 doi:10.20944/preprints202102.0318.v2

Article

Androgen Receptor Binding Category Prediction with Deep Neural Networks and Structure-, Ligand- and Statistically-Based Features

Alfonso T. García-Sosa [1,*]

1. Institute of Chemistry, University of Tartu, Ravila 14a, Tartu 54011, Estonia
* Correspondence: alfonsog@ut.ee ; ORCiD 0000-0003-0542-4446

Abstract: Substances that can modify the androgen receptor pathway in humans and animals are entering the environment and food chain with the proven ability to disrupt hormonal systems and leading to toxicity and adverse effects on reproduction, brain development, and prostate cancer, among others. State-of-the-art databases with experimental data of human, chimp, and rat effects by chemicals have been used to build machine learning classifiers and regressors and evaluate these on independent sets. Different featurizations, algorithms, and protein structures lead to different results, with deep neural networks (DNNs) on user-defined physicochemically-relevant features developed for this work outperforming graph convolutional, random forest, and large featurizations. The results show that these user-provided structure-, ligand-, and statistically-based features and specific DNNs provided the best results as determined by AUC (0.87), MCC (0.47), and other metrics and by their interpretability and chemical meaning of the descriptors/features. In addition, the same features in the DNN method performed better than in a multivariate logistic model: validation MCC = 0.468 and training MCC = 0.868 for the present work compared to evaluation set MCC = 0.2036 and training set MCC = 0.5364 for the multivariate logistic regression on the full, unbalanced set. Techniques of this type may improve AR and toxicity description and prediction, improving assessment and design of compounds. Source code and data are available at https://github.com/AlfonsoTGarcia-Sosa/ML

Keywords: Machine learning; Artificial intelligence; androgen receptor; random forest; deep neural network

1. Introduction

Concerns are rising over endocrine disruptors entering the environment and food chain [1-4]. The Androgen Receptor (AR) is a protein involved in reproduction, brain development, prostate cancer, androgen insensitivity syndromes, spinal and bulbar muscular atrophy, acne, and alopecia [5]. Androgen receptor pathway modulators are compounds that can have an effect on tumors and reproductive systems [1-4]. The CoMPARA challenge was a collaborative modeling effort to predict possible AR modulators based on a wide collection of state-of-the-art experimental data [6]. Different modeling techniques have been attempted, including molecular docking [7], support-vector machines (SVMs), combined structure-based, ligand-based ECFP distances to known compounds, and Naïve Bayesians [8], among others [6].

Toxicity modeling of compounds is important in several ways: compounds that are used in pharmaceutical and industrial applications need to be assessed for possible adverse effects on humans and other organisms, as well as being an important development barrier for new drugs and useful compounds [1-4]. Difficulties include the lack of experimental tests including chronic and different exposure effects, as well as those of metabolites of compounds [1-4].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2

ML and AI are transforming many fields, including the computational chemistry and medicinal chemistry fields [9, 10, 11, 12, 13, 14]. Particular advantages may be realized by ML methods in classification techniques for drug compound analysis and design [10, 15]. A common criticism of ML methods is the potential to include a high number of variables that can have little insight into their physicochemical meaning [10]. Some ML and AI models are being developed with the aim of being 'explainable' [16] and afford interpretability to chemical groups responsible for positive or negative contributions to a prediction.

The present work shows that different modeling techniques can have their advantages and disadvantages for modeling AR modulating compounds. Deep neural networks (DNNs) and graph convolutional neural networks (CNNs) have been used in other modeling studies, usually using featurization included in widely-available packages [17]. Here, an effort was made to build Random Forest (RF) and DNNs with a given set of features that are chemically important based on calculated protein-ligand binding to several targets, chemical fingerprint distances, and other results from statistical techniques. Aiming to improve the predicted categorization of chemical compounds as AR binders using physicochemically- and biologically-relevant features can help in flagging molecules that may have potential to disrupt AR pathways, and thus, may have the potential of toxic effects. In silico prediction of these effects is important given the reduction of animal testing, and the expense of testing, as well as a first, fast complement to testing.

2. Results

2.1 Balancing and initial separations

The provided training dataset in the CoMPARA challenge was highly unbalanced with a large number of non-binders (N = 205) compared to binders (N = 1,468). Bias in datasets can affect strongly the results of ML algorithms and so addressing these issues is recommended [18]. The training set was thus balanced to provide an equal amount of binders and non-binders. Another effect was also apparent after docking calculations for the chimp protein (Figure 1), where distributions of docking scores for chimp androgen receptor for binders and non-binders show that for those compounds that have a nonzero docking score (N = 1,310), the docking scores are stronger for binders than for non-binders. However, for the smaller amount of those compounds with a docking score of zero (N = 363), the simple docking score on its own cannot separate both distributions.

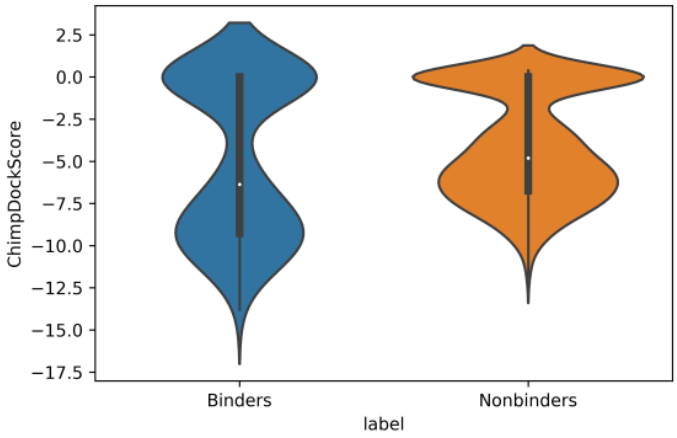
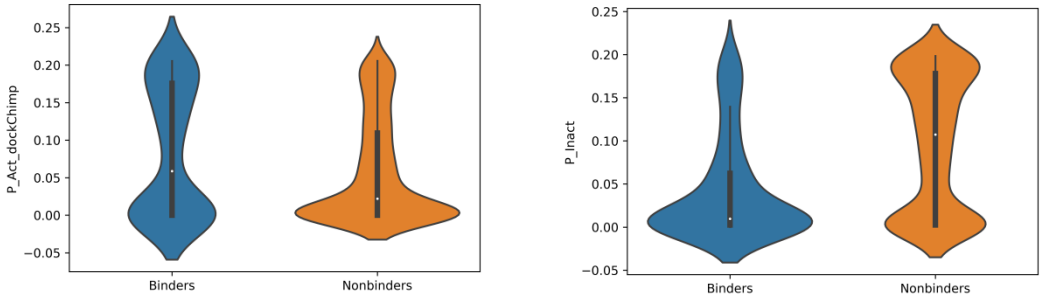


Figure 1. Violin plots of the distributions of docking scores (kcal/mol) for chimp androgen receptor docking scores for binders (blue) and non-binders (orange).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2

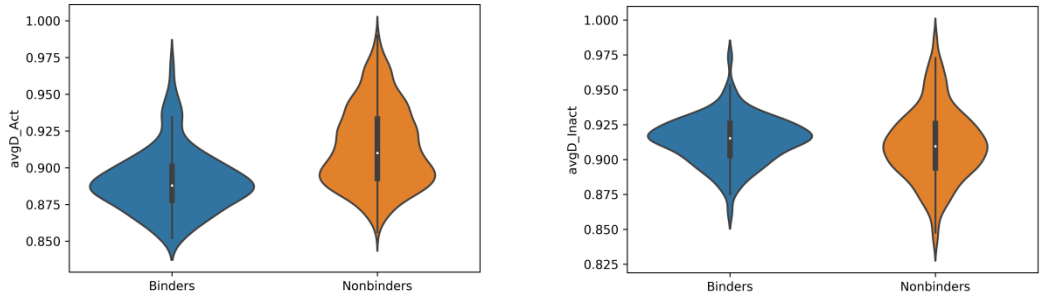
A similar effect is seen for the calculated Bayesian probabilities (Figure 2), where the distribution for binders is higher in probability to belong to the normal distribution of known binders and vice versa for non-binders. Bayesians constructed on the docking scores for both groups showed this separation with means and standard deviations of $\mu = -8.91$ kcal/mol and $\sigma = 1.94$ for binders, and $\mu = -5.97$ kcal/mol and $\sigma = 2.01$ for non-binders.



(a) **(b)**

Figure 2. Violin plots for the distributions of Bayesian probabilities for docking scores for: **a**) binders (P_act_dockChimp) and **b**) nonbinders (P_Inact) for chimp androgen receptor docking scores for binders (blue) and non-binders (orange).

