Random forest classification for predicting lifespan-extending chemical compounds

Sofia Kapsiani
University of Surrey
Brendan J. Howlin (b.howlin@surrey.ac.uk)
University of Surrey

Research Article

Keywords: ageing, anti-ageing drugs, lifespan extension, DrugAge, C. elegans, machine learning, random forest, molecular descriptors, molecular fingerprints, QSAR

Posted Date: March 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-118087/v2

Version of Record: A version of this preprint was published at Scientific Reports on July 5th, 2021. See the published version at https://doi.org/10.1038/s41598-021-93070-6.

Abstract

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 (AUC) score 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 1,738 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 (i) flavonoids, (ii) fatty acids and conjugates, and (iii) organooxygen compounds.

Introduction

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 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]. Harrison *et al.* (2009) showed that rapamycin, an inhibitor of the mTOR pathway, extended the lifespan of both female and male mice[6]. In the same year, Selman *et al.* (2009) reported that genetic deletion of S6 protein kinase 1, a components of the mTOR signalling pathway, increased the lifespan of mice and protected against age-related conditions[7].

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 signalling[8]. Additionally, Putin *et al.* (2016) developed a deep learning neural network that predicted human chronological age from most critical blood markers for determining

chronological age in humans, which were albumin, glucose, alkaline phosphatase, urea and erythrocytes[9]. Moreover, 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 subjects[10]. The findings showed that population-specific ageing clocks were more accurate in predicting chronological age and quantifying biological age than generic ageing clocks[10].

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.

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 properties[4]. 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 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 DrugAge database. The filter-based feature selection method, mutual information, was employed to select the most relevant features. To the best of our knowledge, this is the first application of molecular fingerprints for building a machine learning model 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 1,738 small-molecules.

Random forest models

Random forests are an ensemble of decision trees, where each tree is trained independently using a bootstrap sample and a subset of the features available. This supervised machine learning method was selected as it is robust to overfitting in high-dimensional databases with a small number of entries[4].

The choice of chemical descriptors can significantly impact on the quality and predictions of the QSAR models. Descriptors represent chemical information of the molecules in a digital or numerical way that is suitable for model development and are computer-interpretable[11,12]. In this study, 2D and 3D molecular descriptors were calculated using the Molecular Operating Environment (MOE™) software[13]. 2D descriptors are calculated from the 2D structure of a molecule and provide information related to its structural, topological and physicochemical properties[14]. 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[12,15,16].

Molecular fingerprints are a digital representation of a molecule's structure using binary vectors, where 1 corresponds to a particular feature being present and 0 that it is absent. Herein, extended-connectivity fingerprints (ECFP) of 1,024- and 2,048-bit lengths and RDKit topological fingerprints of 2,048-bit length were generated in the RDKit Python environment[17]. Lastly, the combination of molecular descriptors with ECFPs was tested.

Results And Discussion

Visualization of chemical space

This study involved high-dimensional datasets containing hundreds of molecular fingerprints and descriptors. The PCA algorithm was applied to reduce the chemical space into two-dimensions. The chemical space representations for the ECFP, RDKit fingerprints, molecular descriptors and the combination of ECFP with molecular descriptors produced using the PCA algorithm are shown in Fig. 1.

In chemical space visualisation, structural analogues are positioned nearer to each other than to unrelated compounds[18]. This allows clustering techniques, such as PCA, do identify neighbourhoods with similarly structured molecules[18]. Thus, some degree of clustering was expected to be observed between active compounds.

Among the single descriptor types, shown in Fig. 1(a-d), the highest degree of clustering between active molecules was observed in the chemical space visualisation of the molecular descriptors. An explanation is that the chemical fingerprints used in this study were hashed fingerprints. Hashed fingerprints often involve loss of information due to bit collisions, thus, the distances between the fingerprints may not perfectly correlate to the similarity of the compounds[19]. Interestingly, the chemical space visualisation of the combined feature type, Fig. 1e, is very similar to that of the molecular descriptors shown in Fig. 1d. This indicates that the molecular descriptors have a stronger expressive power than the ECFPs of 1,024-bit length for the chemical space analysis of the DrugAge database.

Feature selection

Feature selection was employed to select the most relevant features for predicting the activity of a molecule in the database. This was performed only for the training set which contained 80% of compounds in the dataset. Feature selection was achieved by applying variance and mutual information-based pre-selection methods. This reduced the number of features used by each model, making computational calculations less expensive. The median AUC scores and standard deviation of 10-fold cross-validation obtained by random forest classification for each feature combination can be found in Supplementary Fig. 1, Additional File 1. For each descriptor type, the feature combination 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 combinations achieved the same AUC score, the combination that had the smallest 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 10-fold cross-validation and on classifying the compounds in the test set are shown in Fig. 2. A summary of the performances of the classifiers on the test and train set as well as the optimal number of variables obtained in feature selection is shown in Supplementary Table 1, Additional File 1.

As illustrated in Fig. 2, the predictive performances of the random forest models 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.

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. 3, compare the performances of the descriptor types for classifying the samples of the test set. Analysis of the ROC curves indicates 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 using only molecular descriptors (MD), had the greatest ability to correctly predict the class of the compounds in the test set. Combining MD with ECFP_1024, the feature type used to obtain the model with second-highest predictive ability, did not result in a model with a higher AUC score. 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.

Confusion matrix

The confusion matrix of the MD model for predicting the class of the molecules in the test set is shown in Fig. 4. The classification accuracy of the model was 0.853 and the AUC score was 0.815.

$$\mathrm{PPV} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\mathrm{PPV} = \frac{21}{21 + 11} = 65.6\ \% \qquad \text{(Equation 1)}$$

$$\mathrm{NPV} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}}$$

$$\mathrm{NPV} = \frac{223}{223 + 31} = 87.8\ \% \qquad \text{(Equation 2)}$$

In binary classification, the PPV and NPV are the percentage of positive and negative values, respectively, that are correctly classified. 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]. This improved the performance of the model, as the PPV for classifying the compounds of the test set increased from 61.1% (value without balancing the class weights) to 65.6% (score achieved after balancing the class weights).

