United States Environmental Protection Agency

Structure-based activity relationships in a high-throughput assay for steroid biosynthesis

Miran J Foster[1,2], Grace Patlewicz[1], Imran Shah[1], Richard S. Judson[1], Katie Paul Friedman[1]

[1]Center for Computational Toxicology and Exposure, US EPA, Research Triangle Park, NC USA; [2]Oak Ridge Associated Universities

Abstract 2379/Poster P117
March 22, 2021
Society of Toxicology Virtual Annual Meeting

Miran J. Foster | foster.miran@epa.gov

This poster does not necessarily reflect EPA policy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Introduction and Data

BACKGROUND

- The high-throughput H295R assay for steroidogenesis (HT-H295R) assay was used to screen chemicals for putative effects on steroid hormone synthesis.
- In this work, we used chemical structure and physiochemical properties to predict bioactivity outcomes in chemicals with no HT-H295R bioactivity data.

DATA

- We used available HT-H295R data, including chemicals evaluated at multi or single concentrations (mc or sc).
- For 653 chemicals with mc data the 11 hormone system is summarized using Mahalanobis distance, converting 11 hormone measurements into 1 more easily interpretable binary outcome (Haggard *et al.*, 2018).
- MC data were highly unbalanced in their outcomes due to a tiered screening approach. Only 51 chemicals tested negative. Artificial negatives were created using the sc chemicals that: perturbed less than 3 hormones, did not perturb an estrogen or androgen hormone, and had a maximum response within 1 standard deviation of the mean.
- Result: 1400 unique structurable chemicals with physicochemical predictions. 845 negative and 555 positive.

Preliminary Structure-Activity Associations

CHEMICAL DESCRIPTORS

- Structures were described with two different sets of binary descriptors: ChemoType ToxPrints (Altamira) and Morgan extended-connectivity fingerprints (ECFP6) (Rogers and Hahn, 2010).
- Physicochemical property predictions from OPERA were obtained from the CompTox Chemicals Dashboard (version 3.5, 2020) and include 13 descriptors: (1) atmospheric hydroxylation rate (AOH), (2) bioconcentration factor (BCF), (3) biodegradability half-life, (4) boiling point, (5) Henry's Law constant, (6) fish biotransformation half-life (KM), (7) octanol: air partition coefficient (KOA), (8) soil adsorption constant (KOC), (9) octanol: water partition coefficient (logP), (10) melting point, (11) vapor pressure, (12) water solubility, and (13) average mass.