The distribution of Tanimoto distances to average values for ECFP fingerprints of binders and nonbinders (Figure 3) show skews in the distributions with nonbinders tending to have a tail skewed towards larger distances to the average of known binders, whilst the binding compounds have slightly larger distances to the average of known nonbinding compounds.



(a) **(b)**

Figure 3. Violin plots of the distributions of average Tanimoto distances to: **a**) known binders (avgD_Act), and to: **b**) known nonbinders (avgD_Inact) for binders (blue) and non-binders (orange).

After balancing, the total numbers for both sets of compounds were: Number of compounds in train set: 410, composed of 205 binders and 205 nonbinders. No balancing was made for the evaluation set, where the number of compounds in validation set was 3,882.

2.2 Comparison to RF and CNN

Classifier and regression machine learning models were built and evaluated (Table 1). For the Random Forest (RF) Classifier (I), the best results obtained after eight-fold cross-validation were: Best hyperparameters: (100, 'sqrt'), giving AUC values of train_score: 1.000000, validation_score: 0.7564. The Graph Convolutional Neural Network Classifier (CNN, VII) using hyperparameters: {'learning_rate': 0.00014031305077588868,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2

'weight_decay_penalty': 2.954578768759171e-06, 'nb_epoch': 40}, gave AUC values of: train_score: 0.827864, validation_score: 0.739908. There is clearly overfitting occurring in RF and CNN methods given the high ROC-AUC values for the training set (Figure 4), with the best metrics for the evaluation score being around 19 runs.

2.3 DNN with different featurizations

The best results for the training and validation sets based on the ROC-AUC metric was the DNN using the supplied 12 features (Table 1, model II), as well as showing less overfitting.

Table 1. Best methods obtained for different algorithms and featurizations. Standard deviations included for best methods.

Method	Train ± s.d.	Valid ± s.d.	Best hyperparameters
RF classifier myfeats (**I**)	AUC 0.9999 ± 0.0009; MCC 0.9951 ± 0.0153; F1 0.9963 ± 0.0153; Prec. 0.9976 ± 0.011; Recall 0.9951 ± 0.0198	AUC 0.7564 ± 0.0105; MCC 0.297435 ± 0.0478; F1 0.5805 ± 0.1041 ($3x10^6$ epochs); Prec. 0.8856 ± 0.0148 ($1.5x10^5$ epochs); Recall 0.4481 ± 0.0866 ($3x10^6$ epochs)	eight-fold cross-validation, (19 runs, 'sqrt'), $2.25x10^6$ epochs
DNN classifier myfeats (**II**)	AUC 0.9424 ± 0.0655; MCC 0.7472 ± 0.1283; F1 0.8608 ± 0.0754; Prec. 0.8732 ± 0.063; Recall 0.8585 ± 0.1092 ($4.5x10^6$ epochs)	**AUC 0.8686 ± 0.0398**; **MCC 0.4685 ± 0.0892**; **F1 0.7943 ± 0.1617** ($4.5x10^6$ epochs); **Prec. 0.9052 ± 0.1988**; **Recall 0.8585 ± 0.2054** ($4.5x10^6$ epochs)	Learning rate: 0.00047, weight decay penalty: 2.637e-6, $2.5x10^6$ epochs
GraphConv CNN (**VII**)	AUC 1.0	AUC 0.7264	(50 runs, 'sqrt')
RF classifier CDDD features (**V**)	AUC 0.9997	AUC 0.7308	(18, 'sqrt')
DNN classifier CDDD features (**VI**)	AUC 0.8498	AUC 0.7563	Learning rate: 0.00067, weight decay penalty: 4.073e-6, $2.5x10^6$ epochs
RF regression myfeats (**III**)	R^2 = 0.8817	R^2 = -0.0520	(10 runs, 'log2')
DNN regression myfeats (**IV**)	R^2 = 0.2721	R^2 = -0.1926	four-fold cross-validation, learning rate: 0.000359 weight decay penalty: 8.831e-6, nb. epochs: 20

The models for regression did not achieve good R^2 values, which is logical due to the awkwardness of fitting a regression to a binary outcome value. For RF in regression mode: Best hyperparameters: (10, 'log2'), giving R^2 values of train_score: 0.8817, validation_score: -0.0520. For CNN in regression mode, after four-fold cross-validation, with the best hyperparameters: {'learning_rate': 0.000359206871754689, 'weight_decay_penalty': 8.830664294504987e-06, 'nb_epoch': 20}, R^2 values were: train_score: 0.2721: validation_score: -0.1926.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2

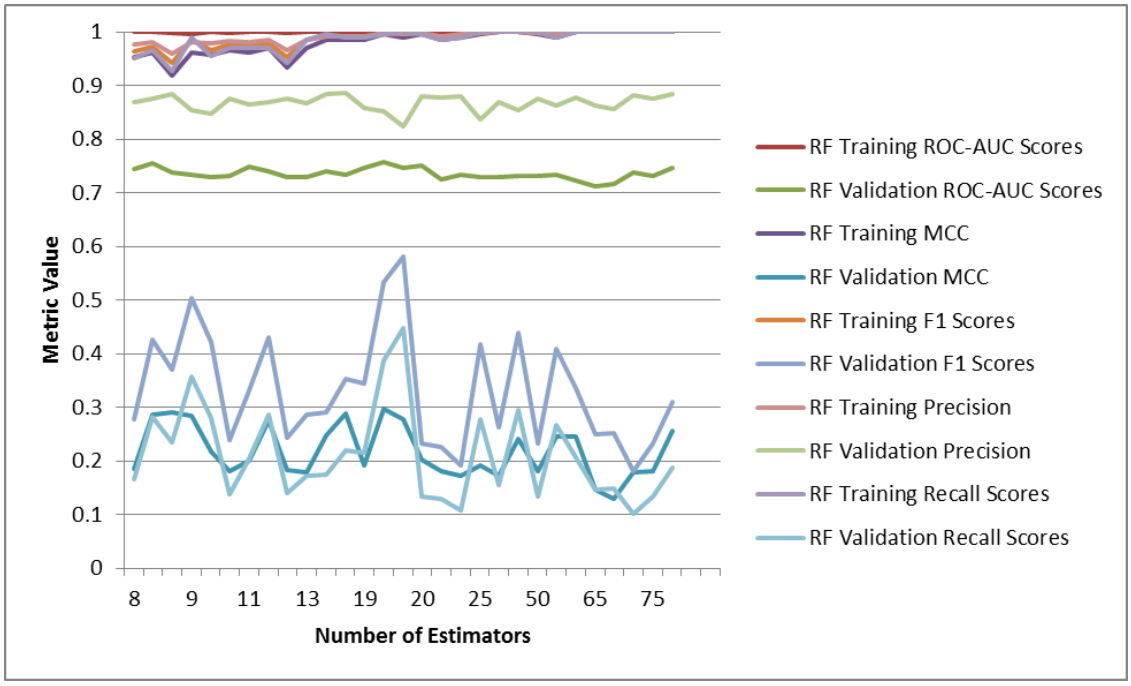


Figure 4. Metrics for different Random Forest models with varying number of epochs. MCC=Mathew's correlation coefficients for balanced training test and validation test.

The results (Figure 4) show that increasing the number of estimators increases the degree of overfitting for the balanced training set. Tests on the full (unbalanced) initial training set show less overfitting, as well as using DNN (Figure 5).

