Journal Title Here, 2024, pp. 1–12

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Predicting the pro-longevity or anti-longevity effect of model organism genes with enhanced Gaussian noise augmentation-based contrastive learning on protein-protein interaction networks

Ibrahim Alsaggaf,[1] Alex A. Freitas[2] and Cen Wan[1,*]

[1]School of Computing and Mathematical Sciences, Birkbeck, University of London, WC1E 7HX, London, United Kingdom and [2]School of Computing, University of Kent, CT2 7FS, Kent, United Kingdom

[*]Corresponding author. cen.wan@bbk.ac.uk

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Ageing is a highly complex and important biological process that plays major roles in many diseases. Therefore, it is essential to better understand the molecular mechanisms of ageing-related genes. In this work, we proposed a novel enhanced Gaussian noise augmentation-based contrastive learning (EGsCL) framework to predict the pro-longevity or anti-longevity effect of four model organisms' ageing-related genes by exploiting protein-protein interaction networks. The experimental results suggest that EGsCL successfully outperformed the conventional Gaussian noise augmentation-based contrastive learning methods and obtained state-of-the-art performance on three model organisms' predictive tasks when merely relying on protein-protein interaction network data. In addition, we use EGsCL to predict 10 novel pro-/anti-longevity mouse genes, and discuss the support for these predictions in the literature.

Key words: contrastive learning, protein-protein interaction networks, Gaussian noise data augmentation, ageing

Introduction

Ageing is a highly complex biological process that involves many genes and biological pathways [50, 8]; and despite significant progress in ageing-biology research, the precise molecular mechanisms of ageing are still not well understood [17, 8, 45]. In addition, ageing research is particularly important because ageing is a major driving factor for many diseases [12, 28, 31]; and so a better understanding of the effects of ageing-related genes could lead to new therapies that would potentially extend not only the longevity, but also the healthspan (period of health life) of individuals [28, 36, 37]. With the help of Artificial Intelligence (more specifically, machine learning), research has been carried out to predict new ageing-related genes or biomarkers, and to identify ageing-related biological pathways or processes [13, 65]. In this work, we focus on predicting the pro-longevity or anti-longevity effect of genes from four model organisms in ageing research (mouse, worm, fly and yeast). We cast this problem as a classification task from the perspective of supervised machine learning, where each instance (example) represents an ageing-related gene, each instance's class label indicates whether that gene has a pro-longevity or anti-longevity effect on the lifespan of an organism [59, 57] – based on such class labels as recorded in the GenAge database [9]. The predictive features are PPI network-based features.

Protein-protein interaction (PPI) networks are a type of biologically meaningful and relevant features that have been widely used in multiple bioinformatics tasks like protein function prediction [55, 61, 58] and disease-gene association prediction [35, 33, 19]. PPI networks have also been used for ageing research. Freitas et al. [16] first exploited PPI networks as a type of features to classify DNA repair genes into ageing-related or non-ageing-related genes. Fang et al. [14] classified ageing-related genes into DNA repair or non-DNA repair-related genes using PPI networks-based features. This type of features were also used for predicting ageing-related genes for flies [51], mice [15] and humans [27]. More recently, Magdaleno et al. [30] exploited PPI network features to predict ageing-related genes' dietary restriction associations, and Ribeiro et al. [46] used PPI network features to predict lifespan-extending chemical compounds for worms.

In this work, we propose a new contrastive learning-based framework to cope with PPI network features by developing two novel contrastive learning algorithms. In general, contrastive learning aims to learn a type of discriminative distribution where similar instances are pulled closer whilst different instances are pushed away. The conventional self-supervised contrastive learning methods like SimCLR [4] first create two

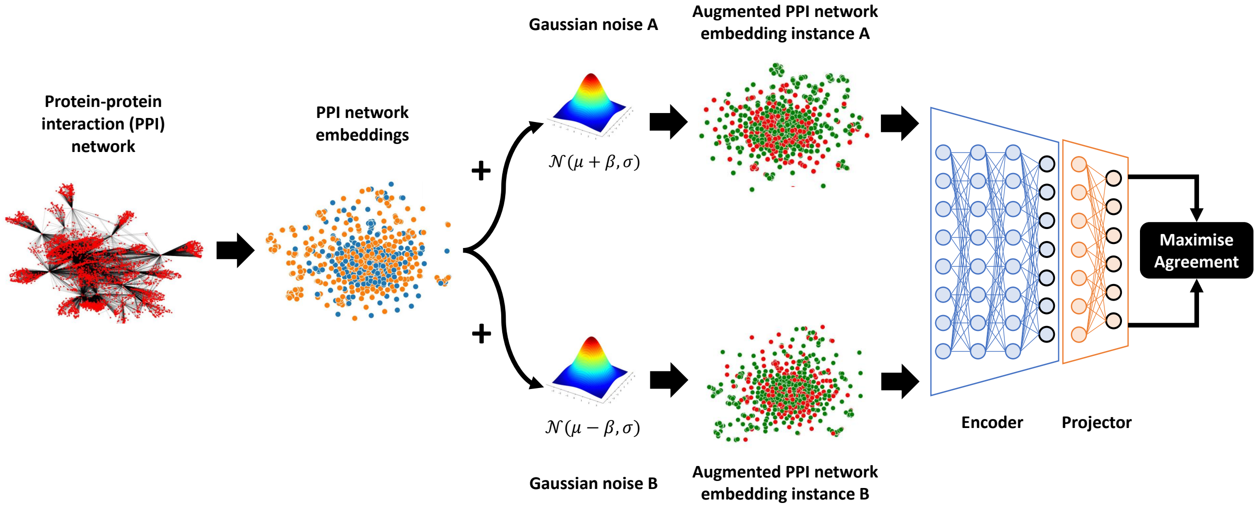
Protein-protein interaction (PPI) network
PPI network embeddings
Gaussian noise A
Augmented PPI network embedding instance A
$\mathcal{N}(\mu+\beta,\sigma)$
+
+
$\mathcal{N}(\mu-\beta,\sigma)$
Gaussian noise B
Augmented PPI network embedding instance B
Encoder
Projector
Maximise Agreement

Fig. 1. The flowchart for the proposed enhanced Gaussian noise augmentation-based contrastive learning (EGsCL) framework based on protein-protein interaction networks.

views for each instance by using different data augmentation strategies. For each target instance, two views that are generated from that target instance are treated as positive views, and all other views that are not generated from that target instance are treated as negative views. Then SimCLR optimises the network parameters to reduce the distance between two positive views, whilst enlarging the difference between positive views and negative views. The self-supervised learning paradigm was further extended to the supervised contrastive learning paradigm [23], where the definition of positive and negative views relies on the original labels of instances. For each target instance, views are considered positive if they are generated from those instances bearing the same class label as that target instance. *Vice versa*, the negative views are generated from the instances bearing different labels to that target instance. This supervised contrastive learning paradigm successfully demonstrated better predictive performance than the self-supervised contrastive learning paradigm.

Data augmentation plays a crucial role on contrastive learning and is usually considered domain-specific. For example, in the computer vision area, the mainstream augmentation methods [4, 5, 23, 22, 47] rely on spatial and colour transformations (e.g. random cropping and Gaussian blur) to create different views of original images. In the natural language processing area, text paraphrasing and word replacement [42] are usually used as augmentation methods. Several works introduced different data augmentation strategies for bioinformatics research. For example, Ciortan and Defrance [6] and Wan et al. [60] used a type of random masking strategy to deal with single-cell RNA-seq expression profiles. Alsaggaf et al. [1] and Xu et al. [62] adopted a noise-addition approach by randomly adding Gaussian noise vectors to gene expression profiles to create different views. In this work, we propose a new Gaussian noise-based data augmentation strategy that adopts a mean-shifting approach to enlarge the difference between views to improve the contrastive learning process.

The remainder of this paper is organized as follows. The materials and methods section introduces the newly proposed enhanced Gaussian noise augmentation-based contrastive learning algorithms, followed by the results section and the discussion section, where a further analysis of the proposed algorithms was conducted. Finally, the conclusion section summaries this paper's major findings and mentions some future research directions.

Materials and methods

Enhanced Gaussian noise augmentation-based contrastive learning.

In general, the proposed enhanced Gaussian noise augmentation-based contrastive learning (EGsCL) framework learns a type of discriminative feature representations based on protein-protein interaction (PPI) networks. As shown in Figure 1, given a protein-protein interaction network, EGsCL first extracts a type of PPI network embedding features using the well-known *node2vec* [18] method. Then the PPI network embedding features were used to create augmented instances (a.k.a. views) by using different Gaussian noises. For one d-dimensional PPI network embedding instance x in a given dataset, EGsCL randomly draws two Gaussian noises from two different Gaussian distributions, i.e. $\mathcal{N}(\mu+\beta,\sigma)$ and $\mathcal{N}(\mu-\beta,\sigma)$, where μ and σ denote the mean and standard deviation of the dataset, whilst β is a shifting hyperparameter that is used to manipulate the differences between those two Gaussian distributions. Those two Gaussian noises are then added with the values of x, leading to two different augmented PPI network embedding instances. After creating a pair of augmented instances for all individual PPI network embedding instances in the dataset, a new sample set that includes all those augmented instances is used as inputs for the contrastive learning networks consisting of an encoder and a projector. The contrastive learning networks optimise the parameters by adopting the conventional supervised or self-supervised contrastive learning strategies, i.e. minimising the dissimilarity between each augmented instance and its corresponding positive augmented instance(s), whilst maximising the dissimilarity to its corresponding negative augmented instances. To cope with the classification tasks in this work, we used the EGsCL-learned feature representations to train support vector machines to predict the pro-longevity or anti-longevity effect of different model organisms' genes. The notations used in this paper are summarised in Table 1.

