Score function of violations and best cutpoint to identify druggable molecules and associated disease targets

I.L. Hudson [a], **S.Y. Leemaqz** [b], **S. Shafi** [c] **and A.D. Abell** [d]

[a] *Dept. of Statistics, Data Science & Epidemiology, Swinburne University of Technology, Melbourne*
[b] *Adelaide Medical School, The University of Adelaide,*
[c] *School of Mathematical and Physical Sciences, University of Newcastle, NSW,*
[d] *Department of Chemistry, The University of Adelaide, SA*
Email: lhudson@swin.edu.au

Abstract: Predicting druggability and prioritising certain disease modifying targets is critical in drug discovery. Expanding the spectrum of disease-relevant targets to pharmacological manipulation is vital to reducing morbidity and mortality. We test a druggability rule, based on 10 molecular parameters (scores counting violations, denoted by score10), which uses cutpoints for each molecular parameter based on mixture clustering discriminant analysis (MC/DA) (Hudson et al., 2014). A total of 1279 small molecules from the DrugBank chem-informatics database (Knox et al., 2011), combining detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with drug disease target information, were analysed and these were shown to be aligned with 173 targets. The score10 function comprised 4 traditional parameters of the rule of five (Ro5) (Lipinski, 2016), plus 5 extra parameters (polar surface area PSA, number of rotatable bonds, rings and halogens, N and O atoms) with an extra candidate of lipophicility, log D (the distribution coefficient) recently suggested by Bhal et al., 2007 as a possible preferable predictor for permeation (Zafar, Hudson et al., 2016, 2013;) to Lipinski's traditional partition coefficient, Log P, a predictor for permeation.

Multivariate skew normal (SN) (Lee and Mc Lachlan 2013) and Gaussian (MN) mixture clustering identified 5 molecule groups based on the 10 predictors, or 9 predictors when the number of halogen atoms was omitted. MN clusters were highly differentiable with 3 of the 5 clusters classified as poor druggable candidates, similarly the SN clusters. Logistic regression was used to determine the best cutpoint, C, for the total number of violations, score10 (< C versus greater or equal to C, for C= 3, 4 or 5) using predictor models containing the molecule's Ro5 status (if Ro5 compliant the molecule is druggable by Lipinski's rule), oral status, and poor vs good druggability grouping based on the clustering. We studied the performance of a support vector machine (SVM) and Recursive partitioning (RP) based on the 10 molecular descriptors, to classify compounds with high or low violator scores (defined by our optimal cutpoint, C). RP was applied to find simple hierarchical rules to classify the high score violators from the low (< C). PRoC analyses (Robin et al., 2011) and logit analyses showed that a cutpoint of 5 is best in partitioning chemo-space. For either partition of the score10 function, logistic models with the MN10 cluster predictor were superior to that of the (SN10). The best model was obtained for a cutpoint of 5 (AIC = 1403.79) and established that molecules with 5 or more violations tended to be non-oral candidates ($p < 0.00001$), MN10 poor ($p < 0.00001$) and be Ro5 violators ($p < 0.00001$), with a significant oral by cluster interaction ($P < 0.03$) found. The SVM classifier of the score10 partition (C=5) gave a Matthews coefficient C= 0.887. PROC analyses gave high values for the area under the curve (AUC) of 98.7%, with 95% CI (98.2%-99.3%), sensitivity (r) and specificity (s), 0.961 and 0.924, respectively for the training set. For the validation set SVM gave an AUC of 98.1%, 95% CI (97%-99.2%), r=0.927, s=0.983 and likewise a high C=0.818. The RP classification gave similar but slightly lower AUC and C values as the SVM. Specifically, the RP classifier for the score10 partition yielded an AUC of 95.1% with 95%CI (93.8%-96.4%), sensitivity of 0.918, specificity 0.936, and C= 0.845 for the training set; for the validation set an AUC of 95.3% with 95% CI (93.1%-97.5%), with r=0.924, s=0.886 and C=0.809. The RP rules to classify the high score violators from the low (< 5) confirmed the value of log D's inclusion in the scoring function and supported the original MC/DA cutpoints established for each molecular descriptor (Hudson et al., 2014). Our work illustrated that SVM used in combination with simple molecular descriptors can provide a reliable assessment of our simple scoring function of counts of violations partition. Moreover, molecules with score10 representing 5 or more violations were shown to be associated with specific disease targets, namely, Anti-Bacterial, Antineoplastic, Antihypertensive and Anti-allergic, within which most of the drugs have a non-oral delivery mode. Target drugs with a median score10 < 5 were Adrenergic, Dietary, Analgesics, Anti-infective, Anesthetics, Adjuvants, Anti-convulsants, Antimetabolites and Antidepressants, all of which, except Dietary and Anesthetics, were non-oral.

