prediction (White 2024). These computational models are the backbone of drug discovery, analogous to language models in natural language processing and image classifiers in computer vision (Seidl et al. 2023). They facilitate the selection and prioritization of molecules for follow-up biological assay and molecular structure optimization towards specific therapeutic targets. Molecular activity prediction and QSAR modeling are thus pivotal to informing the crucial steps of candidate selection and optimization in the drug development pipeline.

Structuring Diverse and High-Dimensional Data

To make machine learning algorithms adequately process and understand this wealth of biological information and correctly make predictions about molecular activity and cellular phenotypes, the data will have to be structured in such a manner as to be readable and understandable by these algorithms. Despite the revolutionary potential of computer modeling in biomedicine, there are considerable challenges in store, namely in extracting features from a wide range of types of biological data (Qi 2024). Representative molecular representation, for instance, remains an elementary challenge in cheminformatics since dozens of interrelated parameters determine the nature of a molecule. Genomic and genetic data sets are frequently framed as extremely high-dimensional, i.e., the number of features easily dwarfs the number of available data samples (Mowlaei & Shi 2023). For bacterial genomics, the high inherent dimensionality of the data and the probability of spurious relationships make identifying truly informative features a daunting challenge (James et al. 2025). Even for image data, the absence of structuring and unevenness of the cellular phenotypes unveiled in microscopy images presents enormous analytical challenges. A key real-world scenario where current methods fall short, and which this thesis aims to resolve, is the prediction of blood-brain barrier (BBB) permeability. Conventional methods that rely on pre-computed molecular descriptors and fingerprints often fail in such complex classification tasks because the chemical differences or heterogeneity between molecules that can and cannot cross the barrier

are minimal yet functionally critical. As demonstrated in this thesis, a descriptive analysis of BBB-permeable and non-permeable compounds reveals a "highly intermixed pattern," indicating that they occupy a very similar chemical space. This low heterogeneity renders traditional feature extraction methods insufficient, creating a significant bottleneck in drug discovery for challenging targets like the brain. This failure necessitates the development of more sophisticated models that can learn directly from a molecule's complete topology to capture the subtle structural relationships that determine its biological activity (Gautam et al. 2022).

Ensuring Biological Relevance and Interpretability

A primary challenge is to extract features not only informative but also of biological relevance to capture the rich underlying mechanistic processes responsible for molecular and cellular behavior. The determination of "biological relevance" is not a simple choice between statistical or experimental methods, but as a multi-stage process of building scientific confidence. The process begins with the selection of input features that have inherent biological meaning, such as physicochemical properties linked to membrane permeability or gene expression profiles that provide a snapshot of cellular activity. Statistical and deep learning models then process these features to generate novel predictions. The biological relevance of these statistical outputs is then established by rigorously testing whether the predictions align with, explain, or are consistent with existing, validated biological knowledge. Earlier attempts at predicting molecular properties with artificial intelligence were based on man-engineered features without the broad generalizability and flexibility appropriate to higher-complexity biological systems (Li et al. 2023). High-level interpretability of the minute visual patterns to identify informative patterns within cellular phenotypes within images is required (Pratapa et al. 2021). A few predictive models based on bacterial genomics have struggled under the challenge of generalization and with flagging false positives among associations, demonstrating limitations regarding relevance in the extracted features (James et al. 2025). The sparsity typically encountered with mutation data will normally preclude legacy machine learning methods at feature extraction from being able to capture subtle patterns that relate to novel, rare genetic changes (Jaksik et al. 2024).

To overcome these challenges and take the maximum benefit of computational modeling in the area of biomedical research, a wide variety of feature types are used in computational biology. These features are designed to capture the inherent characteristics of molecules, cells, and biological systems, thereby allowing the construction of precise predictive models.

Chemical and Physicochemical Features

A basic kind of feature originates from the molecule's chemical structure. Molecules can be symbolically represented as strings, such as the Simplified Molecular Input Line Entry System (SMILES), a compact and practical molecular structure encoding scheme (Dablander 2024). SMILES strings are a universal base element for constructing other molecular representations, such as molecular fingerprints, one-hot encodings, and word embeddings. Although SMILES is a convenient and widely adopted format, its linearity and sensitivity to small mistakes are potential drawbacks (Dablander 2024). Alternatively, molecules can be symbolically represented as graphs, with atoms as nodes and bonds as edges, and hence a richer representation of molecular topology than linear strings. Graph-based representations directly specify the connectivity and spatial arrangement of atoms in a molecule (Guo et al. 2022). Graph Neural Networks (GNNs) are explicitly designed to process directly such molecular graphs, learning representations that inherently contain structural information. Sophisticated approaches, such as learnable graph embeddings, which are created by GNNs, are a powerful tool to learn automatically high-level features from molecular graphs. These embeddings can capture complex structural and chemical features, essential for various downstream tasks in drug discovery and cheminformatics. Self-supervised learning techniques are widely applied to pre-train GNNs, successfully addressing the issue of the absence of labeled data in molecular property prediction (Li et al. 2023). Additionally, combining knowledge graph embeddings with generative models has proven to be an effective approach to novel drug candidate design with pre-specified properties (Malusare & Aggarwal, 2024).

Table 1: Comparison of Molecular Representation Methods

Representation Method	Type	Key Features/Advantages	Limitations
SMILES	String-based	Compact, human-readable, and easily parsable by software	Fragile, doesn't fully capture 3D structure or complex topology directly.
Molecular Graphs	Graph-based	Directly represents atoms and bonds, captures molecular connectivity and topology	Can be computationally more intensive than string-based methods.

Molecular Fingerprints	Feature-based	Fixed-length binary or integer vectors encoding the presence or absence of substructures	Compared to graphs, information loss can be sensitive to the choice of fingerprinting algorithm.
Graph Embeddings	Learned	Dense, low-dimensional vector representations learned by neural networks from graphs	Interpretability can be challenging; performance depends on the training data and model architecture.

A further key class of features comprises the physicochemical properties, which yield quantitative descriptions of a molecule's physical and chemical characteristics (Dablander 2024). These determine the behavior of a molecule within biological systems as well as in interaction with the target proteins.

Table 2: Examples of Physicochemical Properties Used in QSAR/QSPR Models

Property Name	**Abbreviation**	**Description**	**Significance in Drug Design**
Lipophilicity	logP,/logD	Octanol-water partition coefficient (logP); Distribution coefficient (logD)	Membrane permeability, drug absorption, distribution, and interaction with hydrophobic binding sites.
Water Solubility	logS	Solubility of a compound in water	Drug absorption, bioavailability, and formulation development.
Molecular Weight	MW (Da)	The sum of the atomic weights of the atoms in a molecule	Membrane permeability, overall size, and shape of the molecule.

Polar Surface Area	PSA/EPSA	Surface area of a molecule composed of polar atoms (usually oxygen and nitrogen)	Drug permeability across membranes, especially the blood-brain barrier, and interactions with polar environments.
Hydrogen Bond Donors	HBD (count)	Number of hydrogen atoms attached to electronegative atoms (e.g., N, O)	Interactions with biological targets, solubility in aqueous environments.
Hydrogen Bond Acceptors	HBA (count)	Number of electronegative atoms (e.g., N, O) with lone pairs of electrons	Interactions with biological targets, solubility in aqueous environments.
Acid Dissociation Constant	pKa	Measure of the acidity of a molecule	Ionization state at physiological pH, affecting solubility, permeability, and interactions with charged residues in targets.
Highest Occupied Molecular Orbital	HOMO (eV)	Energy of the highest occupied molecular orbital	The electron-donating ability of the molecule, the potential for interactions with electron-deficient regions in targets.
Lowest Unoccupied Molecular Orbital	LUMO (eV)	Energy of the lowest unoccupied molecular orbital	The electron-accepting ability of the molecule, the potential for interactions with electron-rich regions in targets.

Table 3: Feature extractors for chemical-based descriptors		
Feature Extractor	**Feature Type**	**Citation**
Mol2Vec	homogeneous (structure-based)	(Jaeger et al. 2018)
SMILES2Vec	homogeneous (structure-based)	(Goh et al. 2017)
ChemicalBERT	homogeneous (structure-based)	(Qin et al. 2020)
Mol-BERT	homogeneous (structure-based)	(Li & Jiang 2021)
ChemBERTa	homogeneous (structure-based)	(Chithrananda et al. 2020)
SMILES-BERT	homogeneous (structure-based)	(Wang et al. 2019)
Signaturizer	heterogeneous (biology-based)	(Bertoni et al. 2021)
Mordred	Chemical-based	(Moriwaki et al. 2018)

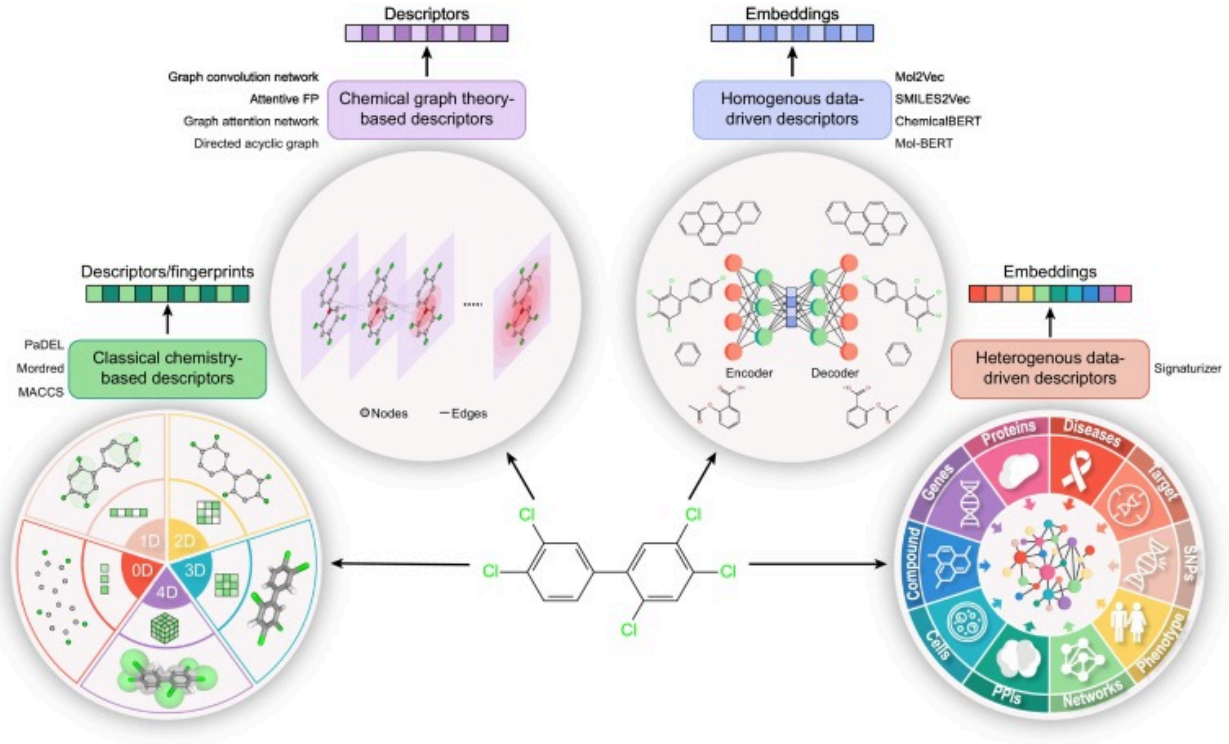


Figure 1.1: Chemical-based features in understanding molecular activity prediction *(Mittal & Ahuja 2023)*

A schematic overview summarizing various molecular representation methods standard in cheminformatics is provided. The classification covers three broad classes: classical chemistry-driven descriptors, chemical graph theory-based approaches, and data-driven descriptors. Data-driven descriptors also fall under subclasses based on input data type. Several methodologies and tools have been developed to compute or extract classical chemistry-driven descriptors. Several methods for classical chemistry-driven descriptors involve 0D, 1D, 2D, 3D, and 4D representations collectively portraying molecular substructures ***(lower left)****. Additionally, several independent packages for computing descriptors exist (Yap 2011; Moriwaki et al. 2018). Chemical graph theory-based approaches are universally used for representing molecules and predicting molecular properties by employing deep learning algorithms. Graph convolution networks, graph attention networks, directed acyclic graphs, and attentive FP are commonly used deep learning methodologies for graph-based modeling (Gautam et al. 2022). Computational tools such as the DeepChem library provide end-to-end capabilities for graph-based modeling. Data-driven descriptors are a recent advancement in cheminformatics for describing molecular structures of chemical compounds. These descriptors employ significant sources of data and self-supervised deep learning models. Mol2Vec (Jaeger et al. 2018), SMILES2Vec (Goh et al. 2017), ChemicalBERT (Qin et al. 2020), and signaturizer (Bertoni et al. 2021)are exemplary examples of data-driven descriptors* ***(upper right)****.*

Image-Based Morphological Features

Image-based features also provide a large amount of data for investigating cellular morphology and phenotypes. Microscopy images provide us with direct visual evidence of cellular conditions and reactions to stimuli (Way et al. 2021). Quantitative features associated with cellular morphology, i.e., size, shape, intensity, and texture information, typical of spatial arrangement and patterns within cells and tissues, are extracted using standard image analysis methods (Du et al. 2022). Geometric features such as area, perimeter, and roundness are readily computed from cell images after segmentation (Kode & Barkana 2023). However, with the recent emergence of deep learning, a dramatic impact has been made in image-based feature extraction with the ability to learn hierarchical representations automatically from image data (Krentzel et al. 2023). Convolutional Neural Networks (CNNs) have been found to possess remarkable abilities in identifying complex and subtle visual patterns in microscopy images and even surpass traditional, manually designed features for cell segmentation and phenotype classification tasks (Krentzel et al. 2023).

Genomic and Transcriptomic Features

Genomic and transcriptomic characteristics offer windows into organisms' and cells' genetic composition and transcriptional activity. Gene expression profiles, derived from single-cell and bulk RNA sequencing

(scRNA-seq), capture a snapshot of active genes in a cell at a single moment in time (Mao et al. 2024). These profiles are very informative for defining various cellular states, interpreting cellular responses to diverse perturbations, and identifying potential drug targets. Genetic variants, like single-nucleotide polymorphisms (SNPs) and insertions/deletions (indels), are sequence differences in the DNA and can be linked with disease susceptibility and drug response (Anon n.d.). Polygenic risk scores, for instance, use data from hundreds of genetic variants to predict an individual's likelihood for developing particular diseases (Anon n.d.). Pharmacogenomics also investigates how these genomic changes affect an individual's response to drug therapy (Li et al. 2024). In addition to genetics, epigenetic markers, such as DNA methylation and histone modification, offer yet another layer of relevant information by revealing how gene expression is regulated without changing the underlying DNA sequence (Koudonas et al. 2024). Such markers can be modulated by the environment and disease state and are therefore potential markers of regulatory processes and valuable predictors of drug response (Koudonas et al. 2024).

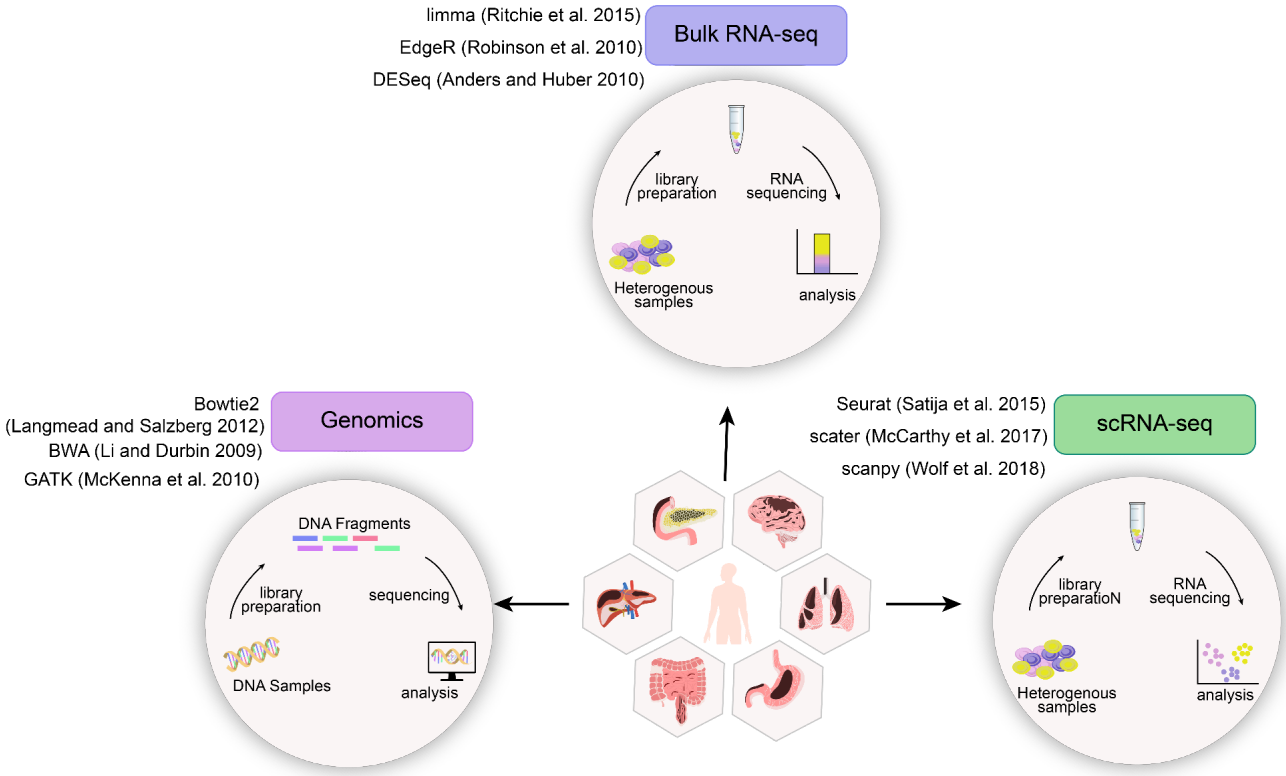


Figure 1.2: Graphical overview of various genomics and transcriptomics approaches to recovering biological attributes from heterogeneous tissue sources.

The figure compares three main approaches: (1) Genomics, aiming at DNA exploration by packages such as Bowtie2 (Langmead & Salzberg 2012), BWA (Li & Durbin 2009), and GATK (McKenna et al. 2010); (2) Bulk RNA-seq, quantifying mean gene expression from mixed samples using analytical packages such as limma (Ritchie et al. 2015), EdgeR (Robinson et al. 2010), and DESeq (Anders & Huber 2010); and (3) Single-cell RNA-seq (scRNA-seq), resolving gene expression at the single-cell level using packages such as Seurat (Satija et al. 2015), scater (McCarthy et al. 2017), and scanpy (Wolf et al. 2018). Each pathway illustrates a general library preparation, sequencing, and computational analysis flow.

Proteomic and Biological Assay Features

Proteomic features focus on proteins, the primary functional molecules in cells, where their sequences provide critical insights into potential functions and properties. For machine learning, these sequences are encoded into numerical formats, enabling models to predict various protein attributes. Additionally, the three-dimensional structures of proteins play a vital role in drug design, particularly through

structure-based drug design (SBDD), which uses atomic arrangements to create small molecules that bind to specific sites and modulate protein activity (Ben David et al. 2023). Techniques like molecular docking predict interactions between molecules and protein binding sites. Protein-protein interaction (PPI) networks further enhance our understanding of cellular processes by mapping interactions between proteins, essential for identifying drug targets (Kurata & Tsukiyama 2022). Complementing these approaches, biological assay data from high-throughput screening (HTS) provide direct experimental measurements of molecular activity and cellular responses. By integrating this data with structural, physicochemical, and omics information, we can develop more robust predictive models that leverage insights from diverse sources (Chen & Wild 2010; Whitehead et al. 2019).

Thesis Scope: A Focused Approach to High-Impact Data Domains

While this thesis acknowledges this vast landscape of biological data, it adopts a focused research strategy to prioritize depth of innovation over breadth of coverage. Rather than a shallow survey of all data types, this work concentrates on three of the most high-impact and challenging data domains, i.e., molecular structures, single-cell transcriptomics, and cellular microscopy, to develop, validate, and deliver complete, end-to-end computational solutions that address significant bottlenecks in their respective fields. Furthermore, a central theme of this work is addressing the critical balance between predictive accuracy and model interpretability. In an era where deep learning models are often criticized as uninterpretable "black boxes," their utility in biomedical research, where understanding the underlying mechanism is as important as the prediction itself, can be limited. This thesis argues that accuracy and interpretability are not a zero-sum trade-off but can exist in a synergistic relationship. The tools developed herein are engineered not only for high performance but also for transparency. This is achieved by integrating specific interpretability features, such as modules that provide direct mechanistic insights into gene regulatory networks (EcTracker) or visually pinpoint the sub-cellular regions a model uses to make its predictions (scCamAge). By generating new, testable hypotheses, these features build trust in the model's accuracy and transform them into genuine scientific discovery tools, a crucial step for their adoption and validation by the broader experimental research community.

1.1 Objective of the Thesis

This thesis aims to develop and prove computational tools for effective feature engineering on different types of biological features to enhance the understanding and prediction of complex biological processes. In particular, the objectives are three: (1) to develop deepGraphh, a user-friendly web platform that uses cutting-edge graph-based methodology for molecular feature representation and Quantitative

Structure-Activity Relationship (QSAR) prediction to enable precise prediction of chemical properties; (2) to develop EcTracker, an R/Shiny web server for single-cell RNA sequencing data analysis that can detect ectopically expressed genes, classify cell types, and regulon-based inference of regulatory networks. Overall, these tools address the challenge of feature engineering in molecular, image-based, and single-cell transcriptomic data in an attempt to facilitate a better understanding of biological systems; and (3) to develop scCamAge, a multimodal deep learning tool that integrates single-cell resolution imaging data with aging-related bioactivities to predict cellular changes, especially in aging, and show its applicability to other organisms.

1.2 Organization of the Chapters

This thesis is organized into five chapters, each presenting the usage of biological features in developing tools and further understanding the complex biological process to gain insights. The organization of the chapters is as follows:

Chapter 1 lays the groundwork for this work by examining the significance of biological features in predicting molecular activity and the methods used in this analysis.

Chapter 2 discusses the core subject of molecular representation in QSAR modeling by presenting deepGraphh, a web service that provides high-powered graph-based modeling to researchers with little programming ability. It discusses the merits of graph-based methods over conventional descriptors and fingerprints and the strengths of deepGraphh in parameter optimization, cross-validation, and model building by different methods.

Chapter 3 introduces EcTracker, a comprehensive web-based tool for single-cell RNA sequencing data analysis designed to assist in distinguishing cellular identities, ectopically expressed genes, and the respective regulatory networks. The chapter also emphasizes the speed and interactive nature of the EcTracker tool and its ability to provide novel insights into single-cell datasets.

Chapter 4 discusses the limitations of existing deep learning-based image analysis systems in identifying complex cellular changes with aging. It presents scCamAge, a context-aware multimodal prediction model that integrates image-derived spatiotemporal features and aging-related bioactivities. The chapter demonstrates scCamAge's performance in validating yeast cell data and discloses its surprising ability to

predict human fibroblast senescence, even when trained on yeast only. It shows that retraining on human data greatly enhances scCamAge's prediction capacity.

Chapter 5 concludes the thesis by summarizing the findings, discussing their implications, and proposing future research directions. It emphasizes the integrative approach taken in this work, combining multiple types of biological features and creating tools to understand and unravel the diverse molecular activities.


Objectives

Chapter 2: To Develop and Optimize Graph-Based Computational Models for Quantitative Structure-Activity Relationship (QSAR) Analysis

2.1 Introduction

To forecast different drug-like qualities, one of the most crucial tasks in any chemoinformatics workflow is to accurately and thoroughly represent the chemical structure. It takes a lot of effort and time to extract significant features for a specific chemical through experimentation (Lo et al. 2018). Furthermore, the rate at which new compounds are created makes it hard to use only experimental methods to extract the properties of this ever-expanding chemical space (Lo et al. 2018). Significantly, a range of computational tools have been developed that satisfy this seeming need and offer a thorough and quick method for chemoinformatics analysis (Yap 2011; Moriwaki et al. 2018; Bertoni et al. 2021). These techniques enable the extraction of useful features from them as well as machine-readable chemical representations and tools. Conventional techniques for generating features make use of 1D, 2D, and 3D chemical structure representations (Capecchi et al. 2020). Furthermore, machine learning techniques for model construction (classification or regression) employ these molecular descriptors as input features. Additionally, the characteristic can typically also be represented as a binary vector with a given string length, referred to as molecular fingerprints (Morgan 1965; Rogers & Hahn 2010). Atom pair fingerprints, Morgan fingerprints (Morgan 1965; Rogers & Hahn 2010), substructure fingerprints, MACCS fingerprinting (Riniker & Landrum 2013), MinHashed fingerprints MHFP6 (Probst & Reymond 2018), and other types of molecular fingerprints are among the many varieties that are available and can be applied to both big and tiny molecules. Even while different fingerprint types vary greatly, frequent use of them in chemoinformatics activities is rather prevalent **(Figure 2.1)** *(Fernández-Torras et al. 2022)*.

Although both molecular fingerprints and descriptors are useful in representing the chemical properties of the input molecules, they often fail in complex classification tasks where heterogeneity between the classes is quite low, which is often the case with biological datasets. In pursuit of developing a method that accounts for both chemically and biologically relevant information, recently, bioactivity-based descriptors have been proposed. Signaturizer is a recently developed method for the generation of bioactivity-based descriptors. It generates 128-dimensional vectors (25 vectors in total) that collectively describe the chemical properties of the molecule, their targets and metabolic genes, network properties of the targets, cell response profiles, drug indications and side effects, among others (Bertoni et al. 2021). Although the methods for feature generation are increasingly diversifying, equal efforts are ongoing for the chemical representation itself that facilitates the automated generation of chemical features. For example, the compounds can be represented as graphs and aligned with the deep learning methods for the

downstream classification/regression tasks (Sun et al. 2020). Consequently, the models in deepGraphh do not inherently handle 3D spatial features or stereochemistry, such as chirality, which can be essential for determining drug-like properties. This is a key distinction from some traditional methods that explicitly compute 3D descriptors and represent a scope definition for the current version of the tool. Importantly, irrespective of the feature generation techniques, all these methods facilitate the rapid generation of chemical information that cannot be achieved using experimental techniques within the stipulated time frame. Despite numerous benefits, one major limitation of these approaches is that not all computationally derived descriptors can be tested by experimental approaches, as some of them are mere mathematical representations.

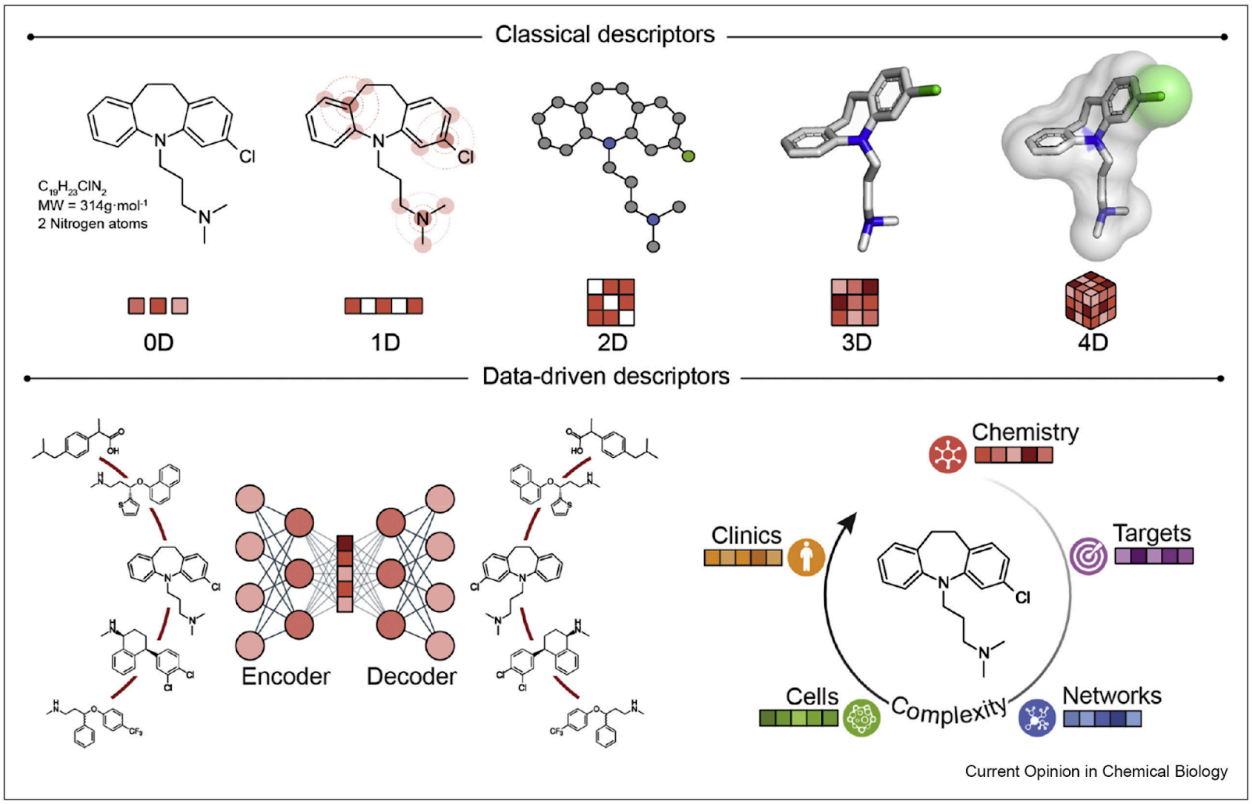


Figure 2.1: Encoding chemical compounds based on their chemical properties and biological activities

Broadly, the models developed by machine learning-based algorithms can be classified as predefined descriptor-based models and automatic models (e.g., graph-based methods). Although the former offers an off-the-shelf and widely used approach for model generation, the latter offers a new means for chemical representation, and due to its astonishingly high performance, it is recently gaining rapid popularity in the field of chemoinformatics (Yang et al. 2019; Xiong et al. 2020). The descriptors-based methods utilize machine learning algorithms such as support vector machine (SVM) (Ma et al. 2015;

Alvarsson et al. 2016), gradient boosting (GBM), logistic regression (LR) (Ren et al. 2016), multilayer perceptron (MLP), random forest (RF) (Svetnik et al. 2003; Zhang & Aires-de-Sousa 2007), extra tree classifier (ET), etc., for classification/regression tasks, whereas the graph-based methods utilize the features of molecular graphs as inputs, and subsequently, specialized graph-based deep learning algorithms are used to train the desired models. Briefly, graph-based methods allow the representation of compounds as graphs G = (V, E), where V represents the atoms and E represents the bonds between the atoms. These chemical graphs are used as input for graph-based classifiers such as graph neural network (GNN), which extract features from these graphs and subsequently use them for downstream analysis (Sun et al. 2020; Jiang et al. 2021). Although the comparative analysis between the traditional descriptors-based and graphs-based methods is highly debatable and data-dependent, a few studies strongly advocate for the outperformance of the graph-based methods (Yang et al. 2019; Xiong et al. 2020; Wu et al. 2018). Here we introduce deepGraphh, an online web service for building classification or regression models using graph-based methods. Importantly, deepGraphh is highly configurable and presents an easy-to-implement graphical user interface (GUI). deepGraphh also supports cross-validation and returns comprehensive results, both in graphical as well as tabular formats. One of the key highlighting features of deepGraphh is that it allows live tracking of the major steps involved in model generation. By utilizing deepGraphh-supported models, we built classification and regression models for the blood-brain barrier (BBB) permeability prediction. We used these models to predict the BBB permeability of humans (Meng et al. 2021) and microbiome-generated metabolites (Cheng et al. 2022). Finally, comparative analysis with traditional chemical-based descriptors revealed the comparable performance of graph-based methods. In summary, deepGraphh is the first open-source web service that supports multi-functional graph-based deep learning frameworks for predicting various drug-like properties

2.2 Methodology

2.2.1 Software Architecture

Front-end: The front-end is built with HTML, CSS, and JavaScript. All HTML pages are primarily static. All of the parameter value sets are fed into a model, which evaluates the results using Python.

Database: A database is constructed to keep track of the user-generated models. It is created with SQL (Structured Query Language), which is used to store, manipulate, and retrieve data. The database holds data for each user, including information on the parameters selected and the model's current stage.

Back-end: DeepGraphh was created using DeepChem libraries. skearn library is used to evaluate the model, including accuracy, precision, recall, Area Under the Receiver Operating Characteristic (ROC)

Curve (AUC-ROC), Matthews correlation coefficient (MCC), F1 score, and Cohen's kappa. The Run.py file contains all of the functions that accept input from pages, feed it into the model, and save the result to the server. It also includes database function calls, which are used to insert data for the building of a new model and display downloaded results.

2.2.2 Mordred Descriptors

Mordred descriptors were computed using the Mordred (1.2.0 version) command-line interface. Mordred produced 1826 two- and three-dimensional molecular descriptors. Mordred-generated descriptors were further preprocessed, including data trimming and handling of missing values (Moriwaki et al. 2018).

2.2.3 Signaturizer Descriptors

The Signaturizer (1.1.11 version) command-line interface was used to generate bioactivity-based descriptors. Signaturizer created a multidimensional vector with a size of 3200. Signaturizer-generated descriptors underwent additional preprocessing, including data trimming and treatment of missing values (Bertoni et al. 2021).

2.2.4 Model Generation

DeepGraphh performs classification and regression tasks using the DeepChem library and four widely used graph-based models: Graph Convolution Network (GCN), Directed Acyclic Graphs (DAG), Graph Attentive Network (GAT), and Attentive FP (AFP). Notably, the DeepChem project offers a suite of high-quality deep learning tools for analysing chemistry datasets. It provides a variety of chemistry datasets and tutorials for doing chemoinformatics activities, rather than being the result of an exhaustive comparative study against all possible GNN architectures like MPNN or DimeNet. Notably, the DeepChem project offers a suite of high-quality deep learning tools for analysing chemistry datasets. It provides a variety of chemistry datasets and tutorials for doing chemoinformatics activities.