Table 1. The list of notations used in this paper.

Notation	Description
x	A d-dimensional PPI network embedding instance.
$\mathcal{N}(\mu, \sigma)$	A Gaussian distribution, where μ and σ denote its mean and standard deviation.
β	A hyperparameter that is used to manipulate a Gaussian distribution by shifting its mean.
$\mathcal{X}$	A training dataset.
$\mathcal{Y}$	A set of class labels.
$\mathcal{B}$	A set of m-sized training batches.
$\mathcal{E}$	A contrastive learning encoder.
$\mathcal{P}$	A contrastive learning projection head.
τ	A temperature hyper-parameter.
b	A m-sized training batch.
$\mathcal{S}$	A set to store two different augmentations of each original instance.
z	A d-dimensional Gaussian noise.
$\tilde{x}$	An augmentation (i.e. view) of the original instance x.
$\mathcal{L}_i^{SL}$	The supervised contrastive loss function value for the i^{th} instance.
$\mathcal{H}_i^{+}$	A set of projections of positive augmented instances w.r.t. $\tilde{x}_i$.
$\mathcal{H}_i$	A set of projections of all positive and negative augmented instances w.r.t. $\tilde{x}_i$.
$\lvert\mathcal{H}_i^{+}\rvert$	The number of positive augmented instances w.r.t. $\tilde{x}_i$.
$\mathcal{F}(\cdot)$	The cosine similarity.
$\mathcal{V}(\tilde{x})$	A variable that maps an augmented instance $\tilde{x}$ to its original instance x.
$\mathcal{L}_i^{SSL}$	The self-supervised contrastive loss function value for the i^{th} instance.

Algorithm 1: Supervised enhanced Gaussian noise augmentation-based contrastive learning (Sup-EGsCL).

```
Input: a training dataset $\mathcal{X}$;
       a class labels set $\mathcal{Y}$ for all individual instances $x$ in $\mathcal{X}$;
       initialise a set of $m$-sized batches $\mathcal{B}$;
       initialise an untrained contrastive learning encoder $\mathcal{E}$;
       initialise an untrained contrastive learning projection head $\mathcal{P}$;
       initialise a temperature hyperparameter $\tau$;
       initialise a mean-shift hyperparameter $\beta$.
Output: a trained contrastive learning encoder $\mathcal{E}^*$.
1  foreach $b \in \mathcal{B}$ do
2      initialise an empty variable $\mathcal{L}_b$ for the loss function value of batch $b$;
3      initialise an empty set $\mathcal{S}$ for the augmented instances of $\mathcal{X}$;
4      foreach $x_i \in b$ do
5          $z_a \sim \mathcal{N}(\mu + \beta, \sigma)$;
6          $z_b \sim \mathcal{N}(\mu - \beta, \sigma)$;
7          $x_{ia} \leftarrow x_i + z_a$;
8          $x_{ib} \leftarrow x_i + z_b$;
9          $\mathcal{S} \leftarrow \mathcal{S} \cup x_{ia}$;
10         $\mathcal{S} \leftarrow \mathcal{S} \cup x_{ib}$;
11     end
12     foreach $\tilde{x}_i \in \mathcal{S}$ do
13         initialise an empty variable $\mathcal{L}_i^{SL}$ for the loss function value for $\tilde{x}_i$;
14         initialise an empty positive augmented instances projection set $\mathcal{H}_i^{+}$ for $\tilde{x}_i$;
15         initialise an empty projection set $\mathcal{H}_i$ for all augmented instances except $\tilde{x}_i$;
16         foreach $\tilde{x}_j \in \mathcal{S}$ do
17             if $\tilde{x}_i \neq \tilde{x}_j$ then
18                 $h_j \leftarrow \mathcal{P}(\mathcal{E}(\tilde{x}_j))$;
19                 $\mathcal{H}_i \leftarrow \mathcal{H}_i \cup h_j$;
20                 if $\mathcal{Y}(\tilde{x}_i) == \mathcal{Y}(\tilde{x}_j)$ then
21                     $\mathcal{H}_i^{+} \leftarrow \mathcal{H}_i^{+} \cup h_j$;
22                 end
23             end
24         end
25         $h_i \leftarrow \mathcal{P}(\mathcal{E}(\tilde{x}_i))$;
26         $\mathcal{L}_i^{SL} \leftarrow$ LossFunction$(h_i, \mathcal{H}_i^{+}, \mathcal{H}_i, \tau)$;
27         $\mathcal{L}_b \leftarrow \mathcal{L}_b + \mathcal{L}_i^{SL}$;
28     end
29     $\mathcal{L}_b \leftarrow \mathcal{L}_b / 2m$;
30     optimize $\mathcal{E}$ and $\mathcal{P}$ by using $\mathcal{L}_b$;
31 end
32 return $\mathcal{E}^*$
```

Algorithms 1 and S1 (in supplementary file 1) show two different pseudocodes of the proposed enhanced Gaussian noise augmentation-based contrastive learning (EGsCL) algorithms working with supervised and self-supervised contrastive learning loss functions, respectively. In Algorithm 1, supervised enhanced Gaussian noise augmentation-based contrastive learning (Sup-EGsCL) takes a training dataset $\mathcal{X}$ and a corresponding class label set $\mathcal{Y}$ as inputs and initialised five variables, i.e. a set of m-sized batches $\mathcal{B}$, an untrained encoder $\mathcal{E}$, an untrained projection head $\mathcal{P}$, a temperature hyperparameter τ and a mean-shift hyperparameter β. From lines 1 to 31, Sup-EGsCL processes each batch of training instances b in turns. It creates an empty variable $\mathcal{L}_b$ to store the loss function value for b and an empty set $\mathcal{S}$ to store the augmented instances (a.k.a. views). For each training instance x_i in b (lines 4 - 11), two d-dimensional Gaussian noises, i.e. z_a and z_b, are randomly drawn from two different Gaussian distributions, i.e. $\mathcal{N}(\mu + \beta, \sigma)$ and $\mathcal{N}(\mu - \beta, \sigma)$, where μ and σ denote the mean and standard deviation of the training dataset $\mathcal{X}$. β is a hyperparameter that is used to adjust the differences between those two Gaussian distributions. Then z_a and z_b are added to x_i to create two different augmented instances, i.e. x_{ia} and x_{ib} (lines 7 - 8). Those two augmented instances are added to the set $\mathcal{S}$ (lines 9 - 10). After obtaining the complete set $\mathcal{S}$ that consists of all the augmented instances for the entire training dataset $\mathcal{X}$, from lines 12 - 28, Sup-EGsCL processes each augmented instance $\tilde{x}_i$ in $\mathcal{S}$ to compute the loss function value. It creates three empty variables, i.e. a variable $\mathcal{L}_i^{SL}$ for storing the supervised loss function value for $\tilde{x}_i$, a set $\mathcal{H}_i^{+}$ for storing the projections of positive augmented instances with respect to $\tilde{x}_i$, and a set $\mathcal{H}_i$ for storing the projections of all positive and negative augmented instances with respective to $\tilde{x}_i$. From lines 16 - 24, EGsCL defines the positive augmented instances according to the pre-defined class labels. Each augmented instance $\tilde{x}_j$ in $\mathcal{S}$ that is different from the target instance $\tilde{x}_i$ is added to $\mathcal{H}_i$ after getting its corresponding projection using the encoder $\mathcal{E}$ and the projector $\mathcal{P}$ (lines 17 - 19). Only the projections of those augmented instances bearing the same class label as $\tilde{x}_i$ will be considered as positive augmented instances with respect to $\tilde{x}_i$ and their projections will be added to $\mathcal{H}_i^{+}$ (lines 20 - 22). *Vice versa*, the negative augmented instances with respect to $\tilde{x}_i$ are defined as those augmented instances bearing different class labels to $\tilde{x}_i$. After obtained the completed sets of $\mathcal{H}_i^{+}$ and $\mathcal{H}_i$, Sup-EGsCL creates the projection of the target instance $\tilde{x}_i$ (line 25). Then Sup-EGsCL computes the loss function value $\mathcal{L}_i^{SL}$ that will then be added to $\mathcal{L}_b$ (lines 26 - 27). After processing all augmented instances in $\mathcal{S}$, the loss function value $\mathcal{L}_b$ will be normalised by $2m$ denoting the total number of augmented instances in $\mathcal{S}$, and both the encoder and the projector will be optimised (lines 29 - 30). The pseudocode will output a trained encoder $\mathcal{E}^*$ after processing all batches (line 32). Equation 1 defines the supervised contrastive loss function for the target instance $\tilde{x}_i$, where $\lvert\mathcal{H}_i^{+}\rvert$ denotes the number of positive augmented instances w.r.t. $\tilde{x}_i$, j denotes the indices of the positive augmented instances, and k denotes the indices of all augmented instances except i. $\mathcal{F}(\cdot)$ denotes the cosine similarity and τ is a temperature hyper-parameter that controls the strength of penalty on positives and negatives.

$$\mathcal{L}_i^{SL} = \frac{-1}{|\mathcal{H}_i^+|} \sum_{h_j \in \mathcal{H}_i^+} \log \frac{e^{\mathcal{F}(h_i,h_j)/\tau}}{\sum_{h_k \in \mathcal{H}_i} e^{\mathcal{F}(h_i,h_k)/\tau}} \qquad (1)$$

Algorithm S1 shows the pseudocode of the self-supervised enhanced Gaussian noise augmentation-based contrastive learning (Self-EGsCL) method, which shares the same initialisation and data augmentation process with the Sup-EGsCL method. The main difference between Algorithms 1 and S1 is the positive augmented instance selection strategy. As shown in lines 9 and 10, Self-EGsCL stores the original instance information for each augmented instance. For example, the value of variable $\mathcal{V}(x_{ia})$ is assigned as x_i, if x_{ia} is the augmented instance of x_i. In lines 22 - 24, for each augmented instance $\tilde{x}_i$, Self-EGsCL treated another augmented instance $\tilde{x}_j$ as a positive augmented instance, if both $\tilde{x}_i$ and $\tilde{x}_j$ are generated by using the same original instance (i.e. $\mathcal{V}(\tilde{x}_i) == \mathcal{V}(\tilde{x}_j)$). All other augmented instances in $\mathcal{S}$ are treated as negative augmented instances. Self-supervised EGsCL uses a similar loss function (Equation S1 in supplementary file 1) as Sup-EGsCL. Because there is only one positive augmented sample w.r.t. one single target augmented instance (i.e. $|\mathcal{H}_i^+| = 1$), Self-EGsCL does not normalise the loss function value $\mathcal{L}_i^{SSL}$.