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™ software[22]. 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 1), which contained both 2D and 3D molecular descriptors.

Table 1
Top 30 features ranked by Gini importance for the MD random forest model. The description of the features was taken from the MOE™ software documentation[22].

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

Gini importance	Feature	Description
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

The highest-ranking features were broadly separated into the following categories (i) atom and bond counts (ii) topological and (iii) 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 drug[23]. 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 database[4]. Nitrogen atoms could have affected the physicochemical properties of the drugs as well as the interactions and binding of the molecules with target residues.

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 edges[24]. These descriptors can provide information on the degree branching of the structure as well as molecular size and shape[24]. Although topological descriptors are extensively used in predictive modelling, they are usually hard to

interpret[25]. 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[26,27]. On the other hand, descriptors prefixed with “Q_” were calculated using the Amber10:EHT force field[22]. 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 1,738 small-molecules obtained from the DrugBank database[28]. The top-ranking compounds with a predictive probability of $\geq$ 0.080 for increasing the lifespan of *C. elegans* are shown in Table 2. 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; (i) flavonoids, (ii) fatty acids and conjugates, and (iii) organooxygen compounds. The compound classification was taken from the category “Class” in the chemical taxonomy section of the DrugBank database (provided by Classyfire) or assigned manually if not available[29].

Table 2 Chemical compounds from the screening database with a predictive probability of 0.80 or above for increasing the of *C. elegans.*

Compound name	Predictive probability
Diosmin	0.96
Gamolenic acid	0.95
Rutin	0.95
Hesperidin	0.94
Lactose	0.89
6''-O-Malonyldaidzin	0.84
Fidaxomicin	0.84
Sucrose	0.83
Lactulose	0.83
Sodium aurothiomalate	0.82
Aloin	0.81
Rifapentine	0.81
Plecanatide	0.80
Calcifediol	0.80
Chlortetracycline	0.80

Flavonoids

Flavonoids are a group of secondary metabolites in plants that are common polyphenols in the human diet[30]. 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 isoflavones[30].

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 sensitivity [30].

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 synthetically [32]. 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 lymphedema[33]. Nevertheless, diosmin has a poor aqueous solubility, which is a challenge for oral administration[34].

 with essential oils showed skin antioxidant, anti-

ageing and sun-blocking effects on mice [34]. 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→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 Figure 5.

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 had no affect the worm's lifespan[39]. Davalli *et al.* (2016) also reported that rutin did not improve the longevity of *C. elegans* [40]. On the other hand, Chattopadhyay *et al.* (2017) showed the rutin promoted longevity in a species of fly, *Drosophila melanogaster (D. melanogaster)*[37].

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 levels[43].

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 stress [44]. Gutierrez-Zepeda *et al.* (2005) found that *C. elegans* fed with soy isoflavone glycitein had an improved resistance towards oxidative stress[45]. However, in comparison to control worms, the lifespan of *C. elegans* fed with glycitein was not significantly affected[45]. The effect of daidzein 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 which extended the worm's lifespan in presence of pathogenic bacteria and heat[46]. 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 yoghurt[47]. 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 conditions[48]. Additionally, lipids have a key role in intercellular and intracellular signalling as well as organelle homeostasis[49]. Research on both invertebrates and mammals suggest that alteration 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. In response to fasting 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 increased the worm's starvation resistance and prolonged its lifespan by stimulating autophagy[51]. Similarly, Qi *et al.* (2017), found that treating *C. elegans* with ω-3 PUFA -linolenic acid in dose-dependent manner extended the worm's lifespan[52]. 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 the worm's lifespan could be a result of the combined effects of the α-linolenic acid and oxylipin metabolites[52]. 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*[53]*.* 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 lifespan[53].

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 position[54]. Rich sources of GLA include evening primrose oil (EPO), black currant oil, and borage oil[55]. In mammals, GLA is synthesized from linoleic acid (dietary) via the action of the enzyme σ -6 desaturase[54,55]. GLA is a precursor for other essential fatty acids such as arachidonic acid[54,55]. Conditions such as hypertension and diabetes as well as stress and various aspects of ageing, reduce the capacity of -6 desaturase to convert linoleic acid to GLA[56]. 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 anomalies[54].

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[29,57]. In preclinical models, sodium aurothiomalate inhibited protein kinase C iota (PKCι) signalling, which is overexpressed in non-small cell lung, ovarian and pancreatic cancers[57].

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 effects[58]. 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 shortened the lifespan of *C. elegans*[59].

Sucrose, with a lifespan increase probability of 0.83, is a disaccharide composed of glucose and fructose[60]. It is used as the main form of transporting carbohydrates in fruits and vegetables[60]. Other sugars such as trehalose, galactose and fructose have been found to extend the lifespan of *C. elegans*[58,61,62]. However, Zheng *et al.* (2017) found the treating *C. elegans* with sucrose had no significant effect on the organism's mean lifespan[62]. In rats, sucrose has been found to shorten the mean lifespan and elevate the blood pressure[63]. Rovenko *et al.* (2015) showed that in *D. melanogaster,* high sucrose consumption decelerated pupation, increased pupa mortality and promoted obesity[64].

Lactulose, with a lifespan increase probability of 0.83, is a synthetic disaccharide composed of monosaccharides lactose and galactose[64]. Lactulose has been to be an effective treatment for chronic constipation in elderly patients as well as improve the cognitive function in patients with hepatic encephalopathy[64,65].

Other classes of compounds

Other compounds with a predictive probability ≥ 0.80 for increasing the lifespan of *C. elegans* included aloin, 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 tuberculosis[66]. Macrolactams are a small class of compounds which consist of cyclic amides having unsaturation or heteroatoms replacing one or more carbon atoms in the ring[29]. Other macrolactams such as rifampicin and rifamycin have been found to increase the lifespan of *C. elegans*[67].