Keywords: *Druggability rules, Beyond the Rule of Five (bRo5), disease targets, machine learning*

1. INTRODUCTION

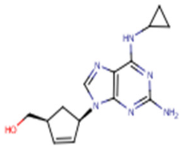
High throughput docking of small molecule ligands (candidate drugs) into high resolution protein structures is now standard in computational approaches to drug discovery (Ursu et al., 2017, 2011). Further details are in Hudson et al. (2016), who investigated Self Organising Map (SOM) artificial neural network procedures as a computational tool for the evaluation of docking experiments of calpain ligands (small drug molecules) for the treatment of cataracts. Predicting druggability and prioritising certain disease modifying targets for drug development is of high practical relevance in pharmaceutical research (Rask-Anderson et al., 2011, Lavecchia et al., 2015). Druggability predictions are important to avoid intractable targets and to focus drug discovery research on sites with preferable prospects and mortality (Lazo & Sharlow, 2016). Recently many targets have been classified as "undruggable" due to their lack of oral bioavailability. 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 Ro5 space, to the so-called beyond Ro5 (bRo5) space, noting that the Ro5 delineated cutpoints for Lipinski's traditional parameters (multiples of 5) by which if 1 violation occurred among the 4 parameters (highlighted in Table 2) the molecule was assumed Ro5 non-compliant. Recent drug studies suggest that cell permeable and orally bioavailable drugs have been discovered far into bRo5 space (Mattson et al., 2016), which affords significantly more possibilities for orally bioavailable and cell permeable compounds (Doak et al., 2016). Too strict an implementation of the Ro5 may well have hampered the pharmaceutical industry in regard to finding novel and more difficult drugs as well as more conventional drug targets as the cutpoints of the first 3 of Lipinski parameters rules were rather lower (Table 2). In this paper we tested a druggability rule based on 10 molecular parameters (scores counting violations, denoted score10) which uses cutpoints for each parameter based on mixture clustering discriminant analysis (MC/DA) (Hudson et al., 2014). The score10 function comprised the 4 traditional parameters of rule of five (Ro5) highlighted in Table 2, plus 5 extra parameters (PSA, number of rotatable bonds, rings, halogens, N and O atoms) with an extra candidate of permeability, log D (the distribution coefficient), suggested recently (Bhal et al., 2007) as a preferable predictor to Lipinski's partition coefficient, Log P.

2. DATA AND METHODS

2.1 Data and Druggability Scoring Rules

We analysed 1279 small molecules from the DrugBank database (Knox et al. 2011), a unique bioinformatics and chem-informatics resource combining detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with drug target (i.e. sequence, structure, pathway) information on 6,711 drug entries, and 1441 FDA-approved small molecule drugs (one candidate molecule example is shown in Table 1). Of the 1279 candidates there are 105 Ro5 violators, 681 and 598 with oral and non-oral delivery modes, respectively.

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 136470-78-5	286.3323 $C_{14}H_{18}N_6O$		Anti-HIV Agents / Nucleoside and Nucleotide Reverse Transcriptase Inhibitors / Reverse Transcriptase Inhibitors	For the treatment of HIV-1 infection, in combination with other antiretroviral agents.

