Check for updates

Advanced Intelligent Discovery

ADVANCED INTELLIGENT DISCOVERY Open Access

RESEARCH ARTICLE OPEN ACCESS

Large-Scale Machine Learning to Screen for Small-Molecule Senolytics

Alexis Dougha[1,2] | Jesus Gil[3,4] | Pedro J. Ballester[1]

[1]Department of Bioengineering, Imperial College London, London, UK | [2]Unité de Biologie Fonctionnelle et Adaptative, INSERM ERL, 1133, CNRS UMR, 8251, Université Paris Cité, Paris, France | [3]MRC Laboratory of Medical Sciences, Hammersmith Hospital Campus, London, UK | [4]Institute of Clinical Sciences (ICS), Faculty of Medicine, Imperial College London, London, UK

Correspondence: Pedro J. Ballester (p.ballester@imperial.ac.uk)

Received: 23 March 2026 | **Revised:** 11 April 2026 | **Accepted:** 20 April 2026

Keywords: graph neural networks | machine learning | phenotypic drug discovery | senolytics | virtual screening

ABSTRACT

Senescent cells are promising therapeutic targets given their key roles in cancer and age-related diseases. Virtual screening of vast chemical spaces for compounds selectively killing senescent cells (senolytics) is a promising approach. However, this requires computational models able to identify senolytics among a much larger proportion of highly diverse non-senolytic compounds. Here, we investigate the application of machine learning to enable such virtual screening. For the first time, we exploited real-valued training data via regression and analyzed the impact of detrimental factors such as data inconsistencies, training-test distribution shifts, and non-early-recognition performance metrics to identify senolytics. To this aim, we leveraged two senolytic datasets. The results show that many learning models cannot discriminate between senolytic and non-senolytic compounds. Furthermore, while graph neural networks lose their efficacy with test molecules chemically different from training molecules, a simple classification and regression tree (CART) algorithm led to a 21% hit rate in this realistic case study. Moreover, as CART is explainable, we discussed the extracted insights into how chemical structures relate to senolytic activity. We make this model available at https://github.com/alexisdougha/virtual-screening-senolytics.

1 | Introduction

Phenotypic drug discovery has emerged as an alternative to biochemistry assays and structure-based drug discovery to target diseases in a target-agnostic fashion [1]. This can enable researchers to speed up the drug discovery process, and above all broaden the range of possibilities for diseases for which the mechanism of action is poorly understood, or for which no experimental structures of a suitable target are available [2, 3]. Instead, it is possible to take advantage of cell-based assays that evaluate the effects of drugs over a given phenotype. In this way, the druggable space can be expanded, new molecular properties can be explored, and novel targets can be identified simultaneously [4]. A roadblock to this approach is to successfully scale up these assays via wet-lab high-throughput screening (HTS), using, for instance, cell-based assays [5]. It is unfeasible to screen the billions of potential drugs that are currently available to synthesize on demand. Computational models is the only viable option to explore these vast chemical spaces [6]. To make reliable predictions, such models need data, which can be generated by HTS [7].

Artificial intelligence (AI) offers unique opportunities in drug discovery [2, 8–12]. In this study, we leverage data from senolytic assays, which compare the viability of senescent and control cells, to identify the best model that will guide virtual screens for senolytic drugs. Cellular senescence is a stress response that avoids the replication of cancerous, damaged, and old cells. Senescent cells undergo a stable cell cycle arrest and undergo multiple phenotypic changes, including the production of a bioactive secretome called the senescence-associated secretory phenotype (SASP) [13]. Induction of senescence is needed for normal development, tissue homeostasis, and protects against cancer

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Advanced Intelligent Discovery* published by Wiley-VCH GmbH.

initiation and fibrosis. However, the aberrant accumulation of senescent cells contributes to aging and disease [14]. Consequently, senescent cells have become an attractive target to fight cancer, aging and many age-related diseases [15]. For instance, the elimination of senescent cells is a promising strategy against Type 1 diabetes [16] or cancer [17].

Two recent prospective studies, by Smer-Barreto et al. and by Wong et al., have shown that AI methods can help to identify senolytics [18, 19]. They both used computational models to virtually screen libraries and identify candidate senolytic compounds. Smer-Barreto et al. carried out a virtual screen of a library combining 4,340 bioactive compounds from the *L3800 Selleck FDA-approved & Passed Phase I* and the *L2100 TargetMol Anticancer* libraries. Wong et al. virtually screened a wider library of 804,959 compounds combining 799140 compounds from an *Extended Broad Institute Library and 5819 compounds from the Broad Institute's Drug Repurposing Hub* [20], which includes marketed drugs. For the virtual screening, Smer-Barreto et al. used Extreme gradient boosting [21] (XGB) with a selected subset of 165 physical and chemical features [22], while Wong et al. used a directed message-passing neural network (D-MPNN) to learn an embedding of the molecules [23], then used as input by a feed-forward neural network to discriminate between senolytic and non-senolytic compounds. Following experimental validation, Smer-Barreto et al. and Wong et al. obtained a hit rate of 14.3% (3/21) and 11.6% (25/216), respectively. Two of the three compounds reported by Smer-Barreto et al. (Periplocin and Oleandrin) are cardiac glycosides, a class of compounds that have been previously identified as senolytics [24, 25]. The other, Ginkgetin, derives from ginkgo biloba extract, previously reported to be senolytic [26]. On the other hand, Wong et al. identified BRD-K56819078, which was dissimilar from other compounds of the training set (max Tanimoto similarity is 0.29) and displayed senolytic activity in an old mice, significantly eliminating senescent cells in the kidneys [19].

Another reason to employ AI to guide virtual screening for senolytics is that this approach has been successful in identifying compounds against many other targets [27, 28]. In phenotypic drug discovery, AI models have been developed to search for molecules with activity in human pathogen cultures [29, 30] or cancer cell line cultures [31, 32]. The models can be adapted to perform virtual screens not only of small molecules but also peptides [33]: They are especially valuable to screen ultra-large libraries that are too large for HTS campaings [34] and to speed up the drug discovery process at various stages [35, 36]. In our study, we use machine learning (ML) algorithms that can identify complex patterns in the input features, and graph neural networks that can learn a relevant feature embedding, valuable to predict the intended properties.

In particular, Quantitative Structure–Activity Relationship (QSAR) is a molecular property prediction problem that can be performed by ML models [37, 38]. While the ML algorithm itself must be correctly selected to build predictive models, the choice of input features is also important [39]. Correctly assessing the quality of the resulting ML model is a significant task: complex algorithms for QSAR have become popular [40], but simple algorithms like random forest combined with a fixed molecular representation (like physical and chemical descriptors) are strong baselines that often outperform their more recent counterparts [41, 42].

In this work, we aim to understand how to best use data from multiple heterogeneous sources to discover new senolytics under realistic and comprehensive scenarios, taking into account data inconsistencies across datasets, experimental uncertainties, training-test distribution shifts, and relevant virtual screening metrics.

2 | Results

2.1 | Workflow Design Supporting Realistic Virtual Screening Benchmarking

Many factors can negatively impact the predictive performance of machine learning models, which can ultimately make them unsuitable for use in real-world virtual screening projects. To highlight these biases and facilitate the understanding of the results, we briefly introduce the designed workflow.

All the experiments are designed following a consistent global framework (Figure 1A) to comprehensively assess the impact of each factor. Each algorithm included in the study is benchmarked in both its regression and classification versions. Regressors directly predict the senolytic ratio (SR) that can be employed to prioritize compounds likely to be senolytics, whereas classifiers output a senolytic probability, that is also used to prioritize compounds and compute early-recognition performance metrics (see Experimental Design section for details about SR and metrics). In this study, we are going to evaluate supervised learning algorithms, including logistic and linear regression (LR), Classification and Regression Trees (CART), Random Forest (RF), Extreme Gradient Boosting (XGB), Support Vector Machine (SVM), Directed Message Passing Neural Network (D-MPNN) and Graph Learning-based Adaptive Machine (GLAM; see details in subsection Supervised Learning Algorithms). To assess the impact of salt processing in SMILES, both raw and salt-stripped SMILES are included. Similarly, the hyperparameters are selected according to two metrics, not to rely on a single criterion, as shown in Figure 1B. This procedure requires more models to be trained, but it allows us to detect the most predictive models and understand which factors are crucial. Given that this study aims to understand how to take advantage of heterogeneous data produced by two different laboratories, we first analyzed the similarities between the datasets. The maximum Tanimoto similarity of each compound from one dataset against any compound from the other dataset is equal to one for 13% of the Collins dataset and for 27% of the Gil dataset (Figure 1C,D, see Datasets subsection for details about cell lines and senescence induction). Inconsistencies between identical compounds were identified (Figure S1), making it more challenging for learning algorithms to extract informative knowledge. In light of these observations, we designed four computational experiments to identify the best models to virtually screen small-molecule senolytics (Figure 2).

