

# Artificial Neural Networks & Random Forest Classification of druggable molecules and disease targets via scoring functions (SFs)

I.L. Hudson<sup>a</sup> , S.Y. Leemaqz<sup>b</sup>  and A.D. Abell<sup>c</sup> 

<sup>a</sup> Department of Mathematical Sciences, College of Science, Engineering and Health, Royal Melbourne Institute of Technology (RMIT), Melbourne, Victoria, Australia, <sup>b</sup> Robinson Research Institute, Adelaide Medical School, University of Adelaide, South Australia, <sup>c</sup> Department of Chemistry, Adelaide Node Director Centre for Nanoscale BioPhotonics (CNBP), University of Adelaide, Adelaide, South Australia  
Email: irene.hudson@rmit.edu.au

**Abstract:** In recent years, machine learning has played an increasing role to help identify druggable molecules. In particular research has shown that random forests (RFs), recursive partitioning (RP), support vector machines (SVMs) and artificial neural networks (ANNs) have been commonly employed in this arena. Expanding disease modifying targets to pharmacological manipulation is vital to human health. Modelling disease targets allow for prediction and prioritisation based on their molecular characteristics and druggability. The aim of this current paper is 2 fold: (i) to propose a computational method to identify druggable disease targets using combinations molecular parameters (MPs) and (ii) to establish which of ANN or RF procedures and which scoring functions best partition molecular and disease target space. Classifications by Artificial Neural Networks (ANNs) and Random Forest (RF) based on 8 molecular parameters (MPs) were performed to classify disease targets with high or low violator scores (using cutpoints 3, 4 or 5), and the 4 traditional parameters of Lipinski's rule of five (Ro5), plus 4 extra parameters (polar surface area (PSA), number of rotatable bonds and rings, N and O atoms, and a choice between 2 alternatives for lipophilicity, the distribution coefficient (log D) and the partition coefficient (log P) (Hudson et al., (2017), Zafar et al., (2013, 2016)).

For the molecule parameter (MP) data RF performed better than ANNs and the log D model of either score 4 or score 5 was optimal compared to the log P model. ANNs however, were superior to the RF models for MP sets containing both log D and log P. For the RF score 4 log D model the most important variables were log D, molecular weight (MW) and number of rotatable bonds (ROT). The next best model via RF was score 5 log D, with its most important variables being PSA, log D and MW, according to mean decrease in gini scores.

Overall, for the target data the RF models performed better than ANNs, with inclusion of log D being important. For the RF target models the score 5 partition performed best, AUC (95% CI) of 0.88 (0.63, 1.0) for all 3 models; with the higher mean decrease gini values (MDGs) attributable to MPs (MW, NATOM, ROT, PSA Hacceptors, NRING). The MP variables then chosen with lower MDGs were (log D, NATOM, NRING, log P, Hdonors), indicating log D is superior to log P (VIs, 2.14 > 1.47). Also the RF score 4 log D, and log P models performed equally well, AUC (95% CI) of 0.85 (0.70, 1.00) - closely followed by the RF score 3 target models, score 3 log D and score 3(log D+log P), which both did well with AUC (95% CI) of 0.84 (0.73, 0.94).

The ANN target based score 4 log D model, achieved best classification, with AUC (95% CI) of 0.89 (0.77, 1.0). In contrast the score 4 log D+Log P model performed the worst, with AUC (95% CI) of 0.69 (0.51, 0.86). Similarly for the RF analysis, the score 4 log D+log P performed worse with AUC (95% CI) of 0.83 (0.68, 0.92), whilst separate score 4 log D or log P models classified equally well (0.85, (0.707, 1.0)). All 3 cutpoint 3 ANN target models, showed PSA to be highly important compared to the MW. In contrast MW is the most important variable for all RF target models and all cutpoints. Log D has greater variable importance (VI) compared to MW in the score 3 log D+log P ANN model (17.31 > 12.60). Also in the score 3 log P ANN model, MW has least VI of 6.46 compared to log P's VI of 17.15. Log D is more important than log P in the score 3 log D+log P model. For the optimal score 4 log D, model top VIs are attributable to (PSA, log D, NRING, Hacceptors, MW), showing strong influence of PSA and Log D compared to the traditional MW.

The RP and ANN rules to classify the high score violators from the low confirmed the value of log D in the scoring function, validating Zafar et al. (2016, 2013) and the original MC/DA cutpoints for each MP by Hudson et al. (2017). Score functions of violations and best cutpoints to identify druggable molecules and targets were confirmed and shown to be associated with specific diseases. Our simple scoring functions of counts of violations partitioned chemospace well, identifying both good/ poor druggable molecules and targets.

