MDPI

Article

Identifying Protein Features and Pathways Responsible for Toxicity using Machine learning, CANDO, and Tox21 datasets: Implications for Predictive Toxicology

Lama Moukheiber [1], William Mangione [2], Saeed Maleki [3], Zackary Falls [2], Mingchen Gao [1*] and Ram Samudrala [2*]

[1] Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA; lamamouk@buffalo.edu;
[2] Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, USA; wmangion@buffalo.edu; zmfalls@buffalo.edu
[3] Department of Mechanical Engineering, University at Buffalo, Buffalo, NY, USA; saeedmal@buffalo.edu
* Correspondence: ram@compbio.org; mgao8@buffalo.edu

Abstract: Humans are exposed to numerous compounds daily, some of which have adverse effects on health. Computational approaches for modeling toxicological data in conjunction with machine learning algorithms have gained popularity over the last few years. Machine learning methods have been used to predict toxicity-related biological activities using chemical structure descriptors. However, toxicity-related proteomic features have not been fully investigated. In this study, we construct a computational model using machine learning for selecting the most important proteins representing features in predicting the toxicity of the compounds in the Tox21 dataset using the multiscale Computational Analysis of Novel Drug Opportunities (CANDO) platform for therapeutic discovery. Tox21 is a highly imbalanced dataset consisting of twelve *in vitro* assays, seven from the nuclear receptor (NR) signaling pathway and five from the stress response (SR) pathway, for more than 10,000 compounds. For our computational model, we employed a random forest (RF) with the combination of Synthetic Minority Oversampling Technique (SMOTE) and Edited Nearest Neighbor (ENN) method, aka SMOTE+ENN, which is resampling method to balance the activity class distribution. Within the NR and SR pathways, the activity of the aryl hydrocarbon receptor (NR-AhR), toxicity mediating transcription factor, and mitochondrial membrane potential (SR-MMP) were two of the top-performing twelve toxicity endpoints with AUROCs of 0.90 and 0.92, respectively. The top extracted features for evaluating compound toxicity were passed into enrichment analysis to highlight the implicated biological pathways and proteins. We validated our enrichment results for the activity of the AhR using a thorough literature search. Our case study showed that the selected enriched pathways and proteins from our computational pipeline are not only correlated with NR-AhR toxicity but also form a cascading upstream/downstream arrangement. Our work elucidates significant relationships between protein and compound interactions computed using CANDO and the associated biological pathways to which the proteins belong, with twelve toxicity endpoints. This novel study uses machine learning not only to predict and understand toxicity but also elucidates therapeutic mechanisms at a proteomic level for a variety of toxicity endpoints.

Keywords: machine learning; random forest; feature selection; structure–activity relationships; high-throughput screening; enrichment analysis; proteomic signature; toxicity; drug behavior

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Exposure to persistent natural and synthetic environmental pollutants continues to be a significant health concern [1]. With technological advances in computational toxicology, systems biology, and bioinformatics, researchers and regulators have access

to tools that allow for rapid assessment of toxic compounds, reducing the usage of low-throughput, expensive, and time-consuming *in vivo* animal testings [2,3]. High-throughput screening (HTS) [4,5] has been utilized in conjunction with computational models to profile compounds for potential adverse effects and assess how compounds interact with biological systems. Furthermore, in the past few years, quantitative high-throughput screening (qHTS) has emerged as a powerful tool to allow the study of toxicological pathways linked with toxicity endpoints [6].

The Toxicology in the 21st Century (Tox21) program has been established as a collaborative effort among federal entities, including the National Center for Advancing Translational Sciences (NCATS), the National Institute of Environmental Health Sciences (NIEHS), the Environmental Protection Agency (EPA), and the Food and Drug Administration (FDA), to advance toxicity assessment. This program has applied quantitative high-throughput screening (qHTS) to profile a library of around 10,000 compounds, including but not limited to environmental hazards, industrial chemicals, drugs, and food additives [7–10]. The Tox21 compound library was run against a panel of seven nuclear receptors (NR) and five stress response (SR) pathway assays, generating the most significant high-quality *in vitro* toxicity data to date [11]. Data generated from Tox21 has been used to identify compounds that interact with specific toxic pathways, including some not previously known [12–15]. This data has also been used, together with chemical structure information, to train predictive machine learning models to provide a prediction of whether a chemical will elicit a particular toxicological outcome based on *in vitro* findings [16–21]. Deep learning has achieved the best prediction performance in toxicity prediction on the Tox21 data. However, one major limitation of deep neural networks is that they are hard to reveal effective chemical descriptors in the toxicity prediction [22].

This paper proposes a novel approach to build a model that predicts the essential proteins *in vivo* toxicity, which are then fed into an enrichment analysis to assess the mechanistic pathways contributing to each toxicity endpoint and provide a high-level biological interpretation of toxicity. To extract important protein feature descriptors which contribute most to the target prediction, many feature selection techniques originating from machine learning have been proposed [23–26]. Random forest (RF) [27], which is an embedded feature selection technique, has emerged as an efficient and robust algorithm that can address feature selection, even with the higher number of variables [28–30]. RF constructs many decision trees and averages the predicted values to obtain the final ones. Due to the random exploration of features at each node in the tree construction, RF lends itself to feature selection. Determining and assessing the most relevant feature descriptors between compounds and toxicity endpoints on a mechanistic and proteomic scale will comprehensively evaluate compound toxicity.

Previous studies using machine learning models on Tox21 data have shown that the data is highly imbalanced, with more inactive toxic assays instances than active ones [31–33]. Highly imbalanced data can be a problem when training machine learning classification models, as the model becomes biased toward the inactive class, resulting in a higher misclassification rate for the active class. Most of the previous classification algorithms using Tox21 data have not handled the imbalanced problem of toxicity prediction explicitly [31,32,34,35]. We propose an improved RF method for feature selection to find the relevant features for toxicity prediction, i.e., RF with the combination of Synthetic Minority Oversampling Technique (SMOTE) and Edited Nearest Neighbor (ENN) method, aka SMOTE+ENN, which is a resampling method that can handle the imbalanced problem.

The Tox21 dataset has been well studied using chemical descriptors as their feature space to understand toxicity [16,33]. In contrast, this study utilizes the Computational Analysis of Novel Drug Opportunities (CANDO) platform to obtain the protein feature descriptors to understand toxicity at the protein pathway level. CANDO is a multiscale shotgun drug discovery, repurposing, and design platform which employs multitarget-

ing to generate proteomic scale interaction signatures for approved and investigational small molecule therapeutics against large libraries of protein structures from various organisms [36–43]. These proteomic signatures are analyzed to computationally assess compound similarity, with the premise that drugs with similar signatures may treat the same diseases. Version 3 of the platform now features several protein-related and drug-related biological entities, such as protein pathways, protein-protein interactions, protein-disease associations, and adverse drug reactions, further enhancing the ability to compare small molecule compounds in the context of biological systems. CANDO has been validated in multiple indications with an overall success rate of 58/163, not including 51 drug candidates with activity against SARS-CoV-2 (out of 275 ranked predictions) extracted from *in vitro* and electronic health record-based studies published in the literature [44]. However, CANDO has never been employed explicitly to predict compound toxicity, despite how conducive this multiscale framework is for that task.

This study aims to improve toxicity prediction by combining the capabilities of machine learning, the CANDO platform, and enrichment analysis to identify the most effective protein structures for predicting 12 toxicity endpoints from the Tox21 data and highlight implicated biological pathways. We begin the study by pre-processing the data, generating protein feature descriptors using CANDO, and performing dataset balancing. Following this, we apply the RF algorithm to select the top N structural protein descriptors and subject them to enrichment analysis to identify high-level biological entities that explain the mechanisms through which the toxicity is induced. We provide a case study of aryl hydrocarbon receptor (AhR) activation, comprehensively demonstrating the pathways potentially responsible for its associated toxicity using only those identified via enrichment analysis. Our approach of combining CANDO, machine learning model, and feature selection allow for a detailed understanding of compound behavior and, therefore, a greater ability to predict not only toxicity but also therapeutic mechanisms.