Recently, Hudson et al. (2014) developed druggability rules (scores counting violations) which use a new cutpoint for each molecular parameter based on a mixture clustering (mclust) (Fraley et al., 2012) discriminant analysis (MC/DA) approach. Table 2 shows that the cut-off values of the Ro5 (Lipinski, 2016), of Veber et al. (2002) and of Hudson et al. (2014) are not in agreement, particularly for MW, PSA and log P. However, Hudson et al.'s cutpoints are much in agreement with recent cutpoints of Ursu et al. (2017), in particular for MW, log P and PSA (refer to columns 4-6 of Table 2). Note the cutpoint for Log D of 3.5 (Hudson et al., 2014) is smaller than 5.5 suggested earlier by Bhal et al. (2007). Work by Zafar et al, (2013, 2016) also found the pH dependent version of permeability, log D at intestinal pH (log D ~ 3.5), superior to the classic parameter log P.

Table 2. Cutpoints for violation of physicochemical parameters: orally formulated drug subset of Drugbank

Property	Lipinski Ro5	Veber et al., 2002	Hudson et al., 2014	Ursu et al., 2017 Median	Ursu et al., 2017 3rd Quartile	Bhal et al., 2007
Molecular weight (MW) (Daltons, Da)	≤ 500	≤ 700	≤ 305	≤ 324.8	≤ 411.5	
Log P	≤ 5	≤ 9	≤ 1.9	≤ 2.43	≤ 4.05	
Hydrogen Bond (HB) donors (HBd)	≤ 5		≤ 4	≤ 1	≤ 2	
HB acceptors (HBa)	≤ 10		≤ 7	≤ 5	≤ 7	
HBd+ HBa		≤ 12	≤ 11	≤ 6	≤ 9	
Polar surface area (PSA) (Å^2 angstroms)		≤ 140	≤ 65	≤ 67.2	≤ 98.8	
Number ROT BONDS		≤ 10	≤ 7	≤ 4	≤ 7	
Number of N, O atoms			≤ 40			
Ring number			≤ 2			
Halogens			≤ 2			
Log D			≤ 3.5			≤ 5.5

2.3 Statistical Methods

Multivariate skew normal (SN) (Lee and Mc Lachlan 2014) and Gaussian (MN) mixture clustering were used to identify molecule groups based on the 10 predictors and to differentiate so-called good versus poor druggable candidates as in Hudson et al., (2014). Logistic regression was performed to determine the best cutpoint, C, for the total number of violations, score10 (< C versus greater or equal to C, for C = 3, 4 or 5). These thresholds for C were suggested by preliminary trends in clustering of Hudson et al., (2014). The logit predictor models analysed here contain the molecule's Ro5 status (if Ro5 compliant, then druggable by Lipinski's rule), oral status, and poor *vs* good druggability based on the MN or SN clustering. Support vector machine (SVM) and Recursive partitioning (RP) based on the above 10 molecular descriptors was then used to classify compounds according to the partition - high or low Score10 values as defined by the optimal cutpoint. The data set was divided into a 959-molecule training set and a 320-molecule test set. RP was used to find simple hierarchical rules to classify the high score10 violators from the low (< C).

3. RESULTS

3.1. Logit analysis

Multivariate skew normal (SN) (Lee & Mc Lachlan 2014) and Gaussian (MN) mixture clustering identified 5 molecular groups based on the 10 predictors, or 7 clusters when based on 9 predictors, when the number of halogen atoms (h) was omitted (MN10-h, SN10-h models). The MN10 groups were highly differentiable with 3 of the 5 obtained clusters classified as poor druggable candidates (in total these 3 clusters contain 350 molecules). For either partition of the score10 function, with and without the number of halogen atoms included as a predictor of the clusters, logit models containing the MN10 cluster predictor were superior to the skew normal SN10 based models. The best MN10 model showed that molecules with 5 or more violations (AIC = 1403.79) tended to be non-oral candidates (p < 0.00001), MN10 poor (p < 0.00001) and be Ro5 violators (p < 0.00001) (Table 3). This optimal MN10 model gave a significant oral by cluster interaction (p < 0.02), in that for Ro5 compliant molecules, the predicted probability of scoring high (≥5) for the MN10 good molecules was 0.328 and 0.477, for oral and non-oral molecules, respectively. The probability of scoring ≥ 5 for the MN10 poor cluster was 0.750 and 0.757 for oral and non-oral Ro5 compliant molecules. For the Ro5 violators the predicted probability of scoring high (≥5) was 0.999 for both the poor and good MN10 classes, irrespective of their oral status. For the MN10 model, in terms of odds, the odds of a ligand scoring ≥ 5 increased by 0.527 for a ligand classified as non-oral and in the MN10 good cluster (with Ro5 status fixed). The odds of a molecule being a high scorer ≥5 increased by 3.014 if the molecule is classed as MN10 poor and oral (with Ro5 status fixed). The odds of a molecule being in the high score10 group increased by 3.13 if it is non oral and in the poor MN10 cluster (with Ro5 status fixed). The odds of a molecule being a high score violator increases by 94.54 if it is Ro5 noncompliant (with oral status and MN10 clustering fixed). Note the SN10 based logit models (where 433 molecules were classified as poor, based on k=7 SN10 clusters) did not demonstrate a significant oral by cluster interaction (not reported here). PRoC analyses (Robin et al., 2011) confirmed that a cutpoint of