2.2.4.1 Graph Convolution Network (GCN)

The GCN method was proposed by Kipf et al. (Kipf & Welling 2016). It follows the propagation rule:

$$H^{(l+1)} = \sigma(\widehat{D}^{-1/2}\widehat{AD}^{-1/2}H^{(l)}W^{(l)})$$

where $H^{(l)}$ and $W^{(l)}$ are the l^{th} neural network layers.σ is the non-linear activation function and $\widehat{D}$ and $\widehat{A}$ are the degree and adjacency matrix, respectively. Convolution means the set of kernels that have weights that slide through the graph and learn features from the neighbors. It aggregates the information to form a representation and then transforms the representation with linear projection and then by non-linear activation (Kipf & Welling, 2016).

2.2.4.2 Graph Attentive Network (GAT)

GAT is an extension of the vanilla GCN but only differs in the method of information aggregation. In GAT, different weights are given to the nodes by the attention mechanism, and then convolution is done by the weighted sum of the information using the formula:

$$H_i^{(l+1)} = \sigma(\Sigma_{j\epsilon N(i)} a_{ij}^{(l)} W^{(l)} H_i^{(l)})$$

where $a_{ij}^{(l)}$ is the attention score for the node i and j for the l^{th} layer, W is the learnable matrix, $N(i)$ is the neighbor of the i node, and σ is the activation function. Unlike GCN, GAT focuses on the meaningful part of the graph (Veličković et al. 2017).

2.2.4.3 Attentive FP (AFP)

AFP was initially proposed by Xiong et al., and it uses a recursive neural network (RNN) to agglomerate the information of the graphs. It focuses on the important part of the graphs and learns non-local interactions between two molecules from a specified task, which helps in gaining the molecular property directly from data that is not interpretable by humans (Xiong et al. 2020).

2.2.4.4 Directed Acyclic Graphs (DAG)

DAG considers the graphs as directed graphs, although chemical molecules are not directed graphs. DAG considers one center atom, and for all other atoms, directions are defined toward the orientation of the center atom. For example, if a molecule has n_a atoms, then n_a DAGs will be generated. In DAG, atom features and bond features are calculated for all graphs and then fed to the classification and regression model (Lusci et al. 2013).

2.2.5 Hyperparameter Tuning

Hyperparameter tuning is used to determine the best settings for a learning algorithm. Machine or deep learning models have varying restrictions or parameter values depending on the dataset. Hyperparameter

tweaking helps to identify the appropriate settings for the best model performance. A tuning grid containing the selected parameters is created within the permissible range to conduct a hyperparameter search. For a training dataset, all of the grid's parameters are run, and model evaluation is performed on each hyperparameter to get the best ideal parameters for the dataset. It is important to note that a standard k-fold cross-validation was used to evaluate the stability of the final, optimized model, rather than a nested cross-validation approach for the hyperparameter tuning process itself. The classification model (GCN) uses the following parameters: n_tasks = 1, batch_size = 32, graph_conv_layers = [128,128], predictor_hidden_feats = 256, learning_rate = 0.01, and predictor_droput = 0. For the regression model (DAG), we used the following parameters: n_tasks = 1, batch_size = 32, layer_sizes_gather = [128,128], layer_sizes = [64,64], and dropout = 0.

2.2.6 Datasets

The curated datasets for the blood-brain barrier utilized in this investigation were obtained from Meng et al. 2021 (Meng et al. 2021). The authors offered two types of datasets: classification and regression analysis. Notably, the categorization dataset includes molecules with and without BBB permeability, totaling 7807 compounds. To account for any experimental uncertainty, we exclusively used B3DB-provided cleaned/processed data (Meng et al. 2021). Notably, during preprocessing, RDKit filtered out 7 compounds that were not transformed into molecular structures. The second dataset for regression analysis contains 1058 chemicals, 7 of which were ruled out using RDKit (Landrum n.d.). The Human Metabolome Database (HMDB) (Wishart et al. 2018) and gutMGene databases (Cheng et al. 2022) were used to obtain information on human metabolites as well as metabolites produced by the gut microbiome. These databases contain 217445 and 186 metabolites, respectively. Notably, during HMDB preprocessing, RDKit removed 137 metabolites because they had not been transformed into molecular structures.

2.3 Results

2.3.1 To design and implement novel graph-based algorithms that capture complex molecular structures and their interactions for accurate activity prediction.

The introduction of artificial intelligence in drug design has resulted in the development of a plethora of new tools and methods that not only improved and simplified the overall QSAR workflow but also added computationally driven, faster, less exhaustive, and fully automated methods of feature extraction from chemical structures. One such technique is to represent compounds as molecular graphs and then use graph-based algorithms (such as GNN) for feature extraction and model development (Sun et al. 2020).

Although graph-based methods in chemoinformatics are gaining popularity, their inherent demand for high programming skills and the lack of publicly available GUI-enabled solutions hinder their adoption by the larger community. To address this, here we introduce deepGraphh, an open-source, first-of-its-kind, and one-stop platform for model generation using graph-based methods. deepGraphh supports four different graph-based methods that include Graph Convolution Network (GCN) (Kipf & Welling 2016), Directed Acyclic Graph (DAG) (Wu et al. 2018; Lusci et al. 2013), Attentive FP (AFP) (Xiong et al. 2020), and Graph Attention Network (GAT) (Veličković et al. 2017) **(Figure 2.2)**. GCN is the first approach supported by deepGraphh, allowing for feature extraction via adaptive learning (Kipf & Welling, 2016). DAG, the second approach implemented in deepGraphh, treats chemical structure as a directed graph. Because chemical structures are not represented as directed graphs, the core atom is chosen, and several directed graphs are formed from the remaining atoms (Wu et al. 2018; Lusci et al. 2013). deepGraphh also supports AFP, which concentrates on the important parts of graphs (Xiong et al. 2020). AFP is not limited to characterizing the local atomic environment by spreading node information from neighboring nodes to further away ones. It does, however, facilitate non-local effects at the intramolecular level through the use of graph attention mechanisms. This aspect of AFP makes it an effective tool for revealing buried important information about molecular topological features. Finally, GAT offers alternative neural network-based designs that operate on graph-structured data and use masked self-attentional layers. **(Figure 2.2)** (Abdelaziz et al. 2016). It is one of the best GNN designs, employing the attention mechanism to perform statistically normalized convolution operations. It assigns weights to surrounding nodes using stacking layers, allowing all nodes to focus on the features in their area. Notably, deepGraphh's complete backend source code is written in Python. DeepGraphh makes use of the DeepChem library's (https://deepchem.io/) four most popular classification/regression procedures. The deepGraphh workflow consists of several steps, beginning with the conversion of a chemical Simplified Molecular-Input Line-Entry System (SMILES) into molecular graphs, followed by the extraction of features from these graphs and their subsequent use for model tuning/building via an elegant and highly configurable GUI support. DeepGraphh also allows users to perform K-Fold Cross Validation (CV) and predictions on the external dataset they provide. In brief, the deepGraphh workflow takes user-supplied training and testing files in the comma-separate variable (CSV) format, where each file contains information about chemical names (real or arbitrary), their SMILES, and their activation status in binary format ('0' for non-activating compounds and '1' for activating compounds). Notably, deepGraphh enables multi-classification, which can be manually selected by the user. In this instance, the activation status consists of numerous positive values (0, 1, 2, 3, etc.), with each integer representing a class. Notably, for the best outcomes, balanced data should be used in the training dataset. The processes described above are the same for all four approaches supported by deepGraphh. However, each technique

differs significantly in the downstream steps, beginning with feature extraction, which in deepGraphh is handled by two algorithms: MolGraphConv Featurizer and ConvMol Featurizer. The DAG technique uses the ConvMol Featurizer, but the other three methods use the MolGraphConv Featurizer, which is a universal featurizer for graph convolution-based approaches **(Figure 2.2)**. This assignment of a specific featurizer to each model type is a fixed implementation detail within the DeepChem library, and deepGraphh does not provide an option to interchange them. As a result, an analysis of model sensitivity to the choice of featurizer was not performed as part of this study. Regardless of the feature, deepGraphh allows users to create either classification or regression models. The time required for model construction is proportional to the dimensions of the input training dataset; thus, deepGraphh provides a live tracking mechanism to keep the user informed about the current operation. DeepGraphh provides users with a zip file under the download button after successful model building, which contains user-supplied files, a log file of the entire run containing information about model parameters selected by the user, a confusion matrix, a raw model file, a prediction outcome on the testing dataset, and the AUC-ROC curve. To facilitate quick learning and familiarization with model parameters, the deepGraphh web server includes a full tutorial component. Furthermore, it includes a number of exemplary datasets for users to test deepGraphh functions (Abdelaziz et al. 2016; Cui et al. 2019; Hua et al. 2021). If users want to store their built model and reuse it later to test different compounds, we provide a Python notebook that allows them to load the saved model and make predictions on fresh datasets. In a word, deepGraphh provides a greatly simplified user interface and a variety of graph-based approaches for QSAR research.

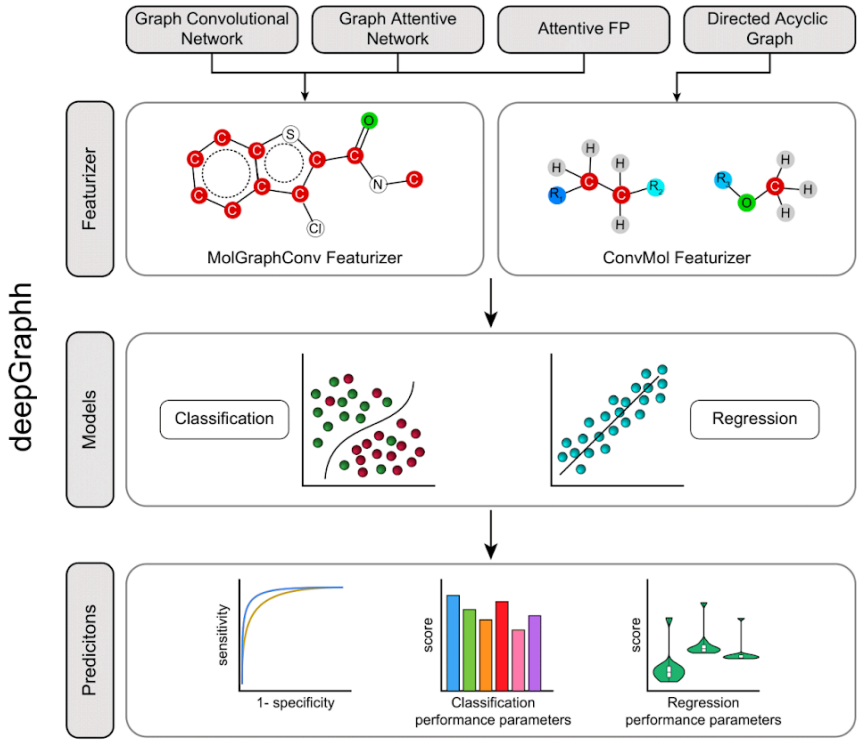


Figure 2.2: deepGraphh, a comprehensive web server for graph-based modeling for QSAR analysis

A schematic representation of the graph-based methods supported by deepGraphh for Quantitative Structure-Activity Relationship modeling. Scheme illustrating four methods supported by deepGraphh, the implemented featurizers for feature extraction, model types, and the output results.

2.3.2 To enhance the deepGraphh web service by integrating machine learning techniques that improve the precision and scalability of QSAR models.

We then used deepGraphh to create a classification model that predicted blood-brain barrier permeability. To accomplish this, we used one of the largest manually constructed datasets available to date. The collection contains 7800 chemicals, with 4949 and 2851 classified as positive (BBB+) and negative (BBB-) for blood-brain barrier permeability, respectively **(Figure 2.3 A)** (Meng et al. 2021). To analyze the chemical makeup of the compounds in each class (BBB+ and BBB-), we first produced their atom-pair fingerprints and then projected them onto the three-dimensional Principal Component Space. The findings show a highly intermixed pattern of BBB+ and BBB- substances in the PCA space, implying a higher degree of chemical similarity between the classes **(Figure 2.3 B)**. We then investigate whether there is any differential enrichment of functional groupings between BBB+ and BBB-. To test this, we

used Bioconductor's ChemmineR program (Cao et al. 2008) to determine the enrichment of the 12 common functional groups across both classes. Our findings point to a partial de-enrichment of primary amines, carboxylic acids, and primary alcohol groups in the BBB+ dataset **(Figure 2.3 C)**.

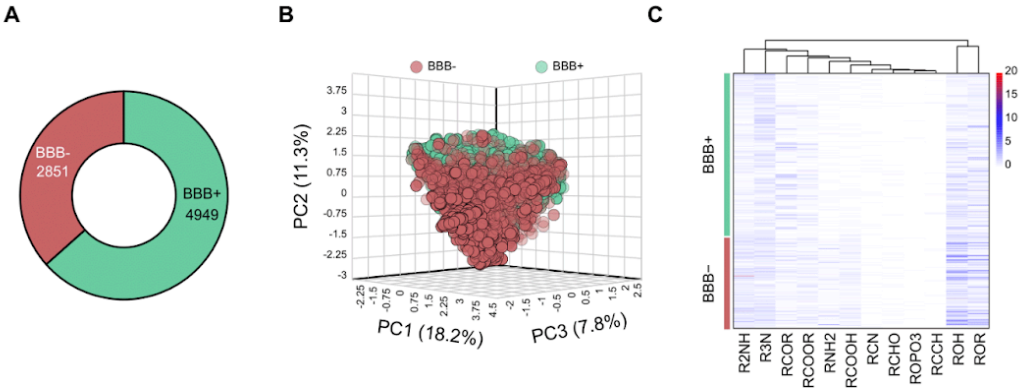


Figure 2.3: Descriptive analysis of blood-brain barrier permeability dataset

(A) *Donut plot depicting the number of compounds with and without blood-brain barrier permeability used as training/testing data.* ***(B)*** *Principal Component Analysis depicting the chemical heterogeneity between BBB+ and BBB- compounds.* ***(C)*** *Heatmap depicting the relative enrichment of the indicated functional groups in the BBB positive (BBB+) and BBB negative (BBB-) dataset.*

Finally, we used machine learning/deep learning to create classification models for BBB+ and BBB-. Notably, in the case of graph-based deep neural networks, we used all four models offered by deepGraphh: Graph Convolution Network (GCN), Directed Acyclic Graph (DAG), Attentive FP (AFP), and Graph Attention Network (GAT) **(Figure 2.4 A, B, D)**. Furthermore, for the descriptor-based machine learning approaches, we used widely used algorithms such as RF (Svetnik et al. 2003; Zhang & Aires-de-Sousa 2007), MLP, KNN (Mitchell 2014), Gaussian Naive Bayes (GNB), Gradient Boosting Classifier (GBC), Extreme Gradient Boosting (XGB), and Logistic Regression (LR) (Ren et al. 2016) **(Figure 2.4 C, E)**. Notably, we used two types of descriptors to generate features for machine learning-based methods: chemistry-based (Mordred) (Moriwaki et al. 2018) **(Figure 2.4 C, E)** and bioactivity-based (Signaturizer) (Bertoni et al. 2021) **(Figure 2.4 C, E)**. To compare the methods, we used the same training and testing datasets for graph-based and descriptor-based methods. We created 18 unique models utilizing the graph-based, Mordred-based, and Signaturizer-based techniques. A comparative investigation of model performance found that graph-based and descriptor-based techniques performed similarly. Furthermore, when the Mordred and Signaturizer-based machine learning models were compared, models created on bioactivity-based descriptors outperformed the other models.

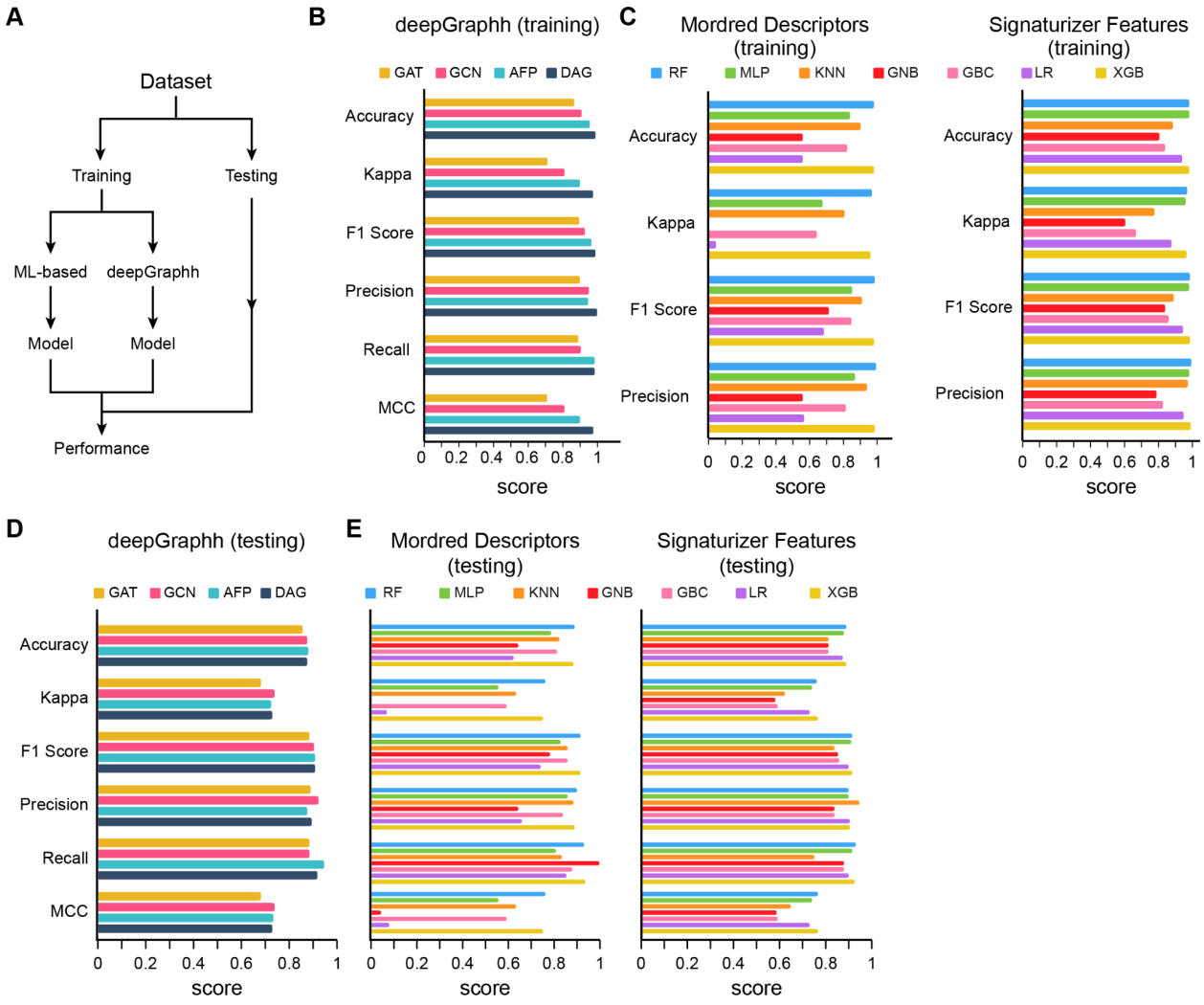


Figure 2.4: Development of classification predicting blood-brain barrier permeability using chemical-based, bioactivity-based, and graph-based methods.

(A) *Flowchart of the methodology used for the analysis.* ***(B-C)*** *Bar graphs depicting the performance parameters such as Accuracy, Cohen Kappa, F1 Score, Precision, Recall, and Matthews correlation coefficient (MCC) on the training dataset.* ***(D-E)*** *Bar graphs depicting the performance parameters such as Accuracy, Cohen Kappa, F1 Score, Precision, Recall, and Matthews correlation coefficient (MCC) on the testing dataset.*

Interestingly, comparing the AUC-ROC values across all models indicated consistency among graph-based methods, unlike other machine learning-based models **(Figure 2.5 A, D)**. Notably, while GCN and AFP have similar testing accuracies of 0.876 and 0.877, respectively, their training accuracies are substantially different, at 0.909 and 0.952. Model overfitting was one of the criteria we used to reject models. Because GCN performed better in both training and testing, we chose it for a rigorous 10-fold

cross-validation (CV) and found stable model performance **(Figure 2.5 B, C, E-G)**. This model was then utilized to do additional downstream analysis **(Table 4)**.

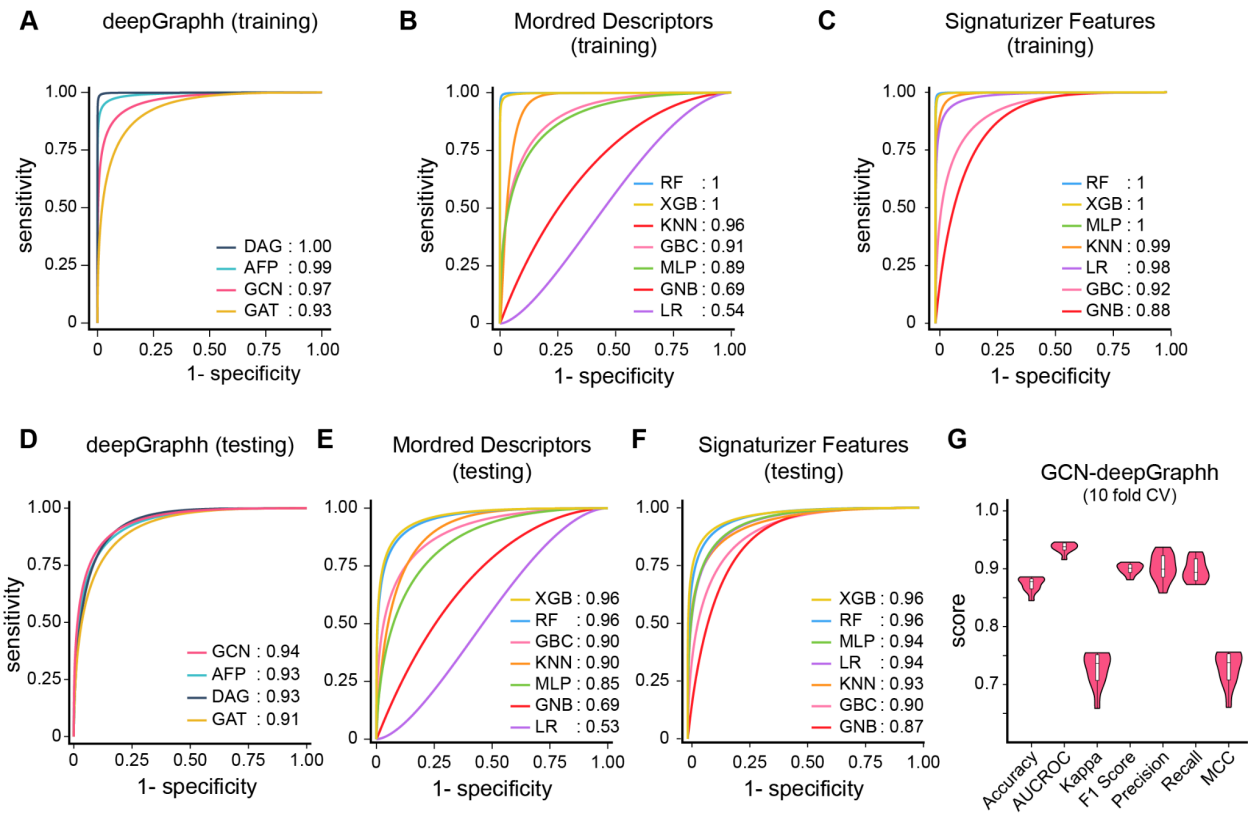


Figure 2.5: Comparison of the AUC-ROC values for classification models for predicting blood-brain barrier permeability using chemical-based, bioactivity-based, and graph-based methods.

(A-C) *AUC (Area under the curve) plots representing the performance of indicated models in classifying the BBB+ and BBB- compounds using the graph-based methods (left side), chemical descriptors computed using Mordred (center), and Bioactivity-based descriptors (right). Classifiers are mentioned on the right of each plot for the training dataset.* ***(D-F)*** *AUC (Area under the curve) plots representing the performance of indicated models in classifying the BBB+ and BBB- compounds using the graph-based methods (left side), chemical descriptors computed using Mordred (center), and Bioactivity-based descriptors (right). Classifiers are mentioned on the right of each plot for the testing dataset.* ***(G)*** *Violin-box plot depicting the distribution and the median of results of 10-fold cross-validation of the best performing classification model i.e., GCN model of deepGraphh.*

We then create the regression model using a subset of the BBB dataset for which logBB values were included in the generated dataset. logBB (logarithm of the Blood-Brain Barrier partition coefficient) quantifies how well a compound can cross the blood-brain barrier (BBB). For regression analysis, we used the same classifiers and features as for classification models. A comparison of regression model's

Mean Squared Errors (MSE), Root Mean Squared Error (RSE), and Mean Absolute Error (MAE) demonstrated that graph-based and descriptor-based techniques performed similarly. Of note, among the graph-based techniques, we discovered minimal errors in the DAG-based model **(Table 4)** that were extremely significant; therefore, we chose this model for the downstream prediction analysis **(Figure 2.6A-G)**.

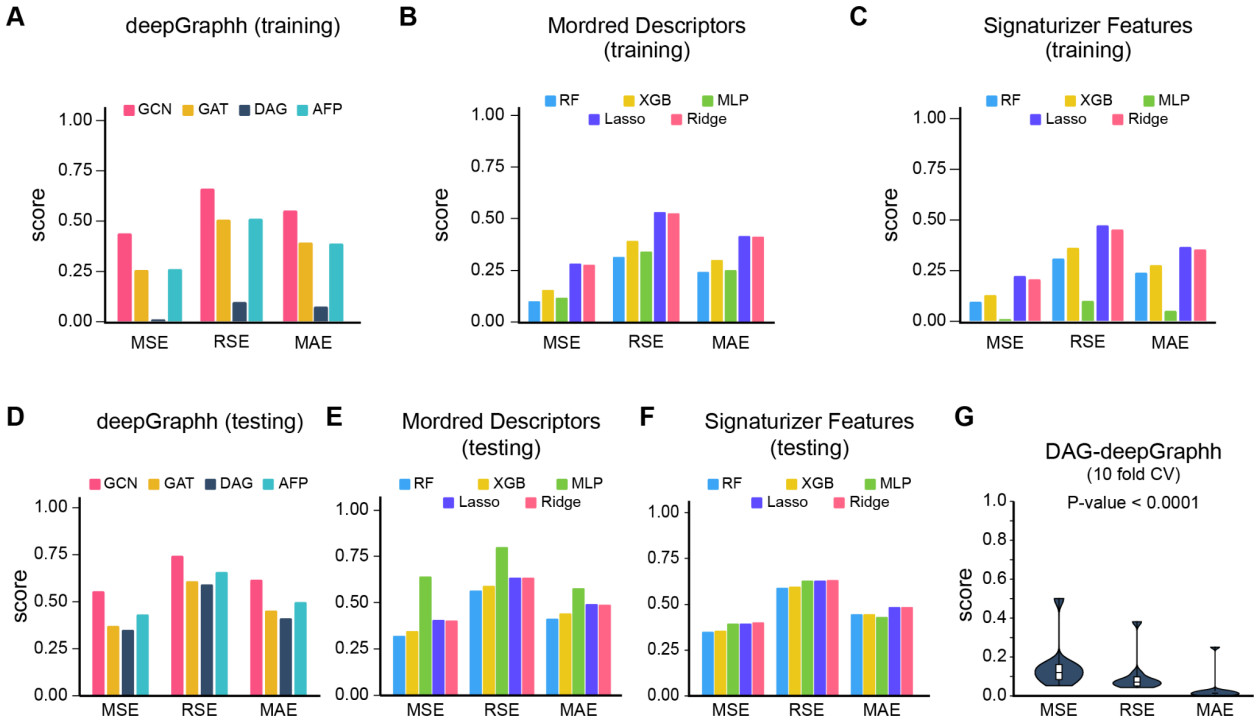


Figure 2.6: Development of regression models for predicting blood-brain barrier permeability using chemical-based, bioactivity-based, and graph-based methods.

(A-C) *Bar graphs depicting the performance parameters of the indicated regression models using the training dataset. Performance is evaluated using Mean Squared Errors (MSE), Root Mean Squared Error (RSE), and Mean Absolute Error (MAE).* ***(D-F)*** *Bar graphs depicting the performance parameters of the indicated regression models using the testing (external) dataset.* ***(G)*** *Violin-box plot depicting the distribution and the median of results of 10-fold cross-validation of the best-performing classification model i.e., DAG model of deepGraphh.*

Table 4: Best Model Parameters for Classification and Regression Model				
Graph Convolution Network			**Directed Acyclic Graph**	
Parameters	Value		Parameters	Value
mode	classification		mode	regression
n_tasks	1		n_tasks	1
batch_size	32		batch_size	32
graph_conv_layers	[128,128]		layer_sizes_gather	[128,128]
predictor_hidden_feats	256		layer_sizes	[64,64]
learning_rate	0.01		dropout	0
predictor_dropout	0			

One of the most important components in maintaining central nervous system homeostasis is the function of the blood-brain barrier (BBB). Multiple variables influence BBB permeability, with metabolites being the least well-characterized biomolecule (Meng et al. 2021). Recent studies have found that the chemicals released by the gut microbiota affect the BBB. However, comprehensive estimates of the BBB permeability of human or microbial-derived metabolites remain lacking. To test this, we collected a list of all human and microbial metabolites from the HMDB and gutMGene databases, respectively, and projected them on our top-performing BBB permeability prediction models (GCN-based model for classification and DAG-based model for regression) (Meng et al. 2021; Cheng et al. 2022). Projection of human and gut microbial metabolites on our classification model revealed that only a subset of human metabolites (7%) have BBB permeability **(Figure 2.7A)**, whereas approximately 37% of microbial metabolites have the potential to permeate the BBB **(Figure 2.7B)**. In the HMDB dataset, 217308 query metabolites were evaluated for BBB permeability potential, with only 15230 projected to be positive (cutoff > 0.5). We then filtered our BBB+ predicted metabolites using rigorous analysis. To accomplish this, we projected the BBB+ anticipated (15230) metabolites onto our DAG-regression model, yielding predicted logBB values. Notably, we employed three distinct cutoffs for identifying BBB permeable/non-permeable/doubtful substances, which is broadly consistent with the literature (Vilar et al. 2010). We identified 788, 3146, and 11296 metabolites using the cutoffs of logBB < -1, logBB ≥ 0.3, and logBB between -1 and 0.3, indicating BBB non-permeable, permeable, and questionable substances, respectively. Since the gut microbiota After applying strict cutoffs, we found 12 metabolites with logBB < -1, 2 metabolites with logBB ≥ 0.3, and 55 metabolites with logBB values between -1 and 0.3. Notably, a literature review indicated that pyruvate (logBB = 0.36), a predicted BBB-permeable metabolite by our model, is known to have BBB permeability (Dopkins et al. 2018; Agus et al. 2018). Finally, we inquired

whether the metabolites with BBB permeability are confined or distributed across discrete metabolic pathways. To test this, we performed functional Over Representation Analysis (ORA) and Pathway Enrichment Analysis on the predicted BBB+ human and microbial metabolites and found selective enrichment for metabolites involved in amino acid metabolism, pyruvate metabolism, steroid hormone biosynthesis, and so on **(Figure 2.7C, D)**. Notably, the functional over-representation analysis with metabolites from the gutMGene database yielded no clear results, most likely because of their small number. Notably, in the case of functional analysis using microbial metabolites, we did not see any discernible pattern, which could be attributed to the small number of mapped compounds.

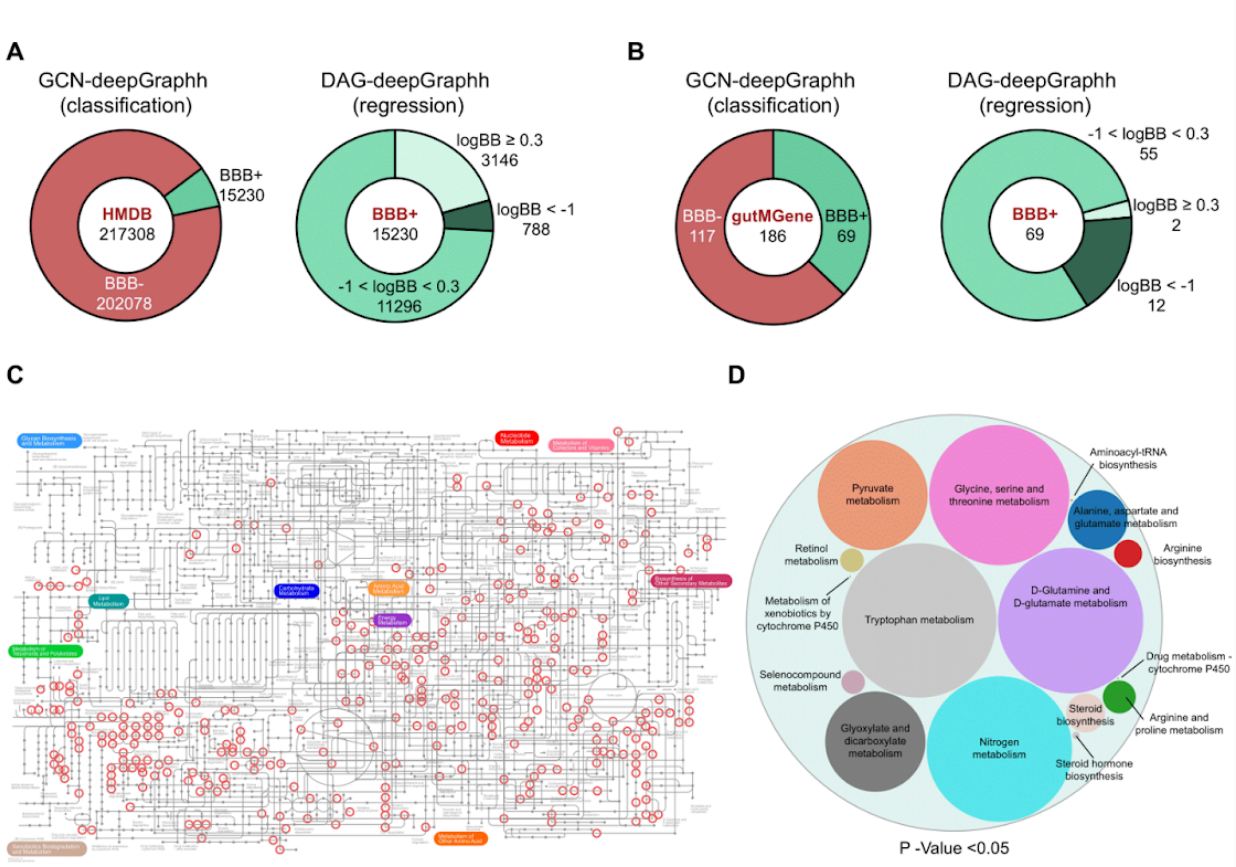


Figure 2.7: Implementation of graph-based methods for predicting blood-brain barrier permeability of human and gut microbial metabolites.