2. Results

In this section, we present (1) a summary of the curated and prepossessed Tox21 datasets, (2) performance metrics of our computational algorithm on the Tox21 data, (3) a comparison between our study and other published Tox21 studies, (4) the enriched pathways for the 12 Tox21 assays, (5) a case study of the NR-AhR toxicity with literature studies supporting our computational model results and enrichment analysis. Figure 1 provides a high-level overview of our study design (see Section 4).

2.1. Preprocessed data

Table 1 shows the preprocessed Tox21 compounds and their activities measured by 12 qHTS *in vitro* assays. The number of active and inactive compounds for each of the 12 qHTS assays was computed, along with the imbalanced ratio (IR), which is the ratio of the majority class (inactive non-toxic compounds) to the minority class (active toxic compounds). The IR varied greatly between the 12 assays, with the inactive compounds being the predominant majority (ratio 10:1 or higher) compared to active compounds. The higher the IR value, the more imbalanced the activity class distribution for the assay. In the training datasets, the IR ranged from 5.52 for the SR mitochondrial membrane potential (SR-MMP) assay up to 37.62 for the NR peroxisome proliferator-activated receptor γ (NR-PPAR-γ) assay. The test datasets had IRs larger than or equivalent to those of their corresponding training datasets, and the IR values ranged from 5.06 for the SR antioxidant response element (SR-ARE) assay up to 69.75 for the NR androgen receptor ligand-binding domain (NR-AR-LBD) assay.

2.2. Our study and other Tox21 studies

This study used the RF classifier to avoid overfitting and provide promising performance. Further, it was a commonly used model by participating teams in the Tox21

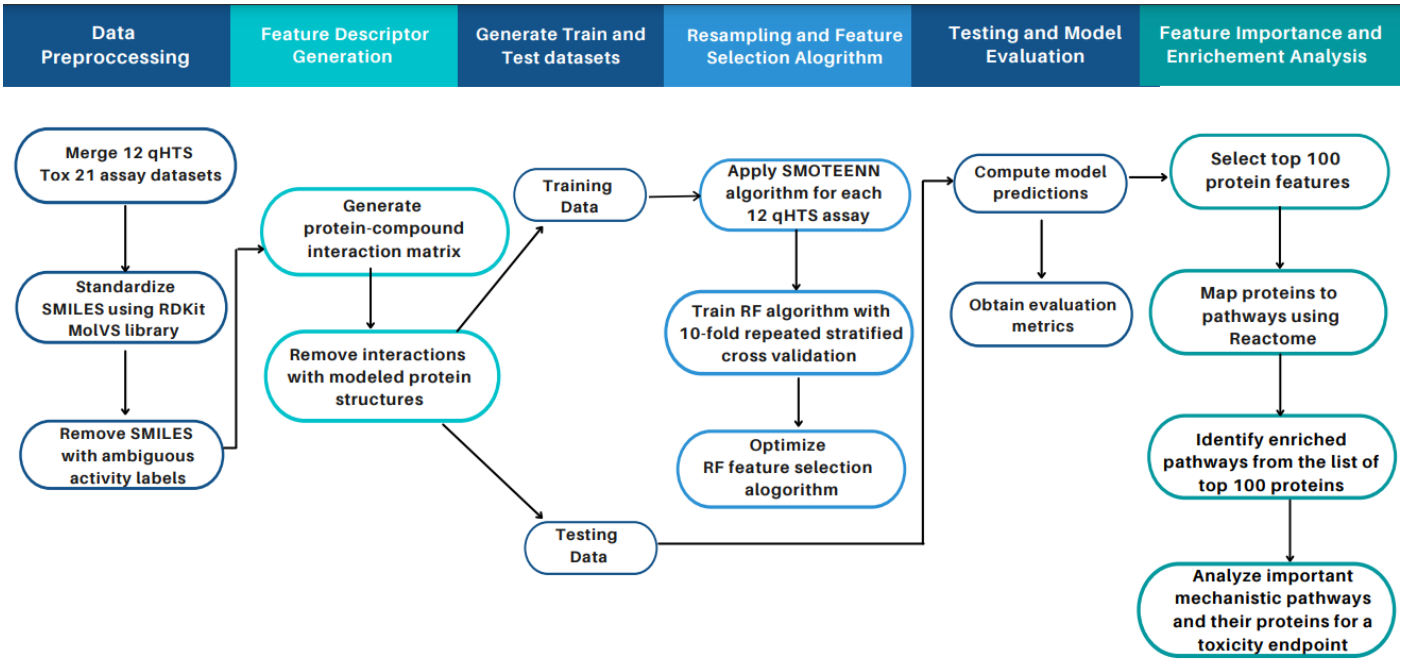

Figure 1. An overview of our study workflow and computational model pipeline. As part of the data pre-processing step, the Tox21 12 assay datasets containing the SMILES and class activity of the compounds were merged. The compounds were normalized and standardized using the MolVs library, built on RDKit. Compound duplicates with ambiguous activity labels (i.e., equally active and inactive outcomes for the same compound) were removed. Model features were generated using a protein-compound interaction matrix, and the features were filtered for interactions with only solved protein structures. The data containing the features and the class activity for the compound was generated for each of the 12 assays. Each data was split into a training and testing set. The SMOTE+ENN algorithm was applied to oversample the minority class and obtain an augmented training subset used to train our RF classifier. The parameters for the RF classifier were selected using tenfold cross-validation to attain optimal model performance. The classifier was evaluated on the testing data, and metrics including F1-score, Recall, Precision, balanced accuracy, AUROC, and AUPRC were calculated to evaluate the classifier's overall performance. RF classifier was then used to obtain the top 100 important features (protein structures). The proteins were associated with pathways annotated in Reactome. The overrepresented pathways in the top 100 proteins via enrichment analysis were identified. The enriched pathways for NR-AhR assay were analyzed as a case study to evaluate the mechanistic understanding of its toxicity

Data Challenge. Two of the winning teams used an RF model to achieve the best performance in predicting compound toxicity against the NR- androgen receptor (NR-AR), NR-aromatase, SR-p53 [20], and NR estrogen receptor alpha ligand-binding domain (NR-ER-LBD) assay [19]. The area under the receiver operating characteristic curve (AUROC) and the winning teams' balanced accuracy (BA) scores was provided as evaluation metrics during the Tox21 Data Challenge.

Banerjee et al. [32] highlighted the importance of using sampling methods when training a classifier on an imbalanced chemical data set, such as the Tox21 dataset. This is important because non-sampling methods on imbalanced data result in poor recall performance due to the classifier favoring the majority inactive class [32]. Another study on a similar methodology employed an RF classifier with different resampling techniques and showed that the RF with SMOTE+ENN classifier performed the best on the Tox21 data. [31].

Therefore, in this study, we implemented the RF with SMOTE+ENN algorithm to train our model for feature selection and handle class imbalance. The SMOTE+ENN was applied to resample the training data, and the RF classifier was then fit to the training data. The trained model was used to make predictions on the test data and was evaluated using classification performance metrics. This approach was applied to each of the twelve assay datasets. Table 2 reports the model evaluation metrics for the RF classification model for 12 qHTS assays. The reported values varied depending

Table 1: **Compound activity measurement distribution and imbalance ratio (IR) for the 12 preprocessed train and test assay datasets from Tox21.** The highest and lowest IRs for the training and test sets are in bold.