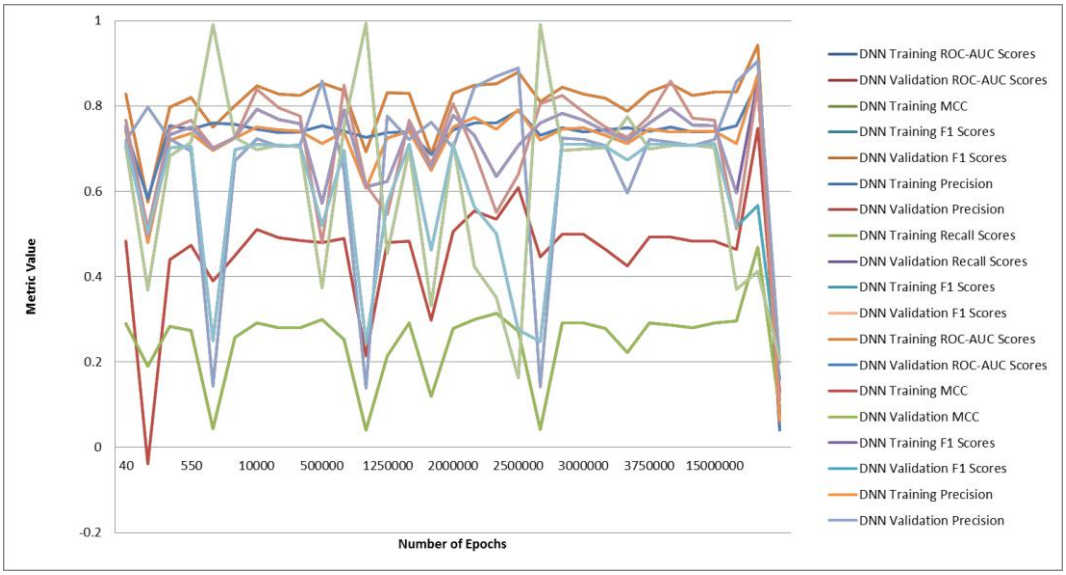
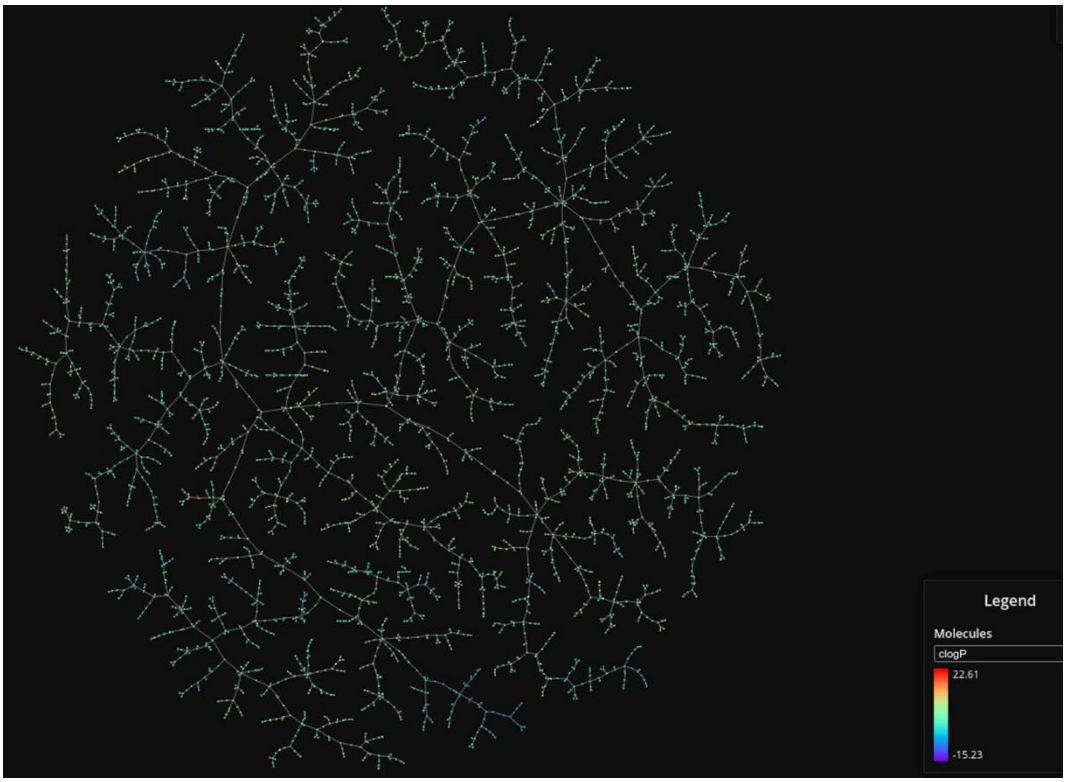


Figure 5. Metrics for different Deep Learning models with varying number of epochs. MCC=Matthews correlation coefficients for balanced training test and validation test.

DNN were better predictors than RF models, for a variety of metrics (Figure 5). The results obtained are good for ROC-AUC, accuracy, F1-scores, and precision metrics for training (balanced), validation, as well as full training set (Table 1 and out.txt at https://github.com/AlfonsoTGarcia-Sosa/ML). The MCC scores obtained are reasonable, considering the lack of balance in the datasets, as well as the lack of distinction in features between binders and non-binders in the evaluation set. This lack of feature distinction can be seen on Figures 6 and 7, where t-maps, histograms and density maps for calculated

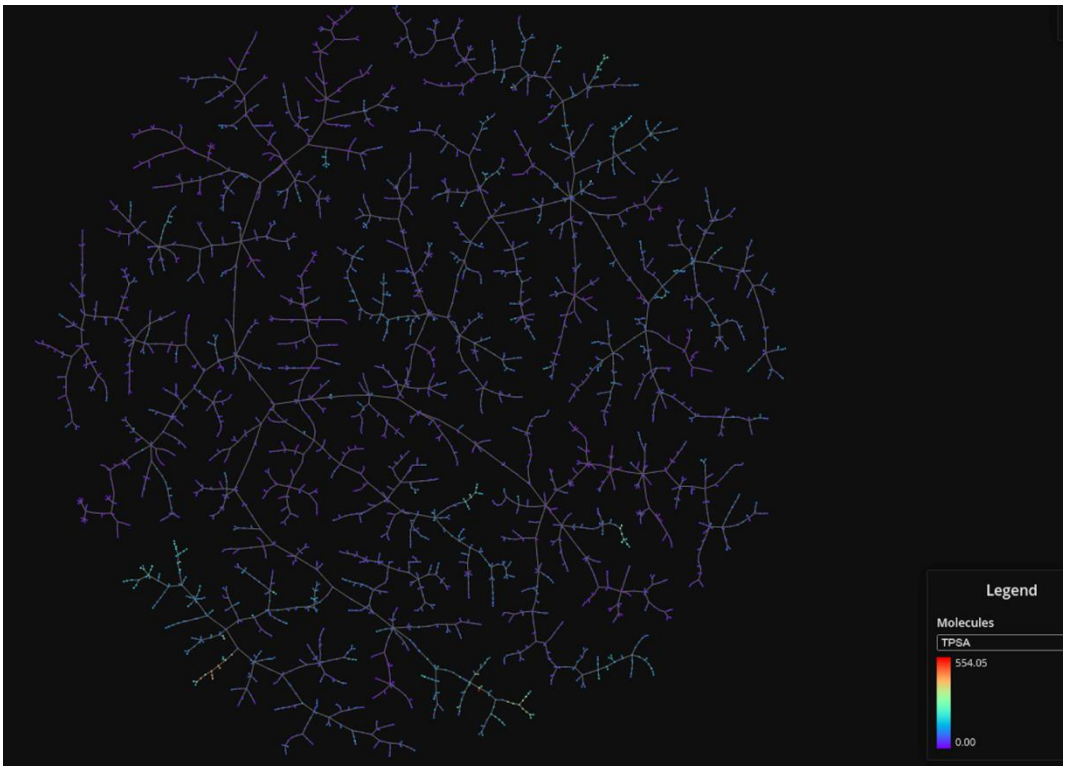
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2

octanol/water partition (clog*P*), topological polar surface area (TPSA, Å^2), number of heterocycles, molecular weight (MW, g/mol), number of rotational bonds (nRotB), number of hydrogen bond donors (nHBDon), number of hydrogen bond acceptors (nHBAcc), and Alog*P* show a highly overlapping distribution for binders as well as nonbinders in the evaluation set, highlighting the difficulty of classification for this dataset.

