



OPEN

Random forest classification for predicting lifespan-extending chemical compounds

Sofia Kapsiani & Brendan J. Howlin

Ageing is a major risk factor for many conditions including cancer, cardiovascular and neurodegenerative diseases. Pharmaceutical interventions that slow down ageing and delay the onset of age-related diseases are a growing research area. The aim of this study was to build a machine learning model based on the data of the DrugAge database to predict whether a chemical compound will extend the lifespan of *Caenorhabditis elegans*. Five predictive models were built using the random forest algorithm with molecular fingerprints and/or molecular descriptors as features. The best performing classifier, built using molecular descriptors, achieved an area under the curve score (AUC) of 0.815 for classifying the compounds in the test set. The features of the model were ranked using the Gini importance measure of the random forest algorithm. The top 30 features included descriptors related to atom and bond counts, topological and partial charge properties. The model was applied to predict the class of compounds in an external database, consisting of 1738 small-molecules. The chemical compounds of the screening database with a predictive probability of ≥ 0.80 for increasing the lifespan of *Caenorhabditis elegans* were broadly separated into (1) flavonoids, (2) fatty acids and conjugates, and (3) organooxygen compounds.

Pharmacological interventions for longevity extension. Ageing is a major health, social and financial challenge, characterised by the deterioration of the physiological processes of an organism^{1,2}. Ageing is a predominant risk factor for many conditions including various types of cancers, cardiovascular and neurodegenerative diseases^{3,4}. Interventions targeting the cellular and molecular process of ageing have the potential to delay and protect against age-related conditions.

Several ageing studies have identified interventions that extend the lifespan of model organisms ranging from nematodes and fruit flies to rodents, using dietary restrictions, genetic modifications and pharmaceutical interventions. Lee et al. (2006) presented the first evidence that long-term dietary deprivation can improve longevity in a multicellular species, *Caenorhabditis elegans* (*C. elegans*)⁵. A few years later, Harrison et al. (2009) found that treating mice with rapamycin, an inhibitor of the mTOR pathway, extended the median and maximum lifespan of the mice⁶. Additionally, Selman et al. (2009) reported that genetic deletion of S6 protein kinase 1, a component of the mTOR signalling pathway, increased the lifespan of mice and protected against age-related conditions⁷.

Ye et al. (2014) developed a pharmacological network to identify pharmacological classes related to the ageing of *C. elegans*⁸. The network showed that resistance to oxidative stress and lifespan extension clustered in a few pharmacological classes, most of them related to intercellular signalling⁸. Moreover, Putin et al. (2016) developed a deep learning neural network that predicted human chronological age from a basic blood test⁹. The study identified the top five most critical blood markers for determining chronological age in humans, which were albumin, glucose, alkaline phosphatase, urea and erythrocytes⁹. Additionally, Mamoshina et al. (2018) developed a deep learning-based haematological ageing clock using blood samples from Canadian, South Korean, and Eastern European populations, with millions of subjects¹⁰. The findings showed that population-specific ageing clocks were more accurate in predicting chronological age and quantifying biological age than generic ageing clocks¹⁰.

Barardo et al. (2017) built a random forest model to predict whether a compound would increase the lifespan of *C. elegans* based on the data of the DrugAge database^{1,4}. The features used to build the model were molecular descriptors and gene ontology terms. Feature selection was performed using random forest's feature importance measure. The best performing model, with an AUC score of 0.80, was applied to predict the class of the compounds in the DGIdb database.

Department of Chemistry, FEPS, University of Surrey, Guildford, Surrey GU2 7XH, UK. email: b.howlin@surrey.ac.uk

Model	Number of selected features	Cross-validation on training set (AUC±SD) (%)	Test set (AUC) (%)
ECFP_1024	55	79.4±4.8	79.3
ECFP_2048	504	78.9±4.2	77.6
RDKit5	654	83.6±5.3	77.7
MD	69	82.3±4.1	81.5
ECFP_1024_MD	33	82.8±4.0	80.6

Table 1. Performances on tenfold cross-validation and the test set achieved by each descriptor type; ECFP (1024-bit length), ECFP (2048-bit length), RDKit fingerprints (2048-bit length), molecular descriptors and the combination of ECFP (1024-bit length) with molecular descriptors. The number of features selected by applying variance and mutual information-based methods is reported.

Purpose of the work. This study builds on the work conducted by Barardo et al. (2017) to further explore the use of the DrugAge database for predicting compounds with anti-ageing properties⁴. Specifically, the random forest algorithm was employed to predict whether a compound will increase the lifespan of *C. elegans*. This was achieved by building five predictive models, each using different descriptor types, based on the data of the DrugAge database published by Barardo et al. (2017)⁴. The features of the models were molecular fingerprints and/or molecular descriptors calculated from the structure of the compounds in the DrugAge database. Feature selection was performed using variance and mutual information-based methods. To the best of our knowledge, this is the first implementation of molecular fingerprints for building machine learning models based on the entries of the DrugAge database. The best performing model was applied to predict the class of the compounds in an external database, consisting of 1738 small-molecules extracted from the DrugBank database¹¹.

Random forest models. Random forest is a supervised machine learning technique, consisted of an ensemble of decision trees, where each tree is trained independently using a random subset of the data¹². The random forest model is widely used in chemo- and bioinformatics related tasks as it is robust to overfitting on high dimensional databases with a small sample sizes⁴.

The choice of chemical descriptors can significantly impact the quality and predictions of quantitative structure–activity relationship (QSAR) models. Descriptors represent chemical information of molecules in a digital or numerical way that is suitable for model development and are computer-interpretable^{13,14}. In this study, 2D and 3D molecular descriptors were calculated using the Molecular Operating Environment (MOE) software¹⁵. 2D descriptors are calculated from the 2D structure of a molecule and provide information related to its structural, topological and physicochemical properties¹⁶. On the other hand, 3D descriptors are generated from the 3D structure of a chemical compound and include electronic parameters (e.g. dipole momentum), quantum–chemical descriptors (e.g. HOMO and LUMO energies), and surface:volume descriptors^{14,17,18}.

Molecular fingerprints represent the molecule's structure using binary vectors, where 1 corresponds to a particular feature or structural group being present and 0 that it is absent. Herein, extended-connectivity fingerprints (ECFP) of 1024- and 2048-bit lengths and RDKit topological fingerprints of 2048-bit length were generated in the RDKit Python environment¹⁹. Lastly, the combination of molecular descriptors with ECFPs was tested.

Results and discussion

Feature selection. Variance and mutual information filter-based methods were employed to select a subset of relevant features for predicting the anti-ageing properties of the molecules in the dataset. This was performed only for the training set which contained 80% of compounds in the dataset. Feature selection reduced the number of variables used by each model, making computational calculations less expensive. The median AUC scores and standard deviation of tenfold cross-validation (on the training set) obtained by random forest classification for each feature subset can be found in Supplementary Fig. 1, Additional File 1. The feature subset with the highest AUC score in 10-fold cross-validation was selected for classifying the compounds in the test set. In cases where two feature subsets achieved the same AUC score, the subset with the lowest standard deviation was used.

Model selection. The test set contained 20% of the data not used in training the models. The performances of the random forest classifiers on the test and train set as well as the optimal number of variables obtained by feature selection are shown in Table 1.

As illustrated in Fig. 1, the predictive performances of the random forest models built with ECFP and MD features did not significantly drop for classifying the compounds in the test set and were compatible with the spread of the AUC scores from cross-validation. This indicated that overfitting was minimised. On the other hand, the performance of the classifier using topological RDKit fingerprints on the test set dropped by approximately 6%. Therefore, the RDKit5 model was not selected for predicting the effect of the compounds in the screening database on the lifespan of *C. elegans*.

The receiver operating characteristic (ROC) curve is the plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) at varying classification thresholds. The ROC curves, displayed in Fig. 2, compare the performances of the descriptor types for classifying the samples of the test set. The figure illustrates that the five random forest models performed better than a random prediction.