In vitro qHTS assay ID	Target/Assay	Total number of compounds	Training set			Test set		
			Inactive	Active	IR	Inactive	Active	IR
NR-AhR	aryl hydrocarbon receptor	7103	5777	734	7.87	521	71	7.34
NR-ER-LBD	estrogen receptor (luciferase assay)	7509	6643	282	23.56	564	20	28.20
NR-ER	estrogen receptor	6630	5474	651	8.41	456	49	9.31
NR-Aromatase	aromatase	6286	5496	274	20.06	479	37	12.94
NR-PPAR-γ	peroxisome proliferatoractivated receptor γ	7039	6283	167	**37.62**	559	30	18.63
NR-AR	androgen receptor	7783	6958	252	27.61	561	12	46.75
NR-AR-LBD	androgen receptor (luciferase assay)	7298	6521	211	30.90	558	8	**69.75**
SR-MMP	mitochondrial membrane potential	6316	4899	888	**5.52**	474	55	8.62
SR-ARE	nuclear factor (erythroid- derived 2)-like 2 antioxidant responsive element	6339	4919	881	5.58	450	89	**5.06**
SR-ATAD5	genotoxicity indicated by ATAD5	7646	6787	256	26.51	569	34	16.73
SR-p53	DNA damage p53-pathway	7358	6351	409	15.53	560	38	14.74
SR-HSE	heat shock factor response element	7040	6144	305	20.14	574	17	33.76

on metrics and assays. Overall the modeling approach achieved decent performance with results achieving strong performance measured by AUROC equal to or higher than 0.7, except for NR-AR-LBD, which had very few active labels, as shown by its high IR ratio of 30.9 in the training set. The AUROC has the highest value of 0.92 for the SR-MMP, with the lowest IR ratio of 5.52 in the training set. This is not surprising as the more active labels, the better the model can learn to discriminate between the active and inactive classes. The SR-MMP assay also had the highest F1-score of 0.488 and the highest recall and high precision. The area under the precision-recall curve (AUPRC) score was also high for SR-MMP, with a score of 0.60, signifying that the model could handle the positive samples correctly.

We compare the prediction performance of our approach with those employed in the Dmlab [20] and Microsomes [19], which are the two winning RF models in the Tox21 Data Challenge. We also compare our results with the best RF-based classifier (SMN model) employed in [31]. Since AUROC and BA are the only accuracy measures in the Tox21 Data Challenge, we only evaluated these two metrics in our results. Table 3 shows the metrics of the other studies with our proposed model. On average, our model performed close to the dmlab, Microsomes, and SMN model approaches regarding AUCROC and BA across most of the twelve assays.

Despite the metric comparisons with other studies, it is not an apple-to-apple comparison since our algorithm is used for feature selection, while the compared RF-based models are employed for prediction. Furthermore, such a comparison requires using the same data pre-processing methodology and feature descriptors when training a model, which was not done herein.

Table 2: **Metrics for evaluating the performance of RF+SMOTEENN with twelve Tox21 assay test datasets.** Average of the metric is computed across all 12 assays.

Assays	F_1	Precision	Recall	AUROC	AUPRC	BA
NR-AhR	0.47	0.31	0.90	0.90	0.55	0.82
NR-ER-LBD	0.34	0.33	0.35	0.81	0.27	0.66
NR-ER	0.42	0.30	0.69	0.81	0.41	0.76
NR-Aromatase	0.32	0.25	0.43	0.80	0.28	0.67
NR-PPAR-γ	0.29	0.31	0.27	0.74	0.24	0.62
NR-AR	0.26	0.27	0.25	0.71	0.20	0.62
NR-AR-LBD	0.00	0.00	0.00	0.62	0.04	0.49
SR-MMP	0.49	0.33	0.93	0.92	0.60	0.86
SR-ARE	0.42	0.31	0.70	0.76	0.40	0.69
SR-ATAD5	0.33	0.28	0.38	0.74	0.23	0.66
SR-p53	0.23	0.16	0.45	0.83	0.20	0.64
SR-HSE	0.29	0.31	0.27	0.76	0.24	0.62
Average	0.32	0.26	0.47	0.78	0.31	0.68

Table 3: **Comparison of performance metrics AUROC and BA between this study and other Tox21 data studies who utilized RF-based models.**

Assays	AUCROC our method	AUCROC SMN [31]	AUCROC Dmlab [20]	AUCROC Microsome [19]	BA our Method	BA SMN [31]	BA Dmlab [20]	BA Microsome [19]
NR-AhR	0.90	0.92	0.78	0.90	0.82	0.82	0.56	0.70
NR-ER-LBD	0.81	0.82	0.77	0.83	0.66	0.69	0.59	0.55
NR-ER	0.81	0.86	0.77	0.78	0.76	0.79	0.66	0.62
NR-Aromatase	0.78	0.85	0.84	N/A	0.67	0.73	0.56	N/A
NR-PPAR-γ	0.74	0.77	0.83	0.72	0.62	0.68	0.55	N/A
NR-AR	0.71	0.68	0.83	N/A	0.62	0.64	0.61	N/A
NR-AR-LBD	0.62	0.80	0.82	N/A	0.49	0.55	0.49	N/A
SR-MMP	0.92	0.93	0.95	N/A	0.86	0.85	0.69	N/A
SR-ARE	0.76	0.89	0.77	0.80	0.69	0.85	0.52	0.61
SR-ATAD5	0.74	0.81	0.80	0.81	0.66	0.70	0.61	0.54
SR-p53	0.83	0.88	0.88	0.83	0.64	0.68	0.58	0.52
SR-HSE	0.76	0.85	0.86	N/A	0.62	0.65	0.49	N/A

2.3. Enrichment analysis

The top 100 protein features determined to be most important by the RF algorithm for each Tox21 endpoint were subjected to enrichment analysis in the context of protein pathways. The number of enriched entities for the 12 toxicity endpoints is summarized in Table 4 with significance being determined by a p-value <= 0.05. We only analyzed the enriched pathways and proteins for NR-AhR, but the same approach can be applied for the other eleven assays. The exact proteins and pathways highlighted for the NR-AhR activation assay are provided in Table 5, including the total number of proteins in the pathway and the calculated p-value.

2.4. Case study for NR-AhR

We selected the NR-AhR as a toxicity case study to evaluate our algorithm's performance by analyzing the selected proteins and enriched pathways implicated in NR-AhR related toxicity. The NR-AhR is a helix-loop-helix ligand-activated transcription factor that mediates the carcinogenic and other toxic effects of a variety of environmental pollutants, such as polycyclic aromatic hydrocarbons (PAHs) and polyhalogenated aromatic hydrocarbons [45]. The latter class of compounds includes halogenated dibenzo-p-dioxins, also known as dioxins. One of these compounds is 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD), which is the most commonly used environmental agent for studying NR-AhR-mediated biochemical and toxic responses because of its high affinity for the NR-AhR [46]. It also causes a wide range of toxic effects, including immunosuppression

Table 4: **Number of enriched pathways for each Tox21 assay.**

Assays	Total number of enriched pathways
NR-AhR	34
NR-AR-LBD	53
NR-Aromatase	33
NR-ER	23
NR-ER-LBD	32
NR-PPAR-γ	13
NR-AR	27
SR-ARE	46
SR-ATAD5	59
SR-HSE	49
SR-MMP	29
SR-p53	42

and tumor promotion [47,48]. Most of the effects caused by TCDD and other PAHs are mediated by NR-AhR [46]. Previous studies have suggested that NR-AhR signaling elicits numerous critical biological processes, including the modification of the cell cycle, cell proliferation, immune responses, and tumorgenesis [49–52].

Based on our enrichment analysis, the NR-AhR signaling pathway is involved in 34 different toxic signaling pathways. We analyzed a handful of them and their corresponding proteins to illustrate the validity of our toxicity analysis. One potential pathway that our analysis suggested is the crosstalk between NR-AhR and DNA damage pathways, including the DNA double-strand break ends and the DNA Double-Strand Break Response. Damage to DNA is well known to contribute to the development of cancer. One form of DNA damage is DNA double-strand breaks (DSBs), which are caused by ionizing radiation and chemical exposure [53]. Our analysis selected breast cancer type 1 susceptibility protein (BRCA1), a protein in the DNA damage pathways, as an important feature in predicting NR-AhR toxicity. BRCA1 is an E3 ubiquitin-protein ligase that plays a central role in DNA repair and is a tumor suppressor gene. Many studies have associated the BRCA-1 gene in cell-cycle control, DNA repair, and human breast cancer [54–56]. A study has shown that TCDD exposure causes cancer by affecting the repair of DSBs mediated by the NR-AhR signaling pathway [57], supporting our result. Ratten et al. showed that TCDD could alter the expression of the tumor suppressor gene BRCA1 by downregulating the BRCA1 promoter activity [58]. A similar study in the literature found that BRCA1 is inhibited by benzo [a]pyrene (BaP), a polycyclic aromatic hydrocarbon ligand known to have a high affinity to NR-AhR [59]. Our analysis also selected the ATM-mediated phosphorylation pathway and G2/M pathway to be involved in NR-AhR toxicity. The implication of these pathways in NR-AhR related toxicity is supported by the work of Tzeh K Foo et al. where they indicated the importance of the ATM-mediated phosphorylation pathway for the DNA repair function of BRCA1 and the maintenance of the G2-M checkpoint pathway; disruptions to BRCA1 mechanisms can have effects on BRCA1-related cancers [60].