5 for score10 is better than a cut-off of 4 to partition chemo-space (not reported here due to space restrictions), this optimal C of 5, was true for both the MN10 and SN10 based models, in agreement with the logit analysis.

Table 3. MN10 and SN10 based logistic models for Score10. NS denotes "not significant".

	C	MN10 included as a predictor				SN10 included as a predictor		
			β coefficient	S.E	p-value	β coefficient	S.E	p-value
Score10	**< 5**	Non-oral	-0.640	0.1617	< 0.00001	-0.418	0.1288	< 0.002
		Ro5-	4.549	1.0111	< 0.000001	4.589	1.0131	< 0.000001
		Poor cluster	1.103	0.1891	< 0.0000001	1.248	0.1756	< 0.00000001
		Oral*Cluster	0.680	0.2849	< 0.02	-	-	NS

3.2. Disease target results

In this study, 173 targets for 1279 small molecules were retrieved from the data and the median score10 value obtained for the molecules in a given target group, where drug targets were associated with at least one FDA approved small drug. For example, if several molecules (say 40) are associated with one target, then the median values of score10 and of each of the 10 molecular descriptors that are used to evaluate score10 for 40 drugs were evaluated for that particular target. Of the 173 targets studied, 99 targets were predominantly oral and 74 non-oral in terms of modes of delivery. In this paper we report the top 12 targets which contain 25 or more molecules. Table 4 gives the disease target name for the top 12 targets, their associated number of molecules, median score10 value, overall oral status, along with the number of oral molecular candidates in the given target. For the twelve most representative targets to which 579 small molecules belong, Figure 1 shows the median values of 4 molecular descriptors, namely, molecular weight (MW) and log P of Lipinski (2016), Polar Surface Area (PSA) of Veber et al (2002); and log D of Bhal et al., (2007), latter used to compare with classical log P candidate for permeability. The targets with a median score10 of 5 or more (C= 5), were Anti-Bacterial, Antineoplastic, Antihypertensive and Anti-allergic, within which most of the drugs have a non-oral delivery mode (Table 4). Figure 1 displays the medians of Log P, MW, PSA and Log D across the top 12 disease targets.

Table 4. Description of the top 12 targets in terms of Score10 and oral status.

Target	Total no. (n) of molecules	Median Score10	Score10 status	Number of oral molecules	Oral status
Adjuvant	25	2	Low score	14	Oral
Anti-anxiety	26	4	Low score	20	Oral
Antihypertensive	35	5	Violator	17	Non-oral
Dietary	40	1	Low score	11	Non-oral
Antineoplastic	57	5	Violator	28	Non-oral
Anti-inflammatory	61	4	Low score	40	Non-oral
Anti-Infective	40	1.5	Low score	23	Oral
Anti-Bacterial	92	6	Violator	42	Non-oral
Anesthetics	30	2	Low score	6	Non-oral
Analgesic	48	3	Low score	30	Oral
Adrenergic	98	3	Low score	62	Oral
Anti-Allergic	27	5	Violator	12	Non-oral

Table 5. Description of the target specific median of molecular parameters, score 10 and oral status for 6 targets