The best performing model, selected by its ability to correctly classify the compounds in the test set, was used for predicting the class of the compounds in the screening dataset. The classifier built solely with molecular

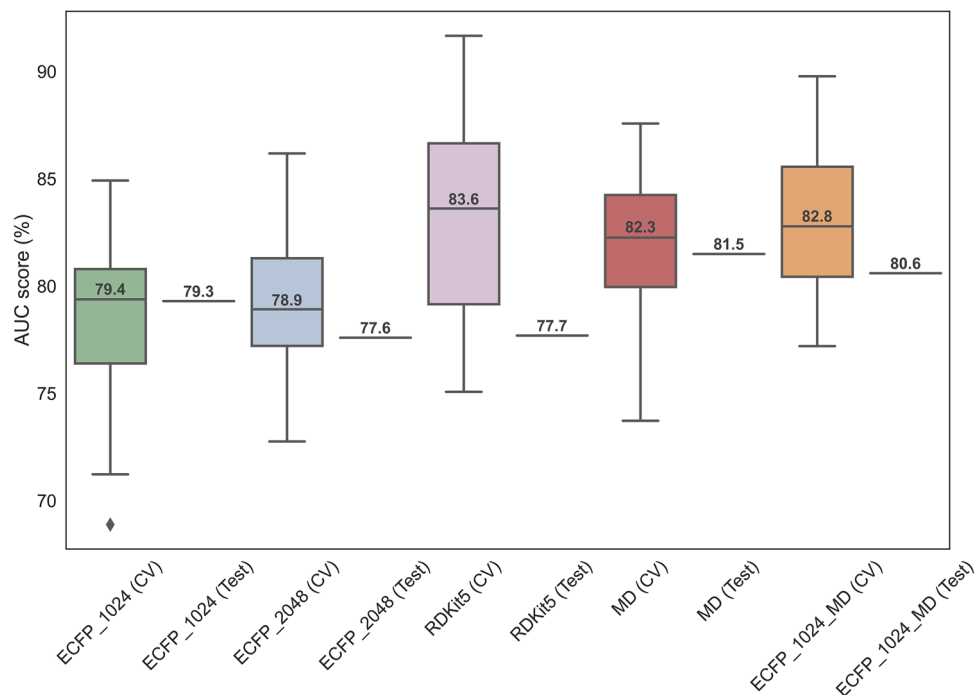


Figure 1. Box-and-whisker plot displaying the distribution of the AUC for the tenfold cross-validation (CV) on the training set and the AUC score for the single measurement taken on the test set, obtained by random forest classification. Each box represents the cross-validation data for a different model, where ECFP of 1024-bit length is shown in green, ECFP of 2048-bit length in blue, RDKit fingerprints in pink, molecular descriptors in red and the combination of ECFP of 1024-bit length with molecular descriptors are represented in orange colour. The value reported within the boxes is the median AUC value of the tenfold cross-validation. The points outside the boxplots represent possible outliers.

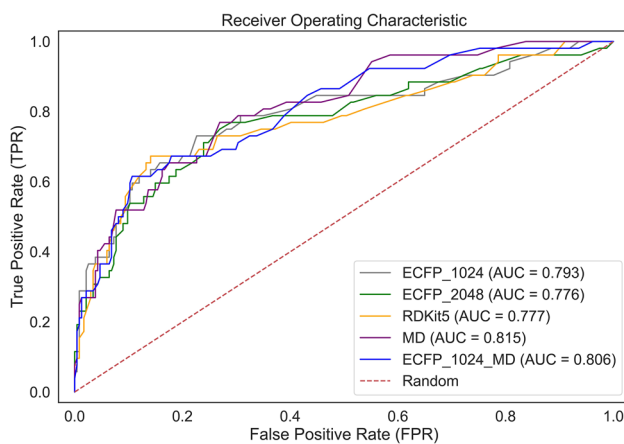


Figure 2. The ROC curves comparing the performances of the ECFP_1024 (in grey), ECFP_2048 (in green), RDKit5 (in orange), MD (in purple) and ECFP_1024_MD (in blue) for classifying the compounds in the test set. The AUC scores are reported for each descriptor type. The red dashed line corresponds to a random classifier, that gives random answers, with an AUC value of 0.5²⁰.

descriptors (MD), achieved the highest AUC score for predicting the class of the compounds in the test set. Combining MD with ECFP_1024, the feature type used to obtain the model with the second-highest predictive ability, did not further improve the performance of the classifier. The ECFP_1024 features could have provided additional information that was not useful to the random forest classifier making the predictions more difficult. Therefore, the MD model, which had an AUC score of 0.815 for classifying the compounds in the test set, was selected for further analysis.

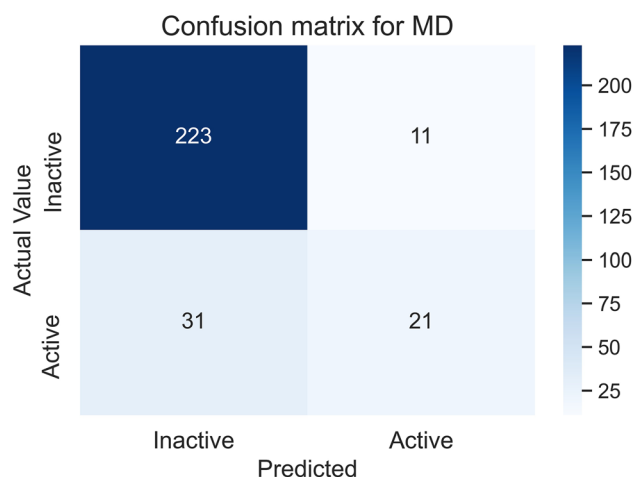


Figure 3. The confusion matrix was obtained from the performance of the MD model on the test set. The test set had a total of 286 compounds, of which 52 were active and 234 were inactive.

Confusion matrix. The confusion matrix of the MD model for predicting the class of the molecules in the test set is shown in Fig. 3. The classification accuracy of the model was 0.853 and the AUC score was 0.815.

The calculation of the Positive Predictive Value (PPV), Equation, and Negative Predictive Value (NPV), Equation, is shown below:

$$\text{PPV} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (1)$$

$$\text{PPV} = \frac{21}{21 + 11} = 65.6 \%$$

$$\text{NPV} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}} \quad (2)$$

$$\text{NPV} = \frac{223}{223 + 31} = 87.8 \%$$

In binary classification, the PPV and NPV are the percentage of correctly classified compounds among all compounds predicted as positives or negatives, respectively. Herein, the PPV and NPV indicate that the random forest model performed better on correctly classifying inactive compounds than active ones. The data used in this study was imbalanced as approximately 79% of the samples were negative entries. Thus, a random prediction that a compound is inactive had a much higher initial probability of being correct. To handle the imbalanced data, the “class_weight” argument of the random forest algorithm was set to “balanced”, which penalises misclassification of the minority class (i.e. the positive samples)²¹. The remaining parameters of the random forest model were left to the default settings of the *scikit-learn* Python library (please refer to the “[Random forest settings](#)” section in the Methods)²². This increased the PPV for classifying the compounds in the test set from 61.1% (value without balancing the class weights) to 65.6% (score achieved after balancing the class weights), while the NPV was reduced by 0.2%.

Feature importance. Investigating which features are considered “more important” by black-box models such as random forest can aid understanding of how these models make predictions. In this experiment, the feature relevance was measured using the “Gini importance” of the random forest algorithm. The selected model, MD, was composed of 69 molecular descriptors calculated by the MOE software²³. The table containing the full feature ranking can be found in Additional File 2. The analysis was focused on the top 30 features with the highest Gini importance (Table 2), which contained both 2D and 3D molecular descriptors.

The highest-ranking features were broadly separated into the following categories (1) atom and bond counts (2) topological and (3) partial charge descriptors.

