BMC Bioinformatics

RESEARCH Open Access

Machine learning-based predictions of dietary restriction associations across ageing-related genes

Gustavo Daniel Vega Magdaleno[1], Vladislav Bespalov[2], Yalin Zheng[3], Alex A. Freitas[4] and Joao Pedro de Magalhaes[1*]

*Correspondence: jp@senescence.info
[1] Integrative Genomics of Ageing Group, Institute of Life Course and Medical Sciences, University of Liverpool, 6 West Derby St, Liverpool L7 8TX, UK
Full list of author information is available at the end of the article

Abstract

Background: Dietary restriction (DR) is the most studied pro-longevity intervention; however, a complete understanding of its underlying mechanisms remains elusive, and new research directions may emerge from the identification of novel DR-related genes and DR-related genetic features.

Results: This work used a Machine Learning (ML) approach to classify ageing-related genes as DR-related or NotDR-related using 9 different types of predictive features: PathDIP pathways, two types of features based on KEGG pathways, two types of Protein–Protein Interactions (PPI) features, Gene Ontology (GO) terms, Genotype Tissue Expression (GTEx) expression features, GeneFriends co-expression features and protein sequence descriptors. Our findings suggested that features biased towards curated knowledge (i.e. GO terms and biological pathways), had the greatest predictive power, while unbiased features (mainly gene expression and co-expression data) have the least predictive power. Moreover, a combination of all the feature types diminished the predictive power compared to predictions based on curated knowledge. Feature importance analysis on the two most predictive classifiers mostly corroborated existing knowledge and supported recent findings linking DR to the Nuclear Factor Erythroid 2-Related Factor 2 (NRF2) signalling pathway and G protein-coupled receptors (GPCR).

We then used the two strongest combinations of feature type and ML algorithm to predict DR-relatedness among ageing-related genes currently lacking DR-related annotations in the data, resulting in a set of promising candidate DR-related genes (*GOT2*, *GOT1*, *TSC1*, *CTH*, *GCLM*, *IRS2* and *SESN2*) whose predicted DR-relatedness remain to be validated in future wet-lab experiments.

Conclusions: This work demonstrated the strong potential of ML-based techniques to identify DR-associated features as our findings are consistent with literature and recent discoveries. Although the inference of new DR-related mechanistic findings based solely on GO terms and biological pathways was limited due to their knowledge-driven nature, the predictive power of these two features types remained useful as it allowed inferring new promising candidate DR-related genes.

Keywords: Dietary restriction, Ageing, Machine learning

Background

Ageing increases the risk of disease and death as it declines homeostasis and decreases the capacity to respond to environmental stimuli [1]. Given the widespread interest in reversing and ultimately preventing the detrimental effects of ageing, considerable effort has been devoted to understanding its underlying biochemical mechanisms [2]. It is known that ageing-related changes are multifactorial and involve a variety of processes, including genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, impaired nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intracellular communication [3].

Dietary Restriction (DR), which involves reducing total dietary energy intake while maintaining adequate vitamin and mineral levels, is currently the most promising intervention for increasing both lifespan and healthspan, as experiments with a variety of species have shown that DR not only induces longevity but also retards the ageing process [2]. Although the mechanism underlying these pro-longevity effects is unknown, evidence suggests that DR: (1) reduces oxidative damage by reducing the production of Reactive Oxygen Species (ROS); (2) decreases circulating insulin and glucose levels, resulting in decreased cell growth and division and a shift toward maintenance and repair; and (3) decreases growth hormone and insulin-like growth factor levels [4].

Additionally, a wealth of publicly available omics data on ageing has emerged thanks to new high-throughput sequencing technologies [5]. Hence, a relatively recent approach for studying the ageing process is based on Machine Learning (ML) techniques that learn patterns about gene or protein functions by analyzing gene or protein features such as Gene Ontology (GO) terms, metabolic pathways, and protein-protein interactions, to name a few [6]. Examples include the association of human genes with ageing-related diseases [7]; prediction of gene deletion effects on yeast longevity [8]; and the determination of blood age [9]; among others [10].

This work aims to identify novel DR-related candidate genes from ageing-related genes while also identifying genetic features which increase the likelihood that certain ageing-related genes become DR-related. To accomplish this, we created 11 datasets based on 9 different types of predictive features and two approaches to combine all those features into an integrated dataset. The 9 used feature types were: PathDIP pathways, two types of features based on KEGG pathways, two types of Protein-Protein Interactions (PPI) features, Gene Ontology (GO) terms, Genotype-Tissue Expression (GTEx) expression features, GeneFriends co-expression features and protein sequence descriptors. These datasets provide a wealth of information for representing each of the ageing-related genes under study (i.e., genes retrieved from the GenAge "Model organisms" database, where genes are considered ageing-related if there is least one wet-lab study where manipulations of the gene result in noticeable changes in the ageing phenotype and/or longevity of the model organism, as further explained in Methods*: Ageing genes and DR labels retrieval*). Then, a ML approach was used to predict whether or not an ageing-related gene is DR-related (i.e., using information from the GenDR database where DR-related genes are selected as those that impair DR-related effects in at least one wet-lab study, as explained in the same *Methods* subsection). The approach involved comparing the predictive accuracy of four different tree-based ensemble ML algorithms across all 11 datasets, with the most predictive ML algorithm, Dataset combinations being then

used for feature importance analysis, which led to the identification of key DR-related genetic attributes. Finally, we used the two best performing classifiers to infer potential under-explored DR-relatedness from ageing-related genes lacking DR-related annotations in the dataset.

Our findings indicate that the most predictive features are based on curated knowledge, such as GO terms and biological pathways. The least predictive features were gene expression and co-expression features. Apart from the well-established DR-related autophagy and longevity regulating pathways, our findings indicate that the G Protein-Coupled Receptor (GPCR) signalling, cellular responses to external stimuli, and Nuclear Factor Erythroid 2-Related Factor 2 (NRF2) pathways were among the most significant features for inferring DR-relatedness. This is consistent with recent evidence indicating that these pathways act as mediators of DR effects. Additionally, the oxidation-reduction reaction was the most predictive feature across all GO terms. Finally, predictions from the strongest classifiers indicate that the ageing-related genes *GOT2*, *GOT1*, *TSC1*, *CTH*, *GCLM*, *IRS2*, and *SENS2* may share an under-explored association with DR.

Methods

Datasets construction

This and the following sections use the following ML terminology: an instance refers to any ageing-related gene/protein included in the datasets; whereas a feature is any observable property or attribute of any instance (*e.g.*, GO term annotations, association with biological pathways, protein sequence descriptors, etc).

Ageing genes and DR labels retrieval

GenAge [11] is a benchmark database of ageing-related genes. GenAge "model organisms" is a subsection of GenAge consisting of genes in model organisms that, if genetically modulated, result in significant changes in the ageing phenotype and/or longevity [12]. The majority of observations have been made on mice, nematodes, fruit flies and budding yeast. The criterion for including a gene in this subsection of GenAge consists on the existence of at least one wet-lab study where manipulations of the gene result in noticeable changes in the ageing phenotype and/or longevity of the model organism. In this work, ageing-related genes from those four organisms were downloaded from GenAge Build 20 (09/02/2020). Then, human orthologs of these genes were retrieved from the OMA Orthology database [13] (2020 Release) using the OMA browser's Genome Pair View, which allows for the download of orthologs between two species (https://omabrowser.org/oma/genomePW/). Since some genes in different organisms are mapped to overlapping human orthologs, only one gene from each set of repeated genes was retained, resulting in 1137 human ageing-related genes.

GenDR [14] is a database of DR-related genes. DR-essential genes are defined in GenDR as those which, if genetically modified, impair DR-mediated lifespan extension in at least one wet-lab study. This criterion applies even if it was shown for only a single DR regimen. In this work, 215 DR-associated genes from the aforementioned four model organisms were retrieved from the "Gene Manipulation" section of the GenDR database, build 4 (24/06/2017), and used as input for another OMA human orthologs query, which led to 152 human DR genes after keeping only one of each repeated ortholog genes

coming from different organisms. The ageing- and DR-related human genes retrieval processes are summarised in Fig. 1a.

Finally, the overlap of the retrieved ageing- and DR-related human orthologs resulted in 115 genes (Fig. 1b) which were labelled as $Ageing_{DR}$-related, while the remaining 1022 ageing-related genes that didn't overlap with DR-related orthologs were labelled as $Ageing_{NotDR}$-related.

PathDIP dataset

This dataset consists of binary features, i.e., each feature value indicates whether or not an ageing-related gene belongs to corresponding specific PathDIP pathway. To accomplish this, we queried the PathDIP [15] database (version 4.0.7.0) to download a dataset in which instances were ageing-related genes, and features were biological pathways coming from a variety of database sources, including ACSN2, Bio-Carta, EHMN, HumanCyc, INOH, IPAVS KEGG, NetPath, OntoCancro, Panther Pathway, PharmGKB, PID, RB-Pathways, REACTOME, stke, systems-biology.org, SignaLink2.0, SIGNOR2.0, SMPDB, Spike, UniProt Pathways, and WikiPathway. The resulting dataset contained 1640 pathways and 986 ageing-related genes: 110 labelled as $Ageing_{CR}$-related and 876 labelled as $Ageing_{NotDR}$-related.

KEGG-pertinence dataset

This dataset consists of binary features, where each feature indicates whether or not an ageing-related gene (instance) belongs to a specific KEGG pathway. To construct this

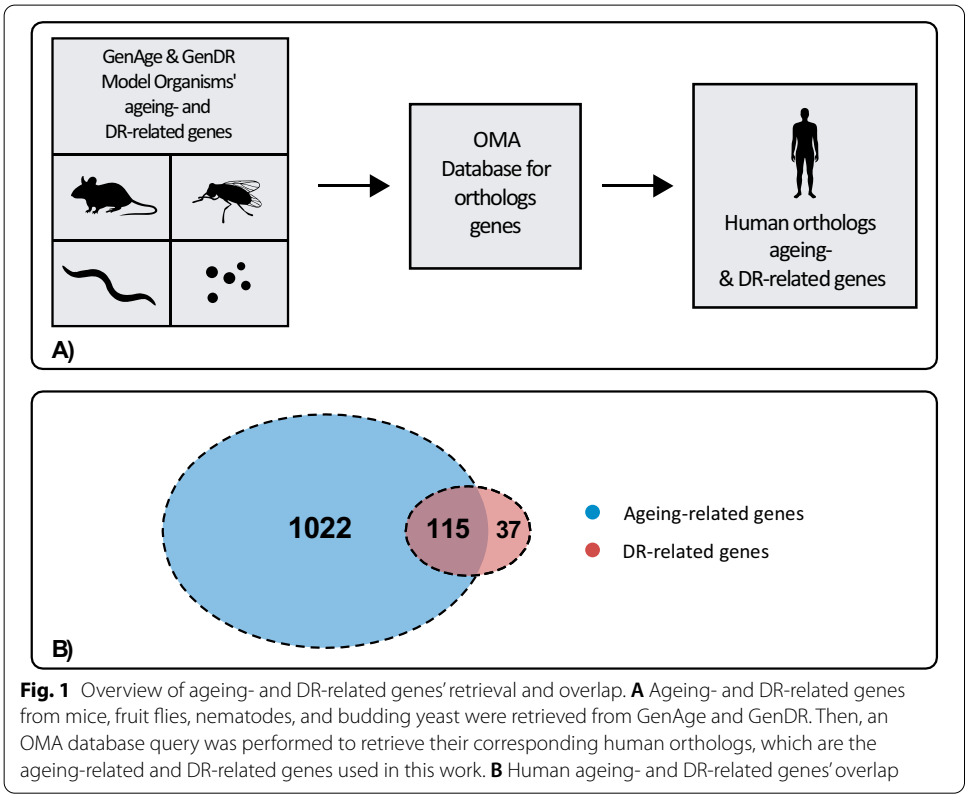

Fig. 1 Overview of ageing- and DR-related genes' retrieval and overlap. **A** Ageing- and DR-related genes from mice, fruit flies, nematodes, and budding yeast were retrieved from GenAge and GenDR. Then, an OMA database query was performed to retrieve their corresponding human orthologs, which are the ageing-related and DR-related genes used in this work. **B** Human ageing- and DR-related genes' overlap

dataset, a data frame relating KEGG pathways with the genes they contain was retrieved by using the *getGeneKEGGLinks* command of the *R*'s Lima package [16]. Then, only the KEGG pathways that were associated with at least one of the ageing-related genes were retained, yielding 312 KEGG pathways. The resulting dataset contained 799 ageing-related genes: 94 labelled as $Ageing_{CR}$-related and 705 labelled as $Ageing_{NotCR}$-related.