Targets (score10)	MW	Log P	logD	HBa	HBd	PSA	Rot	Natoms/ Nrings	Oral status	(n)
Anti-bacterial (6)	453.66	0.20	-1.49	10	3.5	180.82	5	51 (4)	Non-	92
Adrenergic (3)	278.9	2.08	1.8	4	2	66.02	4	42 (2)	Oral	98
Analgesic (3)	285.33	2.3	2.41	3	1	49.55	3.5	43 (3)	Oral	48
Antidepressants (3)	298.36	2.95	2.66	4	1	38.88	4	40 (2)	Oral	16
Antimetabolites (3)	244.2	-1.0	-0.75	7	3	111.18	2	29 (2)	Oral	21
Anticonvulsants (1)	218.25	0.9	1.11	3	1	63.4	2	29 (2)	Oral	24

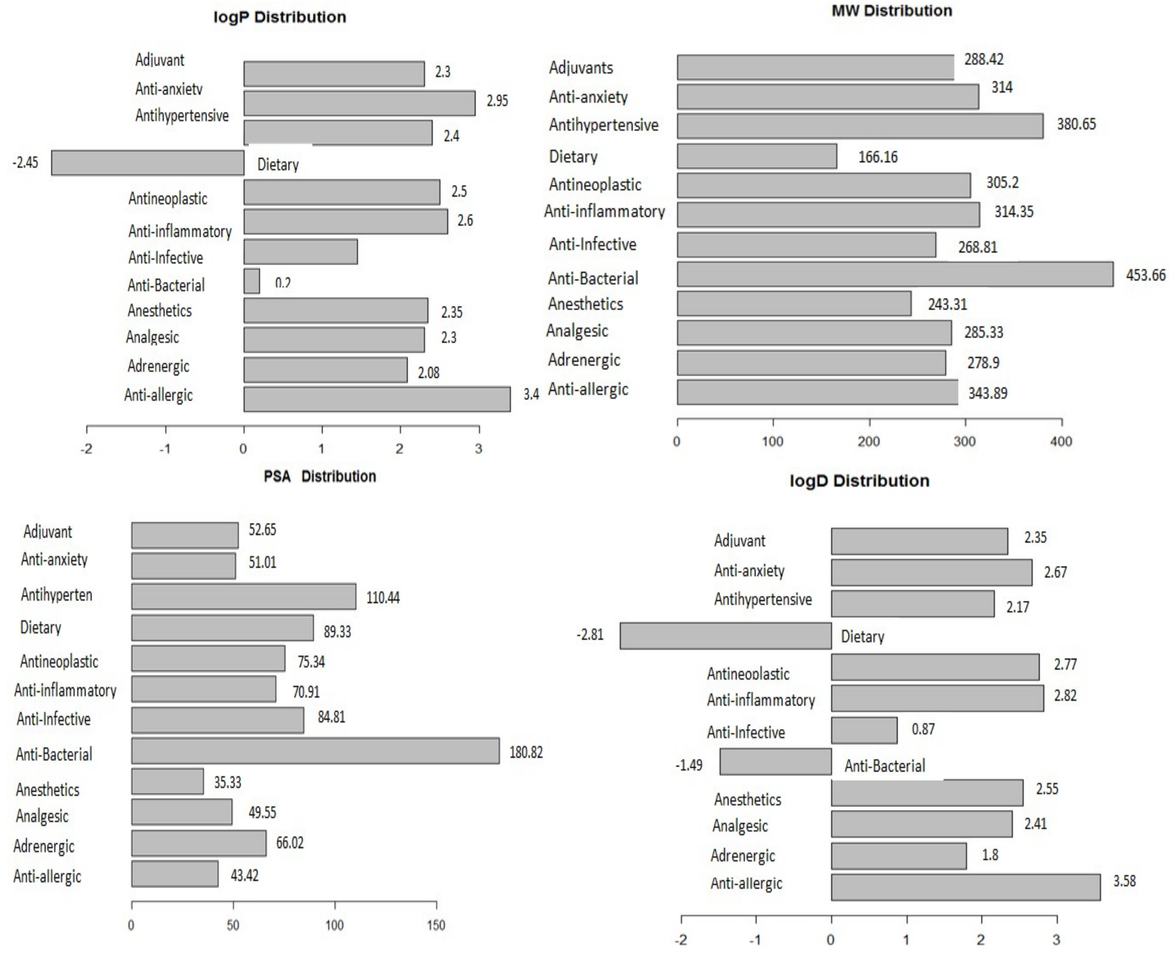


Figure1. Pattern of Log P, MW, PSA and Log D across the top twelve targets.

