HHS Public Access

Author manuscript

Chem Eng Sci. Author manuscript; available in PMC 2024 November 05.

Published in final edited form as:

Chem Eng Sci. 2023 November 05; 281: . doi:10.1016/j.ces.2023.119086.

Machine Learning Methods for Endocrine Disrupting Potential Identification Based on Single-Cell Data

Zahir Aghayev[1,2], **Adam T. Szafran**[3], **Anh Tran**[4,5], **Hari S. Ganesh**[6], **Fabio Stossi**[3,7], **Lan Zhou**[8], **Michael A. Mancini**[3,7], **Efstratios N. Pistikopoulos**[4,5], **Burcu Beykal**[1,2,*]

[1]Department of Chemical and Biomolecular Engineering, University of Connecticut, Storrs, CT

[2]Center for Clean Energy Engineering, University of Connecticut, Storrs, CT

[3]Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX

[4]Artie McFerrin Department of Chemical Engineering, Texas A&M University, College Station, TX

[5]Texas A&M Energy Institute, Texas A&M University, College Station, TX

[6]Discipline of Chemical Engineering, Indian Institute of Technology Gandhinagar, Palaj, Gandhinagar, Gujarat - 382055, India

[7]GCC Center for Advanced Microscopy and Image Informatics, Houston, TX

[8]Department of Statistics, Texas A&M University, College Station, TX

Abstract

Humans are continuously exposed to a variety of toxicants and chemicals which is exacerbated during and after environmental catastrophes such as floods, earthquakes, and hurricanes. The hazardous chemical mixtures generated during these events threaten the health and safety of humans and other living organisms. This necessitates the development of rapid decision-making tools to facilitate mitigating the adverse effects of exposure on the key modulators of the

*Corresponding Author. Tel: +1 (860) 486-2756, burcu.beykal@uconn.edu.

Author Contributions

Conceptualization: Adam T. Szafran, Burcu Beykal, Fabio Stossi, Lan Zhou, Michael A. Mancini, Efstratios N. Pistikopoulos
Methodology: Zahir Aghayev, Burcu Beykal, Adam T. Szafran
Software: Zahir Aghayev, Burcu Beykal
Formal analysis: Zahir Aghayev
Investigation: Zahir Aghayev, Burcu Beykal, Adam T. Szafran, Fabio Stossi, Anh Tran, Hari S. Ganesh
Resources: Michael A. Mancini, Efstratios N. Pistikopoulos, Burcu Beykal
Data curation: Adam T. Szafran.
Writing – original draft: Zahir Aghayev
Writing – review & editing: Adam T. Szafran, Anh Tran, Hari S. Ganesh, Fabio Stossi, Lan Zhou, Michael A. Mancini, Efstratios N. Pistikopoulos, Burcu Beykal
Visualization: Zahir Aghayev
Supervision: Burcu Beykal, Adam T. Szafran, Fabio Stossi, Lan Zhou, Michael A. Mancini, Efstratios N. Pistikopoulos
Funding acquisition: Michael A. Mancini, Efstratios N. Pistikopoulos, Burcu Beykal

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

endocrine system, such as the estrogen receptor alpha (ERα), for example. The mechanistic stages of the estrogenic transcriptional activity can be measured with high content/high throughput microscopy-based biosensor assays at the single-cell level, which generates millions of object-based minable data points. By combining computational modeling and experimental analysis, we built a highly accurate data-driven classification framework to assess the endocrine disrupting potential of environmental compounds. The effects of these compounds on the ERα pathway are predicted as being receptor agonists or antagonists using the principal component analysis (PCA) projections of high throughput, high content image analysis descriptors. The framework also combines rigorous preprocessing steps and nonlinear machine learning algorithms, such as the Support Vector Machines and Random Forest classifiers, to develop highly accurate mathematical representations of the separation between ERα agonists and antagonists. The results show that Support Vector Machines classify the unseen chemicals correctly with more than 96% accuracy using the proposed framework, where the preprocessing and the PCA steps play a key role in suppressing experimental noise and unraveling hidden patterns in the dataset.

Keywords

Machine learning; Endocrine disrupting chemicals; Estrogen receptor activity; Predictive modeling; Classification analysis; High throughput microscopy

1. Introduction

The impact of both natural and human-made chemical compounds on human health and the environment has been analyzed during the past several decades [1]. As a result of endocrine disrupting chemicals' (EDCs') interference with many hormone pathways, they have been associated with numerous detrimental human health outcomes, including altered nervous, immune, respiratory, and cardiovascular system function resulting in increased rates of metabolic disorders (diabetes, obesity, abnormal growth), neurologic disorders (cognitive impairment, learning disabilities), cardiovascular disorders, cancer, and many other patho-physiological effects [2,3]. Likewise, these chemicals impact wildlife because of their high potential to distort endocrine system function, particularly during embryonic stages [4]. EDCs have received both scholarly and societal attention, and they've been regarded as substances of very high concern [5].

One of the key targets of EDCs is the estrogen receptor (ER), a ligand-inducible transcription factor, which belongs to the steroid hormone superfamily of nuclear receptors that mediates the growth, development, and physiology of the reproductive system [6,7]. While compounds with agonistic behavior have the potential of binding to ER and inducing an active confirmation of the receptor (i.e., 17β-estradiol), antagonist compounds can induce an inactive conformation and block other molecules (including endogenous hormones) from binding to the receptor (i.e., tamoxifen). Both antagonists and agonists bind to the same site in the ER ligand-binding domain exerting opposite outcomes which can be clearly distinguished at the molecular and structural level [8]. Due to the potential of toxicants and their unknown complex mixtures binding to the ER, identifying chemicals that will behave

as an agonist or an antagonist plays a crucial role in evaluating the hazardous health effects to their exposure.

Yet, it is a very costly and nearly impossible task to exhaustively evaluate every single chemical and its mixtures for agonistic and/or antagonistic estrogenic behavior through pure experimental strategies [9]. With increased computational power over the last couple of decades, data-driven modeling is shown to be an effective methodology for addressing such challenging problems across many disciplines, including process monitoring, optimization, and control in chemical engineering [10-25], as well as toxicity evaluation and mitigation in environmental sciences [26-31]. Likewise, data-driven analysis poses an untapped potential to aid in modeling biological responses of receptors for facilitated screening of known and unknown environmental toxicants.

To this end, *in vitro* screening and computational techniques are used together to facilitate the identification of the endocrine disrupting potential of chemicals, and to distinguish androgen receptor (AR) [32, 33] and ER binders as agonists and antagonists. Zang et al. [34] developed *in silico* quantitative structure activity relationship (QSAR) model by using ToxCast and Tox21 dataset and machine learning (ML) methods. They handled the imbalanced data distribution by a cluster selection strategy and linearly classified active and inactive compounds with 81.6% total testing accuracy. Furthermore, Chen et al. [35] developed computational models by representing each molecule with seven fingerprints to predict AR and ER binders via k-nearest neighbor (kNN), C4.5 decision tree (C4.5 DT), naïve Bayes (NB), and support vector machine (SVM) algorithms. Their Extended Fingerprint-SVM model resulted being the best model for predicting ER binders with 79% accuracy. Moreover, deep neural networks (DNN) also received significant attention in cheminformatics where Mayr et al. [36] introduced their DeepTox pipeline for predicting toxicity based on Deep Learning (DL). In this study, authors compared DNNs, SVM (Tanimoto kernel), Random Forest (RF), and elastic net methods based on the area under the receiver operating characteristic curve (AUC) performance criterion. Their results showed that the DL outperformed other classification algorithms with a 0.837 average AUC score. Similarly, Chierici et al. [37] offered a framework by building DL and SVM models to predict the endocrine disrupting potential and binding activity of chemical toxicants. In another study by Ciallella et al. [38], classic ML and DL techniques were implemented to model 18 Toxicity Forecaster assays that address ER pathway perturbations. According to the results of this work, consensus forecasts continue to be the best strategy for computational toxicity assessment due to the absence of universal standards for the selection of chemical descriptors and algorithms. Idakwo et al. [39] and Wank et al. [40] notable review articles and Russo et al. [41] comparative study provide a comprehensive summary of the application of ML algorithms for *in silico* toxicity predictions.

Using data collected from a single Toxicity Forcaster assay (Assay ID: A10, OT_ERa_EREGFP_0120), we previously demonstrated that the population-averaged high content imaging data can be used to accurately classify the agonist and antagonist benchmark substances. This *in vitro* assay uses GFP-tagged ER (GFP-ER) and the PRL-HeLa cell line to allow the direct visualization of ER binding to promoter DNA at the level of the single cell and quantifies multiple aspects of the ER signaling pathway [42,43].