First, the Collins dataset is used separately to select, train, and test learning algorithms (Figure 2A). Second, the Gil dataset is also used separately to select, train, and test learning algorithms (Figure 2B). In a third experiment, both datasets are merged before undergoing fivefold cross-validation to assess the ability of machine learning models to handle inconsistent data (Figure 2C). For this experiment, models were trained using default hyperparameters to prevent the additional computational

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aidi.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

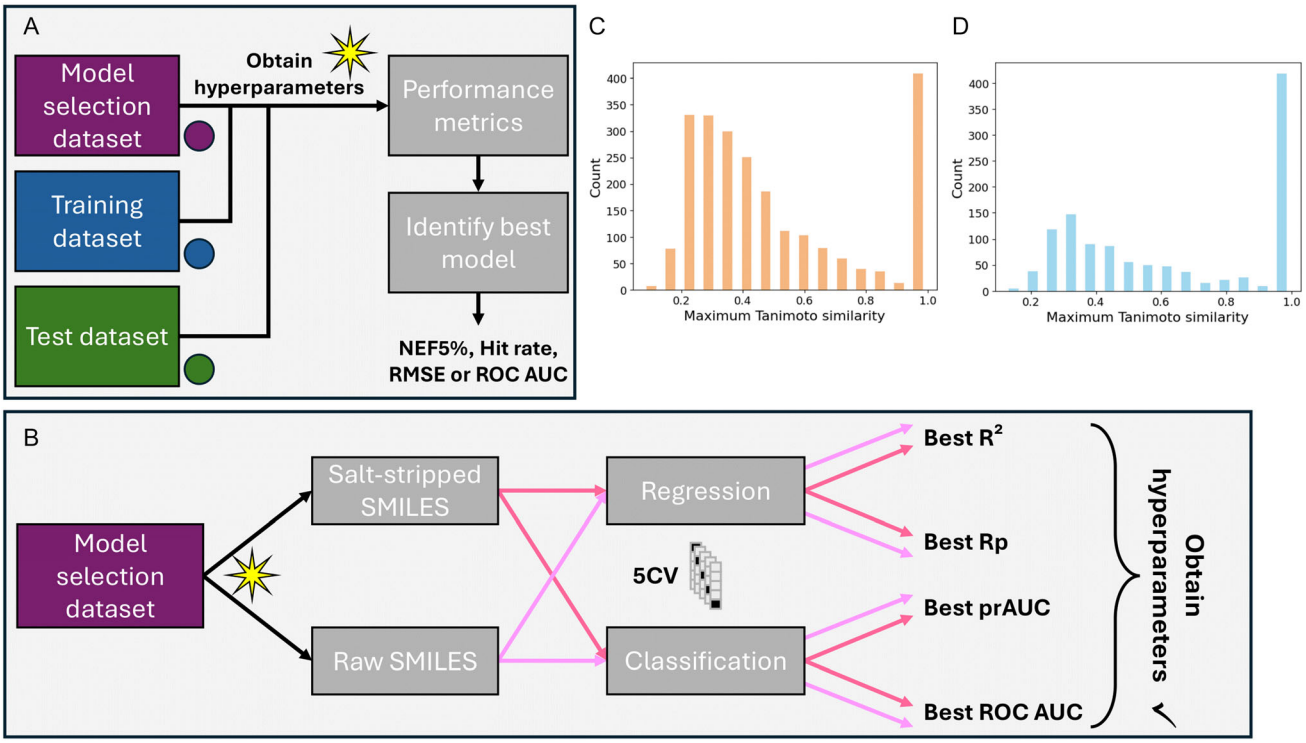


FIGURE 1 | Overall workflow applied for all the experiments reported in the study. The four experiments reported in the study follow a consistent general workflow summarized here. (A) In each experiment, a model selection dataset (purple) is used to find suitable hyperparameters. Then, the models are trained with the training set (blue) and the performance metrics on the test set (green) are reported. Each dataset comprises a set of molecules and their values for the senolytic ratio SR (see Experimental Design section). (B) The process to obtain the hyperparameters is outlined in this figure. For each algorithm, the best hyperparameters are selected using either salt-stripped SMILES or raw SMILES, with the regression or the classification version of the algorithm, and with respect to two metrics for each. This leads to eight sets of hyperparameters. (C) Distribution of maximum Tanimoto similarity of the compounds from the Collins dataset vs any compound from the Gil dataset. (D) Distribution of maximum Tanimoto similarity of the compounds from the Gil dataset vs any compound from the Collins dataset.

cost of repeating the optimization procedure independently for each fold, while avoiding any potential data leakage. Finally, we evaluate the models in a more challenging case, where the aim is to mimic a virtual screening scenario in which the compounds to screen are dissimilar from the training data (Figure 2D). A more comprehensive description of the methodology is provided in the Experimental Design section.

2.2 | Machine-Learning Models can Identify Senolytics Under the Oversimplified Training-Test Distribution Shift-Free Scenario

Virtual screening performance by regression models is typically only evaluated with classification metrics. While evaluating regression models with regression metrics leads to less flattering results, it provides a deeper understanding of performance differences across models [43]. Thus, we are adopting a dual regression-classification evaluation framework where we can see predicted vs measured senolytic ratio (SR, defined as the ratio between control cell viability and senescent cell viability; see Equation (1) in Experimental Design section). The first two experiments focus on using a single dataset at a time: only the Collins dataset (partition in Figure 2A, data distribution in Figure 3A, repeated SR measurements in Figure 3B), and only the Gil dataset (partition in Figure 2B, data distribution in Figure 3C).

Tested with the Collins holdout test set, regressors and classifiers achieve similar performance (Tables S1-4). The best hit rate of 33% is obtained by a XGB classifier (Figure 4A) with a $NEF_{5\%}$ of 0.333, while the best $NEF_{5\%}$ of 0.444 is obtained by a RF regressor (Figure 4B), a XGB regressor (Figure S2), and a CART regressor (Figure S3) with a hit rate of 22%. We used the repeated measurements reported in the Collins dataset to determine the maximum performance that we could obtain on this dataset given, the experimental uncertainty. Considering measurement 1 as the ground truth and measurement 2 as if they were predictions of the same compounds (Figure 3B), we computed the performance metrics of this fictitious model: $NEF_{5\%}$ of 0.500, hit rate of 44%, and RMSE of 0.366. Notably, the RMSE of the best RF regressor is 6% smaller than that of the fictitious model. It supports the idea that this procedure produces results evaluated on easier scenario, whereas prospective virtual screening will see mostly different molecules. So a proper evaluation with dissimilar test molecules will be required.

In the second experiment, we used data exclusively from the Gil dataset (partition in Figure 2B, data distribution in Figure 3C). Contrary to the Collins experiment, the most predictive model in this case is the graph neural network GLAM in regression mode (Figure 4C), achieving $NEF_{5\%}$ of 0.500 and 50% hit rate. It is noteworthy that, depending on the dataset (Collins or Gil), the best selected models are tree-based ensemble algorithms or graph neural networks, although the same protocol was

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aidi.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

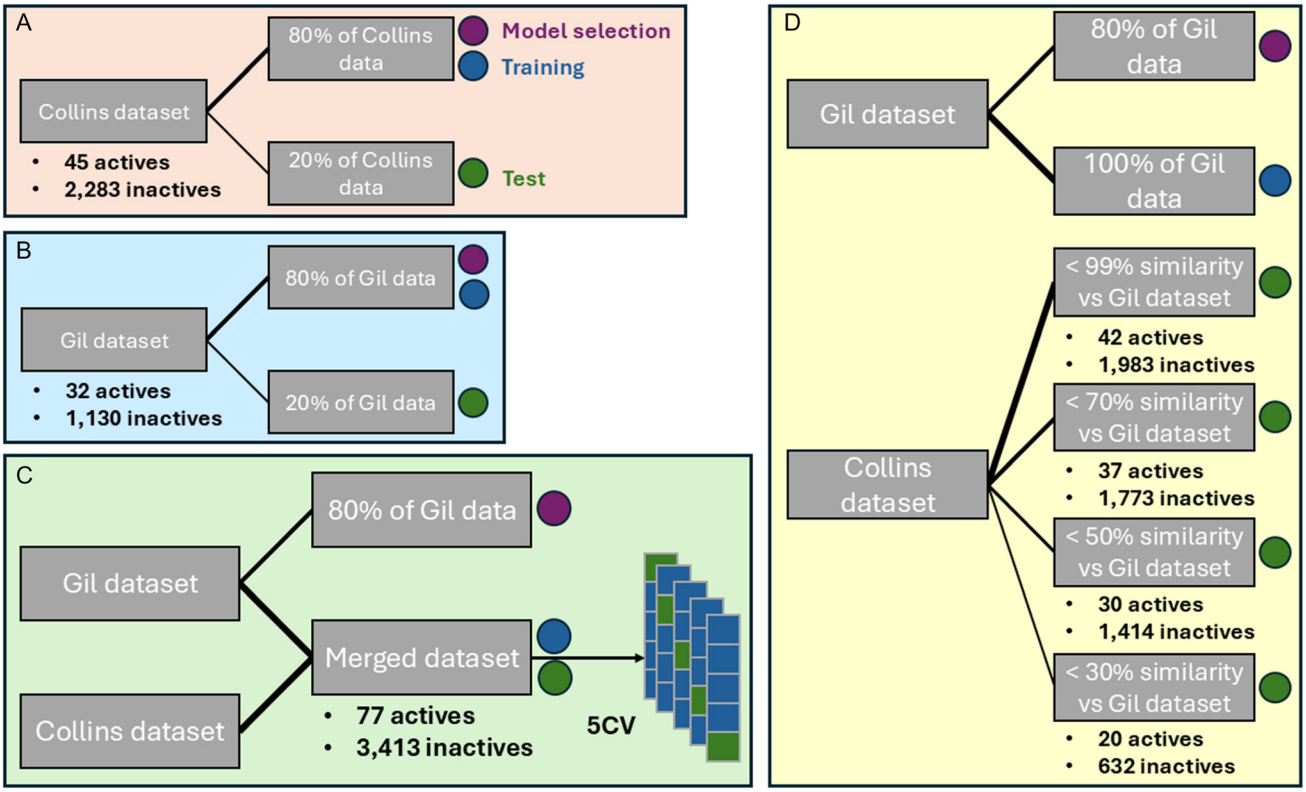


FIGURE 2 | Data partition for each computational experiment. In each experiment, the model selection, training, and test datasets are indicated by purple, blue and green disks, respectively. A more comprehensive description of the stratified splitting strategies is provided in the Experimental Design section. (A) A subset of the Collins dataset is used for both model selection and training, while the holdout test set is used to evaluate the performance of the best models. (B) A subset of the Gil dataset is used for both model selection and training, while the holdout test set is used to evaluate the performance of the best models. (C) The same model selection dataset as in the B panel is used here. Then the performance of the best models is assessed following a fivefold cross-validation procedure combining both Gil and Collins datasets. (D) The same model selection dataset as in the B panel is used here. The Gil dataset acts as the training set while four subsets of the Collins dataset (with increasing dissimilarity from the Gil dataset) are used for testing.

applied in both model selection processes. The best GLAM model obtains an RMSE of 0.986, while the smallest RMSE of 0.828 is obtained by a Chemprop model. This underlines the importance of considering appropriate early-recognition performance metrics since this Chemprop model is substantially outperformed by GLAM in terms of $NEF_{5\%}$ and hit rate, 0.25% and 12.5%, respectively (Figure 4D). By comparing regressors and classifiers, this experiment also highlights the possible advantage of using quantitative data when available. Here, all the classifiers are outperformed by the best regressors when considering $NEF_{5\%}$ and hit rate (Tables S5-8). Moreover, none of the binary classifiers are more predictive than the baseline logistic regression algorithm, suggesting that more sophisticated learning algorithms can fail to model binary data (i.e., instead of real-valued), even in a simple scenario without training-test distribution shift.