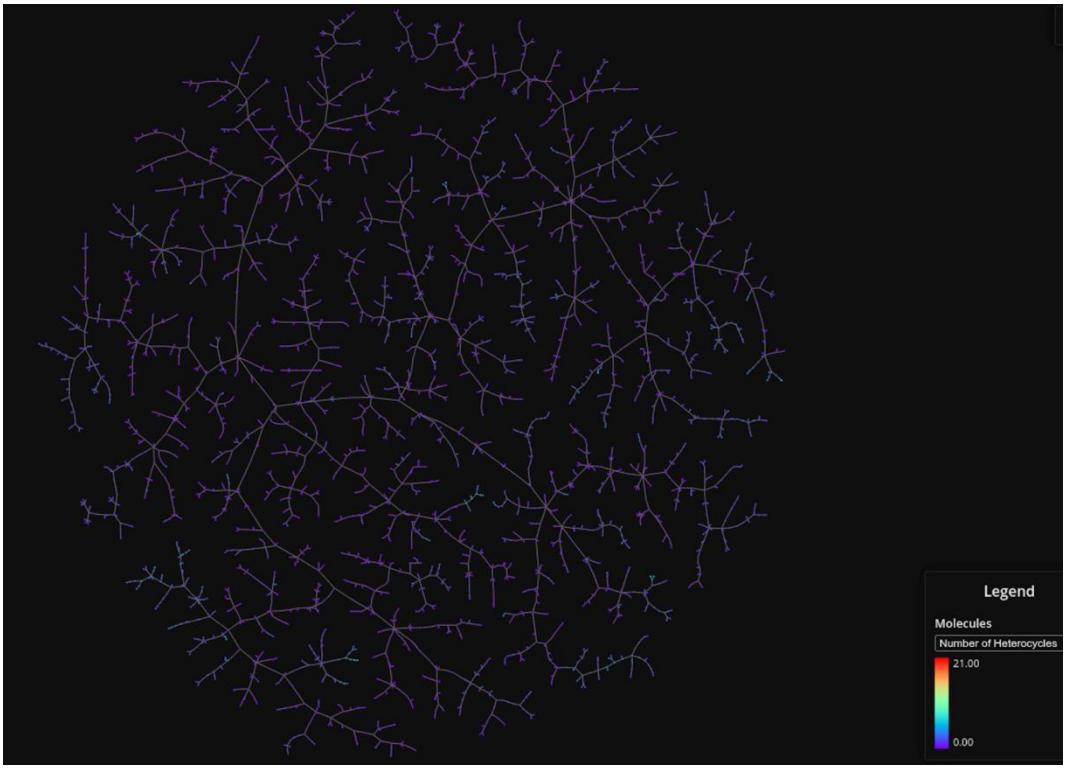


(**a**)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2



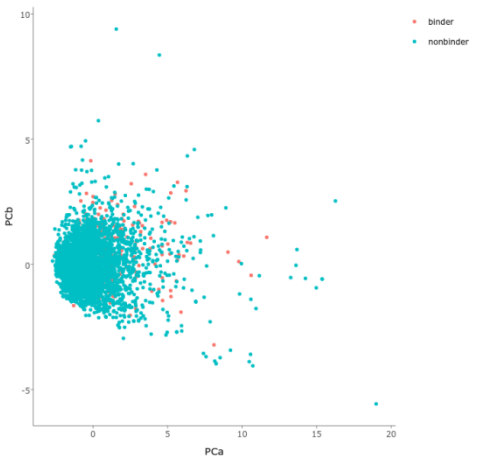
(**b**)



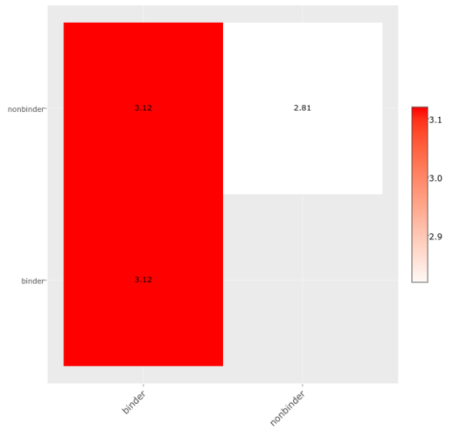
(**c**)

Figure 6. t-Maps for binders and nonbinders in the evaluation set. **a**) clog*P*; **b**) TPSA (Å^2); **c**) Number of heterocycles.

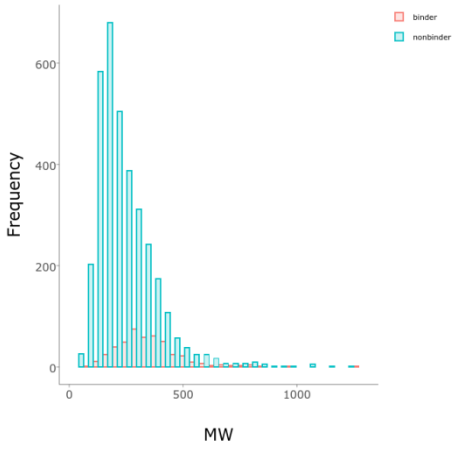
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2



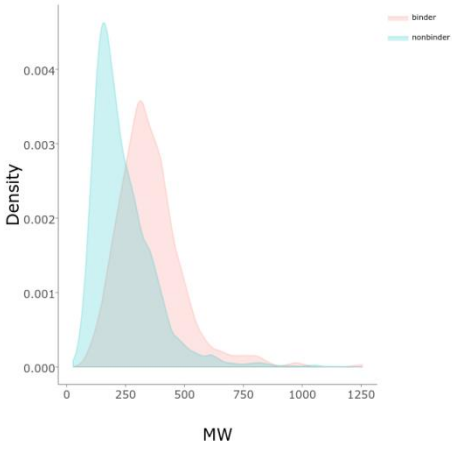
(a)



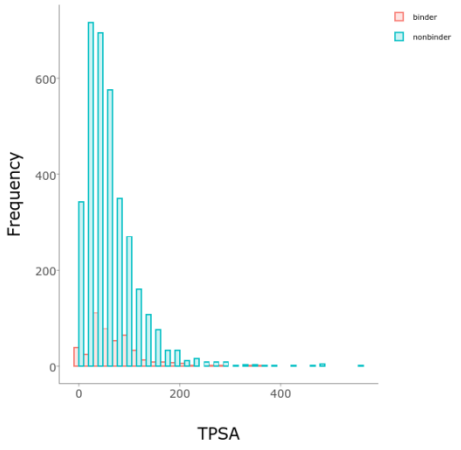
(b)



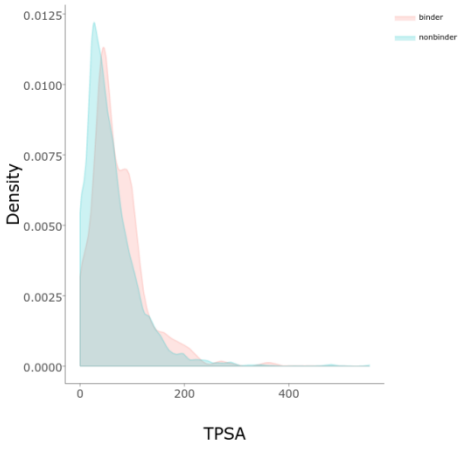
(c)



(d)

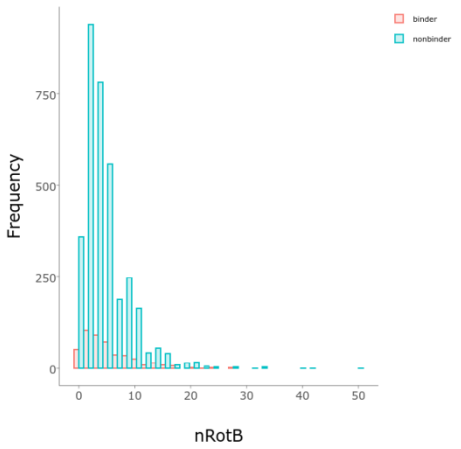


(e)

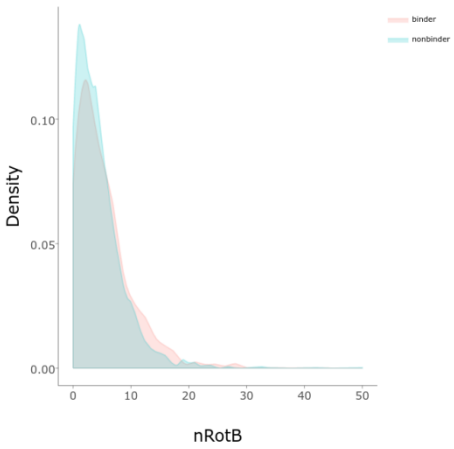


(f)

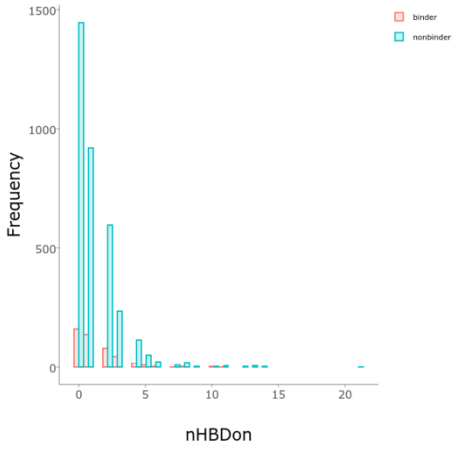
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2



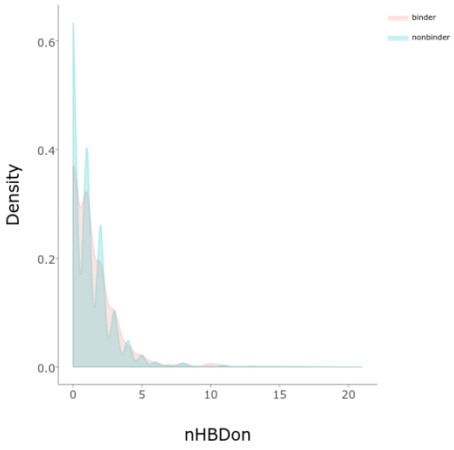
(g)