KEGG-influence dataset

This approach was inspired by the feature-creation method proposed in [17]. Instead of producing binary features like KEGG pertinence, that method examines the internal contents of each KEGG pathway to produce numerical features, where each feature value measures the extent to which each protein influenced all the other proteins in the pathway. Figure 2a illustrates the influence that a reference protein (red node in the graph) exerts on other proteins (the remaining graph's nodes) within a given pathway. In essence, proteins coloured in dark blue can be "reached" only via upstream paths passing

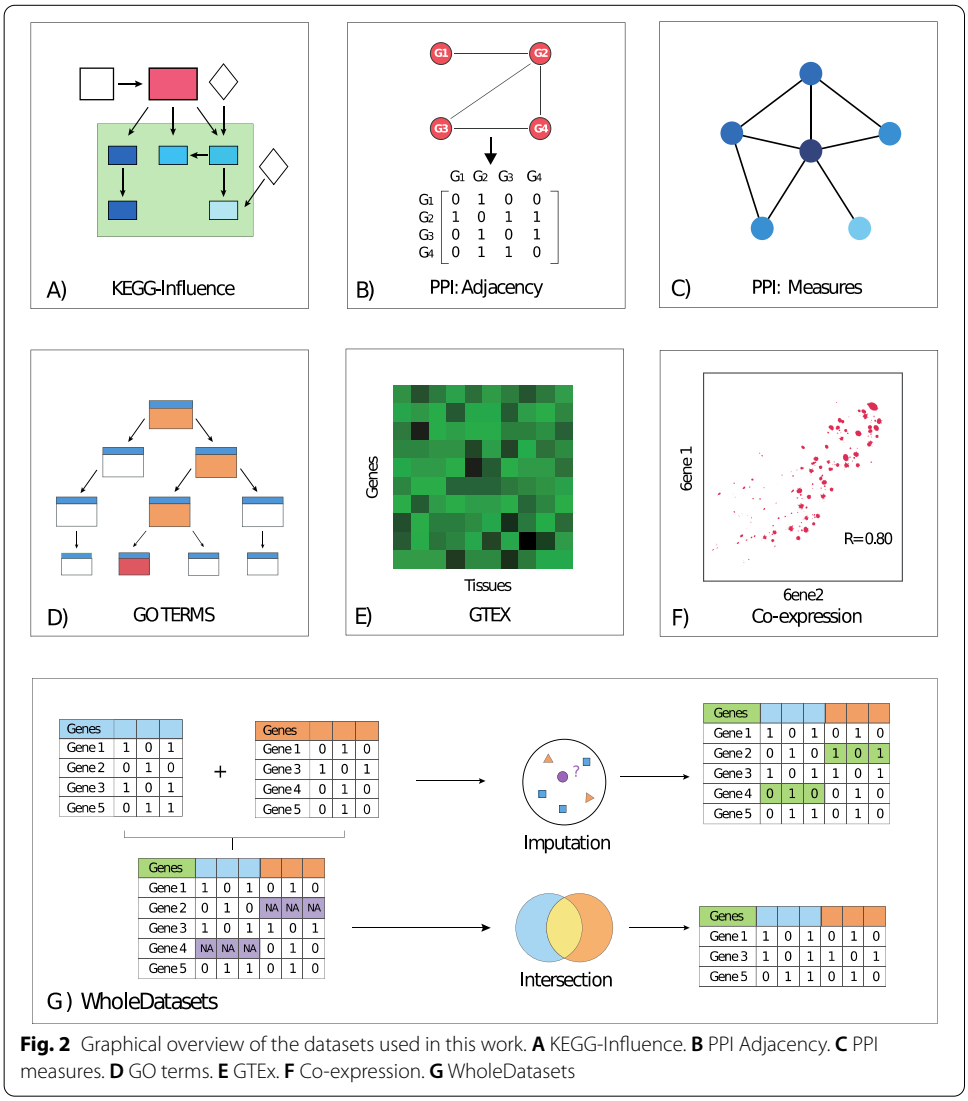

Fig. 2 Graphical overview of the datasets used in this work. **A** KEGG-Influence. **B** PPI Adjacency. **C** PPI measures. **D** GO terms. **E** GTEx. **F** Co-expression. **G** WholeDatasets

through the reference protein and thus are highly influenced by it. As the proteins in the pathway can be reached via more upstream paths not involving the reference protein, they become less influenced by it, and are represented by lighter blue colours. Proteins that receive no influence are coloured in white. Additional file 1: Text S.1.1 contains a complete description of this dataset which contains 1770 features and 799 ageing-related genes, 94 of which are labelled as $Ageing_{DR}$-related and 705 labelled as $Ageing_{NotDR}$ -related.

Protein–protein interaction (PPI) adjacency dataset

The human physical PPI network was downloaded from *BioGrid* (Release 3.5.185): *BIOGRID-MV-Physical*-3.5.181.tab2.*zip* [18]. Interactions were then processed using the R's *igraph* library to create a graph object with a total of 25,292 gene products and 324,892 interactions. After removing loops and repeated edges, the graph consisted of 25,292 nodes and 92,237 edges. This graph contained 850 ageing-related genes, 86 of which were labelled as $Ageing_{DR}$-related while the remaining 764 as $Ageing_{NotCR}$-related. Then, a dataset with binary features was extracted from the PPI-graph's adjacency matrix. The *ij-th* element of this matrix takes the value 1 if the gene products G_i and G_j. are adjacent in the PPI-graph (i.e., if there is an edge connecting them), or 0 otherwise. We retained only ageing-related genes of the adjacency matrix as instances, whereas for features, we kept not only ageing-related genes but also genes not associated with ageing that directly interact with at least one of the ageing-related genes, yielding 5718 features. Figure 2b depicts a representation of this dataset where an adjacency matrix is constructed from a graph.

PPI graph measures dataset

This dataset was created from the same base PPI-graph used to create the PPI-adjacency dataset, but now using as features only 18 graph measures applied to the 850 ageing-related genes within the PPI graph. These measures are classified as centrality- and non-centrality-based and were computed using the *R*'s *igraph* and *Centricerve* libraries [19–21]. Following these libraries' documentation, the centrality measures used in this work were: Leverage, Markov, Maximum neighborhood component, Closeness, Betweenness, Laplacian, Diffusion, Semilocal, Subgraph, Geokpat, Eigenvalue, Eccentricity, Degree and Lobb centralities. In addition, we used the Kcore, DR-ratio, Clustering Coefficient and Topological Coefficient as non-centrality-based measures. A detailed description of all these measures, except DR-ratio, is available on the *Centiserve* webpage [20]. The DR-ratio is the ratio of the number of DR-related direct neighbours of the queried gene over the total number of neighbours of the queried gene (i.e., it describes what percentage of the queried-gene's direct neighbours are DR-related). Figure 2c represents this dataset by displaying a graph whose nodes are coloured based on their degree centrality (the larger the number of neighbours, the higher the degree centrality and the darker the nodes' colour).

GO terms dataset

In this dataset, each binary feature represents a specific GO term with which each ageing-related gene may or may not be associated. To accomplish this, a *BioMart* query

retrieved a list of the Biological Process GO terms associated with the ageing-related genes, yielding 8640 different GO terms. The retrieved GO terms form a hierarchical structure with two properties: (1) if a gene is associated to a specific GO term, it will also be associated with all its ancestors (i.e., more general GO terms); and (2) if a gene is not associated to a given GO term, then the gene is not associated to any of its descendants. The GO term GO:0008150 (named 'Biological Process') is the root of the hierarchy, i.e., it is an ancestor of all other GO terms in the dataset. The *R*'s *GO.db* package [22] was used to retrieve all the ancestors of the originally retrieved GO terms. Those ancestors were then merged to this dataset as new predictive features. This process increased the number of GO terms across all the ageing-related genes from 4877 (i.e., without ancestors) to 8640 (i.e., containing both the original retrieved GO terms and their ancestors). Each GO term (considering both originally retrieved and ancestor GO terms) is represented as a binary feature, indicating whether or not a gene (instance) is annotated with that GO term. Out of the 1137 ageing-related genes, 13 genes were not associated with any GO term in *BioMart* and were removed. This produced a dataset composed of 1124 ageing-related genes: 114 labelled as $Ageing_{DR}$ and 1010 labelled as $Ageing_{NotDR}$. Finally, due to the hierarchical structure of GO terms, any gene associated with a fixed GO term is also associated with all of the term's ancestors. Figure 2d illustrates this phenomenon by indicating that association with a fixed GO term (red node) implies association with all of its ancestors (orange nodes).

GTEx dataset

The median expression levels of human genes across 55 different anatomical tissues were retrieved from the GTEx database [23] (Analysis V8 database) by downloading the file corresponding to the median gene-level Transcripts Per Million (TPM) by tissue. Then, only ageing-related genes were retained, resulting in 1111 ageing-related genes: 114 labelled as $Ageing_{DR}$-related and 997 labelled as $Ageing_{NotDR}$-related. The tissues' median TPM scores were then used as predictive features. A graphical representation of this dataset is illustrated in Fig. 2e through a heatmap of the mean expression of each single gene across different tissues.

Co-expression dataset

The *GeneFriends* database [24] was used to generate co-expression profiles for 1048 ageing-related genes across a set of 44,946 genes that included both the 1048 ageing-related and other genes. The goal was to determine whether the co-expression profile of key genes across the ageing-related genes contributes to the association of certain ageing-related genes with DR. This dataset contained 106 and 942 $Ageing_{DR}$- and $Ageing_{NotDR}$ -related genes, respectively. Figure 2f illustrates a representation of this dataset, where the correlation between the expression of two genes across different samples is depicted.