Our analysis suggested a crosstalk between the NR-AhR, nuclear factor κB (NF-κB) signaling pathway, and the TNF receptor-associated factor 6 (TRAF6). NF-κB is one of the essential transcription factors for osteoclastogenesis [61] and TRAF6 is known to be an essential factor for osteoclastogenesis [62]. A study demonstrated that BaP, which constitutes important carcinogenic components of cigarette smoke, smog, and some over-cooked foods [63], inhibits osteoclastogenesis by activating the NF-κB pathway [64]. Once the NF-κB pathway is activated, the TRAF6 gets recruited together with UBE2N/UBC13, which are ubiquitin-conjugating enzymes, catalyzing the synthesis of Lys-63 linked polyubiquitin chain [65]. This polyubiquitination activates the transforming growth factor β activated kinase-1 (TAK1) complex, which phosphorylates IκB kinase

Table 5: **Important mechanistic pathways and their proteins for the toxicity endpoint of NR-AhR.**

Pathway (Reactome ID)	Total proteins	Selected proteins (UniProt IDs)	p-value
G2/M DNA damage checkpoint	78	O76064,P61088 ,P38398,O96028,P27694	4.64e-05
Processing of DNA double-strand break ends	81	O76064,P61088 ,P38398,O96028,P27694	5.55e-5
Formation of Incision Complex in GG-NER	43	P61088,Q9Y6X2 ,P62877,P27694	6.52e-5
Nonhomologous End-Joining (NHEJ)	52	O76064,P61088 ,P38398,O96028	1.36e-6
Recruitment and ATM-mediated phosphorylation of repair and signaling proteins at DNA double strand breaks	59	O76064,P61088 ,P38398,O96028	2.22e-6
TRAF6 mediated NF-kB activation	24	Q12933,Q9Y4K3,O95786	2.35e-6
DNA Double Strand Break Response	60	O76064,P61088 ,P38398,O96028	2.36e-6
TRAF6 mediated IRF7 activation	28	Q12933,Q9Y4K3,O95786	3.73e-6
Neurofascin interactions	7	Q92823,Q12860	5.28e-6
DDX58/IFIH1-mediated induction of interferon-alpha/beta	77	Q12933,Q96EQ8 ,Q9Y4K3,O95786	6.04e-6
RUNX3 regulates YAP1-mediated transcription	8	P28347,Q15561	7.01e-6
SUMOylation of transcription cofactors	42	Q99496,Q96PU4,Q9Y6X2	1.22e-5
IRAK1 recruits IKK complex	14	Q9Y4K3,P61088	2.21e-5
IRAK1 recruits IKK complex upon TLR7/8 or 9 stimulation	14	Q9Y4K3,P61088	2.21e-5
YAP1- and WWTR1 (TAZ)-stimulated gene expression	14	P28347,Q15561	2.21e-5
TRAF6 mediated IRF7 activation in TLR7/8 or 9 signaling	14	Q9Y4K3,P61088	2.21e-5
TICAM1, RIP1-mediated IKK complex recruitment	19	Q9Y4K3,P61088	4.05e-5
Signal transduction by L1	20	O14786,P13591	4.48e-5
Regulation of FZD by ubiquitination	21	O75581,O75473	4.92e-5
IKK complex recruitment mediated by RIP1	22	Q9Y4K3,P61088	5.39e-5
JNK (c-Jun kinases) phosphorylation and activation mediated by activated human TAK1	22	Q9Y4K3,P61088	5.39e-5
activated TAK1 mediates p38 MAPK activation	23	Q9Y4K3,P61088	5.87e-5
Recognition of DNA damage by PCNA-containing replication complex	31	P62877,P27694	1.03e-4
TAK1 activates NFkB by phosphorylation and activation of IKKs complex	32	Q9Y4K3,P61088	1.10e-4
DNA strand elongation	32	Q9Y248,P27694	1.10e-4
Sialic acid metabolism	33	P16278,Q8TBE9	1.16e-4
Transcriptional Regulation by E2F6	34	Q99496,P38398	1.23e-4
Negative regulators of DDX58/IFIH1 signaling	34	Q96EQ8,O95786	1.23e-4
NOD1/2 Signaling Pathway	35	Q9Y4K3,P61088	1.30e-4
RUNX1 interacts with co-factors whose precise effect on RUNX1 targets is not known	36	Q99496,Q86SE9	1.37e-4
HDR through Single Strand Annealing (SSA)	37	P38398,P27694	1.44e-4
Ovarian tumor domain proteases	38	Q9Y4K3,O95786	1.51e-4
Presynaptic phase of homologous DNA pairing and strand exchange	39	P38398,P27694	1.58e-4
Formation of Fibrin Clot (Clotting Cascade)	39	Q9UNN8,P13224	1.58e-4

(IKK), activating the IKK complex. This demonstrates the link between TRAF6 and IKK in the NF-κB pathway. This study supports our analysis since the "TAK1 activates NF-κB by phosphorylation and activation of IKKs complex" pathway and the TRAF6, and UBE2N/UBC13 proteins were selected as important features for our algorithm.

Additionally, our algorithm selected the p38–mitogen-activated protein kinase (MAPK) as a pathway that interacts with the NR-AhR toxicity pathway. A study in the literature showed that TCDD could activate the p38–MAPK pathway, leading to crosstalk between p38-MAPK signaling and NR-AhR action. It showed that the NR-AhR and p38–MAPK-dependent pathway induces the expression and activity of c-Jun, a proto-oncogene [66], highlighting a novel mode of interaction between the NR-AhR and p38–MAPK pathway in mediating the toxicity and carcinogenicity of the NR-AhR ligand, TCDD. The NF-κB pathway our algorithm selected is connected to the p38-MAPK pathway because TAK1, involved in the (NF-κB) signaling pathway, phosphorylates members of the MKK family, which in turn phosphorylates and activates p38 kinases [67]. Our results are supported as the TRAF6 and UBE2N/UBC13 proteins, associated with the NF-κB pathway, are also linked with the MAPK pathway.

Our computational pipeline predicts some pathways and proteins associated with NR-AhR toxicity. Even though experimental studies from the literature support these predictions, further experimental studies are warranted to confirm their association with NR-AhR toxicity.

3. Discussion

While several studies in the literature have used chemical properties as feature descriptors to predict toxicity [16–21], no one has yet considered using protein descriptors in their model setting. In this study, we looked at toxicity not only from a single protein and target binding but from a pathway perspective, as it is known that when a compound binds to a target, it can trigger multiple pathways or cause a cascade of signaling events. Looking at the broad spectrum of possible pathways enriched for a specific toxicity endpoint enables the investigation of multiple proteins and target binding sites leading to the toxicity. Within the NR and SR pathways, the activity of NR-AhR and SR-MMP were two of the top-performing twelve toxicity endpoints with AUROCs of 0.90 and 0.92, respectively, since these assays had a more balanced activity class distribution in the training set compared to the other assays. Further, from our NR-AhR toxicity analysis, we can see that our algorithm selected proteins and their corresponding pathways that have been shown in the literature to lead to toxicity. Not only that, but these selected proteins have shown to overlap across multiple pathways, showing that toxicity is not a single target binding mechanism but an interconnected network involving many pathways. The overlap in proteins and pathways indicates that our algorithm can potentially provide novel insight into understanding toxicity from a proteomic and pathway perspective. However, further studies are warranted to elucidate the importance of these pathways in NR-AhR toxicity and the other toxicity endpoints. Our novel approach to understanding toxicity at a pathway level would be important to investigate toxicity targets and pathways that have not been further explored and are interconnected with well-known toxicity pathways.