Computational experiments

We evaluated the predictive performance of EGsCL using five different β values, i.e. 0.1, 0.2, 0.3, 0.4 and 0.5. We also compared EGsCL with the conventional Gaussian noise augmentation-based contrastive learning (GsCL) method, which also randomly draws two different Gaussian noises to create a pair of augmented instances for x, but from the same Gaussian distribution, i.e. $\mathcal{N}(\mu, \sigma)$. Therefore, GsCL is equivalent to the case when EGsCL's β value equals 0. We also compared with another GsCL variant with $\mathcal{N}(0, 1)$, which was used in [1] for cell type identification tasks. We used the well-known multi-layer perceptron (MLP) to create the encoder and the projection head of an EGsCL network. The encoder consists of three hidden layers and one output layer (i.e. the representation layer). The projection head consists of one hidden layer and one output layer. The ReLU activation function was used in both MLPs. We used Adam optimiser with a learning rate of 10^{-4} and a weight decay of 10^{-6}. The number of maximum training epochs was set to 1,000. We set the value of τ to 0.1 for the supervised contrastive loss and 0.07 for the self-supervised contrastive loss. Due to the small number of instances, we set the batch size as the same as the number of training instances. The proposed EGsCL methods were implemented by PyTorch [38] and Scikit-learn [40].

We created 12 datasets in total using the ageing-related genes for four model organisms, i.e. mouse, worm, fly and yeast, as reported in the GenAge database [53]. We generated three types of features based on the protein-protein interaction networks deposited in the STRING database (version 12.0) [52]. The first type of features is network embeddings learned by the well-known *node2vec* method [18] leading to a 128-dimensional vector for each individual protein included in the most informative combined score STRING PPI networks. The second type of features is binary PPI features, where the value of 1 denotes protein_a and protein_b have an interaction and the value of 0 means those two proteins do not have an interaction. The third type of features is the combination of both the network embedding and the binary PPI features. The characteristics of all 12 datasets are listed in Table 2. The numbers of instances for four different model organisms range between 124 and 718. The dimensionalities of binary features range between 5,957 and 17,438 and the combined features range between 6,085 and 17,566.

Table 2. Main characteristics of the created datasets.

Model Organisms		Mouse	Worm	Fly	Yeast
# Instances	Total	124	718	186	312
	Pro-longevity	80	239	117	34
	Anti-longevity	44	479	69	278
# Features	Embedding	128	128	128	128
	Binary	17438	16010	11535	5957
	Combined	17566	16138	11663	6085

Each generated dataset was split into two subsets, i.e. 80% of the instances were used for conducting a 10-fold cross-validation, and the remaining 20% of the instances were used to create a validation set for conducting model selection during the contrastive learning process. For each fold of the cross validation, after every 5 training epochs, we froze the encoder $\mathcal{E}$ and used it to transform the training folds, the validation set and the testing fold into the EGsCL feature representations. An SVM classifier was trained on the transformed training folds and then predicted the labels of the transformed validation set. The best encoder was selected according to the highest validation set predictive accuracy. The corresponding SVM classifier was used to predict the predictive accuracy of the transformed testing fold. We measured the predictive performance using three well-known metrics, i.e. Matthews correlation coefficient (MCC), F1 score and average precision (AP) score, which were also used as model selection criteria when reporting corresponding metrics' values.

Results

EGsCL successfully improved the predictive performance of GsCL when using different types of PPI features to predict the pro-longevity or anti-longevity effect of four model organisms' genes.

We first conducted pairwise comparisons between EGsCL and GsCL using supervised and self-supervised settings. In general, both Sup-EGsCL and Self-EGsCL outperformed Sup-GsCL and Self-GsCL, respectively. As shown in Table 3, when using the network embedding features to predict the longevity effects of mouse's genes, Sup-EGsCL with all different β values obtained higher MCC values and AP scores than Sup-GsCL with both $\mathcal{N}(0, 1)$ and $\mathcal{N}(\mu, \sigma)$, denoting by the double up arrows. The former with β values of 0.3 and 0.4 also obtained higher F1 scores than the latter. When using the binary PPI features, Sup-EGsCL with almost all β values except 0.1 obtained higher AP scores than Sup-GsCL. However the latter obtained higher MCC values and F1 scores. When using the combined features, Sup-EGsCL with β values of 0.3 and 0.5 obtained higher MCC values and F1 scores than Sup-GsCL. The former with all β values also outperformed the latter due to higher AP scores. In terms of Self-EGsCL, when using the network embedding features and binary PPI features, it outperformed Self-GsCL with both $\mathcal{N}(0, 1)$ and $\mathcal{N}(\mu, \sigma)$ according to the higher MCC values, F1 and AP scores obtained with different β values, as denoted by the single up arrows. When using the combined features, Self-EGsCL with β values of 0.1 and 0.2 obtained higher MCC values than Self-GsCL. It also obtained higher AP

Table 3. Predictive performance of Sup-EGsCL, Sup-GsCL, Self-EGsCL, Self-GsCL and the benchmark method.

mouse (*Mus musculus*)

Feature Types	Metrics	Sup-EGsCL					Sup-GsCL		Self-EGsCL					Self-GsCL		Benchmark
		β=0.1	β=0.2	β=0.3	β=0.4	β=0.5	$\mathcal{N}(0,1)$	$\mathcal{N}(\mu,\sigma)$	β=0.1	β=0.2	β=0.3	β=0.4	β=0.5	$\mathcal{N}(0,1)$	$\mathcal{N}(\mu,\sigma)$	
Embeddings	MCC	0.309 ⇑	0.366 ⇑	0.380 ⇑	0.397 ⇑	$\underline{\underline{\mathbf{0.427}}}$ ⇑	0.075	0.248	0.176	0.169	0.245 ↑	$\underline{0.285}$ ↑	$\underline{0.285}$ ↑	0.208	0.234	0.146
	F1	0.780	0.783	0.797 ⇑	$\underline{\underline{0.818}}$ ⇑	0.789	0.747	0.796	0.797 ↑	0.792 ↑	0.788	$\underline{0.799}$ ↑	$\underline{0.799}$ ↑	0.738	0.788	0.744
	AP	0.842 ⇑	0.844 ⇑	0.839 ⇑	$\underline{\underline{0.847}}$ ⇑	0.836 ⇑	0.774	0.826	0.837	0.820	0.828	0.811	$\underline{0.845}$ ↑	0.844	0.791	0.764
Binary	MCC	0.237	0.237	0.263	0.263	0.280	$\underline{\underline{0.373}}$	0.237	0.151	0.155	$\underline{0.212}$ ↑	0.168 ↑	0.175 ↑	0.142	0.157	0.325
	F1	0.794	0.800	0.800	0.800	0.800	$\underline{\underline{0.821}}$	$\underline{\underline{0.821}}$	$\underline{0.816}$ ↑	0.809	0.811	0.811	0.811	0.815	0.756	0.811
	AP	0.853	0.855 ⇑	0.855 ⇑	$\underline{\underline{0.856}}$ ⇑	0.855 ↑	0.850	0.853	$\underline{0.838}$ ↑	0.821	0.788	0.827 ↑	0.834 ↑	0.824	0.808	0.805
Combined	MCC	0.237	0.278	0.329 ⇑	0.270	$\underline{\underline{0.402}}$ ⇑	0.309	0.237	$\underline{0.367}$ ↑	0.343 ↑	0.254	0.234	0.288	0.271	0.334	0.371
	F1	0.787	0.792	$\underline{\underline{0.806}}$ ⇑	0.796	0.801 ⇑	0.796	0.787	0.768	0.768	0.771	0.771	0.776	$\underline{0.813}$	0.788	**0.826**
	AP	$\underline{\underline{\mathbf{0.860}}}$ ⇑	0.837 ⇑	0.838 ⇑	0.838 ⇑	0.839 ⇑	0.827	0.836	0.770	$\underline{0.826}$ ↑	0.788	0.783	0.811 ↑	0.794	0.798	0.813

worm (*Caenorhabditis elegans*)