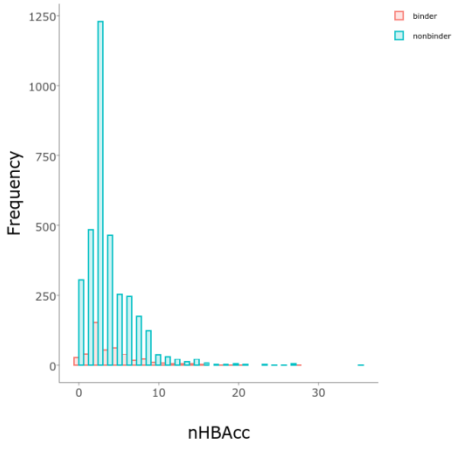
(h)



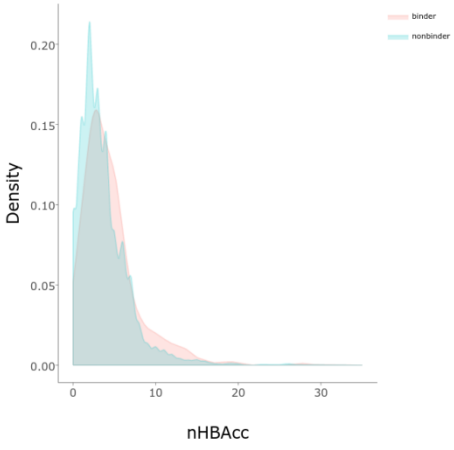
(i)



(j)

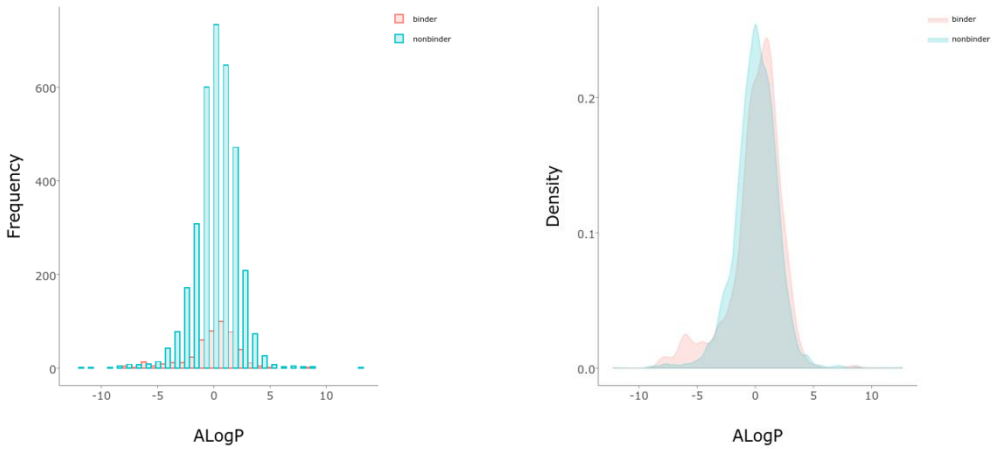


(k)



(l)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2

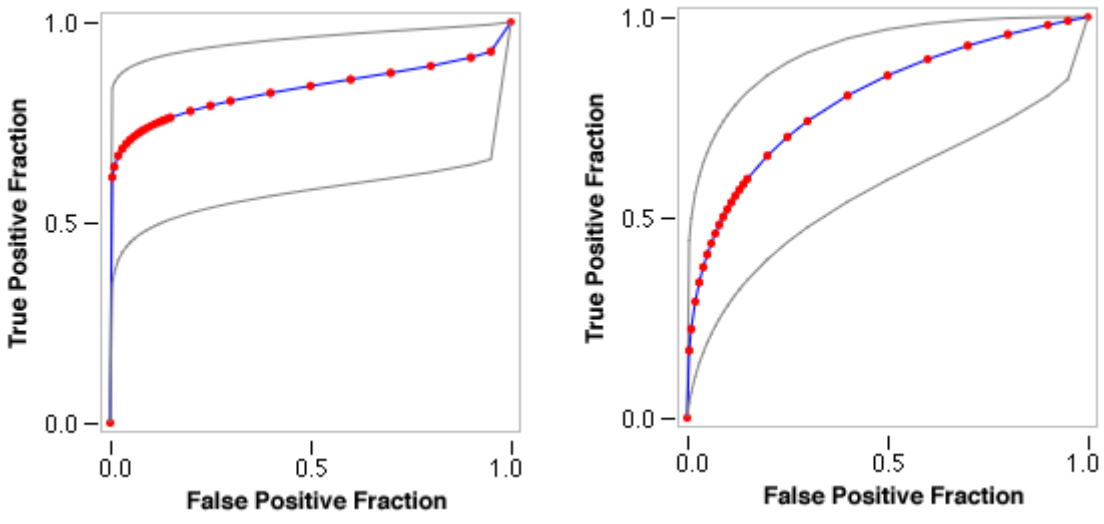


(m) **(n)**

Figure 7. PUMA: **a**) principal component analysis; **b**) balance; and histograms and density plots for binder and nonbinders in the evaluation dataset according to: **c,d**) MW (g/mol); **e,f**) TPSA (Å^2); **g,h**) number of rotational bonds (nRotB); **i,j**) number of hydrogen bond donors (nHBDon); **k,l**) number of hydrogen bond acceptors (nHBAcc); and **m,n**) Alog*P*.

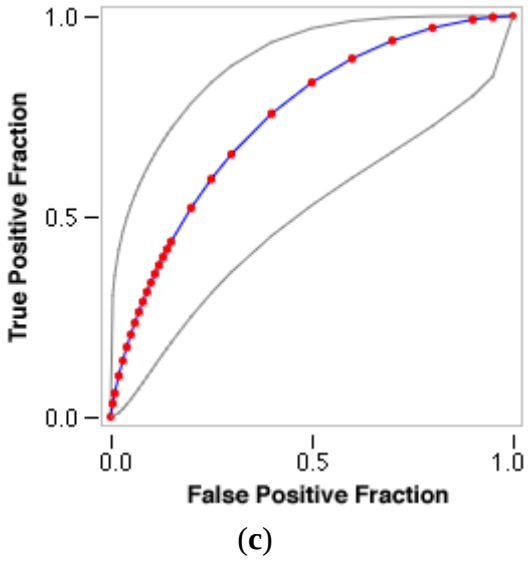
2.4 Validation of docking and comparison to other methods

Decoys were generated for the androgen receptor using the DUD-E database (http://wiki.bkslab.org/index.php/DUDE) and Receiver-Operator Curves (ROCs) and Area Under the Curves (AUCs) for the Human, Chimp, and Rat androgen receptor docking scores were calculated and plot (Figure 8), showing a good separation of true positive from false positives in the most important, *i.e.*, initial parts of the curves. Their values are high, and the chimp protein again shows that it is the most suited with an AUC of 0.832, and an enrichment factor at 1% of 68.92. AUC for human was 0.797, and AUC for rat was 0.744.



(a) **(b)**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2



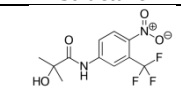
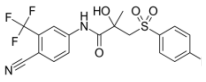
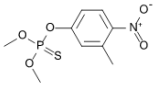
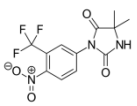
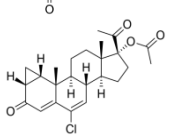
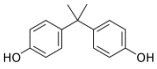
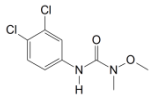
(c)

Figure 8. Receiver-Operator Curves (ROCs) and Area-Under-the Curves (AUCs) for the: **a**) Human; **b**) Chimp; and **c**) Rat androgen receptor docking scores. Chimp AUC = 0.832; Human AUC = 0.797; and Rat AUC = 0.744

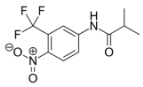
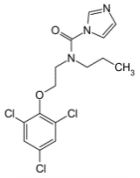
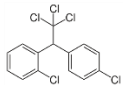
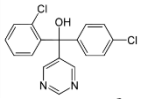
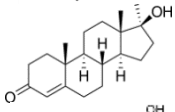
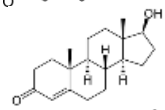
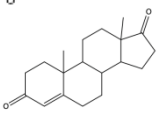
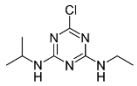
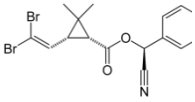
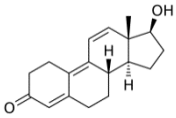
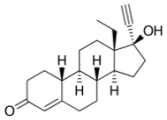
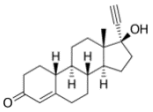
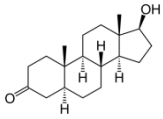
Comparing our results to a structure-based approach by Trisciuzzi et al.[7], they obtained the highest AUC of 0.76 for structures 2pnu and 2hvc, compared to 0.83 for the Chimp AUC in this work.

