Research Paper

Machine learning for predicting lifespan-extending chemical compounds

Diogo G. Barardo[1,*], Danielle Newby[2,*], Daniel Thornton[1], Taravat Ghafourian[3], João Pedro de Magalhães[1,#], Alex A. Freitas[4,#]

[1]Integrative Genomics of Ageing Group, Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, UK
[2]Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, UK
[3]School of Life Sciences, University of Sussex, Falmer, Brighton, UK
[4]School of Computing, University of Kent, Canterbury, UK
[*]Equal contribution
[#] Joint last authors

Correspondence to: João Pedro de Magalhães, Alex A. Freitas; **email:** jp@senescence.info, A.A.Freitas@kent.ac.uk
Keywords: longevity, anti-ageing drugs, pharmaceutical interventions, ageing, bioinformatics, *C. elegans*, machine learning.
Received: June 6, 2017 **Accepted:** July 12, 2017 **Published:** July 18, 2017

ABSTRACT

Increasing age is a risk factor for many diseases; therefore developing pharmacological interventions that slow down ageing and consequently postpone the onset of many age-related diseases is highly desirable. In this work we analyse data from the DrugAge database, which contains chemical compounds and their effect on the lifespan of model organisms. Predictive models were built using the machine learning method random forests to predict whether or not a chemical compound will increase *Caenorhabditis elegans*' lifespan, using as features Gene Ontology (GO) terms annotated for proteins targeted by the compounds and chemical descriptors calculated from each compound's chemical structure. The model with the best predictive accuracy used both biological and chemical features, achieving a prediction accuracy of 80%. The top 20 most important GO terms include those related to mitochondrial processes, to enzymatic and immunological processes, and terms related to metabolic and transport processes. We applied our best model to predict compounds which are more likely to increase *C. elegans*' lifespan in the DGIdb database, where the effect of the compounds on an organism's lifespan is unknown. The top hit compounds can be broadly divided into four groups: compounds affecting mitochondria, compounds for cancer treatment, anti-inflammatories, and compounds for gonadotropin-releasing hormone therapies.

INTRODUCTION

Old age is the greatest risk factor for many diseases, including various types of cancer, inflammatory and neurodegenerative diseases. Traditional medical science combats one disease at a time, instead of combating the underlying biological ageing process that leads to many age-related diseases. From a whole body system's point of view, this traditional one-disease-at-a-time approach focuses on the downstream diseases, rather than considering the underlying mechanisms of age-related functional decline. This approach has limited effectiveness at present and is likely to be less effective in the future, because of an increasingly larger elderly population suffering from multiple age-related diseases. In contrast, interventions that slow down ageing and promote "healthy ageing" could in principle delay the onset of all age-related diseases, with a significant benefit to human health and a large reduction of healthcare costs [1].

Pharmacological interventions are arguably the most practical ageing intervention for humans, avoiding the main problems with genetic interventions (generally unethical in humans) and dietary interventions such as caloric restriction, which are difficult to maintain for the vast majority of people. For instance, there is currently great interest in discovering drugs that mimic the process of caloric restriction (caloric restriction mimetics) [2,3]. In addition, promising research on pharmacological interventions on the ageing process is underway at the National Institute of Aging's Intervention Testing Program (ITP), which consists of administering drugs or chemical compounds to mice under carefully controlled conditions [4,5]. However, as mouse experiments are costly and time consuming, so far only a limited number of drugs or compounds have been evaluated. Thus, using simpler model organisms for evaluating a chemical compound's effect on an organism's lifespan is appealing, and a substantially larger number of studies have administered compounds to *C. elegans* than other organisms. As the ITP for mice, the *Caenorhabditis* Intervention Testing Program has been introduced for assessing longevity variation for chemical compounds [6]. Although *C. elegans* is physiologically different from humans, *C. elegans* is the most studied model organism in ageing research, producing insights that are applicable to other organisms [7], since cellular-level ageing processes are often conserved across distantly-related species [8]. According to the GenAge database [9], *C. elegans* is the animal model with by far the most known ageing-related genes (838 at the time of writing).

In this work we analyse data from the DrugAge database [10], which contains information about chemical compounds and their effect on the lifespan of organisms. DrugAge contains a variety of compounds with anti-ageing properties such as gerosuppressant, geroprotective and senolytic activity [11–13] as well lifespan increasing properties for a specific species. Existing databases with lifespan-extending drugs include AgeFactDB (http://agefactdb.jenage.de/) [14], and Geroprotectors.org [15] (http://geroprotectors.org/). DrugAge incorporates data from these resources and improves on them by providing a more extensive and systematic repertoire of lifespan-extending drugs, compounds and substances. DrugAge is manually curated and features only information relative to lifespan assays conducted in well-controlled studies. DrugAge contains data about several model organisms, and the majority of compounds in DrugAge have been evaluated on *C. elegans*, so we focus on analysing data for this organism.

In order to analyse such data, we use random forests, which is a supervised machine learning method – for a recent review of supervised machine learning applied to the biology of ageing, see [16]. In this work, the random forest builds a classification model to predict whether or not a chemical compound will increase the lifespan of *C. elegans*, based on predictive features describing that compound. We created datasets with two types of predictive features, namely Gene Ontology (GO) terms annotated for proteins interacting with the compounds and chemical descriptors calculated from each compound's chemical structure. In order to evaluate the predictive relevance of these two types of features, we created three different datasets: one using as predictive features only the GO terms, another using as predictive features only the chemical descriptors, and a third dataset using both types of features. In addition, the best model produced by the random the forest method was applied to a screening "external" dataset with compounds from the DGIdb database, where the effect of the compounds on an organism's lifespan is unknown. The predictions of that model were used to identify the "top hit" compounds in the DGIdb dataset, i.e. compounds with higher probabilities of increasing lifespan in *C. elegans*.

There are some related works that performed data analysis on compounds increasing *C. elegans*' lifespan, but without using any predictive machine learning method. In particular, Ziehm et al. used an empirical scoring function combining several different factors to evaluate the relevance of a compound for ageing [17]; and Ye et al. (2014) constructed a pharmacological network in order to reveal pharmacological classes most related to *C. elegans*' ageing [18]. In addition, Calvert et al. 2016 identified drugs which induce gene expression profiles similar to the profiles of genes associated with caloric restriction (CR), and observed that various genes targeted by lifespan-extending drugs are included in CR and longevity networks [3]. Furthermore, Aliper et al. [19] utilised computational tools to carry out signalling pathway analysis of gene expression between young and old stem cells in humans. Based on the signalling pathway results, known compounds were screened and ranked, in order to identify the best compounds to target those pathways and restore a "young" cellular profile. A review of several specific pharmacological classes extending *C. elegans*' lifespan can be found in Carretero et al. 2015 [20], but again with no use of predictive machine learning methods.

To the best of our knowledge, this is the first work to propose the use of a predictive machine learning method (namely Random Forests) to analyse data about the effect of chemical compounds in *C. elegans*' lifespan, as well as the first work to apply machine learning to data about compounds in the DrugAge database.

RESULTS AND DISCUSSION

Predictive accuracy of the models

We have created a DrugAge dataset specifically for studying the classification of compounds into the classes "increase lifespan" or "do not increase lifespan", depending on each compound's effect when administered to *C. elegans*. In this dataset, each compound to be classified belongs to one of the two just-mentioned classes, and is described by a large set of chemical descriptors and biological GO term features.

We use the random forest method as the classification algorithm to analyse this dataset. This type of method was chosen because it is particularly popular in bioinformatics [21,22], it is robust to overfitting in datasets where the number of features is much larger than the number of instances (as with our dataset) [22,23], it is relatively simple to understand and to use, and finally, in contrast to other state-of-the-art classification methods like support vector machines, random forests produce interpretable results based on a variable (feature) importance measure, an interpretation mechanism also exploited in this paper.

Predictive accuracy for the models developed was evaluated by Area Under the ROC curve (AUC). This is a measure between 0 and 1, with 1 indicating perfect (no error) class predictions. The reported predictive accuracy used is the median over the 10 test sets of the external cross-validation. We report the median accuracy, rather than the mean, because the former is more robust to outliers. The median AUC results from each of the different versions of the DrugAge dataset (using either chemical and/or biological descriptors), where for each dataset version we optimised the parameters ntrees and mtry of the random forest method as described in the Methods section.