Feature Types	Metrics	Sup-EGsCL					Sup-GsCL		Self-EGsCL					Self-GsCL		Benchmark
		β=0.1	β=0.2	β=0.3	β=0.4	β=0.5	$\mathcal{N}(0,1)$	$\mathcal{N}(\mu,\sigma)$	β=0.1	β=0.2	β=0.3	β=0.4	β=0.5	$\mathcal{N}(0,1)$	$\mathcal{N}(\mu,\sigma)$	
Embeddings	MCC	0.356	0.369	0.355	0.299	0.363	0.181	$\underline{\underline{0.377}}$	0.275 ↑	0.335 ↑	0.299 ↑	$\underline{0.350}$ ↑	0.306 ↑	0.177	0.269	0.367
	F1	0.550	0.539	0.550	0.548	0.538	0.466	$\underline{\underline{0.561}}$	$\underline{0.526}$ ↑	0.515	0.492	0.487	0.506	0.447	0.524	0.529
	AP	0.692 ⇑	0.696 ⇑	0.695 ⇑	0.697 ⇑	$\underline{\underline{\mathbf{0.698}}}$ ⇑	0.483	0.678	0.593 ↑	$\underline{0.597}$ ↑	0.590 ↑	0.593 ↑	0.579	0.500	0.587	0.685
Binary	MCC	0.367	0.383 ⇑	$\underline{\underline{\mathbf{0.387}}}$ ⇑	0.348	0.350	0.295	0.374	0.308	$\underline{0.346}$ ↑	0.301	0.308	0.313	0.316	0.308	0.377
	F1	$\underline{\underline{0.566}}$ ⇑	0.534	0.559 ⇑	0.553	0.551	0.551	0.555	0.496	0.503	0.494	0.499	0.520	$\underline{0.538}$	0.517	0.530
	AP	0.639	0.641	0.644	0.643	0.629	$\underline{\underline{0.663}}$	0.649	0.598	$\underline{0.638}$ ↑	0.603	0.615 ↑	0.607	0.546	0.607	0.664
Combined	MCC	0.344	0.352	0.338	0.347	$\underline{\underline{0.379}}$ ⇑	0.354	0.354	0.287 ↑	0.289 ↑	$\underline{0.316}$ ↑	0.293 ↑	0.284	0.211	0.285	0.369
	F1	0.578	0.584 ⇑	0.571	0.585 ⇑	$\underline{\underline{\mathbf{0.599}}}$ ⇑	0.544	0.579	0.496	0.493	$\underline{0.549}$ ↑	0.516	0.460	0.492	0.516	0.529
	AP	0.640	0.649	0.649	$\underline{\underline{0.662}}$ ⇑	0.651	0.629	0.658	0.670 ↑	$\underline{0.675}$ ↑	0.670 ↑	0.665 ↑	0.662 ↑	0.562	0.586	0.647

fly (*Drosophila melanogaster*)

Feature Types	Metrics	Sup-EGsCL					Sup-GsCL		Self-EGsCL					Self-GsCL		Benchmark
		β=0.1	β=0.2	β=0.3	β=0.4	β=0.5	$\mathcal{N}(0,1)$	$\mathcal{N}(\mu,\sigma)$	β=0.1	β=0.2	β=0.3	β=0.4	β=0.5	$\mathcal{N}(0,1)$	$\mathcal{N}(\mu,\sigma)$	
Embeddings	MCC	0.260 ⇑	0.231	$\underline{\underline{\mathbf{0.328}}}$ ⇑	0.278 ⇑	0.212	0.038	0.242	0.191	0.135	$\underline{0.198}$ ↑	0.144	0.191	-0.052	0.194	0.134
	F1	0.747	0.747	$\underline{\underline{0.771}}$ ⇑	0.765 ⇑	0.757	0.760	0.752	0.754	0.757	0.754	0.761	0.761	$\underline{0.769}$	0.759	0.725
	AP	0.761	0.760	0.757	0.765	0.741	0.691	$\underline{\underline{0.769}}$	0.779 ↑	0.756	0.762	0.739	$\underline{0.795}$ ↑	0.710	0.764	0.753
Binary	MCC	0.157	0.157	0.157	0.157	0.157	0.147	$\underline{\underline{0.207}}$	0.014	0.038	0.021	0.113 ↑	$\underline{0.135}$ ↑	0.015	0.101	0.270
	F1	0.756	0.756	0.756	0.756	0.756	0.736	$\underline{\underline{0.774}}$	0.733	$\underline{0.745}$ ↑	0.725	0.738 ↑	0.724	0.727	0.733	0.769
	AP	0.802	0.801	0.803	0.804	0.801	$\underline{\underline{0.806}}$	0.802	0.752 ↑	$\underline{0.783}$ ↑	0.751 ↑	0.768 ↑	0.747	0.748	0.722	0.826
Combined	MCC	0.283	0.267	0.283	0.275	$\underline{\underline{0.292}}$ ⇑	0.275	0.283	0.172 ↑	0.197 ↑	0.241 ↑	$\underline{0.244}$ ↑	0.180 ↑	0.116	0.094	0.230
	F1	0.771	0.771	$\underline{\underline{\mathbf{0.782}}}$	0.776	0.774	0.781	$\underline{\underline{\mathbf{0.782}}}$	0.768	0.765	0.762	0.759	0.774	0.767	$\underline{0.777}$	0.760
	AP	0.802	0.806	0.808	0.811	0.804	$\underline{\underline{\mathbf{0.838}}}$	0.802	0.744	0.708	0.689	0.689	0.687	$\underline{0.749}$	0.732	0.821

yeast (*Saccharomyces cerevisiae*)

Feature Types	Metrics	Sup-EGsCL					Sup-GsCL		Self-EGsCL					Self-GsCL		Benchmark
		β=0.1	β=0.2	β=0.3	β=0.4	β=0.5	$\mathcal{N}(0,1)$	$\mathcal{N}(\mu,\sigma)$	β=0.1	β=0.2	β=0.3	β=0.4	β=0.5	$\mathcal{N}(0,1)$	$\mathcal{N}(\mu,\sigma)$	
Embeddings	MCC	0.099	0.154	0.074	0.082	0.114	$\underline{\underline{0.219}}$	0.016	0.034	0.040	0.026	0.004	0.133	$\underline{0.152}$	0.023	**0.274**
	F1	0.130	0.163	0.083	0.090	0.130	$\underline{\underline{0.250}}$	0.130	0.153	0.107	0.090	0.073	0.090	$\underline{0.220}$	0.127	**0.297**
	AP	$\underline{\underline{0.393}}$ ⇑	0.329	0.384 ⇑	0.350	0.323	0.277	0.362	0.315	0.347	0.272	0.285	$\underline{0.444}$ ↑	0.359	0.254	**0.509**
Binary	MCC	0.040	0.103	0.103	0.095	0.103	$\underline{\underline{0.165}}$	0.082	0.010	0.019	0.024	0.073	0.010	0.066	$\underline{0.173}$	0.034
	F1	0.050	0.100	0.100	0.100	0.100	$\underline{\underline{0.167}}$	0.090	0.040	0.040	0.040	0.040	0.040	0.126	$\underline{0.247}$	0.050
	AP	$\underline{\underline{0.469}}$ ⇑	0.430	0.448 ⇑	0.418	0.417	0.408	0.435	$\underline{0.393}$	0.391	0.357	0.380	0.390	$\underline{0.393}$	0.374	0.397
Combined	MCC	0.089	0.089	0.089	0.089	0.089	0.066	$\underline{\underline{0.117}}$	0.171	$\underline{0.180}$ ↑	0.171	0.171	0.163	0.112	0.171	0.034
	F1	0.100	0.100	0.100	0.100	0.100	0.090	$\underline{\underline{0.130}}$	$\underline{0.180}$	$\underline{0.180}$	$\underline{0.180}$	$\underline{0.180}$	$\underline{0.180}$	0.150	$\underline{0.180}$	0.050
	AP	0.379 ⇑	0.367 ⇑	0.357 ⇑	$\underline{\underline{0.385}}$ ⇑	0.374 ⇑	0.321	0.346	0.306	0.310	0.325	0.301	0.287	0.263	$\underline{0.385}$	0.402

[1] ⇑: higher values obtained by Sup-EGsCL compared with Sup-GsCL with both $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,\sigma)$.

[2] ↑: higher values obtained by Self-EGsCL compared with Self-GsCL with both $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,\sigma)$.

[3] $\underline{\underline{\text{Double underline}}}$: the highest value between Sup-EGsCL and Sup-GsCL over all parameters.

[4] $\underline{\text{Underline}}$: the highest value between Self-EGsCL and Self-GsCL over all parameters.

[5] **Bold text**: the overall highest value for the model organism.

scores with β values of 0.2 and 0.5, though Self-GsCL with $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,\sigma)$ obtained higher F1 scores.

When predicting the longevity effects of worm's genes using the network embedding features, Sup-GsCL with $\mathcal{N}(\mu,\sigma)$ outperformed Sup-EGsCL, according to MCC values and F1 scores. However, Sup-EGsCL with all different β values obtained higher AP scores than Sup-GsCL with both $\mathcal{N}(\mu,\sigma)$ and $\mathcal{N}(0,1)$. When using the binary PPI features, Sup-EGsCL obtained higher MCC values with β values of 0.2 and 0.3. It also obtained higher F1 scores with β values of 0.1 and 0.3. However, Sup-GsCL obtained higher AP scores. When using the combined features, Sup-EGsCL with different β

values outperformed Sup-GsCL with both $\mathcal{N}(\mu, \sigma)$ and $\mathcal{N}(0, 1)$, according to the higher MCC values, F1 and AP scores. Analogously, as shown in Table 3, Self-EGsCL with almost all different β values using the network embedding features outperformed Self-GsCL with both $\mathcal{N}(\mu, \sigma)$ and $\mathcal{N}(0, 1)$, according to the higher MCC values and AP scores. It also obtained a higher F1 score with a β value of 0.1. When using the binary PPI features, Self-EGsCL with a β value of 0.2 obtained a higher MCC value and a higher AP score than Self-GsCL, but the latter obtained a higher F1 score with $\mathcal{N}(0, 1)$. When using the combined features, Self-EGsCL with almost all different β values obtained higher MCC values and AP scores. It also obtained a higher F1 score than Self-GsCL with a β value of 0.3.