Golegaonkar *et al.* (2015) showed that rifampicin reduced AGE products and extended the mean lifespan of *C. elegans* by 60%[67]. Advanced glycation end (AGE) products are formed from the non-enzymatic reaction of sugars, such as glucose, with proteins, lipids or nucleic acids[67]. AGE products have been implicated in ageing and age-related diseases such as diabetes, atherosclerosis, and

ns, rifamycin SV and rifaximin, on the worm's

lifespan 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 rifaximin[67]. As shown in Figure 6, this functional group is also present in rifapentine. Experimental testing would be required to investigate whether rifapentine possess similar properties to rifampicin and rifamycin.

Evaluation of the chemical similarity principle

Several of the compounds identified by the random forest model had already been experimentally evaluated for increasing the lifespan of *C. elegans* and other model organisms. In particular, the RDKit fingerprints of rutin are 0.99 (Tanimoto) similar to that of Q3M, an active compound. However, experimental studies found that although it is structurally similar to active compounds, rutin does not extend the lifespan of *C. elegans*[39,40]. Additionally, the Tanimoto coefficient between the RDKit fingerprint representations of lactose and trehalose, an active compound, is 0.85. Nevertheless, *in vivo* studies showed that treatment with lactose reduced the lifespan of *C. elegans*[59]. In these cases, the chemical similarity principle, which states that chemically similar compounds tend to have similar bioactivities, appears to fail. An explanation presented by Martin *et al.* (2002) is that protein structures are complex and flexible systems[68]. Thus, structurally similar chemicals may bind in different orientations to the active site, interact with a different conformation of the protein or even bind to completely different proteins[68].

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 activities.

In this study, the random forest algorithm was applied to analyse the DrugAge database and predict whether a compound would increase the lifespan of *C. elegans*. Five different 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, used 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 1,738 small-molecules from DrugBank. The compounds with a predictive probability of ≥ 0.80 for increasing the lifespan of *C. elegans* were broadly separated into (i) flavonoids, [illegible]mpounds. 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. Future work would include *in vivo* testing of promising compounds such as γ–linolenic acid, aloin and rifapentine to investigate their effect on the lifespan of *C. elegans.*

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 1,392 compounds of which 229 are positive and 1,163 are negative entries[4]. The positive entries of this dataset were obtained from DrugAge database of ageing-related drugs, (Build 2, release date: 01/09/2016), available in 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 organisms[1]. 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 available[1]. 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 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 used [4]. The modified database contained a total of 1,558 compounds with 395 positive entries and 1,163 negative ones. In this study, the term "DrugAge database" refers to the modified dataset with a total of 1,558 compounds.

Representation of chemical compounds

The chemical structures of the DrugAge dataset were converted into canonical SMILES strings using the Python package PubChemPy[70]. The SMILES strings were standardised by the Standardiser tool developed by Francis Atkinson in 2014 [71]. Standardisation removed inorganic compounds, salt/solvent components and metal species as well as neutralised the compounds by adding or removing hydrogen atoms[71]. 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 removed[72]. After standardisation and duplicate removal, the number of molecules in DrugAge database was reduced to a total of 1,430 compounds with 304 positive and 1,126 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[17,22]. The chemical structures were energy minimised in the Energy Minimize General mode of MOE™ using Amber10:EHT force field[22]. 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 strings [17]. ECFP of 1,024-bits and 2,048-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 2,048-bits length were generated with a maximum path length of 5 bonds and denoted as “RDKit5”.

Five random forest models were build 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 2, Additional File 1. The ECFP_1024_MD feature was a combined descriptor type consisting of ECFPs of 1,024 bit-length and molecular descriptors.

Feature selection

Feature selection was implemented in the *scikit-learn* Python library[73]. Features with low variance were removed first, creating three sub-databases 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)[74].

For each of the sub-databases, Adjusted Mutual Information (AMI) was applied using the “adjusted_mutual_info_score” function of *scikit-learn* to order the features based on their AMI score [73]. The following settings were tested: using 5%, 10%, 25%, 50%, 75% and 100% of the features with the highest AMI score [74]. 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.

10-fold Cross-validation

Cross-validation was performed in the *scikit-learn* Python library using the “cross_val_score” function[73]. The predictive database was randomly split into 80% training and 20% test set. The 10-fold 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 outliers[4].

Random forest settings

The random forest classifiers were built in the *scikit-learn* Python module[73]. 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 "roc_auc_score" matrix of *scikit-learn* using the "predict_proba" method[73].

Chemical space implementation

The 2D representations of the chemical space were generated by applying the PCA algorithm in the Python *scikit-learn* library [73]. Visualisation of molecular descriptors required feature scaling as the descriptors had different ranges. Scale difference can negatively impact the performance of the PCA model, as it incorrectly considers some features as more important than others. The resulting molecular descriptors had a standard normal distribution with a mean of zero and a standard deviation of one [73]. Feature scaling was not required for the molecular fingerprints they only consisted of binary values.

Screening database

The best performing model 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)[28]. 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[28,75,76].

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 by more than one molecule, such as vegetable oils. These entries where either removed from the database or replaced by their 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

Tanimoto coefficient and similarity maps

The Tanimoto coefficients and similarity maps were computed in the Python RDKit environment[17]. 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.

Declarations

Acknowledgements

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

Authors' contributions

BJH designed and supervised the study. SK performed data curation, built the predictive models and wrote the manuscript. BJH aided the interpretation of the findings and reviewed the manuscript providing improvements.

Competing interests

The authors declare that there are no conflicts of interest.

Additional information

Supplementary Information can be found in the Additional File 1 and Additional File 2.

Correspondence requests for materials should be addressed to BJH.