Through consideration of the full set of quantified features and the use of linear and nonlinear models, agonist and antagonist chemicals were classified with more than 90% accuracy [44,45]. We further extended this study to assess the predictive capabilities of RF and SVM algorithms when a non-central gamma probability distribution function is used instead of population averaging the high content imaging data [46]. Yet, there are still unanswered research questions to be addressed when coupling ML algorithms with ER activity data: (1) How to improve the accuracy of predictions and achieve more than 90% accuracy?; (2) What are the effects of data preparation on ER activity predictions?; and (3) How to benefit from single-cell level information to develop accurate mathematical representations of the endocrine disruption potential of chemical compounds?

Unlike our previous investigations, here, we address these open research questions by proposing an integrated, data-driven approach for the classification of endocrine disrupting potential of substances using the high-dimensional single-cell level information, principal component analysis (PCA), and nonlinear ML algorithms. Due to the nonlinear and discontinuous nature of the dataset at the single-cell level, 2-class SVM with radial basis kernel and RF classifiers are utilized in this study, while PCA allowed us to identify the components with the highest variability to be used as the predictors in the trained classifiers. We follow rigorous preprocessing and modeling strategies where the unbiased predictive performance of the SVM and RF is shown to achieve more than 96% and 94% correct classification accuracy on benchmark chemicals, respectively, achieving significantly better results compared to the aforementioned studies.

The rest of the manuscript is organized as follows. The experimental and computational methodologies of this study are presented in Section 2. Also, this section provides an extensive discussion of the model, model validation methods, and the performance assessment criteria that were used in this study. Then in Section 3, a comparative analysis of employed classification algorithms and the effect of data preparation on the final classification performance are discussed comprehensively. The final remarks on the manuscript are drawn in Section 4.

2. Materials and Methods

2.1. Reference Chemicals

A set of 45 estrogenic EDC reference chemicals with well-documented *in vivo* activities were provided in DMSO via an agreement with Dr. Keith Houck (US EPA, Research Triangle Park, NC). This library consists of ERα agonists with strong, moderate, weak, or very weak activity, inhibitors of ERα (antagonists/anti-estrogens), and inactive chemicals. To improve the agonist/antagonist balance of the reference chemicals, the library is supplemented with 7 additional chemicals with well-established antagonist/anti-estrogenic activity. For control treatments, samples were treated with alternatively sourced 17β-estradiol (E2, E2-1, E2-2), 4-hydroxytamoxifen (4HT, 4HT-1, 4HT-2), and raloxifene. Finally, all cell treatments included a media control that contains 0.1% DMSO (Media) for a total of 60 unique chemical samples in the experimental data set (Table S30).

2.2. Sample Treatment

To generate the experimental data, GFP-ER:PRL-HeLa cells [42,43] were exposed to each chemical with concentrations ranging between 10 nM and 100 μM (Table S30) for a time of 2 hours. Right before the compound treatment, the stock compound solutions were transferred to 96-well working dilution plates in phenol red-free DMEM with 5% SD-FBS. Working dilutions of compounds were then added in quadruplicate (i.e., 4 technical replicates of each sample) to assay plates containing cells. For all experiments, negative control wells containing 0.1% DMSO and positive control wells containing E2 ranging from 10 nM to 1000 nM were included in the plate layout. All dilutions and cell treatments using reference chemicals were performed using the Beckman Biomek NX liquid-handling robotic platform.

2.3. Sample Processing

After the incubation period was completed, cells were fixed using 4% EM-grade formaldehyde in PEM buffer (80 mM potassium PIPES [pH 6.8], 5 mM EGTA, and 2 mM MgCl2) and quenched with 0.1 M ammonium chloride for 10 min. Cell membranes were disrupted by incubating samples with a 0.5% Triton X-100 solution for 10 minutes. Nuclei were stained using DAPI (1 μg/ml) for 10 minutes. All samples were stored in PBS buffer containing calcium, magnesium, and sodium azide at 4°C prior to imaging.

2.4. Sample Imaging and Quantification

Automated imaging was carried out using a Vala Sciences IC-200 (San Diego, CA) image cytometer. Image acquisition was performed with a 20x objective. To ensure capturing visible arrays, Z-stacks of the GFP channel were captured at 1 μm intervals (for a total of 5 μm). Nuclear array segmentation and automated image analyses were performed using custom measurement and reporting routines incorporated into myImageAnalysis/Pipeline Pilot software [47]. Cell, nucleus, and array (if present) were identified and metrics describing shape and intensity features quantified. Aggregated cells, mitotic cells, and apoptotic cells were removed using filters based on nuclear size, nuclear shape, nuclear intensity, and nucleus-to-nucleus distance. In addition to primary shape and intensity-based extracted features, three secondary features were calculated: (1) the ratio of the array GFP intensity to the nucleoplasm GFP intensity; (2) the fraction of the total GFP intensity in the cell that is localized in the nucleus; and (3) the ratio of the nuclear GFP intensity to the cytoplasmic GFP intensity. This created a dataset size of 162,525 observations × 79 features that is passed to the computational analysis phase.

2.5. Computational Methodology

Our computational analysis pipeline is summarized in Fig 1. First, the multidimensional imaging data from the high throughput microscopy and high content analysis undergo a preprocessing phase to be converted into a more efficient form. The cleaned dataset is then passed to the PCA step, where the experimental noise is suppressed by performing eigendecomposition on the covariance matrix. We then apply and validate two nonlinear classification algorithms (SVM and RF) to model the separation between ER agonists and antagonists using the training compounds. Finally, the predictive performance of these

binary classifiers is quantified and compared using several metrics. Details of each analysis step are provided in Fig 1.

2.5.1. Data Preprocessing—An analysis pipeline was created using Pipeline Pilot software (Biovia) to perform all data normalization, percentile identification, and engineered data set creation. All chemical screening data were normalized to E2-treated control samples using the robust z-score calculation based on single cell data. In experiments that spanned multiple sample plates, the median and MAD values were calculated for each plate, the median MAD value across all plates was determined, and the robust z-score was calculated using the per-plate median value and the overall median MAD value.

We also perform feature engineering to expand the dataset using percentiles. Percentiles were determined by identifying the single cell features falling into the 0 – 10th percentile range (L10), 45th – 55th percentile range (M10), the 90th – 100th percentile range (H10), and the 25th – 75th percentile range (M50) from each sample. Experimental data sets containing all single cell data and engineered data sets containing only the L10 + M10 + H10 or only the M50 single cell features were also divided into 2 subsets depending on the detection of visible GFP-ER nuclear spots (all cells and array positive cells) used for subsequent model development.

Next, each individual dataset was scanned for missing entries and cleaned prior to classification modeling. Missing data are a prevalent interdisciplinary issue, and the way we handle this problem can have a significant influence on the reliability of statistical inferences and the important conclusions drawn from a data analysis. Understanding the pattern of this problem and then taking mitigative actions such as listwise deletion, data amputation, and data recovery are some of the main procedures that should be followed in these cases [48].

Following the missing data check, data cleaning was performed where the redundant features and the inactive compounds were removed prior to the classification analysis (Table 3). In addition, only 17β-Estradiol, 4-hydroxytamoxifen, and raloxifene treatments were retained in the analysis dataset as representatives of control treatments whereas E2, E2-1, E2-2, 4HT, 4HT-1, and 4HT-2 are removed. This results in having 41 active chemical compounds in each dataset, Table 1 shows the data size for all cases while Table 2 shows the list of active chemicals used in this study.

2.5.2. Principal Component Analysis—Once the data preprocessing was completed, each dataset was centered and scaled to perform PCA. The central concept of PCA is to derive a low-dimensional set (i.e., reducing the number of features) of the data while preserving its variability [49]. A set of new orthogonal variables called principal components (PCs) is created by projecting the data onto a new coordinate system and these are expressed as the linear combinations of the original imaging features. The mathematical theory behind this analysis root in the eigendecomposition of the positive semi-definite matrices [50].

The PCs obtained from this analysis are ranked based on the variability represented by each one of them. For example, the greatest variance in the dataset will lie in the first component while the least great variance in the dataset will lie in the last component. To identify a

lower-dimensional set, we observe the proportion of variance explained by each PC. The explained variance values for the top 10 PCs are shown in Fig 2 on a scree plot for each dataset that describes only array positive cells. The scree plot of other datasets may be found in Fig S1.