When using network embedding features to predict the longevity effects of fly's genes, Sup-EGsCL with β values of 0.3 and 0.4 obtained higher MCC values and F1 scores than Sup-GsCL with both $\mathcal{N}(0, 1)$ and $\mathcal{N}(\mu, \sigma)$. But Sup-GsCL with $\mathcal{N}(\mu, \sigma)$ obtained a higher AP score. When using the binary PPI features, Sup-GsCL with $\mathcal{N}(\mu, \sigma)$ outperformed Sup-EGsCL due to the higher MCC value and F1 score. Sup-GsCL with $\mathcal{N}(0, 1)$ also obtained a higher AP score than Sup-EGsCL. When using the combined features, Sup-EGsCL with a β value of 0.5 obtained a higher MCC value than Sup-GsCL. The former with a β value of 0.3 also obtained the same F1 score as the latter with $\mathcal{N}(\mu, \sigma)$. But Sup-GsCL with $\mathcal{N}(0, 1)$ obtained a higher AP score than Sup-EGsCL. In terms of Self-EGsCL, as shown in Table 3, according to MCC values, it outperformed Self-GsCL with a β value of 0.3 using the network embedding features. It also obtained higher AP scores with β values of 0.1 and 0.5, though Self-GsCL with $\mathcal{N}(0, 1)$ obtained a higher F1 score. When using the binary PPI features, Self-EGsCL outperformed Self-GsCL with different β values, according to the higher MCC values, F1 and AP scores. When using the combined features, Self-EGsCL with all different β values obtained higher MCC values, but Self-GsCL obtained higher F1 and AP scores with $\mathcal{N}(\mu, \sigma)$ and $\mathcal{N}(0, 1)$, respectively.

When predicting the longevity effects of yeast's genes, Sup-GsCL with $\mathcal{N}(0, 1)$ obtained higher MCC values and F1 scores than Sup-EGsCL using both the network embedding and binary PPI features. Sup-EGsCL with β values of 0.1 and 0.3 obtained higher AP scores than Sup-GsCL with both $\mathcal{N}(0, 1)$ and $\mathcal{N}(\mu, \sigma)$. It also obtained higher AP scores than Sup-GsCL when using the combined features with all different β values. However, Sup-GsCL with $\mathcal{N}(\mu, \sigma)$ performed better due to the higher MCC value and F1 score. Analogously, when using the network embedding features, Self-GsCL with $\mathcal{N}(0, 1)$ outperformed Self-EGsCL, according to the higher MCC value and F1 score. However, Self-EGsCL with a β value of 0.5 obtained a higher AP score. When using the binary PPI features, Self-GsCL with $\mathcal{N}(\mu, \sigma)$ performed better than Self-EGsCL due to the higher MCC value and F1 score. Self-EGsCL with a β value of 0.1 obtained the same AP score as Self-GsCL with $\mathcal{N}(0, 1)$. When using the combined features, Self-EGsCL with a β value of 0.2 obtained a higher MCC value than Self-GsCL with both $\mathcal{N}(0, 1)$ and $\mathcal{N}(\mu, \sigma)$. Self-EGsCL with all different β values also obtained the same F1 scores as Self-GsCL with $\mathcal{N}(\mu, \sigma)$. However, the latter obtained a higher AP score than the former.

EGsCL successfully obtained state-of-the-art accuracy in predicting the pro-longevity or anti-longevity effect of three model organisms' genes using PPI network-based features.

We further compared EGsCL with the benchmark method that uses raw PPI network features to train SVM classifiers. When predicting mouse genes' longevity effects using the network embedding features, both Sup-EGsCL and Self-EGsCL with all different β values obtained higher MCC values, F1 and AP scores than the benchmark method. Analogously, when working with the binary PPI features, both Sup-EGsCL and Self-EGsCL with almost all different β values obtained higher AP scores, though the benchmark obtained a higher MCC value. In addition, Self-EGsCL with a β value of 0.1 obtained a higher F1 score. When working with the combined features, Sup-EGsCL with a β value of 0.5 obtained a higher MCC value. It also obtained higher AP scores with all different β values, though the benchmark method obtained a higher F1 score. In terms of Self-EGsCL, it failed to obtain any higher MCC value and F1 score, but it obtained a higher AP score with a β value of 0.2.

When predicting worm genes' longevity effects using the network embedding features, Sup-EGsCL with all different β values obtained higher F1 and AP scores than the benchmark method, though the latter obtained a higher MCC value. When working with the binary PPI features, Sup-EGsCL with β values of 0.2 and 0.3 obtained higher MCC values. It also obtained higher F1 scores with all different β values, though the benchmark method obtained a higher AP score. When using the combined features, Sup-EGsCL with a β value of 0.5 obtained a higher MCC value. It also obtained higher F1 and AP scores with almost all different β values than the benchmark method. In terms of Self-EGsCL, it failed to obtain any higher MCC value, F1 and AP scores than the benchmark method using both the network embedding features and the binary PPI features. However, when working with the combined features, it obtained a higher F1 score with a β value of 0.3. It also obtained higher AP scores than the benchmark method with all different β values.

When predicting fly genes' longevity effects, both Sup-EGsCL and Self-EGsCL with almost all different β values obtained higher MCC values, F1 and AP scores than the benchmark method using the network embedding features. However, when using the binary PPI features, the latter obtained higher MCC value, F1 and AP scores. When working with the combined features, Sup-EGsCL with all different β values obtained higher MCC values and F1 scores, though the benchmark method obtained a higher AP score. In terms of Self-EGsCL, it obtained higher MCC values with β values of 0.3 and 0.4. It also obtained higher F1 scores with almost all β values, though the benchmark method obtained a higher AP score.

When predicting yeast genes' longevity effects using the network embedding features, the benchmark method outperformed both Sup-EGsCL and Self-EGsCL due to its higher MCC value, F1 and AP scores. However, when using the binary PPI features, Sup-EGsCL with almost all different β values obtained higher MCC values, F1 and AP scores, but Self-EGsCL failed to obtained higher F1 and AP scores than the benchmark method. When working with the combined features, Sup-EGsCL outperforms the benchmark method with all different β values due to the higher MCC values and F1 scores, though the latter obtained a higher AP score.

Table 4. New predictions about the pro-/anti-longevity effect of mouse genes and their homologous genes from human, fly and worm.

Mouse Gene ID	Mouse Gene Name	Predicted Class	Predicted Probability	Homologous genes from Human (HS), Fly (DM) and Worm (CE)
Pofut1	protein O-fucosyltransferase 1	Pro-longevity	87.8%	POFUT1 (HS), O-fut1 (DM), pfut-1 (CE)
Ints15	integrator complex subunit 15	Pro-longevity	87.7%	INTS15 (HS), CG5274 (DM),)Y56A3A.31 (CE)
Plod2	procollagen lysine, 2-oxoglutarate 5-dioxygenase 2	Pro-longevity	87.7%	PLOD2 (HS), Plod (DM), let-268 (CE)
Arid3a	AT-rich interaction domain 3A	Pro-longevity	87.6%	ARID3A (HS), retn (DM), cfi-1 (CE)
Col3a1	collagen, type III, alpha 1	Pro-longevity	87.3%	COL3A1 (HS)
Grk5	G protein-coupled receptor kinase 5	Anti-longevity	71.3%	GRK5 (HS), Gprk2 (DM), grk-1 (CE)
C2cd4b	C2 calcium-dependent domain containing 4B	Anti-longevity	70.5%	C2CD4B (HS)
Sstr3	somatostatin receptor 3	Anti-longevity	69.6%	SSTR3 (HS), AstC-R1 (DM), npr-24 & npr-16 (CE)
Rab44	RAB44, member RAS oncogene family	Anti-longevity	69.5%	RAB44 (HS), rsef-1 (CE)
Ntsr1	neurotensin receptor 1	Anti-longevity	69.5%	NTSR1 (HS)
‡ Apln	apelin	Anti-longevity	70.2%	APLN (HS)

Analogously, Self-EGsCL also obtained higher MCC values and F1 scores than the benchmark method with all different β values.

Sup-EGsCL is also the overall best method for predicting mouse, worm and fly genes' longevity effects. As denoted by the bold texts in Table 3, in terms of the mouse datasets, Sup-EGsCL with a β value of 0.5 obtained the overall highest MCC value (i.e. 0.427), whilst it also obtained the overall highest AP score (i.e. 0.860) with a β value of 0.1. The overall highest F1 score (i.e. 0.826) was obtained by the benchmark method. Analogously, in terms of the worm datasets, Sup-EGsCL also obtained the overall highest MCC value (i.e. 0.387), F1 score (i.e. 0.599) and AP score (i.e. 0.698) with different β values. The overall highest MCC value (i.e. 0.328) and F1 score (i.e. 0.782) for the fly datasets were obtained by Sup-EGsCL with a β value of 0.3. Sup-GsCL with $\mathcal{N}(\mu, \sigma)$ also obtained the same overall highest F1 score, whilst Sup-GsCL with $\mathcal{N}(0, 1)$ obtained the overall highest AP score (i.e. 0.838). In terms of the yeast datasets, the overall highest MCC value (i.e. 0.274), F1 score (i.e. 0.297) and AP score (i.e. 0.509) were all obtained by the benchmark method.