References

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

2. Qian, M. & Liu, B. Advances in pharmacological interventions of aging in mice. *Transl. Med. Aging.* **3**, 116–120 (2019).
3. Blagosklonny, M. V. Disease or not, aging is easily treatable. *Aging (Albany. NY).* **10**, 3067–3078 (2018).
4. Barardo, D. G. *et al.* Machine learning for predicting lifespan-extending chemical compounds. *Aging (Albany. NY).* **9**, 1721–1737 (2017).
5. Lee, G. D. *et al.* Dietary deprivation extends lifespan in Caenorhabditis elegans. *Aging Cell.* **5**, 515–524 (2006).
6. Harrison, D. E. *et al.* Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature.* **460**, 392–395 (2009).
7. Selman, C. *et al.* Ribosomal Protein S6 Kinase 1 Signaling Regulates Mammalian Life Span. *Science (80-.).* **326**, 140–144 (2009).
8. 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).
9. Putin, E. *et al.* Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging (Albany. NY).* **8**, 1021–1033 (2016).
10. 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).
11. 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).
12. Hong, H. *et al.* Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J. Chem. Inf. Model.* **48**, 1337–1344 (2008).
13. Rinnie, Gaba, V., Rani, K., Gupta, M. K. & Shilpa & 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).
14. 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). doi:https://doi.org/10.1016/B978-0-12-801505-6.00002-8.
15. Lo, Y. C., Rensi, S. E., Torng, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today.* **23**, 1538–1546 (2018).
16. 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).
17. RDKit Open-source cheminformatics. http://www.rdkit.org. Accessed April 2020.
18. Naveja, J. J. & Medina-Franco, J. L. Finding Constellations in Chemical Space Through Core Analysis. *Front. Chem.* **7**, 510 (2019).

19. Hirohara, M., Saito, Y., Koda, Y., Sato, K. & Sakakibara, Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinformatics.* **19**, 526 (2018).
20. Sonego, P., Kocsor, A. & Pongor, S. ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief. Bioinform.* **9**, 198–209 (2008).
21. Chen, C. & Breiman, L. Using Random Forest to Learn Imbalanced Data. *Univ. California, Berkeley* (2004).
22. Chemical Computing Group Inc. Molecular Operating Environment (2019.01) Montreal, Canada. (2019).
23. 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).
24. Gozalbes, R. & Doucet, J. P. D. F. Application of Topological Descriptors in QSAR and Drug Design: History and New Trends. *Infect. Disord. Drug Targets.* **2**, 93–102 (2002).
25. Guha, R. & Willighagen, E. A survey of quantitative descriptions of molecular structure. *Curr. Top. Med. Chem.* **12**, 1946–1956 (2012).
26. Gasteiger, J. & Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron.* **36**, 3219–3228 (1980).
27. Kleinoeder, T. Prediction of Properties of Organic Compounds - Emperical Methods and Management of Property Data. (PhD Thesis, University of Erlangen-Nuernberg 2005).
28. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
29. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 61 (2016).
30. Prasain, J. K., Carlson, S. H. & Wyss, J. M. Flavonoids and age-related disease: risk, benefits and critical windows. *Maturitas.* **66**, 163–171 (2010).
31. Ayaz, M. *et al.* Flavonoids as Prospective Neuroprotectants and Their Therapeutic Propensity in Aging Associated Neurological Disorders. *Front. Aging Neurosci.* **11**, 155 (2019).
32. Ramelet, A. A. Venoactive Drugs. in Sclerotherapy: Treatment of Varicose and Telangiectatic Leg Veins (eds. Goldman, M. P., Guex, J. J. & Weiss, R. A.) 369–377(W.B. Saunders, 2011). doi:https://doi.org/10.1016/B978-0-323-07367-7.00020-0.
33. 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* (ed. Aronson, J. K.) vol. 34 311–316 (Elsevier, 2012).
34. Kamel, R., Abbas, H. & Fayez, A. Diosmin/essential oil combination for dermal photo-protection using a lipoid colloidal carrier. *J. Photochem. Photobiol. B Biol.* **170**, 49–57 (2017).

35. 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), S43–7 (2001).
36. Ganeshpurkar, A. & Saluja, A. K. The Pharmacological Potential of Rutin. *Saudi Pharm. J.* **25**, 149–164 (2017).
37. Chattopadhyay, D. *et al.* Hormetic efficacy of rutin to promote longevity in Drosophila melanogaster. *Biogerontology.* **18**, 397–411 (2017).
38. Riniker, S. & Landrum, G. A. Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminform.* **5**, 43 (2013).
39. Xue, Y. L. *et al.* Isolation and Caenorhabditis elegans lifespan assay of flavonoids from onion. *J. Agric. Food Chem.* **59**, 5927–5934 (2011).
40. Davalli, P., Mitic, T., Caporali, A., Lauriola, A. & D'Arca, D. R. O. S. Cell Senescence, and Novel Molecular Mechanisms in Aging and Age-Related Diseases. Oxid. Med. Cell. Longev. 2016, 3565127 (2016).
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. *Nucleic 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. Qi, W. *et al.* The ω-3 fatty acid α-linolenic acid extends Caenorhabditis elegans lifespan via NHR-49/PPARα and oxidation to oxylipins. *Aging Cell.* **16**, 1125–1135 (2017).
53. 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).
54. 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).
55. 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).
56. 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). doi:https://doi.org/10.1016/B978-0-12-805298-3.00024-4.
57. De Giorgio, R. *et al.* Chronic constipation in the elderly: a primer for the gastroenterologist. *BMC Gastroenterol.* **15**, 130 (2015).
58. Honda, Y., Tanaka, M. & Honda, S. Trehalose extends longevity in the nematode Caenorhabditis elegans. *Aging Cell.* **9**, 558–569 (2010).
59. Xing, S. *et al.* Lactose induced redox-dependent senescence and activated Nrf2 pathway. *Int. J. Clin. Exp. Pathol.* **12**, 2034–2045 (2019).
60. 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). doi:https://doi.org/10.1016/B978-0-12-813278-4.00009-9.
61. Edwards, C. *et al.* Mechanisms of amino acid-mediated lifespan extension in Caenorhabditis elegans. *BMC Genet.* **16**, 8 (2015).
62. Zheng, J. *et al.* Lower Doses of Fructose Extend Lifespan in Caenorhabditis elegans. *J. Diet. Suppl.* **14**, 264–277 (2017).
63. Preuss, H. G. *et al.* Effects of excess sucrose ingestion on the life span of hypertensive rats (SHR). *Geriatr. Nephrol. Urol.* **1**, 13–20 (1991).
64. 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).
65. Yang, N. *et al.* Lactulose enhances neuroplasticity to improve cognitive function in early hepatic encephalopathy. *Neural Regen. Res.* **10**, 1457–1462 (2015).
66. Munsiff, S. S., Kambili, C. & Ahuja, S. D. Rifapentine for the Treatment of Pulmonary Tuberculosis. *Clin. Infect. Dis.* **43**, 1468–1475 (2006).
67. 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).
68. Martin, Y. C., Kofron, J. L. & Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **45**, 4350–4358 (2002).