Using 20 gold-standard reference androgen receptor probe compounds as used by Kleinstreuer et al. [19] shows that there was a good result for predictions of 16/20, i.e., 80% were predicted correct for being a binder to AR (very weak binders were considered as non-binders).

Table 2. AR pathway in vitro reference compounds and their predicted class according to DNN classifier (II).

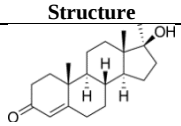
CAS	Name	Structure	Agonist	Antagonist	Predicted	Correct
52806-53-8	hydroxyflutamide		NA	Strong	1	Yes
90357-06-5	bicalutamide		NA	Strong	1	Yes
122-14-5	fenitrothion		NA	Strong	0	X
63612-50-0	nilutamide		Negative	Moderate	0	X
427-51-0	cyproterone acetate		Weak	Moderate	1	Yes
80-05-7	bisphenol a		NA	Moderate/weak	1	Yes
330-55-2	linuron		NA	Moderate/weak	0	X

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2

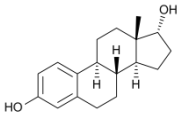
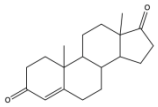
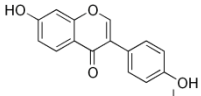
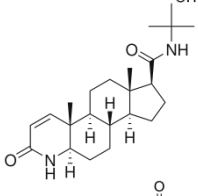
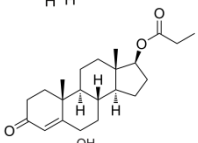
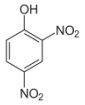
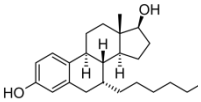
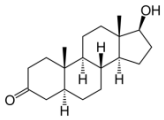
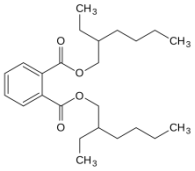
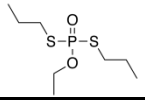
13311-84-7	flutamide		Negative	Moderate/weak	0	X
67747-09-5	prochloraz		Negative	Moderate/weak	1	Yes
789-02-6	*o,p'*-ddt		Negative	Weak	1	Yes
60168-88-9	fenarimol		Negative	Very weak	0	Yes
58-18-4	methyl testosterone		Strong	Negative	1	Yes
58-22-0	testosterone		Strong	Negative	1	Yes
63-05-8	4-androstenedione		Moderate	Negative	1	Yes
1912-24-9	atrazine		Negative	Negative	0	Yes
52918-63-5	deltamethrin		Negative	Negative	0	Yes
10161-33-8	17b-trenbolone		Strong	NA	1	Yes
797-63-7	levonorgestrel		Strong	NA	1	Yes
68-22-4	norethindrone		Strong	NA	1	Yes
521-18-6	5a-dihydrotestosterone		Strong	NA	1	Yes

With respect to well-known compounds that are frequently misclassified [20], the results provided here (Table 3) show four out of 11 compounds correctly predicted compared to three out of 11 reported elsewhere, the difference being the correct prediction of finasteride [20]. Chemical structures in Tables 2 and 3 show several steroid core structures that may be difficult for algorithms to distinguish between actives and inactives, given the strong chemical similarity between them.

Table 3. Chemicals That Were Frequently Predicted Inaccurately by Other Machine Learning Models[20].

CAS	Name	Structure	Agonist	Antagonist	Predicted by II	Correct
58-18-4	methyl testosterone		Strong	Negative	1	Yes

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2

57-91-0	17α-estradiol		Inactive		1	X
63-05-8	4-androstenedione		Moderate	Negative	1	Yes
486-66-8	daidzein			Inactive	1	X
98319-26-7	finasteride			Inactive	0 (RF)	Yes
57-85-2	testosterone propionate		Strong	Inactive	1	X
51-28-5	2,4-dinitrophenol		Negative	Negative	0	Yes
129453-61-8	fulvestrant			Inactive	1	X
84-74-2	dibutyl phtalate		Active		0	X
117-81-7	diethylhexyl phtalate			Active	0	X
13194-48-4	ethoprop			Active	0	X

The loss function diagram for the best DNN (Figure 9, II) shows a relatively stable function with loss for the training set starting around 0.5 and fluctuating moderately up to 0.6 but then decreasing steadily to around 0.27, while the loss function for the validation set fluctuates moderately starting from 0.3 to 0.26.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2

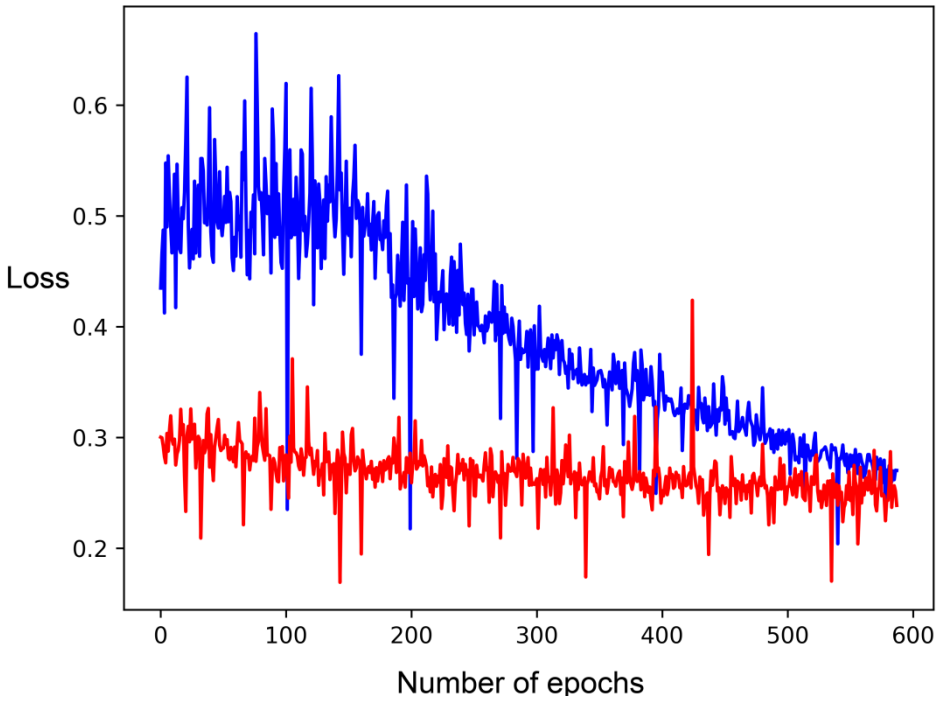


Figure 9. Loss functions of training (blue) and validation (red) datasets by number of epochs in DNN with user-defined features (II).

The approach of using physicochemically and biochemically-relevant user-defined features (12 features, II) is seen in the better metrics performed by the DNN trained and validated on these rather than DNN trained on cddd descriptors (around 500 features, VI), as well as the closely performing RF on the user-defined features (I), and also being better than the CNN using vector featurization of the molecular graph (VII). In addition, the use of ML is warranted in this case, since the same features used in a multivariate logistic regression fashion produced metrics that were not as good, evaluation set MCC = 0.468 and training MCC = 0.868 for the present work compared to evaluation set MCC = 0.2036 and training set MCC = 0.5364 for the multivariate logistic regression on the full, unbalanced set [8].

3. Discussion

Toxicity classification problems can benefit from using DL and specific features that have rationalization on the biochemically and chemically relevant features of the compounds. In this case, predicted binding category to chimp, rat, and human androgen receptor structures, in addition to average Tanimoto distances to known binders and non-binders, as well as Naïve Bayesians as user-provided features to a DNN provided the best results. Unbalanced datasets can be transformed to balanced sets by dropping cases and perform better in training and evaluation. In particular, it is hard to produce toxicity data, especially chronic data, for chemicals with animals and humans, and this lack of data can translate into unbalanced datasets and difficulties for classification and regression techniques. Bias in datasets can be treated with different approaches, such as undersampling [21], as well as distribution following [18]. However, it is clear that more data is beneficial for methods and interpretations of androgen receptor pathway modulators and their toxicity potential. In addition, the most relevant biological assay, be that chimp, rat or human, or their use in consensus, may provide the best experimental setup for classification data.