(A) *Donut plots depicting the number of predicted human metabolites from The Human Metabolome Database (HMDB) classified as BBB+ using the best performing graph-based QSAR model. Donut plots on the right illustrate the number of human metabolites predicted to possess logBB scores < -1, ≥ -0.3, and -1 < logBB < 0.3.* ***(B)*** *Donuts plots depicting the number of predicted gut microbiome-generated metabolites curated from gutMGene databases and classified as BBB+ using the best-performing graph-based QSAR model. Donut plots on the right illustrate the number of gut microbiome-generated metabolites predicted to possess logBB scores < -1,≥ -0.3, and -1 < logBB <*

0.3. ***(C)*** *Schematic diagram depicting the human metabolic map, with the overlaid red dots representing the metabolites, predicted as BBB+.* ***(D)*** *Circle Packing plot depicting the results of the functional enrichment analysis of metabolites predicted as BBB+ from The Human Metabolome Database (HMDB).*

2.4 Discussion

While the use of Artificial Intelligence-based methods in computer-aided drug discovery is growing at an unprecedented rate, similar efforts are underway to develop or improve existing methods that could represent compounds as a feature-enriched entity and allow for the extraction of detailed features (Vamathevan et al. 2019; Paul et al. 2021). Traditionally, feature information from chemical structures has been extracted using machine learning methods and molecular descriptors or fingerprints. Though these methods are extremely standardized across many programming platforms and even web servers, they can suffer significantly when the training dataset is more homogeneous (Capecchi et al. 2020; Morgan 1965; Riniker & Landrum 2013). Recent papers recommend representing compounds as graphs and using downstream deep learning-based approaches for feature extraction, classification, and regression modeling. A comparative investigation of diverse training datasets found that graph-based methods outperformed regular descriptor-based methods. For example, Wu and colleagues proposed that the graph-based strategy outperformed while exploiting a diverse training dataset ranging from quantum mechanics to biology-based challenges (Wu et al. 2018). Korolev et al. found that graph-based approaches like GCN outperformed machine learning models such as SVM and Gradient Boosting (Korolev et al. 2020). In addition to other similar research that argues that graph-based approaches perform best, there are also competing reports that show the contrary (Mayr et al. 2018; Jiang et al. 2021). Regardless of the efficacy of the graph-based methods, one cannot deny that they do add a new dimension to the field of computer-aided drug creation. Furthermore,a realistic evaluation of any approach necessitates rigorous testing utilizing diverse and larger datasets, as it is well understood that the performance of any prediction model is dependent on a variety of factors, ranging from hyperparameters to the composition of the training or testing datasets. All of this suggests that no single strategy can work for all datasets, and that users should experiment with multiple methods/algorithms while developing their models. While there are several centralized web services available for creating descriptor/fingerprint-based machine learning models, such as OCHEM (Online chemical modeling environment), there is currently no single web service available for graph-based QSAR analysis (Sushko et al. 2011). We present deepGraphh, the first graph-based deep learning web service. deepGraphh is a one-stop web service for graph-based chemoinformatics approaches, allowing users to keep their data for up to a month. deepGraphh is open-source and free to use. The source code is available on our laboratory's GitHub repository. DeepGraphh offers four primary approaches of DeepChem: GCN, AFP,

DAG, and GAT. These methods use several featurizers to extract graph-based features and give users the ability to experiment with different parameters for model tuning, as well as build final models for classification or regression tasks. As a proof-of-concept, we use deepGraphh-supported models to predict the blood-brain barrier permeability of human and gut microbiome metabolites (Meng et al. 2021; Cheng et al. 2022). Our findings indicate that metabolites linked with amino acid metabolism have increased blood-brain barrier permeability **(Figure 2.7 C, D)**. This increased permeability may be attributed to the presence of numerous amino acid transport mechanisms on both sides of the BBB (Zaragozá 2020). Surprisingly, we found blood-brain barrier permeability in only a few gut microbiome-derived compounds. In addition, we conducted a comparison analysis with descriptor-based machine learning approaches employing two distinct types of features: chemistry-based (Mordred) (Moriwaki et al. 2018) and bioactivity-based (Signaturizer) (Bertoni et al. 2021). A comparative examination reveals that the descriptor- and graph-based techniques perform similarly. While deepGraphh identified a subset of human or microbial metabolites with potential BBB permeability, rigorous experimental validation is necessary. This can be accomplished with static or monolayer-based BBB models, cone-plate BBB equipment, microfluidic-based, or stem cell (iPSC)-based BBB models (Bagchi et al. 2019). In addition, thorough experimental validations can be performed using in vivo methods such as intravenous injection, cerebral perfusion, positron emission tomography, and microdialysis sampling (Sloan et al. 2012). While the models offer high predictive accuracy, a key area for future development is model interpretability. The GAT and AFP models, for instance, use attention mechanisms that could theoretically be used to identify which atoms or substructures contribute most to a given prediction. However, a current limitation of the deepGraphh web service is the absence of a dedicated module to visualize these attention weights as heatmaps on the molecular structure.

While graph-based methods are clearly developing as a new way for chemical modeling, and their overall performance is comparable to or better than traditional machine learning-based methods, they do have some significant drawbacks. First, graph-based QSAR techniques often perform better on bigger datasets, but suffer significantly when the training dataset is constrained. Second, unlike typical machine learning approaches, graph-based algorithms are resource-intensive and require more computer capacity. Second, a quantitative benchmarking of specific runtime and memory constraints for very large datasets was not performed and remains an important practical consideration for users. Third, while graph-based approaches in chemoinformatics are recent additions, unlike descriptor-based machine learning methods, technological support for their implementation is very restricted. To fill this void, we created deepGraphh, a comprehensive web service for graph-based cheminformatics approaches. Furthermore, deepGraphh is one of the only open-source web servers available today that offers a variety of GUI-based choices for

graph-based QSAR research. Importantly, deepGraphh includes a raw model file in the output directory, as well as a script that enables users to apply their model for prediction analysis on subsequent datasets.

Chapter 3: To Design and Implement Algorithms for Tracking and Elucidating Ectopic Gene Expression Using Large-Scale scRNA-seq Data

3.1 Introduction

Single-cell ribonucleic acid (RNA) sequencing (scRNA-seq) is a novel technique for estimating gene expression, enabling analysis of the transcriptome with exceptional resolution (Hodzic 2016; Lähnemann et al. 2020; Hwang et al. 2018; Hao et al. 2024; Stuart & Satija 2019; Natarajan et al. 2019; Buettner et al. 2015). Unlike bulk or tissue-level RNA sequencing methods, scRNA-seq offers extensive information that may be used to elucidate the mechanisms of cellular identities, functionalities, or fates (Hodzic 2016; Lähnemann et al. 2020; Hwang et al. 2021; Kalra et al. 2020; Kalra et al. 2021). Advancements in single-cell capturing methods and barcoding technologies have facilitated the sequencing of hundreds to thousands of single cells, thereby enabling the construction of a high-resolution transcriptomic map of the respective tissues or organs (Hodzic 2016; Lähnemann et al. 2020; Hwang et al. 2018; Hao et al. 2024; Stuart & Satija 2019; Natarajan et al. 2019; Buettner et al. 2015). Additionally, it facilitates the identification and characterization of rare or marginalized cell types and their associated markers (Jindal et al. 2018; Gupta et al. 2021). The swift implementation and broad adoption of scRNA-seq technology have catalyzed the rapid advancement of chimeric add-on technologies, including high-throughput genome-wide CRISPR/Cas9-mediated genome editing succeeded by scRNA-seq (Datlinger et al. 2017; Dixit et al. 2016; Choi et al. 2020). The emergence of these technologies enhances the potential of scRNA-seq, offering researchers a vital method to elucidate complex biological mechanisms. The availability of computational tools for downstream analysis represents a significant bottleneck in extracting novel and meaningful information from single-cell datasets. Currently, numerous software solutions exist for the efficient, accurate, and rapid analysis of high-dimensional datasets. However, most focus primarily on improving clustering efficiencies, noise removal, or inferring cellular trajectories (Chen et al. 2019; Forcato et al. 2021; Stuart & Satija 2019). Furthermore, the rapid advancement of scRNA-seq technologies, along with the vast availability of datasets in public domains, necessitates concurrent efforts in computational development. This is essential to create frameworks that enable users with limited computational skills to effectively explore and analyze single-cell datasets. This work presents EcTracker, a web server based on R/Shiny that offers various functionalities for the identification, characterization, and network analysis of cell-type or tissue-specific gene signatures, as well as ectopic genes in scRNA-seq datasets. The term "ectopic" is operationally defined as gene expression that deviates from established physiological norms. The criteria for labeling expression as ectopic are not based on a single threshold but on a comparative framework where a gene or signature is

identified as ectopic if it shows statistically significant enrichment in a cell type where it is not expected to be expressed, based on comprehensive reference atlases. EcTracker provides cellular-level information (CellEnrich) and tissue-level information (TissueEnrich) by utilizing data from the recently developed human single-cell landscape, fetal single-cell atlas, Human Protein Atlas (HPA), and Genotype-Tissue Expression (GTEx) Portal. The implementation of Discriminant Regulon Expression Analysis (DoRothEA) in EcTracker facilitates the identification of key driver genes, specifically transcription factors, associated with these gene signatures derived from single-cell datasets. We reanalyzed a recently published Perturb-seq dataset to demonstrate the functionality of EcTracker, focusing on the differentiation of human embryonic stem cells (hESCs) into the endoderm lineage (Genga et al. 2019). EcTracker identified a significant enrichment of adult endothelial cell signatures indicative of endothelial to mesenchymal transition in the *SMAD2* knockout cells, whose identity was previously ambiguous in the original study. EcTracker offers users a comprehensive tool for single-cell analysis and a variety of cellular or tissue resolution gene signatures for enrichment analysis, providing a deeper understanding of cellular functionalities.

3.2 Methodology

EcTracker workflow includes widely used single-cell tools, including as the single-cell analysis toolbox for expression in R (scater) and other Bioconductor packages, the Seurat software suite, AUCell, TissueEnrich, DoRothEA, and VIPER (Satija et al. 2015; Aibar et al. 2017; Alvarez et al. 2016; Amezquita et al. 2020; Holland, Szalai, et al. 2020; Jain & Tuteja 2019). We used these packages because of their smooth performance and widespread popularity in the single-cell community. The following are the advantages of the main software packages implemented in the EcTracker: (1) The Seurat software package is a comprehensive, easy-to-implement, resource-rich, and commonly used approach for single-cell data analysis. Seurat version 3 supports initial QC, standard single-cell data analysis, and interpretation of cellular heterogeneity in a single-cell dataset. (2) The single-cell analysis toolbox for expression in R (scater)-based analysis is used as an alternative to the Seurat approach for standard single-cell data analysis. Two main differences between this technique and Seurat are processing speed and computer resource requirements. It enables for the normalization, grouping, and low-dimensional display of entire datasets. It provides for the computation of the DGE analysis using the pairwise Wilcoxon test (default), pairwise t-test, and pairwise binomial tests. (3) Harmony is an efficient, robust, quicker, and frequently used batch correction approach. It has numerous major features, including the capacity to scale big datasets and discriminate between various subpopulations. (4) AUCell: The use of AUCell in EcTracker enables the estimation of enrichment scores for cell-type-specific gene signature enrichment. In contrast to other approaches, AUCell is quite resilient to noise in both gene sets and

expression data. (5) TissueEnrich: it is provided as a Bioconductor package, making it easy to integrate in the web server. (6) DoRothEA: it comprises a precompiled list of interaction data that was manually picked from a variety of reliable sources.

3.2.1 Input datasets

EcTracker accommodates datasets from many single-cell technologies, including droplet-based and SMART-Seq-based methodologies (Zheng et al. 2017). Methodologically, it utilizes a raw read count matrix produced by any of the previously stated single-cell technologies, with genes and cells arranged in rows and columns, respectively. Furthermore, it accommodates the typical outputs of Cell Ranger, a dedicated software platform for the 10× Genomics dataset. Methodologically, it enables users to input three standard Cell Ranger output files: matrix.mtx, barcodes.tsv, and genes.tsv. It is important to note that the multidimensional single-cell transcriptomic datasets may exhibit batch-specific changes; thus, EcTracker enables users to execute batch correction using two distinct methodologies, primarily contingent upon the workflow chosen for initial data processing. The inclusion of batch correction is of critical importance as it addresses non-biological technical variation between samples that can arise from differences in experimental handling, reagents, or sequencing runs, ensuring that downstream biological comparisons are robust and accurate. Method 1 utilizes the scater R library for preliminary data processing, employing fast mutual nearest neighbors correction (fastMNN) (Haghverdi et al. 2018), a function integrated inside the batchelor Bioconductor package (version 1.6.3) (Haghverdi et al. 2018) for batch correction. The batchelor package possesses a significant advantage for batch correction, as it does not necessitate prior knowledge of the population structure. Method 2 employs the EcTracker workflow, based on the Seurat software suite, which facilitates batch correction with Harmony, a rapid and extensively utilized technique for batch correction. To do batch correction, it is essential to augment the metadata linked to the single-cell expression matrix. Users may manually choose any parameter specified in the metadata for batch elimination.

3.2.2 Processing of single-cell data

EcTracker offers two distinct methodologies for the preliminary processing of scRNA-seq data: the scater-based approach (Method 1) (Amezquita et al. 2020) and the Seurat-based approach (Method 2) (Stuart et al. 2019). Method 1 provides a relatively swift and resource-efficient analysis, while Method 2 conducts a thorough analysis and presents more intermediate steps; consequently, it requires greater processing resources and has an extended runtime. In Method 1 workflow, users can execute initial data

processing processes, including normalization, clustering, and, if necessary, batch correction. Method 1 for differential gene expression (DGE) analysis offers three statistical tests: the pairwise Wilcoxon test (default), the pairwise t-test, and the pairwise binomial test. In contrast, Method 2 provides a broader array of options, including the non-parametric Wilcoxon rank-sum test (default), bimodal analysis, receiver operating characteristic (ROC) analysis, Student's t-test, negative binomial generalized linear model, Poisson generalized linear model, logistic regression (LR), MAST (Finak et al. 2015), and DESeq2 (Love et al. 2014). Besides the previously stated batch-uncorrected DGE analysis approaches, EcTracker additionally facilitates the modeling of batch effects during the computation of differentially expressed genes. The workflow is advanced through the application of the hurdle model from the MAST package (Finak et al. 2015). This feature is located under a distinct tab, allowing users to compute differentially expressed genes with or without batch correction methods, or both. The graphical user interface now enables users to choose the 'condition for DGE analysis,' 'selection of reference level,' and 'choose parameter for coefficient hypothesis,' which it automatically retrieves from the user-provided information. Additionally, users may modify the additional parameters necessary for executing the hurdle model, namely 'frequency of expressed genes', 'fold change threshold (logarithmic scale, base 2)', and 'false discovery rate'. Regardless of the previously indicated methodologies, DGE analysis facilitates the identification of markers peculiar to clusters. Significantly, EcTracker facilitates configurable and straightforward inter-cluster differential gene expression profiling. The web server contains a slider bar for selecting personalized fold change threshold settings. The default fold change threshold value is set to 1 (logarithmic scale, base 2). Regardless of the previously described techniques, for the preliminary scRNA-seq data analysis, users may manually provide parameters to establish the cut-off for the minimum cell count (range: 3–100), decide whether to include or exclude the spike-ins in their datasets, and filter out cells with insufficient sequencing depth. The minimum sequencing depth is established at the commonly advised threshold of 20,000 reads per cell, with an upper limit of 100,000 reads per cell.

3.2.2.1 Method 1 (based on the single-cell analysis tools for expression in R, scater)

It constitutes one of the two alternatives for the preliminary scRNA-seq data processing phase. It is founded on a single-cell analysis toolset for expression in R, specifically utilizing scater-based R libraries. Method 1 first generates a SingleCellExperiment object and subsequently facilitates downstream data processing, which includes initial quality control (QC) through the filtration of low-quality cells based on mitochondrial transcript expression, log normalization, identification of variable genes, and additional analyses such as principal component analysis (PCA) and clustering. The results are illustrated graphically through t-distributed stochastic neighbor embedding (t-SNE) and silhouette plots, with an

option for download in portable document format (PDF). Notably, to eliminate batch-induced variability from single-cell data, Method 1 utilizes fastMNN, a method that connects disparate batches by discovering mutual closest neighbors (MNNs).

3.2.2.2 Method 2 (Utilizing Seurat)

Method 2 of EcTracker first generates a Seurat object, subsequently identifies variable genes, and conducts various downstream analyses, including PCA and initial quality control, which entails filtering out low-quality cells, among others. Additionally, it provides graphical representations of essential outcomes, including JackStrawPlot, elbow plot, and uniform manifold approximation and projection (UMAP) (Stuart et al. 2019; Satija et al. 2015). All these graphs and their corresponding tables are available for download in PDF and comma-separated values (CSV) formats, respectively. EcTracker facilitates the selection and filtration of cells according to the expression levels of mitochondrial genes for quality assessment. The cell-filtering stage is extremely adjustable, allowing users to manually adjust the quality control filtering criteria, including the minimum and maximum nFeature_RNAs and the maximum percentage of mitochondrial content through the graphical user interface. Following filtering, EcTracker does data normalization, scaling, and the identification of highly variable features using default Seurat parameters (Stuart et al. 2019; Satija et al. 2015; Yao et al. 2012). EcTracker employs LogNormalize for data normalization, a standard built-in normalizing function of the Seurat package. This step is essential for analyzing noisy single-cell data that frequently have a non-normal distribution. Specifically, the LogNormalize function is employed to address the challenges of data sparsity and high technical noise (dropout) that are common in scRNA-seq, stabilizing variance across genes with different expression levels. Log normalization is a global scaling strategy that standardizes the expression of each gene across all cells by the total expression, multiplied by a default scale factor of 10,000. It additionally applies a logarithmic transformation to the resulting matrix. Subsequently, to identify highly variable genes among cells, EcTracker calculates mean-variance connections. Significantly, following data scaling, EcTracker does dimensionality reduction with the application of PCA. Moreover, EcTracker generates several graphs, like DimHeatmap, facilitating immediate and simplified exploration of heterogeneity sources within the user dataset, which is crucial for identifying principal components for subsequent analysis (Stuart et al. 2019; Satija et al. 2015; Yao et al. 2012). Furthermore, the JackStrawPlot function facilitates a comparison between the distribution of P-values for each principal component and a uniform distribution (dashed line). This procedure utilizes the Harmony algorithm (Korsunsky et al. 2019) for the integration of single-cell datasets from several sources or batches, providing a rapid, sensitive, and precise method for batch correction.

3.2.3 Submitting Metadata

EcTracker facilitates data interpretation and batch correction by enabling users to submit a metadata file that includes details on cell-type IDs, batch information, replication data, and other variables. EcTracker amalgamates this metadata with the primary table and superimposes it for subsequent analyses. Submission of metadata is necessary during the batch correction phase.

3.2.4 CellEnrich

Recently, extensive atlas-level single-cell transcriptomics has been conducted on fetal and adult human tissues. EcTracker employs an extensive compilation of cell-type-specific markers (cell-type signatures), sourced from these abundant resources, and facilitates the calculation of their enrichment in the user-enhanced scRNA-seq dataset through two distinct statistical methodologies: AUCell (Aibar et al. 2017) and Stouffer's score-based enrichment analysis (SBEA) (Gupta et al. 2021). Both methods are included to provide users with analytical flexibility and the ability to cross-validate findings; obtaining consistent results from two different statistical approaches can increase confidence in the biological interpretation. While the tool does not enforce the use of one over the other, AUCell has been documented as being particularly resilient to noise in both gene sets and expression data, which may make it a preferable first choice for datasets of unknown or variable quality. AUCell is a prevalent software utilized for calculating the enrichment of specified gene sets within a single-cell dataset. AUCell generates numerous qualitative and quantitative charts that depict the enrichment of user-defined gene signatures (Aibar et al. 2017). EcTracker generates a histogram for the ranking and area under the curve (AUC) of the chosen signatures. Additionally, it provides a UMAP illustrating all cells that exceed the threshold for the chosen enrichment. Methodologically, AUCell receives gene sets as input and produces gene set activity for each cell as output (Aibar et al. 2017; Holland, Tanevski, et al. 2020). EcTracker additionally facilitates enrichment analysis with an alternative methodology, SBEA. The supplied read-count expression values are log2-transformed and normalized following the addition of 1 as a pseudo-count. Furthermore, a second round of cell filtering is conducted to preserve cells that express a minimum of 10% of genes within a designated gene set signature. Log-normalized expression estimations for each gene are transformed into Z-scores. A composite Stouffer's Z-score is calculated for each cell as follows:

$$Z \sim \frac{\sum_{i=1}^{N} Z_i}{N}$$

where Zi denotes the cell-specific Z-score corresponding to the ith gene and N denotes the number of genes common between the expression and signatures (Gupta et al. 2021).

3.2.5 TissueEnrich

Like CellEnrich, EcTracker facilitates the implementation of TissueEnrich(Jain & Tuteja 2019), a Bioconductor program designed for assessing tissue-specific signals within datasets. Besides the standard method for enrichment calculation, namely the hypergeometric test (Jain & Tuteja 2019), EcTracker also facilitates enrichment analysis via the SBEA approach (Gupta et al. 2021). The hypergeometric test enables enrichment analysis at the cluster level, allowing users to manually select clusters, while EcTracker facilitates the execution of TissueEnrich analysis on the calculated markers (differentially expressed genes) of such clusters. EcTracker generates a summary table that displays the enrichment scores for eligible tissues and facilitates the presentation of enriched genes on UMAP. TissueEnrich utilizes two principal datasets, HPA (Uhlén et al. 2015) and GTEx (GTEx Consortium 2020), to delineate tissue-specific gene signatures.

The hypergeometric test is employed to assess tissue-specific gene enrichment. The P-value (P) is computed as:

$$P(X > k) = \sum_{i=k+1}^{n} \frac{\binom{K}{i}\binom{N-k}{n-i}}{\binom{N}{n}}$$

and the fold change is calculated as:

$$Fold\ Change = \binom{k}{n} / \binom{K}{N}$$

where N represents the overall number of genes, K denotes the total number of tissue-specific genes for a particular tissue, n signify the number of genes in the input gene set, and k indicates the number of tissue-specific genes within the input gene set. The P-values were adjusted for multiple hypothesis testing employing the Benjamini and Hochberg adjustment (Jain & Tutej, 2019).

3.2.6 Genetic regulatory network

EcTracker facilitates the investigation of gene regulatory networks (GRNs) for differentially expressed genes with user-selected gene signatures. This study can be conducted as a continuation of either CellEnrich or TissueEnrich. To accomplish this, users must designate a cluster, and EcTracker conducts the GRN analysis by synthesizing the knowledge base of established transcription factors and target gene data. GRN results can be represented through two separate interactive graphical methods: (1) transcription factor–gene interaction and (2) gene-signature interaction. GRN analysis offers essential functional and mechanistic insights into the mode of regulation (MOR) of significant genes by employing the user's single-cell expression dataset (Holland, Szalai, et al. 2020; Holland, Tanevski, et al. 2020). Significantly, the incorporation of GRN analysis in EcTracker is facilitated by the DoRothEA package, a reputable resource for transcription factor targets (regulons) that can function as gene sets for inferring transcription factor activity. Additionally, DoRothEA offers insights into the MOR between transcription factors and target interactions (Holland, Szalai, et al. 2020; Holland, Tanevski, et al. 2020). Notably, DoRothEA transcription factors are assigned an empirical confidence rating that reflects the reliability of their regulons. It spans from A (highest confidence) to E (lowest confidence). EcTracker employs confidence levels ranging from A to C to get rigorous results (Holland, Szalai, et al. 2020; Holland, Tanevski, et al. 2020). The transcriptional factor activity in EcTracker is inferred using the VIPER program, a statistical framework designed to estimate protein activity from gene expression data (Alvarez et al. 2016). Significantly, all plots and tables produced by the EcTracker are available for download on the analysis page.

The Perturb-seq dataset of human embryonic stem cells developed into the endoderm lineage was obtained from the Gene Expression Omnibus under accession number GSE127202 (Genga et al. 2019). Perturb-Seq is a cutting-edge genomic technique that combines genetic perturbation with single-cell RNA sequencing to reveal the functions of genes on a massive scale. The expression files of both replicates were retrieved and amalgamated using the integrated capability of the Seurat software suite. The sample dataset in EcTracker was obtained from the Gene Expression Omnibus under accession number GSE73122.

3.2.7 CellEnrich gene sets

Gene signatures for CellEnrich were obtained from the benchmarked datasets, specifically the Human Cell Landscape (Han et al. 2020; Cao et al. 2020; Uhlén et al. 2015; GTEx Consortium 2020; Holland,

Szalai, et al. 2020) and the Human Fetal Cell Atlas (Han et al. 2020; Cao et al. 2020; Uhlén et al. 2015; GTEx Consortium 2020; Holland, Szalai, et al. 2020). The TissueEnrich datasets, specifically the HPA dataset (Han et al. 2020; Cao et al. 2020; Uhlén et al. 2015; GTEx Consortium 2020; Holland, Szalai, et al. 2020) and GTEx datasets (Han et al. 2020; Cao et al. 2020; Uhlén et al. 2015; GTEx Consortium 2020; Holland, Szalai, et al. 2020), were acquired in the Gene Matrix Transposed file format (*.gmt), alongside various gene signatures including ectodermal cell differentiation, ectodermal development, embryonic stem cell, endodermal cell differentiation, endodermal development, fibroblast migration, fibroblast proliferation, mesoderm development, mesodermal cell differentiation, developing heart atrial cardiomyocytes, and developing heart ventricular cardiomyocytes from the Molecular Signatures Database v7.2 (Subramanian et al. 2005).

3.2.8 User Interface Design

Shiny built-in functions are utilized for the front-end design of the EcTracker. The often-utilized functions comprise shinyUI, tabPanel, navbarPage, tabsetPanels, and actionButton. Significantly, the output GUI employs the following functions: verbatimTextOutput, plotOutput, and uiOutput. HTML functions and JavaScript ShinyJS are employed to enhance the interactivity of our web server.

3.2.9 Shiny optimization

EcTracker utilizes many ways to enhance its efficiency and deliver a seamless experience to consumers. The per-step idle time is established at 60 minutes to facilitate the analysis of extensive datasets and to prevent reload issues. Additionally, memoization techniques are employed to enhance the speed and efficiency of EcTracker through improved cache management.

3.3 Results

3.3.1 EcTracker: computational methods for accurately identifying ectopic gene expression patterns from large-scale single-cell RNA sequencing (scRNA-seq) data.

EcTracker is a web server based on R/Shiny that offers a range of established and innovative features for scRNA-seq data analysis **(Figure 3.1, 3.2)**. EcTracker conducts

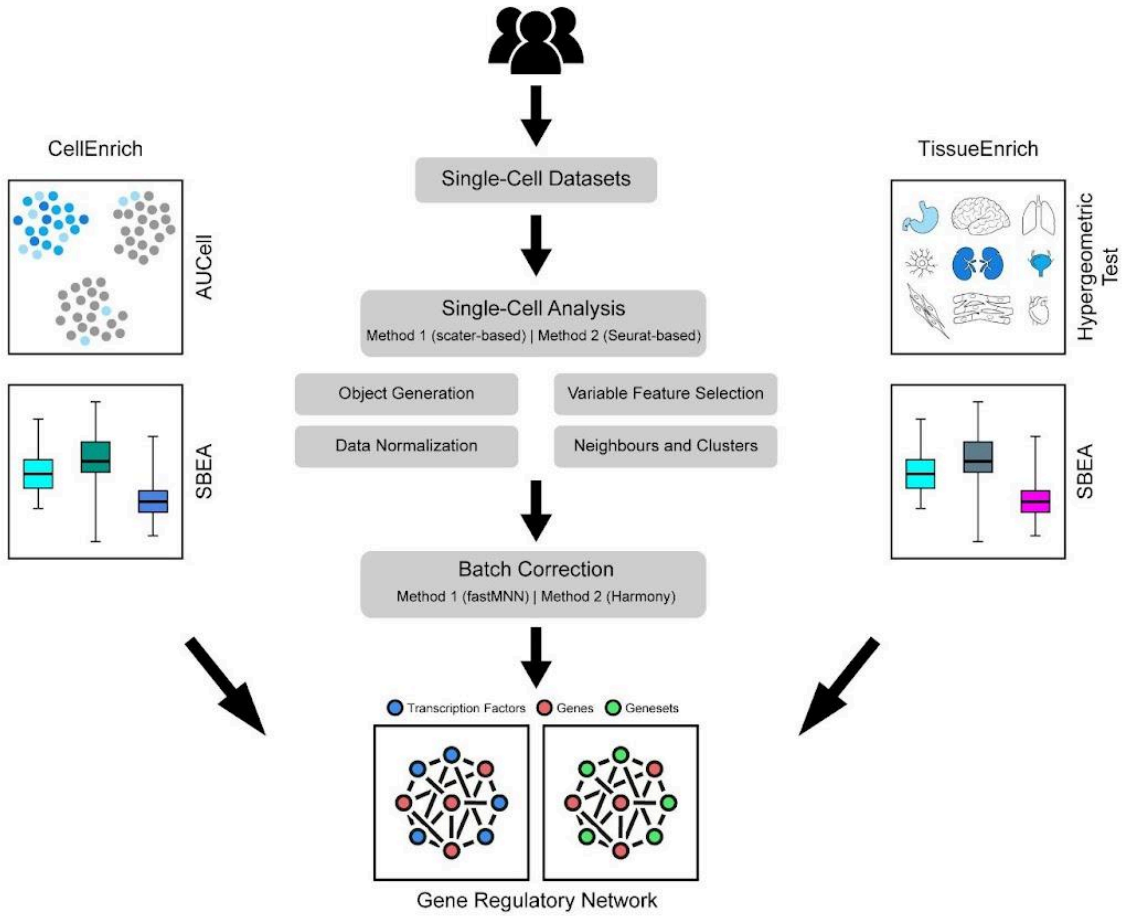


Figure 3.1: EcTracker, a web-based solution for comprehensive analysis of ectopic transcripts in single-cell datasets.

Graphical representation highlighting the key steps of the computational workflow implemented in EcTracker. EcTracker supports expression matrix from Drop-seq as well as Cell Ranger output files. Initial scRNA-seq data processing can be performed using either of the two workflows, annotated as Method 1 and 2. EcTracker supports enrichment analysis of the predefined gene signature at the cellular (CellEnrich) and tissue (TissueEnrich) levels. To gain mechanistic insights involved in the regulation of ectopic gene expression, EcTracker supports Gene Regulatory Network analysis and returns a detailed relationship between query genes within a gene signature and transcription factors.

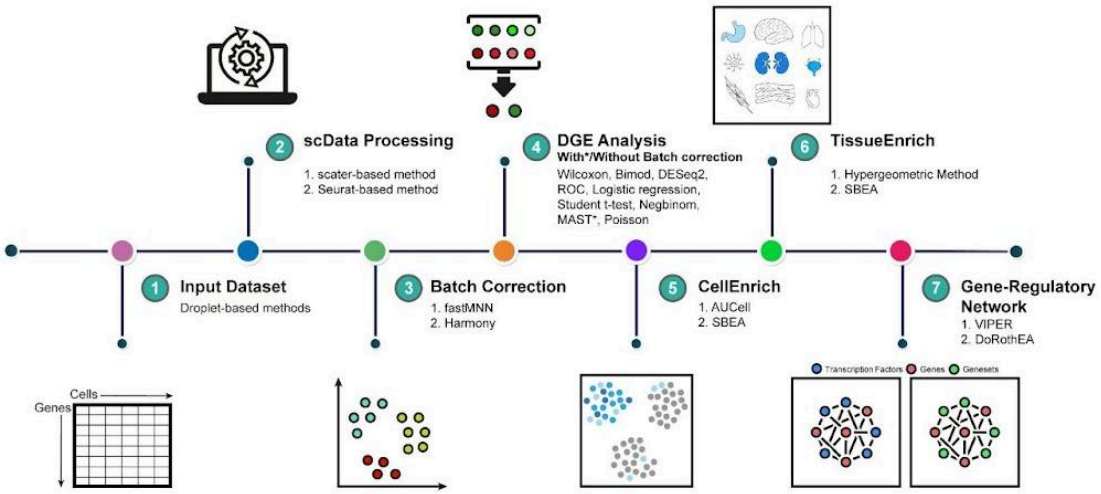


Figure 3.2: Graphical representation depicting the functional nodes of EcTracker.

Standard scRNA-seq analysis on the raw expression matrix to establish the foundation for executing its core functionalities. This tool accommodates single-cell datasets derived from prevalent scRNA-seq technologies; further information can be found in the Methodology section. Users may upload either a single expression matrix with read count data or the three standard output files from Cell Ranger **(Figure 3.2, 3.3)**. The standard analysis encompasses data normalization, quality control, clustering, identification of marker genes, dimensionality reduction, and additional processes. The aforementioned steps are accomplished through either Method 1, which utilizes the single-cell analysis toolkit for expression in R (scater) (Amezquita et al. 2020), or Method 2, which is based on Seurat (Stuart et al. 2019), for initial data processing. Users can select from the two methods for each run; the first method is faster but has limited analysis parameters, while the second method is more comprehensive yet requires more time. EcTracker enables users to conduct batch correction, though this step is not mandatory. It provides two distinct, widely utilized, and resource-efficient methods for batch correction: fastMNN and Harmony, which are implemented in Methods 1 and 2 of EcTracker, respectively. Submission of metadata is required for the batch correction step. Users may select or modify parameters for batch correction according to metadata. EcTracker includes an explanation tab for each graphical and tabular output, offering interpretative details regarding the displayed plot or table. EcTracker facilitates DGE analysis by enabling users to set expression thresholds and download results in a tabular format. This format includes details on the identities of differentially expressed genes, expression fold change, and the P-value. EcTracker provides support for various DGE analysis methods, including the pairwise Wilcoxon test (default), pairwise

Student's t-test, pairwise binomial test, bimodal analysis, ROC analysis, negative binomial generalized linear model, Poisson generalized linear model, LR, MAST (Finak et al. 2015), and DESeq2 (Love et al. 2014). EcTracker also facilitates batch-corrected DGE analysis, enabling the modeling of batch effects through the hurdle model of the MAST package while identifying differentially expressed genes (Finak et al. 2015). During each run, a user may choose from the previously mentioned methods; however, the Wilcoxon test serves as the default method for the batch-uncorrected approach.