**Keywords:** Disease targets, score function druggability rules, machine learning

## 1. INTRODUCTION

In recent years, machine learning has played an increasing role to help determine the classification of druggable molecules (Doak & Kihlberg, 2017). In particular, research has shown that recursive partitioning (RP), random forests (RFs), support vector machines (SVMs) and artificial neural networks (ANNs) have been commonly employed in this arena (Lavecchia, 2015; Kandoi *et al.*, 2015). High throughput docking of small molecule ligands (candidate drugs) into high resolution protein structures is now standard in computational approaches to drug discovery (Ursu *et al.*, 2017). For example Hudson *et al.* (2016) investigated Self Organising Map (SOM) ANNs as a computational tool for the evaluation of docking experiments of calpain ligands (small drug molecules) for the treatment of cataracts. Recently Hudson *et al.*, (2017) also studied the performance of support vector machine (SVM) and Recursive partitioning (RP) based on 10 molecular descriptors, to classify molecules with high or low violator scores (defined by an optimal cutpoint,  $C$ , shown to be 5), based on skew normal and logit models. RP and SVM were then used to classify the high score molecular violators from the low ( $< 5$ ). Hudson *et al.*, (2017) showed that SVM used in combination with simple molecular descriptors provided a reliable assessment of a scoring function of counts of violations of MPs developed by Hudson *et al.*, (2014) to partition molecular chemo-space into druggable molecules (Guan *et al.*, 2019).

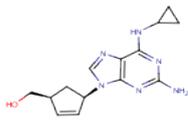
Recently many disease targets have been classified as “undruggable” due to their lack of oral bioavailability (Doak *et al.*, 2016; Gashaw *et al.* 2011). Generally such targets have binding sites which are large, highly lipophilic, flexible, highly polar and featureless. This has in part created a momentum in small molecule drug discovery to move outside the rule of 5 (Ro5) space of Lipinski (2016), to the so-called beyond Ro5 (bRo5) space (Doak *et al.*, 2017). The aim of this current paper is two-fold: (i) to propose a computational method to identify druggable disease targets using combinations of molecular parameters and (ii) to establish which of ANN or RF procedures and which scoring functions best partition disease target space. This has not been formally established for the data set studied by Hudson *et al.*, (2017). Note however that Hudson *et al.* (2017) showed indirectly that molecules with 5 or more violations were empirically associated with specific disease targets, but did not test machine learning methods using the actual target scores associated with the 172 disease targets, as performed in this paper. Recent preliminary work by Leemaqz *et al.* (2018), using SVM and RPs directly on the median scores of molecules associated with disease targets, showed the superiority of SVM to RP to partition targets with a median score of 4 or more violations. The score function evaluated in this current study for both the molecular parameters (MPs) and for the target median scores, comprises the 4 traditional parameters of the rule of five (Ro5) (Lipinski 2016), plus 5 extra molecular parameters, polar surface area (PSA), number of rotatable bonds, rings, number of N and O atoms, and Log P (a measure of lipophilicity) and Log D (the distribution coefficient) suggested by Zafar *et al.*, (2016, 2013) as a preferable predictor for permeation to Lipinski’s classical partition coefficient, Log P (Bhal *et al.*, 2007). Hudson *et al.* (2017, 2014) developed druggability rules for molecules (scores counting violations) that account for physico-chemical properties, and derived novel cutpoints for each of 10 MPs based on a mixture clustering (mclust) discriminant analysis (MC/DA) approach (Fraley *et al.*, 2012). We use these MP cutpoints in this study, which were shown to be much in agreement with recent cutpoints per molecular parameter reported by Ursu *et al.* (2017).

## 2. DATA AND METHODS

### 2.1. Data and Druggability scoring for molecules and disease targets

We analysed 1279 small molecules from the DrugBank database (Law *et al.*, 2013), a unique bioinformatics and chemo-informatics resource combining detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with drug target (i.e. sequence, structure, and pathway) information, containing 6,711 drug entries (a candidate molecule is shown in Table 1). In total there are 105 Ro5 non-compliant molecules, 681 with oral and 598 non-oral delivery modes. In this study 172 targets representing 1279 molecules were found and the median score value obtained for the molecules in the given target; with 99 targets predominantly oral and 73 non-oral in terms of modes of delivery. Specifically in this study MP score functions were based upon the filter scores of Hudson *et al.* (2017), where molecules with violations of less or equal to 4 or 5, were considered a druggable molecule. In this study the score function per molecule sums the number of violations based on 8 or 9 MPs modelled. The filter scores were created as follows for each molecular parameter (MP): lipophilicity (Log P)  $\leq 1.9$ , lipophilicity (Log D)  $\leq 3.5$ , molecular weight (MW)  $\leq 305$ , hydrogen bond donors  $\leq 4$ , hydrogen bond acceptors  $\leq 10$ , polar surface area (PSA)  $\leq 140$ , rotatable bonds  $\leq 7$ , ring numbers  $\leq 2$ , and number of nitrogen and oxygen atoms  $\leq 40$ . Hudson’s cutpoint for Log D of 3.5, is smaller than 5.5, suggested by Bhal *et al.* (2007). Implementation of  $\log D \sim 3.5$ , instead of the classical  $\log P$  parameter for the estimation of the molecule’s lipophilicity was shown to be more optimal (Zafar *et al.*, 2016; 2013), but in this study using ANN and RF we test models including both  $\log D$  and  $\log P$  and those with just one of these lipophilicity variants.