Protein-descriptor dataset

This dataset contains numeric features associated with the proteins encoded by the ageing-related genes. Since each gene may code for either one or more proteins, this dataset differs from others in the sense that it provides information on ageing-related proteins rather than genes. The names and sequences of human proteins were obtained using

the proteins command in *R*'s *ensembldb* library, with the database *EnsDb.Hsapiens.v86* [25, 26] as the source. 1109 ageing-related genes encoded 6180 ageing-related proteins, from where 115 genes and 514 proteins were designated as $Ageing_{DR}$-related, while the remaining 994 genes and 5666 proteins as $Ageing_{NotDR}$-related. Additionally, features were computed from the amino acid sequences of the *proteins* using the *R*'s *protr* and *Peptides* packages [27, 28], which resulted in the features discussed in Additional file 1: Text S.1.2.

Whole-datasets

Two datasets were created that combine features from PathDIP, KEGG-Pertinence, KEGG-Influence, PPI adjacency matrix, PPI graph measures, GO terms, GTEx expression data, and Co-expression datasets. The protein-descriptors dataset was not included as it only provides information on proteins rather than genes, which complicates the gene-based merging process as proteins do not always have a one-to-one relationship with genes. We coined the term 'WholeDataset' to refer to the resulting dataset that combines all of the aforementioned features, yielding a total of 63,099 features and 1,137 ageing-related genes: 115 labelled as $Ageing_{DR}$ and 1022 labelled as $Ageing_{NotDR}$.

Since the merged datasets had different numbers of ageing-related genes, the WholeDataset contained data gaps for ageing-related genes whose features were not annotated across all the datasets. We addressed this issue using two approaches, namely, imputation and intersection, which are described next and illustrated in Fig. 2g, where the combination of datasets results in genes with missing data (purple cells), which are then imputed (green cells) or removed to leave only genes containing features from all the datasets (intersected genes).

Whole-Dataset-imputation in this approach we used a 5 Nearest Neighbors (5NN) imputation method. For each ageing-related gene G that is missing a value for feature F, this method first determines the top five ageing-related genes that have a known value for F and are most similar (have the smallest Euclidean distance) to G in the training set. The Euclidean distance is computed using all the features for which the values of G are known. If F is a continuous feature, its missing value in G is imputed using the mean value of F across the 5NN of G in the training set. If F is binary, the missing value in G is imputed using the mode of F (i.e., its most frequent value) across the 5NN of G in the training set.

Whole-Dataset-intersection In this approach we only retained the ageing-related genes that are present across every single dataset, except the protein-descriptors dataset. This guarantees the absence of any missing feature values. However, information is lost since only about half of the ageing-related genes had known values for all the features. The resulting dataset contained 628 ageing-related genes: 72 labelled as $Ageing_{DR}$-related and 556 labelled as $Ageing_{NotDR}$-related.

Machine learning

This work focuses on decision tree-based ensembles, which are a type of powerful ML technique that combines the predictions of several base learners (decision trees) in order to improve predictive accuracy over a single base learner, reaching state-of-the-art predictive power while achieving relatively high computational efficiency

[29–31]. This type of ensembles is usually categorised into two broad groups: (1) Bagging methods, where each base learner is trained independently from the others—so, the base learners are conceptually trained in parallel. In bagging methods, the predictive accuracy is usually improved due to the reduction of the variance in the ensemble's predictions, by comparison with the variance in the predictions of a single base learner. (2) Boosting methods, where the base learners are trained sequentially, and each base learner is trained with instance weights that are determined in order to correct the errors of previous base learners in the sequence. This tends to reduce the bias in the predictions [32].

One challenge in this work is that $Ageing_{NotDR}$-related genes are roughly tenfold more numerous than $Ageing_{DR}$-related genes, resulting in an imbalanced data that biases ML predictions towards $Ageing_{NotDR}$-related genes. To address this, under-sampling of the majority class ($Ageing_{NotDR}$-related genes) was performed for each of the base learners' training set. Hence, after under-sampling each training set (for each base learner) has the same number of $Ageing_{DR}$-related and $Ageing_{NotDR}$-related genes. Note that, when performing cross-validation, under-sampling was applied to the training set only, i.e., the test set remains with the original, very imbalanced class distribution, in order to reflect as well as possible the challenge of imbalanced classes associated with the target real-world classification problem.

Two of the ML algorithms we used, Balanced Random Forests (BRF) and Easy Ensemble Classifier (EEC), are bagging and boosting methods, respectively. Both BRF and EEC were implemented using the *Python* package *Imbalanced-learn* [33]. Additionally, XGBoost (XGB) [34] and CatBoost (CAT) [35] were used, as they are both high-performance open-source libraries for gradient boosting in decision trees. Figure 3a illustrates these algorithms graphically. The four techniques were run with *random_state* set to 42.

Predictive accuracy calculation

Predictive accuracy was calculated by using a nested cross-validation (CV) procedure (a common approach in ML), as follows. To implement the outer 10-fold cross-validation, the dataset instances (ageing-related genes) are randomly divided into 10 stratified outer folds of roughly the same size. Then each ML algorithm is run 10 times, each using a different outer fold as the testing set and all the other 9 outer folds as the training set. Before each run of an algorithm, however, its hyperparameters are tuned by an inner 5-fold cross-validation. To implement this, each training set of the outer CV is randomly divided into 5 stratified inner folds of roughly the same size. Then, for each candidate configuration of hyperparameter settings of the algorithm, the algorithm is run 5 times, each time using a different inner fold as the validation set (to measure predictive accuracy) and the other 4 inner folds as a reduced training set. The predictive accuracy of each candidate algorithm configuration is computed as the average accuracy over the 5 validations sets, and then the algorithm configuration with the highest average predictive accuracy is chosen as the best configuration for the current iteration of the outer CV. Next, the algorithm with that best configuration is applied to the full training set of the current iteration of the outer CV, and the learned classifier is evaluated on the current testing set. Finally, this whole process is repeated for the 10 iterations of the outer CV, and the predictive accuracy of the algorithm is computed as the average of the 10 values

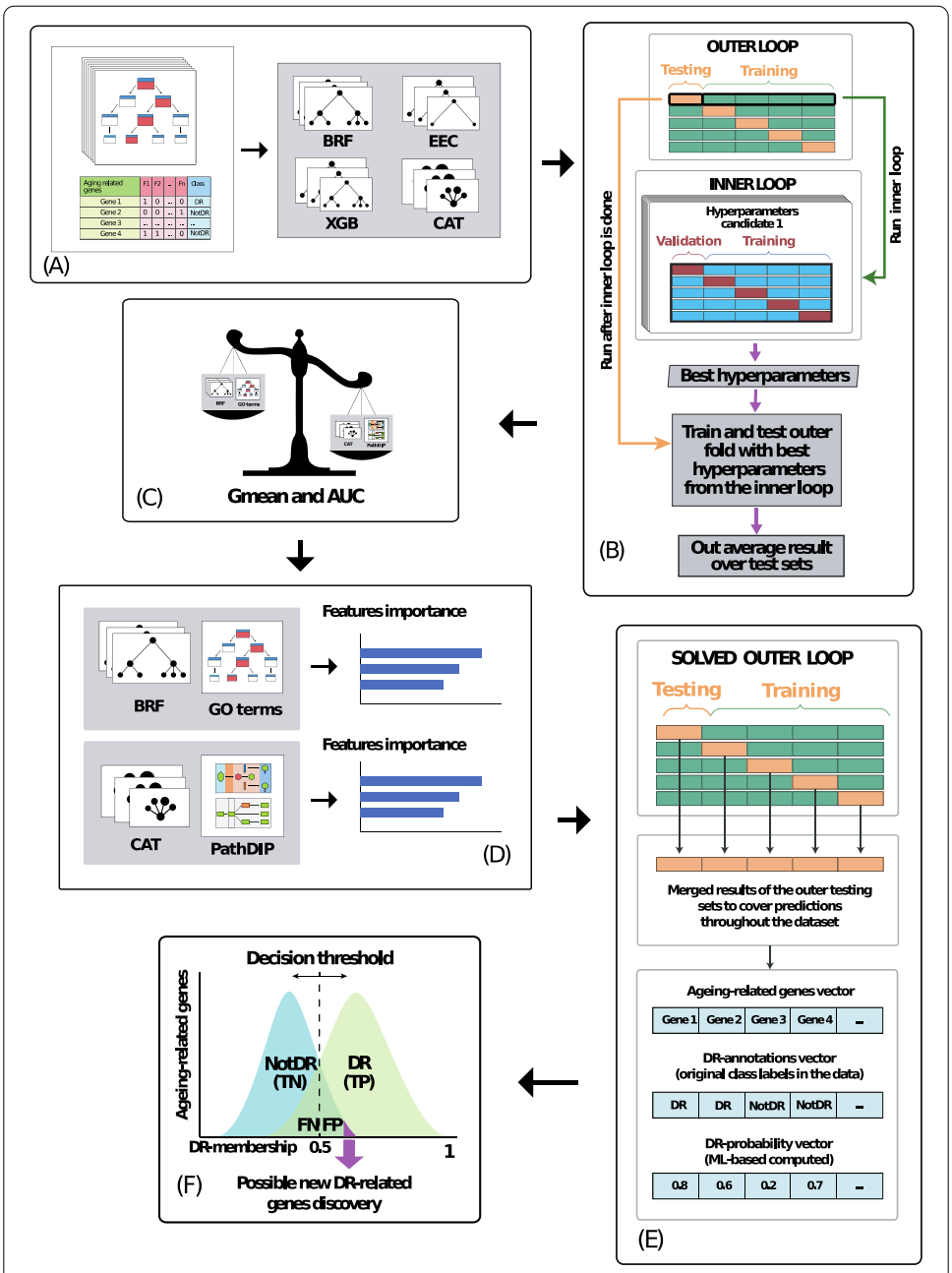

Fig. 3 Summary of ML methods. **A** Selection of each possible {Dataset, ML algorithm} model. **B** Evaluation of each model through the nested-CV. Only 5 outer folds are depicted to facilitate visualization. **C** Comparison of the predictive power of all the {Dataset, ML algorithm} models using Gmean and AUC. **D** Computation of feature importance analysis in the two best performing models from different datasets. **E** Construction of the DR-annotations and the DR-probability vectors. **F** Possible DR-related inference from $Ageing_{NotDR}$ -related genes strongly predicted as DR-related. TN, TP, FN and FP refer to True Negatives, True Positives, False Negatives and False Positives, respectively

of predictive accuracy over the 10 testing sets of the outer CV. Note that hyperparameter optimization is performed by the inner CV using only the training set (i.e., not using the testing set), which is always reserved for measuring generalisation performance.

A graphical representation of the nested CV procedure is depicted in in Fig. 3b (which only displays five outer folds for ease of visualization). In this work, the inner loop was implemented using the *scikit-learn*'s *GridSearchCV* command [29], with *random_state* = 42 and hyperparameters as stated in Additional file 1: Text S.2.

The performance metric used for hyperparameter tuning during the inner-loop of the nested CV was the Geometric Mean (Gmean), defined in equation (1):

$$Gmean = \sqrt{sensitivity \times specificity} \tag{1}$$

where sensitivity is the percentage of $Ageing_{DR}$-related genes (i.e., the minority class) that were correctly predicted as $Ageing_{DR}$-related, and specificity is the percentage of $Ageing_{NotDR}$-related genes (i.e., the majority class) that were correctly predicted as $Ageing_{NotDR}$-related. Since sensitivity and specificity take values in the [0, 1] interval, so does Gmean. Gmean is suited for class-unbalanced problems as this metric measures the balance between classification performances on both the majority and minority classes [36]. The closer Gmean is to 1, the better is the classification.