4. Materials and Methods

4.1. Tox21 datasets

The Tox21 compound structures and activity measurements for 12 different qHTS assays were extracted from the Tox21 Data Challenge [11]. We used the training set, the leaderboard, and the test set, consisting of 11,764, 296, and 647 compounds, respectively. We joined the leaderboard set with the training set to form our training data and used the testing data for model evaluation. The 12 qHTS *in vitro* assays consisted of two categories, seven of which were part of the Nuclear Receptors (NR) and five Stress Response (SR) pathways. The NR assays included the androgen receptor (AR), androgen receptor ligand-binding domain (AR-LBD), aryl hydrocarbon receptor (AhR), aromatase, estrogen receptor (ER), estrogen receptor luciferase assay (ER-LBD), and peroxisome proliferator-activated receptor γ (PPAR-γ). The SR assays included the antioxidant response element (ARE), heat shock factor response element (HSE), p53, mitochondrial

membrane potential (MMP), and ATPase Family AAA Domain Containing 5 (ATAD5). In each assay, the activity of a compound was assigned a class label, where a label of 1 signified the compound was active, and a label of 0 signified the compound was inactive. There were duplicates and inconsistent activity labels for the compounds across the twelve assays.

4.2. UniProt

A human protein library of 19,582 sequences was extracted from UniProt [68], of which 4,966 had at least one solved X-ray crystallography structure in the Protein Data Bank (PDB) [69], including 4,641 with one chain, 298 with two chains, and 27 with three chains, culminating in 5,316 total human proteins after removing two non-viable structures. These protein structures were chosen by matching their UniProt IDs to all corresponding structures in the PDB, filtering for chains with the most significant sequence coverage to the whole sequence, then selecting the chain with the lowest resolution. Proteins were mapped to 2,219 pathways in Reactome [70] with an average of 35.1 structures per pathway.

4.3. Protein-compound interaction scoring protocol

The Computational Analysis of Novel Drug Opportunities (CANDO) drug discovery platform was used to generate protein interaction signatures for each compound [36–43]. Interaction scores between all compounds in the Tox21 dataset and each structure in the human protein library were computed using a rapid in-house bioanalytical docking protocol known as BANDOCK [71]. First, binding sites were predicted for each protein using COACH [72], a consensus method combining structural and sequence similarity to proteins in the PDB [69], with each prediction having a confidence score (BScore) as well as an associated co-crystallized ligand. Depending on the scoring protocol used, the output interaction score for the compound and protein considers both the BScore and the molecular fingerprint similarity between the compound and the associated ligand (CScore). In this study, the scores were determined by multiplying the BScore by the CScore, which itself was the Sorenson-Dice coefficient between the ECFP4 fingerprints (computed using RDKit [73]) of the query compound and binding site ligand. Since the COACH algorithm outputs multiple binding sites (and therefore associated ligands) for each protein, the maximum value of the product of BScore and CScore is the chosen interaction score; these serve as a proxy for binding strength and (or) likelihood.

4.4. Study design

A general overview of the workflow of this study is illustrated in Figure 1. Our study design consisted of data pre-processing, feature generation, resampling, feature selection, and enrichment analysis. Data pre-processing and feature generation was implemented on both the training and testing datasets. Resampling using SMOTE+ENN and feature selection using the RF algorithm were applied separately to each of the twelve assays in the training dataset. This was followed by model optimization using a repeated stratified three-fold cross-validation and model evaluation to assess the prediction results of the SMOTE+ENN with the RF algorithm. Enrichment analysis was conducted on the ranked list of top 100 protein descriptors selected by the algorithm to analyze the Tox21 toxicity assays at a proteomic and pathway level. All former steps are described in detail in the subsequent sections.

4.4.1. Data pre-processing and feature generation

Compound structures were extracted from the Tox21 Data challenge as SMILES files. Compound structure standardization and normalization were implemented using the RDKit MolVS library [74]. We applied a fragmentation step as described in [31], where SMILES with salt fragments, varied resonance structures, and tautomers were removed,

and the valid SMILES were canonicalized by normalizing inconsistent chemical groups. To handle the ambiguous compounds with duplicate activities (0 or 1) for a particular target, we removed compounds with an equal number of active and inactive labels. Compounds with an unequal number of active and inactive labels were retained where the most frequent activity label was selected.

Following normalization and merging of the compounds, the resulting SMILES were used to generate a compound-protein interaction score matrix using the CANDO platform, and the scoring protocol BANDOCK [71]. This provided the feature descriptor input to the model. The matrix consisted of 7,808 compounds for the training set and 628 compounds testing sets. The generated matrix consisted of 8,385 proteins as feature descriptors. The feature space was filtered for interactions with only solved protein structures, leading to 5316 proteins. The elimination of features reduced the computation requirement and the effect of the curse of dimensionality [75]. The resulting CANDO matrix and the activity measurements of the compounds for each of the twelve different qHTS assays were merged, generating twelve training datasets, one per assay. For each assay, only compounds labeled active or inactive are retained.

In order to understand the activity measurement distribution for each of the twelve assays, the number of active and inactive compounds for each of the twelve assays was computed, along with the imbalanced ratio (IR), which is the ratio of majority class (inactive non-toxic compounds) to the minority class (active toxic compounds).

4.4.2. Data resampling for predictive model

As mentioned before, the sparsity in the Tox21 data results in a model that favors the predominant inactive class and misclassifies the minority active class, while the minority class is usually the class of interest. We balanced the class distribution for each training set by implementing the SMOTE+ENN method to address this challenge. SMOTE+ENN combines both oversampling (using SMOTE) and undersampling (using ENN). SMOTE [76] synthesizes samples in the minority class by linear interpolation to increase the number of instances in the minority class. ENN reduces the number of instances in the majority class by removing noisy samples from the majority class, which is inconsistent with its k-nearest neighbors [77]. The SMOTE+ENN algorithm has been shown to deliver promising results when applied to imbalanced data sets with a small number of positive instances, including the Tox21 data [31,78].

4.4.3. Random Forest

Recent work in computational biology has observed increased use of RF, owing to its unique advantages as being non-parametric, interpretable, and high prediction accuracy for modeling in cheminformatics [79,80]. Previous studies using Tox21 have shown the competency of RF in predicting compound toxicity [19–21].

RF is an ensemble learning method, combining decision trees as base learners for increased performance [27]. Each tree is trained by different bootstrap samples having the same size as the training set. By selecting a random subset of features at each node in the tree construction, RF introduces randomization and increased diversity in the ensemble, reducing the variance of the base learners [81]. The construction of RF is described in the following steps:

1) Draw a bootstrap sample: we randomly sample N compounds with replacement from the original data set.
2) Create maximum decision trees: we construct a decision tree for each bootstrap sample by randomly sampling a subset of features at each node and choosing the best split among those features.
3) Construct a forest by repeating Step 1 and Step 2 for N trees.
4) Predicting the outcome: from the built forest, the prediction is obtained by aggregating the predictions of the N trees (i.e., majority votes for classification and average for regression tasks).

Given that the Tox21 dataset is highly dimensional with a large feature space, this can lead to model overfitting. However, the RF algorithm is less susceptible to model overfitting due to the utilization of ensemble strategies and random sampling, and therefore we selected it as our model in this study [30].

4.4.4. Feature Selection and Variable Important Measure and Using Random Forest

The high-dimensional feature space of many tasks in bioinformatics has created needs for feature selection techniques [82]. Feature importance refers to the process of finding parts of the input feature space relevant to a prediction task. The main advantage of using RF compared to other machine learning algorithms is that it directly performs feature selection during model construction.

A commonly used measure to evaluate the importance of each predictor variable from an RF classifier is the Gini importance measure, also known as the mean decrease of impurity (MDI). As a simple feature importance to rank features, RF Gini feature importance has gained popularity in bioinformatics applications [82].