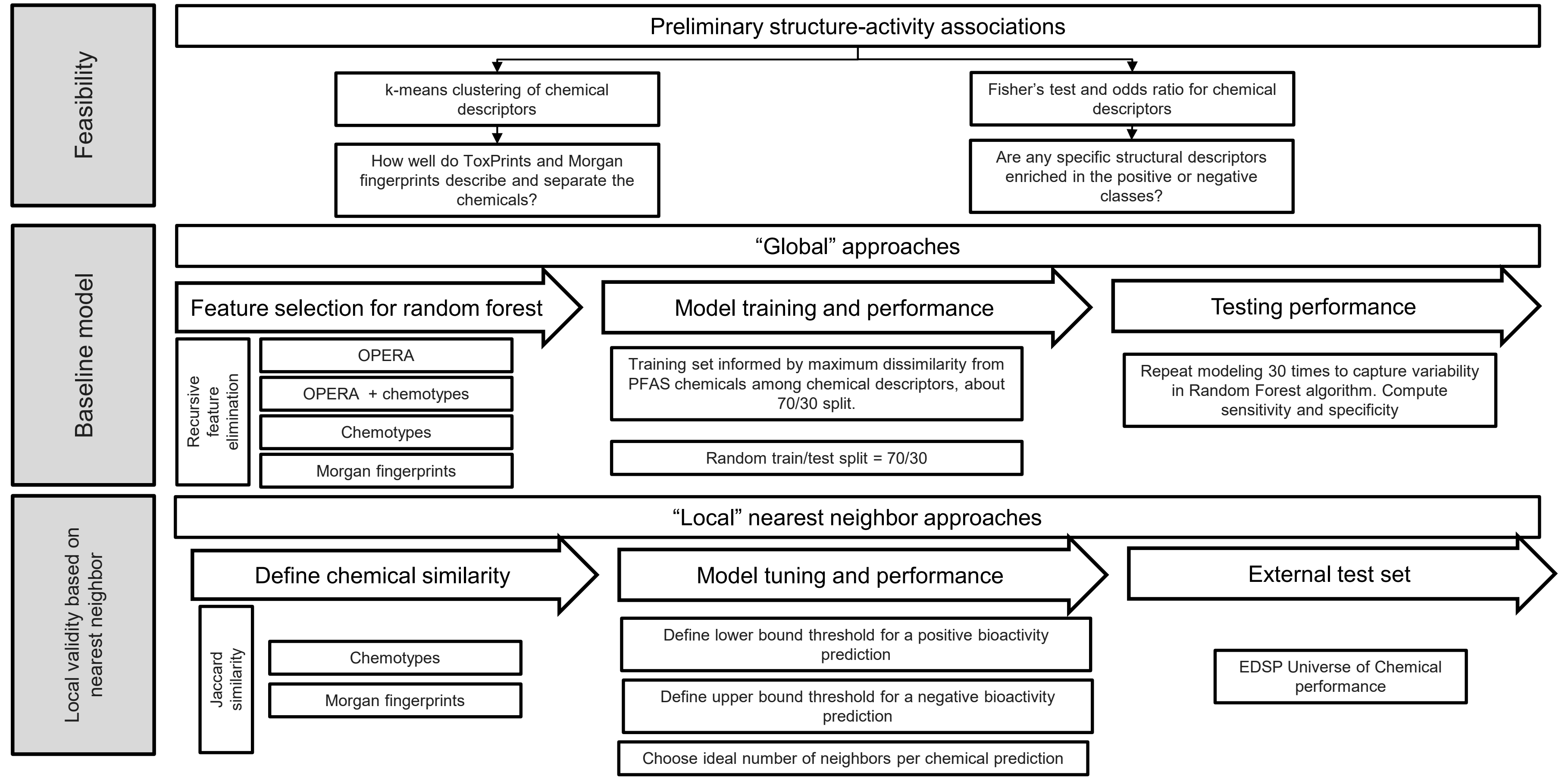
ENRICHMENT

To identify basic structure activity relationships between the bioactivity data and the chemotypes, we performed enrichment analysis. A Fisher's exact test (as many of the cell counts are small) was used to generate odds ratios that indicate the odds of a positive outcome given a present ChemoType. With 515/729 ChemoTypes present, p-values were adjusted using false discovery rate. 59 chemotypes are enriched in the negative space and 55 are enriched in the positive space.

10 ToxPrints with the largest or smallest odds ratios (from the chemical set with HT-H295R bioactivity data)

Odds Ratio	Lower Bound	Upper Bound	p-value	p-value, adjusted	ToxPrint Name
59.7965	10.5659	2345.7769	2.20E-16	2.20E-16	ring.fused_steroid_generic_.5_6_6_6.
14.1378	2.4132	569.2057	9.30E-05	5.24E-04	bond.CX_halide_alkenyl.Cl_dichloro_.1_1..
14.1378	2.4132	569.2057	9.30E-05	5.24E-04	bond.CX_halide_alkenyl.X_dihalo_.1_1..
13.8396	2.3601	557.4708	9.28E-05	5.24E-04	bond.CC..O.C_ketone_alkane_cyclic_.C5.
12.1720	3.2549	102.4057	2.00E-07	3.30E-06	bond.P.O_phosphate_dithio
0.0958	0.0310	0.2526	1.00E-07	1.10E-06	chain.alkaneLinear_tetradecyl_C14
0.1050	0.0243	0.3554	4.04E-05	2.64E-04	group.carbohydrate_ketohexose
0.1152	0.0365	0.3151	2.50E-06	2.32E-05	bond.CX_halide_alkyl.X_ethyl
0.1332	0.0414	0.3765	2.50E-05	1.80E-04	chain.alkeneLinear_mono.ene_vinyl
0.1444	0.0444	0.4168	7.72E-05	4.61E-04	bond.C.O_aldehyde_alkyl

Workflow Outline



Feasibility

Preliminary structure-activity associations

k-means clustering of chemical descriptors

How well do ToxPrints and Morgan fingerprints describe and separate the chemicals?

Fisher's test and odds ratio for chemical descriptors

Are any specific structural descriptors enriched in the positive or negative classes?