**Table 1.** DrugBank3.0 information on one candidate molecule (DB01048 Abacavir).

| DrugBank ID & Name CAS Number   | Molecular Weight Formula                                     | Chemical Structure  | Categories   | Therapeutic Indication   |
|---------------------------------|--|---|--|--|
| DB01048 Abacavir<br>136470-78-5 | 286.3323<br>C <sub>14</sub> H <sub>18</sub> N <sub>6</sub> O |  | Anti-HIV Agents /<br>Nucleoside and<br>Nucleotide Reverse<br>Transcriptase Inhibitors<br>/ Reverse Transcriptase<br>Inhibitors | For the treatment of HIV-1<br>infection, in combination<br>with other antiretroviral<br>agents |

## 2.2. Statistical and program approach

The dataset was partitioned into 70% and 30% for training and testing purposes, respectively, to avoid model overfitting. The success of the model was determined by assessing the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, a measure used to predict model accuracy (Sachs, 2018). For replication, all models were created using a seed of 44. RF models were based on 10-fold cross validation on the training partition with a repetition of 5. As RF is considered an accumulative model algorithm, the only tuning parameter used was the number of trees. The averaged model across the specified number of trees was then assessed on the testing data. Artificial neural network (ANN) models were created using a 10-fold cross validation on the training partition and repetition of 5, with models assessed via the training partition. All variable scores were scaled and centred before being integrated in the model.

The final model was created using a tune grid containing decay (to avoid overfitting via regularisation) parameters between 0.1 to 0.5 by a value of 0.1 and size parameters (the number of hidden nodes in the network) between 1 to 10 used. Each of these steps was conducted across each of the different score functions: score C (log P), score C (log D), score C (log D+log P), for C=3, 4 or 5. This ensured the best machine learning method was identified and the most effective score for classification established. RStudio and Kuhn *et al.*'s (2018) Caret package were used for all classification and Beck's (2018) NeuralNetTools for the visual plots for the neural network models.