2.3 | Models' Performance Substantially Decreases With Inconsistent Data

Encouraged by the results obtained in simpler scenarios, we conducted a third experiment leveraging both datasets (Figure 2C) to evaluate the ability of learning algorithms to handle inconsistent data. In contrast to the experiment using only the Gil dataset, different models achieved the best metrics on the five-fold cross-validation experiment. Notably, the same hyperparameters were used in both experiments, as hyperparameter optimization was performed with the same 80% partition of the Gil dataset (Figure 2B,C). Surprisingly, the most predictive model is a XGB classifier ($NEF_{5\%}$ of 0.299, hit rate of 18.2%; Figure 4E), outperforming all the regressors (Tables S9-12). None of the regressors outperform the linear regression baseline in terms of hit rate (10.4%). We speculate that when the noise due to data inconsistencies is too large, learning algorithms are more impaired when dealing with real-valued data rather than with binary data. Regarding the selection of a regressor or a classifier, a parallel can be drawn between this experience and that of Gil in which a regressor outperforms all the classifiers and no classifier can beat its reference baseline. Similarly, here, a classifier outperforms all the regressors and this is even more significant as no regressor can beat its reference baseline.

2.4 | CART Models are Predictive and Outperform Other Learning Algorithms Under the Most Challenging Scenario

In the last experiment, we wanted to identify the most predictive models under a realistic scenario mimicking a real-world screening program. To this end, attention should be paid to avoid biases

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aid.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

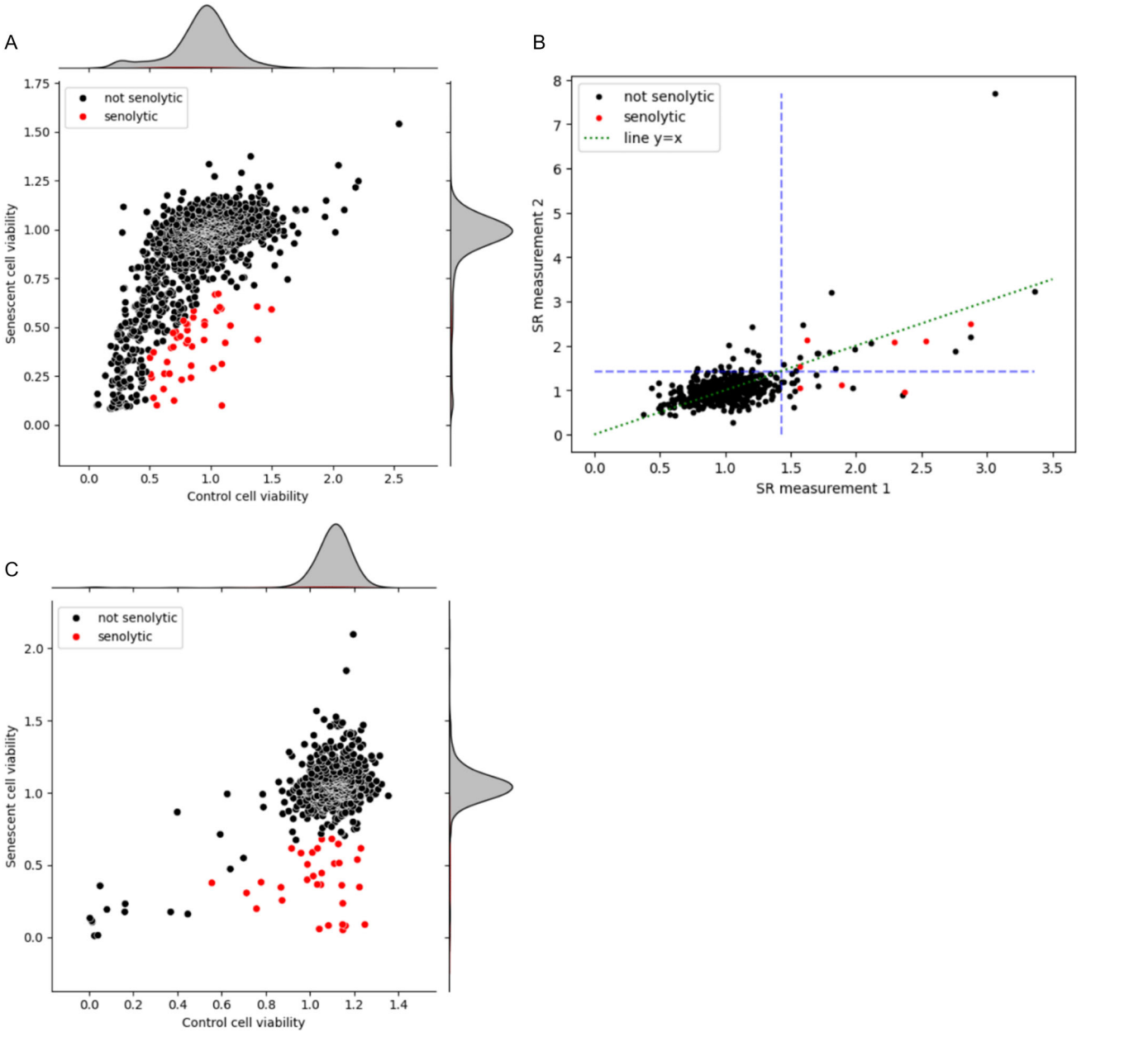


FIGURE 3 | Data distributions. (A) Result of the experimental screen conducted by the Collins laboratory [19]. (B) Moderate correlation between repeated measurements of the senolytic ratio in the Collins test set. Considering measurement 1 as the ground truth and measurement 2 as if they were predictions of the same compounds, the performance metrics are the following: Hit rate = 44.4%, $NEF_{5\%} = 0.500$, RMSE = 0.366. Dotted blue lines indicate the $SR = \frac{1}{0.7}$ cutoff. (C) Result of the experimental screen conducted by the Gil laboratory [24].

highlighted by the other experiments. In particular, the experiment comprising only Collins data (Figure 2A) proved that caution is needed to design the splitting procedure because high train-test similarity is likely to lead to overoptimistic performance metrics (error smaller than experimental uncertainty), while in ultra-large libraries, most compounds will be dissimilar from the training dataset. The experiment, which includes both senolytic datasets, has shown that learning algorithms are impaired by inconsistent training data. Additionally, this experiment may be biased by the model selection procedure, which leverages 80% of the Gil dataset while this subset is also part of the merged dataset used to conduct the fivefold cross-validation. With these observations in mind, we have established a rigorous protocol for evaluating the performance of learning algorithms to inform prospective studies. The Gil dataset is used for both model selection and training, while four subsets of the Collins dataset are used for testing. Each subset gathers all the compounds from the Collins dataset that show less similarity (than a given threshold) to any compound used for training. Using different similarity thresholds (99%, 70%, 50%, and 30%) allows us to comprehensively evaluate the performance of the models on gradually more challenging test sets. The models were trained with five different seeds to enable rigorous statistical comparison of performance.

In these experiments, CART in regression mode consistently outperforms all the other learning algorithms (Tables S13–16). Notably, a CART model using salt-stripped SMILES achieves the best $NEF_{5\%}$ (Table S14) and the best hit rate (Table S16) for all the dissimilarity thresholds. Even in the most challenging

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aidi.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