Baseline model

"Global" approaches

Feature selection for random forest

Recursive feature elimination

OPERA

OPERA + chemotypes

Chemotypes

Morgan fingerprints

Model training and performance

Training set informed by maximum dissimilarity from PFAS chemicals among chemical descriptors, about 70/30 split.

Random train/test split = 70/30

Testing performance

Repeat modeling 30 times to capture variability in Random Forest algorithm. Compute sensitivity and specificity

Local validity based on nearest neighbor

"Local" nearest neighbor approaches

Define chemical similarity

Jaccard similarity

Chemotypes

Morgan fingerprints

Model tuning and performance

Define lower bound threshold for a positive bioactivity prediction

Define upper bound threshold for a negative bioactivity prediction

Choose ideal number of neighbors per chemical prediction

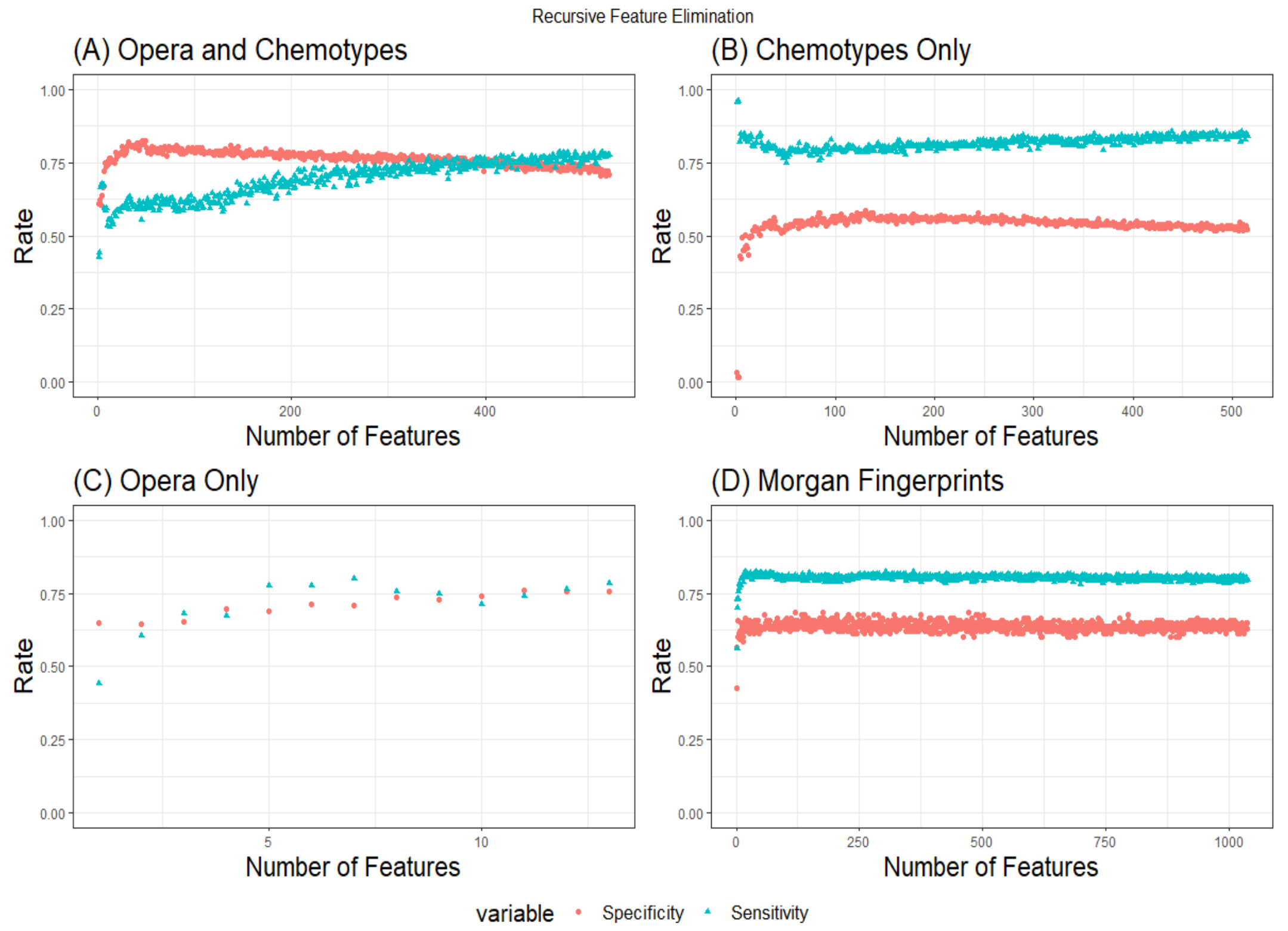
External test set

EDSP Universe of Chemical performance

Global approach: Random Forest Models

To attempt to classify the 1400 chemicals with HT-H295R data into bioactivity groups based on their structure and physicochemical properties, we first attempted random forest modeling. Two different training approaches were used: a simple random split (about 70/30 performs best) and using maximum dissimilarity. Maximum dissimilarity was performed by taking the set of 157 outlying chemicals found in kmeans and finding 900 additional chemicals that are the most dissimilar using the maxDissim function from the caret package in R. For model tuning, ntree was set at 700 trees, and mtry was (default) the square root of the number of features.