## 3. RESULTS

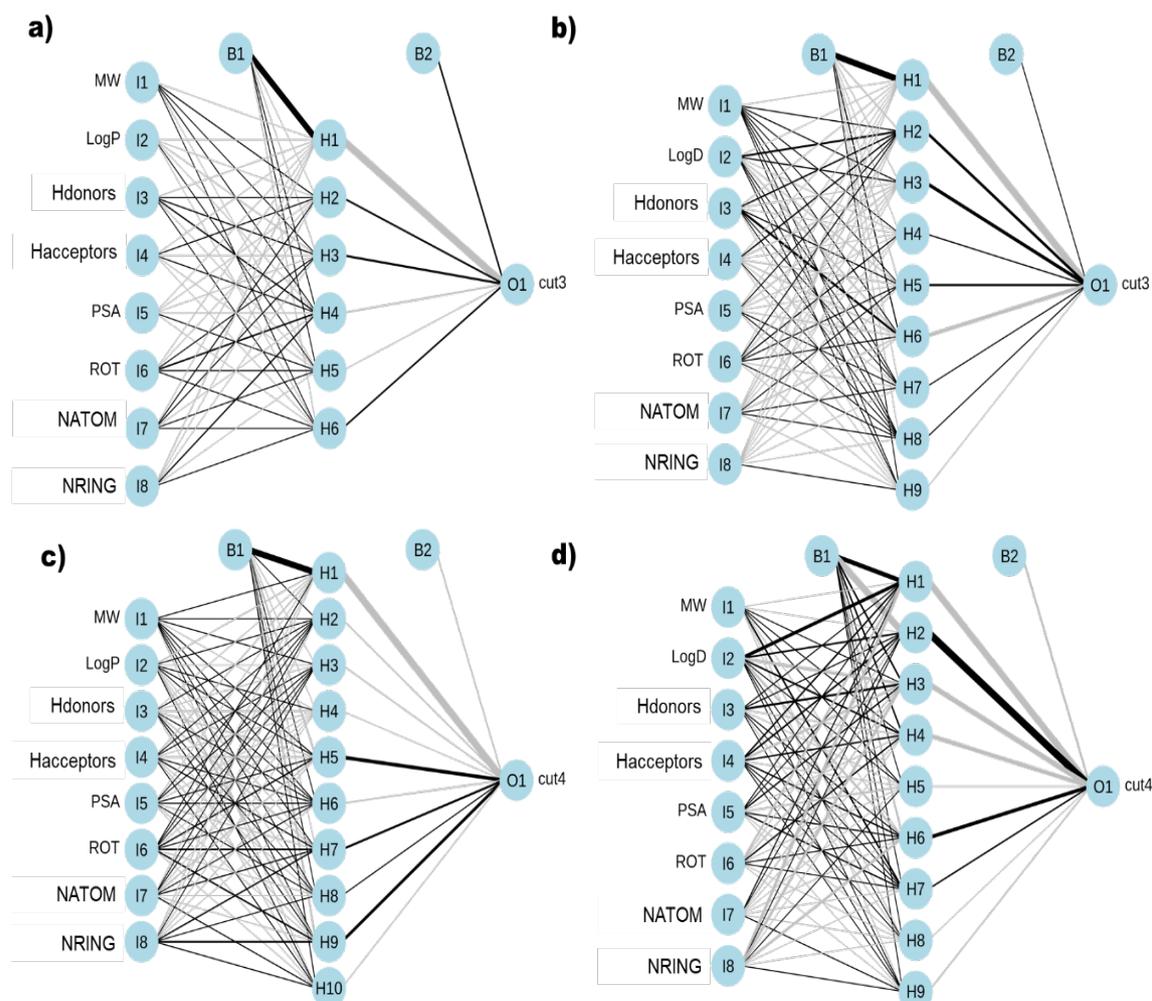
The AUC and 95% CI of each machine learning model (RF and ANN) for both the molecular properties (MPs) and the disease targets data across score functions (with varying cutpoints and model variations founded on whether Log D, Log P or both were included in the model (log D+log P) are given in Table 2. Based on the work of Hudson *et al.*, (2017) only score 4 or 5 cutpoints were tested and valid for the MPs, as reported in Table 2. For the target based score functions cutpoints 3 to 5 were tested (see Leemaqz *et al.*, 2018) (Table 2).

**Table 2.** AUC and 95 % CIs for RF and ANN models for both the MPs and the Disease Target data.

| Score functions    | RF on MPs                  | ANN on MPs                 | RF on Targets            | ANN on Targets           |
|--------------------|----------------------------|----------------------------|--------------------------|--------------------------|
|                    | AUC (95% CI)               | AUC (95% CI)               | AUC (95% CI)             | AUC (95% CI)             |
| Score 3 log P      | -                          | -                          | 0.82 (0.71, 0.93)        | 0.82 (0.71, 0.93)        |
| Score 3 log D      | -                          | -                          | <b>0.84 (0.73, 0.94)</b> | 0.80 (0.69, 0.91)        |
| Score 3 log D+logP | -                          | -                          | <b>0.84 (0.73, 0.94)</b> | <b>0.84 (0.73, 0.94)</b> |
| Score 4 log P      | 99.58 (99.18-99.98)        | 99.29 (98.72-99.85)        | 0.85 (0.70, 1.00)        | 0.72 (0.54, 0.90)        |
| Score 4 log D      | <b>99.70 (99.42-99.99)</b> | <b>99.27 (98.59-99.70)</b> | <b>0.85 (0.70, 1.00)</b> | <b>0.89 (0.77, 1.00)</b> |
| Score 4 log D+logP | 93.05 (92.70-94.20)        | 98.00 (97.60-99.00)        | 0.83 (0.68, 0.92)        | 0.69 (0.51, 0.86)        |
| Score 5 log P      | 99.16 (98.51-99.82)        | 98.86 (98.04-99.69)        | 0.88 (0.63, 1.00)        | 1.00 (1.00, 1.00)        |
| Score 5 log D      | <b>99.33 (98.73-99.92)</b> | <b>99.01 (98.22-99.80)</b> | <b>0.88 (0.63, 1.00)</b> | 0.50 (0.50, 0.50)        |
| Score 5 log D+logP | 93.79 (92.70-95.20)        | 98.10 (97.60-98.60)        | 0.88 (0.63, 1.00)        | 0.50 (0.50, 0.50)        |