Atom and bond counts are simple descriptors that do not provide any information on molecular geometry or atom connectivity. The highest-ranking atom and bond count descriptors were a_nN, b_single, a_count, opr_brigid, and a_nH. While very simplistic, the atom and bond counts outperformed more complex 2D and 3D molecular descriptors. This is because atom and bond counts can partially capture the overall properties of a compound such as size, hydrogen bonding and polarity, which often impact the activity of a drug²⁴. The number of nitrogen atoms, a_nN, was the top-ranking feature of the MD random forest model with a Gini importance score of 0.062. This is consistent with the results of Barardo et al. (2017) where a_nN was also ranked highest for predicting the class of the compounds in the DrugAge database⁴. Nitrogen atoms could have affected the physicochemical properties of the drugs as well as the interactions and binding of the molecules with target residues.

Gini importance	Feature	Description
0.062	a_nN	Number of nitrogen atoms
0.029	PEOE_VSA + 2	Total positive van der Waals surface area of atoms with a partial charge in the range of 0.10 to 0.15
0.026	vsurf_D8	Hydrophobic volume
0.024	h_pKa	The pKa of the reaction that removes a proton
0.023	SMR_VSA6	Sum of van der Waals surface areas such that the molar refractivity contribution is in the range of 0.485 to 0.560
0.023	rsynth	A value in [0,1] indicating the synthetic reasonableness, or feasibility, of the chemical structure
0.022	PEOE_VSA-4	Total positive van der Waals surface area of atoms with a partial charge in the range of - 0.25 to - 0.20
0.021	PEOE_VSA + 4	Total positive van der Waals surface area of atoms with a partial charge in the range of 0.20 to 0.25
0.021	PEOE_VSA-6	Total positive van der Waals surface area of atoms with a partial charge that is less than - 0.30
0.021	PEOE_VSA_PPOS	Total positive van der Waals surface area of atoms with a partial charge that is greater than 0.20
0.020	chi0_C	Carbon connectivity index (order 0)
0.020	Q_VSA_PNEG	Total negative polar van der Waals surface area of atoms of with a partial charge that is less than - 0.20
0.020	PEOE_VSA_POL	Total polar van der Waals surface area of atoms of which the absolute value of their partial charge is greater than 0.20
0.020	chi0v_C	Carbon valence connectivity index (order 0)
0.019	SMR_VSA3	Sum of van der Waals surface areas such that the molar refractivity contribution is in the range of 0.35 to 0.39
0.019	Q_VSA_PPOS	Total positive van der Waals surface area of atoms with a partial charge that is greater than 0.20
0.018	b_single	Number of single bonds
0.018	a_count	Number of atoms
0.018	SlogP_VSA3	Sum of van der Waals surface areas such that the logP(o/w) is in the range of 0.0 to 0.1
0.018	PEOE_VSA_PNEG	Total negative polar van der Waals surface area of atoms of with a partial charge that is less than - 0.20
0.017	TPSA	Topological polar surface area
0.017	zagreb	Zagreb index
0.017	weinerPol	Wiener polarity number
0.017	opr_brigid	The number of rigid bonds
0.017	Kier3	Third kappa shape index
0.016	PEOE_VSA-1	Total positive van der Waals surface area of atoms with a partial charge in the range of - 0.10 to - 0.05
0.016	chi0	Atomic connectivity index (order 0)
0.016	Kier2	Second kappa shape index
0.016	SlogP_VSA2	Sum of van der Waals surface areas such that the logP(o/w) is in the range of - 0.2 to 0.0
0.015	a_nH	Number of hydrogen atoms

Table 2. Top 30 features ranked by Gini importance for the MD random forest model. The description of the features was taken from the MOE software documentation²³.

The highest-ranking topological descriptors included chi0_C, chi0v_C, zagreb, weinerPol, Kier3, chi0 and Kier2. Topological descriptors take into account atom connectivity. The descriptors are computed from molecular graphs, where atoms are represented by vertices and the bonds by edges²⁵. These descriptors can provide information on the degree branching of the structure as well as molecular size and shape²⁵. Although topological descriptors are extensively used in predictive modelling, they are usually hard to interpret²⁶. Topological descriptors may have provided information on how well a molecule fits in the binding site and along with atom counts the interactions with the binding residues.

Top ranking partial charge descriptors were PEOE_VSA + 2, PEOE_VSA-4, PEOE_VSA + 4, PEOE_VSA-6, PEOE_VSA_PPOS, Q_VSA_PNEG, PEOE_VSA_POL, Q_VSA_PPOS and PEOE_VSA_PNEG. The “PEOE_” prefix denotes descriptors calculated using the partial equalization of orbital electronegativity (PEOE) algorithm for quantification of partial charges in the σ -system^{27,28}. On the other hand, descriptors prefixed with “Q_” were calculated using the Amber10:EHT force field²³. In a ligand-receptor system, partial charges can play a key role in the binding properties of the molecule as well as molecular recognition.

Predicting potential lifespan-extending compounds. The MD random forest model was applied to predict the class compounds in an external database, consisting of 1738 small-molecules obtained from the DrugBank database¹¹. The top-ranking compounds with a predictive probability of ≥ 0.80 for increasing the lifespan of *C. elegans* are shown in Table 3. The full ranking of the molecules in the screening database can be found in Additional File 2. The compounds were broadly separated into the following categories; (1) flavonoids, (2) fatty acids and conjugates, and (3) organooxygen compounds. The compounds were classified based on

Compound name	Predictive probability	Compound category
Diosmin	0.96	Flavonoids
Gamolenic acid	0.95	Fatty acids and conjugates
Rutin	0.95	Flavonoids
Hesperidin	0.94	Flavonoids
Lactose	0.89	Organooxygen compounds
6"-O-Malonyldaidzin	0.84	Flavonoids
Fidaxomicin	0.84	Other/Macrolide
Sucrose	0.83	Organooxygen compounds
Lactulose	0.83	Organooxygen compounds
Sodium aurothiomalate	0.82	Fatty acids and conjugates
Alloin	0.81	Other
Rifapentine	0.81	Other/Macrolactams
Plecanatide	0.80	Other/Polypeptides
Calcifediol	0.80	Other/Steroids and steroid derivatives
Chlortetracycline	0.80	Other/Tetracyclines

Table 3. Chemical compounds from the screening database with a predictive probability of 0.80 or above for increasing the of *C. elegans*.

categories “Class” and “Sub Class” of the chemical taxonomy section of the DrugBank database (provided by Classyfire) or assigned manually if not present²⁹.

Flavonoids. Flavonoids are a group of secondary metabolites in plants that are common polyphenols in the human diet³⁰. Major nutritional sources include tea, soy, fruits, vegetables, wine and nuts^{30,31}. Flavonoids are separated into subclasses based on their chemical structure, including flavones, flavonols, flavanones, and isoflavones³⁰.

Flavonoids have been associated with health benefits for age-related conditions such as metabolic diseases, cancer, inflammation and cognitive decline^{30,31}. Possible mechanisms of action include antioxidant activity, scavenging of radicals, central nervous system effects, alteration of the intestinal transport, sequestration and processing of fatty acids, PPAR activation and increase of insulin sensitivity³⁰.

Diosmin was the top-hit molecule in the screening database, with a predictive probability of 0.96. Diosmin is a flavonol glycoside that is either extracted from plants such as Rutaceae or obtained synthetically³². It has anti-inflammatory, free radical scavenging, and anti-mutagenic properties and has been used medically to treat pain and bleeding of haemorrhoids, chronic venous disease and lymphedema³³. Nevertheless, diosmin has poor aqueous solubility, which is a challenge for oral administration³⁴. Kamel et al. (2017) found that a combination of diosmin with essential oils showed skin antioxidant, anti-ageing and sun-blocking effects on mice³⁴. The underlying mechanisms for diosmin's anti-ageing and photo-protective effects include enhancing lymphatic drainage, ameliorating capillary microcirculation inflammation and preventing leukocyte activation, trapping, and migration^{34,35}.