69. Tacutu, R. *et al.* Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.* **41**, D1027–D1033 (2013).
70. PubChemPy. https://pypi.org/project/PubChemPy/. Accessed April 2020.
71. Atkinson, F. L. (2014) Standardiser. https://github.com/flatkinson/standardiser. Accessed on April 2020.
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. Pedregosa, F. *et al.* Scikit-learn. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
74. Fehér, N. K. (2018) Exploring Predicted Drug Metabolism in in silico Toxicity Prediction. Dissertation, University of Cambridge.
75. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2018).
76. Consortium, T. U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2018).

Figures

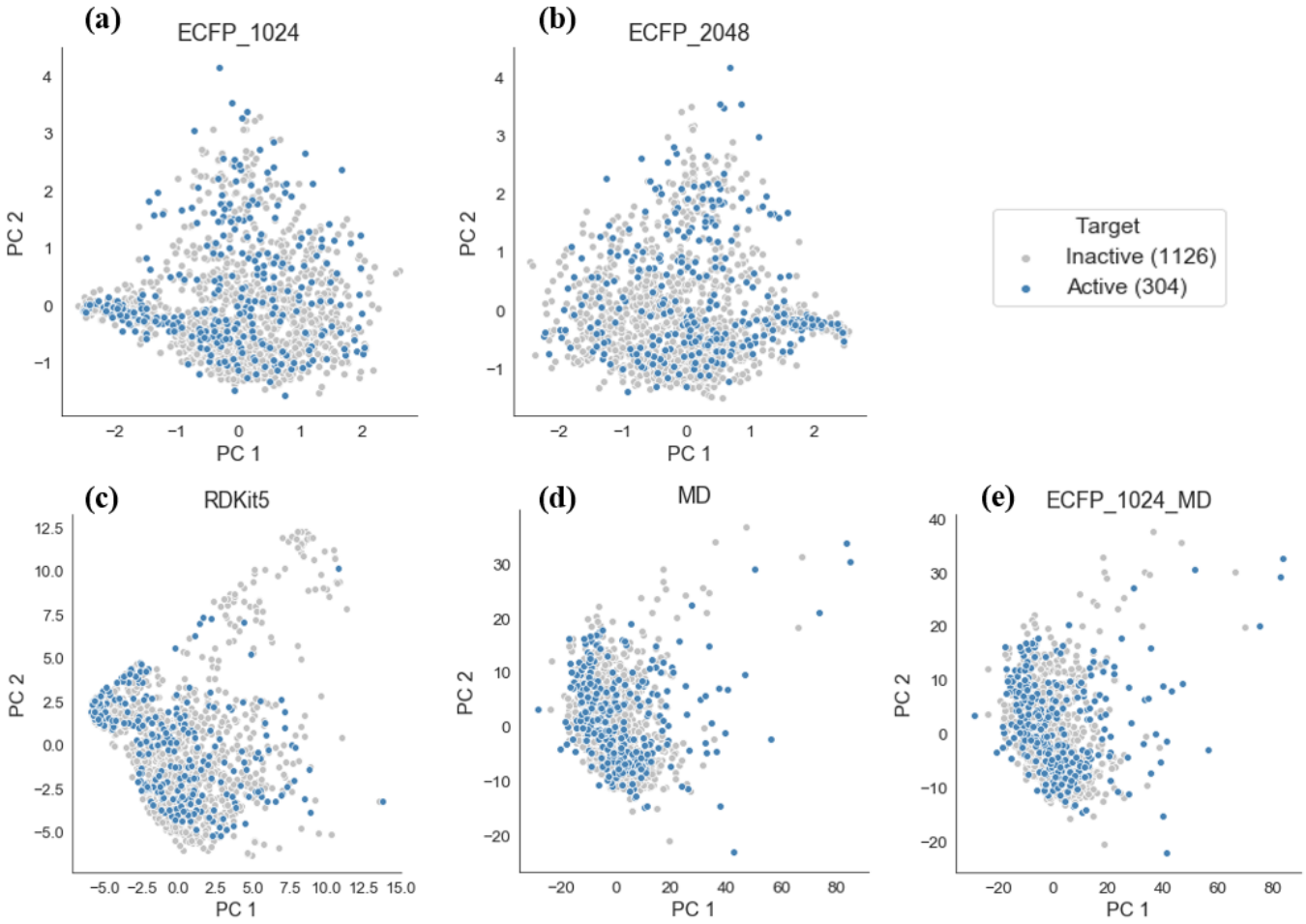

Figure 1

Visual representation of chemical space for 1,430 compounds in the DrugAge dataset computed using the PCA algorithm with different chemical representations: (a) ECFP of 1,024-bit length, (b) ECFP of 2,048-bit length, (c) RDKit fingerprints of 2,048-bit length, (d) molecular descriptors and (e) combination of ECFP of 1,024-bit length with molecular descriptors. The active compounds are represented in blue colour while the inactive molecules are shown in grey.