### 3.1. Results of the analyses based on molecular parameters

From Table 2 the RF and ANN models on the MPs, demonstrated that RF performed better than ANNs and the log D variant of either score 4 or score 5 was optimal. Random forest analysis of the MPs showed that score 4 log D had best performance followed by score 5 log D (Table 2). In the RF score 4 log D model the most important variables were log D, molecular weight (MW) and number of rotatable bonds (ROT), based on gini values (gini values not reported here). The next best model via RF, score 5 log D had polar surface area (PSA), log D and molecular weight (MW) as the most important variables. Overall, the score functions that best classified molecular druggable status included log D. From Table 2 ANNs of the MPs showed that score 4 log P performed best (closely followed by its log D variant), with its most important MP variables being



**Figure 1.** Artificial Neural Network plots for each Drug Target model; a) cutpoint 3 log P, b) cutpoint 3 log D, c) cutpoint 4 log P and d) cutpoint 4 log D.

log D, molecular weight (MW) and number of rotatable bonds (ROT). The next best ANN model was score 5 log D with most important variables being polar surface area (PSA), log D and MW. The ANN log P models, based on score 4 log P and score 5 log P, revealed that MW and log P were important for the score 5 log P model, with higher variable importance (VI) being attributable to MW, hydrogen bond acceptors (Hacceptors) and Log P. However, neither MW nor log P were important to the score 4 log P model, with highest VIs attributable to Hacceptors, ROT and NATOM as MPs (VIs are not reported here).

### 3.2. Results of the analyses based on the target data

The ANN plots for the 4 models of the target data based on score C log P, or score C log D (for C=3, 4) are given in Figure 1. Based on AUCs the ANN models on the disease target data, the log D variant of score 4 was

optimal (Table 2). The following ANN target models for score 3 log D+log P and score 3 log P are optimal with AUC (95% CI) of 0.84 (0.73, 0.94) and 0.82 (0.71, 0.93), respectively (Table 2). Note that RF based target models for score 3 log D+log P and score 3 Log D achieved the same AUC values as the corresponding score 3 log D+Log P ANN models. For the ANN target based C=4 partition, score 4 log D model achieved best classification, with AUC (95% CI) of 0.89 (0.77, 1.0). In contrast the score 4 log D+Log P model performed the worst, with AUC (95% CI) of 0.69 (0.51, 0.86). Similarly for the RF analysis, score 4 log D and score 4 log P classified equally well (0.85, (0.707, 1.0), with again the score 4 log D+log P, worse with AUC (95% CI) of 0.83 (0.68, 0.92). The score 5 ANN models were poor given sparse data and not detailed here.

For all 3 ANN target models based on cutpoint 3 in Table 3, PSA is shown to be highly important in comparison to MW (Table 4). In contrast MW is the most important variable for all the RF target models for all cutpoints. Also Log D has greater variable importance (VI) compared to MW in the log D+log P model (17.31 >12.60) and log D and MW have similar VIs for the log D cutpoint 3 model (13.21~13.05). Interestingly for the log P score 3 model, MW has least VI, with log P and MW VIs being 17.15 and 6.46, respectively. Log D is more important than log P in the score 3 log D+log P model (17.31>8.13) but less so for the score 4 log D+log P model. For the ANN score 3 log P target model, top VIs are attributable to (Hdonors, log P, PSA) with VIs (20.65, 17.15, 14.48) (with MW of least importance with a VI of 6.46). For the most optimal score 4 log D ANN model top VIs are attributable to (PSA, log D, NRING, Hacceptors, MW) with VIs (21.65, 19.53, 13.18, 11.48, 11.06), revealing the stronger contribution of both PSA and Log D, compared to MW (Table 4).

For the RF target models score 5 partition performed the best across all 3 models with AUC (95% CI) of 0.88 (0.63, 1.0) (Table 2). Likewise the RF target score 3 models involving log D (i.e. log D and (log D+log P)) performed equally best with AUC (95% CI) of 0.84 (0.73, 0.94). For the RF target score 4 models both separate log D and log P models performed equally well with AUC (95% CI) of 0.85 (0.70, 1.00). Overall the RF models performed better than ANNs and inclusion of log D was important (Table 2). Table 4 reports the mean decrease gini (MDG) values for the RF target models, showing that for cutpoint 3 the highest MDGs were attributable to (MW, NATOM, PSA, NRING, Hacceptors, ROT), in that order, for all 3 models. The variables then chosen were (log D, Hdonors, log P) with corresponding MDG sequence (3.04, 2.93, 1.84), indicating log D is more important than log P. For the remaining score 3 RF models the sequence of variables and MDGs were (log D, Hdonors) with MDGs (3.86, 2.86) and (Hdonors, log P) with MDGs (3.0, 2.8), indicating log D's superiority. In the RF cutpoint 4 target models the contribution of the higher MDGs were from (MW, NATOM, ROT, Hacceptors, NRING, PSA) in that order for all 3 models (Table 4).