Other flavonoids that ranked high for increasing the lifespan of *C. elegans* were rutin and hesperidin with a predictive probability of 0.95 and 0.94, respectively. Rutin (or quercetin-3-rutinoside), is a flavonol glycoside that is abundant in many plants such as passionflower, apple, tea, buckwheat seeds and citrus fruits^{36,37}. It possesses a range of biological properties including antioxidant, anticancer, neuroprotective, cardio-protective and skin-regenerative activities^{36,37}. Rutin had a high structural similarity to other flavonoids in the DrugAge database and particularly with quercetin 3-O- β -d-glucopyranoside-(4 \rightarrow 1)- β -d-glucopyranoside (Q3M). The Tanimoto coefficient between the RDKit fingerprints of Q3M and rutin was 0.99. The similarity map between the two compounds is shown in Fig. 4.

Q3M is a flavonoid abundant in onion peel that was found to extend the lifespan of *C. elegans*³⁹. In the same study, even although rutin was found to improve the tolerance of *C. elegans* to oxidative stress, which is desirable for longevity, rutin (20 μ g/mL) did not significantly affect the worm's lifespan³⁹. On the other hand, a more recent study by Cordeiro et al. (2020) found that treatment of *C. elegans* with 15, 30 and 120 μ M rutin increased the lifespan of the nematodes from 28 (control) to 30 days⁴⁰.

Hesperidin has shown reactive oxygen species (ROS) inhibition and anti-ageing effects in the yeast species *Saccharomyces cerevisiae*⁴¹. Fernández-Bedmar et al. (2011) found that hesperidin extracted from orange juice had a positive influence on the lifespan of *D. melanogaster*⁴². Wang et al. (2020) showed that orange extracts, where hesperidin was the predominant phenolic compound, increased the mean lifespan of *C. elegans*⁴³. In the same study, orange extracts were also found to promote longevity by enhancing motility and reducing the accumulation of age pigment and ROS levels⁴³.

Soy isoflavones include genistein, glycitein, and daidzein. Genistein, a compound of the DrugAge, has been found to prolong the lifespan of *C. elegans* and increase its tolerance to oxidative stress⁴⁴. Gutierrez-Zepeda et al. (2005) found that *C. elegans* fed with soy isoflavone glycitein had an improved resistance towards oxidative stress⁴⁵. However, in comparison to control worms, the lifespan of *C. elegans* fed with glycitein (100 μ g/ml)

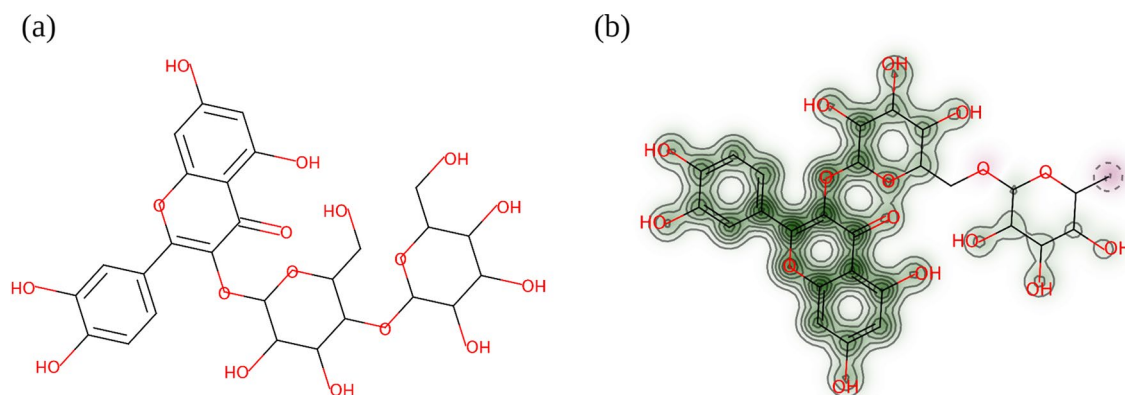


Figure 4. Similarity map for ECFP fingerprint with a default radius of 2 (a) structure of reference molecule Q3M from the DrugAge database (b) similarity map of rutin. In green colour are bits that if removed will decrease the similarity, whereas removing bits represented in pink colour will increase the similarity between the two compounds³⁸. The figure was created in the RDKit Python environment¹⁹.

was not significantly affected⁴⁵. The effect of daidzein (100 μ M) on the lifespan of *C. elegans* in the presence of pathogenic bacteria was investigated by Fischer et al. (2012)⁴⁶. The study found that daidzein had an estrogenic effect that extended the lifespan of the nematode in presence of pathogenic bacteria and heat⁴⁶. Herein, we applied the MD random forest model to predict the effect of 6''-O-malonyldaidzin on the lifespan of *C. elegans*. 6''-O-Malonyldaidzin is an o-glycoside derivative of daidzein found in food products such as soybean, miso, soy milk and soy yoghurt⁴⁷. Its predicted probability for extending the lifespan of the worm was 0.84.

Fatty acids and conjugates. Lipid metabolism has an essential role in many biological processes of an organism. Lipids are used as energy storage in the form of triglycerides and can therefore aid survival under severe conditions⁴⁸. Additionally, lipids have a key role in intercellular and intracellular signalling as well as organelle homeostasis⁴⁹. Research on both invertebrates and mammals indicates that alterations in lipid levels and composition are associated with ageing and longevity^{48,49}.

A recent review by Johnson and Stolzing (2019), on lipid metabolism and its role in ageing, summarised key lipid-related interventions that promote longevity in *C. elegans*⁵⁰. Some of the studies presented in the review are reported here. O'Rourke et al. (2013), showed that supplementing *C. elegans* with the ω -6 polyunsaturated fatty acids (PUFAs) arachidonic acid and di-homo- γ -linoleic acid (DGLA) increased the worm's starvation resistance and prolonged its lifespan by stimulating autophagy⁵¹. Similarly, Shemesh et al. (2017) found that DGLA extended the lifespan of *C. elegans* and maintained protein homeostasis in adulthood⁵². Additionally, Qi et al. (2017), found that treating *C. elegans* with ω -3 PUFA α -linolenic acid extended the nematodes lifespan⁵³. The study indicated that the ω -3 fatty acid underwent oxidation to generate a group of molecules known as oxylipins. The findings suggested that the increase of the worm's lifespan could be a result of the combined effects of the α -linolenic acid and oxylipin metabolites⁵³. Sugawara et al. (2013) found that a low dose of fish oils, which contained PUFAs eicosapentaenoic acid and docosahexaenoic acid, significantly increased the lifespan of *C. elegans*⁵⁴. The authors proposed that a low dose of fish oils induces moderate oxidative stress that extended the lifespan of the organism. In contrast, large amounts of fish oils had a diminishing effect on the worm's lifespan⁵⁴.

Gamolenic acid or γ -linolenic acid (GLA) was the second top-hit molecule of the screening database with a predictive probability of 0.95. GLA is an ω -6 PUFA, composed of an 18-carbon chain with three double bonds in the 6th, 9th and 12th position⁵⁵. Rich sources of GLA include evening primrose oil (EPO), black currant oil, and borage oil⁵⁶. In mammals, GLA is synthesized from linoleic acid (dietary) via the action of the enzyme δ -6 desaturase^{55,56}. GLA is a precursor for other essential fatty acids such as arachidonic acid^{55,56}. Conditions such as hypertension and diabetes as well as stress and various aspects of ageing, reduce the capacity of δ -6 desaturase to convert linoleic acid to GLA⁵⁷. This may lead to a deficiency of long-chain fatty acid derivatives and metabolites of GLA. GLA has been used as a constituent of anti-ageing supplements and has shown to possess various therapeutic effects in humans including improvement of age-related anomalies⁵⁵.

Sodium aurothiomalate, with a lifespan increase probability of 0.82, is a thia short-chain fatty acid used for the treatment of rheumatoid arthritis and has potential antineoplastic activities^{29,58}. In preclinical models, sodium aurothiomalate inhibited protein kinase C iota (PKC ι) signalling, which is overexpressed in non-small cell lung, ovarian and pancreatic cancers⁵⁸.