RECURSIVE FEATURE ELIMINATION



MODEL TRAINING AND PERFORMANCE

	Training			Testing		
	Sensitivity	Specificity	**Balanced Acc**	Sensitivity	Specificity	**Balanced Acc**
Opera	0.7317	0.6540	**0.6928**	0.7648	0.7154	**0.7401**
Opera (random training sample)	0.7989	0.6071	**0.703**	0.8034	0.6079	**0.7057**
Opera+ToxPrints	0.7587	0.6617	**0.7102**	0.7959	0.7254	**0.7606**
ToxPrints	0.7359	0.6200	**0.6780**	0.7679	0.6357	**0.7018**
Opera+Most enriched ToxPrint	0.7332	0.6550	**0.6941**	0.7661	0.7147	**0.7404**
Opera+Parent	0.7667	0.6543	**0.7105**	0.8222	0.7071	**0.7646**
Opera+Most important parent ToxPrint	0.7509	0.6288	**0.6898**	0.8032	0.7301	**0.7667**
Morgan+Opera (random training sample)	0.8069	0.5978	**0.7023**	0.8081	0.5951	**0.7016**
Morgan+Opera	0.7566	0.6617	**0.7091**	0.7970	0.6358	**0.7164**
Only fingerprints	0.7192	0.5865	**0.6528**	0.7601	0.5731	**0.6666**

Several models were developed based on combinations of chemical descriptors. ToxPrints were also modeled at a higher tier, condensing the features to higher a higher "parent" level (515 features become 71). The outcome of 30 model replications for each model are listed. Recursive feature elimination suggests addition of structural descriptors to OPERA physicochemical descriptors resulted in similar performance to the overall performance with OPERA descriptors alone. Adding structural features does very little to improve our model. This may not be because structural information is truly unimportant in informing bioactivity prediction but because random forest cannot efficiently utilize the amount of large binary data that structural features provide to its fullest potential.