Gini importance is derived from the Gini index [27], which is splitting criteria used to determine which feature to split on during tree training. It measures the level of impurity of the samples assigned to a node based on a split. Gini index ranges between 0 and 1, where 0 means all instances belong to one class, and one means that the instances are distributed randomly across the classes. The smaller the Gini index, the purer the node. The Gini index is measured as follows:

$$Gini(P) = \sum_{i=1}^{n} p_i(1 - p_i) = 1 - \sum_{i=1}^{n} p_i^2, \tag{1}$$

where $P = (p_1, p_2, \ldots, p_n)$, p_i is the probability of the class i at a certain node, and n is the number of classes [80].

A feature's Gini importance value is computed as the sum of the Gini indices weighted by the probability of reaching that node averaged among all trees in the forest. A high Gini importance means that the feature is likely to be informative [83].

4.4.5. Model Training and hyperparameters

Following data pre-processing, SMOTE+ENN was applied to the training data to oversample the minority class and obtain an augmented training set to train the RF. The RF was trained with repeated stratified k-fold cross-validation (CV) to optimize the model's hyper-parameters.

Stratified CV is a variation of the k-fold CV, which ensures the folds are generated by maintaining the percentage of samples for each class as in the original data set. During CV, models are trained using the same hyper-parameters and tested on a different subset of data. In the process, the training data is divided into K subsamples, keeping the distribution of the active and inactive classes balanced. K-1 subsamples are used for each fold to train the model, and the last subsample is used for testing. This process is performed until all the folds are selected for testing. CV provides a way to improve the estimated performance of our ML model, as it gives a more reliable model assessment score than performing cross-validation only once, which could be misleading when having a random split.

A 10-fold CV with three repetitions was performed in the evaluation phase to smooth the variability and ensure a robust model performance assessment. The RF model parameters were selected using the randomized search method. This included the maximum depth of the tree and the number of trees in the forest. Other model tuning parameters were set to their respective default values in scikit-learn [84]. After the RF model is optimized, the Gini importance is calculated to extract the top 100 weighted feature descriptors.

The model training is applied to each of the twelve data sets, generating an optimized feature selection model with the optimal feature descriptors per assay. The model

implementation is done using the scikit-learn package [84] and the resampling technique from the Imbalanced-learn package in Python [85].

4.4.6. Performance Evaluation Metrics

We reported both the area under the precision-recall curve (AUPRC) and the area under the receiver-operating characteristic curve (AUROC) given the class imbalance for the task [86]. We do not report accuracy (the proportion of correct predictions to the total number of compounds), as a high accuracy does not translate into a model correctly predicting the rare inactive class and can therefore be misleading for evaluating our model performance [87].

The decision made by the RF classifier can be represented in a 2-by-2 confusion matrix, given that we have a binary classifier. The confusion matrix has four categories: True positives (TP), the number of active compounds that are correctly labeled. False positives, which is the number of incorrectly labeled inactive chemicals. True negatives refer to the number of correctly labeled inactive compounds, and lastly, false negatives (FN) correspond to the number of incorrectly labeled inactive compounds. We utilized imbalanced classification metrics derived from the confusion matrix, including recall, precision, F1-score, and balanced accuracy (BA). Recall is a measure of the accuracy of the active minority class. It gives the proportion of actual positives identified correctly by the model, i.e., the number of correctly labeled active compounds out of the actual active compounds. Precision (also called positive predictive value) gives the proportion of positive identifications that were actually correct, i.e., the number of correctly labeled active compounds out of the compounds predicted as active. F1 score is the harmonic mean of precision and recall. Balanced accuracy (BA) is the average of recall and precision. These evaluation metrics were implemented using the scikit-learn package in python [84]. The formulas of the evaluation metrics are as follows:

$$\text{Recall= Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$F_1\text{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

$$\text{Balanced accuracy} = \frac{Sensitivity + Specificity}{2} \tag{5}$$

4.4.7. Enrichment analysis

The top 100 most essential proteins for predicting each toxicity endpoint for the compounds in Tox21 are the inputs to an enrichment analysis protocol. This enrichment analysis identifies significantly overrepresented pathways based on their mappings to proteins obtained via Reactome. Essentially, it utilizes the hypergeometric distribution to determine the probability that a pathway is significantly overrepresented based on the number of proteins in the top feature set associated with the pathway relative to the whole human proteome and the total number of proteins associated with each other pathway. A p-value is computed using the probability mass function with the number of top features associated with the pathway serving as input. The human proteome used in this study includes 19,582 proteins from UniProt, of which 4,966 have at least one solved X-ray crystallography structure available in the PDB and are mapped to at least one biological pathway.

5. Conclusions

To the best of our knowledge, this is the first computational pipeline that utilizes protein descriptors to extract the important protein structures from the twelve toxicity

endpoints in the Tox21 dataset to evaluate compound toxicity. We employed a combination of the CANDO drug discovery platform, data balancing, feature selection, and enrichment analysis to understand compound toxicity behavior at the protein pathway level. We expect this computational pipeline will provide a novel perspective in evaluating environmental compounds and allow researchers and the pharmaceutical industry to explore the underlying proteomic mechanisms that not only induce toxicity but also potentially assist in developing novel therapeutics to mediate toxicity targets.

Author Contributions: LM, SM, WM, ZK, MG, RS were responsible for conceptualization and investigation of the study. LM was responsible for methodology, data pre-processing, modeling, and original draft preparation and visualizations. LM, SM, WM, ZF, MG, RS helped with manuscript development and editing. MG and RS provided general oversight and mentorship.

Funding: This work was supported in part by the National Institutes of Health Director's Pioneer Award (DP1OD006779), National Institutes of Health Clinical and Translational Sciences Award (UL1TR001412), NIH NLM T15 Award (T15LM012495), NCATS ASPIRE Design Challenge Award, NCATS ASPIRE Reduction-to-Practice Award, and startup funds from the Department of Biomedical Informatics at the University at Buffalo.

Data Availability Statement: All the data and code used to pre-process the data and build the machine learnings model is available at GitHub under https://github.com/lamawmouk/Tox21_FeatureSelection_SMOTEENN_RF

Acknowledgments: The authors would like to acknowledge the support provided by the Center for Computational Research at the University at Buffalo, as well as members of the Samudrala Computational Biology Group.

Conflicts of Interest: The authors have declared that no competing interests exist. 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; nor in the decision to publish the result.

References

1. Binetti, R.; Costamagna, F.M.; Marcello, I. Exponential growth of new chemicals and evolution of information relevant to risk control. *Annali-Istituto Superiore di Sanità* **2008**, *44*, 13.
2. Ekins, S.; Nikolsky, Y.; Nikolskaya, T. Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends in pharmacological sciences* **2005**, *26*, 202–209.
3. Kavlock, R.; Dix, D. Computational toxicology as implemented by the US EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. *Journal of Toxicology and Environmental Health, Part B* **2010**, *13*, 197–217.
4. Inglese, J.; Auld, D.S.; Jadhav, A.; Johnson, R.L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C.P. Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proceedings of the National Academy of Sciences* **2006**, *103*, 11473–11478.
5. Shukla, S.J.; Huang, R.; Austin, C.P.; Xia, M. The future of toxicity testing: a focus on in vitro methods using a quantitative high-throughput screening platform. *Drug discovery today* **2010**, *15*, 997–1007.
6. Sun, H.; Xia, M.; Austin, C.P.; Huang, R. Paradigm shift in toxicity testing and modeling. *The AAPS journal* **2012**, *14*, 473–480.
7. Krewski, D.; Jr., D.A.; Andersen, M.; Anderson, H.; III, J.C.B.; Boekelheide, K.; Brent, R.; Charnley, G.; Cheung, V.G.; Jr., S.G.; Kelsey, K.T.; Kerkvliet, N.I.; Li, A.A.; McCray, L.; Meyer, O.; Patterson, R.D.; Pennie, W.; Scala, R.A.; Solomon, G.M.; Stephens, M.; Yager, J.; Zeise, L.; of Committee on Toxicity Testing, S.; of Environmental Agents, A. Toxicity testing in the 21st century: a vision and a strategy. *Journal of Toxicology and Environmental Health, Part B* **2010**, *13*, 51–138, [https://doi.org/10.1080/10937404.2010.483176]. PMID: 20574894.
8. Kavlock, R.J.; Austin, C.P.; Tice, R. Toxicity testing in the 21st century: implications for human health risk assessment. *Risk analysis: an official publication of the Society for Risk Analysis* **2009**, *29*, 485.