The AUC results are reported in Table 1. Comparing the AUC values across the dataset versions (last column in Table 1), it is clear that, in general, the set of chemical descriptors have a greater ability to predict a compound's class than the set of GO terms. More precisely, the dataset using only chemical descriptors as features has substantially larger AUC than the one using only GO terms as features (0.781 vs. 0.716, respectively). However, the GO term features still offer some positive contribution to the predictive accuracy of random forests, since the dataset version leading to the highest AUC value in Table 1 (0.800) was the one using both GO terms and chemical descriptors as features.

Biological and chemical features for the prediction of longevity compounds in *C. elegans*

One of the benefits of utilizing the random forest method, as well as it being a highly predictive technique, is that for each feature an importance measure can be calculated. This importance measure (often called variable importance) offers the opportunity to interpret the relevance of each feature in the model produced. In this work, using the Boruta and Ranger R packages [21,24] and computing the importance of features in the best model (built using both GO terms and chemical descriptors as features), 93 features – 73 chemical descriptors and 20 GO terms – were selected as statistically significant features (full table Supplemental Data). Recall that the GO term features are derived from the proteins which are targeted by each compound.

The 20 GO terms selected as significant mainly make up biological process GO terms (14 out of 20), five molecular function terms and one defining a cellular component term. Biological process GO terms describe a series of processes as well as specific biological processes such as macromitophagy and macroautophagy, which are among the features with the highest importance

Table 1. Predictive accuracy (median AUC values on 10-fold cross validation) obtained by random forest with parameters optimized for each DrugAge dataset version (each with a different feature type combination).

Dataset features	RF's optimized parameters		Median AUC
	ntrees	mtry	
GO terms only	300	52	0.716
Chemical descriptors only	100	16	0.781
GO terms and chemical descriptors	900	210	0.800

in this work. Molecular function GO terms describe specific activities that occur at the molecular level such as isomerase activity and protein disulfide isomerase activity. Finally, cellular component GO terms describe locations in the cell, e.g. at the level of organelles or macromolecular complexes such as the mitochondrial proton-transporting ATP synthase complex, highlighted as the only significant cellular component GO term feature in this work.

Chemical molecular descriptors are calculated from the chemical structure and are normally used to build predictive models to study the relationship between a compound's chemical structure and its biological and pharmacokinetic properties such as drug distribution and absorption [25,26]. This paper is the first use of chemical molecular descriptors (as well as GO terms) to study the relationship between longevity and the chemical structure of compounds that may affect longevity.

Chemical molecular descriptors can be broadly categorized into three main groups, which describe a compound's chemical structure and its main properties. These groups are: hydrophobic, electronic and steric (size and/or shape) descriptors. Hydrophobicity descriptors describe the hydrophobic character of a chemical compound and how easily it can cross cell membranes, and they may also be important for receptor interactions. Electronic molecular descriptors describe the electron distribution in a chemical compound and its electrostatic interactions, therefore they give an indication of how strongly (in terms of affinity) and how specifically a chemical compound binds to specific receptors. Finally, steric descriptors describe the size and shape of the chemical compound. The size and shape of a compound may influence its binding with an enzyme or receptor binding sites and can also affect other psychochemical properties. Note that a chemical molecular descriptor can belong to more than one of the categories described above.

The top 20 selected features with the highest median variable importance are shown in Table 2. Considering just the top 20 features as shown in Table 2, there are slightly more GO terms (12 out of 20) than chemical molecular descriptors (8 out of 20). Those 12 GO terms include terms related to mitochondrial processes, terms related to enzymatic and immunological processes and terms related to metabolic and transport processes. Furthermore, the eight chemical molecular descriptors in the top 20 features contain descriptors related to electronic and steric (size and shape) effects, but not to hydrophobic effects directly.

It can be seen from the list of important features that the vast majority of the most important features are very specific molecular and biological processes. However, these specific processes are generic in their applicability and occur across many tissues and organs. For example "isomerase activity" covers a broad range of various enzymes that catalyze reactions across many biological processes, such as in glycolysis and carbohydrate metabolism. Although it is evident that isomerase activity is relevant to metabolism (amongst other processes) and hence ageing, this feature is not specific enough to suggest practical targets for pharmacological intervention. In spite of this, some of the specific features have been linked with longevity and ageing processes.

GO terms related to metabolism encompass the vast majority of the GO term features listed in Table 2. These GO terms range from very general metabolism-related properties such as aerobic respiration to more specific processes such as dipeptidase activity, pyruvate metabolic process, fatty acid transport and mitochondrial electron transport from NADH to ubiquinone. Given the involvement of metabolic factors in several theories of ageing such as the free radical theory of ageing, as well as the well-established effect of calorie-restriction on longevity, it is expectable that the compounds that affect ageing do so by interacting with these pathways and processes, as evidenced also by the importance of such features in the random forest model.

One apparent group of features that can be related to longevity and ageing are the GO terms related to autophagy (macroautophagy and macromitophagy) and mitochondrial processes. Macroautophagy is the process where cellular contents are degraded by lysosomes or vacuoles and recycled, and this process controls cytosolic protein and organelle degradation [27,28]. Whereas macromitophagy is the degradation of mitochondrion by macroautophagy and controls mitochondrial quality and quantity [29]. It is known that autophagy in general is associated with ageing processes. This can be evidenced by the occurrence of degenerative changes in mammalian tissues, similar to changes seen with ageing, as a result of genetic inhibition of autophagy. Moreover, pharmacological or genetic manipulations that increase life span in model organisms often stimulate autophagy. In the same way, there is a decrease in autophagy with increasing age in organisms, which leads to accumulation of damage [30] which is thought to be responsible for the functional loss in many biological and physiological processes as ageing occurs [31,32]. In addition to macroautophagy, mitophagy is specifically implicated in ageing. Mitophagy has been shown to be a selective, "non-random" process [33] that is governed by several biological pathways (see [34] for a review of the molecular mechanisms).

Table 2. Top 20 selected features with the highest median variable importance.

Median Variable Importance	Feature	Feature type	Feature Description
14.4	a_nN	MD	Number of nitrogen atoms in the molecule
12.8	isomerase activity	GO	Catalysis of the geometric or structural changes within one molecule
11.8	macromitophagy	GO	Degradation of a mitochondrion by macroautophagy
11.6	macroautophagy	GO	Process in which cellular contents are degraded by lysosomes
11.1	protein disulfide isomerase activity	GO	Catalysis of the rearrangement of both intrachain and interchain disulfide bonds in proteins.
11.0	dipeptidase activity	GO	Catalysis of the hydrolysis of a dipeptide.
9.72	pyruvate metabolic process	GO	The chemical reactions and pathways involving pyruvate
9.47	PEOE_VSA+4	MD	Total positive van der waals surface area of atoms with atomic charge in the range of 0.20-0.25.
9.31	fatty acid transport	GO	The directed movement of fatty acids into, out of or within a cell, or between cells
8.79	mitochondrial electron transport, NADH to ubiquinone	GO	The transfer of electrons from NADH to ubiquinone mediated by the multisubunit enzyme known as complex I
8.64	vsurf_Wp2	MD	Polar volume at -0.5, a descriptor reflecting the polarizability of a molecule
8.57	isotype switching	GO	The switching of activated B cells from IgM biosynthesis to biosynthesis of other isotypes
8.40	translation	GO	The cellular metabolic process in which a protein is formed
8.18	Q_RPC-	MD	Relative negative partial charge, defined as the most negative atomic charge divided by the sum of all negative atomic charges in the molecule.
8.09	aerobic respiration	GO	The enzymatic release of energy from inorganic and organic compounds
7.98	a_IC	MD	Atom information content (total), defined as the entropy of the element distribution in the molecule multiplied by the number of atoms.
7.95	PEOE_VSA_FPPOS	MD	Fractional polar positive vdw surface area
7.86	triglyceride mobilization	GO	The release of triglycerides from storage within cells or tissues, making them available for metabolism.
7.79	chi1v	MD	Valence corrected molecular connectivity index (order 1)
7.70	bpol	MD	Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule

GO: Gene ontology term; MD: Chemical Molecular descriptor

Mitochondrial respiration, and in particular electron transport chain, is the main source of reactive oxygen species. As a result, mitochondrial homeostasis is particularly affected by ageing, as ROS generation in mitochondria leads to mitochondrial protein and mtDNA damage [34]. Therefore, mitophagy can be regarded as a defense against oxidative stress, mitochondrial dysfunction, and ageing. This is supported by findings that along with mitochondrial biogenesis pathways, a key mediator of mitophagy and longevity assurance under conditions of stress in *C. elegans* (DCT-1) is upregulated when mitophagy is

impaired [35]. It is therefore not unexpected to find in this work that chemical compounds that modulated mitophagy are also important promoters of longevity. It is interesting to note that in model organisms such as *C. elegans* disruption of mitochondrial electron transport chain processes can lead to increases in longevity, through genetic [36] or pharmacological interventions [37]. Finally, a related property, aerobic respiration, was also selected by the random forest model. Although aerobic respiration is a very broad term encompassing many processes that lead to the production of cellular energy, it is very well-associated with ageing through the known impact of mitochondrial function and caloric restriction.