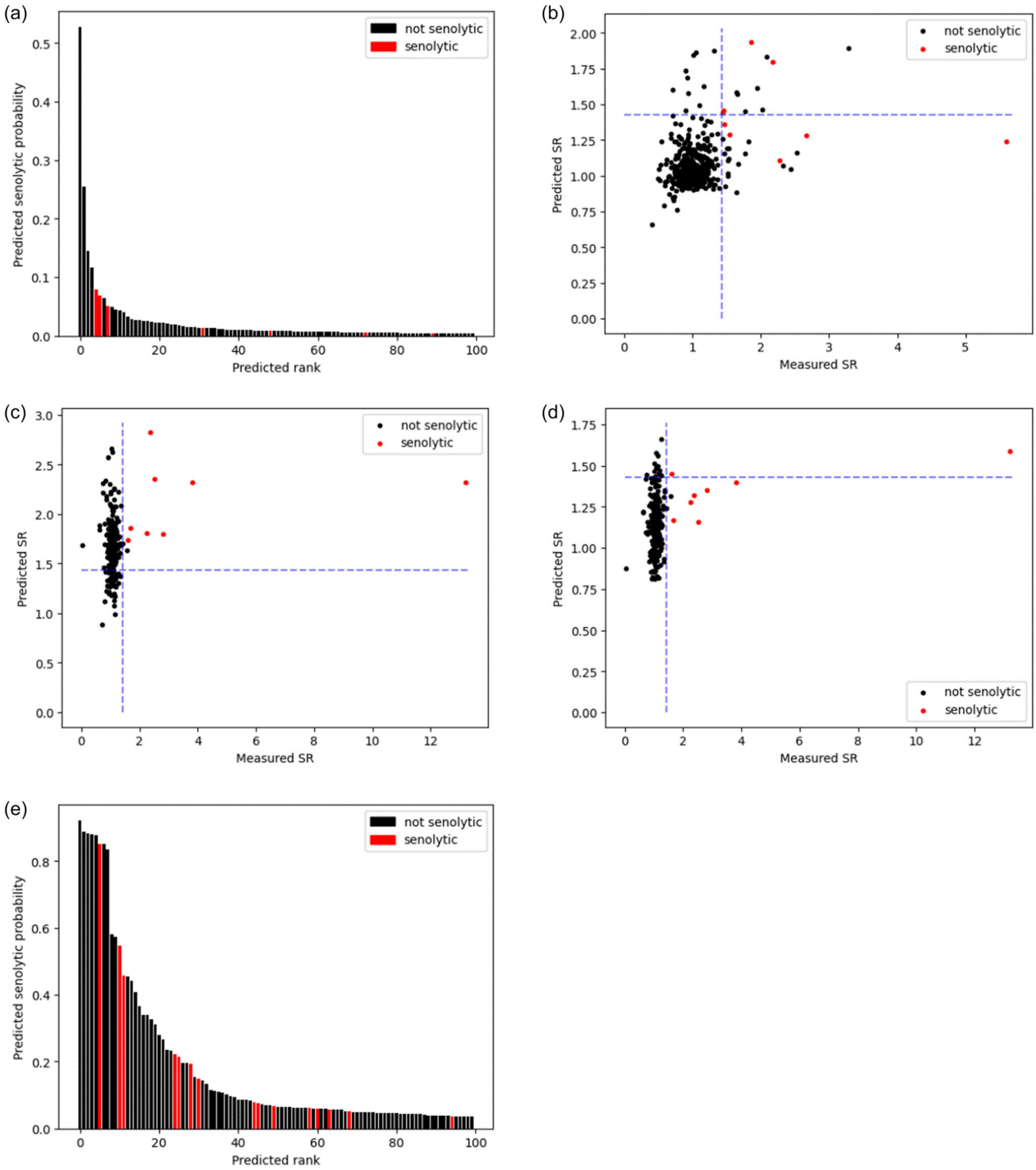


FIGURE 4 | Senolytic test sets predictions. Dotted blue lines indicate the $SR = \frac{1}{0.7}$ cutoff. Not all points to the right of the vertical blue line are senolytics according to Collins definition, only those in red (this small discrepancy comes from taking a single-task regression modeling approach). Points above the second dotted line are molecules predicted senolytic by the model. (A) Test set predicted by the best XGB classifier model (Collins dataset; Figure 2A). Only the top 100 predicted active probabilities are shown. Hit rate = 33.3%, $NEF_{5\%} = 0.333$, ROC AUC = 0.871. The hyperparameters are reported in Table S1. (B) Test set SR predicted by the best RF regressor model (Collins dataset; Figure 2A). Hit rate = 22.2%, $NEF_{5\%} = 0.444$, RMSE = 0.343. The hyperparameters are reported in Table S3. (C) Test set SR predicted by the best GLAM regressor model (Gil dataset; Figure 2B). Hit rate = 50.0%, $NEF_{5\%} = 0.500$, RMSE = 0.986. The hyperparameters are reported in Table S7. (D) Test set SR predicted by the best Chemprop regressor model (Gil dataset; Figure 2B). Hit rate = 12.5%, $NEF_{5\%} = 0.250$, RMSE = 0.828. The hyperparameters are reported in Table S7. (E) Fivefold cross-validation predictions by the best XGB classifier model (Merged dataset; Figure 2C). Only the top 100 predicted active probabilities are shown. Hit rate = 18.2%, $NEF_{5\%} = 0.299$, ROC AUC = 0.794. The hyperparameters are reported in Table S9.

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aid.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

case using the 30% max-similarity test set, this model achieves a $NEF_{5\%}$ of 0.250 and a hit rate of 21% (average over five seeds; Figure 5A,B). Interestingly, classifiers are outperformed by regressors, and the performance gap is even more substantial on more dissimilar test sets (based on the increasing $NEF_{5\%}$ difference between the best regressor and the best classifier for each similarity cutoff; Tables S13–14). Again, RMSE is not able to identify the best learning algorithm for early-stage virtual screening: a SVM model achieves the best RMSE of 0.443 in the most challenging scenario (Figure 5C), but it shows 0 hit rate (Figure 5B) and only a $NEF_{5\%}$ of 0.050 (Figure 5A). Even more striking, the CART regressor's hit rate is greater than all the other learning algorithms (statistically significant; Figure 5B) while its RMSE is simultaneously the worst of non-baseline regressors (statistically significant; Figure 5C). It proves that, contrary to $NEF_{5\%}$, RMSE is not a relevant metric in the case studied here. Remarkably, this trend is confirmed for all the experiments conducted in this study. Indeed, researchers using RMSE to measure the performance of machine learning models for virtual screening assume that the closer the RMSE is to 0, the higher the hit rate, and vice versa (i.e., anti-correlated quantities). However, there is no such anti-correlation; Pearson correlation of 0.014, *p*-value of 0.91 proving no correlation (Figure 5D). Even when measuring the correlation between the RMSE and the hit rate within the same experiment, no anti-correlation is found (Table S17). This observation is consistent with another study involving 60 datasets that showed little correlation between a global metric (ROC AUC) and the hit rate [44].

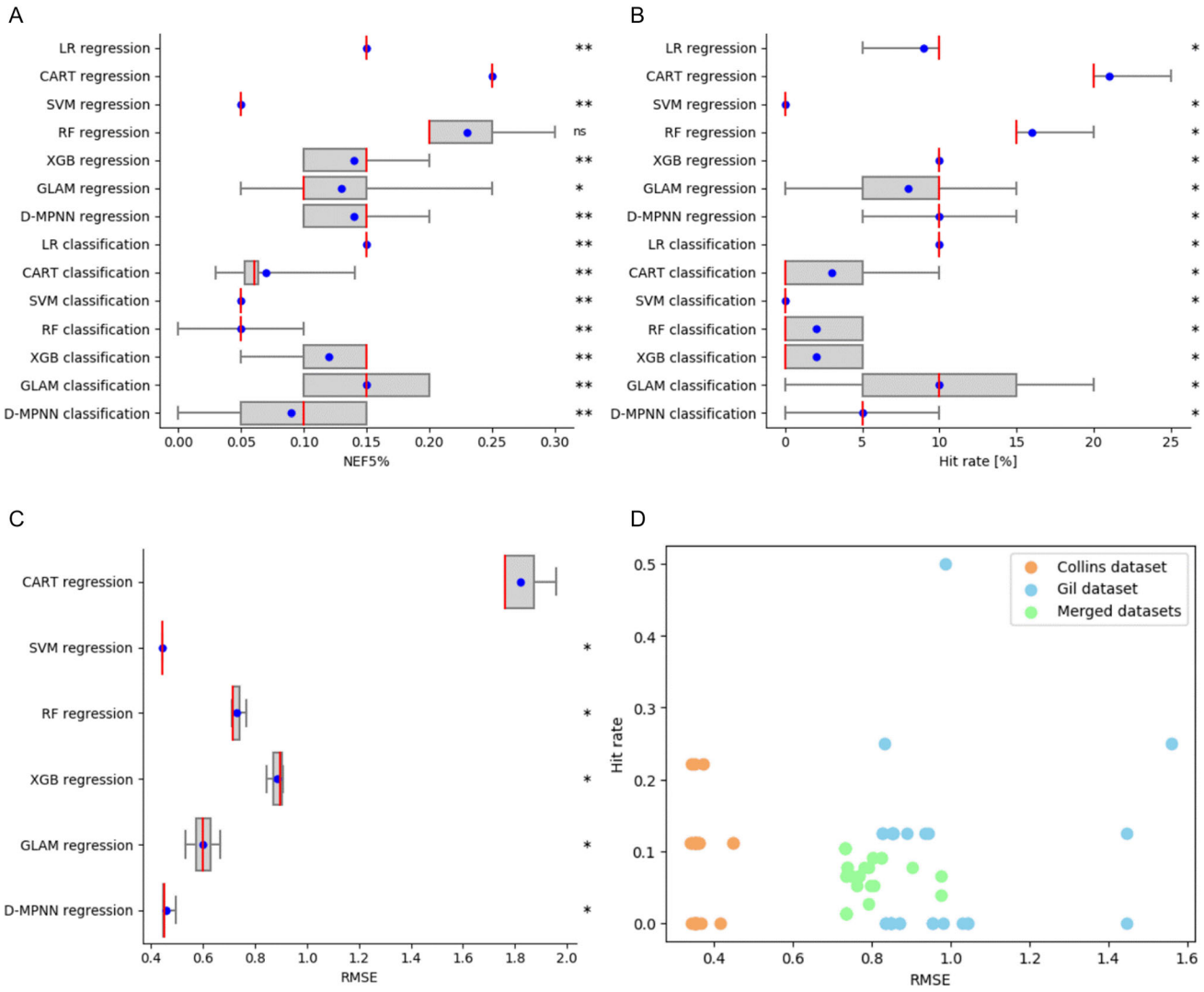


FIGURE 5 | Boxplots of dissimilar test set predictions and RMSE case study. The blue points and the red bars in the boxplots indicate the mean and the median of the reported metric, respectively. The whiskers extend from the minimum to the maximum of the reported metric, with predictions repeated for five seeds. (two-by-two post hoc comparisons CART regression vs each of the other learning algorithm, ns = not significant, *0.01 < = p < 0.05, **0.001 < = p < 0.01, see details in Statistical analysis subsection). (A) Boxplot of $NEF_{5\%}$ for the most dissimilar test set (< 30% similarity; Figure 2D). The mean values and hyperparameters are reported in Tables S13–14. (B) Boxplot of hit rate for the most dissimilar test set (< 30% similarity; Figure 2D). The mean values and hyperparameters are reported in Tables S15–16. (C) Boxplot of RMSE for the most dissimilar test set, all regressors except LR (< 30% similarity; Figure 2D). The mean values and hyperparameters are reported in Table S14. (D) Lack of correlation between the RMSE and the hit rate (n = 72). Contrary to what is expected when using the RMSE as a performance metric, there is no significant correlation with the hit rate, with a Pearson correlation of 0.014 (p-value of 0.91). The points represent the test set predictions in experiments 1 – 3.

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aidi.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.5 | CART Models are Interpretable

Contrary to more sophisticated black-box models, CART is highly interpretable. In the last experiment (dissimilar test sets), the best CART model had a tree depth of five, which makes it manageable to visualize (Figure 5A) and enables the identification of the key features driving senolytic prediction. Importantly, the tree structure remains almost identical across the seeds (Figures S4-7), supporting the robustness of the analysis.