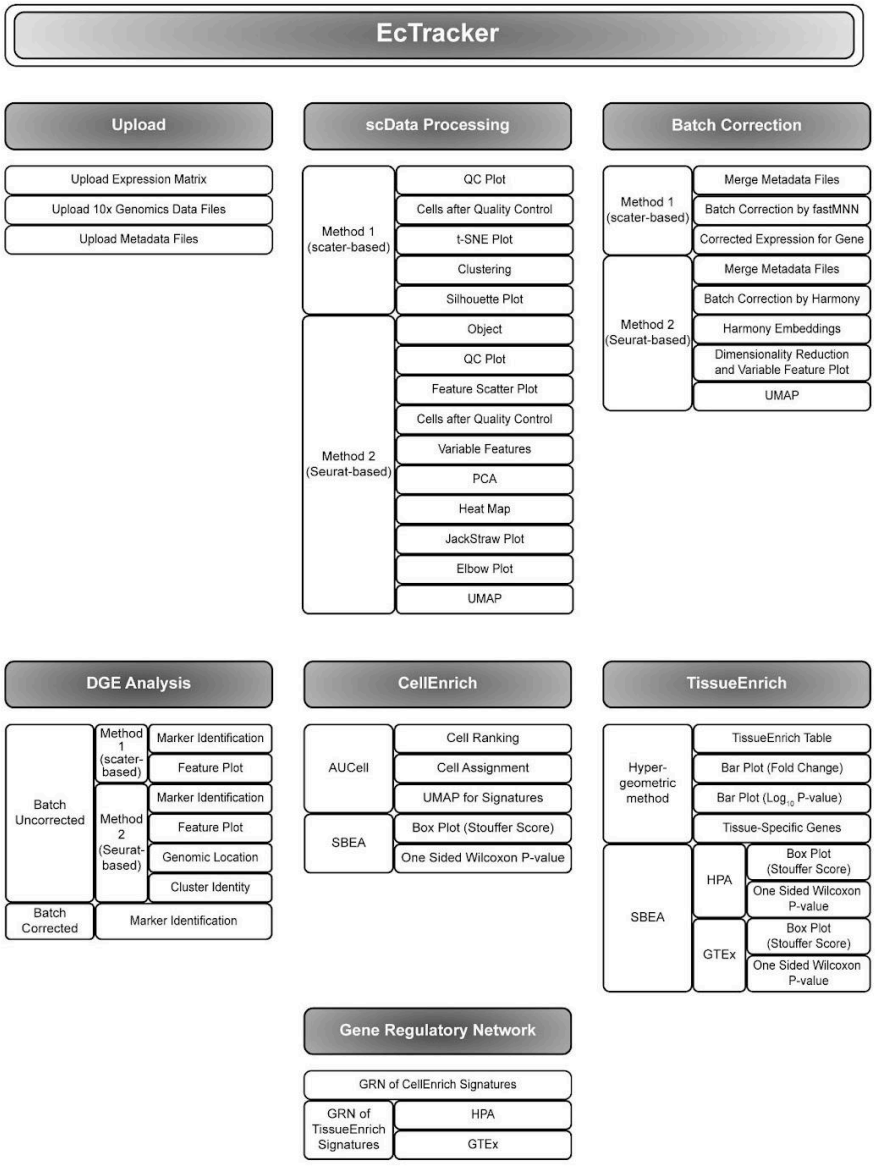


Figure 3.3: Schematic diagram depicting the functional architecture of EcTracker.

3.3.2 Identification of ectopic or cell- and tissue-specific genes

EcTracker employs two distinct approaches for the cluster-wise identification of ectopic transcripts in the single-cell expression dataset. In the initial approach (CellEnrich), EcTracker facilitates the identification of cells and their ectopic transcripts by utilizing cell-type-specific marker information sourced from the extensive Human Cell Landscape (Han et al. 2020)and Fetal Cell Atlas (Cao et al. 2020). Users can visualize cellular expression of precompiled cell or tissue type gene signatures, as well as ectopic genes in their datasets, utilizing the widely adopted AUCell (AUC) (Aibar et al. 2017) and Stouffer's SBEA (Gupta et al. 2021). AUCell is a commonly utilized R package for identifying cells in single-cell datasets that express specific gene signatures. EcTracker employs a one-sided Wilcoxon rank-sum test to assess the enrichment and significance of cell-type-specific signatures in the context of SBEA. Besides CellEnrich, EcTracker employs TissueEnrich (Jain & Tuteja 2019) as an alternative method for calculating tissue-level geneset enrichment in the dataset provided by the user. EcTracker utilizes TissueEnrich, a prominent computational package that employs GTEx and HPA datasets (Uhlén et al. 2015) to enrich tissue-specific gene signatures within the user dataset. Users may choose to utilize the default graphical outputs from TissueEnrich or SBEA for the visualization of signature enrichment, presented as box plots.

3.3.3 Ectopic gene chromosomal bias

EcTracker includes a built-in visualization method for graphically representing the chromosomal locations of user-selected transcripts, typically differentially expressed genes, in a cluster-wise format (Gel & Serra 2017). This feature enables users to assess loci-associated biases for ectopic transcripts, potentially offering mechanistic genomic insights.

3.3.4 Connecting ectopic genes with transcription factors

EcTracker employs DoRothEA, a comprehensive computational framework for regulon identification, to elucidate the mechanisms underlying ectopic expression induction (Holland, Szalai, et al. 2020; Holland, Tanevski, et al. 2020). The integration of DoRothEA with the statistical method VIPER enables the identification of key transcription factors influencing the expression of ectopic transcripts in user datasets. EcTracker produces output in the form of an interactive network graph, along with raw data points that include interaction information, confidence levels, and their MOR, available as a downloadable table.

3.3.5 Case study: EcTracker reveals the cellular identity of *SMAD2* knockout hESCs during endoderm differentiation

We reanalyzed a publicly available single-cell perturb-seq dataset to evaluate the efficacy of EcTracker in identifying cellular identities. Human embryonic stem cells (hESCs) were subjected to treatment with lentiviral particles that contained CRISPR guide RNAs targeting transcription factors identified as potential key drivers of endoderm development (Genga et al. 2019). Post-infection, cells were differentiated into the endoderm lineage utilizing established in vitro protocols. Scrambled guide RNA (gRNA) served as a control (Genga et al. 2019) **(Figure 3.4A)**. The expression files, comprising the expression matrix, cell barcodes, and gene information, were obtained from the Gene Expression Omnibus, specifically from 10× Genomics output. The Cell Ranger output files served as input for the standard single-cell analysis, which was conducted using the default functions. The dataset comprised 17,234 cells and 19,886 features/genes. The initial analysis identified 10 distinct clusters, as opposed to the four major clusters reported in the original study, indicating a greater clustering efficiency of EcTracker **(Figure 3.4B, 3.5)**. An inter-cluster comparison for the DGE was conducted. We collected and overlaid the metadata, including gRNAs linked to each cluster **(Figure 3.4C, D).** We conducted additional tests and observed a negative correlation between the gRNA and the expression of its target gene across all clusters. In the original study, the authors highlighted the ambiguous identity of the *SMAD2* loss-of-function cells. Consequently, we investigated whether EcTracker could provide insights into the potential identity of these cells. The AUCell feature in the CellEnrich module of EcTracker identified a significant enrichment of the adult endothelial cell signature, indicative of endothelial to mesenchymal transition, primarily in the *SMAD2* knockout cells **(Figure 3.4E)**. AUCell identified 1705 cells exhibiting an AUC value exceeding 0.25 **(Figure 3.7A, B)**. Some of the other signatures that showed preferably weaker or non-localized enrichment in the *SMAD2* knockout cluster are fetal lungs (10,946 cells, AUC value > 0.0091), fetal spleen (10,212 cells, AUC value > 0.0061), and fetal erythroblast (1355 cells, AUC value > 0.051). We subsequently validated the expression of several marker genes within this gene signature and noted the restricted expression of *ID1, ID3, HLA-B, HLA-C, RARRES2,* and *CCND1* in the *SMAD2* knockout cluster **(Figure 3.4F, 3.6A-F)**. The Stouffer's SBEA yielded comparable findings **(Figure 3.8A-C)**. Subsequently, we evaluated the TissueEnrich module using this dataset, which demonstrated enrichment for endoderm-derived tissues and organs **(Figure 3.8D)**. To elucidate the mechanism of adult endothelial cell signature activation related to endothelial to mesenchymal transition in *SMAD2* knockout cells, we conducted a GRN analysis utilizing EcTracker. Our results identified a set of transcription factors (*TEAD1, MYBL2, E2F4,* and *MYC*) that were potentially regulating the expression of adult endothelial cells (endothelial to mesenchymal transition) signature-specific genes, such as

CALD1 and *ID3* **(Figure 3.4G)**. EcTracker facilitated the elucidation of the cellular identity of *SMAD2* knockout cells within single-cell datasets.

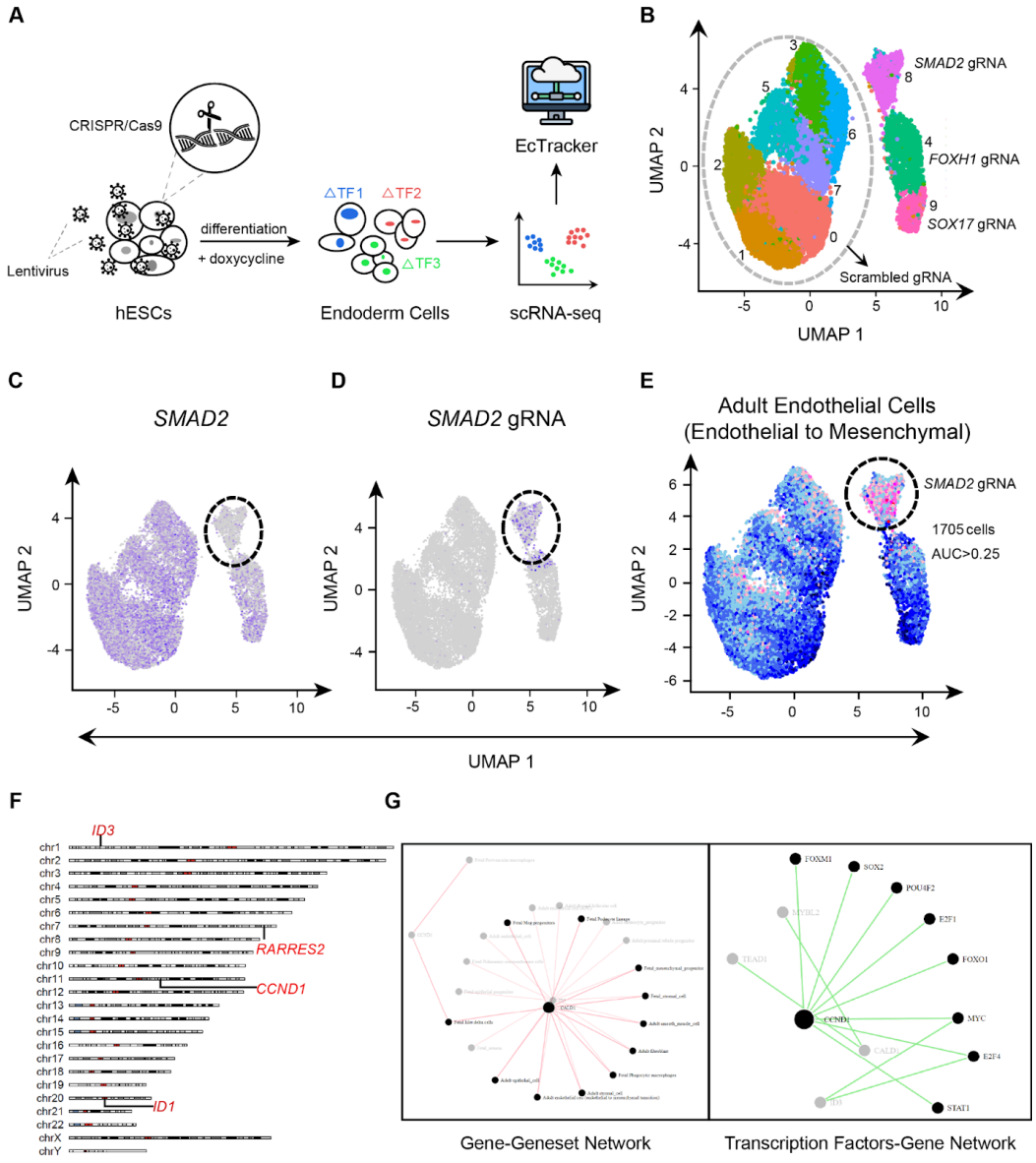


Figure 3.4: EcTracker unraveled the cell-type identity of SMAD2 knockout cells during endoderm differentiation of hESCs.

(A) *Schematic representation of the experimental workflow used in the case study. In brief, human embryonic stem cells were treated with lentiviruses containing guide RNAs against transcription factors, predicted to be the key drivers of endoderm lineage. Post-infection, the cells were activated and differentiated into the endoderm lineage. Post differentiation, the single-cell sequencing is performed, and both the gene expression as well as the enrichment of functional guide RNAs were calculated.* ***(B)*** *Uniform Manifold Approximation and Projection (UMAP) embedding of single-cell expression profiles represents the distinct cell types with indicated genomic interventions using the*

CRISPR/Cas9 genome editing method. ***(C-D)*** *UMAP-based embedding portrays the relative expression of indicated transcripts in the conditions.* ***(E)*** *UMAP-based embedding depicting the AUC scores computed using the AUCell package. Pink-red and black-blue color cells represent the subpopulations above and below the AUC threshold.* ***(F)*** *Karyogram depicting the chromosomal loci of the key gene markers for the SMAD2 knockout cells during endoderm differentiation.* ***(G)*** *Gene regulatory networks depicting the association between differentially enriched transcription factors and the Adult Endothelial Cells (Endothelial to Mesenchymal) gene signature.*

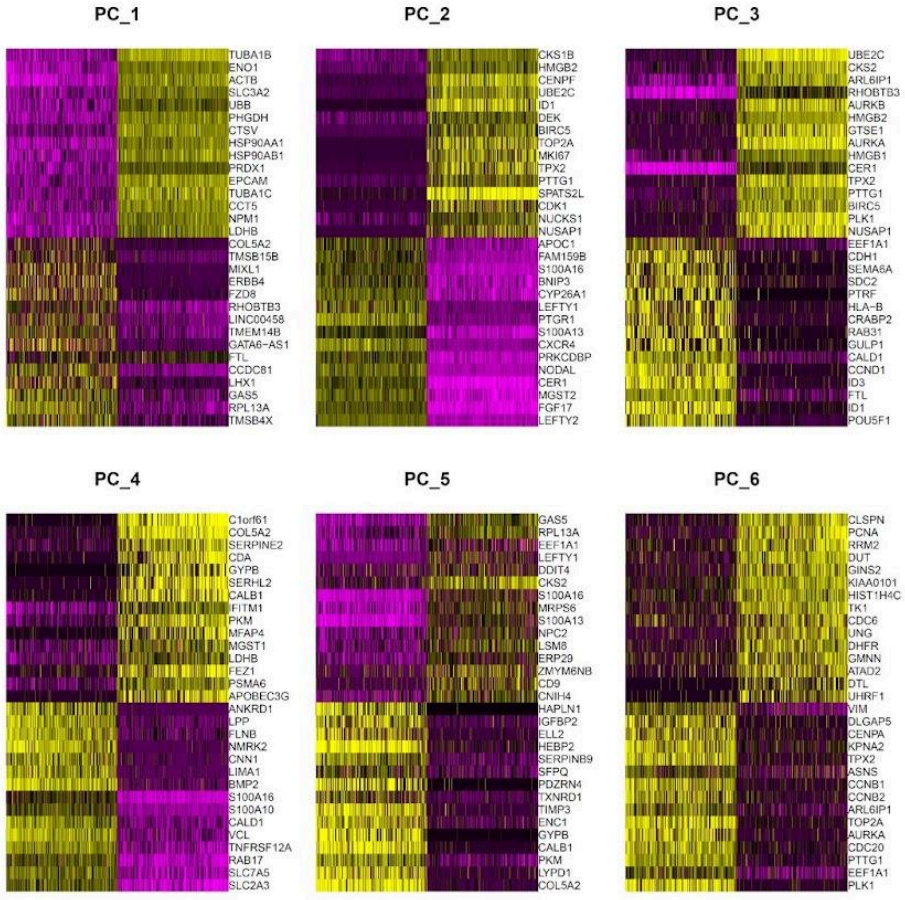


Figure 3.5: Heatmaps from Method 2 of EcTracker depicting the ordered cells and features across different principal components.

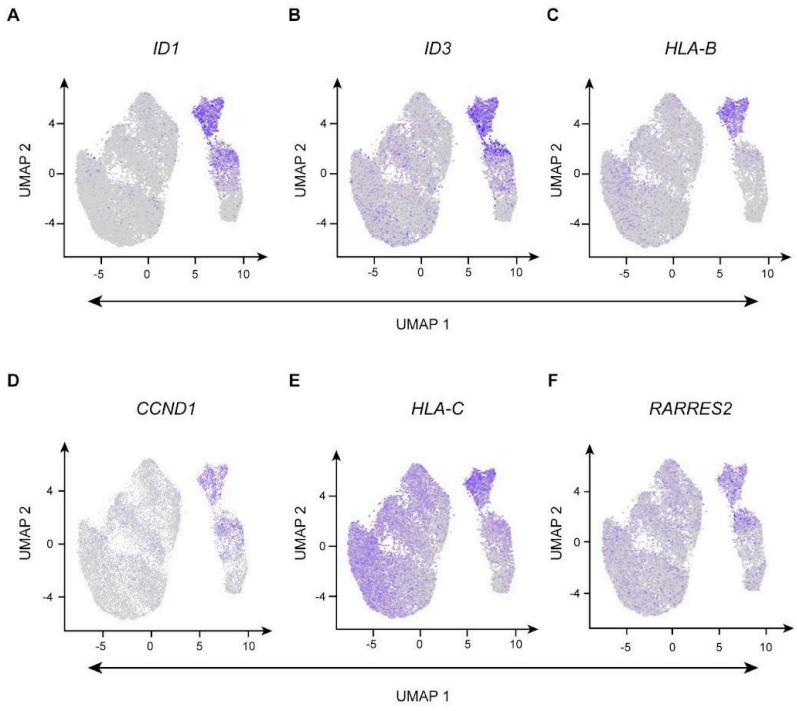


Figure 3.6: UMAP-based embedding portrays the relative expression of some of the selected genes of Adult Endothelial Cells (Endothelial to Mesenchymal) gene signature.

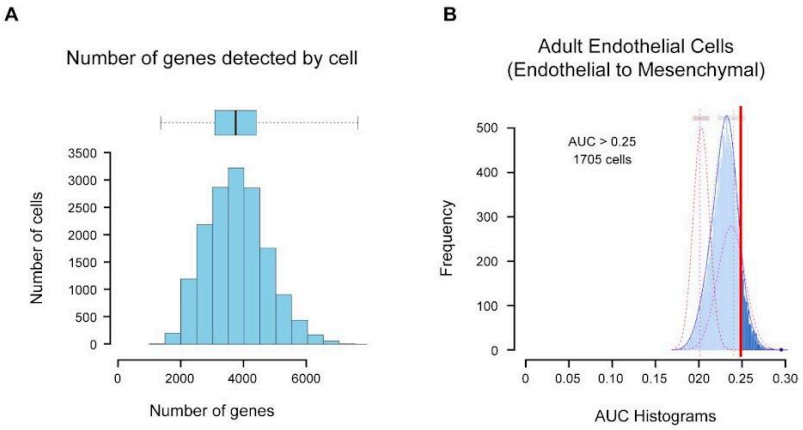


Figure 3.7: Distributions of Gene Expression and AUC Scores Across Cells

(A) *Histogram depicting the number of expressed genes across all cells in the user dataset that can be used to compute AUC scores.* ***(B)*** *AUC histogram depicting the distribution of the AUC scores of selected gene signatures across all cells. It also displays the number of cells that qualify the threshold (indicated as a thick vertical line) for the indicated gene signature.*

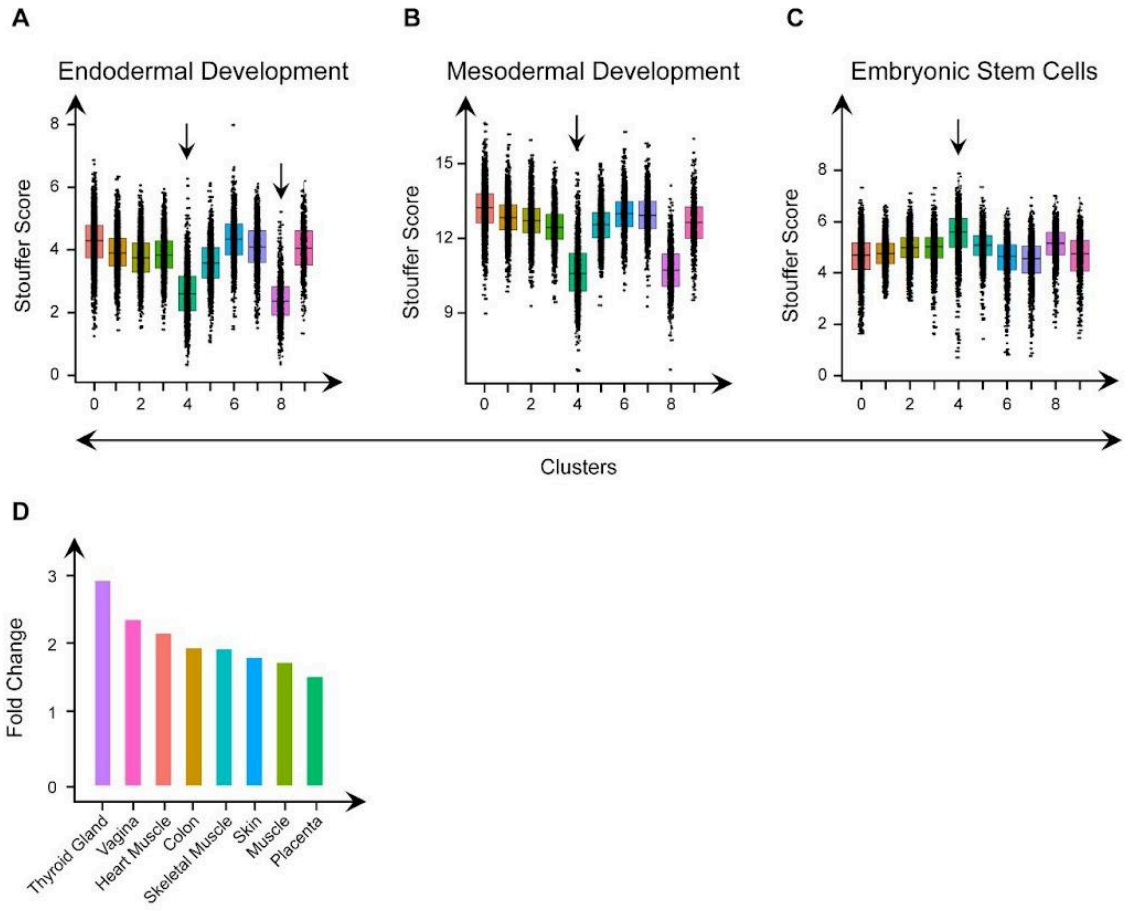


Figure 3.8: Stouffer's Scores and Tissue Signature Enrichment Across Clusters

(A-C) *Box plot depicting Stouffer's score of indicated gene signatures across all clusters. Arrowheads indicate the FOXH1 guide RNA cluster.* ***(D)*** *Bar plot depicting the enrichment (fold change) of indicated tissue signatures in the SMAD2 gRNA-positive cluster cells.*

3.4 Discussion

Rapid and widespread adoption of single-cell genomics techniques has resulted in a massive collection of genomics datasets (Kalra et al. 2020). Furthermore, it sparked much-needed help in the computational domain, which is critical for extracting useful information from complex single-cell datasets. Several attempts have been made to develop user-friendly computational workflows to accomplish the aforementioned task; however, the majority of them are limited to specific subdomains of single-cell data analysis, such as clustering efficiencies, noise removal, or inferring cellular trajectories (Gardeux et al. 2017; Ji et al. 2017; Patel 2018; Zhu et al. 2017; Feng et al. 2019). We provide EcTracker, a simple, web-based computational tool for performing multifactorial analysis on scRNA-seq datasets. The primary

innovation of EcTracker is not the development of novel statistical algorithms, but rather its function as a sophisticated workflow integrator. It seamlessly combines multiple powerful, existing R/Bioconductor packages into a single, cohesive, and accessible web-based platform, thereby lowering the barrier to entry for complex scRNA-seq analysis.

EcTracker offers a simple, web-based interface for doing in-depth analyses of single-cell datasets, including standard analysis, identification of ectopically expressed genes, and gene regulatory analysis. Although a significant number of online servers exist, including ASAP (Gardeux et al. 2017), SCRAT (Ji et al. 2017), iS-CellR (Patel 2018), Granatum (Zhu et al. 2017), Single Cell Explorer (Feng et al. 2019), and Alona (Franzén & Björkegren 2020), each is quite limited and mostly supports basic scRNA-seq analyses. SCRAT, iS-CellR, and Granatum are the only web server alternatives that give the native R/Shiny experience. While a formal quantitative benchmark against other tools in terms of speed or accuracy was not performed, EcTracker's value is demonstrated through its comprehensive feature set and its successful application in the case study, which yielded novel biological insights and a higher clustering efficiency than the original analysis. Furthermore, each of these web servers supports a number of statistical approaches for DGE analysis, which is substantially supported by the EcTracker itself. Importantly, EcTracker is the only tool that supports add-on approaches such as visualizing the genomic location of important genes and in-built gene regulatory analyses. Another notable aspect of the EcTracker is that it is accessible as both a web server and a stand-alone application in case users need to process larger datasets. EcTracker's GitHub website provides a full overview of the deployment on a local server or as a stand-alone version. In conclusion, EcTracker is more feature-rich and has numerous modules that enable the thorough estimation of expressed genes in single-cell datasets, and thus may provide an additional, intensive, and graphical user interface-based solution for scRNA-seq data analysis. The EcTracker offers support for multiple single-cell technologies (Drop-seq and 10× Genomics), as well as speed, interactivity, and comprehensiveness. It also allows for flexible threshold selection and the implementation of CellEnrich, TissueEnrich, and GRN modules for end-to-end data analysis. Despite its numerous advantages, EcTracker has several drawbacks. First, unlike other web servers that support numerous model organisms, EcTracker only accepts human samples, which could be owing to limitations related with the add-on methods. Second, EcTracker lacks pseudotime trajectory analysis capabilities; finally, it does not support pathway analysis modules such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) or other Gene Ontology modules. While the tool provides clear guidance on choosing between the faster Method 1 and the more comprehensive Method 2, a potential area for future enhancement would be the inclusion of a "best practices" guide. Such a guide could offer

recommendations on setting appropriate statistical thresholds for different parameters, helping users with diverse datasets to navigate the tool's flexibility more effectively.

In summary, EcTracker includes all of the modules required to extract relevant biological insights from the Perturb-seq and CROP-seq datasets (CRISPR droplet sequencing), which is a technique that integrates CRISPR-based genetic perturbations with single-cell RNA sequencing (scRNA-seq) to study how gene knockouts or knockdowns affect transcriptomes at single-cell resolution. EcTracker demonstrated the ability to identify the cellular identity of *SMAD2* knockout cells. EcTracker found that *SMAD2* mutant cells have an adult endothelial cell (endothelial to mesenchymal transition) hallmark. Notably, the endothelium to mesenchymal transition is a critical driver of pluripotent stem cell endoderm development into hepatocytes (Li et al. 2017). While any analysis based on inferred activity carries a risk of overinterpretation, the GRN analysis in EcTracker mitigates this by grounding its statistical findings in an established biological context. The identification of a plausible regulatory network that mechanistically explains the observed phenotype in the case study serves as a form of biological validation, reducing the likelihood of spurious conclusions. Despite having numerous advantages over other existing web servers for scRNA-seq analysis, one of EcTracker's main disadvantages is that it is species-specific, with gene signatures limited to human samples only. EcTracker, taken combined, provides the much-needed web-based software solution for identifying and characterizing unappreciated ectopic genes, as well as capabilities for deciphering cellular identities in single-cell datasets. It also enables the enrichment analysis of legitimate cell types and tissue markers in single-cell clusters. Finally, it allows users to undertake regulon analysis, which provides mechanistic insights into the GRNs.

Chapter 4: To Create a Context-Aware Prediction Framework for Aging-Associated Bioactivities and Morphometrics

4.1 Introduction

As cells mature, their chemical composition, phenotypic expression, and morphological characteristics undergo dynamic alterations over time (Phillip et al. 2017; Guo et al. 2022; López-Otín et al. 2023). Diverse methodologies have been utilized to ascertain cellular age and its related health, encompassing both conventional, gold-standard techniques such as genomic approaches and innovative technology like deep learning. Recently, single-cell genomics revealed the molecular alterations associated with aging in cells, elucidating their identity, molecular composition, and lineage on a pseudo-temporal scale (Zhu et al. 2023; Doshida et al. 2023; Nikopoulou et al. 2023). Regrettably, despite its potency, it permits only a picture of the molecular event and hinders reevaluation over an extended temporal scale. An additional standard method involves employing traditional genetic lineage tracing techniques (McKenna & Gagnon, 2019; Wu et al,. 2019). Although potent, these methodologies exhibit constraints that hinder their use and wider utility in aging research (Chen et al. 2022; Hsu 2015). The constraints encompass stochastic and combinatorial labeling, the inability to achieve high spatial and temporal resolution, the difficulty in recording intricate cellular dynamics, and the introduction of exogenous genetic components that may constrain native activity. Recent experiments have demonstrated the ability of deep learning approaches to autonomously recognize and learn age-related morphological traits using either face or cellular microscopic pictures. For instance, research has shown successful age classification from facial images (Garain et al. 2021), presented the Deep Learning-Based Senescence Scoring System by Morphology (Deep-SeSMo) for the identification of senescent cells based on morphological characteristics (Kusumoto et al. 2021), and highlighted the use of deep learning in precisely identifying senescent cells utilizing nuclear morphology or its related features exclusively (Heckenbach et al. 2022; Duran et al. 2024), with later advancements in phenotyping senescent mesenchymal stromal cells (Weber et al. 2023). Although there has been an increase in deep learning techniques for predicting organismal or cellular age and related phenotypes, current solutions are deficient in generalizability, aging-associated bioactivity prediction, and explainability modules, which are essential for enhancing predictions and aiding the development of experimental strategies for subsequent mechanism investigation.

The budding yeast Saccharomyces cerevisiae is an exemplary model organism for aging research. A pertinent and unanswered inquiry is whether the pro- or anti-longevity responses observed in various yeast mutants or through pharmacological stimulation are activated at distinct or same temporal stages

throughout aging. This inquiry is pertinent to mammalian cellular models in aging research, which have historically required numerous biochemical and phenotypic experiments at various chronological intervals, both of which are labor-intensive and time-consuming. A key inquiry is whether the spatiotemporal and morphometric alterations associated with aging are evolutionarily conserved, akin to molecular or pathway-level conservation, despite initial diverse cellular morphologies (Zimmermann et al. 2018).

To tackle these and other previously mentioned technical obstacles, we present scCamAge, a sophisticated transfer learning framework that employs spatiotemporal data from phase-contrast images for accurate single-cell yeast age prediction. For the purposes of this study, "cell age" is defined differently for the two model organisms to reflect distinct biological paradigms. For the budding yeast Saccharomyces cerevisiae, age is defined by its Chronological Lifespan (CLS), which measures the survival of a population of non-dividing cells over time in days. In contrast, for human dermal fibroblasts, age is defined as a distinct cellular state of senescence, which was induced experimentally and validated using the established biomarker Senescence-Associated β-Galactosidase (SA-β-gal) staining. The central premise of this work is that key morphological hallmarks of aging, such as cellular hypertrophy (increase in size), are evolutionarily conserved between these organisms, allowing a model trained on yeast CLS to recognize the state of senescence in human cells. scCamAge encompasses models for predicting aging-related bioactivity and provides concurrent estimations of morphometric parameters. scCamAge underwent extensive testing across many datasets employing pro-longevity pharmaceuticals, genetic loss-of-function mutants exhibiting unique longevity profiles, and stress-induced aging reactions. Finally, we evaluated the applicability of scCamAge on human datasets for senescence and saw a substantial improvement in performance compared to the base model, suggesting the evolutionary conservation of aging-related cellular morphometric and bioactivity characteristics. In summary, scCamAge, comprising the prediction engine, analysis code, and picture datasets, is available at no cost as Docker images and may be accessed online at https://the-ahuja-lab.shinyapps.io/scCamAge/.

4.2 Methodology

4.2.1 Human Cells and Yeast Strains

All model construction and iterative thermal pulsing studies were conducted on the Saccharomyces cerevisiae strain BY4741 (*MATa his3Δ leu2Δ0 met15Δ ura3Δ*). For validation purposes, we utilized yeast genetic knockouts of the *msw1Δ, alt2Δ, gsy2Δ, trx1Δ, sod2Δ, kgd1Δ, adh2Δ, cys4Δ, gre3Δ, ipt1Δ, met17Δ, met2Δ, prs3Δ, pho89Δ,* and *pdx3Δ* genes, based on the BY4741 strain (*MATa xxx::kanMX his3Δ leu2Δ*

ura3Δ0 met15Δ). Yeast cells were cultivated in yeast extract–peptone–dextrose (YPD) media (1% yeast extract, 2% peptone, and 2% dextrose) and Synthetic Complete (SC) Media (2% glucose, 20 mg/L histidine, 120 mg/L leucine, 20 mg/L methionine, 20 mg/L uracil, and 20 mg/L adenine) at 30 °C and 150 rpm.