Fig 2 shows that the individual explained variance of PCs for all datasets drop significantly after the first PC. In Fig 2a, the first PC explains 78.87% of the variance whereas the second PC explains only 9.45% of the variance in the dataset derived from the original features. Similarly, for the L10-M10-H10 case, we observe that the first PC explains 62.57% and the second PC explains 10.98% of the variance in the dataset. For the M50 case, we observe a 56.854% drop in explained variance as we go from the first PC to the second one. To determine the optimal number of PCs to retain in the analysis, we observe the inflection point or the “elbow” of scree plots. Based on Fig 2, the elbows are located at the third PC, where the cumulative proportion of explained variance of the first three PCs is 92.77%, 79.54%, and 91.07%, for the original, L10-M10-H10, and the M50 cases, respectively. Thus, we reduce the dimensionality of the dataset to 3 projected features from the PCA, where the sizes of each data matrix are provided in Table 4.

2.5.3. Classification Model Development—The PCA-transformed dataset then undergoes the compound-wise train-test split procedure prior to the modeling. The number of agonist and antagonist chemicals and their ER potency, as shown in Table 1, were taken into high consideration to have a balanced split across both ER activity classes. Accordingly, two-thirds of the total 12 antagonist chemicals were split randomly and allocated in the training set whereas the remaining one-third of antagonistic compounds are reserved for the testing set. Of the 29 compounds with agonistic behaviors, 18 were allocated in the training set, while the rest were used in the testing set. The train-test split of active chemicals according to their endocrine disrupting potential is summarized in Fig 3 and details on the dimensionality of each set are provided in Table 5. It is important to note that the same set of training and testing chemicals are used for analyzing each dataset to have a fair comparison and the test set is exclusively reserved to quantify the predictive capabilities of the trained classifiers on unseen data.

In this study, we use the two-class c-parametrized SVM with a Gaussian radial basis kernel and RF algorithms to distinguish between the endocrine disrupting potentials of the studied benchmark chemicals. These models are chosen due to the nonlinear and discontinuous nature of the datasets and their easy-to-implement training routines.

An SVM is a machine learning model that allows distinguishing observations between two classes by finding the separating hyperplane between the two classes of information. The linear c-SVM learning problem with soft margins is formulated as a convex optimization problem [51,52] where the objective is to find the maximum margin hyperplane while minimizing the number of misclassified points by the SVM (Eq. 1),

$$\begin{aligned} \min_{w,\, b,\, \xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ s.t. \quad & y_i\left(w^T x_i + b\right) \geq 1 - \xi_i \qquad \forall i = 1, \ldots, N \\ & \xi_i \geq 0 \qquad \forall i = 1, \ldots, N \end{aligned} \tag{1}$$

where x_i are the inputs of each sample i, y_i are the outputs of each sample i, w are the weights of the features, b is the bias of the hyperplane, ξ_i are the slack variables for samples i that are misclassified, N is the total number of samples in the training set, and C is the cost hyperparameter that controls the trade-off between the maximum margin

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \\ s.t. \quad & \sum_{i=1}^{N}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \qquad \forall i = 1, \ldots, N \end{aligned} \tag{2}$$

and the hinge loss. If the value of C is small the trained SVM model will have a wider margin and more tolerance for misclassifications whereas if the C is large, then the SVM will have a smaller margin to keep the misclassification value small. For deriving the nonlinear separating boundary between the two classes of information, the Lagrange dual problem of Eq. 1 with kernel trick is used (Eq. 2) and the decision variables for the SVM training become the dual variables (α_i).

In our work, we use the Gaussian radial basis kernel which also comes with an additional hyperparameter γ. These two hyperparameters that affect the learning process of the SVM model require tuning to achieve maximum predictive capability. This is done using a 6-fold cross-validated grid search where a total of 100 C and γ combinations are explored within the 10 evenly spaced values in the interval of [128, 1280] and [1/1280, 1/128], respectively. In cross validation step, instead of simply randomly splitting and shuffling the cell-level data, cross validation is performed per compound basis. As a result, there are 4 compounds in first 5 folds and 6 compounds in 6th fold (each compound containing their respective cell-level measurements). The median accuracy of the folds is selected to be the accuracy of the corresponding hyperparameter combination. The hyperparameter combination with the highest median accuracy score is deemed optimal and used to construct the final classification model.

As an alternative approach for binary nonlinear classification, we also use the RF model which is an ensemble of unpruned decision trees that are based on a bootstrap sampling of the training set and a randomly chosen subset of variables [53]. This machine learning approach also has two hyperparameters: (1) *ntree* – the number of trees to grow; and (2) *mtry* – the number of features to randomly sample at each split. Similar to the SVM tuning process, we perform 6-fold cross-validated grid search to find the optimum hyperparameter combinations that give the best predictive performance, where a total of 30 hyperparameter combinations are explored within 10 evenly spaced ntree and 3 mtry values in the ranges of [25,250] and [1,3], respectively. The SVM and RF algorithms are implemented in the

R (version 4.1.2) programming language using the `svm` function in the `e1071` library, and the `randomForest` function in the `randomForest` library, respectively. The RMarkdown documentation of the entire computational analysis is also provided in the Supplementary Material.

2.5.4. Model Performance Metrics—To assess the unbiased predictive performance of the trained models, we follow the testing approach that the performance metrics of unseen chemicals are measured on which the model has not been trained. We use a range of performance metrics including, accuracy, sensitivity, specificity, F_1 score, and balanced accuracy as shown in Eqs. 3-7. Here, an agonist compound that is misclassified as an antagonist is defined as a "false positive" or FP, while a misclassified antagonist compound is defined as a "false negative" or FN. Accordingly, the correctly classified agonist and antagonist compounds are defined as "true positive" (TP) and "true negative" (TN) respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

$$F_1\,Score = \frac{2TP}{2TP + FP + FN} \tag{6}$$

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right). \tag{7}$$

3. Results and Discussion

3.1. High-Dimensional Data Visualization

Before applying the classifiers, we visualized our preprocessed dataset to gain insights on the grouping of the studied chemical compounds using hierarchical clustering (Fig 4). The clustering is performed by calculating the pairwise-Euclidean distances and using the complete linkage methodology on per compound-averaged data. Fig 4 illustrates the clustering between agonist and antagonist chemicals in all three cases for the array positive cells. The clustering results of all cells is provided in Fig S2. Fig 4a and 4b indicates that when these dendrogram trees are cut at heights that form two clusters, compounds are not exclusively grouped with respect to their agonistic and antagonistic behavior. However, the distinct grouping of the studied compounds according to their ER potential is observed in

the engineered M50 dataset as shown in Fig 4c, except for the AZD9496, ZK164015, ICI 182780 (Fulvestrant) compounds that grouped with the agonists.

We also observe this controversy in the three-dimensional PCA score plot of the M50 dataset, where there is a discontinuous separation between agonist and antagonist compounds (Fig 5a). We observe that the discontinuity in the antagonist clusters arises from the same three compounds (AZD9496, ZK164015, ICI 182780 (Fulvestrant)) that clustered with the agonist compounds in hierarchical clustering. This is an expected observation because these chemicals are known to destabilize ERα expression through the same pathway activated by ERα agonists [44].

As a result, we use two nonlinear classifiers, namely the two-class *c*-parametrized SVM with a Gaussian radial basis kernel and RF algorithms, to achieve the discontinuous classification of agonist and antagonist compounds. These data visualization tools verify that linear classifiers would lack the ability to achieve desired outcomes. The hyperparameter tuning process to model the discontinuous separation of agonist and antagonist compounds, along with the nonlinear classification results will be extensively presented and discussed in the following section.

3.2. Hyperparameter Tuning and Classification Results

To determine the optimum learning parameters of the SVM and RF algorithms, we use 6-fold cross-validated grid search where an extensive number of hyperparameter combinations are exhaustively explored and the one with the highest accuracy is selected for the final model. Fig 6 summarizes the grid search results for the array positive cells of M50 dataset where the best hyperparameter combinations have been specified with red squares.

Subsequently, the final SVM and RF models are trained using all the training compounds and their selected respective hyperparameter combinations. The predictive performances of the final models are then quantified using the testing compounds (Table 2). Table 6 reports the performance metrics – accuracy, sensitivity, specificity, F_1 score, and balanced accuracy – for both classifiers.