Other GO features with links to longevity and ageing processes are protein disulfide isomerase activity and translation. Protein disulfide isomerase activity refers to the activity of isomerases that are involved in protein folding via formation and breakage of disulfide bonds within proteins in the endoplasmic reticulum (ER) [38,39]. The activity of this enzyme is key to protein folding and quality control in the ER. A number of studies have demonstrated that the levels of disulfide isomerase and their catalytic activity diminish with age [40]. Misfolding of proteins and ER stress are alleviated by the signalling pathway known as the ER stress response or the unfolded protein response, which involves protective measures to limit the protein load. These include up-regulation of ER chaperones involved in the refolding of proteins, activation of pathways leading to reduction of protein translation and degradation of misfolded proteins. Where ER stress cannot be reversed, cellular functions deteriorate and apoptosis will occur [41]. There is evidence in the literature to suggest that disruption of protein disulfide isomerase activity leads to ER stress and accumulation of misfolded proteins, which can give rise to age-related disease pathology [42]. Finally, the GO term translation has a clear biological relevance, since it is well-known that translation inhibition extends lifespan in *C. elegans* [43]. Translation has also been highlighted as a prime category in age-related genes in *C. elegans* in a recent paper by Fernandes et al. (2016) [44]. It is therefore evident that pathways involved in protein translation and folding may be a target of anti-ageing compounds, hence the significance of GO terms such as "translation" and "disulphide isomerase" in the random forest model.

The molecular descriptors in Table 2 indicate the molecular properties that impact the longevity effect of the compounds. From the eight molecular descriptors listed in the table, the majority are electrostatic descriptors such as PEOE_VSA+4, vsurf_Wp2, Q_RPC-, PEOE_VSA_FPPOS and bpol. These electrostatic parameters also carry information regarding the topology of the molecule, and along with steric parameters such as chi1v and a_IC explain the interaction and binding of the compounds with their target sites. These targets/processes are in addition to those already described in the model by the biological features (GO terms).

Overall, even though the used dataset (like any other biological dataset) is somewhat biased by the fact that some genes have been much more studied than others [44], some of the most important features shown in Table 2 can be related to important and known biological processes of ageing and longevity, such as those related to autophagy and mitochondrial processes. Furthermore, the other selected biological and chemical features are a good starting point that warrants further investigation, to further link the chemical and biological features of chemical compounds with longevity and underlying biological ageing processes.

Predictions of novel potential life-extending compounds

The best model built from the DrugAge dataset (using GO terms and chemical descriptors) was used to predict the probability of the class "increase lifespan" for over 6,000 compounds from the DGIdb database v2 [45], where the class label of each compound is unknown. By using the predicted class probabilities we can rank and prioritise those compounds with the highest probability of increasing the lifespan of *C. elegans*. The list of all compounds predicted from the DGIdb dataset and their associated class probabilities can be found in the Supplemental Data, and the class probabilities for the top 20 compounds can be found in Table 3.

As shown in Table 3 the highest predicted class probability for a compound in DGIdb was 0.69. Although not close to 1, this can be considered a relatively high probability, considering that the baseline probability (relative frequency) of the class "lifespan increase" in the DrugAge dataset used to build the model was only 0.20. In this section, we focus on the 50 "top hit" DGIdb compounds, with the highest values of probabilities for the predicted class "lifespan increase". In general, the top hit compounds predicted to have longevity enhancing effects fall into four groups: compounds affecting mitochondria, compounds used in treatments for cancer, anti-inflammatories, and compounds used in gonadotropin-releasing hormone therapies.

Compounds related to mitochondrial processes

Acrolein (lifespan increase class probability = 0.69) was the top hit in our screening dataset. Acrolein is a highly

reactive electrophile and a building block to many other chemical compounds, including the amino acid methionine. This compound has been shown to be an electron transport chain inhibitor, leading to mitochondrial dysfunction [46]. Acrolein is implicated in pathways such as p53 and the NF-κB inflammation pathway [47]. Acrolein is toxic at high concentrations [46], but at lower doses *in vitro* exposure to acrolein inhibits NF-κB activation, suggesting that inhibition of NF-κB gives rise to acrolein's anti-inflammatory properties – however, the evidence is conflicting [48,49]. Therefore, the high probability of lifespan increase predicted by our model, despite the known toxicity of acrolein, may result from the contribution of a large diversity of the pathways affected by this compound, some of which are desirable for longevity.

Table 3. Top 20 chemical compounds with the highest lifespan-increase class probability from the external screening dataset.

Chemical Compound Name	Predicted Probability
acrolein	0.691
valspodar	0.683
ganirelix	0.674
acetaldehyde	0.669
mmk-1	0.667
rdp-58	0.665
cetrorelix	0.657
gal-b5	0.656
m40	0.654
DB03393	0.650
bortezomib	0.650
ro 25-1392	0.650
gv1001	0.650
lactose	0.650
ergotamine	0.650
cardiolipin	0.642
dactinomycin	0.642
abt-510	0.640
aplyronine a	0.637
valinomycin	0.637

Other compounds affecting mitochondrial processes include valinomycin and cardiolipin (both with lifespan increase class probability = 0.64). Valinomycin is a potassium ionophore and causes mitochondrial dysfunction by uncoupling oxidative phosphorylation in the electron transport chain [50]. Cardiolipin is a dimeric phospholipid found in the inner mitochondrial membrane (IMM), where it plays a major role in oxidative phosphorylation. Alterations in the content and composition, and peroxidation of cardiolipin leads to mitochondrial dysfunction [51,52]. Decrease in cardiolipin content has been observed in ageing brain, and in several pathologies including myocardial ischemia, heart failure and Parkinson's disease [53]. Therefore, it is expectable that cardiolipin administration is predicted to promote longevity.

Anti-cancer drugs and longevity

Anti-cancer compounds from our top 50 hits in the DGIdb dataset include drugs such as temsirolimus, valspodar and bortezomib. Interestingly, temsirolimus (lifespan increase class probability = 0.62) is a derivative and pro-drug of sirolimus – also known as rapamycin. Rapamycin was the first pharmacological compound shown to extend lifespan in both genders in mice models [54,55], *C. elegans* [56] and *D. melanogaster* [57]. Numerous studies indicate that inhibition of the TOR (Target of Rapamycin) kinase is implicated in lifespan control [58,59]. Temsirolimus also inhibits mTOR, and this compound has been shown to improve certain cellular phenotypes in accelerated ageing models via increasing autophagy [60].

Valspodar (lifespan increase probability = 0.68), the second top-hit in our screening dataset, is an experimental chemosensitizer drug. Valspodar desensitizes tumor cells making them more vulnerable to anti-cancer drugs, due to its ability to inhibit P-glycoprotein (P-gp), which is overexpressed in many cancer cells. However, possibly of more relevance is the apoptotic effect of valspodar (and its structurally related compound, cyclosporine A) that stems from their disruption of mitochondrial membrane potential leading to mitochondrial dysfunction [61].

Bortezomib (lifespan increase probability = 0.65) is a proteasome inhibitor, and studies have shown that the inhibition of proteasome activity by bortezomib is associated with enhanced apoptosis due to inhibition of NF-κB activity [62,63]. However, this compound also leads to the accumulation of misfolded proteins and ER stress followed by unfolded protein response (UPR) and macroautophagy [64], which may potentially lead to longevity promotion.

Dactinomycin (lifespan increase probability = 0.64) interferes with ribosome biogenesis through the inhibition of RNA polymerase I [65], which leads to the activation of p53 [66]. Inhibition of the mTOR pathway leads to a reduction of ribosome biogenesis and increases lifespan in several species [54,57,67]. mTOR and p53 signalling pathways are connected by a number of different mechanisms, highlighting a complex relationship [66,68,69]. Considering that there are similar signaling molecules involved in both cancer and

ageing [70,71], such as mTOR [72], p53 [69] and NF-kB [73], it is not unexpected to find anti-cancer drugs in our list of top hit compounds. However, this could be due to research bias, where anti-cancer drugs may be overrepresented in datasets (including DrugAge) due to the extensive study of cancer therapies.