4.2.2 Phase Contrast Microscopy

The yeast primary culture was cultivated in YPD medium at 30 °C for 18 hours to produce models and validation datasets. A secondary culture was established using the initial culture under uniform environmental conditions. In subsequent studies, 10 µL of the secondary culture was seeded into Deep-well plates holding 1 mL of SC media, which were then sealed with breathable membranes (Z380059, Sigma-Aldrich). Microscopic imaging sampling occurred on alternate days from day 2 to day 20. Cellular samples were concentrated using centrifugation, and 5 µL of the concentrated sample was deposited onto microscope slides for imaging. Microscopic images were obtained using a Nikon Eclipse Ci-L Fluorescence Microscope equipped with a phase contrast filter. Imaging was conducted at a magnification of 100X, with an exposure duration of 10 ms. All images for the primary analysis were acquired using this standardized microscope setup to ensure internal consistency; however, the study did not explicitly control for variations that would arise from using different microscope types or image resolutions, which remains a consideration for broader applicability.

4.3.3 Fluorescence Microscopy

Wild-type (WT) yeast cells were cultivated in SC media following the previously published protocol. Every two days, 100 µL aliquots were taken from five biological duplicates to assess cellular responses. Thereafter, cells were centrifuged at 6000 rpm for 3 minutes, and the resultant pellet was resuspended in 100 µL of fresh 1X phosphate-buffered saline (PBS) containing the designated working concentrations of fluorescent dyes: FM™ 4-64FX (F34653, Invitrogen) at 5 µM, H2DCFDA (D399, Invitrogen) at 10 µM, and TMRE (87917, Sigma) at 25 nM. Subsequent to the dye treatment, the cells were incubated in darkness at 30 °C for 30 minutes to facilitate dye absorption and labeling. Following incubation, cells were washed and resuspended in fresh 1X PBS to eliminate surplus color. Subsequently, 5 µL samples from each condition were collected and analyzed microscopically using a Nikon Eclipse Ci-L Fluorescence Microscope. An exposure period of 100 ms was utilized for imaging TMRE-labeled cells using the TRITC filter. H2DCFDA (FITC filter) and FM™ 4-64FX (TRITC filter) were photographed with a 2-second exposure to collect fluorescence signals.

4.2.4 Drug Treatments

Wild-type yeast cells were cultivated until secondary culture, then moved to SC media, where varying quantities of MG132 (M7449, Sigma), etoposide (E1383, Sigma), and azacytidine (34476, SRL) were administered to the cells. The cells were treated for 4 hours with azacytidine (10 mM and 30 mM), 1 hour with etoposide (1 µM and 25 µM), and 2 hours with MG132 (25 µM and 100 µM), respectively. Following incubation, phase contrast microscopic pictures of cells were captured at 100X magnification using the Nikon Eclipse Ci-L Fluorescence Microscope with a 10 ms exposure period.

4.2.5 Imaging of yeast genetic knock-outs and externally drug-treated cells

To assess the efficiency of the scCamAge model, we employed yeast genetic knockouts (KOs), which were cultivated as previously described. Nine pro-longevity and six anti-longevity phenotypic yeast genetic knockout strains were selected, and microscopic pictures were captured bi-daily from day 2 to day 20. Additionally, to ascertain whether scCamAge can detect the pro-longevity effects of established pro-longevity pharmaceuticals, metformin (PHR1084, Sigma) at a working concentration of 50 mM, spermine (91710, SRL) at a working concentration of 1 mM, and spermidine (17030, SRL) at a working concentration of 1 mM were administered to the cells under drug-treated conditions, with microscopic images captured every other day from day 2 to day 20.

4.2.6 scCamAge Framework Development

The development of scCamAge entails the compilation of a dataset utilizing phase-contrast photographs of yeast cells gathered from Day 2 to Day 20 (as previously detailed). We employed yeastspotter (Lu et al. 2019), a technique utilizing the Mask Region-based Convolutional Neural Network (MaskR-CNN) for the segmentation of yeast cells in micrographs. Mask R-CNN is an enhanced variant of Faster R-CNN employed for segmentation tasks, such as recognizing objects in an image and precisely delineating their boundaries. The Mask R-CNN methodology comprises several fundamental components. Initially, it retrieves information from the input image via a backbone convolutional neural network (CNN); thereafter, these features are directed to two concurrent subnetworks, one for object detection and the other for generating segmentation masks (He et al. 2017). Upon generating the masked picture, the contours image, specifically individual yeast micrographs, was retrieved utilizing thresholding techniques in conjunction with the masked photos. To compute the area and perimeter of the micrographs, the OpenCV (Open Source Computer Vision Library) routines cv2.contourArea() and cv2.contourPerimeter() were utilized. Following the extraction of all individual yeast micrographs, model construction was undertaken to develop scCamAge.

To develop a resilient scCamAge, we employed various transfer learning models and selected the optimal model for predicting the age of yeast cells. Four primary models were evaluated to develop a robust scCamAge model. VGG19 is an enhancement of VGG16, increasing the depth from 16 to 19 layers in the deep convolutional neural network; these layers denote the weighted or convolutional layers, therefore the designation VGG19. VGG19 processes images of size 224×224 and use small 3×3 convolutional filters to extract information in both horizontal and vertical directions. ResNet50 is a formidable convolutional neural network of 50 layers, which includes 48 convolutional layers, one MaxPool layer, and one average pooling layer. A primary characteristic is the skip connection, which facilitates the retention of information from preceding layers, hence enhancing the representation of input data by incorporating the output of an earlier layer into a subsequent layer (He et al. 2016). GoogLeNet (vision:v0.10.0, googlenet), also referred to as the Inception Network, was created by researchers at Google. The network comprises 22 layers, making it computationally more efficient. It employs an inception module featuring numerous parallel branches utilizing various convolutional filters (1x1, 3x3, 5x5) alongside a max-pooling layer. The output from these branches is linked to the subsequent layer. It comprises two auxiliary classifiers linked to the outputs of Inception (4a) and Inception (4d). These classifiers are employed during training to enhance model performance. Ultimately, it incorporates global average pooling, which mitigates the gradient vanishing issue and decreases the number of trainable parameters by 25. The chosen model was InceptionResNet (vision:v0.10.0, inception_v3), an enhanced convolutional neural network architecture that surpasses its predecessors in multiple aspects. Its renowned precision is designed for micrograph classification tasks. It employs label smoothing, a regularization technique that mitigates model overfitting by softening the labels during training, hence diminishing the confidence of the model's predictions. The 7 x 7 convolution is factorized into two 3 x 3 convolutions, enhancing computing efficiency. To improve convergence and mitigate the vanishing gradient problem in deep networks, the auxiliary classifier is utilized to transmit label information to lower layers of the network. The network employs batch normalization in the sidehead, enhancing the model's stability and convergence during training (Szegedy et al. 2016).

The resulting model's robustness was guaranteed through comprehensive hyperparameter optimization. The dataset was partitioned into a training, validation, and testing split of 70%, 10%, and 20%, respectively. The model underwent training for 150 epochs during the tuning process to ascertain the ideal learning rate and the most efficacious optimization strategy **(Table 5)**. We conducted hyperparameter tuning by testing learning rates of 0.01, 0.001, and 0.0001, with batch sizes of 32 and 64. The ultimate hyperparameter configurations were chosen for their capacity to enhance model performance and

generalizability, hence ensuring a strong and dependable model. Two optimization strategies were employed to achieve the optimal fit for the dataset. Stochastic Gradient Descent (SGD) is employed in deep learning to update network parameters according to the gradients of the loss function relative to those parameters. The Adam Optimizer is an enhancement of Stochastic Gradient Descent (SGD) that employs the moving averages of the first and second moments of gradients to adjust the learning rates for each parameter. Notably, standard data augmentation techniques such as random rotations were used, while contrast jittering was not employed in this study.

4.2.7 Model Building and Testing

The dataset was divided into training data (80%) and testing data (20%) for model development. All models underwent hyperparameter optimization, and the optimal model was chosen. Several strategies were employed to avoid overfitting, especially given the small size of the human fibroblast datasets. The primary strategy was the use of transfer learning from a pre-trained InceptionResNet model, which leverages knowledge from a massive image dataset to improve performance and generalization on smaller, specific datasets. The model architecture itself includes built-in regularization techniques such as label smoothing and batch normalization to prevent overconfidence and improve training stability. To verify robustness, 10-fold cross-validation was conducted on the ultimately selected hyperparameters. The One-vs-All (OvA) technique was employed to assess the AUC-ROC of the scCamAge model. This method entails training a binary classifier, designating one class as positive while categorizing all other classes as negative for each respective class. Subsequently, we calculate the AUC ROC curve for each binary classifier. The aggregate AUC ROC curve for the multi-class model is derived by averaging the AUC ROC curves of each binary classifier.

Features were taken from the scCamAge model to develop the regression and classification models. The identical division of data was employed for both training and testing (80-20%), utilizing various models, including random forest, a supervised learning approach, and bagging methodology that applies ensemble learning for regression models(Ho 2002), Multi-Layer Perceptron (MLP) Regressor: a neural network composed of numerous layers of nodes interconnected with the following nodes (Chi 2002), XGBRegressor; a component of the XGBoost library, known for its efficiency and scalability (Chen & Guestrin 2016), Least Absolute Shrinkage and Selection Operator (LASSO) is a statistical technique that integrates variable selection and regularization, enhancing the accuracy of the final selected model (Tibshirani 2011). Ridge regression is a regularization method that mitigates multicollinearity and overfitting by incorporating a penalty into the cost function. It is more efficacious in datasets with high correlation. The primary influence of the penalty term is governed by the hyperparameter λ; an increase in

λ results in greater coefficient shrinkage, hence reducing overfitting. It is most applicable in computational biology, genetics, and environmental studies, where data exhibits multicollinearity. A comparable methodology was employed for the classification model, namely utilizing random forest, Multi-Layer Perceptron (MLP), support vector machine (SVM), Gaussian Naive Bayes (GNB), Gradient Boosting Classifier (GBC), Stochastic Gradient Descent (SGD), Logistic Regression (LR), and Extra Tree Classifier (ET) (Hoerl & Kennard 2000). The final model was chosen based on the performance of the parameters and was evaluated for robustness through ten-fold cross-validation (CV).

4.2.8 Chronological Lifespan Assay

Pulsed cells from 25 cycles (13 µL) were introduced as an inoculant into deep-well plates containing 1.3 mL of SC media, which were subsequently covered with a Breathe-Easy membrane (Z380059, Sigma). An aliquot of cells was utilized for the propidium iodide-based cell viability quantification experiment on Days 1, 3, 5, 10, 15, and 20. On the specified days, a 50 µL aliquot was mixed with 50 µL of freshly prepared Propidium Iodide (PI; 11195, SRL) staining solution in 1x PBS to get a final PI concentration of 5 µg/mL in 96-well black plates, which were incubated in the dark for 20 minutes. The positive control in this experiment consisted of heat-killed (HK) cells. A Biotek Synergy HTX multi-mode reader was employed to quantify fluorescence at excitation and emission wavelengths of 530/25 nm and 590/25 nm, respectively. OD600 was assessed using an additional 50 µL aliquot from the same plates after a 1:2 dilution with 1X PBS. The fluorescence readings were initially standardized in the data analysis utilizing the respective well's blank normalized OD600 values. The normalization process was subsequently conducted using heat-killed cells, which were calibrated to 100% mortality. To avoid negative percentages, the proportion of death was established at 100% if a sample demonstrated greater fluorescence than the HK cell. The statistics were subtracted from 100% to get the survival percentage, and any outliers were excluded. The survival numbers for each day were subsequently adjusted to the initial day of the experiment, designated as 100% survival. This work was carried out in collaboration with Subhadeep Duari, a member of the same laboratory.

4.2.9 Wheat Germ Agglutinin Staining

Cells were subjected to 25 cycles of thermal pulsing initially. Subsequently, 10 µL of the cell suspension was utilized to initiate a 1 mL culture on the deep-well plate using SC media, maintained at 30 °C. On day 5, a 100 µL aliquot of cells was obtained from five distinct biological replicates. The cells were resuspended in 100 µL of fresh 1X phosphate-buffered saline (PBS). Subsequently, cells were exposed to Wheat Germ Agglutinin (WGA; InvitrogenTM, W11261) at a final concentration of 1 µg/mL and

incubated in the dark for 45 minutes at 30 °C. Subsequent to incubation, the cells were centrifuged at 6000 rpm for 3 minutes, and the resultant pellet was resuspended in 15 µL of fresh 1X PBS. From this, 5 µL was utilized for microscopy (Nikon Eclipse Ci-L Fluorescence Microscope with a Nikon DS-Qi2 camera) employing a FITC filter for green fluorescence. The photos were captured with an exposure duration of 600 ms. Images were analyzed using ImageJ 68, and each green dot was counted manually. This work was carried out in collaboration with Subhadeep Duari, a member of the same laboratory.

4.2.10 Senescence Induction in Human Cells

Human dermal fibroblasts (NHDF-Ad; CC-2511, Lonza Bioscience) were employed for the senescence assay. Cells were cultured in a T-75 flask containing 10% Fetal Bovine Serum (FBS) in Dulbecco's Modified Eagle Medium (DMEM). At 70% confluency, the cells were divided and plated onto coverslips in 6-well plates, with approximately 1500 cells per well. The cells were grown in a CO2 incubator until they attained around 30% confluency (around 2 days). Thereafter, the cells were administered camptothecin (CPT) to induce senescence, as recommended by Kusumoto et al (Kusumoto et al. 2021). We employed a dose gradient of 6.25 nM, 12.5 nM, 25 nM, 50 nM, and 100 nM for the induction of senescence. Furthermore, we employed solvent control (DMSO) and untreated cells as controls for this test, incubating the cells for 48 hours in a CO2 incubator.

We evaluated the effectiveness of CPT in causing senescence at various concentrations using the Senescence Associated β-Galactosidase assay (SA-βgal assay), with image capture conducted via a Creuzet Tissue Culture Trinocular Microscope at 5X magnification. The quantification of senescence in both positive and negative cells was then conducted utilizing ImageJ software (https://imagej.net/software/fiji/) (Schindelin et al. 2012). Additionally, to prepare the CPT dataset, the coverslips from each well were inverted onto DPX mountant on a slide to obtain phase contrast images of human dermal fibroblast cells. Imaging was conducted at 10X magnification with a Nikon Eclipse Ci-L Fluorescence Microscope, with an exposure duration of 3 ms and an analog gain of 64X. This work was carried out in collaboration with Sakshi Arora and Subhadeep Duari, members of the same laboratory.

4.2.11 Computational Analysis of the scCamAge Model on Human Senescence Dataset

To evaluate the efficacy of the scCamAge model in forecasting human senescent cells, we conducted tests on two distinct datasets: initially, the internally developed camptothecin-induced senescence of human fibroblast cells, and subsequently, a publicly accessible dataset of replicative senescence in human fibroblasts (Sturm et al. 2022). We analyzed the image features derived from scCamAge, the features

from the bioactivity prediction module of scCamAge, the combined features of image and bioactivity, and the standard InceptionResNet image features. We employed a Random Forest classifier and conducted 10-fold cross-validation for this comparison. We conducted the One-vs-All analysis on the test predictions for each fold. The model and evaluation were conducted utilizing scikit-learn packages (Pedregosa et al. 2012).

4.3 Results

4.3.1 Development of scCamAge Prediction Engine

Recent advancements in deep learning algorithms for image analysis have markedly improved the accuracy and efficiency of retrieving complex information from microscopic images; yet, a comprehensive and flexible solution, particularly in longevity research, remains mostly absent. We provide scCamAge, a solution based on a multi-model transfer learning technique that enables the capture of age-related phenotypic, biochemical, and morphometric alterations **(Figure 4.1A)**. To create scCamAge, we first produced high-resolution phase-contrast microscopic images of aging yeast cells at ten significant time intervals, ranging from day 2 (young) to day 20 (aged), under optimal growth conditions across three independent large-scale experiments **(Figure 4.1A, B; Figure 4.2A)**. We acquired 472,606 single cells, and after manually filtering out poorly segmented cells, we retained 336,514 single cells, resulting in a loss of approximately 28.8% of the initial count **(Figure 4.1B; Figure 4.2B, C)**. We reused the publicly accessible YeastSpotter model for automatic single-cell segmentation (Lu et al. 2019) **(Figure 4.2D)**. Additionally, we employed single-cell images rather than entire micrographs to enhance longevity forecasts by mitigating background noise, overlapping cells, and stage variability, while accurately capturing essential morphological and phenotypic characteristics. Following image selection, we employed a transfer learning methodology and evaluated four distinct pre-trained models: InceptionResNet (vision: v0.10.0, inception_v3), GoogLeNet (vision: v0.10.0, googlenet), ResNet (version 50), and VGG 19 (version 19) for the development of the Predictor module of scCamAge **(Figure 4.2E)**. The InceptionResNet (vision: v0.10.0, inception_v3) model exhibited exceptional performance, achieving an AUC (Area Under the Curve) value between 72.45 and 83.93 **(Figure 4.1C)**. Consequently, it was chosen for subsequent model reconstruction with hyperparameter adjustment **(Figure 4.2F, G)**. The optimal hyperparameters were utilized for 500 training epochs. The optimal model performance was recorded at the 200th epoch, achieving an AUC of 86.2% and an accuracy of 53.4% (with random accuracy at 10%) **(Figure 4.1D)**. We conducted a thorough assessment of the model's performance utilizing the rigorous AUC-ROC (Area under the Receiver Operating Characteristic Curve) comparison, specifically the One-vs-All methodology, and noted satisfactory outcomes, with AUC values

spanning from 0.77 (day 12) to 0.97 (day 2) **(Figure 4.1E; Figure 4.2H)**. The model trained at the 200th epoch was ultimately designated as the default Predictor model and subsequently referred to as the scCamAge model in future downstream analyses. Furthermore, we assessed the efficacy of this model in forecasting single-cell yeast chronological age utilizing an independent dataset (total single cells - 82,268) and noted a positive correlation (Spearman's correlation coefficient of 0.794), demonstrating the robustness of the scCamAge model **(Figure 4.1F)**.

Our research illustrates the potential of scCamAge to forecast six essential aging-related bioactivities, including reactive oxygen species (ROS) levels, mitochondrial dynamics, vacuolar shape, epigenetic modifications, proteostasis, and genomic stability **(Figure 4.1G)**. The term "bioactivity" in this context refers to these six specific, measurable biological states. They were defined and measured using two distinct experimental methods to generate ground-truth data for the models, namely bioactive dye-based measures and drug exposure experiments, where we systematically recorded and examined aging signs from single-cell micrographs. In the initial approach, we utilized bio-viable fluorescent dyes to directly assess mitochondrial potential, reactive oxygen species levels, and vacuolar dynamics at ten chronological aging time points in yeast. Phase-contrast and fluorescent micrographs of individual cells were captured and paired, while the scCamAge method was employed to produce embeddings from the phase-contrast pictures. Regression models utilizing these embeddings demonstrated robust predictive performance, yielding R^2 values ranging from 0.62 to 0.65, following thorough tuning and 10-fold cross-validation **(Figure 4.1G, H; Figure 4.2A; Figure 4.3A)**. We trained various regression models employing algorithms including Random Forest (RF), eXtreme Gradient Boosting(XGB), Multi-layer Perceptron (MLP), RIDGE, and LASSO (Least Absolute Shrinkage and Selection Operator), ultimately selecting the RIDGE-based models for their superior performance, as indicated by the maximum R^2 value under default settings **(Figure 4.3B, C)**. To facilitate scalable screening, we employed phase-contrast image embeddings to predict aging-related bioactivities, as they efficiently correlate image-based attributes with fluorescent intensities, eliminating the necessity for co-supplementing fluorescence images during testing.

In the second strategy, we subjected cells to sublethal doses of medicines such as etoposide, azacytidine, and MG132 to elicit DNA damage, epigenetic dysregulation, and proteostasis disruption, respectively **(Figure 4.1G, I; Figure 4.2A; Figure 4.3A)**. Classification models utilizing scCamAge embeddings and dosage information demonstrated strong predictive performance, with AUC values between 0.6 and 0.8 **(Figure 4.1I; Figure 4.3D-F)**. We examined the efficacy of eight distinct classification algorithm-based models utilizing image-specific embeddings and treatment dose information. To thoroughly validate the models' efficacy and ascertain whether the observed variations stem from a generalized stress response or

are specific to individual drug treatments, we constructed new binary classifiers for each drug (etoposide, azacytidine, and MG132) and assessed their performance using stringent 10-fold cross-validation. Furthermore, each drug-specific model was evaluated by displaying phase-contrast images of their test data alongside images from yeast cells treated with alternative drugs. Our findings indicated that although drug-specific models attained elevated AUC-ROC values on their designated datasets, predictions on alternative datasets resulted in AUC-ROC values near 0.5, implying random classification. This suggests that the models predominantly acquired spatiotemporal morphological characteristics unique to each medication, rather than universal stress-related attributes. Further tests, including oxidative stress (H_2O_2), heat stress (42 °C), and metabolic stress (hydroxyurea) corroborated these results, as analyses of these stress-response datasets yielded diminished AUC-ROC values **(Figure 4.1J-K; Figure 4.4A-B)**. When evaluating the performance metrics it is expected that training set accuracies will be higher than test set accuracies; a large gap would indicate overfitting. However, the box plots showing the results of 10-fold cross-validation **(Figure 4.3F, Figure 4.4B)** provide a more robust assessment of generalization, with tight distributions indicating stable model performance across different data subsets.

Besides forecasting bioactivities, scCamAge facilitates the automatic calculation of cellular morphometric characteristics, including area, perimeter, and convex hull. During yeast chronological aging, morphometric analysis was conducted solely on live cells to guarantee precision. Employing CLS tests and propidium iodide (PI) labeling, dual-channel imaging facilitated the extraction of phase-contrast pictures for PI-negative (viable) cells **(Figure 4.1L, M; Figure 4.2J)**. Our research indicated a notable age-related augmentation in cellular area, perimeter, and convexity **(Figure 4.1N; Figure 4.2I, K-L)**. Young cells displayed increased variability in area and perimeter distributions, which became more consistent with age. This transition was evident in a leptokurtic kurtosis range of 2.27-2.69, underscoring notable morphological alterations linked to aging. The concluding element of scCamAge is its Explainability module, intended to pinpoint sub-cellular regions essential for the model's age projections. This module functions analogously to attribution techniques like saliency maps, generating a visual overlay on the cell image to highlight the regions most influential to the model's prediction. Utilizing an external collection of single-cell pictures across ten chronological aging time points, we detected a dynamic alteration in regions of interest. In younger cells (day 2), enrichment was localized in the nucleus, whereas in older cells (day 20), it migrated toward the cell wall **(Figure 4.1O)**. In summary, scCamAge amalgamates various functionalities: its Predictor module provides single-cell resolution age estimations, the Bioactivity module forecasts six principal aging-related bioactivities, and the Explainability and Morphometry modules improve interpretability and automate morphometric

assessments. Collectively, these components position scCamAge as an all-encompassing instrument for cellular aging research **(Figure 4.1A)**.

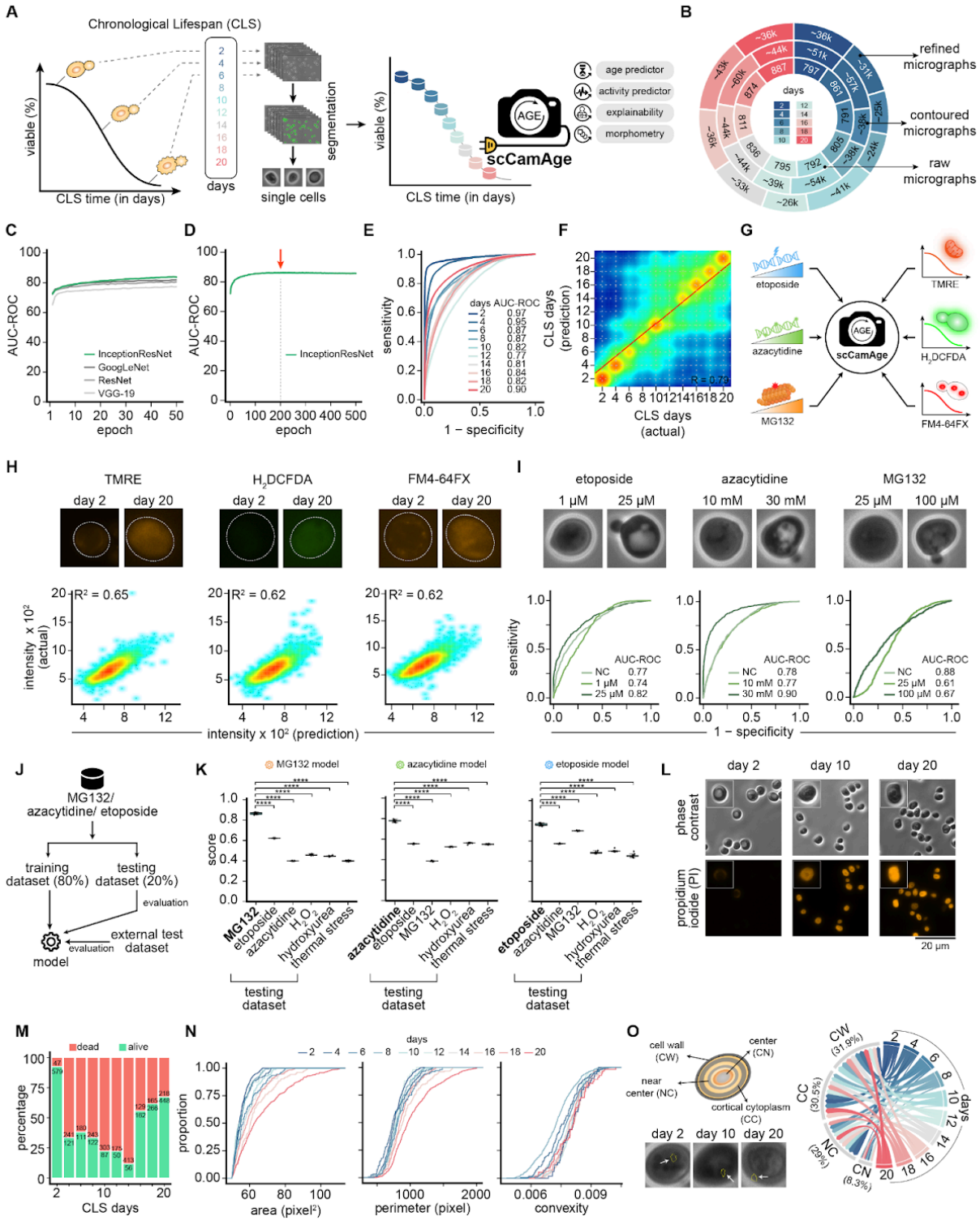


Figure 4.1: Development of scCamAge Deep Learning Architecture

(A) *Schematic diagram presenting the survival curve of yeast chronological aging, emphasizing crucial time points for acquiring micrographs to construct a training dataset for scCamAge modeling. On the right, a diagram outlines the main steps in the scCamAge workflow, including image segmentation to isolate single yeast cells, their utilization in constructing a multi-model for predicting the age of individual yeast cells, and other noted features.* ***(B)*** *Donut pie chart displaying the count of raw, contoured, and manually refined single-yeast cell images at specified chronological time points.* ***(C)*** *Line plot comparing the performance of generalized transfer learning models-InceptionResNet, GoogLeNet, ResNet, and VGG19 on a single-cell yeast chronological aging dataset over fifty epochs.* ***(D)*** *Line plot indicating the performance of the transfer learning model InceptionResNet on a single-cell yeast chronological aging dataset over five hundred epochs.* ***(E)*** *The area under the ROC Curve (AUC-ROC) illustrates the One-vs-All AUC-ROC values under specified conditions.* ***(F)*** *Contour scatter plot with overlaid regression line illustrating the correlation between the scCamAge predicted and actual (ground truth) class labels on the held-out dataset of single yeast cells undergoing chronological aging.* ***(G)*** *Schematic diagram illustrating aging-associated bioactivities, including cellular reactive oxygen species (ROS) levels, mitochondrial abundance and potential, vacuolar organelle morphology and dynamics, epigenetic modifications, alterations in proteostasis, and genomic instability, along with the strategies employed for their inference.* ***(H)*** *Representative micrographic insets depicting the fluorescence of single-cell yeast pre-treated with bioactivity measuring vital dyes. The bottom graphs display the correlation between predicted vs. actual fluorescence intensities in the indicated conditions.* ***(I)*** *Representative micrographic inlets depicting the phase contrast single-cell yeast pre-treated with the indicated drugs. The bottom graphs display the AUC-ROC plots for the indicated conditions.* ***(J)*** *Schematic representation of the overall strategy used for cross-validation to evaluate the specificity of the indicated drug treatment models.* ***(K)*** *Box plots showing the cross-validation performance results, represented as AUC-ROC values (y-axis), for drug-specific models (MG132, Azacytidine, and Etoposide) evaluated on the specified testing datasets (x-axis). The Mann-Whitney U test was used to compute the p-value, with a significance threshold of 0.05. Symbols *, **, ***, and **** represent p-values less than 0.05, 0.01, 0.001, and 0.0001, respectively.* ***(L)*** *Representative micrographs (left) displaying phase-contrast images (top) and corresponding propidium iodide (PI)-stained cells (bottom) at the indicated time points during yeast chronological aging assays. The scale bar corresponds to 20 µm.* ***(M)*** *The bar plot (right) shows the relative proportions of live (green) and dead (red) cells quantified at the respective time points.* ***(N)*** *Empirical cumulative distribution function (ECDF) plots showing the distributions of cell area, perimeter, and convexity of live yeast cells (PI-negative) at the indicated time points during yeast chronological aging assays.* ***(O)*** *A schematic diagram highlighting the key cellular area used for the quantitative analysis of the explainability module of the scCamAge model. Representative micrographs at the bottom depict the scCamAge explainability module guided top-most feature (highlighted in yellow) responsible for chronological age predictions in the indicated conditions. Circos plot depicting the relative cellular localization of the scCamAge explainability module guided top-most features at critical time points of chronological aging.*

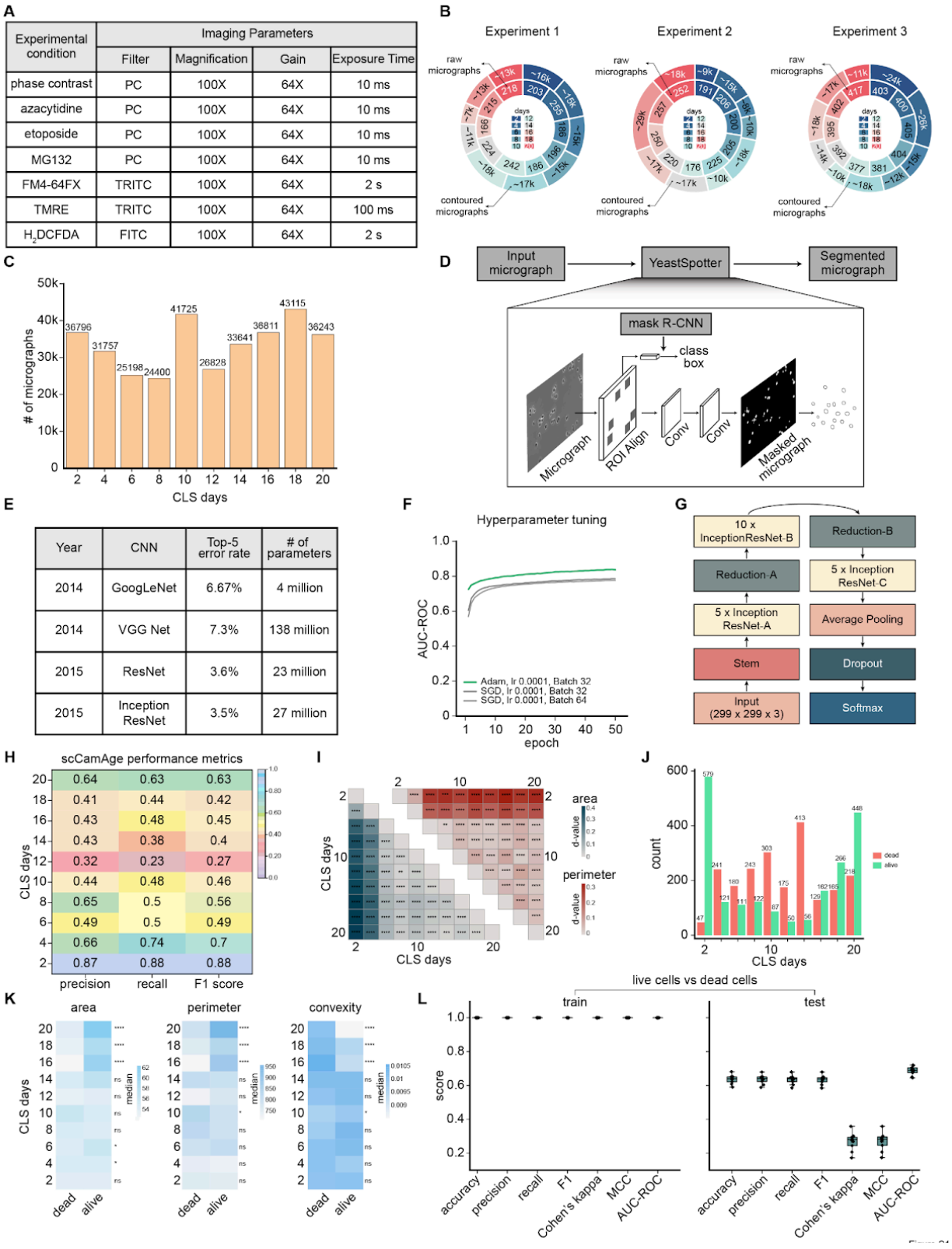


Figure 4.2: Construction of scCamAge Using the Transfer Learning Approach.