The nonlinear classification results of array positive M50 scenario show that 2-class nonlinear SVM performs slightly better than the RF classifier with a correct classification accuracy of 96.36% on the unseen data, whereas the accuracy of the RF model is 94.36%. However, in the training phase, we observe that the RF model learns the patterns in the data perfectly with 100% accuracy, sensitivity, specificity, F_1 score, and balanced accuracy. Yet, we know that the training performances are not representative of the predictive performance of the machine learning models, which is evident in the testing performance of the RF model. We also observe that the sensitivity of the SVM model is higher for the testing compounds which implies that this model classifies more agonist compounds correctly as agonists. This is also presented in the confusion matrices shown in Fig 7. Furthermore, the testing specificity of the final SVM model is 100% which indicates that all antagonists in the testing set, including ZK164015, ICI 182780 (Fulvestrant), are correctly classified. Hence, the SVM model can learn the discontinuous regions of antagonist compounds in the dataset and performs better on predicting the ER potential of unseen compounds. The RF model

also has favorable testing performance for the M50 dataset where all testing performance metrics scored above 90%.

A deeper insight to the model performances can also be obtained from analyzing the per cell classification of the SVM and RF models (Tables S26-S29). The results show that SVM misclassifies 0.64% of cells (1 out of 156 cells) treated with 17β-Estradiol, 50.77% of cells (33 out of 65 cells) treated with Dicofol, 1.66% of cells (4 out of 241 cells) treated with 4-Cumylphenol, and 23.9% of cells (65 out of 272 cells) treated with AZD9496 in the training phase. However, in the testing phase we observe that SVM can perfectly classify the antagonist compounds with 100% accuracy whereas several cells treated with very weak, weak, and strong agonist compounds are misclassified. This can be attributed to the similarity of chemicals based on their Euclidean distances in space. If we take a closer look at hierarchical clustering, the two correctly classified antagonists – ZK164015, ICI 182780 (Fulvestrant) – are grouped together, while AZD9496 is clustered among the agonist chemicals (Fig 4c). As a result, it is more challenging to classify all the cells treated with AZD9496 with 100% accuracy. Similar observation is also made for Dicofol and Kaempferol where SVM poorly predicts their estrogenic potential with 49.23% and 65.15% accuracy, respectively. These two compounds are grouped together at the final branch of the dendrogram as shown in Fig 4c.

With regards to the RF, this nonlinear classifier mapped the discontinuous separation defined by the antagonist chemicals in the agonist region with over 90% accuracy. We observe that 5.69% of cells (12 out of 211 cells) treated with ZK164015 are misclassified in the testing set whereas the misclassification rate for ICI 182 780 (Fulvestrant) treated cells is higher with 7.88% of cells (29 out of 368 cells) being misclassified as agonists. This decision tree-based classifier also had difficulties with correctly classifying the ER activity of Kaempferol like the SVM, where only 77.27% of the cells are classified as agonists and other agonists that lowered the overall predictive performance of the RF model.

3.3. The Effect of Data Preparation to the Model Performance

Although in the previous section, we extensively analyzed the results of the M50 array positive dataset, the same modeling procedure is also applied to the other aforementioned datasets (Table 1) to compare the effects of different preprocessing and feature engineering strategies on the classification of agonist and antagonist compounds. In Fig 8, the testing accuracies of nonlinear classifiers are reported for 6 differently constructed datasets.

According to the classification results, the datasets with engineered features have higher accuracies than the original experimental data. Also, it is noteworthy to state that the 2-class nonlinear SVM model outperforms the RF model in 5 cases, except for the LMH10 data for array positive cells, where the RF classifier maps the separation with 84.21% compared to the SVM accuracy of 77.95%. In addition, our classification framework can achieve slightly higher accuracy compared to M50 data set with array positive cells when LHM10 dataset with all cells used in the SVM model. Nonetheless, the results indicate that feature engineering can reveal patterns that can be hidden in the original experimental data and allow for more than 96% classification accuracy of agonist and antagonist compounds with the SVM model.

4. Conclusions

In this study, we introduced a framework for modeling the endocrine disrupting potential of benchmark chemicals using single-cell level information and machine learning algorithms. After getting multidimensional high throughput microscopy and high content analysis-based data, we performed series of preprocessing steps to reveal the underlying patterns in datasets and constructed accurate classification models. Later, we used Principal Component Analysis (PCA) to reduce the dimensionality of our datasets and selected optimal number of PCs as the descriptive features for the nonlinear classifiers. We used two-class nonlinear Support Vector Machine model and Random Forest algorithm to map the separation between the endocrine disrupting potential of chemical compounds. Our results showed that both classification models were capable of predicting the endocrine disruption potential of compounds with high accuracy, where the SVM model achieved more than 96% and the RF model achieved more than 94% classification accuracy on unseen compounds. Additionally, our analysis indicated that feature engineering improved the classification accuracy by 20 points. We believe that our framework will be beneficial for facilitating the decision-making process during and after environmental emergencies for determining the possible biological hazards of unknown chemicals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health [NIH P42-ES027704] and University of Connecticut. Portions of this research were conducted with the advanced computing resources provided by the University of Connecticut Storrs High Performance Computing facility. Imaging for this project was supported by the Integrated Microscopy Core at Baylor College of Medicine and the Center for Advanced Microscopy and Image Informatics (CAMII) with funding from NIH (DK56338, CA125123, ES030285), and CPRIT (RP150578, RP170719). The manuscript contents are solely the responsibility of the grantee and do not necessarily represent the official views of the NIH. Further, NIH does not endorse the purchase of any commercial products or services mentioned in the publication.

References

1. Colborn T, 1995. Environmental estrogens: health implications for humans and wildlife. Environmental Health Perspectives, 103(suppl 7), pp. 135–136.
2. Schantz SL and Widholm JJ, 2001. Cognitive effects of endocrine-disrupting chemicals in animals. Environmental health perspectives, 109(12), pp.1197–1206. [PubMed: 11748026]
3. Endocrine Society."Endocrine-Disrupting Chemicals (EDCs) | Endocrine Society." Endocrine.org, Endocrine Society, 13 July 2022, https://www.endocrine.org/patient-engagement/endocrine-library/edcs
4. Encarnação T, Pais AA, Campos MG and Burrows HD, 2019. Endocrine disrupting chemicals: Impact on human health, wildlife and the environment. Science progress, 102(1), pp.3–42. [PubMed: 31829784]
5. Kassotis Christopher D., Trasande Leonardo, Chapter One - Endocrine disruptor global policy, Editor(s): Vandenberg Laura N., Turgeon Judith L., Advances in Pharmacology, Academic Press, Volume 92, 2021, Pages 1–34 10.1016/bs.apha.2021.03.005. [PubMed: 34452684]
6. Lee HR, Kim TH and Choi KC, 2012. Functions and physiological roles of two types of estrogen receptors, ERα and ERβ, identified by estrogen receptor knockout mouse. Laboratory animal research, 28(2), pp.71–76. [PubMed: 22787479]