Chemical compounds with anti-inflammageing effects

Ageing has been characterized by chronic, low-grade inflammation, also labeled as "inflammageing" [74]. Human studies have shown that suppression of chronic inflammation is a major determinant of successful longevity, over a very wide age range up to extreme old age [75,76].

The compound rdp-58 (lifespan increase class probability = 0.67), tested for the treatment of the inflammatory disorder ulcerative colitis [77,78], leads to a reduction of proinflammatory (tumor necrosis factor alpha) TNF-α and interleukins (ILs) such as interferon-γ, IL-2, IL-6, and IL-12 [79].

Ergotamine (lifespan increase probability = 0.65), a vasoconstrictor used for the treatment of migraines, has also been shown to reduce the level of proinflammatory TNF-α [80]. Dihydroergotamine methanesulfonate increases longevity in *C.elegans* [18] and was used to build our models. Dihydroergotamine methanesulfonate is a derivative of ergotamine, so this can explain the predicted pro-longevity effects for ergotamine.

The compound ro 25-1392 (lifespan increase probability = 0.65) is a type II vasoactive intestinal peptide receptor (VIPR2) agonist [81]. ro 25-1392 is an analogue of vasoactive intestinal peptide (VIP), which binds to both VIPR1 and VIPR2, leading to protection in models of inflammatory and autoimmune conditions [82,83].

Reproductive hormone factors and longevity

Gonadotropin-releasing hormone (GnRH) is responsible for the release of follicle-stimulating hormone (FSH) and luteinizing hormone (LH) in the pituitary gland, promoting the production of testosterone and estrogen. It is a part of the hypothalamic–pituitary–gonadal axis, which helps in the regulation of reproductive and immune systems [84].

In our list of top hit compounds there are examples of GnRH antagonists, such as ganirelix [85] and cetrorelix [86] (lifespan increase class probabilities 0.67 and 0.66, respectively); and agonists such as nafarelin [87] and histrelin [88,89] (lifespan increase class probabilities 0.63 and 0.62, respectively). Both antagonists and agonists (whose continued use leads to desensitisation of GnRH receptors) of GnRH receptors lead to the reduction of FSH and LH.

The decline in GnRH has been shown to contribute to ageing-related changes such as bone fragility and reduced neurogenesis in mice. Zhang [90] showed in mice that activation of NF-κB in the hypothalamus led to a reduced production of GnRH by neurons and that continued activation led to accelerated ageing, whereas GnRH treatment reduced neurogenesis and decelerated ageing. These findings suggest a link between inflammation and ageing related to GnRH. However, whether this relationship involving GnRH applies to humans and primates is questionable, as it appears that female primates have higher levels of GnRH with increasing age [91], whereas in Norway rats GnRH levels decreased with increasing age [92]. It is therefore apparent that GnRH has some role in longevity independent of its role in reproduction.

CONCLUSIONS

In this work we analysed data from the DrugAge database [10], which contains information about chemical compounds and their effect on the lifespan of organisms. We focused on compounds administered to *C. elegans*, since the majority of compounds in DrugAge have been evaluated in this model organism. For our data analysis, we used the machine learning method random forests, which builds a classification model to predict whether or not a chemical compound will increase the lifespan of *C. elegans*, based on predictive features describing that compound. We built three types of classification models, using either chemical descriptors or Gene Ontology terms, or both types of features. The dataset with both types of features led to the highest predictive accuracy in our experiments.

We used a score calculated by the random forest method to identify the most relevant features. Among the 20 highest score features, there are several GO terms which have a well-established association with the ageing process such as "macromitophagy" and "macroautophagy". The high score of these GO terms is consistent with the fact that pharmacological or genetic interventions that increase lifespan in model organisms often stimulate autophagy [44]. Another example of a relevant GO term in the top 20 features was "translation". It is well-known that translation inhibition extends lifespan in *C. elegans* [43]. The interpretation of the chemical features in the top 20 features is more difficult, since they refer to low-level chemical properties rather than broader biological processes – in general, those chemical features refer to electronic, size and shape effects of the compounds.

Furthermore, we applied the best classification model built by the random forest to a screening "external"

dataset with compounds from the DGIdb database, where the effect of the compounds on an organism's lifespan is unknown. The predictions of that model were used to identify the "top hit" compounds in the DGIdb dataset, i.e. compounds with higher probabilities of increasing lifespan in *C. elegans*. We observed that these top hit compounds can be broadly divided into four groups: compounds affecting mitochondria, compounds for cancer treatment, anti-inflammatories, and compounds for gonadotropin-releasing hormone therapies.

In conclusion we have built, using machine learning, a model to predict the longevity effects of chemical compounds in *C.elegans*, using the recently published DrugAge dataset. The list of top-hit compounds and their analysis contributes to our knowledge of likely longevity-extending compounds, and experimental confirmation of these predictions would be an interesting direction for future research.

METHODS

Dataset creation

Chemical compounds that increased longevity in *C. elegans* were extracted from the DrugAge database (Build 2, release date: 01/09/2016) [10], available from the Human Ageing Genomic Resources website [9]. These compounds were assigned a positive class label (i.e. increased lifespan). Additionally, compounds that were found not to increase or had no effect on longevity in *C. elegans* were collected from the literature and were assigned a negative class label. The sets of positive and negative labelled compounds were combined to form the dataset for modelling. For ease, hereafter reference to the DrugAge dataset for modelling describes the positive entries from DrugAge plus the negative class label compounds. The number of positive and negative entries obtained were 229 and 1163 respectively, after dataset curation. The list of negative entries is present in the Supplemental Data. Compound entries from the DGIdb database v2 [45] were used to test and prioritise chemical compounds for longevity effects from the classification models built from the DrugAge dataset. The DGIdb dataset is used as our independent screening (or "external") dataset, where the compounds' longevity class labels are unknown.

Calculation of chemical molecular descriptors for the datasets used

For calculation of chemical molecular descriptors for chemical compounds, SMILES (Simplified Molecular-Input Line-Entry System) codes, which are line notations encoding the chemical structure, were extracted using PubChem [93] or ChemSpider (http://www.chemspider.com). For compounds where the chemical structure was not available, the structure was drawn directly from the literature reference (if available) and the SMILES code extracted. Compounds were removed at this stage if there was no SMILES code available, contained heavy inorganic metals, were duplicate compounds or were compound mixtures. Additionally, if there were chemical isomers with the same class label, only one entry was kept. For the DGIdb dataset, compounds were removed if they had a molecular weight greater than 3000 Daltons due to software restraints.

For molecular descriptor calculation, MOE (Chemical Computing Group Inc.) v2013.0.8 and Advanced Chemistry Development ACD Labs/LogD Suite v12 were used to calculate 2D and 3D molecular descriptors using the desalted and minimised chemical structure (using the semi-empirical method MOPAC PM6) of each compound. For the DrugAge dataset, molecular descriptors were removed if they had greater than 98% constant values, resulting in a final number of 268 molecular descriptors. The same 268 molecular descriptors were also calculated for the compounds in the DGIdb dataset. Due to software limitations, some molecular descriptors could not be calculated for some compounds. Using the "missForest" R package v1.4 [94], missing values were imputed for chemical molecular descriptors where this occurred, in both the DrugAge and the DGIdb datasets. A total of 1392 compound entries had molecular descriptors calculated for the DrugAge dataset (229 positive and 1163 negative entries). For the DGIdb dataset, 6802 entries had molecular descriptors calculated.

Computation of biological descriptors for the datasets used

Biological descriptors for each compound in each of the datasets were obtained by extracting drug-gene interactions using the DGIdb v2 database [45] and drug-protein interactions using the STITCH v4 database [95]. For drug-protein interactions using STITCH, only the top 100 interactions with a confidence score greater than 0.450 (considered a 'medium confidence strength' in STITCH) were used. The drug–gene/protein interactions obtained were annotated using GO terms (biological process, molecular function and cellular component terms) using the ClueGO plugin [96] in Cytoscape v3.3.0 [97]. For ClueGO, the parameters selected were "GO term fusion" and the entire "GO tree interval" using a background of *Homo sapiens* as the reference set. *Homo sapiens* annotations were used rather than *C. elegans* due to the poor representation of GO terms for this model organism. There were 10757 GO terms that were created as categorical biological

features for the datasets. For each GO term, for each compound a categorical "yes" or "no" feature value was provided for each compound, indicating whether or not, respectively, the protein interacting with that compound was annotated with that GO term.