Organooxygen compounds. Lactose, with a lifespan increase probability of 0.89, is a disaccharide found in milk and other dairy product. In the human intestine, lactose is hydrolysed to glucose and galactose by the enzyme lactase. Out of the compounds in the DrugAge database, lactose had the highest structural similarity with trehalose. Trehalose has been found to increase the mean lifespan of *C. elegans* by over 30%, without showing any side effects⁵⁹. The Tanimoto coefficient between the RDKit fingerprint representations of trehalose and lactose was 0.85. Even though lactose has a high (Tanimoto) similarity to trehalose, Xing et al. (2019) found that lactose treatment at 10, 25, 50 and 100 mM concentrations shortened the lifespan of *C. elegans*⁶⁰.

Sucrose, with a lifespan increase probability of 0.83, is a disaccharide composed of glucose and fructose⁶¹. It is used as the main form of transporting carbohydrates in fruits and vegetables⁶¹. Other sugars such as trehalose, galactose and fructose have been found to extend the lifespan of *C. elegans*^{59,62,63}. Zheng et al. (2017) found the treating *C. elegans* with sucrose (55 μ M, 111 μ M, or 555 μ M) had no significant effect on the organism's mean lifespan⁶³. On the other hand, a more recent study by Wang et al. (2020) found that treatment with 50 μ M sucrose significantly increased the lifespan of the nematodes, while a concentration of 400 μ M significantly shortened their lifespan⁶⁴.

Lactulose, with a lifespan increase probability of 0.83, is a synthetic disaccharide composed of monosaccharides lactose and galactose⁶⁵. Lactulose has shown to be an effective treatment for chronic constipation in elderly patients as well as improve the cognitive function in patients with hepatic encephalopathy^{65,66}.

Other classes of compounds. Other compounds with a predictive probability of ≥ 0.80 for increasing the lifespan of *C. elegans* included alloin, a constituent of *aloe vera* with a predictive probability of 0.81, as well as the antibiotics fidaxomicin (predictive probability = 0.84), rifapentine (predictive probability = 0.81) and chlortetracycline (predictive probability = 0.80).

Rifapentine is a macrolactam antibiotic approved for the treatment of tuberculosis⁶⁷. Macrolactams are a small class of compounds that consist of cyclic amides having unsaturation or heteroatoms replacing one or more carbon atoms in the ring²⁹. Other macrolactams such as rifampicin and rifamycin have been found to increase the lifespan of *C. elegans*⁶⁸.

Golegaonkar et al. (2015) showed that rifampicin reduced AGE products and extended the mean lifespan of *C. elegans* by 60%⁶⁸. Advanced glycation end (AGE) products are formed from the non-enzymatic reaction of sugars, such as glucose, with proteins, lipids or nucleic acids⁶⁸. AGE products have been implicated in ageing and age-related diseases such as diabetes, atherosclerosis, and neurodegenerative diseases⁶⁸. The effect of two other macrolactams, rifamycin SV and rifaximin, on the lifespan of the nematode was also investigated. Rifamycin SV was found to exhibit similar activity to rifampicin, while rifaximin lacked anti-glycating activity and did not extend the lifespan of *C. elegans*. The authors suggested that the anti-glycation properties of rifampicin and rifamycin could be attributed to the presence of a para-dihydroxyl moiety, which was not present in rifaximin⁶⁸. As shown in Fig. 5, this functional group is also present in rifapentine. Experimental testing would be required to investigate whether rifapentine possesses similar properties to rifampicin and rifamycin.

Conclusions

Pharmaceutical interventions that modulate ageing-related genes and pathways are considered the most effective approach for combating human ageing and age-related diseases. Widely used strategies for identifying active compounds include screening existing drugs with potential anti-ageing properties.

In this study, the random forest algorithm was built to predict whether a compound would increase the lifespan of *C. elegans* using the entries of the DrugAge database, which contains molecules with known anti-ageing properties such as metformin, spermidine, bacitracin, and taxifolin. Our results provide an update on the findings of Barardo et al. (2017), by employing the latest version of the DrugAge database, which includes many more entries. Specifically, five random forest models were built using molecular fingerprints and/or molecular descriptors as features. Feature selection and dimensionality reduction were performed using variation and mutual information-based pre-selection methods. The best performing classifier, the MD model, was built using molecular descriptors and achieved an AUC score of 0.815 for classifying the compounds in the test set. Combining molecular descriptors with ECFPs did not further improve the model's performance. The features of the MD model were ranked using random forest's Gini importance measure. Among the 30 highest important features were molecular descriptors related to atom and bond counts, topological and partial charge properties.

The highest performing model was applied to predict the class of the compounds in the screening database which consisted of 1738 small-molecules from DrugBank. The compounds with a predictive probability of ≥ 0.80 for increasing the lifespan of *C. elegans* were broadly separated into (1) flavonoids, (2) fatty acids and conjugates, and (3) organooxygen compounds.

This study elucidated several molecules such as orange extracts, rutin, lactose and sucrose, that have been experimentally evaluated on *C. elegans* but were not entries of the predictive database. A limitation of our algorithm is that it does not consider the substance's dose. For example, at certain concentrations sugars such as sucrose can promote longevity in *C. elegans*, whereas at higher concentrations such sugars have a detrimental effect on the lifespan of the nematodes⁶⁴. Moreover, lactose, which received a predictive probability of 0.89 for increasing the lifespan of *C. elegans* by our model, was found to reduce the lifespan of *C. elegans* at 10–100 mM concentrations by Xing et al. (2019)⁶⁰. Nevertheless, the compound could have a beneficial health effect at a different concentration.

Future work would involve in vivo testing of promising compounds such as γ -linolenic acid, rutin, lactulose and rifapentine to investigate their effect on the lifespan of *C. elegans*, as well as, reevaluate the effect of lactose at lower concentrations. Finally, further work would also explore how the predicted probability of lifespan increase is affected when testing two structurally similar compounds that promote longevity at different concentrations.

Methods

Dataset for predicting lifespan-extending compounds. The dataset published in the study by Barardo et al. (2017) contains positive entries, which are compounds that "increase the lifespan of *C. elegans*" and negative entries, compounds that "do not increase the lifespan of *C. elegans*"⁴. In particular, the dataset contains 1392 compounds of which 229 are positive and 1163 are negative entries⁴. The positive entries of this dataset were obtained from the DrugAge database of ageing-related drugs, (Build 2, release date: 01/09/2016), available