The predictive performance of the final models on the testing sets of the outer CV was evaluated by two measures (Fig. 3c): (a) Gmean of sensitivity and specificity, and (b) Area Under the Receiver Operating Characteristic Curve (AUC), which is an overall summary of predictive accuracy. AUC also takes values in [0, 1], where 1 is the ideal value (indicating that all predictions were correct), and an AUC value of 0.5 corresponds to random predictions.

A special form of nested CV procedure was applied to the protein-descriptors dataset, as follows. The sequences of ageing-related proteins encoded by a single ageing-related gene are highly correlated. As a result, the testing and training sets of the proteins dataset are also likely to be highly correlated, impeding a fair measure of predictive accuracy. To address this issue, the inner and outer folds of the nested-CV were performed at the gene level rather than at the protein level. This was accomplished by directly applying the nested-CV splitting to all the ageing-related genes containing proteins in the *EnsDb.Hsapiens.v86* database [28, 29]. Following that, the corresponding proteins for each of these ageing-related genes were retrieved. Proteins encoded by the genes in the outer training, outer testing, inner training and inner validation sets were then used to create their corresponding data subsets of the protein-descriptors dataset.

Preprocessing

Two distinct preprocessing methods were applied, depending on the type of data to be computed. For binary features (i.e., PathDIP, KEGG-Pertinence, PPI-Adjacency, GO terms), a minimum threshold of association occurrences was tuned as follows: we defined the 'Thresholds_Set' $Ts = \{3, 4, 5\}$ as the candidate minimum number of instances (i.e., genes) to which a feature must be associated in order to be retained for the ML model's training and testing. The value that maximizes the average prediction performance across all the inner loops of the nested CV is then selected for training the corresponding outer loop.

On the other hand, continuous features (i.e., KEGG-Influence, GTEx, and Proteins descriptors) were filtered based on a correlation criterion where one of each two features with correlation greater than 99% was eliminated. The co-expression dataset, despite

continuous, was preprocessed differently from other continuous datasets due to its remarkably larger number of features that made correlation analysis highly demanding in terms of computational power and time. To address this, an F-statistic-based univariate feature filtering algorithm was applied using the *SelectKBest* command of the *sklearn* feature_selection python's library to retrieve the 1000 genes that co-express the most with the DR-related target labels. Then, similarly to other continuous datasets, the 99% correlation-based filtering was applied to the remaining 1000 genes.

WholeDatasets' features were classified into two categories: binary, which underwent the 'Thresholds_Set' preprocessing step $Ts = \{3, 4, 5\}$; and continuous, to which the univariate filter was applied to reduce the number of continuous features to 1000, followed by the 99% correlation step. Some KEGG-Influence-based features had only two values and thus were preprocessed as binary within the WholeDatasets, even though this dataset was created by trying to compute all its features as continuous. This implies that some KEGG Influence-based features within the WholeDatasets were preprocessed by the minimum occurrence threshold criterion, while others by the univariate feature plus the correlation criterion.

Feature importance calculation

We calculated the feature importance for the best learned models (Fig. 3d). To accomplish this, we used 100% of the ageing-related genes (instances) of the dataset under study as the training set, which ensured that the features importance were calculated using all of the data available, maximising the quality of the feature importance calculation. No instances were withheld for testing purposes, as this task's objective was not to determine predictive accuracy (already determined by the nested-CV procedure) but to compute the features' predictive relevance.

The importance of BRF's features is determined by using the default Gini index of class impurity, which calculates how well a split separates the samples of the two classes in each node of a decision tree. A feature's importance is basically given by the weighted average of the reduction of the Gini index across the tree nodes labelled with that feature, with weights proportional to the number of instances split by that feature [37]. On the other hand, EEC, XGB, and CAT calculate a feature's relevance through the permutation method, which compares the model's predictive accuracy on the original data vs the model's accuracy on a dataset with a random permutation of that feature's values, so that the extent of the drop in the model's accuracy after the random permutation indicates how much the model is dependent on the feature. Finally, in order to compare the results of the two feature importance methods, the resulting feature rankings were scaled from their original values to the [0, 100] interval, where 100 represents the most important feature and 0 represents no relevance.

New DR-related genes inference

Each learned classification model outputs a probability that a gene belongs to the $Ageing_{DR}$-related class. For converting a predicted probability into a class label we use a classification threshold of 0.5, i.e., any gene with a predicted $Ageing_{DR}$-related probability less than 0.5 is predicted to be $Ageing_{NotDR}$-related, whereas any gene with a predicted probability greater than 0.5 is predicted to be $Ageing_{DR}$-related.

Although DR-related predictions are binarized by the threshold, the DR-related probabilities remain informative as a measure of the prediction's certainty, from the model's viewpoint. For instance, if two given genes, *A* and *B*, have DR-related probabilities of 0.6 and 0.9, respectively, both are classified as DR-related; but from the model's perspective, gene *B* is more reliably related with DR.

Hence, it is possible to infer novel $Ageing_{DR}$-related genes by identifying, among all the genes annotated in the dataset as $Ageing_{NotDR}$-related genes, which ones have the greatest predicted $Ageing_{DR}$-related probabilities, which are the strongest false positives (FP) genes. After all, the $Ageing_{NotDR}$-related class label annotation in the dataset is not very reliable in general, because it basically means that there is no evidence for $Ageing_{DR}$-relatedness in the literature, and absence of evidence is not the same as evidence for absence of $Ageing_{DR}$-relatedness. Hence, the strongest FP genes are good candidate targets for future experiments to determine $Ageing_{DR}$-related genes.

Keeping in mind that the nested-CV's outer testing sets do not overlap and that they collectively include all the genes (instances) in the dataset, we created, for each of the best performing Dataset, ML algorithm combinations (i.e., the best models), two vectors: a *DR-probability vector* combining the predicted DR-related probabilities from all the outer testing splits, and a *DR-annotations vector* by combining the original annotations of class labels in the data from all the outer splits (Fig. 3e). Next, based on these two vectors, we retained only $Ageing_{NotDR}$-related genes with DR-probability equal to or greater than 0.5 (i.e., FP genes), meaning that they were classified as DR-related by the model but lack a DR-related annotation in the dataset (based on the literature). We identified the top 10 FP genes with greater DR-probability for each of the best performing models and discussed their potential as candidate DR-related genes for confirmation in further wet-lab experiments (Fig. 3f).

Finally, we looked for common top DR-related genes candidates among the shared ageing-related genes between the two strongest models, namely {GO terms, BRF} and {PathDIP, CAT}, as described in the *Results* section. To do so, we originally computed a common-ranking by averaging the DR-probability scores of common genes in both models and then sorted the genes in descendent order based on the averaged score. This approach had, however, one issue as the DR-probability density distributions of both strongest models lied in different intervals ([0.35,0.75] in {GO terms, BRF} while [0, 1] in {PathDIP, CAT}, as described in Results). Consequently, each computed average was biased towards the distribution with the most extreme values (i.e., genes with the greatest/lowest DR-probability scores in {PathDIP, CAT} will have stronger influence when averaging than genes with greatest/lowest DR-membership score in {GO terms, BRF}). With the aim to provide a similar comparison scale for the top DR-related candidates in both models while considering the distribution shape, we mapped the DR-membership scores of all genes in both {GO terms, BRF} and {PathDIP, CAT} models to the [0, 1] interval and then retrieved common genes in both models to compute DR-probability arithmetic averages, highlighting $Ageing_{NotDR}$-related genes with the highest probabilities of being $Ageing_{DR}$-related genes from both best models' perspectives.

Statistical analysis

To report on the most important features, two statistical analysis tests were used (with a significance level of 0.01): a Two-Proportions Z-Test for binary features and a T-test for continuous features. The use of the test for binary features is based on the concept of a feature's positive value, which is defined as the presence of annotation (*e.g.*, a GO term annotation) for a gene, as opposed to the absence of that annotation (the negative value of the feature).

The test for binary features was designed to determine whether the proportion of $Ageing_{DR}$-related genes associated with a particular relevant feature's positive value (i.e., the ratio of $Ageing_{DR}$-related genes associated with the relevant feature's positive value over the total number of $Ageing_{DR}$-related genes in the dataset) is significantly different than the proportion of $Ageing_{NotDR}$-related genes associated with the same relevant feature's positive value. For continuous features, the test determined whether the mean value of the feature across all $Ageing_{DR}$-related genes was significantly different than the mean value of the feature across all $Ageing_{NotDR}$-related genes. The resulting p-values were adjusted for multiple tests using the Benjamini–hochberg correction.

Results

This section first highlights the best performing combinations of ML algorithms and datasets. Top relevant features are then presented. Finally, the predicted top DR-associated genes candidates are reported.

Predictive accuracy results

The AUC and Gmean values obtained by BRF, EEC, XGB, CAT are compared in Table 1. For each dataset (feature type) and predictive accuracy measure, the best result from the four algorithms is highlighted in bold face. This bold face meaning also holds in Table 2. GTEx and co-expression were the least predictive feature types, yielding results comparable to random predictions (AUC close to 0.5), whereas GO terms and PathDIP were the most predictive: GO terms had the highest average AUC (0.83) and the second highest average Gmean (0.75) across the four algorithms; whilst PathDIP had the highest average Gmean (0.76) and the second highest average AUC (0.81). BRF got the highest AUC values overall, whereas the highest Gmean values were more distributed across all algorithms, with higher means for EEC and CAT.

By defining a model as a combination of a dataset and the classification algorithm that runs over it {Dataset, ML algorithm}, the model with the highest Gmean (0.77) was {PathDIP, CAT}, closely followed by {GO Terms, BRF} and {PathDIP, BRF}, both with a Gmean of 0.76. Regarding AUC results, the best model was {GO Terms, BRF} with an AUC of 0.84, closely followed by {GO terms, EEC} and {PathDIP, CAT}, both with an AUC of 0.83. Since {PathDIP, CAT} and {GO Terms, BRF} achieved complementary and notably close first and second places regarding Gmean and AUC, they are both the most predictive models overall.