Evidently, there are considerations to be taken about how to classify kinetic data [21, 22], in many cases of biological interest, e.g., in antibody interactions, complex formation steady-state is not reached [21]; to distinguish between binding sites; and that analyzing interaction data from biosensor instruments is based on the simplified assumption that

 doi:10.20944/preprints202102.0318.v2

larger biomolecules interactions are homogeneous [22]. Also, that for the CoMPARA challenge [6], the organizers (EPA) used the thresholds determined in the CERAPP project and applied them to AR concentration-response values (AC_{50}) from the literature, using the following scheme among several possible:

- Strong: Activity concentration < 0:09 μM
- Moderate: Activity concentration 0.09 - 0:18 μM
- Weak: Activity concentration 0.18 - 20 μM
- Very Weak: Activity concentration 20 - 800 μM
- Inactive: Activity concentration > 800 μM

The use of ML in the form of DNN with user-specified features on a balanced set provides better results as compared to the same features in a multivariate logistic fashion, as well as purely structure- or ligand-based approaches, as seen by better AUC, MCC, and other values.

4. Materials and Methods

4.1. Training Set

The training set of compounds was provided during the CoMPARA challenge for predicting androgen receptor activity for chemicals [6], and included curated data with SMILES strings. This training set was composed of state-of-the-art experimental data from ToxCast [24], Tox21 [25], and DrugBank [26] databases, amounting to 1,673 chemical compounds with 205 positives (binders), and 1,468 negatives (non-binders). Binders were coded as actives (“1”), non-binders were coded as inactives (“0”). The SMILES strings were used as present in the files. Given that the training set was heavily unbalanced, the training set was balanced using pandas tools [27].

4.2. Independent Evaluation Set

The evaluation data set was also provided in the CoMPARA challenge [6] from different databases being completely independent from the training set: EPA’s NCCT collected and curated PubChem data (64 sources) [6, 28]. After including only binding data, there were 3,882 compounds in the evaluation data set, composed of 446 positives (binders) and 3,437 negatives (non-binders). No balancing was performed for the evaluation set for an unbiased evaluation of the models.

4.3. Features

The structures for the human, chimp, and rat androgen receptor were downloaded from the PDB [29] (codes 3v49, 1t7r, and 3g0w) based on their resolution (1.4, 1.7, and 1.95 Å, resp.), completeness of sequence, and relevance of the complex. Protein X-ray crystal structures were preprocessed with the Protein Preparation Wizard from Schrödinger [30]. Docking scores were generated with Glide XP [31] centered on the orthosteric site of AR using 15 Å inner box and 40 Å outer search boxes, that different to default settings. The results of each docking run were used as structure-based features 'HumDockScore', 'ChimpDockScore', and 'RatDockScore' for the docking scores in kcal/mol of the human, chimp, and rat AR structures, respectively. The average of the docking scores for the three protein targets was also computed and stored as feature ‘AVG’[32,33].

For ligand-based features, Extended connectivity fingerprints (ECFP), circular topology-based representations of compounds, were calculated with ChemAxon [36]. Distances between compound fingerprints were calculated by Tanimoto fingerprints using OpenBabel [37], giving ligand-based features named 'avgD_Act' and 'avgD_Inact', respectively, for the calculated distance to the average of known active and inactive compounds of the ECFP fingerprint for each compound.

Naïve Bayesians (NBs) were constructed using the means and standard deviations of the docking scores of actives to the chimp receptor, and the probability given for each

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2021 doi:10.20944/preprints202102.0318.v2

group 'P_Act_dockChimp', and 'P_Inact', respectively, were used as statistical features, as well as their ratio and Bayesian prediction (feature 'PredBayes') corresponding to which value of P calculated to each distribution was greater, with binding = 1, and non-binding = 0, corresponding to ratio > 0.5 and ratio < 0.5, respectively.

Another feature created was 'predMLogR', the probability from a multivariate logistic classifier using these variables, as calculated in [8]. 'PredBindingClass' is also a features, defined as the binary value for this predicted probability, i.e. 'PredBindingClass' = 1 if 'predMLogR' > 0.5, or else 'PredBindingClass' = 0. This is a distinct feature from the NB prediction above.

For comparison, the cddd group of latent-space encoded ligand-based descriptors was also used as described in the original publication [37].

4.4. Models

Three types of model were run: Deep Neural Networks (DNN), and for comparison, Random Forest (RF), and Graph Convolutional Neural Networks (CNN). Two types of featurization were used for DNN and RF: the cddd groups of ligand-based descriptors; and our own, user-specified features from structure-based (docking), ligand-based (fingerprint distances), and statistically-based features. RF and DNN were run both as Classifiers (models I and II, respectively), and for comparison, as Regressors too (III and IV) with deepchem [17]. Two types of featurizations were used: 1) user-specified features calculated for the compounds ('HumDockScore', 'RatDockScore', 'ChimpDockScore', 'AVG', 'P_Act_dockChimp', 'P_Inact', 'PredBayes', 'ratio', 'avgD_Act', 'avgD_Inact', 'PredBindingClass', 'predMLogR'; see section *4.3 Features*, above); and 2) the cddd groups of ligand-based descriptors [37] (RF_cddd V, DNN_cddd, VI). The cddd featurization (512 features exploring the continuous descriptor space) have been reported to give good results for ML models for prediction of compound properties such as solubility and quantitative structure–activity relationships, as well as ligand-based virtual screening tasks [37]. CNN were also employed using atom-based featurization (VII). CNN models tend to be largely used for graphical data, with pixels or vector representations, for example. They have also been reported to give good results on compound property predictions; a study found that CNNs where atom properties are used instead of pixels are more accurate than DNN for predicting quantum chemical energies [10]. Batch size was 128. 10 fold cross-validation was used, as well as the ROC-AUC as guiding metric. The models I and II probed features "max_features": ["auto", "sqrt", "log2", None]. Number of epochs was also varied for the DNN models, from 1 to 30,000,000.

4.5. Metrics.

In all cases, the task classification or regressor was the "binding Class" status of the compounds, that is, coded 1 for binders and 0 for nonbinders that represented experimental actives and inactives for androgen receptor. Validation metrics included Area-Under-the Curve (AUC) measurements of the Receiver-Operator Curve of true positive and false positive rates, that range from 0 (complete misclassification) to 1.0 (complete classification), precision, recall, Matthews correlation coefficient (MCC), F1-score, and accuracy as determined by sklearn [38].

A confusion matrix has four fields: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Specificity (SP) is calculated as: TN/(TN+FP), Precision: TP/(TP+FP), Recall (sensitivity, SE): TP/(TP+FN), Accuracy (Acc.): (TP+TN)/(TP+FP+FN+TN), and Matthews correlation coefficient (MCC). MCC is calculated as:

$$MCC=(TP\cdot TN-FP\cdot FN)/\sqrt{((TN+FN)\cdot(TN+FP)\cdot(TP+FN)\cdot(TP+FP))} \quad (1)$$

F1 score is calculated as:

 doi:10.20944/preprints202102.0318.v2

$$F1 = 2\ ((Precision \times Recall))/((Precision + Recall)) \quad (2)$$

Dataset diversity analysis and visualization were performed with PUMA: Platform for Unified Molecular Analysis, Version 1.0 [39], as well as with t-map [40] that uses MHFP6 fingerprints [41].

All data and code were run on jupyter notebooks and python, deposited, and made available on github at: https://github.com/AlfonsoTGarcia-Sosa.

5. Conclusions

DNNs with 12 user-specified structure-, ligand- and statistically-based features were found to perform best at categorizing AR binders and nonbinders. They outperformed DNN with cddd features, as well as RF, and CNN methods, as well as regressors (expectedly, given the sharp category bins), as well as the same features in a multivariate logistical fashion, as well as simple docking. Implications are that explainability in ML features is important, as physicochemically- and biologically-relevant descriptors can perform best at the categorization for this particular AR dataset. In addition, different ML techniques may be best suited for different application tasks, with DNN performing better than RF given the overtraining seen in the RF models. CNN models may require more information, such as the protein-ligand binding pose or trajectories. The cddd featurizations may well perform better for property prediction and virtual screening. In the present work, the Chimp structure-based features performed better than other protein-derived features in this work and others published elsewhere. Improvement is still possible, given that the MCC can be higher for the evaluation compounds even if predictions obtained were good and improved on predictions for a golden standard of AR reference compounds. Better data, that is, less unbalanced and with better structural diversity may help improve future predictions, as could be combining the present features with other ML and non-ML techniques, such as boosting, or molecular dynamics simulations, respectively.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, project administration, funding acquisition, A.T.G.-S. All authors have read and agreed to the published version of the manuscript.