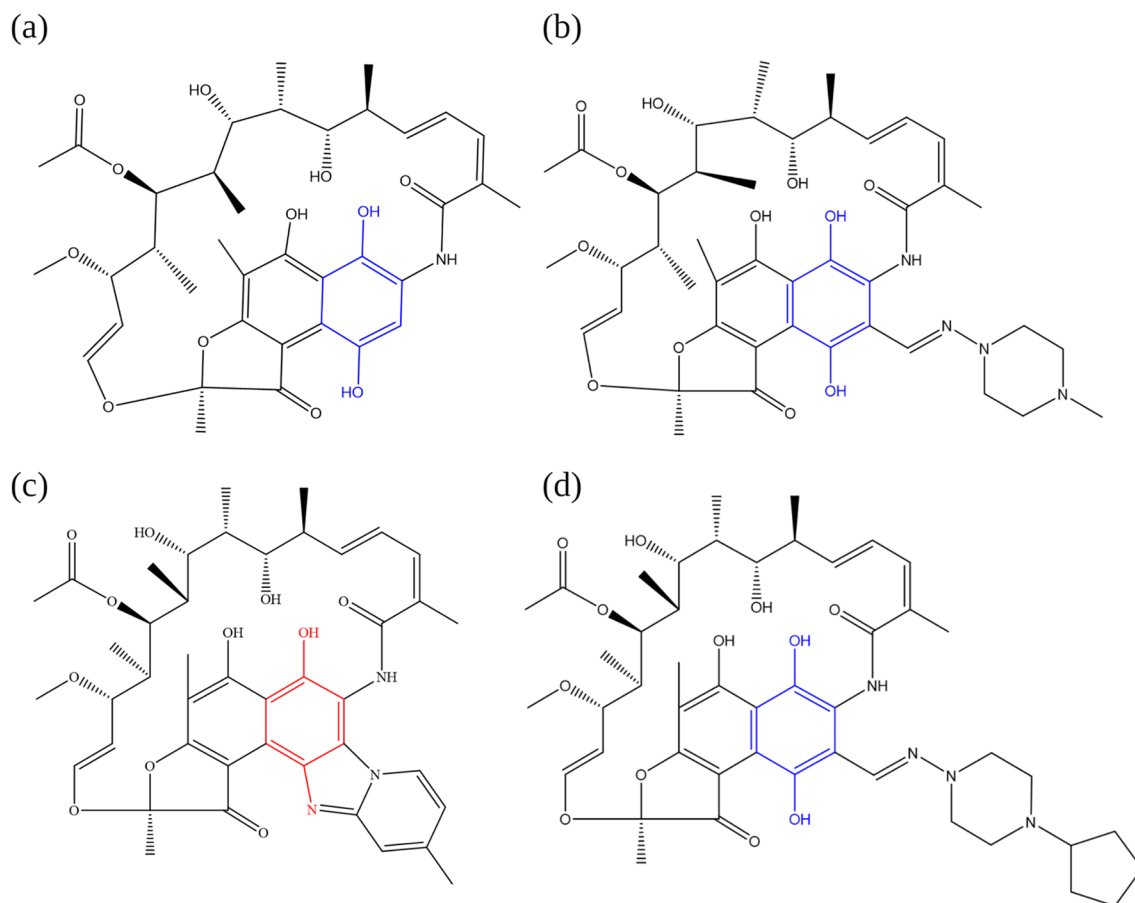


Figure 5. Chemical structure of (a) rifamycin SV (b) rifampicin (c) rifaximin and (d) rifapentine. The para-dihydroxynaphthyl moiety possessed by rifamycin SV, rifampicin and rifapentine is highlighted in blue. Rifaximin possesses a para-aminophenyl moiety incorporated in a ring system, highlighted in red⁶⁸. The figure was designed in ChemDraw and redrawn from Golegaonkar et al. (2015)⁶⁸.

on the Human Ageing Genomic Resources website^{1,69}. DrugAge provides information on drugs, compounds and supplements with anti-ageing properties that have been found to extend the lifespan of model organisms¹. The species include worms, mice and flies, with the majority of data representing *C. elegans*⁴. Data has been obtained from studies performed under standard conditions and contain information relevant to ageing, such as average/median lifespan, maximum lifespan, strain, dosage and gender where available¹. The negative entries of the database used in the study of Barardo et al. (2017) were obtained from the literature.

At the time of writing, the latest version of the DrugAge database, Build 3 (release date: 19/07/2019), corrects for small errors and adds hundreds of new entries. Herein, the positive entries in the database used in Barardo et al. (2017) were replaced with the data from the newest version of DrugAge, Build 3. The same negative entries as Barardo et al. (2017) were used⁴. The modified database contained a total of 1558 compounds with 395 positive entries and 1163 negative ones. In this study, the term “DrugAge database” refers to the modified dataset with a total of 1558 compounds.

Representation of chemical compounds. The chemical structures of the DrugAge dataset were converted into canonical SMILES strings using the Python package PubChemPy⁷⁰. The SMILES strings were standardised by the Standardiser tool developed by Francis Atkinson in 2014⁷¹. Standardisation removed inorganic compounds, salt/solvent components and metal species as well as neutralised the compounds by adding or removing hydrogen atoms⁷¹. Stereoisomers, even if biologically may have different activities, were treated as duplicates as they had identical SMILES strings. For two or more stereoisomers in the same class, only one was kept. For duplicates in different classes, both were removed⁷². After standardisation and duplicate removal, the number of molecules in the DrugAge database was reduced to a total of 1430 compounds with 304 positive and 1126 negative entries. The predictive database used in this study can be found in Additional File 2.

Molecular descriptor generation. The standardised SMILES strings were converted into mol files in the RDKit environment and opened in the MOE software^{19,23}. The chemical structures were energy minimised in the Energy Minimize General mode of MOE using Amber10:EHT force field²³. A total of 354 descriptors were calculated including all 2D, internal i3D and external x3D coordinate depended on 3D descriptors. Due to software limitation, few 3D descriptors (‘AM1_E’, ‘AM1_Eele’, ‘AM1_HF’, ‘AM1_HOMO’, ‘AM1_IP’, ‘AM1_LUMO’,

'MNDO_E', 'MNDO_Eele', 'MNDO_HF', 'MNDO_HOMO', 'MNDO_IP', 'MNDO_LUMO', 'PM3_E', 'PM3_Eele', 'PM3_HF', 'PM3_HOMO', 'PM3_IP', 'PM3_LUMO') could not be calculated for ten chemical structures. The missing values were replaced with the average value of the remaining chemical structures for the given descriptor.

Molecular fingerprint generation. Molecular fingerprints were generated in the Python RDKit environment from the standardised SMILES strings¹⁹. ECFP of 1024-bits and 2048-bits length were calculated with an atomic radius of 2. These were represented as “ECFP_1024” and “ECFP_2048”, respectively. In addition to the ECFPs, RDKit topological fingerprints of 2048-bits length were generated with a maximum path length of 5 bonds and denoted as “RDKit5”.

Five random forest models were built using five different feature types and trained with the data of the DrugAge database. The feature types explored in this study, ECFP_1024, ECFP_2048, RDKit5, MD and ECFP_1024_MD, are summarised in Supplementary Table 1, Additional File 1. The ECFP_1024_MD feature was a combined descriptor type consisting of ECFPs of 1024 bit-length and molecular descriptors.

Feature selection. Feature selection was implemented in the *scikit-learn* Python library²². Features with low variance were removed first, creating three feature subsets var_100, var_95 and var_90. The filters removed features with the same value in all entries (var_100), features that had greater than 95% of constant values (var_95) and features with more than 90% constant values, respectively (var_90)⁷³.

Mutual Information (MI) is a filter-based feature selection method used to measure the dependency between two variables. The MI between variables X and Y is given by (Eq. 3):

$$MI(X; Y) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (3)$$

where $p(x, y)$ is the joint probability mass function and $p(x)$ and $p(y)$ are the marginal probability mass functions for X and Y , respectively⁷⁴. Herein, Adjusted Mutual Information (AMI) was calculated between each feature and the class labels. AMI was used as it is a variation of MI that adjusts for chance⁷⁵.

For each of the feature subset, AMI was applied using the “adjusted_mutual_info_score” function of *scikit-learn* to order the features based on their AMI score²². The following settings were tested: using 5%, 10%, 25%, 50%, 75% and 100% of the features with the highest AMI score⁷³. For example, if var_100 for MD contained 349 features, the database with 5% of the features would consist only of the 17 highest-ranking features. This process is outlined in Supplementary Fig. 2, Additional File 1.

Tenfold cross-validation. Cross-validation was performed in the *scikit-learn* Python library using the “cross_val_score” function²². The predictive database was randomly split into 80% training and 20% test set. The tenfold cross-validation was performed only on the training set. The performance of the models was evaluated using the AUC measure. Cross-validation was repeated 10 times, yielding 10 AUC scores. The predictive accuracy reported was the median AUC value of the 10 measurements obtained by cross-validation. The median, rather than average, AUC score was calculated as the former is more robust to outliers⁴.

Random forest settings. The random forest classifiers were built in the *scikit-learn* Python module²². To handle the unbalanced data used in this study, the random forest parameter “class_weight” was set to “balanced”. The remaining parameters of the random forest classifier were set to their default settings. The models were run with 100 estimators (number of trees in the forest) and the maximum number of features considered in each tree node was the square root of the total number of features. The AUC scores were calculated with the “roc_auc_score” matrix of *scikit-learn* using the “predict_proba” method²².

Screening database. The best performing model on the test set was applied to predict the class of the compounds in an external database, where the effect of the compounds on the lifespan of *C. elegans* was mostly unknown. The external database consisted of small-molecules obtained from the External Drug Links database of DrugBank (version 5.1.5, released on 2020-01-03)¹¹. The External Drug Links database contained a list of drugs and links to other databases, such as PubChem and UniProt, providing information on these compounds^{11,76,77}.