For this work, classification models were built using datasets with different combinations of chemical and biological descriptors (features) from the original DrugAge dataset. The different datasets used were: Firstly, a dataset using only biological descriptors (GO terms) as features. Secondly, a dataset using only chemical descriptors as features. Thirdly, a dataset using both biological and chemical descriptors as features. A summary of compound numbers for each of the different versions of the DrugAge dataset and the DGIdb dataset can be found in Table 4. Datasets DrugAge_1 and DrugAge_3 have fewer compounds than dataset DrugAge2 because they use GO terms as features, and compounds were discarded because their interacting proteins had no GO term annotation.

Random forests

In this work we used a random forest algorithm [98]. For our classification task, a random forest algorithm builds a classification model consisting of a set of decision trees, where each tree predicts a class label for each new compound. The predictions from all of the trees are then counted, and the class label assigned to a new compound is the label (positive or negative) with the highest number of votes from all of the decision trees in the forest.

Random forest training, including parameter optimization (explained in more detail below), was performed using the "mlr" R package (developer version 2.9) [99], which is a general machine learning interface that works as a wrapper for a plethora of learning algorithms available in distinct R packages. We have trained the random forests that the mlr package imported from the "ranger" R package [21].

After building a random forest model, a measure of variable importance can be computed in order to identify the most relevant input variables (features) for predicting the class variable. We used a permutation-based method for measuring variable importance. In order to evaluate the predictive power of a feature, for each tree in the forest, this method computes the predictive accuracy of that tree using two versions of the data: with random permutation of the values of the variable being evaluated, and without random permutation (i.e., using the original data). These differences of predictive accuracies are averaged over all trees in the random forest to give the feature's permutation-based importance value. In this work the variable importance values were computed using the Boruta R package [21,24] with the unscaled, unconditional permutation-based variable importance measure [100], performing the analysis on 100 permutation-based random forests (varying the random seed used to generate the permutations).

Measuring predictive accuracy

We use the Area Under the Receiver Operating Characteristic Curve (AUC) as the predictive accuracy measure in our experiments. This is a popular measure of predictive accuracy in both machine learning and bioinformatics, and copes well with imbalanced class distributions (such as in our datasets). The AUC value varies from 0 to 1, with 1 indicating a perfect classifier, which would correctly classify every instance; 0.5 indicating a classifier that randomly guesses the class (positive or negative label) for each instance, and 0 indicating the worst possible classifier, which would systematically misclassify every instance.

Table 4. Compound numbers for the DGIdb dataset and different combinations of DrugAge datasets using different combinations of chemical and biological descriptors used in this work.

Dataset	n Positive	n Negative	n Total	Type of features used
DrugAge_1	190	783	973	GO terms ONLY
DrugAge_2	229	1163	1392	Chemical Descriptor ONLY
DrugAge_3	190	783	973	GO terms + Chemical Descriptors
DGIdb	-	-	6802	GO terms + Chemical Descriptors

Notation used in the table: n – number of compounds; Positive – increases longevity; Negative – no effect or decrease in longevity; Biological descriptors – GO terms (all three types); Chemical descriptors – molecular descriptors calculated from the chemical structure of compound entries using cheminformatics software.

Nested cross-validation and random forest parameter optimization

To measure the predictive accuracy of the models developed, we used a nested cross-validation procedure. First, the DrugAge dataset was randomly divided into 10 non-overlapping folds with approximately the same number of compounds in each fold. The external cross-validation procedure performs 10 iterations of the classification algorithm (random forest), each time using one of the folds as the test set and the other 9 folds as the training set. In each of these 10 external cross-validation iterations, an internal 10-fold cross-validation procedure was applied to the training set. That is, the training set was randomly partitioned into 10 folds of approximately the same size, and 10 iterations were performed, using one of the training folds as a validation set and the other 9 training folds as the learning set from which a random forest model is built. Hence, in total 100 iterations were performed.

This nested cross-validation structure was used to perform parameter optimization in a strict way, using only the training set and not the test set in each external cross-validation iteration. This is important because parameter optimization is part of the training of a classification algorithm, and it has to be done using the training set only. The test set is reserved purely for measuring generalization ability, i.e. the ability to correctly predict the classes of compounds not observed during training.

A random forest algorithm has two major parameters which are often optimized for the target dataset, namely: the number of trees in the forest (ntrees) and the number of candidate features evaluated to select the best feature in each tree node (mtry) [101]. In order to optimize these parameters, we tested five settings for the ntrees parameter, namely ntrees = 100, 300, 500, 700, and 900; and three settings for the mtry parameter, namely: the square root of the number of features in the dataset (the default setting in [23,102]), as well as the half and the double of that default setting. For other parameters, their default values in the "mlr" R package were used.

In the above nested scheme, in each iteration of the external cross-validation procedure, parameters are optimized as follows: we ran the random forest algorithm 15 times, each time with a different combination of parameter settings (5 ntrees settings times 3 mtry settings), and each time performing an internal cross-validation in the training set. The parameter setting combination producing the best median AUC value across the 10 internal cross-validation iterations was chosen as the optimized parameter settings for the current external cross-validation iteration, and then a random forest algorithm with those optimized parameter settings was ran using the entire training set available at the current external iteration, with its predictive accuracy being evaluated on the test set for that iteration. The final measure of predictive accuracy reported in the Results section is the median AUC value across the 10 test sets in the external cross-validation procedure.

Evaluation methodology

We evaluate the results of the random forest in three ways. First, we measure its predictive accuracy, using the well-known cross-validation procedure that is commonly used in supervised machine learning. Second, we identify the GO terms most relevant for predicting a compound's effect on *C. elegans*' lifespan, according to a feature score calculated by the random forest, and discuss the relevance of such GO terms to the biology of ageing research. Third, we apply the best classification model built by the random forest to a screening "external" dataset with compounds from the DGIdb database, where the effect of the compounds on an organism's lifespan is unknown. That model's predictions are then used to identify the "top hit" compounds in the DGIdb dataset which have more potential as a pharmacological intervention against ageing in *C. elegans*.

AUTHOR CONTRIBUTIONS

AAF & JPM conceived and coordinated the project. DGB, DT & DN contributed to dataset curation, calculation of dataset descriptions. DGB analysed the dataset. TG contributed to result interpretation and to discussion of manuscript. AAF and DN wrote the main manuscript. All authors reviewed and contributed to the manuscript.

ACKNOWLEDGEMENTS

DN would like to thank the Medway School of Pharmacy, Universities of Kent and Greenwich, for the use of software to calculate molecular descriptors.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

FUNDING

This work was supported by a Wellcome Trust grant (104978/Z/14/Z) to J.P.M, and a Leverhulme Trust research Grant (RPG-2016-015) to J.P.M. and A.A.F.

REFERENCES

1. Olshansky SJ, Perry D, Miller RA, Butler RN. Pursuing the longevity dividend: scientific goals for an aging world. Ann N Y Acad Sci. 2007; 1114:11–13. https://doi.org/10.1196/annals.1396.050

2. de Magalhães JP, Wuttke D, Wood SH, Plank M, Vora C. Genome-environment interactions that modulate aging: powerful targets for drug discovery. Pharmacol Rev. 2012; 64:88–101. https://doi.org/10.1124/pr.110.004499

3. Calvert S, Tacutu R, Sharifi S, Teixeira R, Ghosh P, de Magalhães JP. A network pharmacology approach reveals new candidate caloric restriction mimetics in C. elegans. Aging Cell. 2016; 15:256–66. https://doi.org/10.1111/acel.12432

4. Nadon NL, Strong R, Miller RA, Nelson J, Javors M, Sharp ZD, Peralba JM, Harrison DE. Design of aging intervention studies: the NIA interventions testing program. Age (Dordr). 2008; 30:187–99. https://doi.org/10.1007/s11357-008-9048-1

5. Strong R, Miller RA, Antebi A, Astle CM, Bogue M, Denzel MS, Fernandez E, Flurkey K, Hamilton KL, Lamming DW, Javors MA, de Magalhães JP, Martinez PA, et al. Longer lifespan in male mice treated with a weakly estrogenic agonist, an antioxidant, an α-glucosidase inhibitor or a Nrf2-inducer. Aging Cell. 2016; 15:872–84. https://doi.org/10.1111/acel.12496