(A) *Table displaying imaging parameters utilized to capture micrographs for specific datasets.* ***(B)*** *Donut charts showing the total counts of single-cell images collected over three independent experiments to study the chronological aging of yeast cells.* ***(C)*** *Bar chart illustrating the number of micrographs obtained at various chronological age time points during chronological lifespan assays.* ***(D)*** *Diagram outlining the methodology for segmenting individual yeast cells from phase contrast microscopy images.* ***(E)*** *Table listing the transfer learning models used in this research, including their release year, top-5 error rate, and parameter count.* ***(F)*** *Line graph demonstrating improved performance of selected models following adjustments in hyperparameters.* ***(G)*** *A flowchart illustrates the deep learning framework employed to develop the scCamAge model.* ***(H)*** *Heatmap showing the precision, recall, and F1 scores of the scCamAge model at different chronological age points.* ***(I)*** *Heatmap depicting the pairwise d-values from the Kolmogorov-Smirnov statistical test under specified conditions. The lower triangle of the heatmap represents d-values for the area, while the upper triangle shows d-values for the perimeter. The test provides two key outputs: the D-value, which represents the maximum difference between the cumulative distribution functions (CDFs) of the two groups, and the p-value, which indicates the statistical significance of the observed difference. A smaller p-value (<0.05) signifies a significant difference in the distributions. Symbols *, **, ***, and **** refer to p-values <0.05, <0.01, <0.001, and <0.0001, respectively.* ***(J)*** *Bar plot depicting the number of dead and alive cells in CLS days.* ***(K)*** *Heatmap showing median values of area, perimeter, and convexity. The Mann-Whitney U test was used to compute the p-value, with a significance level of 0.05, and *, **, ***, and **** denote p-values of <0.05, <0.01, <0.001, and <0.0001, respectively.* ***(L)*** *Box plot showing the accuracy, AUC-ROC, Cohen's kappa, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC) from 10-fold cross-validation of live versus dead cells.*

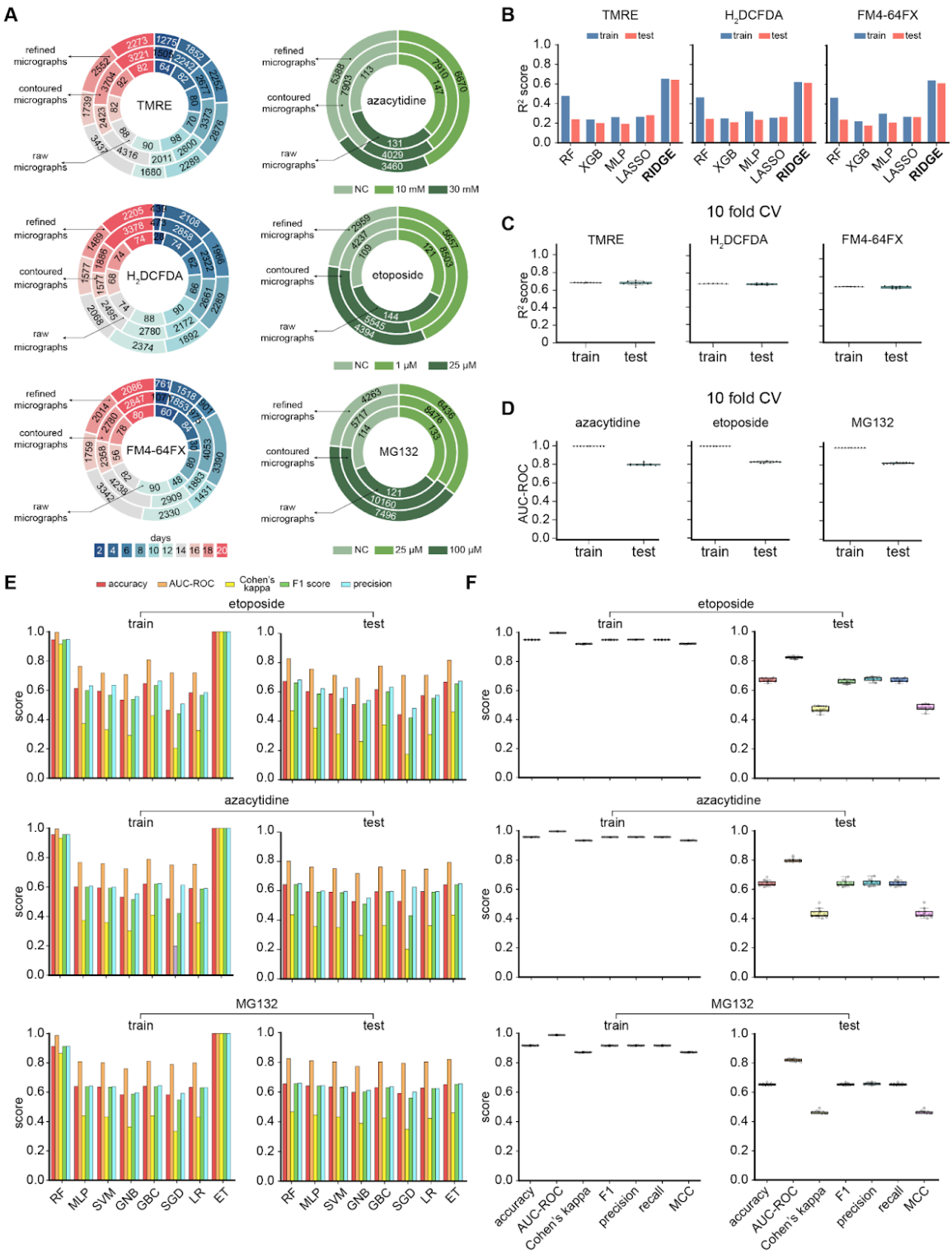


Figure 4.3: scCamAge Utilizes Various Datasets for Predicting Aging-Associated Bioactivities

(A) *Donut charts showing the absolute counts of single-cell yeast images acquired under various conditions.* ***(B)*** *Bar charts displaying the R² values for different regression models on the training and testing datasets aimed at*

predicting aging-associated bioactivities from phase contrast images. ***(C)*** *Box plot illustrating R² values obtained from 10-fold cross-validation in specified conditions.* ***(D)*** *(Repeated; consider removal or modification if intentional) Box plot depicting R² values from 10-fold cross-validation in the specified conditions.* ***(E)*** *Bar chart comparing the accuracy, AUC-ROC, Cohen's kappa, F1 score and, precision, across indicated training and test datasets.* ***(F)*** *Box plot showing the accuracy, AUC-ROC, Cohen's kappa, precision, recall, and Matthews Correlation Coefficient (MCC) from 10-fold cross-validation under specific conditions.*

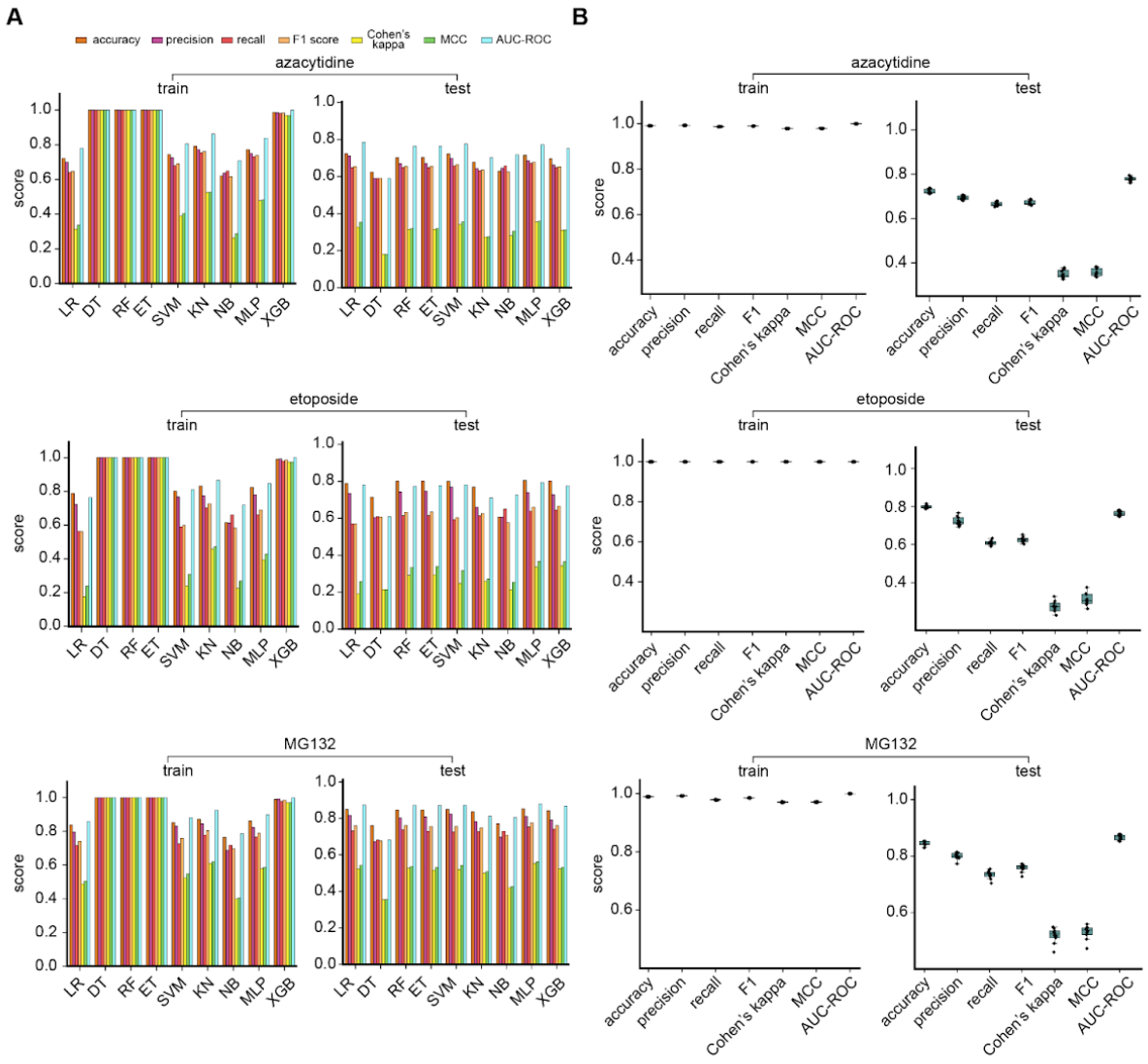


Figure 4.4: Evaluation of Multiple Classification Algorithms for Building scCamAge Bioactivity Prediction Models

(A) Bar chart comparing the accuracy, precision, recall, F1 score, Cohen's kappa, Matthews Correlation Coefficient (MCC), and AUC-ROC across indicated training and test datasets for bioactivity models. (B) Box plot showing the accuracy, precision, recall, F1 score, Cohen's kappa, Matthews Correlation Coefficient (MCC), and AUC-ROC across indicated training and test datasets for bioactivity mode.

4.3.2 scCamAge Pinpoints Temporal Longevity Activation upon Pro-Longevity Drug Treatments

Utilizing scCamAge, we subsequently determined the temporal initiation of pro-longevity responses in yeast cells subjected to metformin, spermine, and spermidine, recognized pro-longevity agents **(Figure 4.5A, B)**. Aliquots were collected at ten chronological age intervals, and phase-contrast micrographs from treated cells (metformin: 20,783 cells; spermine: 23,504 cells; spermidine: 21,827 cells) were examined using scCamAge **(Figure 4.5B)**. Our predictions indicated a strong pro-longevity effect in metformin-treated cells, demonstrated by a leftward shift on the prediction axis, beginning as early as day 4 of the CLS experiment **(Figure 4.5C; Figure 4.5A)**. Treatments with spermidine and spermine shown a postponed buildup of senescent cells relative to controls. Pairwise Spearman correlations among chronological time points revealed the most significant lifespan response with metformin ($R = 0.17$), succeeded by spermine ($R = 0.18$) and spermidine ($R = 0.24$) **(Figure 4.5C, D)**. These findings correspond with earlier documented pro-longevity benefits of these medications (Barardo et al. 2017).

We investigated the effects of metformin, spermine, and spermidine on aging-related bioactivities by contrasting their impacts with those of untreated controls. Dye-based bioactivity tests indicated increased mitochondrial potential, reactive oxygen species (ROS) levels, and intracellular vacuoles in the treated groups **(Figure 4.5E; Figure 4.6E)**. Significantly, although the general trend is consistent among bio-viable dyes, pairwise comparisons of the median distributions distinctly demonstrated statistical significance at various chronological time intervals. Using pharmacological treatment-based models, we noted a significant decrease in the percentage of cells exhibiting increased DNA damage (class: 25 μM etoposide) with aging, indicating that these treatments enhance genomic stability **(Figure 4.6A)**. Likewise, a reduced number of cells exhibited age-related epigenetic dysregulation across treatments, underscoring their protective benefits against epigenetic modifications (classes: 10 mM and 30 mM azacytidine) **(Figure 4.6A)**. Nevertheless, no substantial variations in proteostasis loss were detected between treated and untreated cells. Morphometric analysis demonstrated a notable rightward shift in the distributions of cellular area and perimeter in metformin-treated cells, signifying an enlargement in cell morphology with aging. For spermidine, similar shifts were ephemeral and noted solely at particular time

intervals (CLS days 6 and 8) **(Figure 4.6B-D)**. Utilizing scCamAge's explainability module, we discerned critical cellular areas affecting model predictions. Regions of interest (ROIs) exhibited enrichment at the cell wall (35-36%) and cortical cytoplasm (30-35%) in treated cells, whereas untreated controls displayed greater cytoplasmic enrichment (23-26%) **(Figure 4.5F)**. These data indicate that pro-longevity medicines affect both bioactivity and cellular architecture during the aging process.

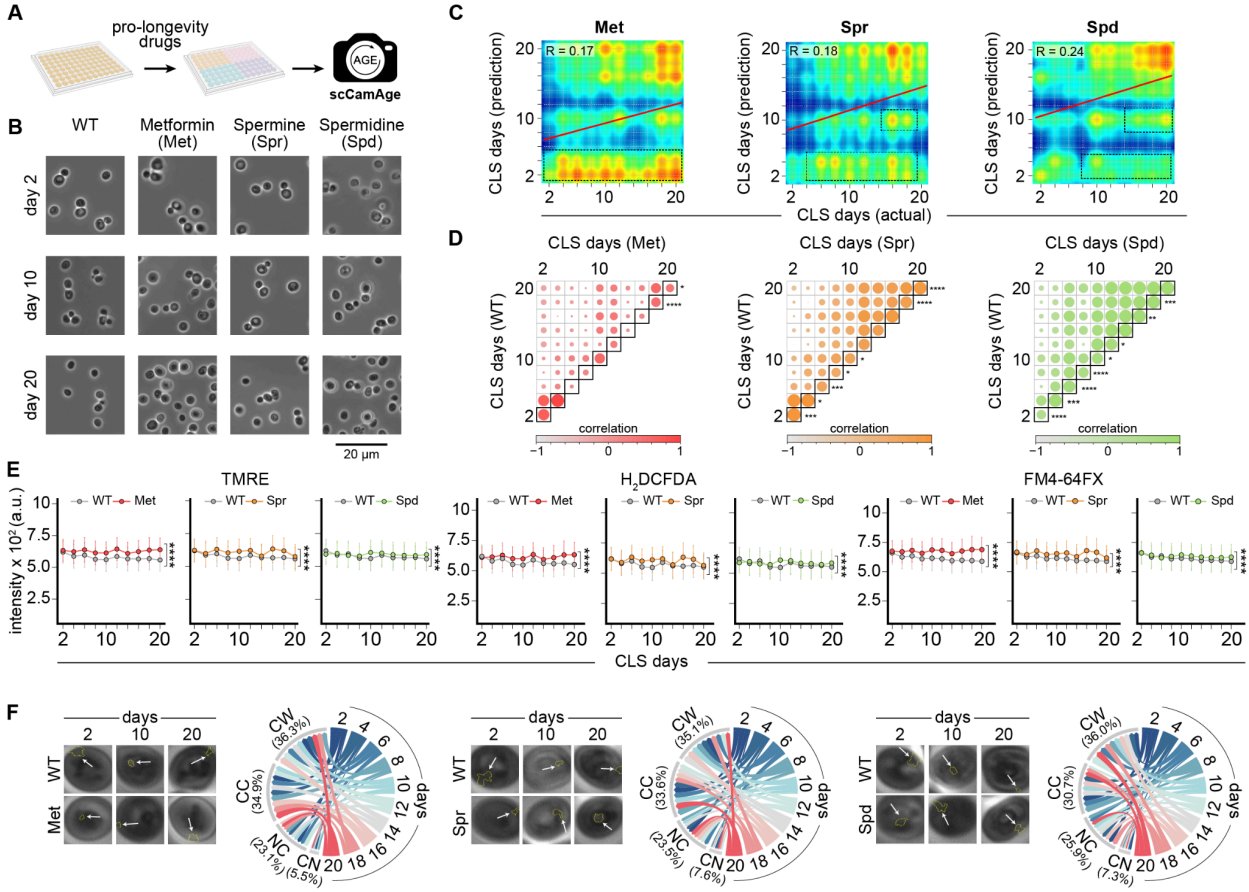


Figure 4.5: scCamAge Reveals Longevity Effects of Pro-Longevity Drugs using Micrograph Analysis

(A) *Schematic representation depicting the experimental workflow used to validate scCamAge performance using well-established pro-longevity drugs on chronological lifespan assays.* ***(B)*** *Representative micrographs showing the phase contrast microscopic images of wild-type (BY4741) yeast cells treated with pro-longevity drugs at specified chronological time points. The scale bar corresponds to 20 µm.* ***(C)*** *Contour scatter plot with overlaid regression line illustrating the correlation between the scCamAge predicted (y-axis) and actual (ground truth; x-axis) class labels on pro-longevity drug-treated yeast cultures at indicated chronological age time points. Of note, overlaid dotted rectangles represent critical regions responsible for pro-longevity effects.* ***(D)*** *Correlation plot depicting pairwise (chronological time points) correlation between untreated and pro-longevity drug-treated conditions. The KS test was performed to compare the distributions of features between the test and control groups. The test provides*

two key outputs: the D-value, which represents the maximum difference between the cumulative distribution functions (CDFs) of the two groups, and the p-value, which indicates the statistical significance of the observed difference. ***(E)*** *Line plots depicting trends of predicted bioactivities across ten chronological time points for specified conditions.* ***(F)*** *Representative micrographs depicting the scCamAge explainability module guided top-most feature (highlighted in yellow) responsible for chronological age predictions in specified conditions. Circos plot showing the relative cellular localization of the scCamAge explainability module guided top-most feature at critical time points of chronological aging in the indicated drug-treatment conditions.*

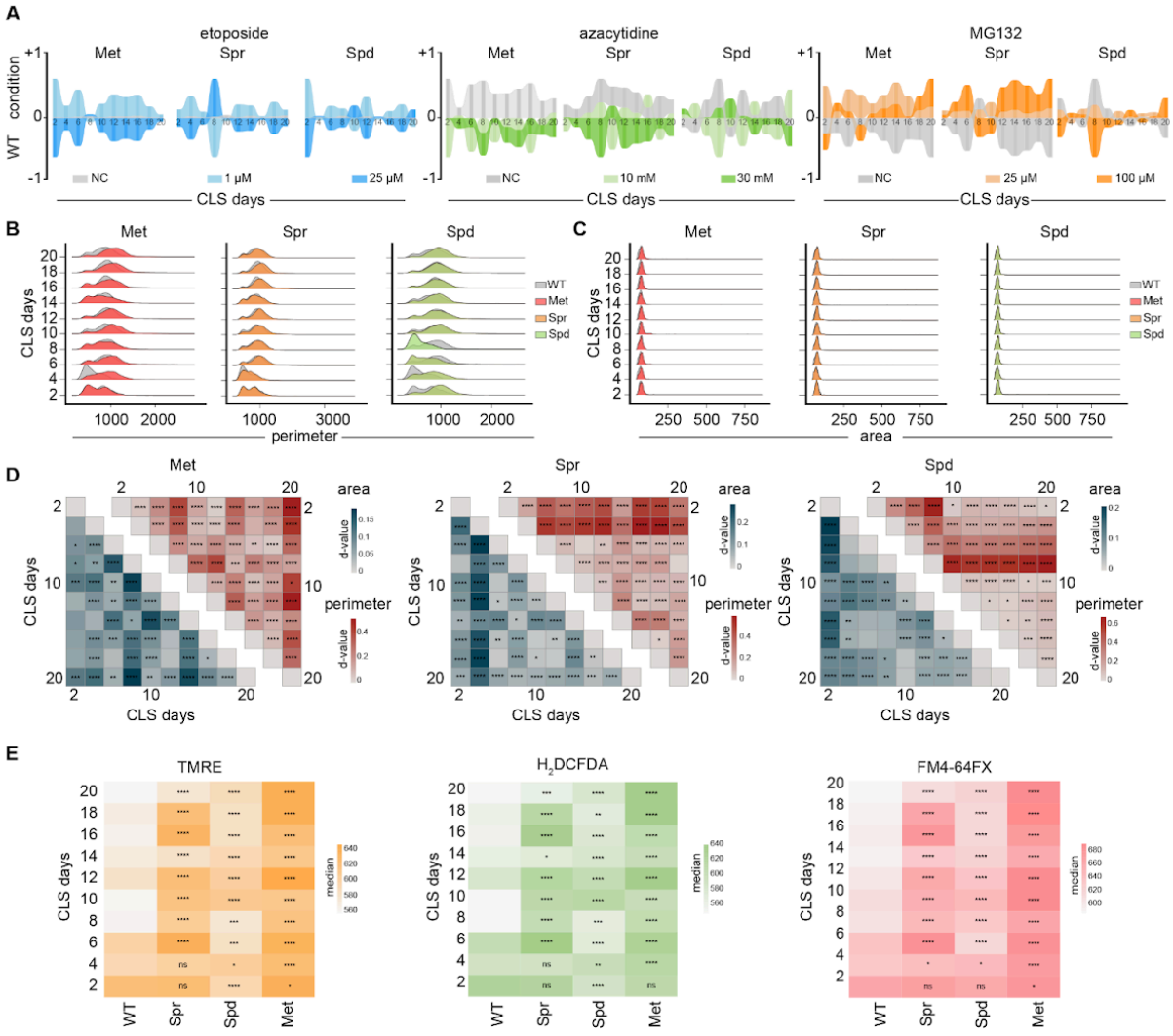


Figure 4.6: scCamAge Reveals Variable Aging-Associated Bioactivity Responses to Interventions

(A) *Ribbon plots illustrate yeast cells' predicted relative enrichment or de-enrichment in various drug treatment conditions compared to untreated controls.* ***(B)*** *Overlaid density plots showing population-level distributions of the perimeters of single yeast cells under specified conditions during chronological aging.* ***(C)*** *Overlaid density plots present population-level distributions of single yeast cell areas under specified conditions during chronological*

aging. ***(D)*** *Heatmaps showing the pairwise d-values from the Kolmogorov-Smirnov statistical test in the indicated conditions. The KS test was performed to compare the distributions of features between the test and control groups. The test provides two key outputs: the D-value, which represents the maximum difference between the cumulative distribution functions (CDFs) of the two groups, and the p-value, which indicates the statistical significance of the observed difference. A smaller p-value (<0.05) signifies a significant difference in the distributions. The p-value cutoff used is 0.05. *, **, ***, and **** refer to p-values <0.05, <0.01, <0.001, and <0.0001, respectively.* ***(E)*** *Heatmaps depicting the median fluorescence intensities of the indicated bioactivity dyes (TMRE, H_2DCFDA, and FM4-64FX) at the indicated time points during chronological aging. Pairwise statistical comparisons were performed between the test and their respective control groups to evaluate significance at the 95% confidence level. The Mann-Whitney U test was applied to the original data distributions to calculate p-values, while only the median values were used to plot the heatmaps. The p-value cutoff used in this study is 0.05. *, **, ***, and **** refer to p-values <0.05, <0.01, <0.001, and <0.0001, respectively.*

4.3.3 scCamAge Pinpoints pro or anti-longevity responses in various yeast knockouts

In yeast aging studies, many mutants have been identified for their impact on longevity by chronological and replicative lifespan assays (Zimmermann et al. 2018; He et al. 2018). We subsequently questioned if the activation of pro-longevity responses in the well-characterized loss-of-function mutants is consistent or occurs at varying temporal intervals during chronological aging. To investigate this, we conducted chronological lifespan (CLS) experiments on documented pro-longevity mutants and noted diverse temporal activation of longevity responses **(Figure 4.7A-D; Figure 4.8A-C; Figure 4.9A-C; Figure 4.8A-E, Figure 4.10A-E)**. In the instance of *pdx3Δ* cells, the pro-longevity response commenced as early as chronological days two or four. Conversely, a delayed onset was noted in the pho89Δ and prs3Δ cells. Diverse temporal activation of pro-longevity responses was similarly found in the *met17Δ, met2Δ, cys4Δ, gre3Δ, adh2Δ,* and *ipt1Δ* knockout mutants **(Figure 4.9A-D)**. These genetic knockouts exhibit an average lifespan increase of 40% to 100% in chronological lifespan experiments in budding yeast **(Figure 4.9B)**.

Given that scCamAge conceptually facilitates the assessment of both pro- and anti-longevity responses, we conducted analogous experiments on well-characterized anti-longevity mutants, including *trx1Δ, kgd1Δ, sod2Δ, alt2Δ, gsy2Δ,* and *msw1Δ*, and noted a comparatively early manifestation of aging phenotypes **(Figure 4.7C-D; Figure 4.8D, E, H, I)**. Subsequently, we employed the Bioactivity module of scCamAge and noted increased mitochondrial potential, reactive oxygen species (ROS) levels, and an abundance of intracellular vacuoles in yeast genetic knockouts **(Figure 4.7E; Figure 4.8E-F; Figure 4.9F, J; Figure 4.10F-G)**. Upon comparing pro-longevity and anti-longevity knockout yeast cells, we noted dynamic responses to azacytidine, MG132, and etoposide across various chronological time points,

underscoring the biological significance of these established aging-related bioactivities in facilitating either anti-longevity or pro-longevity responses **(Figure 4.8D; Figure 4.9G, K)**.

Subsequently, we projected individual cells from pro- or anti-longevity yeast knockouts onto the scCamAge Explainability module, noting variations in relative enrichment at the cell wall region based on the primary selected feature: pro-longevity mutants (*pho89Δ, prs3Δ,* and *pdx3Δ*) exhibited modulation across chronological time points, ultimately resulting in equivalent relative counts on day 2 and day 20, whereas anti-longevity mutants (*trx1Δ, kgd1Δ,* and *sod2Δ*) demonstrated a more significant disparity **(Figure 4.7F).** The prediction results of aging-associated bioactivities indicated notable differences between the wild-type yeast cells and the pro-or-anti-longevity knockout cells. These findings substantiate the efficacy of the scCamAge models in monitoring yeast cell aging at single-cell resolution and facilitating the prediction of age-related bioactivities.

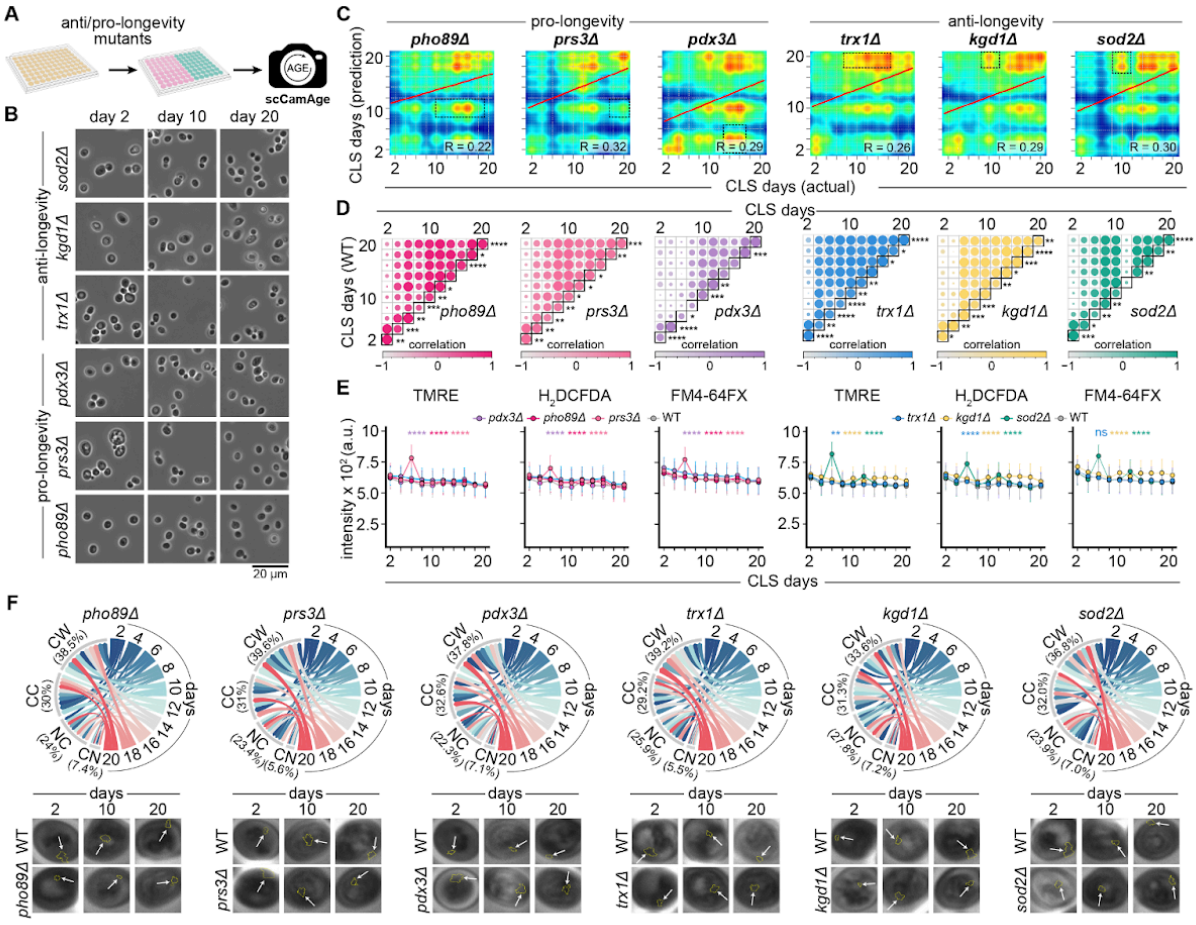


Figure 4.7: scCamAge Tracks Pro-or-Anti-Longevity Responses in Yeast Knockouts

(A) *Schematic representation illustrating the experimental workflow used to validate scCamAge performance using well-established pro- and anti-longevity genetic yeast knockouts.* ***(B)*** *Representative micrographs displaying the phase contrast microscopic images of indicated genetic knockout yeast cells harboring pro- or anti-longevity effects in chronological lifespan assays.* ***(C)*** *Contour scatter plot with overlaid regression line illustrating the correlation between the scCamAge predicted and actual (ground truth) chronological time points in chronological lifespan assays of specified pro- or anti-longevity genetic knockouts.* ***(D)*** *Correlation plot depicting pairwise correlation between wild-type and indicated genetic mutants with reported pro- or anti-longevity phenotypes in yeast chronological lifespan assays.* ***(E)*** *Line plots depicting trends of predicted bioactivities across ten chronological time points for specified conditions.* ***(F)*** *Representative micrographs depicting the scCamAge explainability module guided top-most feature (highlighted in yellow) responsible for chronological age predictions in specified conditions. Circos plot showing the relative cellular localization of the scCamAge explainability module guided top-most feature at critical time points of chronological aging.*