7. Welboren WJ, Stunnenberg HG, Sweep FC and Span PN, 2007. Identifying estrogen receptor target genes. Molecular oncology, 1(2), pp. 138–143. [PubMed: 19383291]
8. Brzozowski AM, Pike AC, Dauter Z, Hubbard RE, Bonn T, Engström O, Öhman L, Greene GL, Gustafsson JÅ and Carlquist M, 1997. Molecular basis of agonism and antagonism in the oestrogen receptor. Nature, 389(6652), pp.753–758. [PubMed: 9338790]
9. Meigs L, Smirnova L, Rovida C, Leist M and Hartung T, 2018. Animal testing and its alternatives-The most important omics is economics. ALTEX-Alternatives to animal experimentation, 35(3), pp.275–305.
10. Beykal B, Boukouvala F, Floudas CA and Pistikopoulos EN, 2018. Optimal design of energy systems using constrained grey-box multi-objective optimization. Computers & chemical engineering, 116, pp.488–502. [PubMed: 30546167]
11. Beykal B, Boukouvala F, Floudas CA, Sorek N, Zalavadia H and Gildin E, 2018. Global optimization of grey-box computational systems using surrogate functions and application to highly constrained oil-field operations. Computers & Chemical Engineering, 114, pp.99–110.
12. Beykal B, Onel M, Onel O and Pistikopoulos EN, 2020. A data-driven optimization algorithm for differential algebraic equations with numerical infeasibilities. AIChE Journal, 66(10), p.e16657. [PubMed: 32921798]
13. Bi K, Beykal B, Avraamidou S, Pappas I, Pistikopoulos EN and Qiu T, 2020. Integrated modeling of transfer learning and intelligent heuristic optimization for a steam cracking process. Industrial & engineering chemistry research, 59(37), pp. 16357–16367. [PubMed: 33041499]
14. Avraamidou S, Beykal B, Pistikopoulos IP and Pistikopoulos EN, 2018. A hierarchical food-energy-water nexus (FEW-N) decision-making approach for land use optimization. In Computer Aided Chemical Engineering (Vol. 44, pp. 1885–1890). Elsevier.
15. Beykal B, Avraamidou S, Pistikopoulos IP, Onel M and Pistikopoulos EN, 2020. Domino: Data-driven optimization of bi-level mixed-integer nonlinear problems. Journal of Global Optimization, 78(1), pp.1–36. [PubMed: 32753792]
16. Beykal B, Avraamidou S and Pistikopoulos EN, 2022. Data-driven optimization of mixed-integer bi-level multi-follower integrated planning and scheduling problems under demand uncertainty. Computers & Chemical Engineering, 156, p. 107551. [PubMed: 34720250]
17. Beykal B, Avraamidou S and Pistikopoulos EN, 2021. Bi-level Mixed-Integer Data-Driven Optimization of Integrated Planning and Scheduling Problems. In Computer Aided Chemical Engineering (Vol. 50, pp. 1707–1713). Elsevier.
18. Beykal B, Aghayev Z, Onel O, Onel M and Pistikopoulos EN, 2022. Data-driven Stochastic Optimization of Numerically Infeasible Differential Algebraic Equations: An Application to the Steam Cracking Process. In Computer Aided Chemical Engineering (Vol. 49, pp. 1579–1584). Elsevier.
19. Qin SJ and Chiang LH, 2019. Advances and opportunities in machine learning for process data analytics. Computers & Chemical Engineering, 126, pp.465–473.
20. Venkatasubramanian V, 2019. The promise of artificial intelligence in chemical engineering: Is it here, finally?. AIChE Journal, 65(2), pp.466–478.
21. Thebelt A, Wiebe J, Kronqvist J, Tsay C and Misener R, 2022. Maximizing information from chemical engineering data sets: applications to machine learning. Chemical Engineering Science, 252, p. 117469.
22. Pan I, Mason LR and Matar OK, 2022. Data-centric Engineering: integrating simulation, machine learning and statistics. Challenges and opportunities. Chemical Engineering Science, 249, p. 117271.
23. Schweidtmann AM, Esche E, Fischer A, Kloft M, Repke JU, Sager S and Mitsos A, 2021. Machine learning in chemical engineering: a perspective. Chemie Ingenieur Technik, 93(12), pp.2029–2039.
24. Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens CV and Van Geem KM, 2021. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. Engineering, 7(9), pp.1201–1211.

25. Lee JH, Shin J and Realff MJ, 2018. Machine learning: Overview of the recent progresses and implications for the process systems engineering field. Computers & Chemical Engineering, 114, pp. 111–121.
26. Onel M, Beykal B, Wang M, Grimm FA, Zhou L, Wright FA, Phillips TD, Rusyn I and Pistikopoulos EN, 2018. Optimal chemical grouping and sorbent material design by data analysis, modeling and dimensionality reduction techniques. In Computer Aided Chemical Engineering (Vol. 43, pp. 421–426). Elsevier.
27. Onel M, Beykal B, Ferguson K, Chiu WA, McDonald TJ, Zhou L, House JS, Wright FA, Sheen DA, Rusyn I and Pistikopoulos EN, 2019. Grouping of complex substances using analytical chemistry data: A framework for quantitative evaluation and visualization. PloS one, 14(10), p.e0223517. [PubMed: 31600275]
28. Orr AA, Wang M, Beykal B, Ganesh HS, Hearon SE, Pistikopoulos EN, Phillips TD and Tamamis P, 2021. Combining experimental isotherms, minimalistic simulations, and a model to understand and predict chemical adsorption onto montmorillonite clays. ACS omega, 6(22), pp.14090–14103. [PubMed: 34124432]
29. Chen Z, Liu Y, Wright FA, Chiu WA and Rusyn I, 2020. Rapid hazard characterization of environmental chemicals using a compendium of human cell lines from different organs. ALTEX-Alternatives to animal experimentation, 37(4), pp.623–638.
30. Roman-Hubers AT, Cordova AC, Rohde AM, Chiu WA, McDonald TJ, Wright FA, Dodds JN, Baker ES and Rusyn I, 2022. Characterization of compositional variability in petroleum substances. Fuel, 317, p.123547. [PubMed: 35250041]
31. House JS, Grimm FA, Klaren WD, Dalzell A, Kuchi S, Zhang SD, Lenz K , Boogaard PJ, Ketelslegers HB, Gant TW and Rusyn I, 2022. Grouping of UVCB substances with dose-response transcriptomics data from human cell-based assays. ALTEX-Alternatives to animal experimentation.
32. Kleinstreuer NC, Ceger P, Watt ED, Martin M, Houck K, Browne P, Thomas RS, Casey WM, Dix DJ, Allen D and Sakamuru S, 2017. Development and validation of a computational model for androgen receptor activity. Chemical research in toxicology, 30(4), pp.946–964. [PubMed: 27933809]
33. Li J and Gramatica P, 2010. Classification and virtual screening of androgen receptor antagonists. Journal of chemical information and modeling, 50(5), pp.861–874. [PubMed: 20405856]
34. Zang Q, Rotroff DM and Judson RS, 2013. Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure–activity relationship and machine learning methods. Journal of chemical information and modeling, 53(12), pp.3244–3261. [PubMed: 24279462]
35. Chen Y, Cheng F, Sun L, Li W, Liu G and Tang Y, 2014. Computational models to predict endocrine-disrupting chemical binding with androgen or oestrogen receptors. Ecotoxicology and environmental safety, 110, pp.280–287. [PubMed: 25282305]
36. Mayr A, Klambauer G, Unterthiner T and Hochreiter S, 2016. DeepTox: toxicity prediction using deep learning. Frontiers in Environmental Science, 3, p.80.
37. Chierici M, Giulini M, Bussola N, Jurman G and Furlanello C, 2018. Machine learning models for predicting endocrine disruption potential of environmental chemicals. Journal of Environmental Science and Health, Part C, 36(4), pp.237–251.
38. Ciallella HL, Russo DP, Aleksunes LM, Grimm FA and Zhu H, 2021. Predictive modeling of estrogen receptor agonism, antagonism, and binding activities using machine-and deep-learning approaches. Laboratory investigation, 101(4), pp.490–502. [PubMed: 32778734]
39. Idakwo G, Luttrell J, Chen M, Hong H, Zhou Z, Gong P and Zhang C, 2018. A review on machine learning methods for in silico toxicity prediction. Journal of Environmental Science and Health, Part C, 36(4), pp.169–191.
40. Wang MW, Goodman JM and Allen TE, 2020. Machine learning in predictive toxicology: recent applications and future directions for classification models. Chemical Research in Toxicology, 34(2), pp.217–239. [PubMed: 33356168]

41. Russo DP, Zorn KM, Clark AM, Zhu H and Ekins S, 2018. Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. Molecular pharmaceutics, 15(10), pp.4361–4370. [PubMed: 30114914]

42. Ashcroft FJ, Newberg JY, Jones ED, Mikic I and Mancini MA, 2011. High content imaging-based assay to classify estrogen receptor-α ligands based on defined mechanistic outcomes. Gene, 477(1-2), pp.42–52. [PubMed: 21256200]

43. Sharp ZD, Mancini MG, Hinojos CA, Dai F, Berno V, Szafran AT, Smith KP, Lele TT, Ingber DE and Mancini MA, 2006. Estrogen-receptor-α exchange and chromatin dynamics are ligand-and domain-dependent. Journal of cell science, 119(19), pp.4101–4116. [PubMed: 16968748]

44. Mukherjee R, Beykal B, Szafran AT, Onel M, Stossi F, Mancini MG, Lloyd D, Wright FA, Zhou L, Mancini MA and Pistikopoulos EN, 2020. Classification of estrogenic compounds by coupling high content analysis and machine learning algorithms. PLoS computational biology, 16(9), p.e1008191. [PubMed: 32970665]

45. Mukherjee R, Onel M, Beykal B, Szafran AT, Stossi F, Mancini MA, Zhou L, Wright FA and Pistikopoulos EN, 2019. Development of the texas A&M superfund research program computational platform for data integration, visualization, and analysis. In Computer Aided Chemical Engineering (Vol. 46, pp. 967–972). Elsevier.

46. Ganesh HS, Beykal B, Szafran AT, Stossi F, Zhou L, Mancini MA and Pistikopoulos EN, 2021. Predicting the Estrogen Receptor Activity of Environmental Chemicals by Single-Cell Image Analysis and Data-driven Modeling. In Computer Aided Chemical Engineering (Vol. 50, pp. 481–486). Elsevier.