6. Lucanic M, Plummer WT, Chen E, Harke J, Foulger AC, Onken B, Coleman-Hulbert AL, Dumas KJ, Guo S, Johnson E, Bhaumik D, Xue J, Crist AB, et al. Impact of genetic background and experimental reproducibility on identifying chemical compounds with robust longevity effects. Nat Commun. 2017; 8:14256. https://doi.org/10.1038/ncomms14256

7. Kenyon C. The first long-lived mutants: discovery of the insulin/IGF-1 pathway for ageing. Philos Trans R Soc Lond B Biol Sci. 2011; 366:9–16. https://doi.org/10.1098/rstb.2010.0276

8. Bartke A. Single-gene mutations and healthy ageing in mammals. Philos Trans R Soc Lond B Biol Sci. 2011; 366:28–34. https://doi.org/10.1098/rstb.2010.0281

9. Tacutu R, Craig T, Budovsky A, Wuttke D, Lehmann G, Taranukha D, Costa J, Fraifeld VE, de Magalhães JP. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. Nucleic Acids Res. 2013; 41:D1027–33. https://doi.org/10.1093/nar/gks1155

10. Barardo D, Thornton D, Thoppil H, Walsh M, Sharifi S, Ferreira S, Anžič A, Fernandes M, Monteiro P, Grum T, Cordeiro R, De-Souza EA, Budovsky A, et al. The DrugAge database of aging-related drugs. Aging Cell. 2017; 16:594–97. https://doi.org/10.1111/acel.12585

11. Matic I, Revandkar A, Chen J, Bisio A, Dall'Acqua S, Cocetta V, Brun P, Mancino G, Milanese M, Mattei M, Montopoli M, Alimonti A. Identification of Salvia haenkei as gerosuppressant agent by using an integrated senescence-screening assay. Aging (Albany NY). 2016; 8:3223–40. https://doi.org/10.18632/aging.101076

12. Moskalev A, Chernyagina E, de Magalhães JP, Barardo D, Thoppil H, Shaposhnikov M, Budovsky A, Fraifeld VE, Garazha A, Tsvetkov V, Bronovitsky E, Bogomolov V, Scerbacov A, et al. Geroprotectors.org: a new, structured and curated database of current therapeutic interventions in aging and age-related disease. Aging (Albany NY). 2015; 7:616–28. https://doi.org/10.18632/aging.100799

13. Wang Y, Chang J, Liu X, Zhang X, Zhang S, Zhang X, Zhou D, Zheng G. Discovery of piperlongumine as a potential novel lead for the development of senolytic agents. Aging (Albany NY). 2016; 8:2915–26. https://doi.org/10.18632/aging.101100

14. Hühne R, Thalheim T, Sühnel J. AgeFactDB--the JenAge Ageing Factor Database--towards data integration in ageing research. Nucleic Acids Res. 2014; 42:D892–96. https://doi.org/10.1093/nar/gkt1073

15. Moskalev A, Chernyagina E, de Magalhães JP, Barardo D, Thoppil H, Shaposhnikov M, Budovsky A, Fraifeld VE, Garazha A, Tsvetkov V, Bronovitsky E, Bogomolov V, Scerbacov A, et al. Geroprotectors.org: a new, structured and curated database of current therapeutic interventions in aging and age-related disease. Aging (Albany NY). 2015; 7:616–28. https://doi.org/10.18632/aging.100799

16. Fabris F, Magalhães JP, Freitas AA. A review of supervised machine learning applied to ageing research. Biogerontology. 2017; 18:171–88. https://doi.org/10.1007/s10522-017-9683-y

17. Ziehm M, Kaur S, Ivanov DK, Ballester PJ, Marcus D, Partridge L, Thornton JM. Drug repurposing for aging research using model organisms. Aging Cell. 2017. https://doi.org/10.1111/acel.12626

18. Ye X, Linton JM, Schork NJ, Buck LB, Petrascheck M. A pharmacological network for lifespan extension in Caenorhabditis elegans. Aging Cell. 2014; 13:206–15. https://doi.org/10.1111/acel.12163

19. Aliper A, Belikov AV, Garazha A, Jellen L, Artemov A, Suntsova M, Ivanova A, Venkova L, Borisov N, Buzdin A, Mamoshina P, Putin E, Swick AG, et al. In search for geroprotectors: in silico screening and in vitro validation of signalome-level mimetics of young healthy state. Aging (Albany NY). 2016; 8:2127–52.

https://doi.org/10.18632/aging.101047

20. Carretero M, Gomez-Amaro RL, Petrascheck M. Pharmacological classes that extend lifespan of Caenorhabditis elegans. Front Genet. 2015; 6:77. https://doi.org/10.3389/fgene.2015.00077

21. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017; 77:1-17. https://doi.org/10.18637/jss.v077.i01

22. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform. 2013; 14:315–26. https://doi.org/10.1093/bib/bbs034

23. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7:3. https://doi.org/10.1186/1471-2105-7-3

24. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. J Stat Softw. 2010; 36: 1-13. https://doi.org/10.18637/jss.v036.i11

25. Freitas AA, Limbu K, Ghafourian T. Predicting volume of distribution with decision tree-based regression methods using predicted tissue:plasma partition coefficients. J Cheminform. 2015; 7: 6. https://doi.org/10.1186/s13321-015-0054-x

26. Newby D, Freitas AA, Ghafourian T. Coping with unbalanced class data sets in oral absorption models. J Chem Inf Model. 2013; 53:461–74. https://doi.org/10.1021/ci300348u

27. Yang Z, Klionsky DJ. Eaten alive: a history of macroautophagy. Nat Cell Biol. 2010; 12:814–22. https://doi.org/10.1038/ncb0910-814

28. Feng Y, He D, Yao Z, Klionsky DJ. The machinery of macroautophagy. Cell Res. 2014; 24:24–41. https://doi.org/10.1038/cr.2013.168

29. Richard VR, Leonov A, Beach A, Burstein MT, Koupaki O, Gomez-Perez A, Levy S, Pluska L, Mattie S, Rafesh R, Iouk T, Sheibani S, Greenwood M, et al. Macromitophagy is a longevity assurance process that in chronologically aging yeast limited in calorie supply sustains functional mitochondria and maintains cellular lipid homeostasis. Aging (Albany NY). 2013; 5:234–69. https://doi.org/10.18632/aging.100547

30. Kurz T, Terman A, Brunk UT. Autophagy, ageing and apoptosis: the role of oxidative stress and lysosomal iron. Arch Biochem Biophys. 2007; 462:220–30. https://doi.org/10.1016/j.abb.2007.01.013

31. Rubinsztein DC, Mariño G, Kroemer G. Autophagy and aging. Cell. 2011; 146:682–95. https://doi.org/10.1016/j.cell.2011.07.030

32. Martinez-Lopez N, Athonvarangkul D, Singh R. Autophagy and aging. Longevity Genes. Springer; 2015. p. 73–87.

33. Kim I, Rodriguez-Enriquez S, Lemasters JJ. Selective degradation of mitochondria by mitophagy. Arch Biochem Biophys. 2007; 462:245–53. https://doi.org/10.1016/j.abb.2007.03.034

34. Palikaras K, Tavernarakis N. Mitophagy in neurodegeneration and aging. Front Genet. 2012; 3:297. https://doi.org/10.3389/fgene.2012.00297

35. Palikaras K, Lionaki E, Tavernarakis N. Coordination of mitophagy and mitochondrial biogenesis during ageing in C. elegans. Nature. 2015; 521:525–28. https://doi.org/10.1038/nature14300

36. Dancy BM, Sedensky MM, Morgan PG. Effects of the mitochondrial respiratory chain on longevity in C. elegans. Exp Gerontol. 2014; 56:245–55. https://doi.org/10.1016/j.exger.2014.03.028

37. Maglioni S, Arsalan N, Franchi L, Hurd A, Opipari AW, Glick GD, Ventura N. An automated phenotype-based microscopy screen to identify pro-longevity interventions acting through mitochondria in C. elegans. Biochim Biophys Acta. 2015; 1847:1469–78. https://doi.org/10.1016/j.bbabio.2015.05.004

38. Wilkinson B, Gilbert HF. Protein disulfide isomerase. Biochim Biophys Acta. 2004; 1699:35–44. https://doi.org/10.1016/S1570-9639(04)00063-9

39. Watanabe MM, Laurindo FR, Fernandes DC. Methods of measuring protein disulfide isomerase activity: a critical overview. Front Chem. 2014; 2:73. https://doi.org/10.3389/fchem.2014.00073

40. Naidoo N. ER and aging-Protein folding and the ER stress response. Ageing Res Rev. 2009; 8:150–59. https://doi.org/10.1016/j.arr.2009.03.001

41. Sano R, Reed JC. ER stress-induced cell death mechanisms. Biochim Biophys Acta. 2013; 1833:3460–70. https://doi.org/10.1016/j.bbamcr.2013.06.028

42. Grek C, Townsend DM. Protein Disulfide Isomerase Superfamily in Disease and the Regulation of Apoptosis. Endoplasmic Reticulum Stress Dis. 2014; 1:4–17. https://doi.org/10.2478/ersc-2013-0001

43. Kenyon CJ. The genetics of ageing. Nature. 2010; 464:504–12. https://doi.org/10.1038/nature08980

44. Fernandes M, Wan C, Tacutu R, Barardo D, Rajput A, Wang J, Thoppil H, Thornton D, Yang C, Freitas A, de Magalhães JP. Systematic analysis of the gerontome reveals links between aging and age-related diseases.