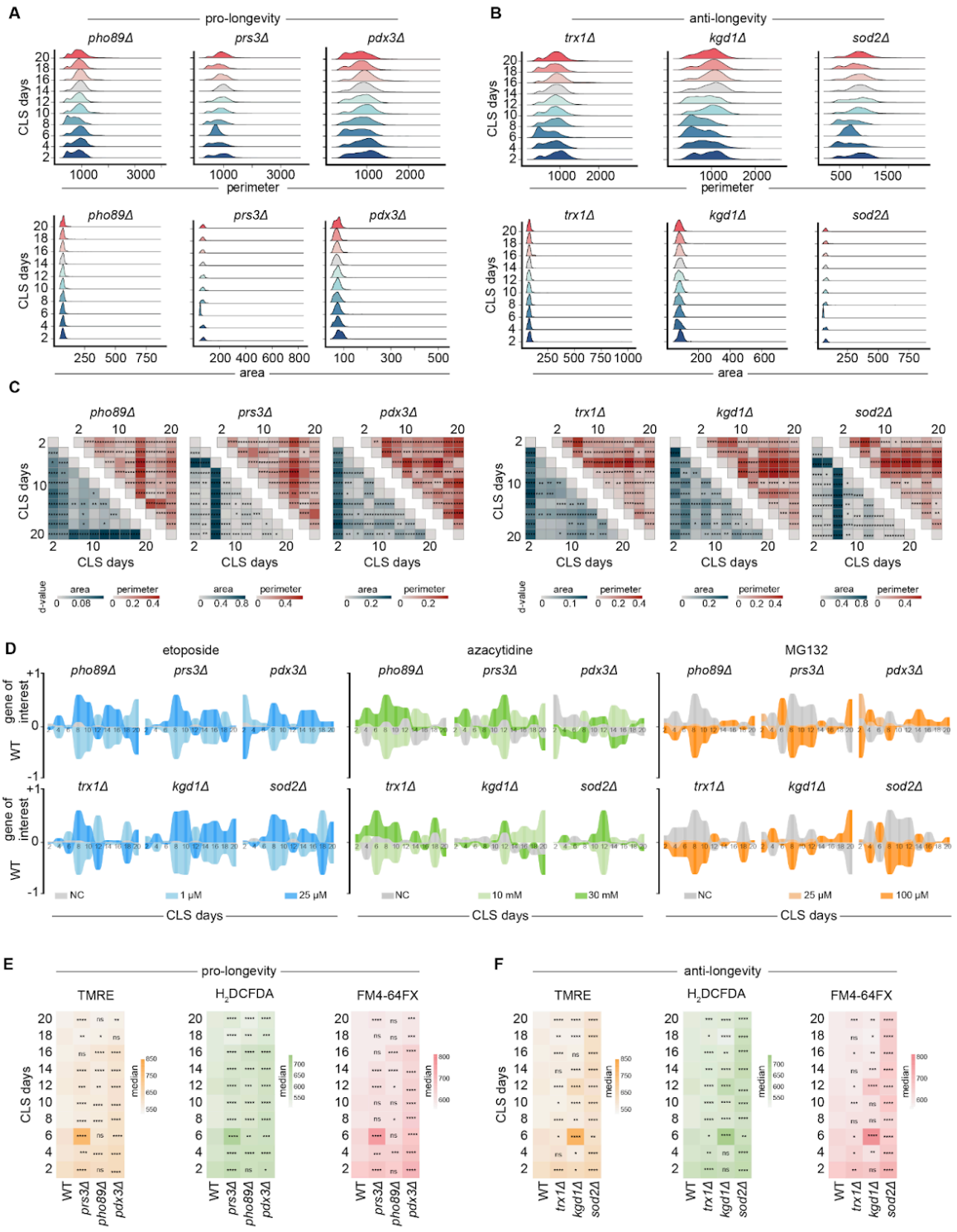
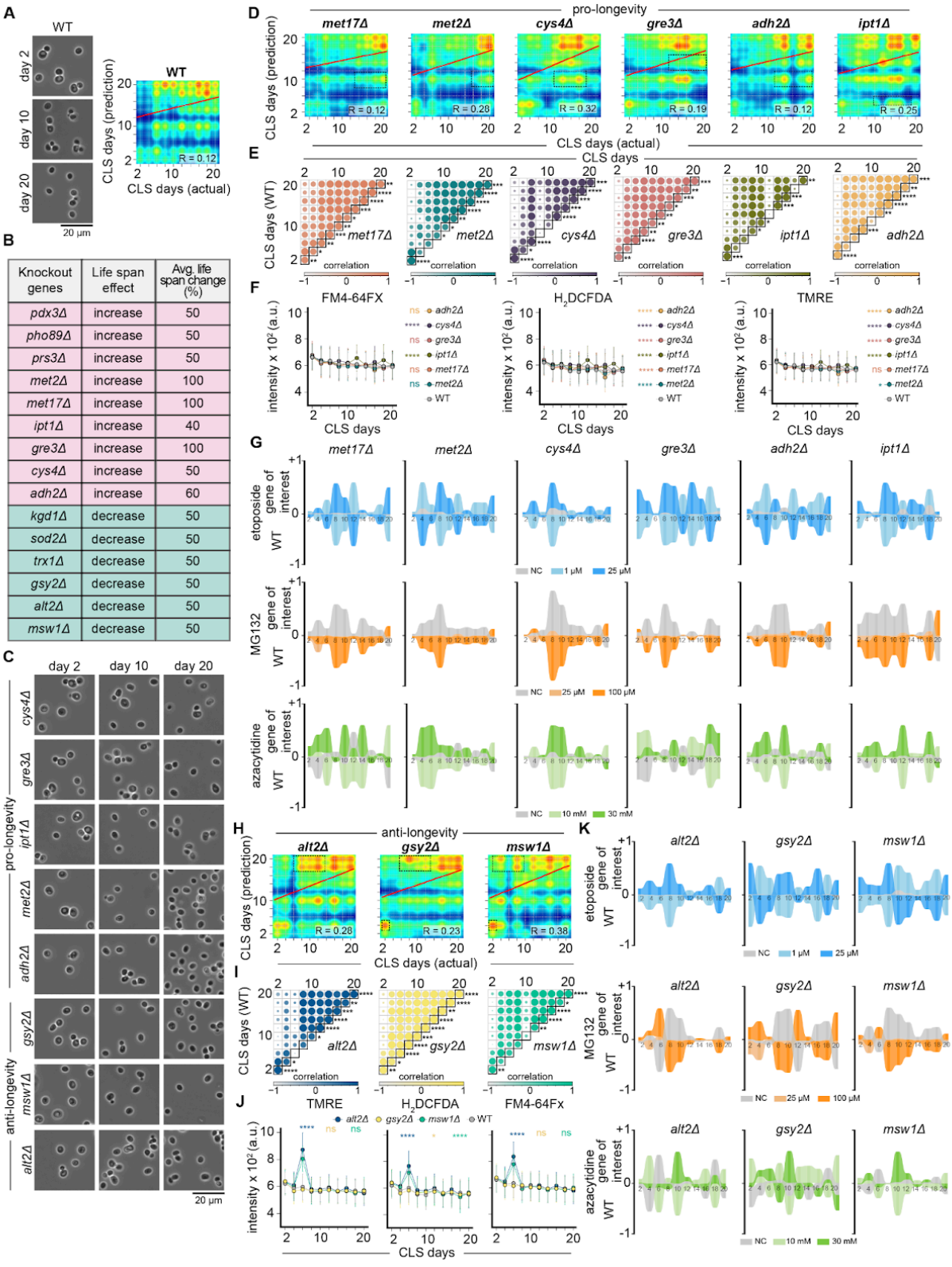


Figure 4.8: scCamAge Facilitates Population-Level Assessment of Aging-Induced Shifts in Cellular Morphometrics

(A) *Overlaid density plots showing the population-level distributions of perimeters and areas of single yeast cells in loss-of-function mutants with reported pro-longevity phenotypes.* ***(B)*** *Overlaid density plots illustrate the*

population-level distributions of perimeters and areas of single yeast cells in loss-of-function mutants with reported anti-longevity phenotypes. ***(C)*** *Heatmaps display the pairwise d-values from the Kolmogorov-Smirnov statistical test under specified conditions.* ***(D)*** *Ribbon plots depicting the predicted relative enrichment or de-enrichment of yeast cells in various loss-of-function mutants (pro- or anti-longevity) compared to wild-type controls.* ***(E)*** *Heatmaps showing the median fluorescence intensities of the indicated bioactivity dyes (TMRE, H_2DCFDA, and FM4-64FX) at the indicated time points during chronological aging of yeast mutants known to harbor pro-longevity phenotypes.* ***(F)*** *Heatmaps showing the median fluorescence intensities of the indicated bioactivity dyes (TMRE, H_2DCFDA, and FM4-64FX) at the indicated time points during chronological aging of yeast mutants known to harbor anti-longevity phenotypes.*



Knockout genes	Life span effect	Avg. life span change (%)
pdx3Δ	increase	50
pho89Δ	increase	50
prs3Δ	increase	50
met2Δ	increase	100
met17Δ	increase	100
ipt1Δ	increase	40
gre3Δ	increase	100
cys4Δ	increase	50
adh2Δ	increase	60
kgd1Δ	decrease	50
sod2Δ	decrease	50
trx1Δ	decrease	50
gsy2Δ	decrease	50
alt2Δ	decrease	50
msw1Δ	decrease	50

Figure 4.9: scCamAge Predicts Longevity Phenotypes from Yeast Mutant Micrographs

(A) *Representative micrographs showcasing phase contrast images of wild-type (BY4741) yeast cells at specific chronological time points, with a contour scatter with overlaid regression line plot to the right illustrating the correlation between scCamAge predictions and actual (ground truth) class labels in wild-type yeast cultures across these time points.* ***(B)*** *Table listing loss-of-function mutants and their reported pro-or-anti-longevity phenotypes from yeast chronological lifespan assays.* ***(C)*** *Representative micrographs displaying phase contrast images of specific loss-of-function yeast mutant cells were reported to exhibit pro- or anti-longevity phenotypes at given chronological time points. The scale bar corresponds to 20 μm.* ***(D)*** *Contour scatter plot with overlaid regression line showing the correlation between scCamAge predictions and actual (ground truth) class labels in pro-longevity loss-of-function yeast mutants at designated chronological age time points.* ***(E)*** *Correlation plot depicting pairwise correlations (by chronological time points) between wild-type and specified genetic mutants.* ***(G)*** *Ribbon plots depicting the predicted relative enrichment or de-enrichment of yeast cells in reported pro-longevity loss-of-function mutants compared to wild-type controls.* ***(H)*** *Contour scatter plot with overlaid regression line showing the correlation between scCamAge predictions and actual (ground truth) class labels in anti-longevity loss-of-function yeast mutants at designated chronological age time points.* ***(I)*** *Correlation plot depicting pairwise correlations (by chronological time points) between wild-type and indicated genetic mutants.* ***(K)*** *Ribbon plots depicting the predicted relative enrichment or de-enrichment of yeast cells in reported anti-longevity loss-of-function mutants compared to wild-type controls.*

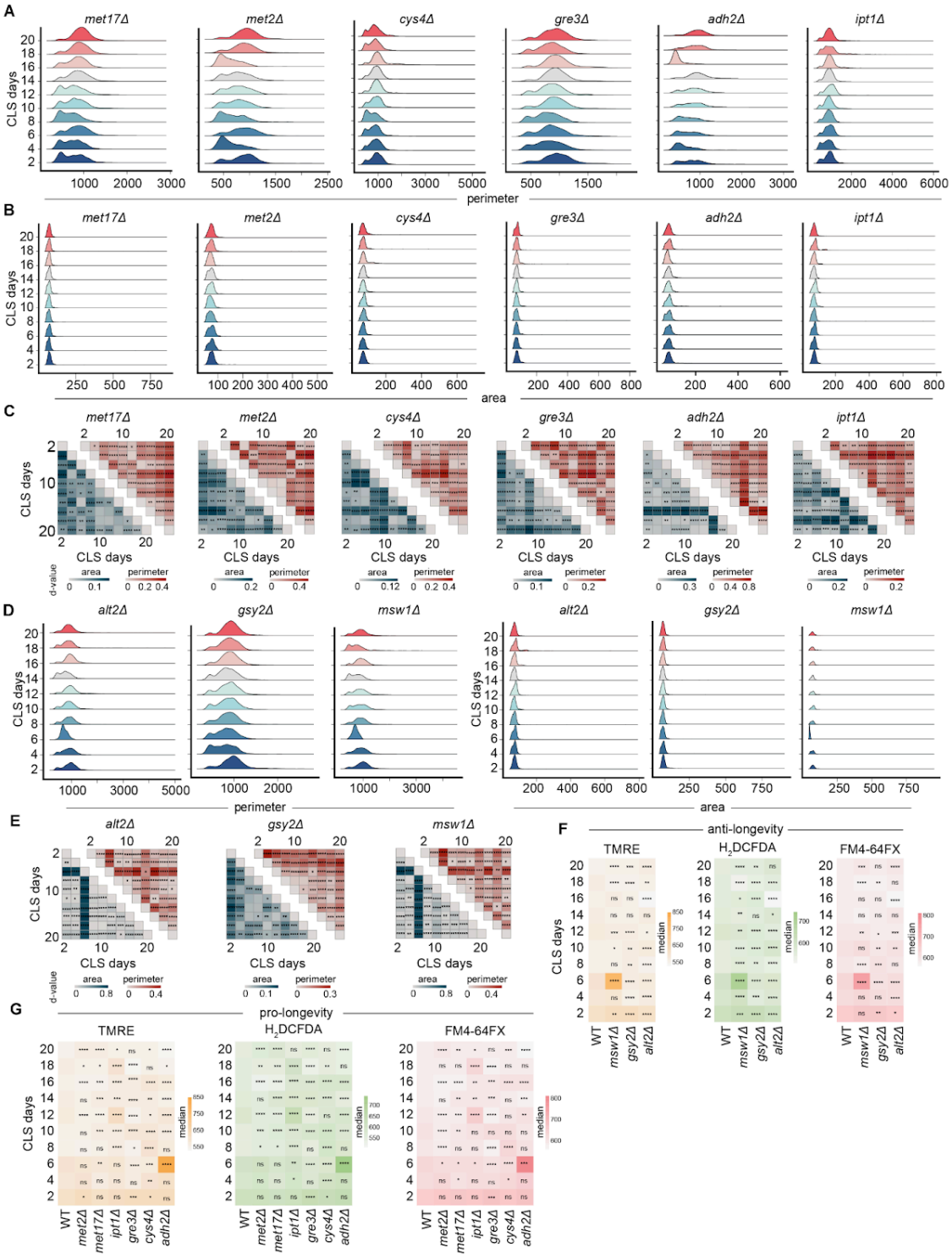


Figure 4.10: scCamAge Reveals Aging-Induced Cellular Morphometric Changes in Reported Pro- or Anti-Longevity Mutants.

(A) *Overlaid density plots illustrating the population-level distributions of single yeast cell perimeters in loss-of-function mutants reported to harbor pro-longevity phenotypes.* ***(B)*** *Overlaid density plots presenting the*

population-level distributions of single yeast cell areas in loss-of-function mutants reported to harbor pro-longevity phenotypes. ***(C)*** *Heat maps display the pairwise d-values from the Kolmogorov-Smirnov statistical test under specified conditions.* ***(D)*** *Overlaid density plots showing the population-level distributions of single yeast cell areas and perimeters in loss-of-function mutants reported to harbor anti-longevity phenotypes.* ***(E)*** *Heatmaps display pairwise d-values from the Kolmogorov-Smirnov test under specified conditions, with a p-value cutoff of 0.05, where *, **, ***, and **** denote p-values <0.05, <0.01, <0.001, and <0.0001, respectively.* ***(F)*** *Heatmaps showing the median fluorescence intensities of the indicated bioactivity dyes (TMRE, H_2DCFDA, and FM4-64FX) at the specified time points during chronological aging of yeast knockouts harboring anti-longevity phenotypes.* ***(G)*** *Heatmaps showing the median fluorescence intensities of the indicated bioactivity dyes (TMRE, H_2DCFDA, and FM4-64FX) at the specified time points during chronological aging of yeast knockouts harboring pro-longevity phenotypes.*

4.3.4 scCamAge Unveiled Evolutionary Conservation of Aging-Related Morphometrics and Bioactivities

Aging-related cellular processes and morphometric alterations, including cellular hypertrophy, modified mitochondrial structure, and diminished nucleolar size, are evolutionarily conserved from yeast to humans. Utilizing these common characteristics, we posited that the scCamAge model, trained on yeast chronological aging datasets, could proficiently categorize senescent cells in human cell cultures. To evaluate this, we created a dataset of camptothecin (CPT)-induced senescence utilizing primary human fibroblasts subjected to different CPT concentrations **(Figure 4.11A-D)**. The quantification of Senescence-Associated β-Galactosidase (SA-β-gal) staining verified elevated populations of senescent cells at escalating dosages of CPT. Images obtained by phase-contrast microscopy of unstained fibroblasts were subsequently segmented to extract individual cells and classified according to CPT exposure **(Figure 4.11C)**. Notably, the yeast-trained scCamAge model accurately recognized and classified senescent fibroblasts in both datasets without further training or adaption, exhibiting its ability to generalize across species and aging situations through shared morphometric characteristics **(Figure 4.11E, F)**. The scCamAge model was constructed on the InceptionResNet architecture utilizing a transfer learning methodology, and when implemented without further training, scCamAge markedly surpassed the baseline InceptionResNet model, confirming the evolutionary conservation of these morphometric characteristics **(Figure 4.11F, I, J)**. Furthermore, the incorporation of bioactivity characteristics significantly improved classification precision, underscoring the evolutionary preservation of aging-related bioactivities **(Figure 4.11G)**. Ultimately, to further assess these results, we applied scCamAge to an independent replicative senescence dataset from Sturm et al **(Figure 4.11D, H)** (Sturm et al. 2022). ScCamAge surpassed the baseline InceptionResNet model once more, without further training or weight reallocation. The combined embeddings (scCamAge and bioactivities) yielded the most precise

predictions **(Figure 4.11F, H, K)**. These data indicate that aging-induced cellular phenotypes, their morphometric markers, and related bioactivities are significantly preserved despite a billion years of evolutionary divergence between yeast and human fibroblasts.

4.3.5 Improving the Generalizability of scCamAge Using Human Fibroblast Senescence Models

We enhanced the functionalities of scCamAge by modifying its yeast-specific predictive models for applications in human senescence through a transfer learning methodology **(Figure 4.11L)**. The yeast-trained scCamAge model underwent additional training on the CPT-induced senescence dataset of human fibroblasts, exhibiting significantly enhanced proficiency in identifying senescence characteristics in human cells **(Figure 4.11L)**. For this adaptation, we employed phase-contrast micrographs obtained at five CPT doses previously validated for senescence induction via SA-β-gal staining. Transfer learning was utilized, incorporating scCamAge's foundational model with its Bioactivity and Morphometry modules, and trained on the CPT-induced human fibroblast senescence dataset. The model reached optimal performance at the 135th epoch of 500, with a One-vs-All AUC of 0.72-0.91, an accuracy of 52.4% in multi-class classification (random accuracy = 16.7%), and almost 86% accuracy in the binary classification of senescent versus non-senescent cells. The precision and recall scores for binary classification were approximately 92% and 88%, respectively **(Figure 4.11M-O)**. These findings confirm scCamAge's capacity to generalize across species, establishing it as a reliable instrument for detecting senescence in human cells.

To facilitate extensive study, we created a specialized library consisting of single-cell image datasets from yeast (chronological aging, bioactivity predictions, heat stress) and human fibroblasts (CPT- and replication-induced senescence). This resource, together with hyperparameter-optimized models (yeast scCamAge, scCamAge including Bioactivity and Morphometry, and human-specific scCamAge models), is accessible as Docker images and Jupyter notebooks for replication and benchmarking **(Figure 4.11P)**. In conclusion, our extensive resource, comprising varied single-cell imaging datasets and optimized models, will function as an essential toolset for enhancing research on cellular aging and senescence.

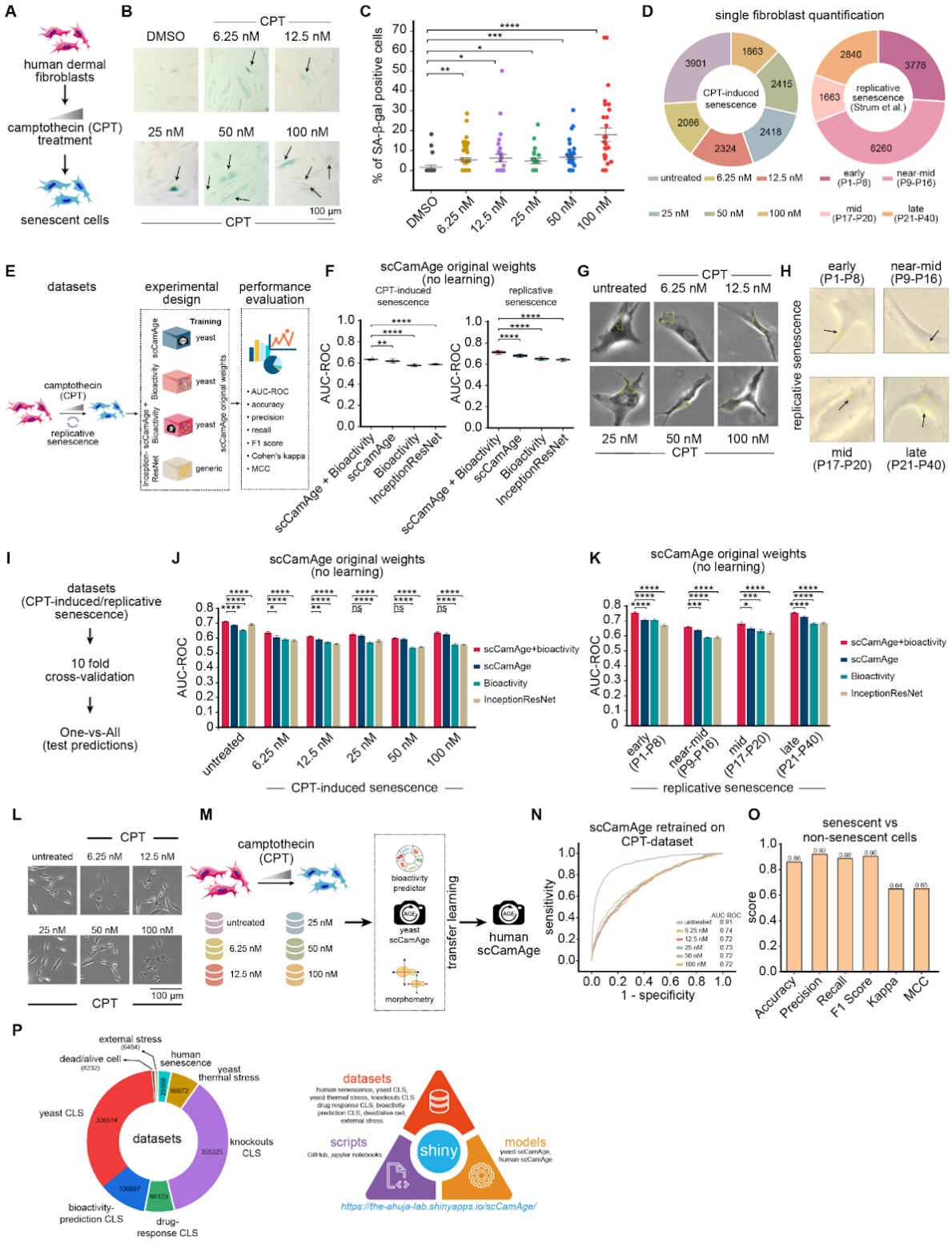

A
human dermal fibroblasts
camptothecin (CPT) treatment
senescent cells
B
CPT
DMSO
6.25 nM
12.5 nM
25 nM
50 nM
100 nM
100 µm
C
% of SA-β-gal positive cells
DMSO
6.25 nM
12.5 nM
25 nM
50 nM
100 nM
D
single fibroblast quantification
CPT-induced senescence
replicative senescence (Strum et al.)
untreated
6.25 nM
12.5 nM
25 nM
50 nM
100 nM
early (P1-P8)
near-mid (P9-P16)
mid (P17-P20)
late (P21-P40)
E
datasets
experimental design
performance evaluation
camptothecin (CPT)
replicative senescence
Training
yeast
scCamAge
Bioactivity
Inception-ResNet
scCamAge original weights
generic
AUC-ROC
accuracy
precision
recall
F1 score
Cohen's kappa
MCC
F
scCamAge original weights (no learning)
CPT-induced senescence
replicative senescence
AUC-ROC
scCamAge + Bioactivity
scCamAge
Bioactivity
InceptionResNet
G
CPT
untreated
6.25 nM
12.5 nM
25 nM
50 nM
100 nM
H
early (P1-P8)
near-mid (P9-P16)
mid (P17-P20)
late (P21-P40)
replicative senescence
I
datasets (CPT-induced/replicative senescence)
10 fold cross-validation
One-vs-All (test predictions)
J
scCamAge original weights (no learning)
AUC-ROC
scCamAge+bioactivity
scCamAge
Bioactivity
InceptionResNet
untreated
6.25 nM
12.5 nM
25 nM
50 nM
100 nM
CPT-induced senescence
K
scCamAge original weights (no learning)
AUC-ROC
scCamAge+bioactivity
scCamAge
Bioactivity
InceptionResNet
early (P1-P8)
near-mid (P9-P16)
mid (P17-P20)
late (P21-P40)
replicative senescence
L
CPT
untreated
6.25 nM
12.5 nM
25 nM
50 nM
100 nM
100 µm
M
camptothecin (CPT)
untreated
6.25 nM
12.5 nM
25 nM
50 nM
100 nM
bioactivity predictor
yeast scCamAge
morphometry
transfer learning
human scCamAge
N
scCamAge retrained on CPT-dataset
sensitivity
1 - specificity
AUC-ROC
untreated 0.91
6.25 nM 0.74
12.5 nM 0.72
25 nM 0.73
50 nM 0.72
100 nM 0.72
O
senescent vs non-senescent cells
score
Accuracy
Precision
Recall
F1 Score
Kappa
MCC
P
external stress
dead/alive cell
human senescence
yeast thermal stress
yeast CLS
knockouts CLS
bioactivity-prediction CLS
drug-response CLS
datasets
scripts
models
shiny
https://the-ahuja-lab.shinyapps.io/scCamAge/

Figure 4.11: scCamAge Reveals Evolutionarily Conservation of Aging-mediated Cellular Morphological Changes

(A) *Schematic diagram depicting the key steps used for camptothecin (CPT)-induced senescence in primary human fibroblasts.* ***(B)*** *Representative micrographs depicting camptothecin (CPT)-induced senescent cells marked using the Senescence Associated Beta Galactosidase assay (SA-βgal assay) staining on primary human fibroblasts. Different concentrations of camptothecin were used, and dimethylsulfoxide (DMSO) served as a solvent control.* ***(C)*** *The mean whisker plot depicted the percentage of SA-β-gal positively stained cells per image in the indicated conditions. Asterisks indicate statistical significance based on the Wilcoxon Rank Sum test, with a p-value cutoff of 0.05, where *, **, ***, and **** correspond to p-values <0.05, <0.01, <0.001, and <0.0001, respectively.* ***(D)*** *Donut pie chart representing the number of single-fibroblast cell images in CPT-induced and replicative senescence datasets in the indicated conditions.* ***(E)*** *Schematic diagram depicting the analysis workflow used to compare the performance of scCamAge, aging-associated bioactivities alone, scCamAge plus bioactivities, and generic InceptionResNet on the CPT-induced senescence dataset or replicative senescence dataset.* ***(F)*** *Box plot depicting the AUC of the testing CPT-induced senescence dataset and replicative senescence dataset. Asterisks denote statistical significance using the Wilcoxon Rank Sum test. The p-value cutoff used is 0.05. *, **, ***, and **** refer to p-values <0.05, <0.01, <0.001, and <0.0001, respectively.* ***(G)*** *Representative micrographs showcasing the top feature (highlighted in yellow) identified by the scCamAge explainability module for CPT-induced senescence.* ***(H)*** *Representative micrographs showcasing the top feature (highlighted in yellow) identified by the scCamAge explainability module in the replicative senescence dataset.* ***(I)*** *Workflow representing the method used to calculate the AUC scores using the One-vs-All method.* ***(J)*** *Bar plots depicting the AUC-ROC values across each class of CPT-induced senescence.* ***(K)*** *Bar plots depicting the AUC-ROC values across each class of replicative senescence. Data are represented as mean ± SEM. The p-value was calculated using the Mann-Whitney U test with a default significance cutoff of 0.05, where *, **, ***, and **** represent p-values <0.05, <0.01, <0.001, and <0.0001, respectively.* ***(L)*** *Representative micrographs depicting camptothecin (CPT)-induced senescent cells at the indicated concentrations. The scale bar corresponds to 100 µm.* ***(M)*** *Schematic diagram depicting the analysis workflow used to create human scCamAge in which yeast scCamAge, alongside bioactivity and morphometry predictions, were used.* ***(N)*** *The area under the ROC Curve (AUC-ROC) illustrates the One-vs-All AUC-ROC values under specified conditions.* ***(O)*** *Bar plot depicting various evaluation parameters depicting the prediction power of human-scCamAge.* ***(P)*** *The schematic diagram represents the offerings of the scCamAge resources.*

4.4 Discussion

Despite the genetic difference observed among different species and the resultant variability in their aging mechanisms, it is remarkable that the molecular systems governing aging exhibit significant conservation between Saccharomyces cerevisiae and humans (He et al. 2018; Weber et al. 2023). This conservation is especially significant given the approximately 1 billion years of evolutionary divergence between these groups. Budding yeast is a prominent eukaryotic model system employed in aging research and the

evaluation of anti-aging therapies, noted for its brief lifespan and compatibility with high-throughput screenings. The extensive acceptance is due to its suitability for genetic modification and chemical alterations, as well as its effectiveness in large-scale experimental configurations (Zimmermann et al. 2018; Barardo et al. 2017; Li et al. 2020; Fehrmann et al. 2013). Several anti-aging drugs, such as spermidine, rapamycin, resveratrol, and metformin, currently in clinical studies, have exhibited significant anti-aging benefits in yeast. This research emphasizes the common aging pathways between yeast and humans and shows the utility of yeast as an experimental model for discovering and evaluating new anti-aging therapies. Additional compelling evidence encompasses the age-related activation of molecular processes identified as hallmarks of human aging (López-Otín et al. 2023). Common aging phenotypic markers shared between humans and yeast encompass the accumulation of reactive oxygen species (ROS), an increase in damaged organelles and proteins, a decline in genomic stability and plasma membrane integrity, telomeric alterations, and the induction of apoptosis or programmed cell death, among others (Janssens & Veenhoff 2016). These considerations unequivocally underscore the significance and pertinence of budding yeast in preliminary large-scale screenings to identify prospective chemical anti-aging agents.

This work presents scCamAge, a context-aware prediction engine utilizing transfer learning for the image-based estimation of chronological age, assessment of aging-related biological activities, and identification of morphometric alterations. scCamAge elucidates the cellular areas essential for predicting chronological age, providing insights into aging mechanisms using explainable artificial intelligence (XAI). This novel methodology addresses existing constraints by providing a more thorough comprehension of aging. scCamAge's primary strength lies in its capacity to elucidate the intricate relationships among aging-related processes, including as genomic instability, reactive oxygen species generation, and mitochondrial composition and dynamics. Although prior deep learning systems have exhibited exceptional efficacy in identifying senescent cells in both cultured environments and tissue sections, to our knowledge, none have offered predictions of bioactivities directly from micrographs (Heckenbach et al. 2022; Kusumoto et al. 2021). This supplementary feature acts as a preliminary proof of concept for forthcoming frameworks to create image-based prediction models by concurrently utilizing bioactivities. This methodology, while concentrating on the aging process, may also be applicable to other scientific disciplines.

We meticulously evaluated scCamAge by validating it using separate and reserved datasets, as well as on datasets pertaining to pro-longevity pharmacological therapies and genetic knockouts(Barardo et al. 2017; de Magalhães & Toussaint 2004). All these investigations validate the precision and dependability of

scCamAge in forecasting cellular age and identifying aging-related alterations. Moreover, scCamAge's capacity to disclose substantial molecular changes and anti-aging reactions in yeast cells exposed to temperature stimuli highlights its value in investigating dynamic cellular processes. The scCamAge model, developed using yeast datasets, demonstrates its capacity to accurately predict human fibroblast senescence (whether chemically or replication-induced), underscoring its cross-species relevance and the evolutionary conservation of aging-related morphometric alterations. This feature broadens the utility of scCamAge and emphasizes the importance of investigating aging processes in various organisms to reveal shared biological principles. scCamAge's capability to deliver rapid, high-resolution single-cell age predictions, comprehensive molecular insights, and cellular morphometric analyses serves as a significant resource for investigating aging-related phenomena across various biological settings.

Upon applying the unmodified scCamAge model to human fibroblast senescence datasets without further training iterations, we noted markedly enhanced performance relative to the foundational InceptionResNet model (vision: v0.10.0, inception_v3) (Szegedy et al. 2016). This signifies that, besides catching yeast-specific characteristics related to the cell wall, scCamAge also detects other non-peripheral traits in human fibroblasts. Recent advancements employing state-of-the-art AI methodologies have concentrated on identifying subcellular compartments that exhibit significant spatiotemporal alterations during senescence induction. However, these techniques predominantly capture morphological changes through deep learning, neglect aging-related bioactivities in their predictions, frequently lack publicly accessible datasets or finalized models, and are not generalized concerning the model system. scCamAge delivers a comprehensive repository of high-quality single-cell image datasets and final prediction models (yeast and human scCamAge) to the community, together with the necessary toolkit for a thorough knowledge of cellular aging and its conserved pathways. In conclusion, scCamAge serves as the inaugural proof of concept in aging research, providing a comprehensive methodology for examining image-based cellular aging dynamics.

Limitations of the study

Despite its formidable capabilities, scCamAge possesses numerous restrictions. The yeast scCamAge model was initially created exclusively using CLS datasets, which may restrict its applicability to RLS assays. Nonetheless, its structure can be modified using transfer learning to incorporate RLS data. Secondly, the bioactivity prediction models were evaluated using a limited spectrum of extreme drug doses, which may not adequately represent aging-related physiological values. This constraint can be circumvented by integrating more extensive information with diverse geroprotector concentrations and genetic biosensors like Rad52-GFP (Styles et al. 2016; Mattiazzi Usaj et al. 2021) for the assessment of

DNA damage. Third, the ROIs in chronologically aged yeast cells, typically situated in the cell wall, may indicate yeast-specific characteristics such as budding activity, hence restricting their applicability to mammalian cells. The medications and genetic loss-of-function mutants employed to validate the scCamAge model are chosen for their substantial influence on lifespan, guaranteeing observable effects. Nonetheless, these pronounced effects may constrain the model's capacity to precisely identify and quantify more nuanced variations in longevity. The model can efficiently detect significant changes resulting from these interventions, but it may lack sensitivity to modest, gradual modifications in the aging process. Fifth, although transfer learning has demonstrated efficacy in adapting scCamAge from yeast to human datasets, it possesses intrinsic limits. The efficacy of transfer learning is largely contingent upon the resemblance between source and target domains; therefore, substantial disparities in cellular shape or bioactivities between yeast and human cells may restrict its efficiency. The neural network weights trained on yeast chronological aging datasets may not adequately encompass the intricacies of human senescence phenotypes, potentially resulting in diminished model sensitivity to nuanced characteristics. Furthermore, some of the learned features identified by the Explainability module, such as the focus on the yeast cell wall, may be yeast-specific and not directly translatable to mammalian cells. The model was also validated using interventions with strong, established effects on longevity, which may limit its sensitivity in detecting more subtle or nuanced variations in the aging process. The sixth notable restriction of scCamAge is the lack of comparable models and datasets for benchmarking purposes. Our literature survey revealed a limited number of studies employing brightfield imaging for the prediction of senescent cells, with none specifically focusing on chronological aging in yeast. Moreover, current human cell-based resources do not provide publicly accessible models or corresponding raw data, such as phase-contrast or brightfield pictures, hence hindering direct comparisons with scCamAge (Heckenbach et al. 2022; Kusumoto et al. 2021; Duran et al. 2024). A further disadvantage of scCamAge is the reliance on YeastSpotter for cell segmentation, leading to an approximate loss of 28.8% of cells due to suboptimal segmentation; this problem could be alleviated by employing sophisticated techniques such as Cellpose (Stringer et al. 2021) or StarDist (Weigert & Schmidt 2022) for enhanced precision. Ultimately, a restricted quantity of aging-related bioactivities was incorporated, although these markedly improved the model's efficacy on human datasets. Incorporating indicators for telomere length, cell cycle activity, and autophagy could enhance scCamAge’s prediction capability across various biological systems.