Sup-EGsCL successfully predicted novel mouse genes with the pro-/anti-longevity effect.

We then used one of the trained Sup-EGsCL-based classifiers during the 10-fold cross-validation to predict the pro-/anti-longevity effect of all the mouse genes included in the STRING database. The pro-longevity genes are defined as those genes whose decreased expression reduces lifespan and/or their overexpression extends lifespan. *Vice versa*, the anti-longevity genes are defined as those genes whose overexpression reduces lifespan and/or their decreased expression extends lifespan [9].

We focus on predicting novel mouse genes for several reasons, as follows. First, the predictive models for mouse data are the most accurate models in general, across the models for the 4 organisms. Second, mice are much closer to humans than the other 3 model organisms investigated (with results for mice being more useful as evidenced from pre-clinical studies). Third, experiments with mice are much slower and more time consuming than experiments with the other 3 types of organisms investigated, so it is particularly important to use machine learning methods to prioritise mouse genes for further testing via wet-lab experiments.

Table 4 shows the top-ranked mouse genes that were most likely to bear pro-/anti-longevity labels according to their probabilities predicted by the trained Sup-EGsCL-based classifier. Those genes are considered potentially novel pro-/anti-longevity genes because they are not included in the GenAge database (and so, they are not in the datasets used to learn our Sup-EGsCL-based classifiers). The table also includes information about homologous genes from human, fly and worm according to the Alliance of Genome Resources database [34] with the stringent homolog information deposit criterion. The complete list of mouse genes that are included in both STRING [52] and NCBI [7] databases with their predicted probabilities of bearing the pro-/anti-longevity effect is included in Supplementary File 2. Other genes might also be considered potentially exhibiting a pro-/anti-longevity effect if their predicted probabilities are no less than a certain threshold, which can be specified by each researcher based on their research requirements.

For example, in order to identify the small sets of top-ranked genes reported in Table 4, we consider that a mouse gene is likely to have a pro-longevity effect if its corresponding predicted probability is no less than 85%; whilst a mouse gene is likely to have an anti-longevity effect if its corresponding probability is no less than 67%. We consider a somewhat smaller probability threshold for identifying potentially novel anti-longevity genes due to the fact that, overall, the degree of confidence (predicted probabilities) for the predicted anti-longevity genes is substantially smaller than the degree of confidence for the predicted pro-longevity genes.

Regarding the predicted pro-longevity genes in Table 4, there is support in the literature for their pro-longevity role, as follows. As the top-ranked pro-longevity gene, Pofut1 and its homologous genes from human, fly and worm play important roles in the well-known ageing-related notch pathway [3]. It has been found in mice that this gene's deletion is linked to multiple muscle ageing-related phenotypes [67] and promotes colorectal cancer cell apoptosis [11]. Ints15 is another top-ranked mouse gene predicted to have a pro-longevity effect. It is known to be related to RNA polymerase II - another well-known ageing-related factor in multiple species [10]. Recent research on mice's Ints15 gene [2] also confirmed its crucial role in cell survival – the knockout of Ints15 induces cell apoptosis. Analogously, Plod2 and its corresponding homologous human genes play an

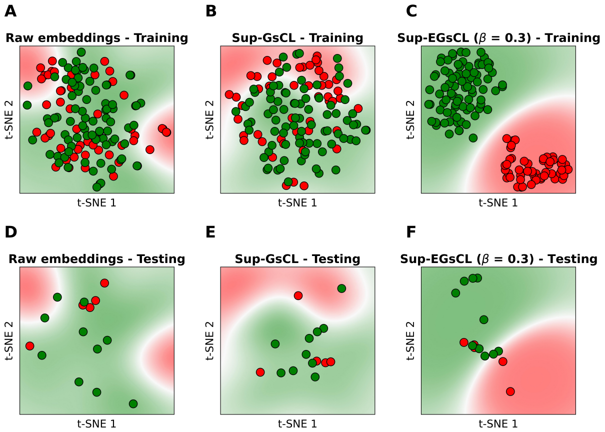

Fig. 2. 2D *t*-SNE visualisations of the training and testing datasets for fly genes using the network embedding features (A, D) and the feature representations learned by Sup-GsCL (B, E) and Sup-EGsCL – β=0.3 (C, F), respectively.

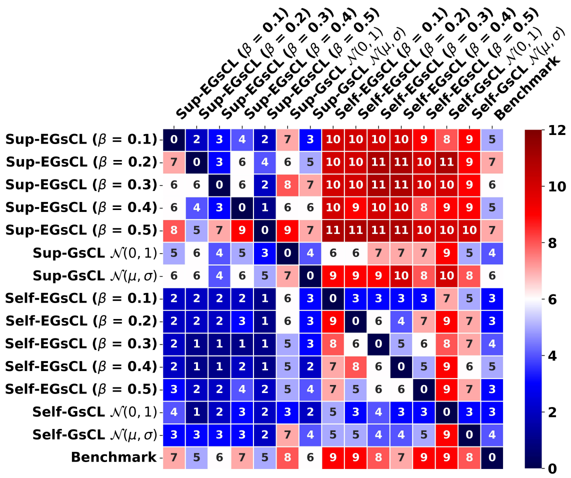

Fig. 3. A heatmap showing the numbers of datasets where the methods on the rows obtained higher MCC values than the methods on the columns.

important role in responses to hypoxia [49], which could extend the lifespan of mice [48]. Arid3a and its homologous genes from human, fly and worm are another group of genes that are linked to RNA polymerase II-related transcription regulations. It has been revealed that the loss of Arid3a gene leads to defects in hematopoiesis [44] – a common pattern observed in aged individuals [43]. Col3a1 and its human homolog are linked with type III collagen, which plays a crucial role in normal collagen I fibrillogenesis in the cardiovascular system, and the deletion of Col3a1 shortens the lifespan of mouse [29].

Among those predicted mouse genes that have an anti-longevity effect as shown in Table 4, Grk5 regulates responses to inflammatory factors [54] – a key factor leading to senescence [26]. Recent research in human and mouse has revealed that silencing the Grk5 gene could suppress inflammatory factors [54]. C2cd4b is linked with reactive oxygen species, which is a well-known ageing-related factor [56]. The overexpression of C2cd4b leads to an increased risk of type 2 diabetes [64, 25], but inhibition of C2CD4B expression prevents hyperglycemia-induced oxidative stress [41]. Sstr3 and its homologs are linked with the G protein-coupled receptor (GPCR) signalling pathway. It has been found that GPCRs play important roles in T-cell-related ageing processes [32], and the blockade of SSTR3 in human cells can reduce T-cell responses [63]. Rab44 is also closely associated with immunosenescence. The knockout of Rab44 in mice diminishes anaphylaxis [21], which is a process involving a large number of mast cells releasing a wide range of inflammatory mediators [39]. Ntsr1 has also been found to regulate apoptotic processes – the inhibition of NTSR1 in human breast cancer cell lines leads to reduced ERK 1/2 phosphorylation [20], which induces apoptotic processes [24]. However, among the top-ranked genes that are predicted to have an anti-longevity effect, Apln was actually found to be associated with the pro-longevity effect, since accelerated senescence was observed in Apln knockout mice [66]. This shows that of course even highly accurate models like our Sup-EGsCL-based classifiers can occasionally make wrong predictions; and so experiments measuring mouse lifespan need to be done, in future work, to determine whether the novel pro-/anti-longevity genes predicted in this work really have their predicted effect.

Discussion

Sup-EGsCL successfully learns discriminative feature representations based on network embedding features leading to better decision boundaries.

We compared the raw network embedding features and two types of feature representations learned by Sup-EGsCL and Sup-GsCL, respectively. Figure 2 shows the 2D *t*-SNE visualisation of the training and testing datasets for fly genes including the learned SVM decision boundaries. As shown in Figures 2.A and 2.D, when using the raw network embedding features, both the training and testing instances bearing different class labels are distributed in overlapping areas. The learned decision boundary also failed to distinguish the red and green dots denoting two different class labels. As shown in Figures 2.B and 2.E, Sup-GsRL failed to learn discriminative feature representations since the instances bearing different class labels were still distributed in the overlapping areas. Analogously, the learning SVM decision boundaries also failed to separate the majority of the red and green dots. In contrast, Sup-EGsCL with a β value 0f 0.3 shows better sample distributions. As shown in Figures 2.C and 2.F, both the training and testing instances are grouped into two separate areas, whilst the learned SVM decision boundaries successfully distinguished more red and green dots.

Augmentation with noises sampled from two different Gaussian distributions leads to higher predictive accuracy.

We further discussed the differences in augmentation approaches between EGsCL and GsCL. The former samples noises from two different Gaussian distributions, i.e. $\mathcal{N}(\mu+\beta,\sigma)$ and $\mathcal{N}(\mu-\beta,\sigma)$, whilst the latter samples two noises from one single Gaussian distribution, e.g. $\mathcal{N}(\mu,\sigma)$. In general, noises sampled from two different Gaussian distributions lead to higher predictive accuracy, compared with using noises sampled from one single Gaussian distribution. Figure 3 shows a heatmap for the pairwise comparisons between different methods according to their MCC values obtained by 12 datasets, i.e. 4 model organisms' ageing-related genes described by 3 different feature types. Sup-EGsCL with both β values of

0.3 and 0.5 obtained higher MCC values in more datasets (i.e. 7 out of 12) than Sup-GsCL with $\mathcal{N}(\mu, \sigma)$, whilst Self-EGsCL with almost all different β values except 0.1 also obtained higher MCC values than Self-GsCL with $\mathcal{N}(\mu, \sigma)$ in more datasets. Sup-EGsCL with a β value of 0.4 obtained higher MCC values in the same number of datasets as Sup-GsCL with $\mathcal{N}(\mu, \sigma)$, which obtained higher MCC values in more datasets than Sup-EGsCL with β values of 0.1 and 0.2.