Funding: Please add: This research was funded by Haridus- ja teadusministeerium, grant IUT34-14

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: All of the data used is provided in text file format, all jupyter notebooks and python code used are also provided and are available at the public repository: https://github.com/AlfonsoTGarcia-Sosa

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Sifakis, D.; Androutsopoulos, V. P.; Tsatsakis, A. M.; Spandidos, D. A. Human exposure to endocrine disrupting chemicals: effects on the male and female reproductive systems. *Environ. Toxicol. Pharmacol*. **2017**, *51*, 56-70, https://doi.org/10.1016/j.etap.2017.02.024

2. Cheung, A.; Zajac, J.; Grossmann, M. Muscle and bone effects of androgen deprivation therapy: current and emerging therapies. *Endocr.-Relat. Cancer* **2014**, *21*, R371-R394. https://erc.bioscientifica.com/view/journals/erc/21/5/R371.xml
3. Manolagas, S. C.; O'Brien, C. A.; Almeida. M. The role of estrogen and androgen receptors in bone health and disease. *Nat. Rev. Endocrinol*. **2013**, *9*, 699–712. https://doi.org/10.1038/nrendo.2013.179
4. Mendelsohn, M. E.; Karas, R. H. Molecular and Cellular Basis of Cardiovascular Gender Differences. *Science* **2005**, *308*, 1583—1587. https://doi.org/ 10.1126/science.1112062
5. Nadal, M.; Prekovic, S.; Gallastegui, N. et al. Structure of the homodimeric androgen receptor ligand-binding domain. *Nat. Commun.* **2017,** *8*, 14388. https://doi.org/10.1038/ncomms14388
6. Mansouri K. et al. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environmental Health Perspectives* **2020**, *128* (2), 027002
7. Trisciuzzi, D.; Alberga, D.; Mansouri, K.; Judson, R.; Novellino, E.; Mangiatordi, G. F.; Nicolotti. O. Predictive Structure-Based Toxicology Approaches to Assess the Androgenic Potential of Chemicals. *J. Chem. Inf. Model*. **2017**, *57*, 2874-2884
8. García-Sosa, A.T.; Maran, U. Combined docking, chemical fingerprints, and Naïve Bayesian classifiers for the androgen receptor binding activity of the CoMPARA data of substances of environmental and health concern. *submitted*.
9. Schneider, P.; Walters, W.P.; Plowright, A.T.; Sieroka, N.; Listgarten, J.; Goodnow, R.A. et al. Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery* **2020**, *19*, 353–364
10. Peña-Guerrero, J.; Nguewa, P.A.; García-Sosa A.T. Machine Learning, Artificial Intelligence, and Data Science Breaking into Drug Design and Neglected Diseases. *WIREs Comput. Mol. Sci*. **2021**, e1513, https://doi.org/10.1002/wcms.1513
11. Raj, R. J. S.; Shobana, S. J.; Pustokhina, I. V.; Pustokhin, D. A.; Gupta, D.; Shankar, K. Optimal feature selection-based medical image classification using deep learning model in internet of medical things. *IEEE Access* **2020**, *8*, 58006-58017
12. Pustokhina, I. V.; Pustokhin, D. A.; Rodrigues, J. J. P. C.; Gupta D., Khanna, A.; Shankar, K. et al. Automatic vehicle license plate recognition using optimal k-means with convolutional neural network for intelligent transportation systems. *IEEE Access*, **2020**, *8*, 92907-92917
13. Khamparia, A.; Pandey, B.; Tiwari, S.; Gupta, D.; Khanna, A.; Rodrigues, J. J. P. C. An integrated hybrid CNN-RNN model for visual description and generation of captions. *Circuits Syst Signal Process.* **2020**, *39*, 776–788
14. Varela-Santos, S.; Melin, P. A new approach for classifying coronavirus COVID-19 based on its manifestation on chest X-rays using texture features and neural networks. *Inf Sci.* **2021**, *545*, 403–414
15. Yosipof, A.; Guedes, R.C.; García-Sosa, A.T. Data Mining and Machine Learning Models for Predicting Drug Likeness and their Disease or Organ Category. *Frontiers in Chemistry* **2018**, *6*:162 https://doi.org/10.3389/fchem.2018.00162
16. Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2020,** *2*, 573–584. https://doi.org/10.1038/s42256-020-00236-4
17. DeepChem. https://github.com/deepchem/deepchem
18. García-Sosa, A.T. Benford's Law in Medicinal Chemistry: Implications for Drug Design. *Future Medicinal Chemistry* **2019**, *11* (17):2247-2253 https://doi.org/10.4155/fmc-2019-0006
19. Kleinstreuer, N. C.; Ceger, P.; Watt, E. D.; Martin, M.; Houck, K.; Browne, P.; Thomas, R. S.; Casey, W. M.; Dix, D. J.; Allen, D.; Sakamuru, S.; Xia, M.; Huang, R.; Judson, R. Development and Validation of a Computational Model for Androgen Receptor Activity. *Chem. Res. Toxicol*. **2017**, *30*, 946–964 https://doi.org/10.1021/acs.chemrestox.6b00347
20. Zorn, K.M.; Foil, D.H., Lane, T.R.; Hillwalker, W.; Feifarek, D.J.; Jones, F.; Klaren, W.D.; Brinkman, A.M.; Ekins, S. Comparison of Machine Learning Models for the Androgen Receptor. *Environmental Science & Technology* Article ASAP, DOI: 10.1021/acs.est.0c03984
21. Martinez-Mayorga, K.; Madariaga-Mazon, A.; Medina-Franco, J.L.; Maggiora, G. The impact of chemoinformatics on drug discovery in the pharmaceutical industry. *Expert Opinion on Drug Discovery* **2020**, *15* (3), 293-306. https://doi.org/10.1080/17460441.2020.1696307
22. Forssén, P.; Multia, E.; Samuelsson, J.; Andersson, M.; Aastrup, T.; Altun, S., et al. Reliable strategy for analysis of complex biosensor data. *Anal Chem*. **2018**, *90*, 5366-5374
23. Forssén, P.; Samuelsson, J.; Lacki, K., Fornstedt, T. Advanced analysis of biosensor data for SARS-CoV-2 RBD and ACE2 interactions. *Anal Chem.* **2020**, *92*, 11520-11524
24. Dix, D.J.; Houck, K.A.; Martin, M.T.; Richard, A.M.; Setzer, R.W.; Kavlock, R.J. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* **2007,** *95*(1):5–12, PMID: 16963515, 10.1093/toxsci/kfl103
25. Tice, R.R.; Austin, C.P.; Kavlock, R.J.; Bucher, J.R. Improving the human hazard characterization of chemicals: a Tox21 update. *Environ. Health Perspect.* **2013,** *21*(7):756–765, PMID: 23603828, 10.1289/ehp.1205784
26. Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672. doi: 10.1093/nar/gkj067
27. Pandas tools. Python Data Analysis Library. https://pandas.pydata.org/
28. PubChem. National Institutes of Health (NIH). https://pubchem.ncbi.nlm.nih.gov
29. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H., et al. The protein data bank. *Nucleic Acids Res.* **2000** *28*, 235–242. doi: 10.1093/nar/28.1.235
30. Protein Preparation Wizard; Schrödinger, LLC: New York, NY, USA, **2019**.
31. Virtual Screening Workflow; Schrödinger, LLC: New York, NY, USA, **2019**.

 doi:10.20944/preprints202102.0318.v2

32. García-Sosa, A.T.; Sild, S.; Maran U. Docking and Virtual Screening Using Distributed Grid Technology. *SQER* **2009**, *28*(8):815-821 https://doi.org/10.1002/qsar.200810174
33. Viira, B.; Selyutina, A.; García-Sosa, A.T.; Karonen, M.; Sinkkonen, J.; Merits, A.; Maran, U. Design, Discovery, Modelling, Synthesis, and Biological Evaluation of Novel and Small, Low Toxicity s-Triazine Derivatives as HIV 1 Nonnucleoside Reverse Transcriptase Inhibitors. *Bioorg. Med. Chem.* **2016**, *24*(11):2519-2529 https://doi.org/10.1016/j.bmc.2016.04.018
34. Glisic, S.; Sencanski, M.; Perovic, V.; Stevanovic, S.; García-Sosa, A.T. Arginase Flavonoid Anti-Leishmanial in Silico Inhibitors Flagged against Anti-Targets. Molecules 2016, 21, 589 https://doi.org/10.3390/molecules21050589
35. Marvin Beans, version 5.3.8; ChemAxon: Budapest, Hungary, **2010**. http://www.chemaxon.com
36. Open Babel. An Open Chemical Toolbox. http://openbabel.org (accessed 1 Nov, **2020**)
37. Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **2019,** *10*:1692–1701. https://doi.org/10.1039/c8sc04175j
38. scikit-learn. Machine Learning in Python. https://scikit-learn.org/stable/
39. Gonzalez-Medina, M.; Medina-Franco, J.L. Platform for Unified Molecular Analysis: PUMA. *J. Chem. Inf. Model.* **2017**; *57* (8):1735-40.
40. t-map. Reymond group. https://tmap.gdb.tools/index.html
41. MHFP6 fingerprints https://github.com/reymond-group/mhfp/tree/master/mhfp