Table 2 reports sensitivity and specificity results, as measures of predictive accuracy for $Ageing_{DR}$-related genes and $Ageing_{NotDR}$-related genes, respectively. The highest mean sensitivity and specificity values across all four algorithms were obtained by

Table 1 Gmean and AUC scores across all {ML algorithm, Dataset} models

Feature type	Num. of features	Num. of instances	Gmean				AUC				Mean	
			BRF	EEC	XGB	CAT	BRF	EEC	XGB	CAT	Gmean	AUC
PathDIP	1640	986 genes	0.76	0.75	0.75	**0.77**	0.81	0.8	0.8	**0.83**	**0.76**	0.81
KEGG-Pertinence	312	799 genes	0.73	0.7	**0.75**	0.72	**0.8**	0.71	0.76	0.73	0.73	0.75
KEGG-Influence	1770	799 genes	**0.7**	**0.7**	0.67	0.67	**0.78**	0.71	0.69	0.68	0.69	0.72
PPI-measures	18	850 genes	0.64	0.59	**0.65**	**0.65**	**0.69**	0.6	0.65	0.67	0.63	0.65
PPI-adjacency	5718	850 genes	0.62	0.61	0.59	**0.65**	**0.72**	0.63	0.64	0.66	0.62	0.66
GO terms	8640	1124 genes	**0.76**	0.75	0.74	0.74	**0.84**	0.83	0.83	0.81	0.75	**0.83**
GTEx	55	1111 genes	0.5	**0.52**	0.5	0.51	0.5	**0.53**	0.52	0.52	0.51	0.52
Co-expression	44,946	1048 genes	0.5	**0.57**	0.42	0.56	0.52	**0.57**	0.54	0.54	0.51	0.54
Proteins-descriptors	156	6180 Proteins (from 1109 genes)	0.45	0.61	**0.62**	0.6	0.65	**0.67**	**0.67**	0.63	0.57	0.66
Whole-dataset imputation	63,099	1137 genes	**0.67**	0.65	0.6	0.66	**0.72**	0.66	0.65	0.67	0.65	0.68
Whole-dataset intersection	63,099	628 genes	0.61	**0.66**	0.56	0.64	0.64	**0.67**	0.63	0.66	0.62	0.65
Mean	–	–	0.63	**0.65**	0.62	**0.65**	**0.70**	0.67	0.67	0.67	0.64	0.68

Gmean and AUC scores for the BRF, EEC, XGB, and CAT algorithms across all the 11 datasets (feature types). The average Gmean and AUC of each algorithm across all the datasets is shown in the last row; and the average Gmean and AUC of each feature type across all four algorithms is shown in the last two columns

Table 2 Sensitivity and specificity values across all {ML algorithm, Dataset} models

Feature type	Num. of features	Num. of instances	Sensitivity				Specificity				Mean	
			BRF	EEC	XGB	CAT	BRF	EEC	XGB	CAT	Sens	Spec
PathDIP	1640	986 genes	**0.87**	0.77	0.58	0.77	0.67	0.7	**0.93**	0.77	**0.75**	0.77
KEGG-Pertinence	312	799 genes	0.74	0.70	**0.75**	0.72	0.7	0.71	**0.76**	0.73	0.73	0.73
KEGG-Influence	1770	799 genes	**0.75**	0.70	0.67	0.67	0.66	**0.71**	0.69	0.68	0.70	0.69
PPI-measures	18	850 genes	**0.64**	0.62	0.6	0.57	0.65	0.57	0.71	**0.77**	0.61	0.68
PPI-adjacency	5718	850 genes	**0.65**	0.52	0.5	0.65	0.62	0.74	**0.78**	0.67	0.58	0.70
GO terms	8640	1124 genes	**0.85**	0.80	0.48	0.69	0.67	0.72	**0.94**	0.78	0.71	0.78
GTEx	55	1111 genes	0.54	0.55	0.45	**0.57**	0.48	0.51	**0.57**	0.48	0.53	0.51
Co-expression	44,946	1048 genes	0.58	**0.61**	0.21	0.56	0.45	**0.86**	0.85	0.54	0.49	0.68
Proteins-Descriptors	156	6180 Proteins (from 1109 genes)	0.25	0.43	0.45	**0.48**	0.86	**0.91**	0.88	0.77	0.40	**0.86**
Whole-Dataset Imputation	63,099	1137 genes	0.69	0.68	0.43	**0.71**	0.65	0.63	**0.87**	0.64	0.63	0.70
Whole-Dataset Intersection	63,099	628 genes	0.60	0.72	0.39	**0.75**	0.63	0.62	**0.87**	0.56	0.62	0.67
Mean	–	–	**0.65**	**0.65**	0.50	0.65	0.64	0.70	**0.80**	0.67	0.61	0.70

Sensitivity ($Ageing_{DR}$ prediction quality) and specificity ($Ageing_{NotDR}$ prediction quality) values for the BRF, EEC, XGB, and CAT algorithms across all the datasets (feature types). The average sensitivity and specificity of each algorithm across all the datasets is shown in the last row; and the average Gmean and AUC of each feature type across all four algorithms is shown in the last two columns

Table 3 Top-five most predictive GO terms in the {GO terms, BRF} model

Feature	Definition	Score	$Ageing_{DR}$ (%)	$Ageing_{NotDR}$ (%)	Adjusted *p* value
GO:0055114	Oxidation–reduction process	100	19.3% {22/114}	11.58% {117/1010}	1
GO:0007188	Adenylatecyclase-modulating G protein-coupled receptor signalling pathway	70.965	13.16% {15/114}	0.3% {3/1010}	**1.52e−20**
GO:1,904,659	Glucose transmembrane transport	67.831	10.53% {12/114}	0% {0/1010}	**3.76e−20**
GO:0007186	G protein-coupled receptor signalling pathway	53.516	15.79% {18/114}	2.77% {28/1010}	**1.21e-07**
GO:0008643	Carbohydrate transport	51.516	9.65% {11/114}	0.1% {1/1010}	**3.51e−16**

Table 4 Top-five most predictive pathways in the {PathDIP, CAT} model

Feature	Definition	Score	$Ageing_{DR}$ (%)	$Ageing_{NotDR}$ (%)	Adjusted *p* value
KEGG.2	Autophagy (animal)	100	18.18% {20/110}	4.34% {38/876}	**3.05e−0.5**
KEGG.30	Longevity regulating pathway (multiple. species)	45.349	12.73% {14/110}	2.63% {23/876}	**8.71e−04**
NetPath.23	*BDNF*	38.953	0.91% {1/110}	1.14% {10/876}	1
REACTOME.10	Cellular responses to external stimuli	37.791	20.91% {23/110}	9.13% {80/876}	3.90e−01
WikiPathways.37	NRF2	34.884	14.55% {16/110}	1.14% {10/876}	**2.58e−12**

PathDIP and Proteins-Descriptors, respectively, as shown in the last two columns of the table. There was no strong winner algorithm in terms of sensitivity, but XGB obtained by far the worst sensitivity values overall, as shown in the last row of the table. On the other hand, XGB achieved in general the highest specificity values, implying more accurate predictions for $Ageing_{NotDR}$-related genes, the majority class. ROC curves of all the ML algorithms in the two strongest datasets, as well as confusion matrices of the two strongest models, {GO Terms, BRF} and {PathDIP, CAT}, are depicted in Additional file 1: Figures S1 and S2, respectively.

Feature importance results

The top-5 most relevant features in each of the two most predictive models, {GO terms, BRF} and {PathDIP, CAT}, are shown in Tables 3 and 4, respectively. The column *Score* in these tables indicates the relative importance of the features, in the range from 0 (no relevance) to 100 (maximal relevance).

The columns $Ageing_{DR}$ and $Ageing_{NotDR}$ denote the percentage of $Ageing_{DR}$- and $Ageing_{NotDR}$-related genes with the GO term or pathway in the corresponding row (i.e., the percentage of genes with the feature's positive value). In addition, a proportion is provided in brackets where the numerator indicates the number of $Ageing_{DR}$- or

$Ageing_{NotDR}$-related genes with the positive feature value, while the denominator indicates the total number of $Ageing_{DR}$- or $Ageing_{NotDR}$-related genes, in the dataset.

The *Adjusted p-value* columns in Tables 3 and 4 indicate the results of the statistical tests applied to detect whether the values in the $Ageing_{DR}$ column are significantly different from those in the $Ageing_{NotDR}$ column as explained in Methods: Statistical analysis. Significant p-values are denoted by bold text.

Interestingly, among the GO terms in Table 3, the most relevant one, oxidation-reduction process, was the only one that failed to achieve a significant difference in terms of the proportion of $Ageing_{DR}$- and $Ageing_{NotDR}$-related genes. Nonetheless, this term is worth highlighting due to it having the highest proportion of occurrence (19.3%) in $Ageing_{DR}$-related genes among all 5 GO terms in this table. The remaining GO terms, which are related to GPCRs and carbohydrate transport, clearly have a stronger association with $Ageing_{DR}$-related genes, as each of them occurred in about 10%–16% of the $Ageing_{DR}$-related genes while occurring in less than 3% of the $Ageing_{NotDR}$-related genes.

Table 4 reports the top PathDIP features. Note that the score for the second best pathway, longevity regulating pathway, is much lower than the score for the best pathway, autophagy. The only features with no significant difference in their percentage of occurrence in $Ageing_{DR}$- and $Ageing_{NotDR}$-related genes were 'cellular responses to external stimuli' and 'brain-derived neurotrophic factor' (BDNF). Even so, the cellular responses to external stimuli pathway contained the greatest proportion of occurrence (20.9%) in $Ageing_{DR}$-genes.

We also provide feature importance results for {KEGG-Pertinence, XGB} and {KEGG-Influence, BRF} in Additional file 1: Text S.3.1 and S.3.2 since these two models also got relatively high-performance scores, while providing complementary insights to the feature importance results of the most predictive models.

DR-associated gene prediction

The two most predictive models, {GO terms, BRF} and {PathDIP, CAT}, were learned from datasets with 1124 and 986 ageing-related genes, respectively. Figure 4a, b show the distribution of $Ageing_{DR}$- and $Ageing_{NotDR}$-related genes across different DR-probability values for {GO terms, BRF} and {PathDIP, CAT}, respectively. The DR-probability distributions in {GO terms, BRF} span a much narrower window, near [0.35, 0.70], than distributions in {PathDIP, CAT}, near [0, 1]. Additionally, both models' $Ageing_{NotDR}$-related genes share a maximal density point that is relatively close to the 0.5 threshold, yet on the $Ageing_{NotDR}$ predicted side (0.45 points). This point is notably denser (about 6 folds) and narrower in {GO terms, BRF} than it is in {PathDIP, CAT}. Regarding the top DR candidate genes, further analysis of Fig. 4a, b, indicates that, for both models, the top DR-probabilities across $Ageing_{NotDR}$-related genes achieved similar scores to the top DR-probabilities in $Ageing_{DR}$-related genes, but are much less numerous. Figure 4c, d depict the top ten DR-candidate genes in the {GO terms, BRF} and {PathDIP, CAT} models, respectively. Even if some of the top genes predicted by both models may have a proclivity for not detecting unknown DR-relationship, it is remarkable that among their top ten DR-candidate genes, only *GOT2* overlapped.