47. Szafran AT and Mancini MA, 2014. The myImageAnalysis project: a web-based application for high-content screening. Assay and drug development technologies, 12(1), pp.87–99. [PubMed: 24547743]

48. Enders CK, 2022. Applied missing data analysis. Guilford Publications.

49. James G, Witten D, Hastie T and Tibshirani R, 2013. An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

50. Abdi H and Williams LJ, 2010. Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), pp.433–459.

51. Cortes C and Vapnik V, 1995. Support-vector networks. Machine learning, 20(3), pp.273–297.

52. Schölkopf B, Smola AJ and Bach F, 2002. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

53. Breiman L, 2001. Random forests. Machine learning, 45(1), pp.5–32.

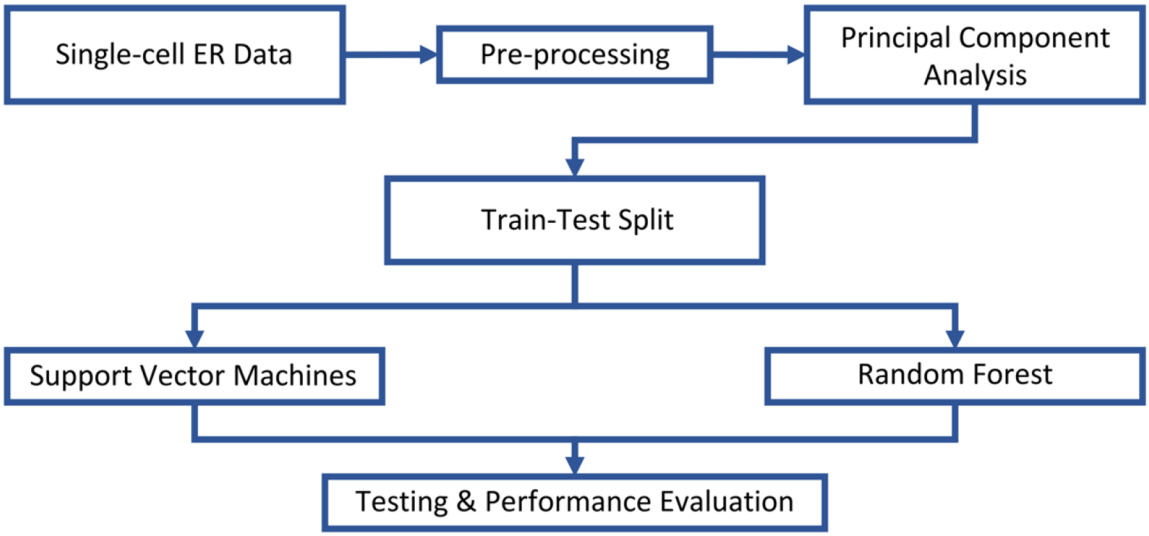

Fig 1.
The estrogenic activity classification flowchart with single-cell level high throughput imaging data.

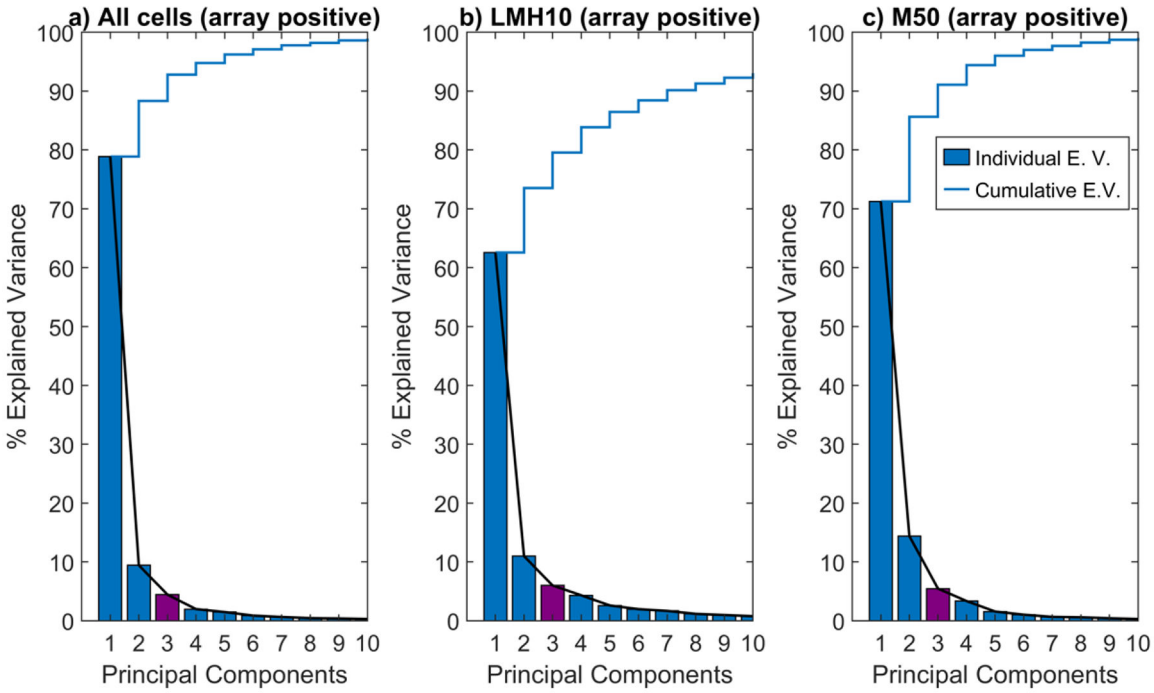

Fig 2.

Scree plot of explained variance and principal components for datasets with visible nuclear spots. The purple bar indicates the “elbow” of the scree plot which is the optimal number of PCs to retain in the analysis.

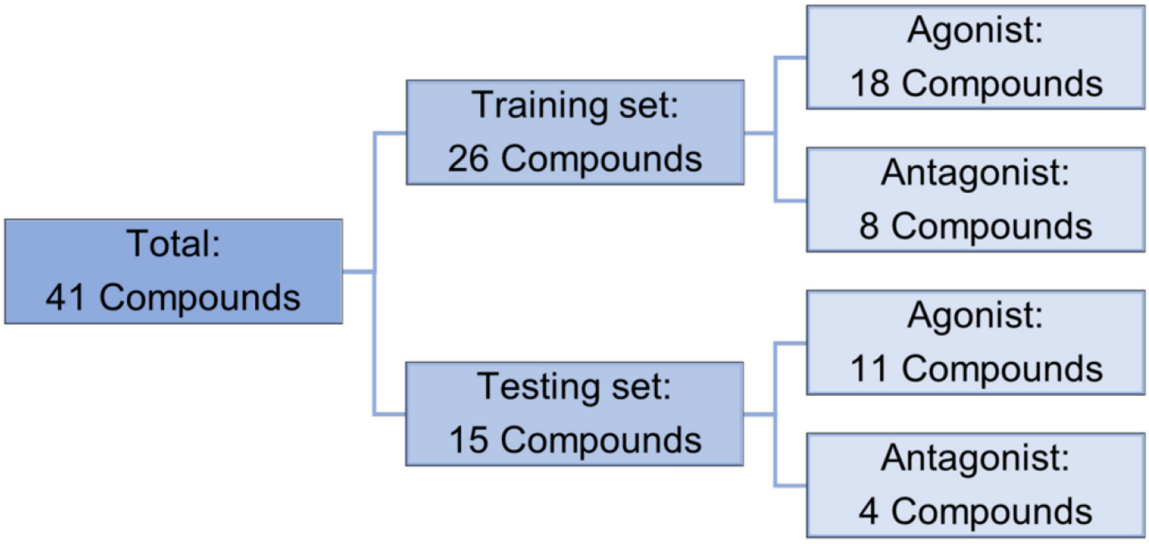

Fig 3.
The Train-test split of the active benchmark chemicals.