The variables then chosen with lower MDGs were (log D, log P, Hdonors) with MDG sequence (2.03, 1.77, 1.69), indicating both log D and log P are of importance. For cutpoint 5 the contribution of the higher MDGs were similarly (MW, NATOM, ROT, PSA Hacceptors, NRING) for all 3 models. The variables then chosen with lower MDGs were (log D, NATOM, NRING, log P, Hdonors) with MDG sequence (2.14, 1.84, 1.73, 1.47, 1.25), indicating log D is superior to log P in importance. For the remaining score 5 models the most important contributors were (NATOM, log D, NRING, Hdonors) with MDGs (2.57, 2.54, 1.81, 1.48), and (NATOM, log P, NRING, Hdonors) with MDGs (2.58, 2.06, 1.95, 1.43), indicating the value of NATOM and either log D or log P in the score 5 RF models.

#### 4. CONCLUSION

The RP and ANN rules to classify the high score violators from the low confirmed the value of log D's inclusion in the scoring function and supported the original MC/DA cutpoints established for each MP (Hudson *et al.*, 2017). Score function of violations and best cutpoints to identify druggable molecules and disease targets were confirmed. Overall the RF models performed better than ANNs and inclusion of log D was important. The most important MPs that influence molecular classification, determined by RF, were MW, NATOM, log D and PSA. It was found that log D was a better measure of lipophilicity for classification as shown in all MP analyses. The ANN target models indicated that PSA and log D were highly important, while the RF models showed MW, NATOM and PSA to be more important than Log D or Log P. Our work illustrates that simple scoring functions of counts of violations can partition chemospace and help identify both good and poor druggable molecules and targets. Moreover, molecules with score functions above 4 or 5 were shown by Hudson *et al.* (2017) to be associated with specific disease targets, e.g., Anti-Bacterial, Antineoplastic, Antihypertensive and Anti-allergic, within which most molecules have a non-oral delivery mode. Target drugs with a low median score were e.g., Adrenergic, Dietary, Analgesics, Anesthetics, Adjuvants, Anti-convulsants, Antimetabolites and Antidepressants, most of which were non-oral. Targets were shown to correlate with our score C log D or log P partitions (C = 3 to 5). Future work will compare SVM, RF, ANN and RP methods using models based on novel factor analytic MP and target constructs and related score functions and will also test new cutpoints.

**Table 3.** Variable Importance values (VI): ANNs on targets: C= 3-5

| ANN Variable Importance values (VI) on target data |       |             |       |             |       |
|--|-------|-------------|-------|-------------|-------|
| Cutpoint 3   |       |             |       |             |       |
| logD+LogP  | VI    | logD        | VI    | logP        | VI    |
| PSA  | 20.87 | PSA         | 16.77 | Hdonors     | 20.65 |
| <b>LogD</b>  | 17.31 | NATOM       | 15.77 | <b>LogP</b> | 17.15 |
| Hacceptors   | 14.13 | <b>LogD</b> | 13.21 | PSA         | 14.48 |
| <b>MW</b>  | 12.60 | <b>MW</b>   | 13.05 | NRING       | 12.80 |
| Hdonors  | 9.75  | Hacceptors  | 12.12 | ROT         | 10.79 |
| <b>LogP</b>  | 8.13  | Hdonors     | 11.73 | NATOM       | 10.62 |
| NATOM  | 7.76  | NRING       | 9.52  | Hacceptors  | 7.04  |
| NRING  | 5.73  | ROT         | 7.82  | <b>MW</b>   | 6.46  |
| ROT  | 3.72  |             |       |             |       |
| Cutpoint 4   |       |             |       |             |       |
| NATOM  | 15.49 | PSA         | 21.65 | <b>MW</b>   | 31.05 |
| PSA  | 15.46 | <b>LogD</b> | 19.53 | PSA         | 14.36 |
| ROT  | 13.34 | NRING       | 13.18 | Hacceptors  | 10.36 |
| <b>LogD</b>  | 13.16 | Hacceptors  | 11.48 | NATOM       | 10.22 |
| <b>LogP</b>  | 12.80 | <b>MW</b>   | 11.06 | <b>LogP</b> | 9.97  |
| Hacceptors   | 11.66 | NATOM       | 9.50  | NRING       | 9.14  |
| NRING  | 6.43  | ROT         | 7.52  | ROT         | 9.07  |
| Hdonors  | 6.27  | Hdonors     | 6.08  | Hdonors     | 5.84  |
| MW   | 5.38  |             |       |             |       |
| Cutpoint 5   |       |             |       |             |       |
| <b>MW</b>  | 56.94 | <b>MW</b>   | 30.65 | <b>MW</b>   | 34.90 |
| PSA  | 31.87 | NATOM       | 14.38 | PSA         | 16.40 |
| NATOM  | 7.69  | PSA         | 11.19 | Hacceptors  | 11.30 |
| Hacceptors   | 1.87  | Hacceptors  | 10.98 | NATOM       | 9.57  |
| <b>LogD</b>  | 0.79  | ROT         | 10.88 | Hdonors     | 9.02  |
| <b>LogP</b>  | 0.67  | <b>LogD</b> | 9.95  | ROT         | 7.73  |
| ROT  | 0.11  | Hdonors     | 7.21  | <b>LogP</b> | 5.77  |
| Hdonors  | 0.06  | NRING       | 4.77  | NRING       | 5.31  |
| NRING  | 0.01  |             |       |             |       |