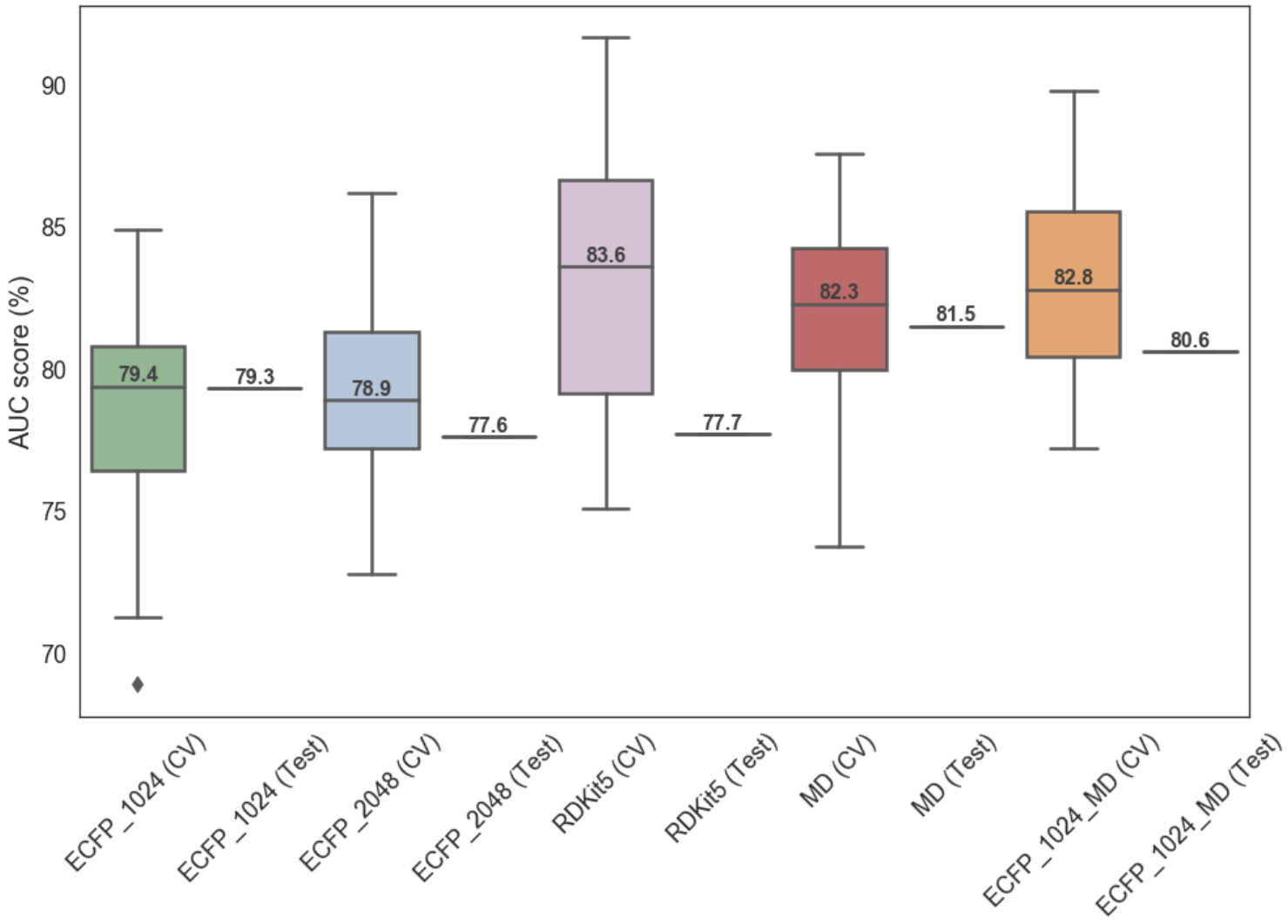

Figure 2

Box-and-whisker plot displaying the distribution of the AUC scores for the 10-fold cross-validation (CV) 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 1,024-bit length is shown in green, ECFP of 2,048-bit length in blue, RDKit fingerprints in pink, molecular descriptors in red and the combination of ECFP of 1,024-bit length with molecular descriptors are represented in orange colour. The value reported within the boxes is the median AUC value of the 10-fold cross-validation. The points outside the boxplots represent possible outliers.