Supervised contrastive learning paradigm leads to higher predictive accuracy than self-supervised contrastive learning paradigm.

In terms of the differences between supervised and self-supervised paradigms, the former leads to higher predictive accuracy for both EGsCL and GsCL methods. As shown in the top right area of Figure 3, Sup-EGsCL with all different β values obtained higher MCC values than Self-EGsCL with all different β values in the vast majority of the datasets. Analogously, Sup-GsCL with $\mathcal{N}(\mu, \sigma)$ obtained higher MCC values than Self-GsCL with $\mathcal{N}(\mu, \sigma)$ in 8 out of 12 datasets, whilst Sup-GsCL with $\mathcal{N}(0, 1)$ also outperformed Self-GsCL with $\mathcal{N}(0, 1)$ in 9 out of 12 datasets.

In terms of the differences between two Gaussian distribution settings, i.e. $\mathcal{N}(\mu, \sigma)$ and $\mathcal{N}(0, 1)$, the former outperformed the latter using either supervised or self-supervised settings. As shown in Figure 3, Sup-GsCL with $\mathcal{N}(\mu, \sigma)$ obtained higher MCC values than Sup-GsCL with $\mathcal{N}(0, 1)$ in 7 out of 12 datasets, whilst Self-GsCL with $\mathcal{N}(\mu, \sigma)$ also outperformed Self-GsCL with $\mathcal{N}(0, 1)$ in 9 out of 12 datasets.

Conclusion

In summary, we proposed two new contrastive learning methods, i.e. Sup-EGsCL and Self-EGsCL, which successfully learn a type of discriminative representations based on protein-protein interaction network data, leading to state-of-the-art accuracy in predicting pro-longevity or anti-longevity effect of model organisms' genes. In addition, we have used Sup-EGsCL to predict 10 novel pro-/anti-longevity mouse genes, and have discussed the support for these predictions in the literature. An interesting future research direction would be to propose new contrastive learning methods for other features like Gene Ontology terms or their corresponding hierarchy embeddings.

Acknowledgement

The authors acknowledge the support of the School of Computing and Mathematical Sciences and the Birkbeck GTA programme.

Data availability

The datasets used in this work and the pretrained encoders can be downloaded from `https://doi.org/10.5281/zenodo.12143797`

Code availability

Source code is available at `https://doi.org/10.6084/m9.figshare.26227532` and at `https://github.com/ibrahimsaggaf/EGsCL`.

References

1. Ibrahim Alsaggaf, Daniel Buchan, and Cen Wan. Improving cell type identification with Gaussian noise-augmented single-cell RNA-seq contrastive learning. *Briefings in Functional Genomics*, page elad059, 2024.
2. Noriyuki Azuma, Tadashi Yokoi, Taku Tanaka, Emiko Matsuzaka, Yuki Saida, Sachiko Nishina, Miho Terao, Shuji Takada, Maki Fukami, Kohji Okamura, Kayoko Maehara, Tokiwa Yamasaki, Jun Hirayama, Hiroshi Nishina, Hiroshi Handa, and Yuki Yamaguchi. Integrator complex subunit 15 controls mrna splicing and is critical for eye development. *Human Molecular Genetics*, 32:2032–2045, 2023.
3. Carmela Rita Balistreri, Rosalinda Madonna, Gerry Melino, and Calogero Caruso. The emerging role of notch pathway in ageing: Focus on the related mechanisms in age-related diseases. *Ageing Research Reviews*, 29:50–65, 2016.
4. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
5. Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255, 2020.
6. Madalina Ciortan and Matthieu Defrance. Contrastive self-supervised clustering of scRNA-seq data. *BMC bioinformatics*, 22(1):1–27, 2021.
7. NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 44:D7–D19, 2016.
8. João Pedro de Magalhães. Distinguishing between driver and passenger mechanisms of aging. *Nature Genetics*, 56:204–211, 2024.
9. João Pedro de Magalhães, Zoya Abidi, Gabriel Arantes Dos Santos, Roberto A. Avelar, Diogo Barardo, Kasit Chatsirisupachai, Peter Clark, Evandro A. De-Souza, Emily J. Johnson, Inês Lopes, Guy Novoa, Ludovic Senez, Angelo Talay, Daniel Thornton, and Paul Ka Po To. Human ageing genomic resources: updates on key databases in ageing research. *Nucleic Acids Research*, 52:D900–D908, 2024.
10. Cédric Debès, Antonios Papadakis, Sebastian Grönke, Özlem Karalay, Luke S. Tain, Athanasia Mizi, Shuhei Nakamura, Oliver Hahn, Carina Weigelt, Natasa Josipovic, Anne Zirkel, Isabell Brusius, Konstantinos Sofiadis, Mantha Lamprousi, Yu-Xuan Lu, Wenming Huang, Reza Esmaillie, Torsten Kubacki, Martin R. Späth, Bernhard Schermer, Thomas Benzing, Roman-Ulrich Müller, Adam Antebi, Linda Partridge, Argyris Papantonis, and Andreas Beyer. Ageing-associated changes in transcriptional elongation influence longevity. *Nature*, 616:814–821, 2023.
11. Yuheng Du, Daojiang Li, Nanpeng Li, Chen Su, Chunxing Yang, Changwei Lin, Miao Chen, Runliu Wu, Xiaorong Li, and Gui Hu. Pofut1 promotes colorectal cancer development through the activation of notch1 signaling. *Cell Death and Disease*, 9:1–12, 2018.
12. Handan Melike Dönertaş, Daniel K. Fabian, Matías Fuentealba, Linda Partridge, and Janet M. Thornton. Common genetic associations between age-related diseases. *Nature Aging*, 1(4):400–412, 2021.

13. Fabio Fabris, João Pedro de Magalhães, and Alex A. Freitas. A review of supervised machine learning applied to ageing research. *Biogerontology*, 18:171–188, 2017.
14. Yaping Fang, Xinkun Wang, Elias K. Michaelis, and Jianwen Fang. Classifying aging genes into dna repair or non-dna repair-related categories. *Intelligent computing theories and technology, lecture notes in computer science*, page 20–29, 2013.
15. Kai Feng, Xin Song, Fei Tan, Yan-Hui Li, Yuan-Chun Zhou, and Jian hui Li. Topological anaylysis and prediction of aging genes in mus musculus. In *2012 International Conference on Systems and Informatics (ICSAI)*, page 2268–2271, 2012.
16. Alex A Freitas, Olga Vasieva, and João Pedro de Magalhães. A data mining approach for classifying dna repair genes into ageingrelated or non-ageing-related. *BMC Genomics*, 12:27, 2011.
17. David Gems and João Pedro de Magalhães. The hoverfly and the wasp: a critique of the hallmarks of aging as a paradigm. *Aging Research Reviews*, 70(101407), 2023.
18. Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.
19. Emre Guney and Baldo Oliva. Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PLOS One*, 7:e43557, 2012.
20. Yasser Heakal, Matthew P Woll, Todd Fox, Kelly Seaton, Robert Levenson, and Mark Kester. Neurotensin receptor-1 inducible palmitoylation is required for efficient receptor-mediated mitogenic-signaling within structured membrane microdomains. *Cancer Biology and Therapy*, 12:427–435, 2011.
21. Tomoko Kadowaki, Yu Yamaguchi, Mizuho A Kido, Takaya Abe, Kohei Ogawa, Mitsuko Tokuhisa, Weiqi Gao, Kuniaki Okamoto, Hiroshi Kiyonari, and Takayuki Tsukuba. The large gtpase rab44 regulates granule exocytosis in mast cells and ige-mediated anaphylaxis. *Cellular and Molecular Immunology*, 17:1287–1289, 2020.
22. Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021.
23. Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673, 2020.
24. Tingting Kong, Minghui Liu, Bingyuan Ji, Bo Bai, Baohua Cheng, and Chunmei Wang. Role of the extracellular signal-regulated kinase 1/2 signaling pathway in ischemia-reperfusion injury. *Frontiers in Physiology*, 10:1038, 2019.
25. Ina Kycia, Brooke N Wolford, Jeroen R Huyghe, Christian Fuchsberger, Swarooparani Vadlamudi, Romy Kursawe, Ryan P. Welch, Ricardo d'Oliveira Albanus, Asli Uyar, Shubham Khetan, Nathan Lawlor, Mohan Bolisetty, Anubhuti Mathur, Johanna Kuusisto, Markku Laakso, Duygu Ucar, Karen L. Mohlke, Michael Boehnke, Francis S. Collins, Stephen C. J. Parker, and Michael L. Stitzel. A common type 2 diabetes risk variant potentiates activity of an evolutionarily conserved islet stretch enhancer and increases c2cd4a and c2cd4b expression. *The American Journal of Human Genetics*, 102:620–635, 2018.
26. Xia Li, Chentao Li, Wanying Zhang, Yanan Wang, Pengxu Qian, and He Huang. Inflammation and aging: signaling pathways and intervention therapies. *Signal Transduction And Targeted Therapy*, 8:1–29, 2023.
27. Yan-Hui Li, Gai-Gai Zhang, and Zheng Guo. Computational prediction of aging genes in human. In *2010 International Conference on Biomedical Engineering and Computer Science*, page 1–4, 2010.
28. Zhe Li, Zhenkun Zhang, Yikun Ren, Yingying Wang, Jiarui Fang, Han Yue, Shanshan Ma, and Fangxia Guan. Aging and age-related diseases: from mechanisms to therapeutic strategies. *Biogerontology*, 22:165–187, 2021.
29. Xin Liu, Hong Wu, Michael Byrne, Stephen Krane, and Rudolf Jaenisch. Type iii collagen is crucial for collagen i fibrillogenesis and for normal cardiovasculardevelopment. *Proceedings of the National Academy of Sciences of the United States of America*, 94:1852–1856, 1997.
30. Gustavo Daniel Vega Magdaleno, Vladislav Bespalov, Yalin Zheng, Alex A. Freitas, and Joao Pedro De Magalhaes. Machine learning-based predictions of dietary restriction associations across ageing-related genes. *BMC bioinformatics*, 23:1–28, 2022.
31. Gustavo Daniel Vega Magdaleno and João Pedro de Magalhães. Pleiotropy and disease interactors: the dual nature of genes linking ageing and ageing-related diseases. *bioRxiv*, 2021.
32. Maria Mittelbrunn and Guido Kroemer. Hallmarks of t cell aging. *Nature Immunology*, 22:687–698, 2021.
33. Saket Navlakha and Carl Kingsford. The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26:1057–1063, 2010.
34. Alliance of Genome Resources Consortium. Updates to the alliance of genome resources central infrastructure. *Genetics*, 227:iyae049, 2024.
35. Csaba Ortutay and Mauno Vihinen. Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Research*, 37:622–628, 2009.
36. Andrey A. Parkhitko, Elizabeth Filine, Stephanie E. Mohr, Alexey Moskalev, and Norbert Perrimon. Targeting metabolic pathways for extension of lifespan and healthspan across multiple species. *Aging Research Reviews*, 64(101188), 2020.
37. Andrey A. Parkhitko, Elizabeth Filine, and Marc Tatar. Combinatorial interventions in aging. *Nature Aging*, 3:1187–1200, 2023.
38. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035. Curran Associates, Inc., 2019.
39. Lucia Pedicini, Jessica Smith, Sinisa Savic, and Lynn McKeown. Rab46: a novel player in mast cell function. *Discovery Immunology*, 3:kyad028, 2024.
40. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