CART interpretability can be exemplified by the case study of the senolytic compounds A1155463 (Figure 5B) and A-1331852 (Figure 5C) that are both correctly predicted as active. Indeed, these compounds share much of their structure (highlighted in red), which is probably related to their senolytic activity. Actually, A1155463 and A-1331852 are BCL-X_L inhibitors [45, 46], an anti-apoptotic protein that can be targeted to eliminate senescent cells [47]. Both studies reported that the common structure is important for BCL-X_L binding (the crystal structure 4QVX of BCL-X_L bound to A1155463 shows interactions with the P2 pocket), but they also showed that the potency is modulated by the complementary part of the compounds that can be optimized to additionally bind the P4 pocket [46]. Interestingly, CART does not simply act as a substructure matching algorithm (that would only recognize the pattern highlighted in red), but it also relies on more subtle features: although both compounds are predicted to be active, they do not follow the same path in the decision tree, and this difference can be explained notably by the distinct parts of their structure. A-1331852 ends up in a leaf that predicts actives (SR = 4.88), while it would have been predicted inactive if it ended up in the sibling node (SR = 1.23). This final splitting criterion, based on the BalabanJ descriptor [48] is very stringent (only one training compound out of 85 is below the BalabanJ cutoff), but this rule proved to be effective in the case of A-1331852. On the other hand, A1155463 would not have satisfied this criterion. It shows that the impact of the distinct region on the BalabanJ is key for CART to predict A-1331852 active. Consistently, BalabanJ is a topological index related to the shape of the molecule (e.g., cycles, branching), which must be impacted by the adamantane moiety in A-1331852. In a similar way, A1155463 ends up in the node leaf that predicts actives (SR = 17.03). The final criterion is based on VSA_EState9 (i.e., the sum of the electrotopological state [49] of atoms whose surface area is in the ninth bin defined by RDKit). The only atoms that contribute to this descriptor are the carbons in the methyl groups attached to the tertiary amine of A1155463 (i.e., not part of the substructure match). Therefore, the distinct region of this compound is also critical for the CART prediction. These examples show how it is possible to interpret CART predictions to provide relevant insights for medicinal chemists in real-world applications Figure 6.

3 | Discussion

Based on a benchmark encompassing a simplified, as well as a more challenging case, we have exemplified the advantage of using quantitative data with regression to predict senolytics in a realistic scenario. To the best of our knowledge, this study is the first one to exploit real-valued senolytic data, which contains more information than their binary counterparts, and to comprehensively compare it with classification, while previously reported studies only focused on binary classification [18, 19]. We have shown that, especially under the training-test distribution shift case, regressors outperform binary classifiers. In addition, this study has provided strong evidence that early-recognition virtual screening performance metrics, including $NEF_{5\%}$ and hit rate, must be used instead of global metrics like RMSE. The most prominent example supporting this is the SVM model that achieves the lowest RMSE on the most dissimilar test set in the last experiment while it achieves a hit rate of 0 and a $NEF_{5\%}$ five times smaller than the best CART model. Comparing the results obtained in the different scenarios also provides an interesting perspective on the performance of GNNs. In the second scenario (involving only the Gil dataset), GLAM outperforms tree-based algorithms while CART is significantly more predictive under the most challenging scenario. This suggests that CART is best suited for senolytics virtual screening. Other studies have also reported that traditional ML can work better than deep-learning algorithms with some datasets [41, 50, 51]. In contrast, biased benchmarks with similar compounds in the train and test datasets are not able to select the best performing model. On top of that, such benchmarks overestimate the performance: we recommend having at least an estimate of the experimental uncertainty of the predicted quantity before starting to model the data in any property prediction task. Thus, it can be used as a proxy to check for overfitting as we have shown with the test error smaller than the experimental error in the Collins dataset-only experiment. Both raw SMILES and salt-stripped SMILES were used in combination with each learning algorithm, but no clear trend could be identified. This suggests that no single rule is better in all situations (i.e. dataset-dependent), but it could be a way to augment data or to build consensus models (i.e. averaging the predictions of models using both representations). As opposed to black-box algorithms, explainable artificial intelligence aims to make AI predictions interpretable. CART is interpretable by design, which makes it a valuable algorithm for virtual screening given the need for transparency and evidence-based decisions in drug discovery [52, 53]. In this study, we decided to simply regress the SR instead of regressing both cellular viabilities but in the future, it could be relevant to assess the performance of multitask models that predict both properties at the same time. Finally, the next step will be to use these models for prospective screening to identify candidates to be experimentally validated.

4 | Conclusion

In this study, we leveraged the Gil dataset for virtual screening for the first time [24]. We have shown that even under the most challenging scenario, the interpretable algorithm CART is predictive, thus can be used to perform virtual screening of small-molecule senolytics. In our case study, we have demonstrated that contrary to global metrics such as the RMSE, early-recognition performance metrics such as $NEF_{5\%}$ are suitable for selecting the best models, producing the greatest hit rate. When tested on a dataset from a different laboratory, comprising compounds very dissimilar to the training set, CART could still achieve 21% hit rate, outperforming all the benchmarked algorithms. Interestingly, this holds true for regression, while binary classifiers that were used in previous studies [18, 19] do not perform well, further demonstrating the value of quantitative data for senolytics discovery. More broadly, the results of this study

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aidi.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

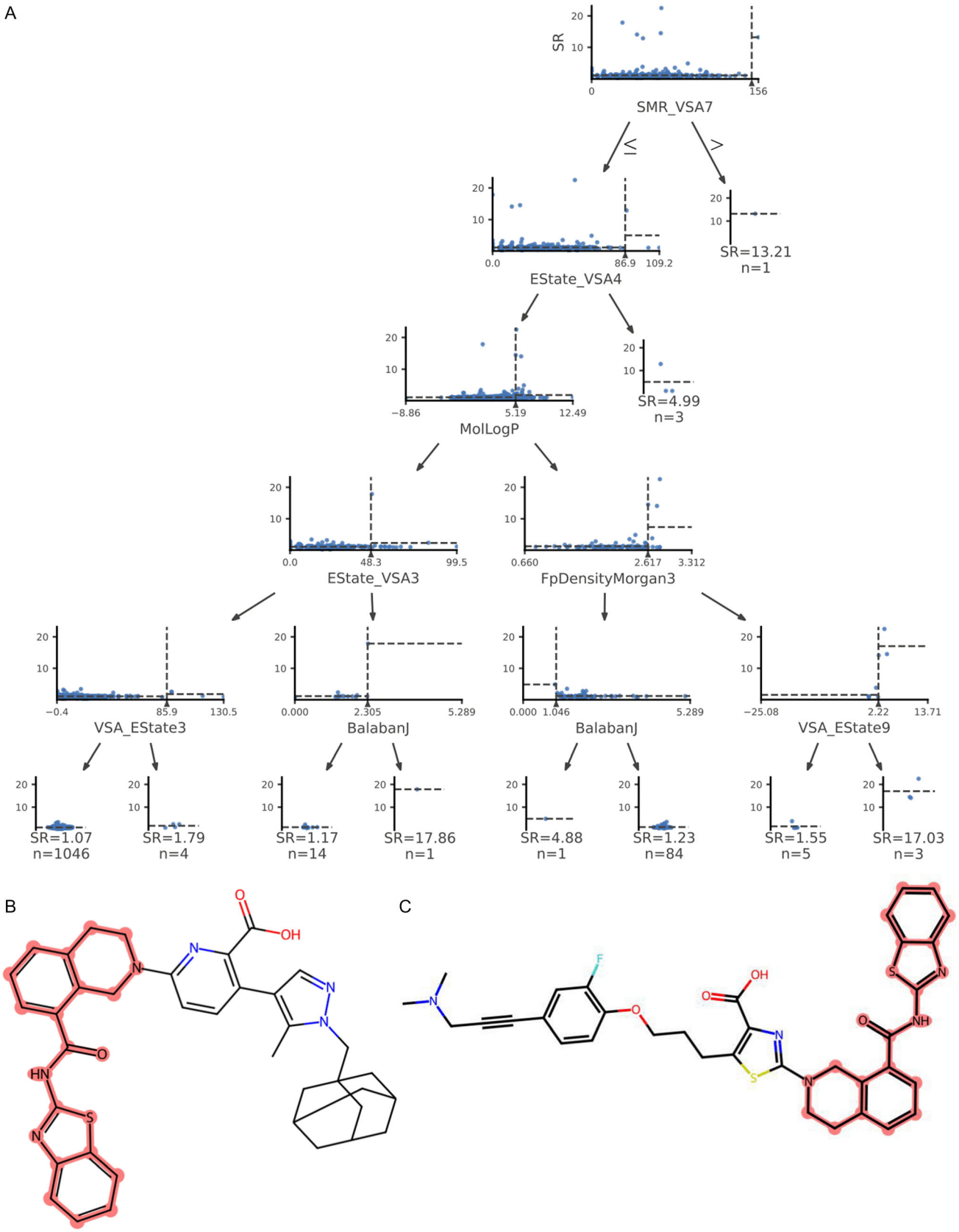


FIGURE 6 | Visualization and interpretation of CART. (A) Tree structure visualization of CART algorithm trained on the Gil dataset. It was generated with seed = 0, other CART trees generated with four other seeds are reported in Figures S4-7. In each branch node, a scatter plot shows the training compound distribution of the senolytic ratio (*y*-axis) against the feature selected (*x*-axis) by the algorithm to split the node further. The vertical dashed line indicates the cutoff value of this feature. Leaf nodes show the senolytic ratio value that will be predicted for the test compounds that follow the decision path leading to this leaf. (B) Senolytic compound A-1331852 from the Collins < 30% similarity test set. The substructure match with A1155463 is shown in red. (C) Senolytic compound A1155463 from the Collins < 30% similarity test set. The substructure match with A-1331852 is shown in red.

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aidi.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

illustrate the importance of establishing robust benchmarks in drug discovery that consider data biases, experimental uncertainty and performance metrics that are suitable for the real-world application.