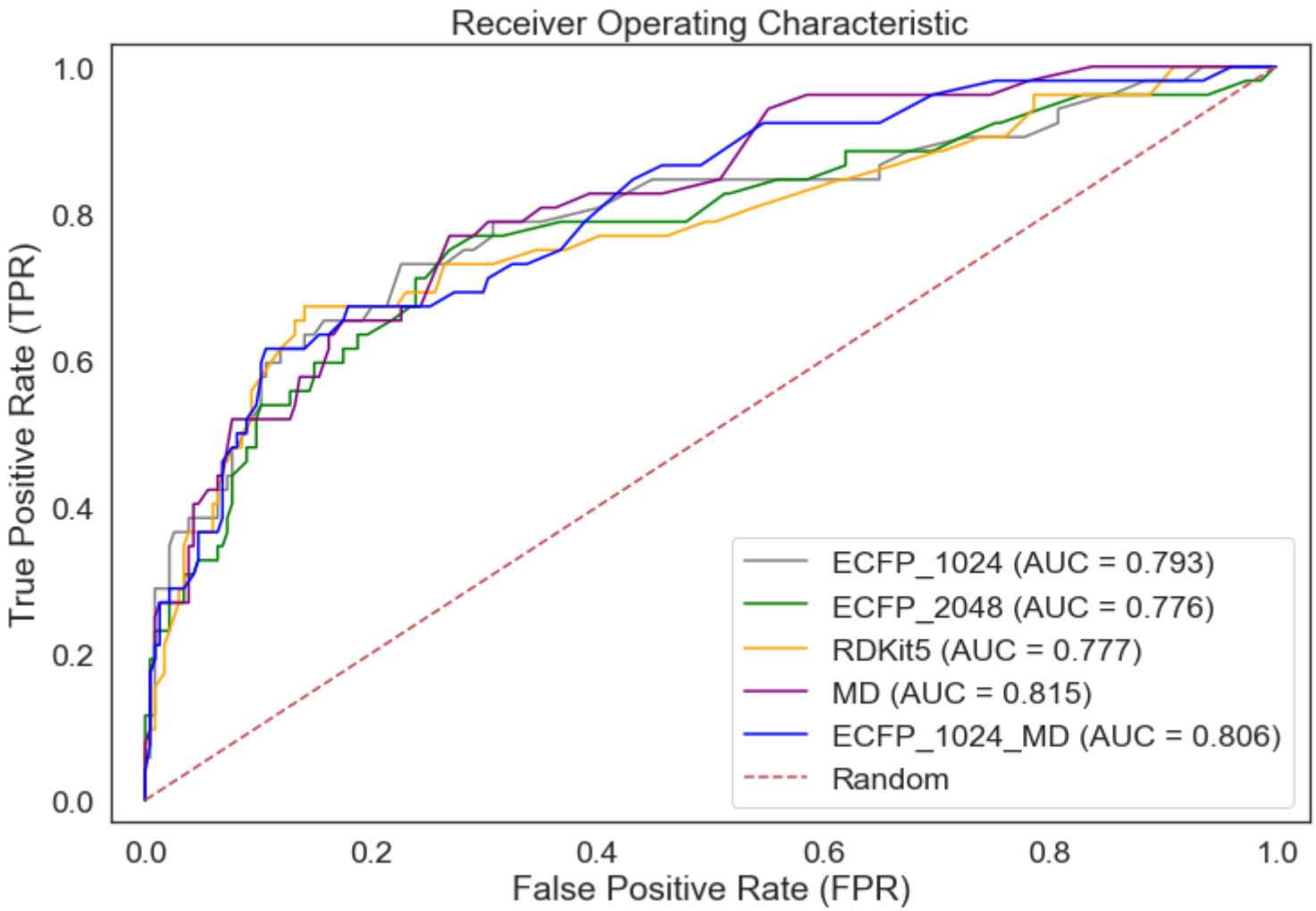

Figure 3

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.

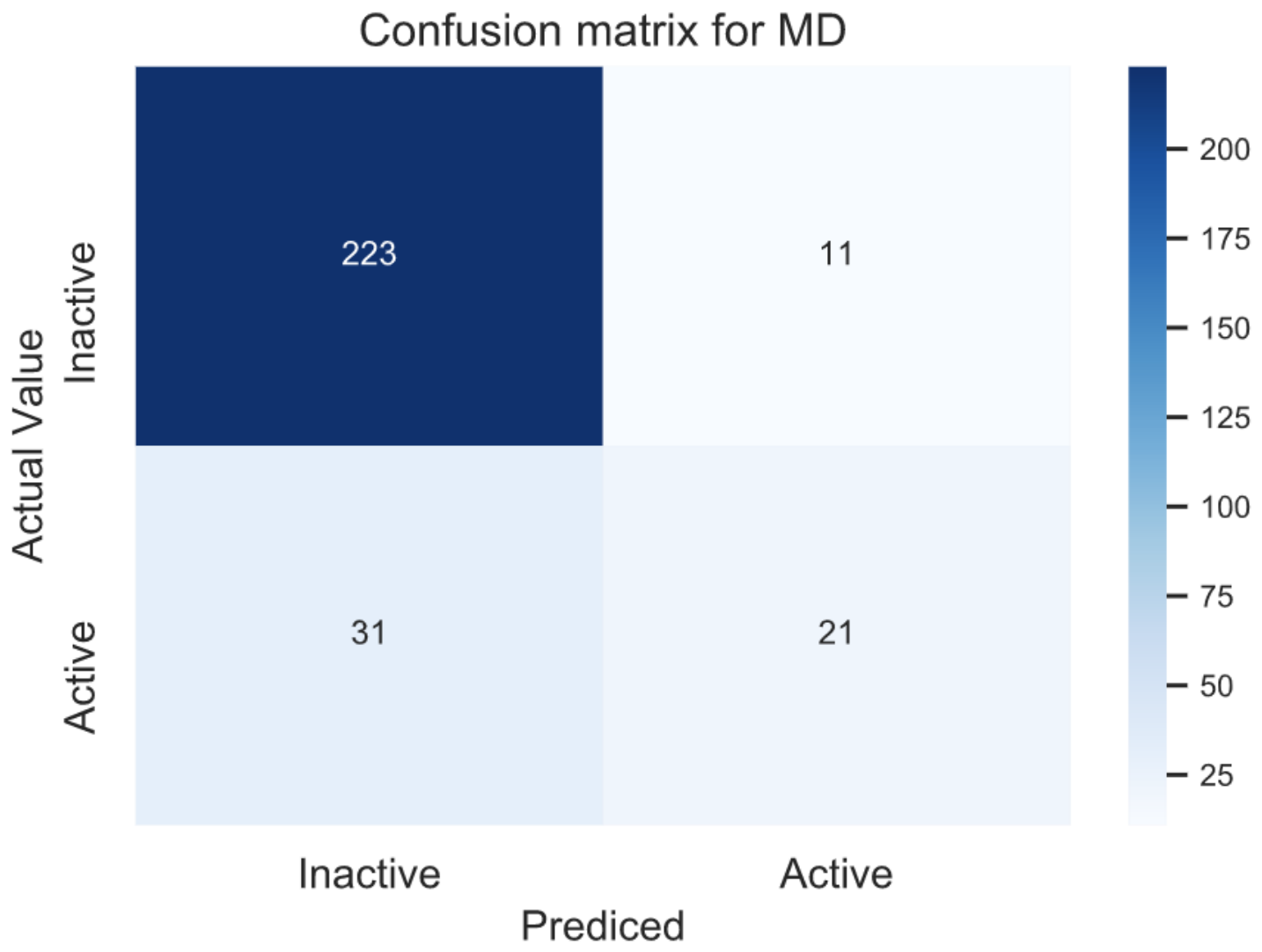

Figure 4

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.