Hum Mol Genet. 2016; 25:4804–18.

45. Wagner AH, Coffman AC, Ainscough BJ, Spies NC, Skidmore ZL, Campbell KM, Krysiak K, Pan D, McMichael JF, Eldred JM, Walker JR, Wilson RK, Mardis ER, et al. DGIdb 2.0: mining clinically relevant drug-gene interactions. Nucleic Acids Res. 2016; 44:D1036–44. https://doi.org/10.1093/nar/gkv1165

46. Moghe A, Ghare S, Lamoreau B, Mohammad M, Barve S, McClain C, Joshi-Barve S. Molecular mechanisms of acrolein toxicity: relevance to human disease. Toxicol Sci. 2015; 143:242–55. https://doi.org/10.1093/toxsci/kfu233

47. Tilstra JS, Clauson CL, Niedernhofer LJ, Robbins PD. NF-κB in Aging and Disease. Aging Dis. 2011; 2:449–65.

48. Kehrer JP, Biswal SS. The molecular effects of acrolein. Toxicol Sci. 2000; 57:6–15. https://doi.org/10.1093/toxsci/57.1.6

49. Comer DM, Elborn JS, Ennis M. Inflammatory and cytotoxic effects of acrolein, nicotine, acetylaldehyde and cigarette smoke extract on human nasal epithelial cells. BMC Pulm Med. 2014; 14:32. https://doi.org/10.1186/1471-2466-14-32

50. Han J, Goldstein LA, Hou W, Froelich CJ, Watkins SC, Rabinowich H. Deregulation of mitochondrial membrane potential by mitochondrial insertion of granzyme B and direct Hax-1 cleavage. J Biol Chem. 2010; 285:22461–72. https://doi.org/10.1074/jbc.M109.086587

51. Shi Y. Emerging roles of cardiolipin remodeling in mitochondrial dysfunction associated with diabetes, obesity, and cardiovascular diseases. J Biomed Res. 2010; 24:6–15. https://doi.org/10.1016/S1674-8301(10)60003-6

52. Paradies G, Paradies V, De Benedictis V, Ruggiero FM, Petrosillo G. Functional role of cardiolipin in mitochondrial bioenergetics. Biochim Biophys Acta. 2014; 1837:408–17. https://doi.org/10.1016/j.bbabio.2013.10.006

53. Chicco AJ, Sparagna GC. Role of cardiolipin alterations in mitochondrial dysfunction and disease. Am J Physiol Cell Physiol. 2007; 292:C33–44. https://doi.org/10.1152/ajpcell.00243.2006

54. Harrison DE, Strong R, Sharp ZD, Nelson JF, Astle CM, Flurkey K, Nadon NL, Wilkinson JE, Frenkel K, Carter CS, Pahor M, Javors MA, Fernandez E, Miller RA. Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. Nature. 2009; 460:392–95. https://doi.org/10.1038/nature08221

55. Miller RA, Harrison DE, Astle CM, Baur JA, Boyd AR, de Cabo R, Fernandez E, Flurkey K, Javors MA, Nelson JF, Orihuela CJ, Pletcher S, Sharp ZD, et al. Rapamycin, but not resveratrol or simvastatin, extends life span of genetically heterogeneous mice. J Gerontol A Biol Sci Med Sci. 2011; 66:191–201. https://doi.org/10.1093/gerona/glq178

56. Robida-Stubbs S, Glover-Cutter K, Lamming DW, Mizunuma M, Narasimhan SD, Neumann-Haefelin E, Sabatini DM, Blackwell TK. TOR signaling and rapamycin influence longevity by regulating SKN-1/Nrf and DAF-16/FoxO. Cell Metab. 2012; 15:713–24. https://doi.org/10.1016/j.cmet.2012.04.007

57. Bjedov I, Toivonen JM, Kerr F, Slack C, Jacobson J, Foley A, Partridge L. Mechanisms of life span extension by rapamycin in the fruit fly Drosophila melanogaster. Cell Metab. 2010; 11:35–46. https://doi.org/10.1016/j.cmet.2009.11.010

58. Ehninger D, Neff F, Xie K. Longevity, aging and rapamycin. Cell Mol Life Sci. 2014; 71:4325–46. https://doi.org/10.1007/s00018-014-1677-1

59. Hansen M, Kennedy BK. Does Longer Lifespan Mean Longer Healthspan? Trends Cell Biol. 2016; 26:565–68. https://doi.org/10.1016/j.tcb.2016.05.002

60. Gabriel D, Gordon LB, Djabali K. Temsirolimus Partially Rescues the Hutchinson-Gilford Progeria Cellular Phenotype. PLoS One. 2016; 11:e0168988. https://doi.org/10.1371/journal.pone.0168988

61. Bustamante J, Caldas Lopes E, Garcia M, Di Libero E, Alvarez E, Hajos SE. Disruption of mitochondrial membrane potential during apoptosis induced by PSC 833 and CsA in multidrug-resistant lymphoid leukemia. Toxicol Appl Pharmacol. 2004; 199:44–51. https://doi.org/10.1016/j.taap.2004.03.021

62. Dai Y, Rahmani M, Grant S. Proteasome inhibitors potentiate leukemic cell apoptosis induced by the cyclin-dependent kinase inhibitor flavopiridol through a SAPK/JNK- and NF-kappaB-dependent process. Oncogene. 2003; 22:7108–22. https://doi.org/10.1038/sj.onc.1206863

63. Van Waes C, Chang AA, Lebowitz PF, Druzgal CH, Chen Z, Elsayed YA, Sunwoo JB, Rudy SF, Morris JC, Mitchell JB, Camphausen K, Gius D, Adams J, et al. Inhibition of nuclear factor-kappaB and target genes during combined therapy with proteasome inhibitor bortezomib and reirradiation in patients with recurrent head-and-neck squamous cell carcinoma. Int J Radiat Oncol Biol Phys. 2005; 63:1400–12. https://doi.org/10.1016/j.ijrobp.2005.05.007

64. Milani M, Rzymski T, Mellor HR, Pike L, Bottini A, Generali D, Harris AL. The role of ATF4 stabilization and autophagy in resistance of breast cancer cells treated with Bortezomib. Cancer Res. 2009; 69:4415–

23. https://doi.org/10.1158/0008-5472.CAN-08-2839

65. Burger K, Mühl B, Harasim T, Rohrmoser M, Malamoussi A, Orban M, Kellner M, Gruber-Eber A, Kremmer E, Hölzel M, Eick D. Chemotherapeutic drugs inhibit ribosome biogenesis at various levels. J Biol Chem. 2010; 285:12416–25. https://doi.org/10.1074/jbc.M109.074211

66. Goudarzi KM, Nistér M, Lindström MS. mTOR inhibitors blunt the p53 response to nucleolar stress by regulating RPL11 and MDM2 levels. Cancer Biol Ther. 2014; 15:1499–514. https://doi.org/10.4161/15384047.2014.955743

67. Powers RW 3rd, Kaeberlein M, Caldwell SD, Kennedy BK, Fields S, Fields S. Extension of chronological life span in yeast by decreased TOR pathway signaling. Genes Dev. 2006; 20:174–84. https://doi.org/10.1101/gad.1381406

68. Tucci P. Caloric restriction: is mammalian life extension linked to p53? Aging (Albany NY). 2012; 4:525–34. https://doi.org/10.18632/aging.100481

69. Hasty P, Christy BA. p53 as an intervention target for cancer and aging. Pathobiol Aging Age Relat Dis. 2013; 3:3. https://doi.org/10.3402/pba.v3i0.22702