Conclusion

Chapter 5: Conclusion and Summary

The research presented in this thesis has successfully demonstrated the sheer significance of expert biological feature extraction in unraveling complex biological processes through the development of three unique computational tools: deepGraphh, scCamAge, and EcTracker. Each tool addresses a unique data modality, i.e., molecular structures, cellular images, and gene expression data, highlighting the power of tailored feature engineering in enhancing our understanding of biological systems. Using graph-based deep learning in deepGraphh to extract complex molecular features enabled accurate prediction of compound activity, highlighting the highest significance of structural representation in chemoinformatics. The innovative combination of single-cell imaging and bioactivity data in scCamAge to extract spatiotemporal features was critical in predicting cellular aging and identifying conserved morphological markers of senescence. Similarly, EcTracker's ability to extract biologically relevant features from single-cell RNA sequencing data, such as ectopic gene expression and regulatory networks, provided revolutionary insights into cell identity and response to genetic alterations.

The effectiveness of these tools in combination emphasizes the utmost importance of identifying and taking advantage of informative biological properties. With advanced computational techniques, this research has overcome the limitations of traditional approaches, enabling a deeper and more complex understanding of biological systems. The ability to evaluate different types of data with the help of these specialized tools emphasizes the advantage of a multi-omics approach to answering complex biological questions.

While this thesis presents powerful and accessible web-based tools, users must be aware of their inherent limitations. The primary limitations of deepGraphh are related to scalability and model robustness; as a graph-based deep learning tool, it is computationally resource-intensive and performs optimally on larger, diverse datasets, with its predictions requiring subsequent experimental validation. For EcTracker, the main constraints are dataset availability and scope, as it currently accepts only human samples and lacks modules for pseudotime trajectory or pathway analysis, limiting its broader application. The limitations of scCamAge concern model robustness and generalizability; its remarkable cross-species performance relies on evolutionarily conserved aging features that may not be present in all biological systems, and its validation on interventions with strong longevity effects may limit its sensitivity to more subtle aging variations. Future research should focus on the further advancement of the utility and capabilities of these tools. For deepGraphh, this could involve studies in explainable artificial intelligence methods to better understand the structural properties underlying activity predictions and broadening its application to span a wider range of biological endpoints. For scCamAge, future research should focus on incorporating dynamic imaging data to better capture temporal morphological change and the application of the model

to study other cellular processes beyond the aging domain, potentially in disease settings. For EcTracker, future research could involve the incorporation of spatial transcriptomics data to understand gene expression in tissue context and the creation of more sophisticated methodologies for deducing regulatory networks.

In short, this thesis adds to the increasing literature that emphasizes the vital connection between successful biological feature extraction and the progress of biological knowledge. The tools constructed here are precious assets to the scientific community that facilitate a better understanding of molecular functions, cellular characteristics, and the complex regulatory mechanisms that guide life. Ongoing improvement and application of these and related tools present tremendous opportunities for future advancement in a broad range of fields in biology and medicine.

Glossary

Glossary of Key Terms

Blood-Brain Barrier (BBB): A highly selective, protective layer of cells that lines the blood vessels in the brain. It controls the passage of substances from the blood into the central nervous system, posing a significant challenge for drug delivery to the brain.

Cheminformatics: A field of science that uses computational and informational techniques to solve a range of problems in the field of chemistry, particularly in designing and discovering new drugs.

Convolutional Neural Network (CNN): A type of deep learning model specialized in analyzing visual data. CNNs are highly effective at recognizing patterns in images, making them ideal for tasks like identifying cellular features in microscopy images.

Deep Learning: A subfield of machine learning based on artificial neural networks with multiple layers (deep architectures). These models can automatically learn complex patterns and features from large datasets like images, text, or molecular structures.

Feature Engineering: The process of using domain knowledge to select, transform, or create the most relevant input variables (features) for a machine learning model. Good feature engineering is critical for creating accurate predictive models.

Graph Neural Network (GNN): A specialized type of neural network designed to operate directly on graph-structured data. In this thesis, GNNs are used to process molecules as graphs, where atoms are nodes and chemical bonds are edges, to learn their structural properties.

High-Dimensional Data: Datasets where the number of features (e.g., genes, molecular descriptors) is vastly larger than the number of samples (e.g., patients, compounds). This presents a major challenge for traditional statistical and machine learning methods.

High-Throughput Screening (HTS): An automated process used in drug discovery that allows researchers to rapidly test tens of thousands of chemical compounds for a specific biological activity, such as inhibiting a target protein.

Interpretability: In the context of AI and machine learning, this refers to the ability to explain why a model made a certain prediction. An interpretable model is not a "black box," allowing researchers to understand its decision-making process and gain biological insights.

Pharmacogenomics: The study of how an individual's genetic makeup affects their response to drugs. It is a key component of personalized medicine, aiming to tailor treatments based on a patient's genetic profile.

Physicochemical Properties: The physical and chemical characteristics of a molecule, such as its size (molecular weight), solubility in water (logS), and affinity for fatty environments (logP). These properties govern how a drug is absorbed, distributed, and metabolized in the body.

QSAR (Quantitative Structure-Activity Relationship): A computational modeling approach that aims to find a mathematical relationship between the chemical structure of a molecule and its biological activity. QSAR models are widely used to predict the activity of new, untested compounds.

scRNA-seq (Single-cell RNA sequencing): A powerful genomic technology that measures the gene expression levels in thousands of individual cells simultaneously. This provides a high-resolution view of cellular heterogeneity, function, and responses to perturbations within a tissue or sample.

Senescence: A biological process in which a cell ages and permanently stops dividing but does not die. Cellular senescence is a key driver of the aging process and is implicated in many age-related diseases.

SMILES (Simplified Molecular Input Line Entry System): A standardized text-based notation for representing chemical structures using a simple string of ASCII characters. It provides a compact and machine-readable way to encode molecules.

Transfer Learning: A machine learning technique where a model trained on one large task (e.g., recognizing objects in general images) is repurposed and fine-tuned for a second, more specific task (e.g., identifying features in cell images). This is particularly useful when data for the specific task is limited.

Appendix

Figure 2.7C: Schematic diagram depicting the human metabolic map, with the overlaid red dots representing the metabolites, predicted as BBB+ (larger version)

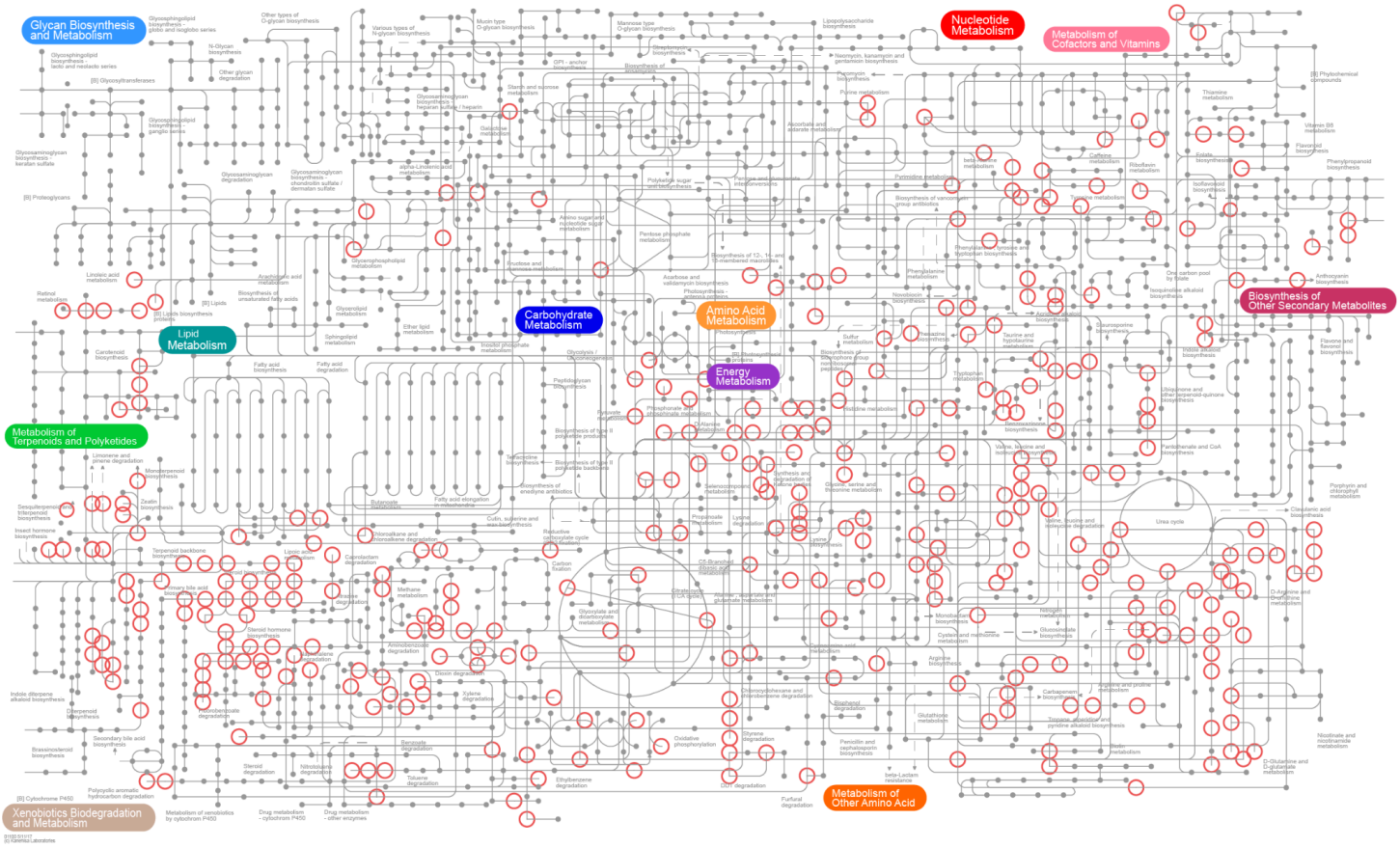


Table 5: Tabular representation containing details about the hyperparameter tuning performed for augmenting the model performances	
link for the table	https://docs.google.com/spreadsheets/d/1bZ98h0MI-YFobFsmYuRCCRSHeHrHGsSR/edit?usp=sharing&ouid=117433293510948571996&rtpof=true&sd=true

References

Abdelaziz, A. et al., 2016. Consensus modeling for HTS assays using in silico descriptors calculates the best balanced accuracy in Tox21 challenge. *Frontiers in environmental science*, 4. Available at: http://dx.doi.org/10.3389/fenvs.2016.00002.

Aibar, S. et al., 2017. SCENIC: Single-cell regulatory network inference and clustering. *bioRxiv*. Available at: http://dx.doi.org/10.1101/144501.

Alvarez, M.J. et al., 2016. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature genetics*, 48(8), pp.838–847.

Alvarsson, J. et al., 2016. Large-scale ligand-based predictive modelling using support vector machines. *Journal of cheminformatics*, 8(1), p.39.

Amezquita, R.A. et al., 2020. Publisher Correction: Orchestrating single-cell analysis with Bioconductor. *Nature methods*, 17(2), p.242.

Anders, S. & Huber, W., 2010. Differential expression analysis for sequence count data. *Nature precedings*. Available at: http://dx.doi.org/10.1038/npre.2010.4282.2.

Anon, *Genetic Algorithms Predict an Individual's Risk for a Specific Disease, Behavior, or Physical Trait*,

Bagchi, S. et al., 2019. In-vitro blood-brain barrier models for drug screening and permeation studies: an overview. *Drug design, development and therapy*, 13, pp.3591–3605.

Barardo, D. et al., 2017. The DrugAge database of aging-related drugs. *Aging cell*, 16(3), pp.594–597.

Ben David, G. et al., 2023. Can epigenetics predict drug efficiency in mental disorders? *Cells (Basel, Switzerland)*, 12(8). Available at: http://dx.doi.org/10.3390/cells12081173.

Bertoni, M. et al., 2021. Bioactivity descriptors for uncharacterized chemical compounds. *Nature communications*, 12(1), p.3932.

Buettner, F. et al., 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2), pp.155–160.

Cao, J. et al., 2020. A human cell atlas of fetal gene expression. *Science (New York, N.Y.)*, 370(6518), p.eaba7721.

Cao, Y. et al., 2008. ChemmineR: a compound mining framework for R. *Bioinformatics (Oxford, England)*, 24(15), pp.1733–1734.

Capecchi, A., Probst, D. & Reymond, J.-L., 2020. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of cheminformatics*, 12(1), p.43.

Chen, B. & Wild, D.J., 2010. PubChem BioAssays as a data source for predictive models. *Journal of molecular graphics & modelling*, 28(5), pp.420–426.

Chen, C., Liao, Y. & Peng, G., 2022. Connecting past and present: single-cell lineage tracing. *Protein & cell*, 13(11), pp.790–807.

Cheng, L. et al., 2022. gutMGene: a comprehensive database for target genes of gut microbes and microbial metabolites. *Nucleic acids research*, 50(D1), pp.D795–D800.

Chen, G., Ning, B. & Shi, T., 2019. Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in genetics*, 10, p.317.

Chen, T. & Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1603.02754.

Chithrananda, S., Grand, G. & Ramsundar, B., 2020. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/2010.09885.

Chi, Z., 2002. MLP classifiers: overtraining and solutions. In *Proceedings of ICNN'95 - International Conference on Neural Networks*. ICNN'95 - International Conference on Neural Networks. IEEE, pp. 2821–2824.

Choi, J.R. et al., 2020. Single-cell RNA sequencing and its combination with protein and DNA analyses. *Cells (Basel, Switzerland)*, 9(5), p.1130.

Cui, X. et al., 2019. In silico prediction of drug-induced rhabdomyolysis with machine-learning models and structural alerts. *Journal of applied toxicology: JAT*, 39(8), pp.1224–1232.

Dablander, M., 2024. Investigating graph neural networks and classical feature-extraction techniques in activity-cliff and molecular property prediction. *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/2411.13688.

Datlinger, P. et al., 2017. Pooled CRISPR screening with single-cell transcriptome readout. *Nature methods*, 14(3), pp.297–301.

Dixit, A. et al., 2016. Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7), pp.1853–1866.e17.

Doshida, Y. et al., 2023. Single-cell RNA sequencing to detect age-associated genes that identify senescent cells in the liver of aged mice. *Scientific reports*, 13(1), p.14186.

Duran, I. et al., 2024. Detection of senescence using machine learning algorithms based on nuclear features. *Nature communications*, 15(1), p.1041.

Du, X. et al., 2022. Morphological components detection for super-depth-of-field bio-micrograph based on deep learning. *Microscopy (Oxford, England)*, 71(1), pp.50–59.

Fehrmann, S. et al., 2013. Aging yeast cells undergo a sharp entry into senescence unrelated to the loss of mitochondrial membrane potential. *Cell reports*, 5(6), pp.1589–1599.

Feng, D. et al., 2019. Single Cell Explorer, collaboration-driven tools to leverage large-scale single cell RNA-seq data. *BMC genomics*, 20(1), p.676.

Fernández-Torras, A. et al., 2022. Connecting chemistry and biology through molecular descriptors. *Current opinion in chemical biology*, 66(102090), p.102090.

Finak, G. et al., 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology*, 16(1), p.278.

Forcato, M., Romano, O. & Bicciato, S., 2021. Computational methods for the integrative analysis of single-cell data. *Briefings in bioinformatics*, 22(1), pp.20–29.

Franzén, O. & Björkegren, J.L.M., 2020. alona: a web server for single-cell RNA-seq analysis. *Bioinformatics (Oxford, England)*, 36(12), pp.3910–3912.

Garain, A. et al., 2021. GRA_net: A deep learning model for classification of age and gender from facial images. *IEEE access: practical innovations, open solutions*, 9, pp.85672–85689.

Gardeux, V. et al., 2017. ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics* , 33(19), pp.3123–3125.

Gautam, V. et al., 2022. deepGraphh: AI-driven web service for graph-based quantitative structure-activity relationship analysis. *Briefings in bioinformatics*, 23(5). Available at: http://dx.doi.org/10.1093/bib/bbac288.

Gel, B. & Serra, E., 2017. karyoploteR: an R/Bioconductor package to plot customizable linear genomes displaying arbitrary data. *bioRxiv*. Available at: http://dx.doi.org/10.1101/122838.

Genga, R.M.J. et al., 2019. Single-cell RNA-sequencing-based CRISPRi screening resolves molecular drivers of early human endoderm development. *Cell reports*, 27(3), pp.708–718.e10.

Goh, G.B. et al., 2017. SMILES2Vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv [stat.ML]*. Available at: http://arxiv.org/abs/1712.02034.

GTEx Consortium, 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (New York, N.Y.)*, 369(6509), pp.1318–1330.

Guo, J. et al., 2022. Aging and aging-related diseases: from molecular mechanisms to interventions and treatments. *Signal transduction and targeted therapy*, 7(1), p.391.

Guo, Z. et al., 2022. Graph-based molecular representation learning. *arXiv [q-bio.QM]*. Available at: http://arxiv.org/abs/2207.04869.

Gupta, K. et al., 2021. The Cellular basis of loss of smell in 2019-nCoV-infected individuals. *Briefings in bioinformatics*, 22(2), pp.873–881.

Haghverdi, L. et al., 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5), pp.421–427.

Han, X. et al., 2020. Construction of a human cell landscape at single-cell level. *Nature*, 581(7808), pp.303–309.

Hao, Y. et al., 2024. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology*, 42(2), pp.293–304.

Heckenbach, I. et al., 2022. Nuclear morphology is a deep learning biomarker of cellular senescence. *Nature aging*, 2(8), pp.742–755.

He, C., Zhou, C. & Kennedy, B.K., 2018. The yeast replicative aging model. *Biochimica et biophysica acta. Molecular basis of disease*, 1864(9 Pt A), pp.2690–2696.

He, K. et al., 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. Available at: http://dx.doi.org/10.1109/cvpr.2016.90.

He, K. et al., 2017. Mask R-CNN. *arXiv [cs.CV]*. Available at: http://arxiv.org/abs/1703.06870.

Hodzic, E., 2016. Single-cell analysis: Advances and future perspectives. *Udruzenje basicnih mediciniskih znanosti [Bosnian journal of basic medical sciences]*, 16(4), pp.313–314.

Hoerl, A.E. & Kennard, R.W., 2000. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics: a journal of statistics for the physical, chemical, and engineering sciences*, 42(1), p.80.

Holland, C.H., Tanevski, J., et al., 2020. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome biology*, 21(1), p.36.

Holland, C.H., Szalai, B. & Saez-Rodriguez, J., 2020. Transfer of regulatory knowledge from human to mouse for functional genomics analysis. *Biochimica et biophysica acta. Gene regulatory mechanisms*, 1863(6), p.194431.

Ho, T.K., 2002. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*. 3rd International Conference on Document Analysis and Recognition. IEEE Comput. Soc. Press, pp. 278–282 vol.1.

Hsu, Y.-C., 2015. Theory and practice of lineage tracing. *Stem cells (Dayton, Ohio)*, 33(11), pp.3197–3204.

Hua, Y. et al., 2021. In silico prediction of chemical-induced hematotoxicity with machine learning and deep learning methods. *Molecular diversity*, 25(3), pp.1585–1596.

Hwang, B., Lee, J.H. & Bang, D., 2021. Author Correction: Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 53(5), p.1005.

Hwang, B., Lee, J.H. & Bang, D., 2018. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8), pp.1–14.

Jaeger, S., Fulle, S. & Turk, S., 2018. Mol2vec: Unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58(1), pp.27–35.

Jain, A. & Tuteja, G., 2019. TissueEnrich: Tissue-specific gene enrichment analysis. *Bioinformatics (Oxford, England)*, 35(11), pp.1966–1967.

Jaksik, R. et al., 2024. Multiomics-based feature extraction and selection for the prediction of lung cancer survival. *International journal of molecular sciences*, 25(7), p.3661.

James, T. et al., 2025. Whole-genome phenotype prediction with machine learning: Open problems in bacterial genomics. *arXiv [q-bio.GN]*. Available at: http://arxiv.org/abs/2502.07749.

Janssens, G.E. & Veenhoff, L.M., 2016. Evidence for the hallmarks of human aging in replicatively aging yeast. *Microbial cell (Graz, Austria)*, 3(7), pp.263–274.

Jiang, D. et al., 2021. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of cheminformatics*, 13(1), p.12.

Jindal, A. et al., 2018. Discovery of rare cells from voluminous single cell expression data. *Nature communications*, 9(1), p.4719.

Ji, Z., Zhou, W. & Ji, H., 2017. Single-cell regulome data analysis by SCRAT. *Bioinformatics (Oxford, England)*, 33(18), pp.2930–2932.

Kalra, S. et al., 2020. Analysis of single-cell transcriptomes links enrichment of olfactory receptors with cancer cell differentiation status and prognosis. *Communications biology*, 3(1), p.506.

Kalra, S. et al., 2021. Challenges and possible solutions for decoding extranasal olfactory receptors. *The FEBS journal*, 288(14), pp.4230–4241.

Kipf, T.N. & Welling, M., 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1609.02907.

Kode, H. & Barkana, B.D., 2023. Deep learning- and expert knowledge-based feature extraction and performance evaluation in breast histopathology images. *Cancers*, 15(12). Available at: http://dx.doi.org/10.3390/cancers15123075.

Korolev, V., Mitrofanov, A. & Korotcov, A., 2020. Graph Convolutional Neural Networks as "General-Purpose" Property Predictors: The Universality and Limits of Applicability. *J. Chem. Inf. Model*, 60, pp.22–28.

Korsunsky, I. et al., 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods*, 16(12), pp.1289–1296.

Koudonas, A. et al., 2024. DNA methylation as drug sensitivity marker in RCC: A systematic review. *Epigenomes*, 8(3), p.28.

Krentzel, D., Shorte, S.L. & Zimmer, C., 2023. Deep learning in image-based phenotypic drug discovery. *Trends in cell biology*, 33(7), pp.538–554.

Kurata, H. & Tsukiyama, S., 2022. ICAN: interpretable cross-attention network for identifying drug and target protein interactions. *bioRxiv*. Available at: http://dx.doi.org/10.1101/2022.08.04.502877.

Kusumoto, D. et al., 2021. Anti-senescent drug screening by deep learning-based morphology senescence scoring. *Nature communications*, 12(1), p.257.

Lähnemann, D. et al., 2020. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1), p.31.

Landrum, G., RDKit. Available at: https://www.rdkit.org [Accessed August 28, 2025].

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), pp.357–359.

Li, H. et al., 2023. A knowledge-guided pre-training framework for improving molecular representation learning. *Nature communications*, 14(1), p.7568.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), pp.1754–1760.

Li, J. & Jiang, X., 2021. Mol‐BERT: An effective molecular representation with BERT for molecular property prediction. *Wireless communications and mobile computing*, 2021(1), pp.1–7.

Li, P. et al., 2024. Improving drug response prediction via integrating gene relationships with deep

learning. *Briefings in bioinformatics*, 25(3). Available at: http://dx.doi.org/10.1093/bib/bbae153.

Li, Q. et al., 2017. A sequential EMT-MET mechanism drives the differentiation of human embryonic stem cells towards hepatocytes. *Nature communications*, 8(1), p.15166.

Li, Y. et al., 2020. A programmable fate decision landscape underlies single-cell aging in yeast. *Science (New York, N.Y.)*, 369(6501), pp.325–329.

López-Otín, C. et al., 2023. Hallmarks of aging: An expanding universe. *Cell*, 186(2), pp.243–278.

Love, M.I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *bioRxiv*. Available at: http://dx.doi.org/10.1101/002832.

Lo, Y.-C. et al., 2018. Machine learning in chemoinformatics and drug discovery. *Drug discovery today*, 23(8), pp.1538–1546.

Lu, A.X. et al., 2019. YeastSpotter: accurate and parameter-free web segmentation for microscopy images of yeast cells. *Bioinformatics (Oxford, England)*, 35(21), pp.4525–4527.

Lusci, A., Pollastri, G. & Baldi, P., 2013. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *Journal of chemical information and modeling*, 53(7), pp.1563–1575.

de Magalhães, J.P. & Toussaint, O., 2004. GenAge: a genomic and proteomic network map of human ageing. *FEBS letters*, 571(1-3), pp.243–247.

Ma, J. et al., 2015. Deep neural nets as a method for quantitative structure-activity relationships. *Journal of chemical information and modeling*, 55(2), pp.263–274.

Malusare, A. & Aggarwal, V., 2024. Improving molecule generation and drug discovery with a knowledge-enhanced generative model. *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/2402.08790.

Mao, Y. et al., 2024. Phenotype prediction from single-cell RNA-seq data using attention-based neural networks. *Bioinformatics (Oxford, England)*, 40(2). Available at: http://dx.doi.org/10.1093/bioinformatics/btae067.

Mattiazzi Usaj, M. et al., 2021. Single-cell image analysis to explore cell-to-cell heterogeneity in isogenic populations. *Cell systems*, 12(6), pp.608–621.

Mayr, A. et al., 2018. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical science (Royal Society of Chemistry: 2010)*, 9(24), pp.5441–5451.

McCarthy, D.J. et al., 2017. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics (Oxford, England)*, 33(8), pp.1179–1186.

McKenna, A. et al., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9), pp.1297–1303.

McKenna, A. & Gagnon, J.A., 2019. Recording development with single cell dynamic lineage tracing. *Development (Cambridge, England)*, 146(12). Available at: http://dx.doi.org/10.1242/dev.169730.

Meng, F. et al., 2021. A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Scientific data*, 8(1), p.289.

Mitchell, J.B.O., 2014. Machine learning methods in chemoinformatics. *Wiley interdisciplinary reviews. Computational molecular science*, 4(5), pp.468–481.

Mittal, A. & Ahuja, G., 2023. Advancing chemical carcinogenicity prediction modeling: opportunities and challenges. *Trends in pharmacological sciences*, 44(7), pp.400–410.

Morgan, H.L., 1965. The generation of a unique machine description for chemical structures-A technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2), pp.107–113.

Moriwaki, H. et al., 2018. Mordred: a molecular descriptor calculator. *Journal of cheminformatics*, 10(1), p.4.

Mowlaei, M.E. & Shi, X., 2023. FSF-GA: A Feature Selection Framework for Phenotype Prediction Using Genetic Algorithms. *Genes*, 14(5). Available at: http://dx.doi.org/10.3390/genes14051059.

Natarajan, K.N. et al., 2019. Comparative analysis of sequencing technologies for single-cell transcriptomics. *Genome biology*, 20(1), p.70.

Nikopoulou, C. et al., 2023. Spatial and single-cell profiling of the metabolome, transcriptome and epigenome of the aging mouse liver. *Nature aging*, 3(11), pp.1430–1445.

Patel, M.V., 2018. iS-CellR: a user-friendly tool for analyzing and visualizing single-cell RNA sequencing data. *Bioinformatics (Oxford, England)*, 34(24), pp.4305–4306.

Paul, D. et al., 2021. Artificial intelligence in drug discovery and development. *Drug discovery today*, 26(1), pp.80–93.

Pedregosa, F. et al., 2012. Scikit-learn: Machine Learning in Python. *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1201.0490.

Phillip, J.M. et al., 2017. Biophysical and biomolecular determination of cellular age in humans. *Nature biomedical engineering*, 1(7). Available at: http://dx.doi.org/10.1038/s41551-017-0093.

Pratapa, A., Doron, M. & Caicedo, J.C., 2021. Image-based cell phenotyping with deep learning. *Current opinion in chemical biology*, 65, pp.9–17.

Probst, D. & Reymond, J.-L., 2018. A probabilistic molecular fingerprint for big data settings. *Journal of cheminformatics*, 10(1), p.66.

Qin, L., Dong, G. & Peng, J., 2020. Chemical-protein interaction extraction via ChemicalBERT and attention guided graph convolutional networks in parallel. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE. Available at: http://dx.doi.org/10.1109/bibm49941.2020.9313234.

Qi, Z., 2024. Cellular Phenotypes: Insights into the Mechanisms of Cell Function and Behaviour. *Biochemistry & Pharmacology: Open Access*, 13(4), pp.1–2.

Ren, Y.Y. et al., 2016. Predicting the aquatic toxicity mode of action using logistic regression and linear discriminant analysis. *SAR and QSAR in environmental research*, 27(9), pp.721–746.

Riniker, S. & Landrum, G.A., 2013. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics*, 5(1), p.26.

Ritchie, M.E. et al., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7), p.e47.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), pp.139–140.

Rogers, D. & Hahn, M., 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5), pp.742–754.

Satija, R. et al., 2015. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5), pp.495–502.

Schindelin, J. et al., 2012. Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7), pp.676–682.

Seidl, P. et al., 2023. Enhancing activity prediction models in drug discovery with the ability to understand human language. *arXiv [q-bio.BM]*. Available at: http://arxiv.org/abs/2303.03363 [Accessed March 31, 2025].

Sloan, K., Nandi, C.D. & Linz, P., 2012. Analytical and biological methods for probing the blood-brain barrier. *Annu. Rev. Anal. Chem*, 5, pp.505–531.

Stringer, C. et al., 2021. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1), pp.100–106.

Stuart, T. et al., 2019. Comprehensive integration of single-cell data. *Cell*, 177(7), pp.1888–1902.e21.

Stuart, T. & Satija, R., 2019. Integrative single-cell analysis. *Nature reviews. Genetics*, 20(5), pp.257–272.

Sturm, G. et al., 2022. A multi-omics longitudinal aging dataset in primary human fibroblasts with mitochondrial perturbations. *Scientific data*, 9(1), p.751.

Styles, E.B. et al., 2016. Exploring quantitative yeast phenomics with single-cell analysis of DNA damage foci. *Cell systems*, 3(3), pp.264–277.e10.

Subramanian, A. et al., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp.15545–15550.

Sun, M. et al., 2020. Graph convolutional networks for computational drug development and discovery. *Briefings in bioinformatics*, 21(3), pp.919–935.

Sushko, I. et al., 2011. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *Journal of computer-aided molecular design*, 25(6), pp.533–554.

Svetnik, V. et al., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, 43(6), pp.1947–1958.

Szegedy, C. et al., 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. Available at:

http://dx.doi.org/10.1109/cvpr.2016.308.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 73(3), pp.273–282.

Uhlén, M. et al., 2015. Proteomics. Tissue-based map of the human proteome. *Science (New York, N.Y.)*, 347(6220), p.1260419.

Vamathevan, J. et al., 2019. Applications of machine learning in drug discovery and development. *Nature reviews. Drug discovery*, 18(6), pp.463–477.

Veličković, P. et al., 2017. Graph Attention Networks. *arXiv [stat.ML]*. Available at: http://arxiv.org/abs/1710.10903.

Vilar, S., Chakrabarti, M. & Costanzi, S., 2010. Prediction of passive blood-brain partitioning: straightforward and effective classification models based on in silico derived physicochemical descriptors. *Journal of molecular graphics & modelling*, 28(8), pp.899–903.

Wang, S. et al., 2019. Smiles-Bert. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. BCB '19: 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. New York, NY, USA: ACM. Available at: http://dx.doi.org/10.1145/3307339.3342186.

Way, G.P. et al., 2021. Predicting cell health phenotypes using image-based morphology profiling. *Molecular biology of the cell*, 32(9), pp.995–1005.

Weber, L. et al., 2023. Phenotyping senescent mesenchymal stromal cells using AI image translation. *Current research in biotechnology*, 5(100120), p.100120.

Weigert, M. & Schmidt, U., 2022. Nuclei Instance Segmentation and Classification in Histopathology Images with Stardist. In *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. 2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC). IEEE. Available at: http://dx.doi.org/10.1109/isbic56247.2022.9854534.

Whitehead, T.M. et al., 2019. Imputation of assay bioactivity data using deep learning. *Journal of chemical information and modeling*, 59(3), pp.1197–1204.

White, J.J., 2024. New methods for predicting drug molecule activity using deep learning. *Bioscience methods*, 15(0). Available at: https://bioscipublisher.com/index.php/bm/article/view/3800 [Accessed March 31, 2025].

Wishart, D.S. et al., 2018. HMDB 4.0: the human metabolome database for 2018. *Nucleic acids research*, 46(D1), pp.D608–D617.

Wolf, F.A., Angerer, P. & Theis, F.J., 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1). Available at: http://dx.doi.org/10.1186/s13059-017-1382-0.

Wu, S.-H.S., Lee, J.-H. & Koo, B.-K., 2019. Lineage tracing: Computational reconstruction goes beyond the limit of imaging. *Molecules and cells*, 42(2), pp.104–112.

Wu, Z. et al., 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* , 9(2), pp.513–530.

Xiong, Z. et al., 2020. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *Journal of medicinal chemistry*, 63(16), pp.8749–8760.

Yang, K. et al., 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8), pp.3370–3388.

Yao, F., Coquery, J. & Lê Cao, K.-A., 2012. Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. *BMC bioinformatics*, 13(1), p.24.

Yap, C.W., 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, 32(7), pp.1466–1474.

Zadorozhny, K. et al., 2025. Similarity-quantized relative difference learning for improved molecular activity prediction. *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/2501.09103 [Accessed March 31, 2025].

Zaragozá, R., 2020. Transport of amino acids across the blood-brain barrier. *Frontiers in physiology*, 11, p.973.

Zhang, Q.-Y. & Aires-de-Sousa, J., 2007. Random forest prediction of mutagenicity from empirical physicochemical descriptors. *ChemInform*, 38(15). Available at: http://dx.doi.org/10.1002/chin.200715208.

Zheng, G.X.Y. et al., 2017. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1), p.14049.

Zhu, H. et al., 2023. Human PBMC scRNA-seq-based aging clocks reveal ribosome to inflammation balance as a single-cell aging hallmark and super longevity. *Science advances*, 9(26), p.eabq7599.

Zhu, X. et al., 2017. Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *bioRxiv*. Available at: http://dx.doi.org/10.1101/110759.

Zimmermann, A. et al., 2018. Yeast as a tool to identify anti-aging compounds. *FEMS yeast research*, 18(6). Available at: http://dx.doi.org/10.1093/femsyr/foy020.