**Table 4.** Mean Decrease Gini (MDG) values for RF targets: C= 3 -5

| RF Mean Decrease Gini (MDG) values on target data |       |             |       |             |             |
|---|-------|-------------|-------|-------------|-------------|
| Cutpoint 3  |       |             |       |             |             |
| logD+LogP   | MDG   | logD        | MDG   | LogP        | MDG         |
| MW  | 20.62 | MW          | 18.73 | MW          | 18.62       |
| NATOM   | 12.95 | NATOM       | 13.30 | NATOM       | 12.86       |
| PSA   | 6.28  | PSA         | 5.86  | PSA         | 6.45        |
| NRING   | 4.78  | NRING       | 5.30  | NRING       | 6.15        |
| Hacceptors  | 4.26  | Hacceptors  | 5.12  | Hacceptors  | 5.02        |
| ROT   | 3.11  | ROT         | 4.50  | ROT         | 4.61        |
| <b>LogD</b>                                       | 3.04  | <b>LogD</b> | 3.86  | Hdonors     | 3.00        |
| Hdonors   | 2.63  | Hdonors     | 2.86  | <b>LogP</b> | <b>2.80</b> |
| <b>LogP</b>                                       | 1.84  |             |       |             |             |
| Cutpoint 4  |       |             |       |             |             |
| MW  | 14.32 | MW          | 12.72 | MW          | 12.54       |
| NATOM   | 8.18  | NATOM       | 8.95  | NATOM       | 8.31        |
| ROT   | 4.16  | ROT         | 4.55  | ROT         | 5.04        |
| Hacceptors  | 3.70  | PSA         | 4.36  | Hacceptors  | 4.41        |
| NRING   | 3.45  | Hacceptors  | 4.11  | NRING       | 4.24        |
| PSA   | 3.22  | NRING       | 3.94  | PSA         | 3.80        |
| <b>LogD</b>                                       | 2.03  | <b>LogD</b> | 2.59  | <b>LogP</b> | <b>2.57</b> |
| <b>LogP</b>                                       | 1.77  | Hdonors     | 1.69  | Hdonors     | 1.93        |
| Hdonors   | 1.69  |             |       |             |             |
| Cutpoint 5  |       |             |       |             |             |
| MW  | 5.07  | MW          | 4.57  | MW          | 4.59        |
| Hacceptors  | 4.47  | PSA         | 4.16  | PSA         | 4.37        |
| PSA   | 3.99  | Hacceptors  | 4.06  | Hacceptors  | 4.19        |
| ROT   | 2.77  | ROT         | 3.25  | ROT         | 3.02        |
| LogD  | 2.14  | NATOM       | 2.57  | NATOM       | 2.58        |
| NATOM   | 1.94  | <b>LogD</b> | 2.54  | <b>LogP</b> | <b>2.06</b> |
| NRING   | 1.79  | NRING       | 1.81  | NRING       | 1.95        |
| <b>LogP</b>                                       | 1.53  | Hdonors     | 1.48  | Hdonors     | 1.43        |
| Hdonors   | 1.26  |             |       |             |             |