5 | Experimental Design

5.1 | Datasets

In this study, datasets from two different sources were used (Table 1). A compound is defined as active if it selectively kills senescent cells. Only the compounds that were successfully featurized were kept (see Featurization subsection). The Gil dataset described the screen of 1162 compounds from the LOPAC library (Library Of Pharmacologically Active Compounds) and assessed the viability of control and etoposide-induced senescent human IMR90 fibroblasts. This has been previously reported in [24]. The Collins dataset comprises data of 2328 compounds screened in a similar fashion [19]. The combination of both datasets will help to understand how to efficiently merge senolytic data from various sources reporting in similar therapy-induced senescence assays [54–56].

5.2 | Senolytics

Senolytics are compounds able to selectively kill senescent cells (i.e., killing much fewer non-senescent cells). More generally, we will also refer to senolytics as active compounds or just actives, and to those without selective senolytic activity as inactives (Table 1). Since both the Collins dataset and the Gil dataset contain quantitative values for senescent and non-senescent cell viabilities, a clear criterion could be chosen following Wong et al. [19]. In our study, we defined the senolytic ratio (SR) as in Equation (1), leveraging the cell viability values gathered in the dataset instead of more demanding IC_{50} values that are not available here. A compound is considered active if $SR > \frac{1}{0.7}$, control cell viability > 0.5 and senescent cell viability < 0.7, and inactive otherwise. This definition is used for all the experiments.

$$SR = \frac{Control\ cell\ viability}{Senescent\ cell\ viability} \quad (1)$$

We chose to regress the SR because it enables us to capture the selectivity of the compounds with only one continuous value ($SR > 1$ means it is more potent against senescent cells than against control cells). It is important to note that for binary classification experiments, we directly use the definition of active that we define in this subsection. On the other hand, regression experiments simply predict the SR while these models are evaluated for their ability to predict active compounds. There is a trade-off between missing information on a compound's potency when performing classification, and missing information by not considering cell viability thresholds when performing regression.

5.3 | Data Processing and Features

Smer-Barreto et al. used 200 RDKit descriptors [22] as input features, and Wong et al. also used them to augment the learnt representation of molecules. Given that their model was successful prospectively, we decided to also use the RDKit descriptors as input features for our machine learning (ML) models. These features contain physical, chemical and topological information about the compounds, including molecular weight, subdivided surface areas, atom and bond counts, adjacency matrix derived descriptors, and other diverse ones. The features were then normalized with the Min–Max scaler to keep all the values in the same [0,1] interval. Smer-Barreto et al. computed the RDKit descriptors directly on the raw SMILES, while in many pipelines, a preprocessing step includes removing salts from the compounds [57, 58]. We decided to include both methods in our benchmark (i.e., using the salt remover module from RDKit to process SMILES or using raw SMILES) since data preprocessing can significantly impact the quality of ML models trained with heterogeneous data sources [59]. Duplicates and compounds that failed to pass through this pipeline were removed from this study, and thus were not included in Table 1. In particular, this applies to inorganic compounds (for instance, calcium chloride in the Collins dataset).

5.4 | Supervised Learning Algorithms

Learning algorithms chosen to perform SR regression and their hyperparameter optimization methodology are described here. Their classification counterpart is also included in the benchmark. When hyperparameters were tuned, the best models were selected based on their coefficient of determination [60] and Pearson correlation for regressors (R^2 and R_p, respectively, as defined in performance metrics subsection). The best classifiers were selected based on their prAUC (area under the precision–recall curve) and ROC AUC (area under the receiver operating curve) scores. The final models were additionally scored with RMSE and early-recognition virtual screening performance

TABLE 1 | Senolytics datasets. Contrary to heterogeneous and qualitative senolytics data previously assembled in the literature [18], the Collins and the Gil datasets only gather quantitative data produced following a consistent protocol. Here we leveraged the Gil dataset for virtual screening for the first time.

Information	Collins dataset	Gil dataset
Introduced in	Wong et al. [19]	Guerrero et al. [24]
Actives plus inactives	2,328	1,162
Actives	45	32
Cell line	IMR90	IMR90
Testing concentration	10 μM	10 μM
Cellular Senescence Induction	Etoposide (Therapy-Induced Senescence)	Etoposide (Therapy-Induced Senescence)

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aidi.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

metrics (see details in performance metrics subsection), evaluated on a holdout test set.

Linear regression can only capture linear relationships between the input features and the regression objective SR. We use it as a simple baseline, together with the logistic regression.

Regression tree is considered because it is a weak learner compared to more complex learning algorithms used in this study. It recursively splits the training data to minimize the residual reduction. We used the CART implementation from scikit-learn [60] with hyperparameters tuned following a grid search procedure in the configuration space reported in Table 2. In particular, the max_depth hyperparameter is kept below six in order to build decision trees of reasonable size and obtain interpretable models.

Random forest (RF) is an ensemble technique which aggregates the predictions of a given number of decision trees. Each decision tree is trained with a subset of the input features, which is supposed to limit overfitting compared with individual trees. The hyperparameters were tuned following a grid search procedure in the configuration space reported in Table 2.

Support vector regression is a support vector machine (SVM) algorithm able to perform regression. The hyperplane from the classification version is replaced by a tube of a given width. Data points outside are penalized according to their distance from the tube. We tuned the kernel hyperparameter, testing the following values: "linear", "poly", "rbf", "sigmoid". Applying non-linear kernels enables us to capture non-linear relationships.

Extreme gradient boosting (XGB) is an implementation of gradient boosting [21]. Contrary to RF, XGB iteratively corrects the error of previous trees to improve the final prediction. The hyperparameters were tuned following a grid search procedure in the configuration space reported in Table 2, as recommended to build QSAR models [62].

Chemprop implements a directed message-passing neural network (D-MPNN), a type of graph neural network (GNN) which uses directed edges (i.e., two for each covalent bond) to pass messages [23, 63]. It was used as a classifier by Wong et al. to prospectively identify senolytics [19]. Chemprop was also successful in diverse recent applications, including antibiotics discovery [64–66], chemical reactivity [67, 68], toxicity [69], and ADME property prediction [70]. First, a graph is generated from the input SMILES, in which node (atom) and edge (bond) features are obtained with RDKit. Then, messages are passed for a given number of steps, and the last node representations are aggregated into the learnt embedding vector. Finally, a feedforward neural network outputs the senolytic ratio. It is possible to concatenate the learnt embedding with a vector containing the 200 RDKit descriptors, which have been used to capture global senolytic properties [64], so we tested both possibilities. Given the large configuration space, hyperparameters were tuned with Bayesian optimization for an efficient search with the automated function implemented by Chemprop. The range of the hyperparameters tuned is reported in Table 2.

Graph learning-based adaptive machine (GLAM) is a state-of-the-art GNN which outperforms other methods on a large range of property prediction tasks [71]. We performed a grid search in the configuration space reported in Table 2, always combined with the message passing block NNConv [72, 73]. This allows us to extend our benchmark by adding another GNN that using a compute-intensive pipeline to find suitable hyperparameters.

5.5 | Data Splits

Careful splitting procedures are applied for each of the experiments in this study (Figure 2). In the first experiment, the Collins dataset is divided into a training and a test set, comprising

TABLE 2 | Hyperparameter search space for all the tuned algorithms. CART, RF, XGB, SVM, and GLAM were tuned with a grid search while the D-MPNN was tuned with Bayesian Optimization (Hyperopt [61] implemented in Chemprop). Note that for the D-MPNN, linked_hidden_size simultaneously represents hidden_size, and ffn_hidden_size, assigning the same value for both hyperparameters.

Algorithm	Hyperparameter	Evaluated values
CART	max_depth	2, 3, 4, 5, 6
	min_samples_split	2, 4, 6
RF	n_estimators	50, 100, 150
	min_samples_split	2, 6
	max_features	$\frac{1}{3}$, $\frac{2}{3}$, None
XGB	max_depth	4, 5, 6, 7, 8, 9, 10
	colsample_bytree	0.50, 0.75, 1.0
	num_round	250, 500, 750, 1000
SVM	kernel	linear, poly, rbf, sigmoid
D-MPNN	depth	2, 3, 4, 5, 6
	ffn_num_layers	1, 2, 3
	dropout	0.0, 0.05, ..., 0.4
	linked_hidden_size	300, 400, ..., 2400
GLAM	learning rate	10^{-3}, 10^{-4}
	batch size	4, 32, 258
	embedding dimension	256, 1024

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aidi.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

80% and 20% of the data, respectively. The split is stratified with respect to the senescent cell viability to make sure that its distribution is similar in the training and test subsets (i.e., compounds were ordered by senescent cell viability, and every fifth compound was assigned to the test set). We chose this data splitting strategy because senescent cell viability is the most important quantity in the datasets that can be measured experimentally, directly related to senolytic potency. The training set also acts as the model selection dataset (i.e., used to identify the best combinations of hyperparameters for each learning algorithm, as shown in Figure 1B).

In the second experiment, the same protocol is applied to the Gil dataset.

In the third experiment, the Collins and the Gil dataset are merged and divided into fivefolds, ensuring that each fold contains a similar proportion of actives and inactives, as well as compounds from both datasets with a stratified random split. The hyperparameters identified with the model selection subset from the Gil dataset are reused.

In the last experiment, the Gil dataset acts as the training set while four test sets are assembled from the Collins dataset. Each dissimilar test set contains all the compounds from the Collins dataset that show less than X% Tanimoto similarity [74] to any molecule in the Gil dataset, acting as the training set: four dissimilar test sets are created with X in {99, 70, 50, 30} to gradually make it more challenging to predict actives. The Tanimoto similarity between two compounds is computed as the Jaccard index between the Morgan fingerprint of these compounds (2,048 bits, radius of 2). The hyperparameters identified with the model selection subset from the Gil dataset are also reused. In this experiment, this does not add any bias since the models are evaluated with dissimilar data.