70. Budovsky A, Tacutu R, Yanai H, Abramovich A, Wolfson M, Fraifeld V. Common gene signature of cancer and longevity. Mech Ageing Dev. 2009; 130:33–39. https://doi.org/10.1016/j.mad.2008.04.002

71. Blagosklonny MV. Selective anti-cancer agents as anti-aging drugs. Cancer Biol Ther. 2013; 14:1092–97. https://doi.org/10.4161/cbt.27350

72. Cornu M, Albert V, Hall MN. mTOR in aging, metabolism, and cancer. Curr Opin Genet Dev. 2013; 23:53–62. https://doi.org/10.1016/j.gde.2012.12.005

73. Cartwright T, Perkins ND, L Wilson C. NFKB1: a suppressor of inflammation, ageing and cancer. FEBS J. 2016; 283:1812–22. https://doi.org/10.1111/febs.13627

74. Franceschi C, Campisi J. Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. J Gerontol A Biol Sci Med Sci. 2014 (Suppl 1); 69:S4–9. https://doi.org/10.1093/gerona/glu057

75. Woods JA, Wilund KR, Martin SA, Kistler BM. Exercise, inflammation and aging. Aging Dis. 2012; 3:130–40.

76. Arai Y, Martin-Ruiz CM, Takayama M, Abe Y, Takebayashi T, Koyasu S, Suematsu M, Hirose N, von Zglinicki T. Inflammation, But Not Telomere Length, Predicts Successful Ageing at Extreme Old Age: A Longitudinal Study of Semi-supercentenarians. EBioMedicine. 2015; 2:1549–58. https://doi.org/10.1016/j.ebiom.2015.07.029

77. Travis S, Yap LM, Hawkey C, Warren B, Lazarov M, Fong T, Tesi RJ, and RDP Investigators Study Group. RDP58 is a novel and potentially effective oral therapy for ulcerative colitis. Inflamm Bowel Dis. 2005; 11:713–19. https://doi.org/10.1097/01.MIB.0000172807.26748.16

78. Murthy S, Flanigan A, Coppola D, Buelow R. RDP58, a locally active TNF inhibitor, is effective in the dextran sulphate mouse model of chronic colitis. Inflamm Res. 2002; 51:522–31. https://doi.org/10.1007/PL00012423

79. De Vry CG, Valdez M, Lazarov M, Muhr E, Buelow R, Fong T, Iyer S. Topical application of a novel immunomodulatory peptide, RDP58, reduces skin inflammation in the phorbol ester-induced dermatitis model. J Invest Dermatol. 2005; 125:473–81. https://doi.org/10.1111/j.0022-202X.2005.23831.x

80. Filipov NM, Thompson FN, Stuedemann JA, Elsasser TH, Kahl S, Stanker LH, Young CR, Dawe DL, Smith CK. Anti-Inflammatory Effects of Ergotamine in Steers. Proc Soc Exp Biol Med. 2000; 225:136–42.

81. Xia M, Sreedharan SP, Bolin DR, Gaufo GO, Goetzl EJ. Novel cyclic peptide agonist of high potency and selectivity for the type II vasoactive intestinal peptide receptor. J Pharmacol Exp Ther. 1997; 281:629–33.

82. Delgado M, Abad C, Martinez C, Leceta J, Gomariz RP. Vasoactive intestinal peptide prevents experimental arthritis by downregulating both autoimmune and inflammatory components of the disease. Nat Med. 2001; 7:563–68. https://doi.org/10.1038/87887

83. Delgado M, Gomariz RP, Martinez C, Abad C, Leceta J. Anti-inflammatory properties of the type 1 and type 2 vasoactive intestinal peptide receptors: role in lethal endotoxic shock. Eur J Immunol. 2000; 30:3236–46. https://doi.org/10.1002/1521-4141(200011)30:11<3236::AID-IMMU3236>3.0.CO;2-L

84. Millar RP, Lu ZL, Pawson AJ, Flanagan CA, Morgan K, Maudsley SR. Gonadotropin-releasing hormone receptors. Endocr Rev. 2004; 25:235–75. https://doi.org/10.1210/er.2003-0002

85. Borm G, Mannaerts B, and The European Orgalutran Study Group. Treatment with the gonadotrophin-releasing hormone antagonist ganirelix in women undergoing ovarian stimulation with recombinant follicle stimulating hormone is effective, safe and convenient: results of a controlled, randomized, multicentre trial. Hum Reprod. 2000; 15:1490–98. https://doi.org/10.1093/humrep/15.7.1490

86. Leroy I, d'Acremont M, Brailly-Tabard S, Frydman R, de Mouzon J, Bouchard P. A single injection of a

gonadotropin-releasing hormone (GnRH) antagonist (Cetrorelix) postpones the luteinizing hormone (LH) surge: further evidence for the role of GnRH during the LH surge. Fertil Steril. 1994; 62:461–67. https://doi.org/10.1016/S0015-0282(16)56932-5

87. Letassy NA, Thompson DF, Britton ML, Suda RR Sr. Nafarelin acetate: a gonadotropin-releasing hormone agonist for the treatment of endometriosis. DICP. 1990; 24:1204–09. https://doi.org/10.1177/106002809002401212

88. Spitz IM, Chertin B, Lindenberg T, Farkas A. Long-acting gonadotropin-releasing hormone implant to maintain medical castration for two years in men with prostate cancer. N Engl J Med. 1999; 340:1439–1439. https://doi.org/10.1056/NEJM199905063401814

89. Eugster EA, Clarke W, Kletter GB, Lee PA, Neely EK, Reiter EO, Saenger P, Shulman D, Silverman L, Flood L, Gray W, Tierney D. Efficacy and safety of histrelin subdermal implant in children with central precocious puberty: a multicenter trial. J Clin Endocrinol Metab. 2007; 92:1697–704. https://doi.org/10.1210/jc.2006-2479

90. Zhang G, Li J, Purkayastha S, Tang Y, Zhang H, Yin Y, Li B, Liu G, Cai D. Hypothalamic programming of systemic ageing involving IKK-β, NF-κB and GnRH. Nature. 2013; 497:211–16. https://doi.org/10.1038/nature12143

91. Gore AC, Windsor-Engnell BM, Terasawa E. Menopausal increases in pulsatile gonadotropin-releasing hormone release in a nonhuman primate (Macaca mulatta). Endocrinology. 2004; 145:4653–59. https://doi.org/10.1210/en.2004-0379

92. Gruenewald DA, Naai MA, Marck BT, Matsumoto AM. Age-related decrease in hypothalamic gonadotropin-releasing hormone (GnRH) gene expression, but not pituitary responsiveness to GnRH, in the male Brown Norway rat. J Androl. 2000; 21:72–84.

93. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound databases. Nucleic Acids Res. 2016; 44:D1202–13. https://doi.org/10.1093/nar/gkv951

94. Stekhoven DJ, Bühlmann P. MissForest--nonparametric missing value imputation for mixed-type data. Bioinformatics. 2012; 28:112–18. https://doi.org/10.1093/bioinformatics/btr597

95. Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, von Mering C, Jensen LJ, Bork P. STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res. 2014; 42:D401–07. https://doi.org/10.1093/nar/gkt1207

96. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pagès F, Trajanoski Z, Galon J. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009; 25:1091–93. https://doi.org/10.1093/bioinformatics/btp101

97. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–504. https://doi.org/10.1101/gr.1239303

98. Breiman L. Random forests. Mach Learn. Springer. 2001; 45:5–32. https://doi.org/10.1023/A:1010933404324

99. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Jones Z, Casalicchio G. mlr: Machine Learning in R. R package version 2.9 https://CRAN.R-project.org/package=mlr. 2016.

100. Nicodemus KK, Malley JD, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics. 2010; 11:110. https://doi.org/10.1186/1471-2105-11-110

101. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform. 2013; 14:315–26. https://doi.org/10.1093/bib/bbs034

102. Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics. 2008; 9:319. https://doi.org/10.1186/1471-2105-9-319

SUPPLEMENTARY MATERIAL

Please browse Full text version to see the Supplementary Tables of this manuscript.

Table S1. Table of chemical compounds that increase or have no effect/decrease longevity in C.elegans used in this work,

Table S2. Table of GO terms and chemical descriptors highlighting those selected via feature selection from the best random forest model.

Table S3. Table showing the predicted probabilities for Longevity effects for the external dataset DGIdb compounds.