Local Approach: Nearest Neighbor Models

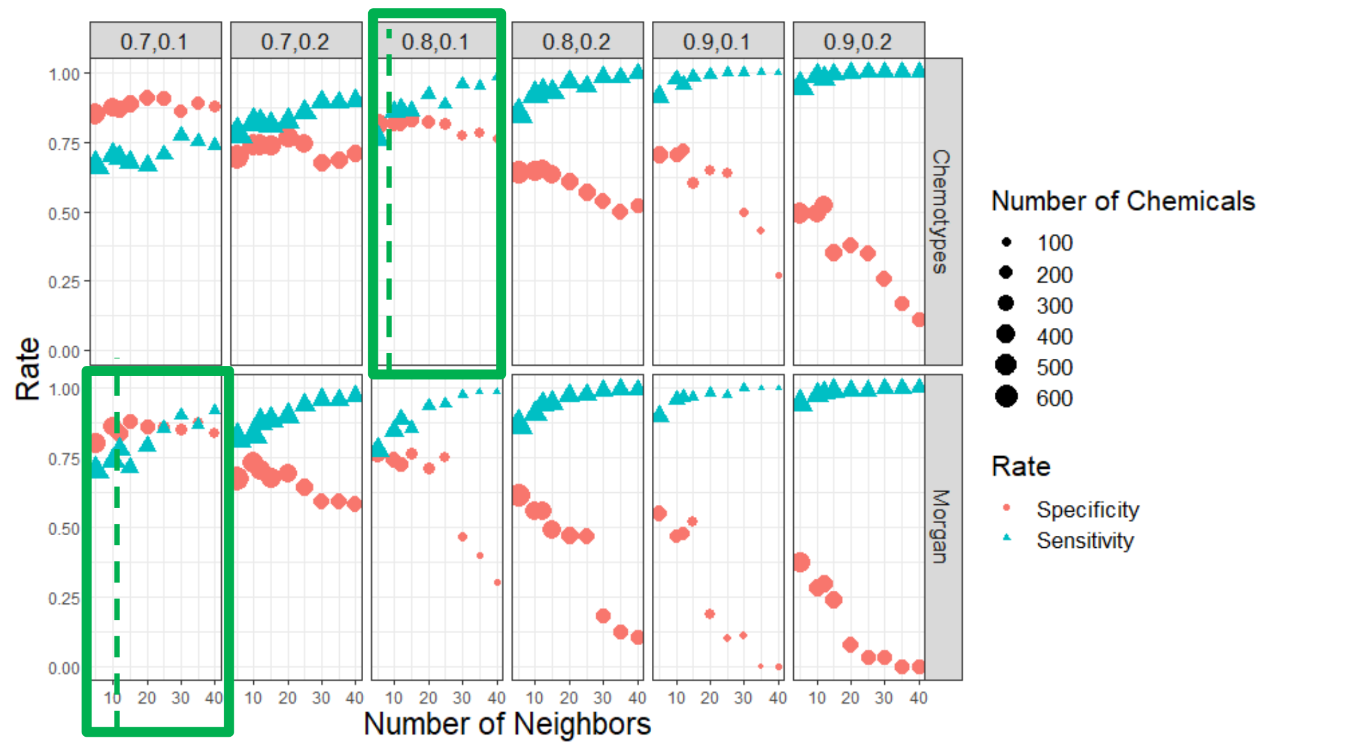
Nearest neighbor models better utilize the binary structural features that do not seem well utilized in random forest modeling. The method is similar to that of generalized read across or "GenRA." Jaccard similarities between all chemicals were generated based on their ToxPrints or Morgan fingerprints. For each chemical, n number of nearest neighbors with the highest Jaccard similarities were selected and applied with the formula below to create bioactivity predictions.

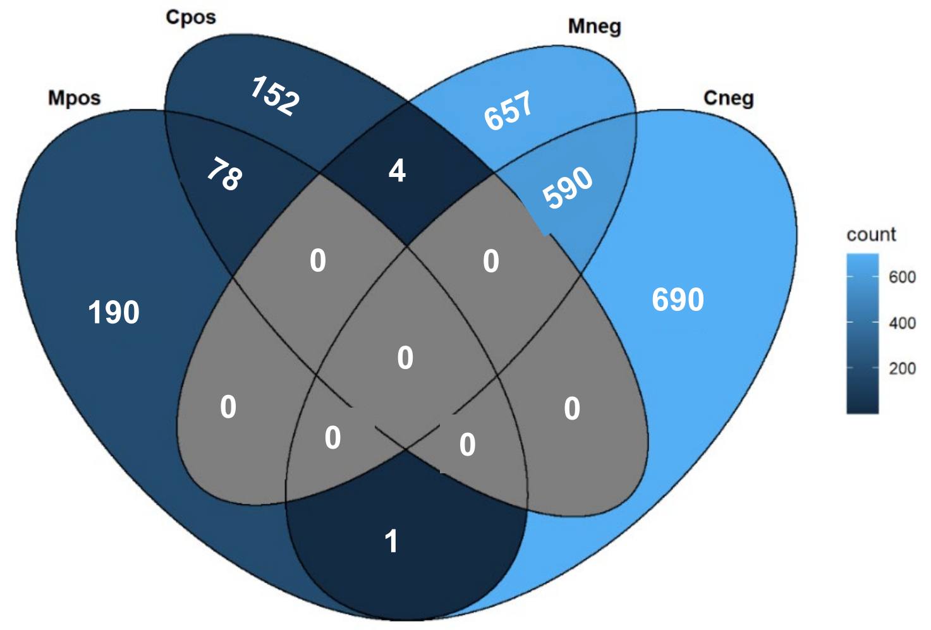
$$p_k = \frac{\sum_{i=1}^{n} j_{ik}\, x_i}{\sum_{i=1}^{n} j_{ik}}$$

Where j_{ik} is the Jaccard similarity between chemical k (prediction chemical) and i (nearest neighbor chemical). x_i represents the bioactivity outcome from HT-H295R (1 or 0) for chemical i, representing the i of n nearest neighbors selected. Finally p_k is the prediction for chemical k.