Hypothesising that pertinence to the common set of top-10 DR-candidate genes in both models increases likelihood of accurate DR-relatedness prediction, we performed

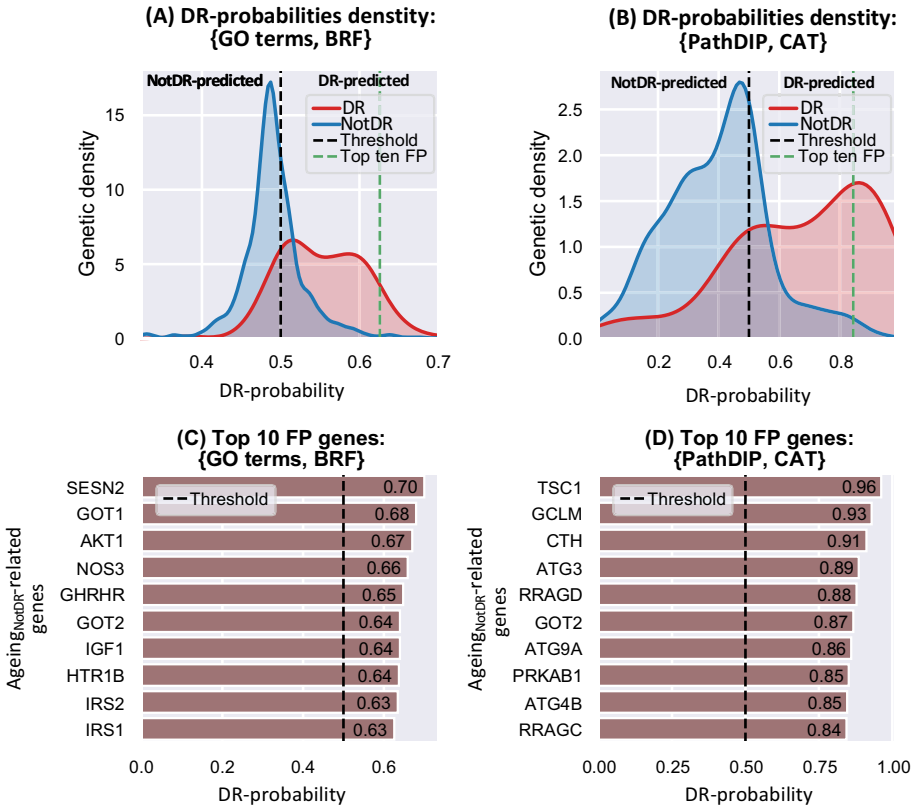

Fig. 4 DR-probabilities assignations of {GO terms, BRF} and {PathDIP, CAT}. **A**, **B** DR-probabilities density distribution of {GO terms, BRF} and {PathDIP, CAT}, respectively. The green dashed line represents the beginning of the top 10 False Positives (FP), which are identified as candidate DR-associated genes. **C**, **D** Top 10 FP genes in {GO terms, BRF} and {PathDIP, CAT}, respectively. The classification threshold (0.5) is represented by the dashed black line

a joint analysis of the top-10 DR-candidate genes (*Methods: New DR-related genes inference*). Briefly, we normalised the range of both models' DR-probability distributions and then retained common ageing-related genes (976 genes, 872 of which are $Ageing_{NotDR}$ -related) in order to compute, for each common ageing-related gene, an arithmetic average of both normalised DR-probabilities. Then, we sorted genes under a criterion that considers both a similar DR-membership range for the two models and their distribution shapes. The correlation between the normalised DR-probabilities assigned by the {GO terms, BRF} and {PathDIP, CAT} models was only moderated, being smaller across $Ageing_{NotDR}$-related genes, increasing throughout $Ageing_{DR}$-related genes and yielding the highest score when using all ageing-related genes (Fig. 5).

Table 5 depicts both models' common $Ageing_{NotDR}$-related genes whose averaged normalised DR-probability exceed 0.8. Note that all of these genes, namely, *GOT2*, *GOT1*, *TSC1*, *CTH*, *GCLM*, *IRS2*, and *SENS2*; appeared in the top-10 DR-related candidates of at least one of the two most predictive models. The top gene in Table 5, Glutamic-Oxaloacetic Transaminase 2 (*GOT2*), appeared within the set of top six ranking genes of both models, indicating that its possible DR-relatedness could be similarly inferred from either biological pathways or biological processes GO terms.

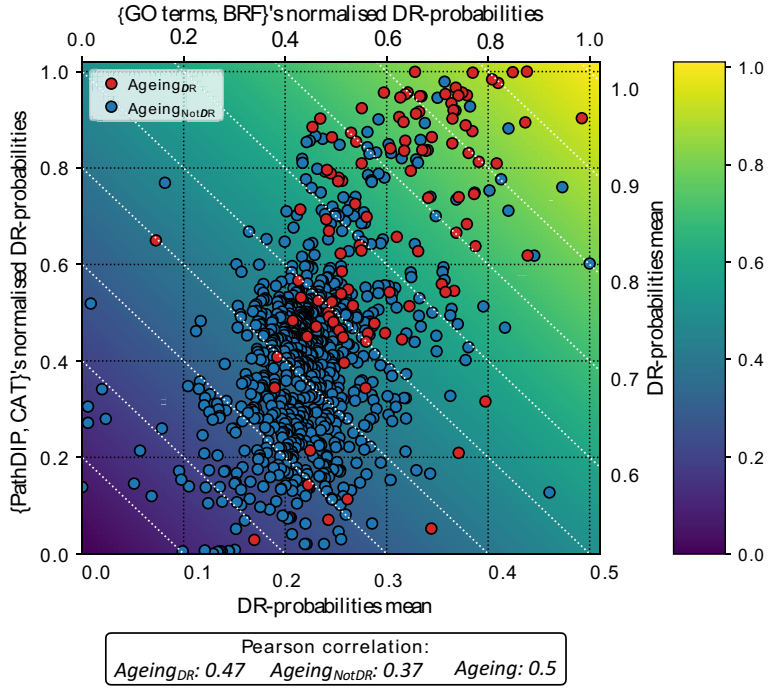

Fig. 5 Correlation of normalised DR-probabilities assigned by {GO terms, BRF} and {PathDIP, CAT} models. The arithmetic mean of both models' normalised DR-probabilities is depicted through the background color while the white dotted lines represent points in the plane with fixed mean values indicated by the grey ticks. The Pearson correlation score between the normalised DR-probabilities computed by the two models is depicted in the text box at the bottom of the figure

Table 5 Top DR-related gene candidates

Joint rank	Gene	Normalised DR-probability			Ranking	
		Mean score	{GO terms, BRF}	{PathDIP, CAT}	{GO terms, BRF}	{PathDIP, CAT}
1	*GOT2*	0.861	0.840	0.882	5th	6th
2	*GOT1*	0.853	0.946	0.760	2th	35th
3	*TSC1*	0.847	0.714	0.979	19th	1th
4	*CTH*	0.846	0.764	0.928	11th	3th
5	*GCLM*	0.823	0.700	0.946	23th	2th
6	*IRS2*	0.801	0.826	0.777	8th	31th
7	*SESN2*	0.801	1	0.602	1th	80th

Top DR-related gene candidates jointly defined by the two strongest models. Top-association ranking is provided relative to the number of common $Ageing_{NotDR}$-related genes

Insulin Receptor Substrate 2 (*IRS2*) and especially Glutamic-Oxaloacetic Transaminase 1 (*GOT1*) got high DR-probabilities in the {GO terms, BRF} model, and a moderately high probability in {PathDIP, CAT}. Moreover, Sestrin 2 (*SENS2*), the gene with the highest DR-probability in the {GO terms, BRF} model, also reached the list but it was, by far, the gene with lowest DR-probability in {PathDIP, CAT}, among all genes

in Table 5. Finally, TSC Complex Subunit 1 (*TSC1*), glutamate-cysteine ligase regulatory subunit (*GCLM*) and Cystathionine Gamma-Lyase (*CTH*) got the three highest DR-probability scores in {PathDIP, CAT}, and also relatively high probabilities in {GO terms, BRF}.

Discussion

This study demonstrated that the most powerful predictors of DR-relatedness across ageing-related genes rely on features heavily based on curated biological knowledge (from the literature), hereafter called 'knowledge-based features', such as GO terms and biological pathways. However, one caveat of these features is the difficulty to directly produce new findings, as they are based on existing knowledge. Nonetheless, the predictive power of these types of features enabled the extraction of additional biological insights from ageing-related genes strongly predicted to be DR-related but lacking current annotation of DR-relatedness.

Features not so heavily based on curated biological knowledge, particularly those based on gene expression and co-expression, were the least predictive, predicting almost randomly and implying that DR-relatedness is unlikely to be explained by gene expression analysis across tissues or by co-expression of ageing-related genes with other genes.

Upon merging the 9 datasets with specific feature types into two "whole" datasets (using two merging approaches), the predictive power was found to diminish compared to the strongest feature type-specific datasets. This occurred despite the use of a simple, univariate feature filtering algorithm. This suggests the importance of exploring the use of more sophisticated feature selection techniques, which is out of the scope of this work and left for future research.

The joint analysis of the top DR-candidate genes in {PathDIP, CAT} and {GOterms, BRF} models highlighted genes whose DR-relatedness can be explored from both GO terms and biological pathways perspectives. The strengths of both models were complementary, as {PathDIP, CAT} was more suited for classifying $Ageing_{NotDR}$-related genes while {GO terms, BRF} performed better with $Ageing_{CR}$-related genes. It is also noticeable that both models achieved similar predictive accuracies despite exhibiting only a moderate correlation of DR-related class predictions. In addition, top-ranked $Ageing_{NotDR}$-related genes with high DR-probabilities were mostly surrounded by $Ageing_{DR}$-related genes in the plane defined by the normalised DR-probabilities in {PathDIP, CAT} and {GO terms, BRF} models, suggesting biological process and pathway similarities between the top DR-related candidates and currently established DR-related genes.

One limitation of this work is that our most predictive models ({GO terms, BRF} and {PathDIP, CAT}) may not always be optimal for predicting DR-association of genes outside GenAge. This occurs because the input features used for learning these two models (and our other models learned in this work in general) have features whose values were computed based on ageing-related genes. That is, since the ML models based on our datasets were trained based on data containing exclusively ageing-related instances, the models were optimized for predicting whether or not ageing-related genes (included in GenAge) belong to the DR-relatedness class; the models were not optimized for predicting whether or not a non-ageing-related gene (not included in GenAge) belongs to the DR-relatedness class.

Features importance

GO term features

The oxidation–reduction process was the most significant GO term for discriminating between the presence or absence of DR-relatedness across ageing-related genes. This term was also notable as it was associated with the greatest number of ageing-related genes across all top features in both PathDIP and GO term feature sets. Nonetheless, it did not demonstrate any significant preference for $Ageing_{DR}$- nor $Ageing_{NotDR}$-related genes, indicating that its effects on ageing could be mediated in both DR and NotDR-dependent ways. Evidence indicates that DR improves redox state [38], though this may not be the mechanism by which DR prolongs life [39]. It has been observed, however, that low levels of ROS may actually be beneficial as mediators of redox signalling [39].

Notably, ageing-related genes associated with glucose and carbohydrate transport were almost exclusively $Ageing_{DR}$-related, which at first sight may partially suggest a relationship between ageing, DR and glucose transport. Nevertheless, this relationship, if existing, is not straightforward as abnormal glucose metabolism is a common but not necessary feature of ageing [40]. The most significant changes in glucose metabolism are due to ageing-related insulin dysfunction [41]. This phenomenon appears, however, to be a consequence rather than a cause of ageing, as the improvement in insulin sensitivity induced by DR was not required for the effects of DR on fitness and longevity [42].

Ageing-related genes within the GPCR signalling pathway were also significantly related with DR. Some GPCRs have emerged as promising targets for reversing senescence and thus ageing [43]. In this regard, one of the few studies discussing the relationship between a GPCR, namely *TGR5*, and DR [44] demonstrated that DR benefits on renal function ageing can be partially explained by up-regulation of *TGR5*; as a result, the authors of that study proposed up-regulation of *TGR5* as a possible DR-mimetic candidate for renal function.