5.6 | Model Selection

In each experiment, combinations of hyperparameters are used to evaluate the model with performance metrics. For CART, RF, XGB, SVM, and GLAM, we perform a grid search to find the best hyperparameters for both the classifier and the regression versions of the learning algorithm. Only the model selection dataset (specified for each experiment, Figure 2) is used to identify the best hyperparameters following a fivefold cross-validation protocol. This process is conducted with respect to two metrics per algorithm (Figure 1B), with or without salt removal from the SMILES. Consequently, this procedure leads to the final evaluation of eight models from the five algorithms. As we do not expect competitive results from the baseline algorithms, logistic regression and linear regression, we used default values for their hyperparameters. The hyperparameters of the D-MPNN are tuned with Bayesian optimization as implemented in the Chemprop package [23]. The best hyperparameters of each model for all the experiments are reported in Tables S1-16.

5.7 | Performance Metrics

Our aim is to build computational models to predict senolytic compounds, that is, identifying subsets of compounds with high probability of being senolytic to test in vitro. The metrics chosen to evaluate our models must take into account this context: only the top predictions of the models really matter, so it is important to assess the quality of the top predictions. To do this, we report the $NEF_{5\%}$ score (Net Enrichment Factor at 5%) [75, 76] and the hit rate, which is the proportion of senolytics among the top predicted compounds. These metrics can be compared across various test sets and are precisely defined in Equation (5) and 6, where N_{seno} is the number of senolytics in the test set. In this study, $NEF_{5\%}$ is selected as the primary metric since it is most aligned with the virtual screening goal. Additionally, other classical metrics are also reported to assess their relevance.

$$EF_{5\%} = \frac{number\ of\ senolytics\ in\ top\ 5\%\ predicted\ compounds}{0.05 \cdot N_{seno}} \tag{2}$$

$$EF_{max,5\%} = \frac{\min\left(0.05 \cdot number\ of\ compounds\ in\ test\ set,\ N_{seno}\right)}{0.05 \cdot N_{seno}} \tag{3}$$

$$NEF_{5\%} = \frac{EF_{5\%}}{EF_{\max,5\%}} \in [0,\ 1] \tag{4}$$

$$Hit\ rate_{top\ N_seno} = \frac{number\ of\ senolytics\ in\ top\ N_{seno}\ predicted\ compounds}{N_{seno}} \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_{i,predicted} - y_{i,measured}\right)^2}{N}} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_{i,\mathrm{predicted}} - y_{i,\mathrm{measured}}\right)^2}{\sum_{i=1}^{N}\left(\overline{y}_{\mathrm{measured}} - y_{i,\mathrm{measured}}\right)^2} \tag{7}$$

$$R_p = \frac{\sum_{i=1}^{N}\left(y_{i,\mathrm{predicted}} - \overline{y}_{\mathrm{predicted}}\right)\cdot\left(y_{i,\mathrm{measured}} - \overline{y}_{\mathrm{measured}}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_{i,\mathrm{predicted}} - \overline{y}_{\mathrm{predicted}}\right)^2 \cdot \sum_{i=1}^{N}\left(y_{i,\mathrm{measured}} - \overline{y}_{\mathrm{measured}}\right)^2}} \tag{8}$$

5.8 | Statistical Analysis

Pre-processing of data is described in the Data processing and features subsection. A statistical analysis was conducted to measure the correlation between the hit rate and the RMSE (Figure 5D and Table S17), in which the sample size is indicated in the caption with n. The two-tailed *p*-value of the Student t-test is reported. Kruskal–Wallis test employed in Figure 5A–C showed significant difference between the algorithms ($p < 0.0001$ for the three metrics), and post-hoc Mann–Whitney U tests between CART and the other algorithms with Benjamini–Hochberg correction are reported in the figure. Statistical analyses were carried out in Python 3.8.18 with scipy 1.10.1 and statsmodels 0.14.1.

Acknowledgments

This work was supported by the Wolfson Foundation and the Royal Society for a Royal Society Wolfson Fellowship [P.J.B.]. Core support from MRC (MC_U120085810) and grants from CRUK (C15075/A28647 and DRCPGM-Jun25/100003.) funded research in J. Gil's laboratory.

Funding

This work was supported by the Royal Society, Wolfson Foundation, MRC and CRUK.

Ethics Declarations

J. G. has acted as a consultant for Unity Biotechnology, Geras Bio, Myricx Pharma Ltd., and Merck KGaA; owns equity in

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aid.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Geras Bio and share options in Myricx Pharma Ltd. and is a named inventor in a patent WO2022090746A1 pending, unrelated to the work described here. J.G.'s lab received Pfizer and Unity Biotechnology unrelated to the work described here.

Conflicts of Interest

The authors declare no conflicts of interest.

Code Availability

The code to train and use the best model is available at https://github.com/alexisdougha/virtual-screening-senolytics. It also contains the weights of the models.

Data Availability Statement

The Gil dataset is available at https://github.com/alexisdougha/virtual-screening-senolytics. The Collins dataset [19] is available at https://www.nature.com/articles/s43587-023-00415-z#Sec39.

References

1. D. C. Swinney, "Phenotypic vs. Target-Based Drug Discovery for First-in-Class Medicines," *Clinical Pharmacology and Therapeutics* 93 (2013): 299–301.

2. S. Malandraki-Miller and P. R. Riley, "Use of Artificial Intelligence to Enhance Phenotypic Drug Discovery," *Drug Discovery Today* 26 (2021): 887–901.

3. Ó. Álvarez-Machancoses and J. L. Fernández-Martínez, "Using Artificial Intelligence Methods to Speed up Drug Discovery," *Expert Opinion on Drug Discovery* 14 (2019): 769–777.

4. F. Vincent, A. Nueda, J. Lee, et al., "Phenotypic Drug Discovery: Recent Successes, Lessons Learned and New Directions," *Nature Reviews Drug Discovery* 21 (2022): 899–914.

5. V. Blay, B. Tolani, S. P. Ho, and M. R. Arkin, "High-Throughput Screening: Today's Biochemical and Cell-Based Approaches," *Drug Discovery Today* 25 (2020): 1807–1821.

6. M. Korn, C. Ehrt, F. Ruggiu, M. Gastreich, and M. Rarey, "Navigating Large Chemical Spaces in Early-Phase Drug Discovery," *Current Opinion in Structural Biology* 80 (2023): 102578.

7. A. V. Sadybekov and V. Katritch, "Computational Approaches Streamlining Drug Discovery," *Nature* 616 (2023): 673–685.

8. P. J. Ballester, "The AI Revolution in Chemistry Is Not that Far Away," *Nature* 624 (2023): 252–252.

9. K. Zhang, X. Yang, W. Yifei, et al., "Artificial intelligence in drug development," *Nature Medicine* 31 (2025): 45–59.

10. G. Ghislat, S. Hernandez-Hernandez, C. Piyawajanusorn, and P. J. Ballester, "Data-Centric Challenges with the Application and Adoption of Artificial Intelligence for Drug Discovery," *Expert Opinion on Drug Discovery* 19 (2024): 1297–1307.

11. F. Ren, X. Ding, M. Zheng, et al., "AlphaFold Accelerates Artificial Intelligence Powered Drug Discovery: Efficient Discovery of a Novel CDK20 Small Molecule Inhibitor," *Chemical Science* 14 (2023): 1443–1452.

12. I. Wallach, D. Bernard, K. Nguyen, et al., "AI Is a Viable Alternative to High Throughput Screening: A 318-Target Study," *Scientific Reports* 14 (2024): 7526.

13. V. Gorgoulis, P. D. Adams, A. Alimonti, et al., "Cellular Senescence: Defining a Path Forward," *Cell* 179 (2019): 813–827.

14. D. Muñoz-Espín and M. Serrano, "Cellular Senescence: From Physiology to Pathology," *Nature Reviews Molecular Cell Biology* 15 (2014): 482–496.

15. W. Huang, L. J. Hickson, A. Eirin, J. L. Kirkland, and L. O. Lerman, "Cellular Senescence: The Good, the Bad and the Unknown," *Nature Reviews Nephrology* 18 (2022): 611–627.

16. P. J. Thompson, A. Shah, V. Ntranos, et al., "Targeted Elimination of Senescent Beta Cells Prevents Type 1 Diabetes," *Cell Metabolism* 29 (2019): 1045–1060.e10.

17. J. -W. Zhang, D. Zhang, and B. -P. Yu, "Senescent Cells in Cancer Therapy: Why and How to Remove Them," *Cancer Letters* 520 (2021): 68–79.

18. V. Smer-Barreto, A. Quintanilla, R. J. R. Elliott, et al., "Discovery of Senolytics Using Machine Learning," *Nature Communications* 14 (2023): 3445.

19. F. Wong, S. Omori, N. M. Donghia, E. J. Zheng, and J. J. Collins, "Discovering Small-Molecule Senolytics with Deep Neural Networks," *Nature Aging* 3 (2023): 734–750.

20. S. M. Corsello, J. A. Bittker, Z. Liu, et al., "The Drug Repurposing Hub: A Next-Generation Drug Library and Information Resource," *Nature Medicine* 23 (2017): 405–408.

21. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794 (Association for Computing Machinery, 2016), https://doi.org/10.1145/2939672.2939785.

22. G. Landrum, P. Tosco, K. Brian, et al., "Rdkit/Rdkit:2024_03_5 (Q1 2024 Release)," Zenodo (2024), https://doi.org/10.5281/ZENODO.591637.

23. E. Heid, K. P. Greenman, Y. Chung, et al., "Chemprop: A Machine Learning Package for Chemical Property Prediction," *Journal of Chemical Information and Modeling* 64 (2024): 9–17.

24. A. Guerrero, B. S. Nás Herranz, et al., "Cardiac Glycosides Are Broad-Spectrum Senolytics," *Nature Metabolism* 1 (2019): 1074–1088.

25. F. Triana-Martínez, P. Picallos-Rabina, S. Da Silva-Álvarez, et al., "Identification and Characterization of Cardiac Glycosides as Senolytic Compounds," *Nature Communications* 10 (2019): 4731.

26. W. Li, L. Qin, R. Feng, et al., "Emerging Senolytic Agents Derived From Natural Products," *Mechanisms of Ageing and Development* 181 (2019): 1–6.

27. D. Boldini, L. Friedrich, D. Kuhn, and S. A. Sieber, "Machine Learning Assisted Hit Prioritization for High Throughput Screening in Drug Discovery," *ACS Central Science* 10 (2024): 823–832.