41. Paola Di Pietro, Angela Carmelita Abate, Valeria Prete, Antonio Damato, Eleonora Venturini, Maria Rosaria Rusciano, Carmine Izzo, Valeria Visco, Michele Ciccarelli, Carmine Vecchione, and Albino Carrizzo. C2cd4b evokes oxidative stress and vascular dysfunction via a pi3k/akt/pkc–signaling pathway. *Antioxidants*, 13:101, 2024.
42. Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Jiawei Han, and Weizhu Chen. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. *arXiv:2010.08670*, 2020.
43. Michelle L. Ratliff, Joshua Garton, Judith A. James, and Carol F. Webb. Arid3a expression in human hematopoietic stem cells is associated with distinct gene patterns in aged individuals. *Immunity and Ageing*, 17:1–15, 2020.
44. Michelle L. Ratliff, Troy D. Templeton, Julie M. Ward, and Carol F. Webb. The bright side of hematopoiesis: regulatory roles of arid3a/bright in human and mouse hematopoiesis. *Frontiers in Immunology*, 5:1–8, 2014.
45. Suresh Rattan. Seven knowledge gaps in modern biogerontology. *Biogerontology*, 25:1–8, 2024.
46. Caio Ribeiro, Christopher K. Farmer, João Pedro de Magalhães, and Alex A. Freitas. Predicting lifespan-extending chemical compounds for c. elegans with machine learning and biologically interpretable features. *Ageing*, 15:6073–6099, 2023.
47. Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv:2010.04592*, 2020.
48. Robert S. Rogers, Hong Wang, Timothy J. Durham, Jonathan A. Stefely, NorahA.Owiti, Andrew L. Markhard, Lev Sandler, Tsz-Leung To, and Vamsi K. Mootha. Hypoxia extends lifespan and neurological function in a mouse model of aging. *PLoS Biology*, 21:e3002117, 2023.
49. Tamara Rosell-García, Oscar Palomo-Álvarez, and Fernando Rodríguez-Pascual. A hierarchical network of hypoxia-inducible factor and smad proteins governs procollagen lysyl hydroxylase 2 induction by hypoxia and transforming growth factor 1hif and smad signaling pathways induce plod2 expression. *Journal of Biological Chemistry*, 294:14308–14318, 2019.
50. Tomas Schmauck-Medina, Adrian Molière, Sofie Lautrup, Jianying Zhang, Stefan Chlopicki, Helena Borland Madsen, Shuqin Cao, Casper Soendenbroe, Els Mansell, Mark Bitsch Vestergaard, Zhiquan Li, Yosef Shiloh, Patricia L Opresko, Jean-Marc Egly, Thomas Kirkwood, Eric Verdin, Vilhelm A Bohr, Lynne S. Cox, Tinna Stevnsner, Lene Juel Rasmussen, and Evandro F. Fang. New hallmarks of ageing: a 2022 copenhagen ageing meeting summary. *Aging*, 14(16):6829–6839, 2022.
51. Xin Song, Yuan-Chun Zhou, Kai Feng, Yan-Hui Li, and Jian-Hui Li. Discovering aging-genes by topological features in drosophila melanogaster protein-protein interaction network. In *2012 IEEE 12th International Conference on Data Mining Workshops*, page 94–98, 2012.
52. Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023.
53. Robi Tacutu, Daniel Thornton, Emily Johnson, Arie Budovsky, Diogo Barardo, Thomas Craig, Eugene Diana, Gilad Lehmann, Dmitri Toren, Jingwei Wang, Vadim E. Fraifeld, and João Pedro de Magalhães. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Research*, 46(D1):D1083–D1090, 2018.
54. Masakazu Toya, Yukio Akasaki, Takuya Sueishi, Ichiro Kurakazu, Masanari Kuwahara, Taisuke Uchida, Tomoaki Tsutsui, Hidetoshi Tsushima, Hisakata Yamada, Martin K. Lotz, and Yasuharu Nakashima. G protein-coupled receptor kinase 5 deletion suppresses synovial inflammation in a murine model of collagen antibody-induced arthritis. *Scientific Reports*, 11:1–11, 2021.
55. Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, 21:697–700, 2003.
56. Caroline Maria Oliveira Volpe, Pedro Henrique Villar-Delfino, Paula Martins Ferreira dos Anjos, and José Augusto Nogueira-Machado. Cellular death, reactive oxygen species (ros) and diabetic complications. *Cell Death and Disease*, 9:1–9, 2018.
57. Cen Wan. *Hierarchical Feature Selection for Knowledge Discovery: Application of Data Mining to the Biology of Ageing*. Springer Cham, Switzerland, 2019.
58. Cen Wan, Domenico Cozzetto, Rui Fa, and David T. Jones. Using deep maxout neural networks to improve the accuracy of function prediction from protein interaction networks. *PLOS One*, 14:e0209958, 2019.
59. Cen Wan, Alex A. Freitas, and João Pedro de Magalhães. Predicting the pro-longevity or anti-longevity effect of model organism genes with new hierarchical feature selection methods. *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, 12(2):262–275, 2015.
60. Hui Wan, Liang Chen, and Minghua Deng. scNAME: Neighborhood contrastive clustering with ancillary mask estimation for scRNA-seq data. *Bioinformatics*, 38(6):1575–1583, 01 2022.
61. Wei Xiong, Hui Liu, Jihong Guan, and Shuigeng Zhou. Protein function prediction by collective classification with explicit and implicit edges in protein-protein interaction networks. *BMC Bioinformatics*, 14:1–13, 2013.
62. Yang Xu, Priyojit Das, and Rachel Patton McCord. SMILE: Mutual information learning for integration of single-cell omics data. *Bioinformatics*, 38(2):476–486, 2022.
63. Bo Zhang, Huiru Feng, Hui Lin, and Rui Li. Somatostatin-sstr3-gsk3 modulates human t-cell responses by inhibiting oxphos. *Frontiers in Immunology*, 15:1322670, 2024.
64. Martin Jinye Zhang, Angela Oliveira Pisco, Spyros Darmanis, and James Zou. Mouse aging cell atlas analysis reveals global and cell type-specific aging signatures. *Elife*, 13:e62293, 2021.
65. Alex Zhavoronkov, Polina Mamoshina, Quentin Vanhaelen, Morten Scheibye-Knudsen, Alexey Moskalev, and Alex Aliper. Artificial intelligence for aging and longevity research: recent advances and perspectives. *Aging Research Reviews*, 49:49–66, 2019.
66. Qionglin Zhou, Linxi Chen, Mingzhu Tang, Yu Guo, and Lanfang Li. Apelin/apj system: A novel promising target for anti-aging intervention. *Clinica Chimica Acta*, 487:233–240, 2018.

67. Deborah A. Zygmunt, Neha Singhal, Mi-Lyang Kim, Megan L. Cramer, Kelly E. Crowe, Rui Xu, Ying Jia, Jessica Adair, Isabel Martinez-Pena y Valenzuel, Mohammed Akaaboune, Peter White, Paulus M. Janssen, and Paul T. Martina. Deletion of pofut1 in mouse skeletal myofibers induces muscle aging-related phenotypes in cis and in trans. *Molecular and Cellular Biology*, 37:e00426–16, 2017.