PathDIP features

The strongest predictive feature was autophagy, which is responsible for the disposal and recycling of metabolic macromolecules and damaged organelles [45]. The second most significant feature, the longevity regulating pathway, is also associated with autophagy via a well-characterized signalling cascade [45]. The fact that the $Ageing_{DR}$-related genes in this study were significantly strongly associated with these autophagy-related pathways supports current hypotheses that some of the DR-related anti-aging effects are mediated by autophagy [46, 47].

The BDNF pathway is presumably involved in brain ageing. Moreover, it is well established that DR enhances *BDNF* in a currently unknown manner [48–50], and that *BDNF* declines with age [51]. The possible relationship between this gene and DR could be explored through the well-characterized DR-related protein kinase B (Akt) pathway, as *BNDF* and *Akt* indirectly interact [52]. However, in the context of our ML algorithms, it is possible that the BDNF pathway favoured NotDR-related predictions, as only one of its 11 ageing-related genes was predicted as DR-related. *NRF2* is absent from GenDR. Nevertheless, our results suggest a strong association between the NRF2's pathway and DR because, while the ratio of $Ageing_{DR}$- related to $Ageing_{NotDR}$-related genes is

approximately 1/10 in the overall dataset, this pathway demonstrated the much greater 16/10 ratio; and this pathway has by far the greatest proportion of $Ageing_{DR}$-related genes, compared to $Ageing_{NotDR}$-related genes across all top PathDIP pathways. This finding could be supported by recent evidence linking *NRF2* and DR [53].

The "cellular responses to external stimuli" pathway is notable for the large number of ageing-related genes associated with it, only outnumbered, across all the most relevant features discussed in this work, by the "oxidation-reduction process" GO term. However, similar to the redox GO term, the relative distribution of ageing-related genes in this pathway did not achieve statistical significance in direction of neither DR nor NotDR. External stimuli responses include responses to metal ions [54], from where a metal ion theory of ageing was constructed. This theory has been poorly explored and opens opportunities for novel DR-research directions as it has been shown that DR decreases the level of certain metal ions in cells [55].

DR-related candidate genes

GOT1 and *GOT2* are genes whose products are involved in the amino acid metabolism that exist in cytoplasmic and mitochondrial forms, respectively [56]. *GOT1* 's expression has been shown to change with age [57], but evidence linking it to DR is far scarcer. To our knowledge, only one study [58] has demonstrated this relationship and proposed the role of *GOT1* as a significant metabolic feature associated with hepatic response to DR that is representative of differences in mediating amino acid influx into the gluconeogenic pathway. While there is lack of evidence linking *GOT2* with DR, one study [59] suggested that either *GOT1* or *GOT2* may impact H2S homeostasis, opening a window for further DR-related insights, as H2S signalling cascade has been observed to promote DR-like pro-longevity effects [60].

The TSC complex is a critical negative regulator of mTORC1 [56], the inhibition of which is associated with DR-like benefits [61]. In this regard, even if the TSC complex's role in regulating mTORC1 in vivo remains under-explored, one study [62] provides insights that indirectly link this gene with DR as it demonstrated that improved insulin sensitivity following short-term protein restriction (PR) required TSC1 for facilitating increased pro-survival signalling after injury, and contributed to PR-mediated resistance to clinically significant hepatic ischemia reperfusion injury.

CTH is a gene that produces endogenous hydrogen sulphide (H2S) as a signalling molecule [56]. Evidence linking *CTH* to DR is scarce. To our knowledge, only one recent study [63] reveals a positive correlation between *CTH* expression and DR application. This increased expression may have potential contributions to DR pro-longevity effects as inhibition of *CTH* is associated with about 15% lifespan reduction in worms. Moreover, *CTH* is a gene that promotes production of H2S, a potential DR-mimetic candidate, suggesting an approach for further studies linking *CTH* with DR.

The Glutamate-Cysteine Ligase Regulatory subunit (*GCLM*) is a gene that regulates the expression of antioxidant enzymes [56]. Its role in DR is not explicitly stated in literature. However, a recent study [64] showed its increased expression during fasting in *PASK*-deficient mice. Since fasting and intermittent fasting are associated to DR-like benefits, a link between *GCLM* and DR can be investigated from this perspective. If such

a link does not exist, the outcome may remain informative by providing insights on differences between DR- and fasting-related beneficial signalling cascades.

IRS2 is a cytoplasmic signaling molecule that mediates the effects of insulin, insulin-like growth factor 1, and other cytokines. A homolog of this gene is present within GenDR (Gene Manipulations) [14], as "*chico*" in fruit fly, however, it was not detected as an ortholog by the OMA database [13] and thus was not considered a DR-related gene. This gene has a further independent entry within GenDR (Gene Expression) where 174 mice genes that significantly change their expression during DR are reported [14, 65]. Out of our 7 top DR-candidate genes, *IRS2* was the only one overlapping these 174 genes as further explained in Additional file 1: Text S.4. This suggests that *IRS2* may not only be differentially expressed during DR but also could have the potential to regulate DR-associated pro-longevity effects. Sestrin 2 (SESN2) is an intracellular leucine sensor that negatively regulates the TORC1 signaling pathway. This gene was out of the scope of DR-relatedness until a recent work highlighted its role as a novel molecular link that mediates the effects of dietary amino acid restriction on TORC1 activity in stem cells of the fly gut, thereby maintaining gut health and ensuring longevity [66]. Hence, although the DR-probability of this gene was relatively low in the {PathDIP, CAT} model, it was the highest in the {GO terms, BRF} model; and the averaged DR-related prediction of this gene is supported by recent evidence.

Conclusions

To our knowledge, this is one of the pioneering studies applying ML algorithms to DR research in the context of ageing. This work demonstrated the strong potential of ML-based techniques to identify DR-associated features as our findings are consistent with literature and recent discoveries. GO terms and PathDIP pathways were the most predictive types of features. Due to their curated knowledge-driven (literature-based) nature, the use of these feature types in the most predictive models has on one hand mostly corroborated existing knowledge (rather than directly generating new knowledge), but has on the other hand provided statistical support associating DR with the NRF2 pathway and GPDRs, which have been recently accumulating evidence towards DR in the literature, and so are worth further exploring.

Inference of novel DR-related features may be easier to accomplish from feature types not biased by curated knowledge. However, our work found an obstacle to this inference due to the low or even null predictive power of such feature types, implying that either (1) their features did not contain relevant information for predicting DR-relatedness; or (2) the used tree-based ensemble algorithms were not suitable for our classification problem with the unbiased feature types used in this work, especially expression and co-expression data; or (3) the number of currently known ageing-related genes was not large enough for our ML algorithms to find complex patterns in our unbiased data, leading to poor predictive accuracy in such datasets. In future work, the application of deep learning techniques could potentially increase the predictive power of unbiased feature types, which could provide novel insights on possible DR-related protein properties and interactions as well as DR-related gene expression and co-expression signatures.

Further insights were taken from genes annotated as $Ageing_{NotDR}$-related genes in the dataset but strongly predicted as DR-related genes based on GO terms and PathDIP

pathways. This analysis revealed a list of genes outside GenDR that are prone to be related with DR despite lacking such annotation. Most of these genes were consistent with some preliminary DR-related experiments, which makes them worth exploring for further wet-lab experiments to get a deeper understanding of their relationship with DR. Among these genes, *GOT2* was the only $Ageing_{NotDR}$–related gene present within the top six stronger false positives in models learned with both PathDIP and GO term features. Other DR-related gene candidates strongly predicted by both most predictive ML models were *GOT1*, *TSC1*, *CTH*, *GCLM*, *IRS2* and *SENS2*, which, together with *GOT2*, remain to be validated in further lab-based experiments.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04523-8.

Additional file 1: Text S.1.1. KEGG-Pertinence Dataset construction. **Text S.1.2**. Protein descriptors. **Text S.2**. Hyperparameters Grid. **Text S.3**. Analysis of GenDR (Gene expression) genes based on DR-association probability. **Text S.4**. KEGG-Pertinence. **Text S.5**. KEGG-Influence. **Table S1**. Classification of amino acids into three possible functional groups according to their corresponding physicochemical properties. **Table S2**. Ensemble algorithms' hyperparameters grid. Ensemble algorithms' hyperparameters grid for sampling strategy, maximal number of features allowed in each base estimator, number of base estimators, class weight and binary threshold. The definition of a_{us} is as expressed in equation (S.1). The number n in Max features is the same as the total number of features in the dataset of interest. The last column shows the algorithm(s) that use each of the hyperparameters mentioned in the first column. **Table S3**. DR-probability analysis of the subset of ageing-related genes that are differentially expressed during DR. **Table S4**. Top-seven most predictive pathways by the {KEGG-Pertinence, BRF} model. **Table S5**. Top-ten most predictive nodes by the {KEGG-Influence, XGB} model. **Figure S1**. ROC curve and AUC of the ML algorithms on the two most predictive datasets. **Figure S2**. Comparison of the two most predictive models. A and B, confusion matrices of {GO terms, BRF} and {PathDIP, CAT}, respectively. **Figure S3**. Overlapping between ageing-related, -related and -related genes.

Acknowledgements
Not applicable.

Authors' contributions
GDVM, JPM and AAF conceived and planned the experiments. GDVM collected the data. GDVM and VB carried out the ML experiments. JPM contributed to the biological interpretation of the results. AAF and YZ contributed to the computational interpretation of the results. GDVM took the lead in writing the manuscript. AAF, JPM and YZ provided critical feedback and helped shape the research, analysis and manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by a Biotechnology and Biological Sciences Research Council (BB/R014949/1) to J.P.M., a Wellcome Trust Grant (208375/Z/17/Z) to J.P.M., a Leverhulme Trust research Grant (RPG-2016-015) to J.P.M. and A.A.F., a CONACyT sponsorship (2019-000021-01EXTF-00468) to G.D.V.M, and an University of Guadalajara loan (V/2020/449) to G.D.V.M..

Availability of data and materials
The datasets and code generated and/or analysed during the current study are available in the GusDany3691/ML-based-prediction-of-DR-related-genes repository, https://github.com/GusDany3691/ML-based-prediction-of-DR-related-genes. The datasets "Protein Descriptors", "Coexpression" and "WholeDatasets" are only available by request to gusdany@liverpool.ac.uk as they exceed GitHub's maximal single-file' storage capacity.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
J.P.M. is CSO of Centaura, a company that aims to prevent and reverse aging, an advisor for the Longevity Vision Fund, YouthBio and NOVOS as well as the founder of Magellan Science Ltd, a company providing consulting services in longevity science.

Author details
[1]Integrative Genomics of Ageing Group, Institute of Life Course and Medical Sciences, University of Liverpool, 6 West Derby St, Liverpool L7 8TX, UK. [2]School of Computer Technologies and Controls, ITMO University, Kronverkskiy Prospekt

49, 197101 St Petersburg, Russia. [3]Department of Eye and Vision Science, Institute of Life Course and Medical Sciences, University of Liverpool, 6 West Derby St, Liverpool L7 8TX, UK. [4]School of Computing, University of Kent, Canterbury CT2 7NF, UK.