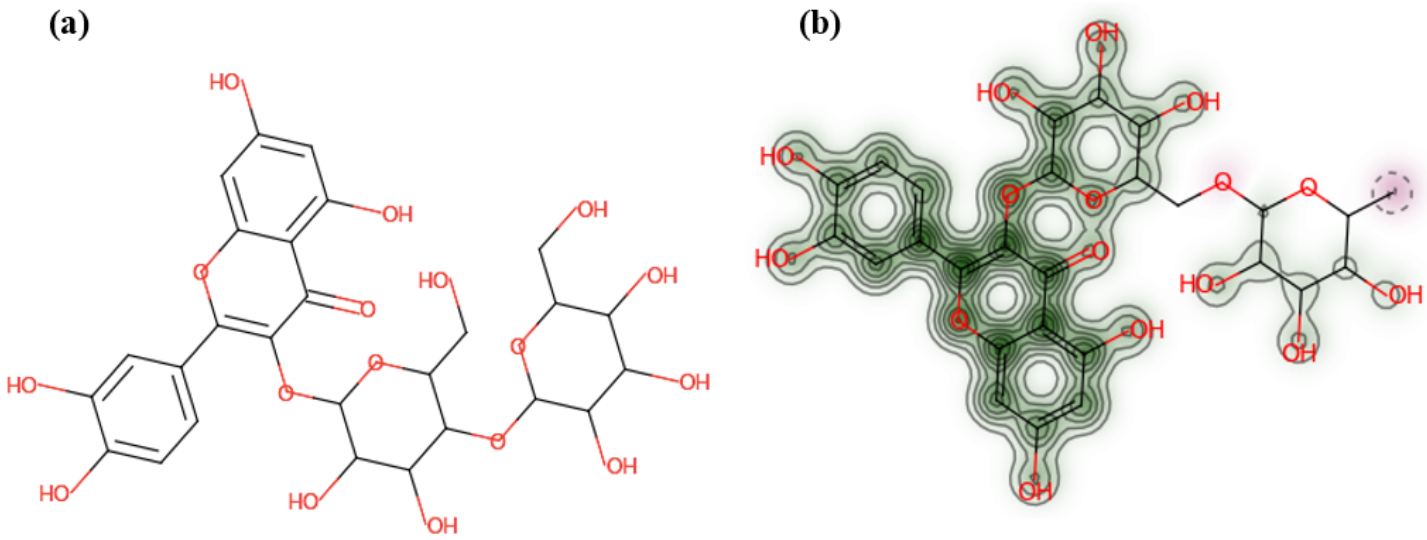

Figure 5

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.

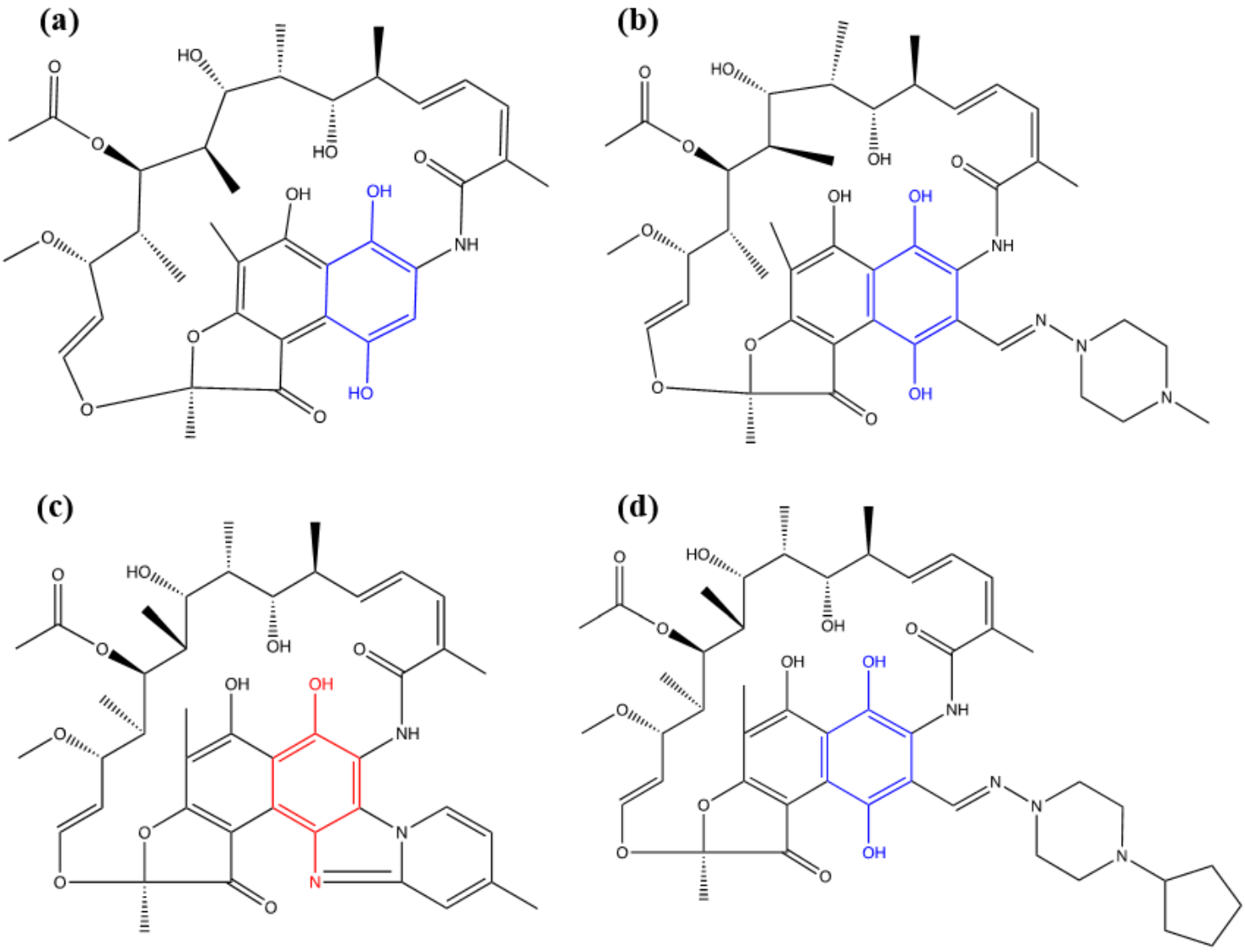

Figure 6

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 red. The figure was designed on ChemDraw™ and redrawn from Golegaonkar et al. (2015)67.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.pdf
- AdditionalFile2.xlsx