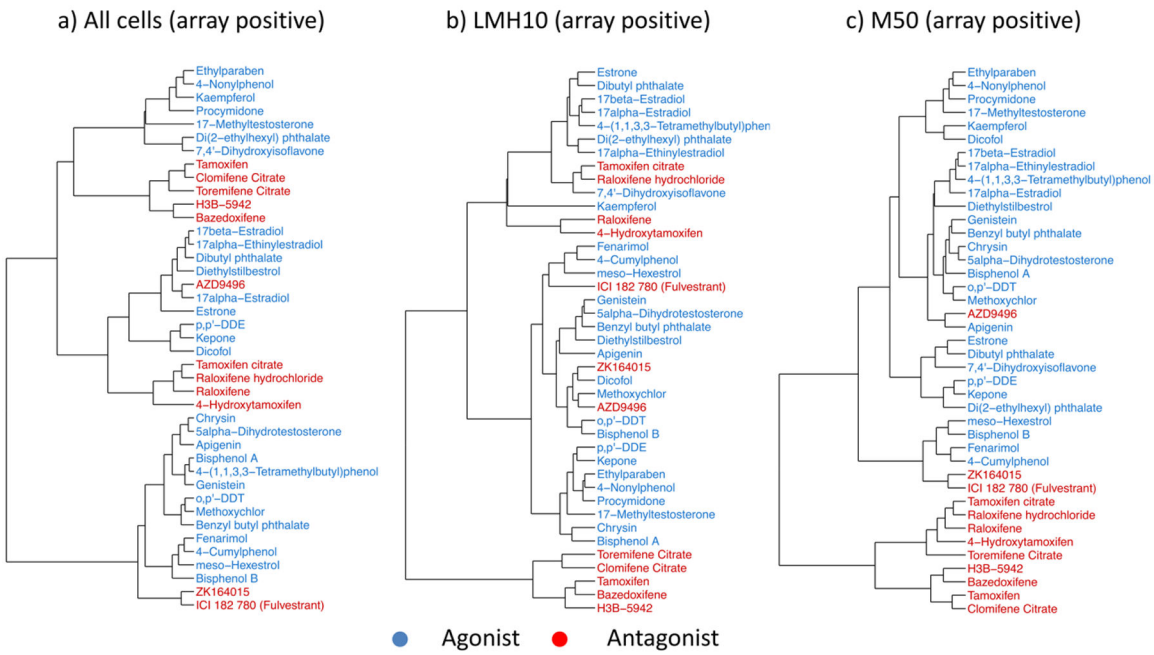

Fig 4.

Euclidean distance complete linkage hierarchical clustering for datasets with visible nuclear spots.

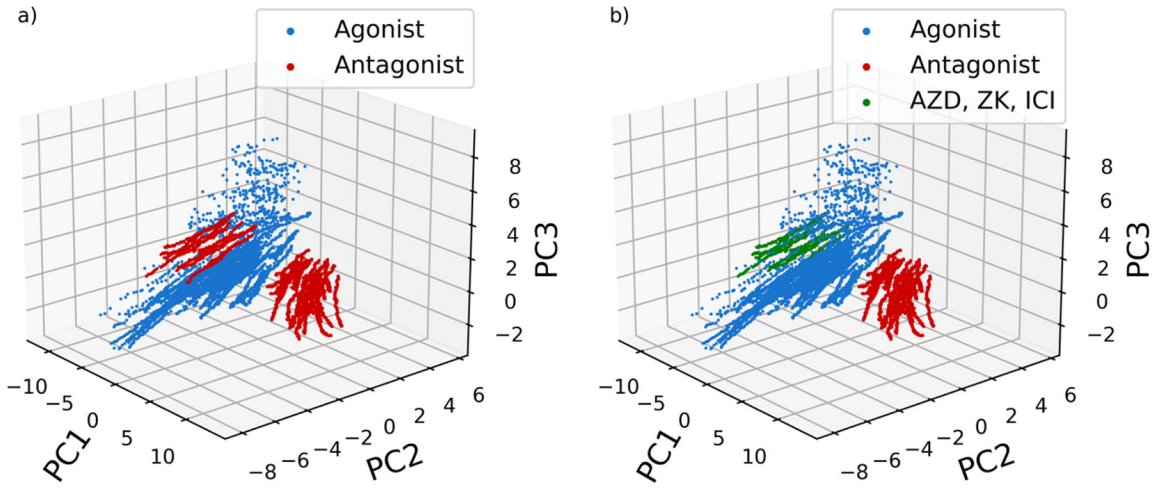

Fig 5.

3D PCA plot for the array positive M50 scenario. Green points specify the overlap in the agonist region by AZD, ZK, ICI compounds.

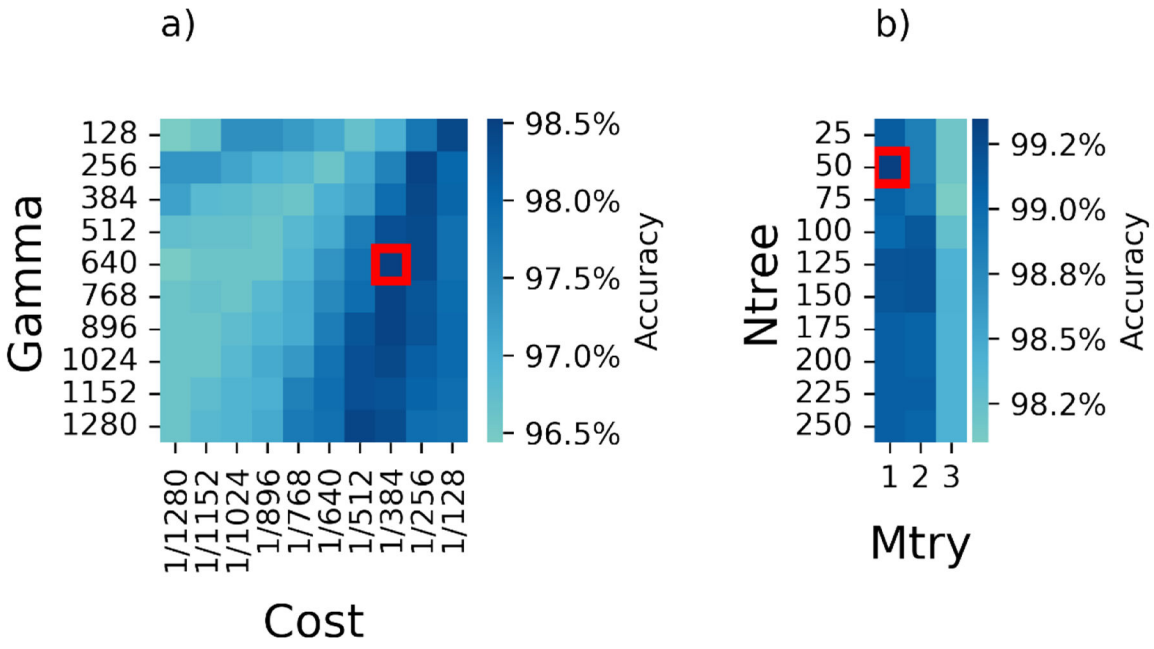

Fig 6.

Heatmaps of grid search results for the nonlinear classifiers of array positive M50 scenario: a) SVM, b) RF. The hyperparameter combination with the highest accuracy (indicated with red) is selected to train the final model.

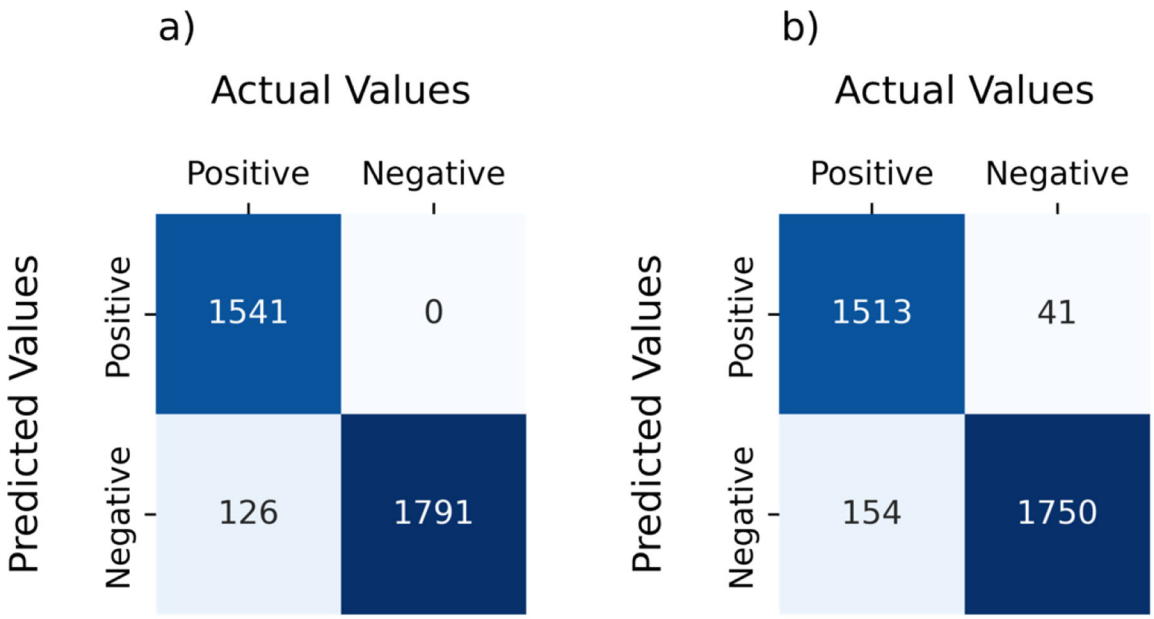

Fig 7.