28. A. M. Mroz, A. R. Basford, F. Hastedt, et al., "Cross-Disciplinary Perspectives on the Potential for Artificial Intelligence across Chemistry," *Chemical Society Reviews* 54 (2025): 5433–5469.

29. S. Zoffmann, M. Vercruysse, F. Benmansour, et al., "Machine Learning-Powered Antibiotics Phenotypic Drug Discovery," *Scientific Reports* 9 (2019): 5013.

30. F. Wong, C. de la Fuente-Nunez, and J. J. Collins, "Leveraging Artificial Intelligence in the Fight against Infectious Diseases," *Sciencee* 381 (2023): 164–170.

31. S. Hernández-Hernández, S. Vishwakarma, and P. Ballester, "Conformal prediction of small-molecule drug resistance in cancer cell lines," in Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications, 92-108 (PMLR, 2022).

32. S. Vishwakarma, S. Hernandez-Hernandez, and P. J. Ballester, "Graph Neural Networks Are Promising for Phenotypic Virtual Screening on Cancer Cell Lines," *Biology Methods & Protocols* 9, no. 1 (2024): bpae065.

33. M. C. R. Melo, J. R. M. A. Maasch, and C. de la Fuente-Nunez, "Accelerating Antibiotic Discovery through Artificial Intelligence," *Communications Biology* 4 (2021): 1–13.

34. C. Gorgulla, A. Boeszoermenyi, Z. -F. Wang, et al., "An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens," *Nature* 580 (2020): 663–668.

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aidi.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

35. D. Paul, G. Sanap, S. Shenoy, et al., "Artificial Intelligence in Drug Discovery and Development," *Drug Discovery Today* 26 (2021): 80–93.

36. P. Schneider, W. P. Walters, A. T. Plowright, et al., "Rethinking Drug Design in the Artificial Intelligence Era," *Nature Reviews Drug Discovery* 19 (2020): 353–364.

37. J. Mao, J. Akhtar, X. Zhang, et al., "Comprehensive Strategies of Machine-Learning-Based Quantitative Structure-Activity Relationship Models," *iScience* 24 (2021): 103052.

38. S. K. Niazi and Z. Mariam, "Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review," *International Journal of Molecular Sciences* 24 (2023): 11488.

39. Danishuddin, and A. U. Khan, "Descriptors and Their Selection Methods in QSAR Analysis: Paradigm for Drug Desig" *Drug Discovery Today* 21 (2016): 129–1302.

40. A. Tropsha, O. Isayev, A. Varnek, G. Schneider, and A. Cherkasov, "Integrating QSAR Modelling and Deep Learning in Drug Discovery: The Emergence of Deep QSAR," *Nature Reviews Drug Discovery* 23 (2024): 141–155.

41. J. Deng, Z. Yang, H. Wang, et al., "A Systematic Study of Key Elements Underlying Molecular Property Prediction," *Nature Communications* 14 (2023): 6395.

42. T. Janela and J. Bajorath, "Simple Nearest-Neighbour Analysis Meets the Accuracy of Compound Potency Predictions Using Complex Machine Learning Models," *Nature Machine Intelligence* 4 (2022): 1246–1255.

43. Q. Guo, S. Hernandez-Hernandez, P. J. Ballester, et al., Scaffold Splits Overestimate Virtual Screening Performance, *Artificial Neural Networks and Machine Learning – ICANN*, eds. Wand M., Malinovska K., Schmidhuber J., and Tetko I. V. (Springer Nature, 2024): 58–72, https://doi.org/10.1007/978-3-031-72359-9_5.

44. Q. Guo, S. Hernandez-Hernandez, and P. J. Ballester, "UMAP-Based Clustering Split for Rigorous Evaluation of AI Models for Virtual Screening on Cancer Cell Lines*," *Journal of Cheminformatics* 17 (2025): 1.

45. Z. -F. Tao, L. Hasvold, L. Wang, et al., "Discovery of a Potent and Selective BCL-XL Inhibitor with In Vivo Activity," *ACS Medicinal Chemistry Letters* 5 (2014): 1088–1093.

46. L. Wang, G. A. Doherty, A. S. Judd, et al., "Discovery of A-1331852, a First-in-Class, Potent, and Orally-Bioavailable BCL-XL Inhibitor," *ACS Medicinal Chemistry Letters* 11 (2020): 1829–1836.

47. R. Yosef, N. Pilpel, R. Tokarsky-Amiel, et al., "Directed Elimination of Senescent Cells by Inhibition of BCL-W and BCL-XL," *Nature Communications* 7 (2016): 11190.

48. A. T. Balaban, "Highly Discriminating Distance-Based Topological Index," *Chemical Physics Letters* 89 (1982): 399–404.

49. L. H. Hall, Brian Mohney, and L. B. Kier, "The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs," *Journal of Chemical Information and Computer Sciences* 31 (1991): 76–82.

50. L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?," in Proceedings of the 36th International Conference on Neural Information Processing Systems, 507-520 (Curran Associates Inc., 2022).

51. D. Jiang, Z. Wu, C. -Y. Hsieh, et al., "Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models," *Journal of Cheminformatics* 13 (2021): 12.

52. J. Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug Discovery with Explainable Artificial Intelligence," *Nature Machine Intelligence* 2 (2020): 573–584.

53. K. K. Kırboğa, S. Abbasi, and E. U. Küçüksille, "Explainability and White Box in Drug Discovery," *Chemical Biology & Drug Design* 102 (2023): 217–233.

54. G. A. Landrum and S. Riniker, "Combining IC50 or Ki Values From Different Sources Is a Source of Significant Noise," *Journal of Chemical Information and Modeling* 64 (2024): 1560–1567.

55. C. Kramer, T. Kalliokoski, P. Gedeck, and A. Vulpetti, "The Experimental Uncertainty of Heterogeneous Public Ki Data," *Journal of Medicinal Chemistry* 55 (2012): 5165–5173.

56. T. Kalliokoski, C. Kramer, A. Vulpetti, and P. Gedeck, "Comparability of Mixed IC50 Data – A Statistical Analysis," *PLoS ONE* 8 (2013): e61007.

57. K. Mansouri, Jé T. Moreira-Filho, C. N. Lowe, et al., "Free and Open-Source QSAR-Ready Workflow for Automated Standardization of Chemical Structures in Support of QSAR Modeling," *Journal of Cheminformatics* 16 (2024): 19.

58. D. Dolciami, E. Villasclaras-Fernandez, C. Kannas, et al., "CanSAR Chemistry Registration and Standardization Pipeline," *Journal of Cheminformatics* 14 (2022): 28.

59. A. P. Bento, A. Hersey, F. Eloy, et al., "An Open Source Chemical Structure Curation Pipeline Using RDKit," *Journal of Cheminformatics* 12 (2020): 51.

60. F. Pedregosa, G. Varoquaux, and A. Gramfort, "Scikit-learn: Machine Learning in Python. Mach. Learn," *Journal of Machine Learning Research* 12 (2011): 2825–2830.

61. J. Bergstra, D. Yamins, and D. Cox, "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures," in Proceedings of the 30th International Conference on Machine Learning, 115-123 (PMLR, 2013).

62. R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford, "Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships," *Journal of Chemical Information and Modeling* 56 (2016): 2353–2360.

63. K. Yang, K. Swanson, W. Jin, et al., "Analyzing Learned Molecular Representations for Property Prediction," *Journal of Chemical Information and Modeling* 59 (2019): 3370–3388.

64. J. M. Stokes, K. Yang, K. Swanson, et al., "A Deep Learning Approach to Antibiotic Discovery," *Cell* 180 (2020): 688–702.e13.

65. G. Liu, D. B. Catacutan, K. Rathod, et al., "Deep Learning-Guided Discovery of an Antibiotic Targeting Acinetobacter Baumannii," *Nature Chemical Biology* 19 (2023): 1342–1350.

66. F. Wong, E. J. Zheng, J. A. Valeri, et al., "Discovery of a Structural Class of Antibiotics with Explainable Deep Learning," *Nature* 626 (2024): 177–185.

67. Y. Guan, C. W. Coley, H. Wu, et al., "Regio-Selectivity Prediction with a Machine-Learned Reaction Representation and on-the-Fly Quantum Mechanical Descriptors," *Chemical Science* 12 (2021): 2198–2208.

68. K. A. Spiekermann, L. Pattanaik, and W. H. Green, "Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy," *Journal of Physical Chemistry A* 126 (2022): 3976–3986.

69. M. Colomba, S. Benedetti, D. Fraternale, et al., "Nrf2-Mediated Pathway Activated by Prunus Spinosa L. (Rosaceae) Fruit Extract: Bioinformatics Analyses and Experimental Validation," *Nutrients* 15 (2023): 2132.

70. K. Swanson, P. Walther, J. Leitz, et al., "ADMET-AI: A Machine Learning ADMET Platform for Evaluation of Large-Scale Chemical Libraries," *Bioinformatics* 40 (2024): btae416.

71. Y. Li, C. -Y. Hsieh, R. Lu, et al., "An Adaptive Graph Learning Method for Automated Molecular Interactions and Properties Predictions," *Nature Machine Intelligence* 4 (2022): 645–651.

72. M. Simonovsky and N. Komodakis, "Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs, (2017)," https://doi.org/10.48550/arXiv.1704.02901 Preprint at.

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aidi.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

73. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural Message Passing for Quantum Chemistry, (2017), https://doi.org/10.48550/arXiv.1704.01212 Preprint at.

74. D. Bajusz, A. Rácz, and K. Héberger, "Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations?," *Journal of Cheminformatics* 7 (2015): 20.

75. S. Liu, M. Alnammi, S. S. Ericksen, et al., "Practical Model Selection for Prospective Virtual Screening," *Journal of Chemical Information and Modeling* 59 (2019): 282–293.

76. V. -K. Tran-Nguyen, M. Junaid, S. Simeon, and P. J. Ballester, "A Practical Guide to Machine-Learning Scoring for Structure-Based Virtual Screening," *Nature Protocols* 18 (2023): 340–3511.

Supporting Information

Additional supporting information can be found online in the supporting information section.

29439981, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aidi.202600013 by Imperial College London, Wiley Online Library on [08/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License