## REFERENCES

- Beck, M. W. (2018). Visualisation and Analysis Tools for Neural Networks. Package 'NeuralNetTools'. <https://cran.r-project.org/web/packages/NeuralNetTools/NeuralNetTools.pdf>
- Bhal, S. K., Kassam, K., Peirson, I. G. & Pearl, G. M. (2007). The rule of five revisited: applying log D in place of log P in drug-likeness filters. *Molecular Pharmaceutics*, 4(4), 556-560.
- Doak B. C. and Kihlberg J. (2017) Drug discovery beyond the rule of 5 - Opportunities and challenges. *Expert Opinion on Drug Discovery*, 12:2, 115-119.
- Doak, B. C., Zheng, J., Dobritsch, D. and Kihlberg, J. (2016). How beyond rule of 5 drugs and clinical candidates bind to their targets. *Journal of Medicinal Chemistry*, 59(6), 2312-2327.
- Fraley, C., Raftery, A. E., Murphy, B. and Scrucca, L. (2012). Mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation: University of Washington.
- Gashaw, I., Ellinghaus, P., Sommer, A. and Asadullah, K. (2011), What makes a good drug target? *Drug Discovery Today*, 16, 1037-1043.
- Guan, L., Yang, H., Cai, Y., Sun, L., Di, P., Li, W., Liu, G. and Tang, Y. (2019) ADMET-score—a comprehensive scoring function for evaluation of chemical drug-likeness, *Med Chem Comm*, 10(1), 148-157.
- Hudson, I.L., Leemaqz, S.Y., Shafi, D. and Abell, A.D. 2017. Score function of violations and best cutpoint to identify druggable molecules and associated disease targets. In Syme, G., Hatton MacDonald, D., Fulton, B. and Piantadosi, J. (eds) MODSIM2017, 22nd International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2017, pp. 487-393.
- Hudson, I. L., Leemaqz, S. Y., Neffe, A. T. and Abell, A. D. (2016). Classifying calpain inhibitors for the treatment of cataracts: A self organising map (som) ANN//KM approach in drug discovery. In S. Shanmuganathan & S. Samarasinghe (Eds.), *Artificial neural network modelling* (pp. 161-212). Springer International Publishing.
- Hudson, I. L., Shafi, S. and Abell, A. (2014). Drug-likeness: statistical tools, chemico-biology space, cartesian planes, drug databases: a case study. Paper presented at the Sixth Annual ASEARC Conference, February 2014. University of Wollongong, Australia.
- Kandoi, G., Acencio, M. L. and Lemke, N. (2015) Prediction of druggable proteins using machine learning and systems biology: a mini-review. *Frontiers in Physiology*, 6, 366.
- Kuhn, M., Wing, J., et al., (2018). Classification and regression training. 'Caret Package'. <https://cran.r-project.org/web/packages/caret/caret.pdf>
- Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today*. 20(3):318-31.
- Law, V., Knox, C., and Wishart, D. S. (2014). Drugbank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Research*, 42(D1), D1091-D1097. doi: 10.1093/nar/gkt1068.
- Leemaqz, S.Y., Hudson I.L., Abell A. (2018). Classification of disease targets to identify druggable molecules by score function of violations. The 11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2018), University of Pisa, Italy, 14-16 December 2018.
- Lipinski, C. A. (2016). Rule of five in 2015 and beyond: Target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions. *Advanced Drug Delivery Reviews*, 101, 34-41.
- Matsson, P., Doak, B. C., Over, B. and Kihlberg, J. (2016). Cell permeability beyond the rule of 5. *Advanced Drug Delivery Reviews*, 101, 42-61. doi: 10.1016/j.addr.2016.03.013.
- Mignani, S., Rodrigues, J., et al, (2018) Present drug-likeness filters in medicinal chemistry during the hit and lead optimization process: how far can they be simplified? *Drug Discovery Today*, 23(3), 605-615.
- Sachs, M. C. (2018). Generate ROC curve charts for print and interactive use. <https://cran.r-project.org/web/packages/plotROC/vignettes/examples.html>
- Ursu, O., Holmes, J., Knockel, J., Bologa, C. G., Yang, J. J., Mathias, S. L., Nelson, S. J. and Oprea, T. I. (2017). Drugcentral: Online drug compendium. *Nucleic Acids Research*, 45(D1), D932-D939. doi: 10.1093/nar/gkw993.
- Zafar, S., Hudson, I.L., Beh, E.J., Hudson, S.A. and Abell, A. (2016). A non-iterative approach for ordinal log-linear models: investigation of logD in estimating drug-likeness. In 31st International Workshop on Simulation and Modeling, Institute National des Sciences Appliquees. Rennes, France, July 4-8, Volume II, pages 163-166.
- Zafar, S., Cheema, S.A., Beh, E.J., Hudson, I.L., Hudson, S.A. and Abell, A. (2013). Linking ordinal log-linear models with Correspondence Analysis: an application to estimating drug-likeness in the drug discovery process. In Piantadosi, J., Anderssen, R.S. and Boland J. (eds) MODSIM2013, 20th International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2013, pp. 1945-1951.