Confusion matrices for nonlinear classifiers for the testing set of array positive M50 scenario: a) SVM, b) RF.

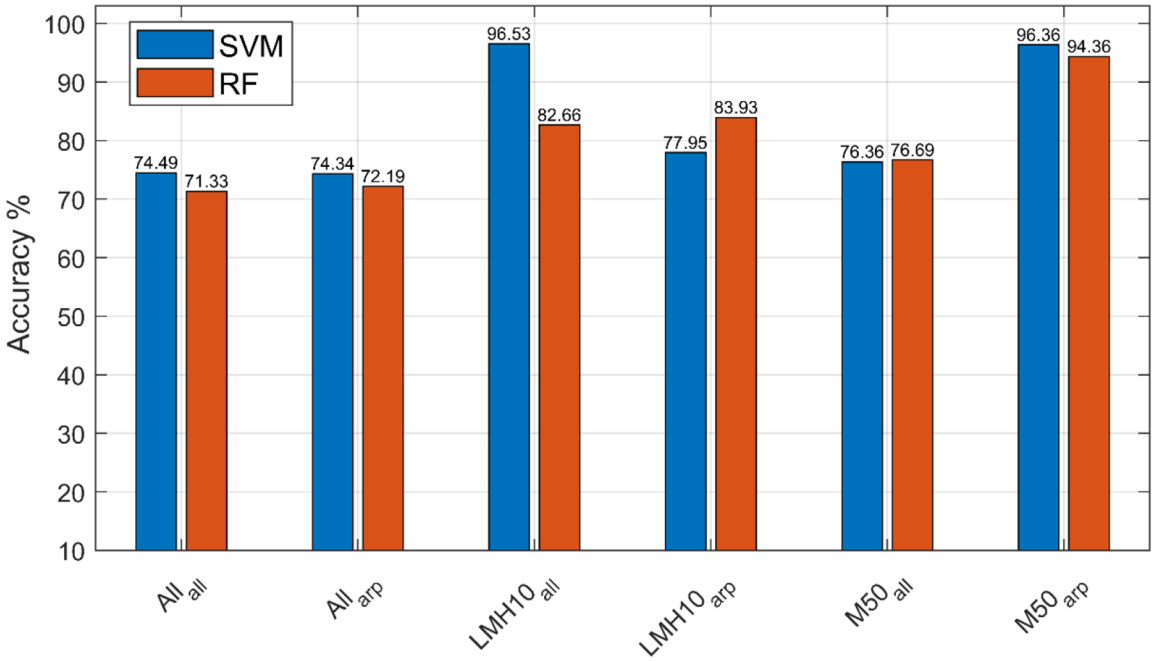

Fig 8.

Testing accuracies with 15 unseen chemicals over 6 different data preparation techniques for the 2-class SVM and RF classifiers.

Table 1.

Dimensionality of the original and preprocessed data for all cases.

Dataset	Original data size	Preprocessed data size
All cells	162525 × 79	48943 × 37
All cells (array positive)	58491 × 79	20748 × 37
L10+M10+H10 cells	15157 × 218	4451 × 108
L10+M10+H10 cells (array positive)	4990 × 218	1914 × 108
M50 cells	73073 × 78	22116 × 36
M50 cells (array positive)	22286 × 78	9420 × 36

Table 2.

The active chemical compounds used in training and testing of the classification algorithms.

	Compound Name	ER Activity	Potency
Training Set	17-Methyltestosterone	Agonist	Very Weak
	4-(1,1,3,3-Tetramethylbutyl)phenol	Agonist	Very Weak
	4-Nonylphenol	Agonist	Very Weak
	Apigenin	Agonist	Very Weak
	Dibutyl phthalate	Agonist	Very Weak
	Dicofol	Agonist	Very Weak
	Fenarimol	Agonist	Very Weak
	Methoxychlor	Agonist	Very Weak
	4-Cumylphenol	Agonist	Weak
	5alpha-Dihydrotestosterone	Agonist	Weak
	Bisphenol A	Agonist	Weak
	Bisphenol B	Agonist	Weak
	Genistein	Agonist	Weak
	Kepone	Agonist	Weak
	o,p'-DDT	Agonist	Weak
	Estrone	Agonist	Moderate
	17β-Estradiol	Agonist	Strong
	Diethylstilbestrol	Agonist	Strong
	4-Hydroxytamoxifen	Antagonist	-
	AZD9496	Antagonist	-
	H3B-5942	Antagonist	-
	Raloxifene	Antagonist	-
	Raloxifene hydrochloride	Antagonist	-
	Tamoxifen	Antagonist	-
	Tamoxifen citrate	Antagonist	-
	Toremifene Citrate	Antagonist	-
Testing Set	Benzyl butyl phthalate	Agonist	Very Weak
	Di(2-ethylhexyl) phthalate	Agonist	Very Weak
	p,p'-DDE	Agonist	Very Weak
	Chrysin	Agonist	Very Weak
	Ethylparaben	Agonist	Very Weak
	Procymidone	Agonist	Very Weak
	Kaempferol	Agonist	Very Weak
	7,4'-Dihydroxyisoflavone	Agonist	Weak
	meso-Hexestrol	Agonist	Strong
	17α-Ethinylestradiol	Agonist	Strong
	17α-Estradiol	Agonist	Strong

	Compound Name	ER Activity	Potency
	ICI 182 780 (Fulvestrant)	Antagonist	-
	ZK164015	Antagonist	-
	Bazedoxifene	Antagonist	-
	Clomifene Citrate	Antagonist	-

Table 3.

The chemical compounds that were removed after the preprocessing step.

	Compound Name	ER Activity	Potency
Inactive Compounds	Media	Inactive	-
	Progesterone	Inactive	-
	Reserpine	Inactive	-
	Cycloheximide	Inactive	-
	Haloperidol	Inactive	-
	Corticosterone	Inactive	-
	Linuron	Inactive	-
	Flutamide	Inactive	-
	Ketoconazole	Inactive	-
	Hydroxyflutamide	Inactive	-
	Spironolactone	Inactive	-
	Atrazine	Inactive	-
	Phenobarbital sodium	Inactive	-
Control Chemicals	E2	Agonist	Strong
	E2-1	Agonist	Strong
	E2-2	Agonist	Strong
	4HT	Antagonist	-
	4HT-1	Antagonist	-
	4HT-2	Antagonist	-

Table 4.

Proportion of variance of first three PCs and the cumulative variance for all 6 scenarios.

Dataset	PC1	PC2	PC3	Cumulative variance
All cells	0.4603	0.1874	0.1114	0.7590
All cells (array positive)	0.7887	0.0945	0.0446	0.9277
L10+M10+H10 cells	0.5315	0.1397	0.0956	0.7668
L10+M10+H10 cells (array positive)	0.6257	0.1098	0.0600	0.7954
M50 cells	0.6499	0.1516	0.1068	0.9083
M50 cells (array positive)	0.7123	0.1438	0.0546	0.9107

Table 5.

Data size of training and testing sets for all cases.

Dataset	Training set	Testing set
All cells	31841 × 3	17102 × 3
All cells (array positive)	13272 × 3	7476 × 3
L10+M10+H10 cells	2897 × 3	1557 × 3
L10+M10+H10 cells (array positive)	1211 × 3	703 × 3
M50 cells	14387 × 3	7729 × 3
M50 cells (array positive)	5962 × 3	3458 × 3

Table 6.

Performance metrics for 2-class SVM and RF metrics of array positive M50 scenario.

Classifier	Set	Accuracy	Sensitivity	Specificity	F_1 Score	Balanced Accuracy
Support Vector Machines	Testing	96.36%	92.44%	100.00%	96.07%	96.22%
	Training	98.27%	98.66%	97.92%	98.19%	98.29%
Random Forest	Testing	94.36%	90.76%	97.71%	93.95%	94.24%
	Training	100.00%	100.00%	100.00%	100.00%	100.00%