Received: 28 July 2021 Accepted: 8 December 2021
Published online: 04 January 2022

References

1. MacNee W. Is chronic obstructive pulmonary disease an accelerated aging disease? Ann Am Thorac Soc. 2016;13(5):S429–37.
2. de Magalhaes JP, Wuttke D, Wood SH, Plank M, Vora C. Genome-environment interactions that modulate aging: powerful targets for drug discovery. Pharmacol Rev. 2012;64(1):88–101.
3. Lopez-Otın C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell. 2013;153(6):1194–217.
4. Gillespie ZE, Pickering J, Eskiw CH. Better living through chemistry: Caloric restriction (CR) and CR mimetics alter genome function to promote increased health and lifespan. Front Genet. 2016;7:142.
5. Wieser D, Papatheodorou I, Ziehm M, Thornton JM. Computational biology for aging. Philos Trans R Soc Lond B Biol Sci. 2011;366(1561):51–63.
6. Fabris F, de Magalhaes JP, Freitas AA. A review of supervised machine learning applied to ageing research. Biogerontology. 2017;18(5):171–88.
7. Fabris F, Palmer D, Salama KM, de Magalhaes JP, Freitas AA. Using deep learning to associate human genes with age-related diseases. Bioinformatics. 2020;36(7):2202–8.
8. Huang T, Zhang J, Xu Z-P, et al. Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches. Biochimie. 2012;94(4):1017–25.
9. Weidner CI, Lin Q, Koch CM. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol. 2014;15:24.
10. Zhavoronkov A, Mamoshina P, Vanhaelen Q, Scheibye-Knudsen M, Moskalev A, Aliper A. Artificial intelligence for aging and longevity research: recent advances and perspectives. Ageing Res Rev. 2019;49:49–66.
11. Tacutu R, Thornton D, Johnson E, et al. Human ageing genomic resources: new and updated databases, build 20 (09/02/2020). Nucleic Acids Res. 2018;46:1.
12. de Magalhaes JP, Budovsky A, Lehmann G, Costa J, Li Y, Fraifeld V, Church GM. The human ageing genomic resources: online databases and tools for biogerontologists. Aging Cell. 2009;8:65–72.
13. Train CM, Glover NM, Gonnet GH, Altenhoff AM, Dessimoz C. Orthologous matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. Bioinformatics. 2017;33(14):75–82.
14. Wuttke D, Connor R, Vora R, et al. Dissecting the gene network of dietary restriction to identify evolutionarily conserved pathways and new functional genes, build 4 (24/06/2017). PLoS Genet. 2012;8(8):1002834.
15. Rahmati S, Abovsky M, Pastrello C, Jurisica I. pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis. Nucleic Acids Res. 2017;45:419–26.
16. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):47.
17. Fabris F, Freitas AA. New KEGG pathway-based interpretable features for classifying ageing-related mouse proteins. Bioinformatics. 2016;32(19):2988–95.
18. Stark C, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(1):535–9.
19. Csardi G, Nepusz T. The igraph software package for complex network research. InterJ Complex Syst. 2006;1695:56.
20. Jalili A, Salehzadeh-Yazdi A, Asgari Y, Arab SS, Yaghmaie M, Ghavamzadeh A, Alimoghaddam K. Centiserver: a comprehensive resource, web-based application and R package for centrality analysis. PLoS ONE. 2015;10(11):589–98.
21. Jalili M, Salehzadeh-Yazdi A, Gupta S, Wolkenhauer O, Yaghmaie M, Resendis-Antonio O, Alimoghaddam K. Evolution of centrality measurements for the detection of essential proteins in biological networks. Front Physiol. 2016;7:375.
22. Carlson M. GO.db: a set of annotation maps describing the entire gene ontology; 2020. R Package Version 3.11.4
23. Carithers LJ, Ardlie K, Barcus M. A novel approach to high-quality postmortem tissue procurement: the GTEx project. Biopreserv Biobank. 2015;13(5):311–9.
24. Dam SV, Craig T, de Magalhaes JP. Genefriends: a human rna-seq-based gene and transcript co-expression database. Nucleic Acids Res. 2015;43:1124–32.
25. Rainer J. EnsDb.Hsapiens.v86: Ensembl Based Annotation Package; 2017. R package version 2.99.0
26. Rainer J, Gatto L, Weichenberger CX. Ensembldb: An R package to create and use ensembl-based annotation resources. Bioinformatics. 2019;35(17):3151–3.
27. Xiao N, Cao D-S, Zhu M-F, Xu Q-S. protr/protrweb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics. 2015;31(11):1857–9.
28. Osorio D, Rondon-Villarreal P, Torres R. Peptides: a package for data mining of antimicrobial peptides. R J. 2015;7(1):4–14.
29. Fernandez-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15:3133–81.
30. Zhang C, Liu C, Zhang X, Almpanidis G. An up-to-date comparison of state-of-the-art classification algorithms. Expert Syst Appl. 2017;82:128–50.

31. Hamori A, Kawai M, Kume T, Murakami Y, Watanabe C. Ensemble learning or deep learning? Application to default risk analysis. J Risk Financ Manag. 2018;11(1):12.
32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
33. Lemaˆıtre G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18(17):1–5.
34. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16. 785–94; 2016. ACM, New York, NY, USA.
35. Dorogush AV, Ershov V, Yandex AG. Catboost: gradient boosting with categorical features support. arXiv:1706.09516 1 (2018)
36. Ri J, Kim H. G-mean based extreme learning machine for imbalance learning. Digital Signal Process. 2020;98:102637.
37. Menze BH, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinform. 2009;10(213):25.
38. Kowaltowski AJ. Caloric restriction and redox state: does this diet increase or decrease oxidant production? Redox Rep. 2011;16(6):237–41.
39. Lennicke C, Cocheme HM. Redox signalling and ageing: insights from drosophila. Biochem Soc Trans. 2020;48(2):367–77.
40. Kalyani RR, Egan JM. Diabetes and altered glucose metabolism with aging. Endocrinol Metab Clin North Am. 2013;42(2):333–47.
41. Boemi M, Furlan G, Luconi MP. Molecular basis of nutrition and aging: a volume in the molecular nutrition series. Academic Press, UOC Malattie Metaboliche e Diabetologia, INRCA-IRCCS, Ancona, Italy; 2016.
42. Dommerholt MB, Dionne DA, Hutchinson DF, Kruit JK, Johnson JD. Metabolic effects of short-term caloric restriction in mice with reduced insulin gene dosage. Redox Rep. 2018;237(1):59–71.
43. Santos-Otte P, et al. G protein-coupled receptor systems and their role in cellular senescence. Comput Struct Biotechnol J. 2019;8(17):1265–77.
44. Wang XX, et al. A dual agonist of farnesoid x receptor (fxr) and the g protein–coupled receptor tgr5, int-767, reverses age-related kidney disease in mice. Comput Struct Biotechnol J. 2017;292(29):12018–24.
45. Chung KW, Chung HY. The effects of calorie restriction on autophagy: Role on aging intervention. Ageing Res Rev. 2019;11(12):2923.
46. Donati A, Recchia G, Cavallini G, Bergamini E. Relevance of autophagy induction by gastrointestinal hormones: focus on the incretin-based drug target and glucagon. J Gerontol Ser A. 2008;63(6):550–5.
47. Manco M, Mingrone G. Effects of weight loss and calorie restriction on carbohydrate metabolism. Curr Opin Clin Nutr Metab Care. 2005;8(4):431–9.
48. Manchishi SM, Cui RJ, Zou XH, Cheng ZQ, Li BJ. Effect of caloric restriction on depression. J Cell Mol Med. 2018;22(5):2528–35.
49. Budni J, Bellettini-Santos T, Mina F, Garcez ML, Zugno AI. The involvement of BDNF, NGF and GDNF in aging and Alzheimer's disease. J Cell Mol Med. 2015;6(5):331–41.
50. Garcıa-Prieto CF, Fernandez-Alfonso MS. Caloric restriction as a strategy to improve vascular dysfunction in metabolic disorders. Circ Re. 2016;8(6):370.
51. Erickson KI, et al. Brain-derived neurotrophic factor is associated with age-related decline in hippocampal volume. J Cell Mol Med. 2010;30(15):5368–75.
52. Chen Y. Aging-induced akt activation involves in aging-related pathologies and aβ-induced toxicity. Aging Cell. 2019;18(4):12989.
53. Pomatto LCD, et al. Deletion of nrf2 shortens lifespan in c57bl6/j male mice but does not alter the health and survival benefits of caloric restriction. Free Radical Biol Med. 2020;152:650–8.
54. Kultz D. Molecular and evolutionary basis of the cellular stress response. Annu Rev Physiol. 2005;67:225–57.
55. Sharma PK, Mittal N, Deswal S, Roy N. Calorie restriction up-regulates iron and copper transport genes in saccharomyces cerevisiae. Mol Biosyst. 2011;7(2):394–402.
56. Stelzer G et al. The genecards suite: From gene data mining to disease genome sequence analysis. Curr Protocols Bioinform; 2016.
57. Craig T, et al. The digital ageing atlas: integrating the diversity of age-related changes into a unified resource. Nucleic Acids Res. 2015;43:873–8.
58. Song YM, et al. Metformin alleviates hepatosteatosis by restoring sirt1-mediated autophagy induction via an amp-activated protein kinase-independent pathway. Autophagy. 2015;11(1):46–59.
59. Miller RA, Buehner G, Chang Y, Harper JM, Sigler R, Smith-Wheelock M. Methionine-deficient diet extends mouse lifespan, slows immune and lens aging, alters glucose, T4, IGF-I and insulin levels, and increases hepatocyte MIF levels and stress resistance. Aging Cell. 2005;4(3):119–25.
60. Ng LT, Gruber J, Moore PK. Is there a role of H2S in mediating health span benefits of caloric restriction? Biochem Pharmacol. 2018;149:91–100.
61. Madeo F, Carmona-Gutierrez D, Hofer SJ, Kroemer G. Caloric restriction mimetics against age-associated disease: targets, mechanisms, and therapeutic potential. Cell Metab. 2019;29(3):592–610.
62. Harputlugil E, Hine C, Vargas D, Robertson L, Manning BD, Mitchell JR. The tsc complex is required for the benefits of dietary protein restriction on stress resistance in vivo. Cell Rep. 2014;8:1160–70.
63. Derous D, et al. The effects of graded levels of calorie restriction: evaluation of the main hypotheses underpinning the life extension effects of CR using the hepatic transcriptome. Aging (Albany NY). 2017;9(7):1770–804.
64. Lettieri-Barbato D, Minopoli G, Caggiano R, Izzo R, Santillo M, Aquilano K, Faraonio R. Fasting drives nrf2-related antioxidant response in skeletal muscle. Int J Mol Sci. 2020;21:7780.

65. Plank M, Wuttke D, van Dam S, Clarkeab SA, de Magalhaes JP. A meta-analysis of caloric restriction gene expression profiles to infer common signatures and regulatory mechanisms. Mil BioSyst. 2012;9:1339–49.
66. Lu J, Temp U, Muller-Hartmann A, Esser J, Gronke S, Partridge L. Sestrin is a key regulator of stem cell function and lifespan in response to dietary amino acids. Nat Aging. 2021;1:60–72.

Publisher's Note