Generation of SMILES strings, standardisation and descriptor calculation was performed in the same method used for the training (DrugAge) database, described in the above sections. Some of the entries of the DrugBank database were substances composed of more than one molecule, such as vegetable oils. These entries were either removed from the database or replaced by one of their main active ingredients. For example, “borage oil” was replaced with “gamolenic acid”. In the case of “soy isoflavones”, the major soy isoflavones (genistein, glycitein, and daidzein) had already been experimentally evaluated on the lifespan of *C. elegans*. Therefore, the entry was replaced with “6”-O-malonyldaidzin, a derivative of daidzein with unknown activity. Stereoisomers were treated as duplicates and only one of them was kept. Substances and stereoisomers present in both the DrugBank and DrugAge databases were removed from the screening database. The resulting database consisted of a total of 1738 small-molecules.

Tanimoto coefficient and similarity maps. The Tanimoto coefficients and similarity maps were computed in the Python RDKit environment¹⁹. The Tanimoto similarity is calculated between a reference molecule, which is known to be active, and a compound of interest with unknown activity.

Herein, the reference molecules were the positive entries of the DrugAge database. The compound with unknown activity was a selected entry from the screening database. The Tanimoto coefficient between the compound of interest with each of the reference molecules was calculated. The highest score achieved and the reference molecule used to obtain that score was reported. The Tanimoto coefficients were computed using the RDKit fingerprint representations of the compounds. Similarity maps were generated using ECFP fingerprint representations.

Received: 23 February 2021; Accepted: 18 June 2021

Published online: 05 July 2021

References

- Barardo, D. *et al.* The DrugAge database of aging-related drugs. *Aging Cell* **16**, 594–597 (2017).
- Qian, M. & Liu, B. Advances in pharmacological interventions of aging in mice. *Transl. Med. Aging* **3**, 116–120 (2019).
- Blagosklonny, M. V. Disease or not, aging is easily treatable. *Aging* **10**, 3067–3078 (2018).
- Barardo, D. G. *et al.* Machine learning for predicting lifespan-extending chemical compounds. *Aging* **9**, 1721–1737 (2017).
- Lee, G. D. *et al.* Dietary deprivation extends lifespan in *Caenorhabditis elegans*. *Aging Cell* **5**, 515–524 (2006).
- Harrison, D. E. *et al.* Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature* **460**, 392–395 (2009).
- Selman, C. *et al.* Ribosomal protein S6 Kinase 1 Signaling regulates mammalian life span. *Science* **326**, 140–144 (2009).
- Ye, X., Linton, J. M., Schork, N. J., Buck, L. B. & Petrascheck, M. A pharmacological network for lifespan extension in *Caenorhabditis elegans*. *Aging Cell* **13**, 206–215 (2014).
- Putin, E. *et al.* Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging* **8**, 1021–1033 (2016).
- Mamoshina, P. *et al.* Population specific biomarkers of human aging: A big data study using South Korean, Canadian, and Eastern European patient populations. *J. Gerontol. A. Biol. Sci. Med. Sci.* **73**, 1482–1490 (2018).
- Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucl. Acids Res.* **46**, D1074–D1082 (2018).
- Schlager, S., Zheng, G., Li, S. & Szekely, G. *Statistical Shape and Deformation Analysis* (Elsevier, 2017). <https://doi.org/10.1016/C2015-0-06799-5>.
- Winter, R., Montanari, F., Noé, F. & Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **10**, 1692–1701 (2019).
- Hong, H. *et al.* Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J. Chem. Inf. Model.* **48**, 1337–1344 (2008).
- Gaba, V., Rani, K. & Gupta, M. K. QSAR study on 4-alkynyldihydrocinnamic acid analogs as free fatty acid receptor 1 agonists and antidiabetic agents: Rationales to improve activity. *Arab. J. Chem.* **12**, 1758–1764 (2019).
- Roy, K., Kar, S. & Das, R. N. Chapter 2: Chemical Information and Descriptors. in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* 47–80 (Academic Press, 2015). <https://doi.org/10.1016/B978-0-12-801505-6.00002-8>.
- Lo, Y. C., Rensi, S. E., Tornig, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **23**, 1538–1546 (2018).
- Perkins, R., Fang, H., Tong, W. & Welsh, W. J. Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem.* **22**, 1666–1679 (2003).
- RDKit: Open-source cheminformatics, accessed April 2020; <http://www.rdkit.org>
- Sonego, P., Kocsor, A. & Pongor, S. ROC analysis: Applications to the classification of biological sequences and 3D structures. *Brief. Bioinform.* **9**, 198–209 (2008).
- Chen, C. & Breiman, L. *Using Random Forest to Learn Imbalanced Data* (Univ. California, 2004).
- Pedregosa, F. *et al.* Scikit-learn. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- Chemical Computing Group Inc. Molecular Operating Environment (2019.01) Montreal, Canada. (2019).
- Bender, A. & Glen, R. C. A discussion of measures of enrichment in virtual screening: Comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **45**, 1369–1375 (2005).
- Gozalbes, R. & Doucet, J. P. Application of topological descriptors in QSAR and drug design: History and new trends. *Infect. Disord. Drug Targets* **2**, 93–102 (2002).
- Guha, R. & Willighagen, E. A survey of quantitative descriptions of molecular structure. *Curr. Top. Med. Chem.* **12**, 1946–1956 (2012).
- Gasteiger, J. & Marsili, M. Iterative partial equalization of orbital electronegativity—A rapid access to atomic charges. *Tetrahedron* **36**, 3219–3228 (1980).
- Kleinoeder, T. Prediction of Properties of Organic Compounds: Empirical Methods and Management of Property Data. (PhD Thesis, University of Erlangen-Nuernberg., 2005).
- Djoubou Feunang, Y. *et al.* ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 61 (2016).
- Prasain, J. K., Carlson, S. H. & Wyss, J. M. Flavonoids and age-related disease: Risk, benefits and critical windows. *Maturitas* **66**, 163–171 (2010).
- Ayaz, M. *et al.* Flavonoids as prospective neuroprotectants and their therapeutic propensity in aging associated neurological disorders. *Front. Aging Neurosci.* **11**, 155 (2019).
- Ramelet, A. A. Venoactive drugs. In *Sclerotherapy: Treatment of Varicose and Telangiectatic Leg Veins* (eds Goldman, M. P. *et al.*) 369–377 (W.B. Saunders, 2011). <https://doi.org/10.1016/B978-0-323-07367-7.00020-0>.
- Mangoni, A. A. Drugs acting on the cerebral and peripheral circulations. In *A Worldwide Yearly Survey of New Data in Adverse Drug Reactions and Interactions* Vol. 34 (ed. Aronson, J. K.) 311–316 (Elsevier, 2012).
- Kamel, R., Abbas, H. & Fayeze, A. Diosmin/essential oil combination for dermal photo-protection using a lipid colloidal carrier. *J. Photochem. Photobiol. B Biol.* **170**, 49–57 (2017).
- Bergan, J. J., Schmid-Schönbein, G. W. & Takase, S. Therapeutic approach to chronic venous insufficiency and its complications: Place of Daflon 500 mg. *Angiology* **52**(Suppl 1), S43–S47 (2001).
- Ganeshpurkar, A. & Saluja, A. K. The pharmacological potential of rutin. *Saudi Pharm. J.* **25**, 149–164 (2017).
- Chattopadhyay, D. *et al.* Hormetic efficacy of rutin to promote longevity in *Drosophila melanogaster*. *Biogerontology* **18**, 397–411 (2017).
- Riniker, S. & Landrum, G. A. Similarity maps: A visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminform.* **5**, 43 (2013).
- Xue, Y. L. *et al.* Isolation and *Caenorhabditis elegans* lifespan assay of flavonoids from onion. *J. Agric. Food Chem.* **59**, 5927–5934 (2011).