MODEL TUNING AND PERFORMANCE

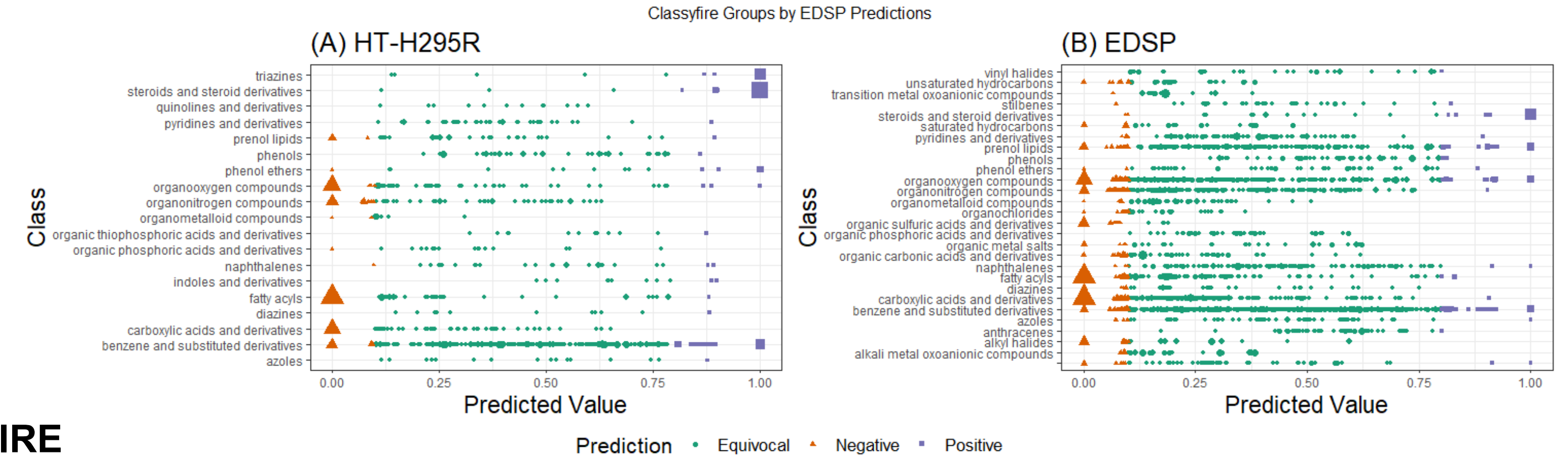
3 parameters adjusted: (1) # neighbors to choose when for predictions; (2) proportion positive neighbors for positive prediction; and, (3) proportion negative neighbors for negative prediction. Note that cutting the proportions off higher means that chemicals with proportions closer to 0.5 get left in the "equivocal space" without any prediction. This resulted in: ToxPrints: 10 neighbors wih cutoffs of 0.8 and 0.1, generating a sensitivity of 0.856 and specificity of 0.817, on 296 chemicals; Morgan: 12 neighbors at 0.7 and 0.1, yielding a sensitivity of 0.780 and specificity of 0.837 on 308 chemicals.





EDSP CHEMICAL UNIVERSE PREDICTIONS

The EDSP chemical universe from the CompTox Chemicals Dashboard contains 6302 structurable chemicals lacking HT-H295R data. Jaccard distances for each of these chemicals with the original 1400 with assay data were generated for prediction generation. The Venn diagram shows the two models (labeled M for Morgan and C for chemotype) and how many chemicals were positive (Pos) and negative (Neg). The substances Pos in both models may have the highest confidence predictions.



CLASSYFIRE

Classyfire taxonomy identified structure-based chemical groups for positive, negative, or equivocal predictions in the nearest neighbor models. Increased point size indicates more chemicals were predicted at that value. The "steroid and steroid derivatives" is the group that tested positive most often in both data sets. Many chemical groups fall somewhere along the equivocal spectrum.

Conclusions

In the two modeling approaches used, there is a trade off between performance and the number of chemicals that yield predictions. A global approach with random forest modeling performed best with OPERA predictors (a balanced accuracy ~74%). In a nearest neighbor approach, better accuracy is achieved, (80-84%), but fewer chemicals have non-equivocal predictions. Using both approaches depending on context may inform gaps in screening data or prioritize chemicals for additional screening.