9. Tice, R.R.; Austin, C.P.; Kavlock, R.J.; Bucher, J.R. Improving the human hazard characterization of chemicals: a Tox21 update. *Environmental health perspectives* **2013**, *121*, 756–765.
10. Collins, F.S.; Gray, G.M.; Bucher, J.R. Transforming environmental health protection. *Science (New York, NY)* **2008**, *319*, 906.
11. Centre, I.I. The Tox21 data Challenge, 2014.
12. Chen, S.; Hsieh, J.H.; Huang, R.; Sakamuru, S.; Hsin, L.Y.; Xia, M.; Shockley, K.R.; Auerbach, S.; Kanaya, N.; Lu, H.; others. Cell-based high-throughput screening for aromatase inhibitors in the Tox21 10K library. *Toxicological Sciences* **2015**, *147*, 446–457.
13. Attene-Ramos, M.S.; Huang, R.; Michael, S.; Witt, K.L.; Richard, A.; Tice, R.R.; Simeonov, A.; Austin, C.P.; Xia, M. Profiling of the Tox21 chemical collection for mitochondrial function to identify compounds that acutely decrease mitochondrial membrane potential. *Environmental health perspectives* **2015**, *123*, 49–56.
14. Huang, R.; Sakamuru, S.; Martin, M.T.; Reif, D.M.; Judson, R.S.; Houck, K.A.; Casey, W.; Hsieh, J.H.; Shockley, K.R.; Ceger, P.; others. Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Scientific reports* **2014**, *4*, 1–9.
15. Paul-Friedman, K.; Martin, M.; Crofton, K.M.; Hsu, C.W.; Sakamuru, S.; Zhao, J.; Xia, M.; Huang, R.; Stavreva, D.A.; Soni, V.; others. Limited chemical structural diversity found to modulate thyroid hormone receptor in the Tox21 chemical library. *Environmental health perspectives* **2019**, *127*, 097009.
16. Huang, R.; Xia, M.; Nguyen, D.T.; Zhao, T.; Sakamuru, S.; Zhao, J.; Shahane, S.A.; Rossoshek, A.; Simeonov, A. Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science* **2016**, *3*, 85.
17. Stefaniak, F. Prediction of compounds activity in nuclear receptor signaling and stress pathway assays using machine learning algorithms and low-dimensional molecular descriptors. *Frontiers in Environmental Science* **2015**, *3*, 77.
18. Judson, R.S.; Magpantay, F.M.; Chickarmane, V.; Haskell, C.; Tania, N.; Taylor, J.; Xia, M.; Huang, R.; Rotroff, D.M.; Filer, D.L.; others. Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor. *Toxicological Sciences* **2015**, *148*, 137–154.
19. Uesawa, Y. Rigorous selection of random forest models for identifying compounds that activate toxicity-related pathways. *Frontiers in Environmental Science* **2016**, *4*, 9.
20. Barta, G. Identifying biological pathway interrupting toxins using multi-tree ensembles. *Frontiers in Environmental Science* **2016**, *4*, 52.
21. Koutsoukas, A.; St Amand, J.; Mishra, M.; Huan, J. Predictive toxicology: modeling chemical induced toxicological response combining circular fingerprints with random forest and support vector machine. *Frontiers in Environmental Science* **2016**, *4*, 11.
22. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science* **2016**, *3*, 80.
23. Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L. *Feature Extraction: Foundations and Applications*; Studies in Fuzziness and Soft Computing, Springer Berlin Heidelberg, 2008.
24. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering* **2014**, *40*, 16–28.
25. Jaiswal, J.K.; Samikannu, R. Application of random forest algorithm on feature subset selection and classification and regression. 2017 World Congress on Computing and Communication Technologies (WCCCT). IEEE, 2017, pp. 65–68.
26. Remeseiro, B.; Bolon-Canedo, V. A review of feature selection methods in medical applications. *Computers in biology and medicine* **2019**, *112*, 103375.
27. Breiman, L. Random Forests. *Machine learning* **2001**, *45*, 5–32.
28. Chen, R.C.; Dewi, C.; Huang, S.W.; Caraka, R.E. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data* **2020**, *7*, 1–26.
29. Díaz-Uriarte, R.; De Andres, S.A. Gene selection and classification of microarray data using random forest. *BMC bioinformatics* **2006**, *7*, 1–13.
30. Rogers, J.; Gunn, S. Identifying feature relevance using a random forest. International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection". Springer, 2005, pp. 173–184.

31. Idakwo, G.; Thangapandian, S.; Luttrell, J.; Li, Y.; Wang, N.; Zhou, Z.; Hong, H.; Yang, B.; Zhang, C.; Gong, P. Structure–activity relationship-based chemical classification of highly imbalanced Tox21 datasets. *Journal of cheminformatics* **2020**, *12*, 1–19.
32. Banerjee, P.; Dehnbostel, F.O.; Preissner, R. Prediction is a balancing act: importance of sampling methods to balance sensitivity and specificity of predictive models based on imbalanced chemical data sets. *Frontiers in chemistry* **2018**, *6*, 362.
33. Ring, C.; Sipes, N.S.; Hsieh, J.H.; Carberry, C.; Koval, L.E.; Klaren, W.D.; Harris, M.A.; Auerbach, S.S.; Rager, J.E. Predictive modeling of biological responses in the rat liver using in vitro Tox21 bioactivity: Benefits from high-throughput toxicokinetics. *Computational Toxicology* **2021**, *18*, 100166.
34. Zhang, J.; Mucs, D.; Norinder, U.; Svensson, F. LightGBM: An effective and scalable algorithm for prediction of chemical toxicity–application to the Tox21 and mutagenicity data sets. *Journal of chemical information and modeling* **2019**, *59*, 4150–4158.
35. LI, J.C. Imbalanced Toxicity Prediction Using Multi-Task Learning and Over-Sampling. 2020 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE, 2020, pp. 1–7.
36. Chopra, G.; Samudrala, R. Exploring polypharmacology in drug discovery and repurposing using the CANDO platform. *Current pharmaceutical design* **2016**, *22*, 3109–3123.
37. Sethi, G.; Chopra, G.; Samudrala, R. Multiscale modelling of relationships between protein classes and drug behavior across all diseases using the CANDO platform. *Mini reviews in medicinal chemistry* **2015**, *15*, 705–717.
38. Minie, M.; Chopra, G.; Sethi, G.; Horst, J.; White, G.; Roy, A.; Hatti, K.; Samudrala, R. CANDO and the infinite drug discovery frontier. *Drug discovery today* **2014**, *19*, 1353–1363.
39. Mangione, W.; Samudrala, R. Identifying protein features responsible for improved drug repurposing accuracies using the CANDO platform: Implications for drug design. *Molecules* **2019**, *24*, 167.
40. Mangione, W.; Falls, Z.; Chopra, G.; Samudrala, R. cando. py: Open Source Software for Predictive Bioanalytics of Large Scale Drug–Protein–Disease Data. *Journal of Chemical Information and Modeling* **2020**, *60*, 4131–4136.
41. Schuler, J.; Samudrala, R. Fingerprinting CANDO: Increased Accuracy with Structure-and Ligand-Based Shotgun Drug Repurposing. *ACS omega* **2019**, *4*, 17393–17403.
42. Schuler, J.; Falls, Z.; Mangione, W.; Hudson, M.L.; Bruggemann, L.; Samudrala, R. Evaluating the performance of drug-repurposing technologies. *Drug Discovery Today* **2021**. doi: https://doi.org/10.1016/j.drudis.2021.08.002.
43. Hudson, M.L.; Samudrala, R. Multiscale virtual screening optimization for shotgun drug repurposing using the CANDO platform. *Molecules* **2021**, *26*, 2581.
44. Mangione, W.; Falls, Z.; Melendy, T.; Chopra, G.; Samudrala, R. Shotgun drug repurposing biotechnology to tackle epidemics and pandemics. *Drug Discovery Today* **2020**, *25*, 1126–1128.
45. Swanson, H.I.; Bradfield, C.A. The AH-receptor: genetics, structure and function. *Pharmacogenetics* **1993**, *3*, 213–230.
46. Mimura, J.; Fujii-Kuriyama, Y. Functional role of AhR in the expression of toxic effects by TCDD. *Biochimica et Biophysica Acta (BBA)-General Subjects* **2003**, *1619*, 263–268.
47. Vorderstrasse, B.A.; Steppan, L.B.; Silverstone, A.E.; Kerkvliet, N.I. Aryl hydrocarbon receptor-deficient mice generate normal immune responses to model antigens and are resistant to TCDD-induced immune suppression. *Toxicology and applied pharmacology* **2001**, *171*, 157–164.
48. Huff, J.; Lucier, G.; Tritscher, A. Carcinogenicity of TCDD: experimental, mechanistic, and epidemiologic evidence. *Annual review of pharmacology and toxicology* **1994**, *34*, 343–372.
49. Marlowe, J.L.; Puga, A. Aryl hydrocarbon receptor, cell cycle regulation, toxicity, and tumorigenesis. *Journal of cellular biochemistry* **2005**, *96*, 1174–1184.
50. Dietrich, C.; Kaina, B. The aryl hydrocarbon receptor (AhR) in the regulation of cell–cell contact and tumor growth. *Carcinogenesis* **2010**, *31*, 1319–1328.
51. Quintana, F.J.; Sherr, D.H. Aryl hydrocarbon receptor control of adaptive immunity. *Pharmacological reviews* **2013**, *65*, 1148–1161.
52. Winans, B.; Humble, M.C.; Lawrence, B.P. Environmental toxicants and the developing immune system: a missing link in the global battle against infectious disease? *Reproductive toxicology* **2011**, *31*, 327–336.
53. Jackson, S.P. Sensing and repairing DNA double-strand breaks. *Carcinogenesis* **2002**, *23*, 687–696.