40. Cordeiro, L. M. *et al.* Rutin protects Huntington's disease through the insulin/IGF1 (IIS) signaling pathway and autophagy activity: Study in *Caenorhabditis elegans* model. *Food Chem. Toxicol.* **141**, 111323 (2020).
41. Sun, K. *et al.* Anti-Aging effects of hesperidin on *saccharomyces cerevisiae* via inhibition of reactive oxygen species and UTH1 gene expression. *Biosci. Biotechnol. Biochem.* **76**, 640–645 (2012).
42. Fernández-Bedmar, Z. *et al.* Role of citrus juices and distinctive components in the modulation of degenerative processes: Genotoxicity, antigenotoxicity, cytotoxicity, and longevity in *drosophila*. *J. Toxicol. Environ. Heal. Part A* **74**, 1052–1066 (2011).
43. Wang, J. *et al.* Effects of orange extracts on longevity, healthspan, and stress resistance in *Caenorhabditis elegans*. *Molecules* **25**, 1–17 (2020).
44. Lee, E. B. *et al.* Genistein from *vigna angularis* extends lifespan in *caenorhabditis elegans*. *Biomol. Ther. (Seoul)* **23**, 77–83 (2015).
45. Gutierrez-Zepeda, A. *et al.* Soy isoflavone glycitein protects against beta amyloid-induced toxicity and oxidative stress in transgenic *Caenorhabditis elegans*. *BMC Neurosci.* **6**, 54 (2005).
46. Fischer, M. *et al.* Phytoestrogens genistein and daidzein affect immunity in the nematode *Caenorhabditis elegans* via alterations of vitellogenin expression. *Mol. Nutr. Food Res.* **56**, 957–965 (2012).
47. Wishart, D. S. *et al.* HMDB 4.0: The human metabolome database for 2018. *Nucl. Acids Res.* **46**, D608–D617 (2018).
48. Papsdorf, K. & Brunet, A. Linking lipid metabolism to chromatin regulation in aging. *Trends Cell Biol.* **29**, 97–116 (2019).
49. Han, S. *et al.* Mono-unsaturated fatty acids link H3K4me3 modifiers to *C. elegans* lifespan. *Nature* **544**, 185–190 (2017).
50. Johnson, A. A. & Stolzing, A. The role of lipid metabolism in aging, lifespan regulation, and age-related disease. *Aging Cell* **18**, e13048 (2019).
51. O'Rourke, E. J., Kuballa, P., Xavier, R. & Ruvkun, G. ω -6 Polyunsaturated fatty acids extend life span through the activation of autophagy. *Genes Dev.* **27**, 429–440 (2013).
52. Shemesh, N., Meshnik, L., Shpigel, N. & Ben-Zvi, A. Dietary-induced signals that activate the gonadal longevity pathway during development regulate a proteostasis switch in *caenorhabditis elegans* adulthood. *Front. Mol. Neurosci.* **10**, 254 (2017).
53. Qi, W. *et al.* The ω -3 fatty acid α -linolenic acid extends *Caenorhabditis elegans* lifespan via NHR-49/PPAR α and oxidation to oxylipins. *Aging Cell* **16**, 1125–1135 (2017).
54. Sugawara, S., Honma, T., Ito, J., Kijima, R. & Tsuduki, T. Fish oil changes the lifespan of *Caenorhabditis elegans* via lipid peroxidation. *J. Clin. Biochem. Nutr.* **52**, 139–145 (2013).
55. Khan, S. A., Haider, A., Mahmood, W., Roome, T. & Abbas, G. Gamma-linolenic acid ameliorated glycation-induced memory impairment in rats. *Pharm. Biol.* **55**, 1817–1823 (2017).
56. Knauf, V. C., Shewmaker, C., Flider, F., Emlay, D. & Ray, E. Safflower with Elevated Gamma-Linolenic Acid. US Patent 2011/0129428A1, Jun. 2, 2011. (2011).
57. Rezapour-Firouzi, S. Chapter 24: Herbal oil supplement with hot-nature diet for multiple sclerosis. In *Nutrition and Lifestyle in Neurological Autoimmune Diseases* (eds Watson, R. R. & Killgore, W. D. S.) 229–245 (Academic Press, 2017). <https://doi.org/10.1016/B978-0-12-805298-3.00024-4>.
58. De Giorgio, R. *et al.* Chronic constipation in the elderly: A primer for the gastroenterologist. *BMC Gastroenterol.* **15**, 130 (2015).
59. Honda, Y., Tanaka, M. & Honda, S. Trehalose extends longevity in the nematode *Caenorhabditis elegans*. *Aging Cell* **9**, 558–569 (2010).
60. Xing, S. *et al.* Lactose induced redox-dependent senescence and activated Nrf2 pathway. *Int. J. Clin. Exp. Pathol.* **12**, 2034–2045 (2019).
61. Yahia, E. M., Carrillo-López, A. & Bello-Perez, L. A. Carbohydrates. In *Postharvest Physiology and Biochemistry of Fruits and Vegetables* (ed. Yahia, E. M.) 175–205 (Woodhead Publishing, 2019). <https://doi.org/10.1016/B978-0-12-813278-4.00009-9>.
62. Edwards, C. *et al.* Mechanisms of amino acid-mediated lifespan extension in *Caenorhabditis elegans*. *BMC Genet.* **16**, 8 (2015).
63. Zheng, J. *et al.* Lower doses of fructose extend lifespan in *caenorhabditis elegans*. *J. Diet. Suppl.* **14**, 264–277 (2017).
64. Wang, X. *et al.* Effects of excess sugars and lipids on the growth and development of *Caenorhabditis elegans*. *Genes Nutr.* **15**, 1 (2020).
65. Rovenko, B. M. *et al.* High sucrose consumption promotes obesity whereas its low consumption induces oxidative stress in *Drosophila melanogaster*. *J. Insect Physiol.* **79**, 42–54 (2015).
66. Yang, N. *et al.* Lactulose enhances neuroplasticity to improve cognitive function in early hepatic encephalopathy. *Neural Regen. Res.* **10**, 1457–1462 (2015).
67. Munsiff, S. S., Kambili, C. & Ahuja, S. D. Rifampentine for the treatment of pulmonary tuberculosis. *Clin. Infect. Dis.* **43**, 1468–1475 (2006).
68. Golegaonkar, S. *et al.* Rifampicin reduces advanced glycation end products and activates DAF-16 to increase lifespan in *Caenorhabditis elegans*. *Aging Cell* **14**, 463–473 (2015).
69. Tacutu, R. *et al.* Human Ageing Genomic Resources: Integrated databases and tools for the biology and genetics of ageing. *Nucl. Acids Res.* **41**, D1027–D1033 (2013).
70. PubChemPy, accessed April 2020; <https://pypi.org/project/PubChemPy/>
71. Atkinson, F. L. Standardiser, accessed April 2020; <https://github.com/flatkinson/standardiser>. (2014).
72. Kotsampasakou, E. & Ecker, G. F. Predicting drug-induced cholestasis with the help of hepatic transporters—an in silico modeling approach. *J. Chem. Inf. Model.* **57**, 608–615 (2017).
73. Fehér, N. K. Exploring Predicted Drug Metabolism in in silico Toxicity Prediction. Dissertation, University of Cambridge (2018).
74. Cover, T. M. & Thomas, J. A. Entropy, relative entropy, and mutual information. in *Elements of Information Theory* 13–55 (John Wiley & Sons, 2006).
75. Vinh, N. X., Epps, J. & Bailey, J. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary? in *Proceedings of the 26th Annual International Conference on Machine Learning* 1073–1080 (Association for Computing Machinery, 2009). <https://doi.org/10.1145/1553374.1553511>.
76. Kim, S. *et al.* PubChem 2019 update: Improved access to chemical data. *Nucl. Acids Res.* **47**, D1102–D1109 (2018).
77. Consortium, T. U. UniProt: A worldwide hub of protein knowledge. *Nucl. Acids Res.* **47**, D506–D515 (2018).

Acknowledgements

We are grateful to the members of the Department of Chemistry at the University of Surrey for their support throughout the study.

Author contributions

S.K. and B.H. wrote the main manuscript text. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interest.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-93070-6>.

Correspondence and requests for materials should be addressed to B.J.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021