54. MacLachlan, T.K.; Somasundaram, K.; Sgagias, M.; Shifman, Y.; Muschel, R.J.; Cowan, K.H.; El-Deiry, W.S. BRCA1 effects on the cell cycle and the DNA damage response are linked to altered gene expression. *Journal of Biological Chemistry* **2000**, *275*, 2777–2785.
55. Gowen, L.C.; Avrutskaya, A.V.; Latour, A.M.; Koller, B.H.; Leadon, S.A. BRCA1 required for transcription-coupled repair of oxidative DNA damage. *Science* **1998**, *281*, 1009–1012.
56. Rosen, E.M.; Fan, S.; Pestell, R.G.; Goldberg, I.D. BRCA1 gene in breast cancer. *Journal of cellular physiology* **2003**, *196*, 19–41.
57. Chan, C.Y.; Kim, P.M.; Winn, L.M. TCDD affects DNA double strand-break repair. *Toxicological Sciences* **2004**, *81*, 133–138.
58. Rattenborg, T.; Gjermandsen, I.; Bonefeld-Jørgensen, E.C. Inhibition of E2-induced expression of BRCA1 by persistent organochlorines. *Breast Cancer Research* **2002**, *4*, 1–7.
59. Jeffy, B.D.; Schultz, E.U.; Selmin, O.; Gudas, J.M.; Bowden, G.T.; Romagnolo, D. Inhibition of BRCA-1 expression by benzo [a] pyrene and its diol epoxide. *Molecular Carcinogenesis: Published in cooperation with the University of Texas MD Anderson Cancer Center* **1999**, *26*, 100–118.
60. Foo, T.K.; Vincelli, G.; Huselid, E.; Her, J.; Zheng, H.; Simhadri, S.; Wang, M.; Huo, Y.; Li, T.; Yu, X.; others. ATR/ATM-Mediated Phosphorylation of BRCA1 T1394 Promotes Homologous Recombinational Repair and G2–M Checkpoint Maintenance. *Cancer Research* **2021**, *81*, 4676–4684.
61. Hayden, M.S.; Ghosh, S. Signaling to NF-κB. *Genes & development* **2004**, *18*, 2195–2224.
62. Tanaka, S.; Nakamura, K.; Takahasi, N.; Suda, T. Role of RANKL in physiological and pathological bone resorption and therapeutics targeting the RANKL–RANK signaling system. *Immunological reviews* **2005**, *208*, 30–49.
63. Hankinson, O. Role of coactivators in transcriptional activation by the aryl hydrocarbon receptor. *Archives of biochemistry and biophysics* **2005**, *433*, 379–386.
64. Voronov, I.; Li, K.; Tenenbaum, H.; Manolson, M. Benzo [a] pyrene inhibits osteoclastogenesis by affecting RANKL-induced activation of NF-κB. *Biochemical pharmacology* **2008**, *75*, 2034–2044.
65. Deng, L.; Wang, C.; Spencer, E.; Yang, L.; Braun, A.; You, J.; Slaughter, C.; Pickart, C.; Chen, Z.J. Activation of the IκB kinase complex by TRAF6 requires a dimeric ubiquitin-conjugating enzyme complex and a unique polyubiquitin chain. *Cell* **2000**, *103*, 351–361.
66. Weiss, C.; Faust, D.; Dürk, H.; Kolluri, S.K.; Pelzer, A.; Schneider, S.; Dietrich, C.; Oesch, F.; Göttlicher, M. TCDD induces c-jun expression via a novel Ah (dioxin) receptor-mediated p38–MAPK-dependent pathway. *Oncogene* **2005**, *24*, 4975–4983.
67. Ninomiya-Tsuji, J.; Kishimoto, K.; Hiyama, A.; Inoue, J.i.; Cao, Z.; Matsumoto, K. The kinase TAK1 can activate the NIK-IκB as well as the MAP kinase cascade in the IL-1 signalling pathway. *Nature* **1999**, *398*, 252–256.
68. Consortium, T.U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **2020**, *49*, D480–D489, [https://academic.oup.com/nar/article-pdf/49/D1/D480/35364103/gkaa1100.pdf]. doi:10.1093/nar/gkaa1100.
69. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic acids research* **2019**, *47*, D520–D528.
70. Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; others. The reactome pathway knowledgebase. *Nucleic acids research* **2020**, *48*, D498–D503.
71. Falls, Z.; Mangione, W.; Schuler, J.; Samudrala, R. Exploration of interaction scoring criteria in the CANDO platform. *BMC research notes* **2019**, *12*, 1–6.
72. Yang, J.; Roy, A.; Zhang, Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, *29*, 2588–2595.
73. Deric.; Sforna, G.; Landrum, G.; Winter, H.D. rdkit github page, 2012, [https://github.com/rdkit].
74. Deric.; Sforna, G.; Landrum, G.; Winter, H.D. RDKit MolVS python package, 2016, [https://molvs.readthedocs.io/en/latest/].
75. Bellman, R. Dynamic programming. *Science* **1966**, *153*, 34–37.
76. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **2002**, *16*, 321–357.
77. Beckmann, M.; Ebecken, N.F.; de Lima, B.S.P.; others. A KNN undersampling approach for data balancing. *Journal of Intelligent Learning Systems and Applications* **2015**, *7*, 104.

78. Batista, G.E.; Prati, R.C.; Monard, M.C. A Study of The Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD explorations newsletter* **2004**, *6*, 20–29.
79. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences* **2003**, *43*, 1947–1958.
80. Qi, Y. Random forest for bioinformatics. In *Ensemble machine learning*; Springer, 2012; pp. 307–323.
81. Dietterich, T.G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning* **2000**, *40*, 139–157.
82. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. *bioinformatics* **2007**, *23*, 2507–2517.
83. Nembrini, S.; König, I.R.; Wright, M.N. The revival of the Gini importance? *Bioinformatics* **2018**, *34*, 3711–3718.
84. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
85. Lemaıtre, G.; Nogueira, F.; Christos, K. Aridas. 2017."Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning.". *J. Mach. Learn. Res*, *18*, 559–563.
86. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves **2006**. pp. 233–240.
87. Provost, F.; Fawcett, T.; Kohavi, R. The case against accuracy estimation for comparing induction algorithms 1998. Proceedings of the 15th international conference on machine learning ICML-98 Morgan Kaufmann. San Mateo, CA.