The pattern of molecular profiles of these 12 targets is shown in Figure 1. For example, the Anti-Bacterial target drugs (with a median score10 of 6) were large in size (median MW = 453.66 daltons) compared to other targets and particularly compared to the Dietary target (median MW = 166.16 daltons). The median values of 3 descriptors representing hydrogen bond donors and acceptors (HBd, HBa)) or electrostatic features, HBa, HBd and PSA, were also high for the Anti-Bacterial target (10, 3.5 and 180.82, respectively) (Table 4-5). In addition, the median values for the Antihypertensive (non-oral) target drugs were high for HBa (6) and PSA (110.4) as was the median score10 of 5 (Fig. 1). The Dietary target had the lowest median score10, median MW and negative medians for both log P = -2.45 and log D = -2.81 (Table 5). Notably most drugs of these 3 targets (Anti-bacterial, Dietary, Antihypertensive) were non-oral but Ro5 compliant.

The Antimetabolites target had a low median score10 of 1, low median MW = 244.2, but high PSA = 111.18, and negative median values for log P = -1.0 and log D = -0.75. Negative median values indicate higher hydrophobicity for both the Dietary and Antimetabolites target drugs. Noteworthy, the median values of log P = 3.4 and log D = 3.58 for the Anti-allergic target drugs (non-oral, with a median score10 of 5) were higher than the other 11 target groups (Figure 1) evidencing that the anti-allergic drugs may prefer not to transport hydrophobic molecules; followed then by the Antihypertensive and Anti-inflammatory targets (non-oral) with respective median score10 of 5 and 4. Interestingly, median Log D for Antibacterial target drugs was negative -1.49 versus 0.20 for log P, indicating that log D and log P contain different information as suggested by Bhal et al., (2007).

3.3. SVM and RP results

The SVM classifier for our score10 partition (with cutpoint 5) yielded a Matthews coefficient C= 0.887. ROC analyses gave high values for the area under the curve (AUC) of 98.7%, with 95% CI (98.2%-99.3%), sensitivity (r) and specificity (s), 0.961 and 0.924, respectively for the training set. For the validation set SVM gave an AUC of 98.1%, 95% CI (97%-99.2%), r=0.927, s=0.983 and likewise a high C=0.818. The RP classification gave similar but slightly lower AUC and C values than the SVM. Specifically, the RP classifier for the score10 partition yielded an AUC of 95.1% with 95%CI (93.8%-96.4%), sensitivity of 0.918, specificity 0.936, and C= 0.845 for the training set; for the validation set an AUC of 95.3% and 95% CI (93.1%-97.5%), with r=0.924, s=0.886 and C=0.809 was obtained. The multivariate RP rules obtained to classify the high score violators from the low (< 5) confirms the univariate MC/DA cutpoints of Hudson (compare Figure 2 with Table 2). The numbers and labels inside the rounded rectangles delineate the number at each recursive partitioning step that satisfy or not the node cutpoint , the ligands are then split into the good and poor partitions based on score 10 at the next recursive step. For example in the validation set of 320 ligands with the partition good C < 5, *vs* poor C > 5) 197 satisfy MW < 342 and are classed good (RHS Fig. 2), and 123 plased in the poor partition, for which MW exceeds 342. Of the 197 good ligands, 183 satisfy MW < 306 - at the next step, these are classed as good, of the 27 remaining molecules (183-156= 27) molecules (those with log D < 4.3 are in good score10 partition, 13 with log D > 4.3 are classed as poor. Note 156+27 = 183 and 1 + 13 = 14. The RP classification tree also supports the value of using log D < 3.5 (< 4.3) (left hand arms of Fig. 2 for the good ligands) and note that Log D < 0.066 (in the training set, right hand arm, poor ligands, Figure 2) captures negative Log D, in the Anti-bacterial and Dietary target values for Log D or Log P (Figure 1).

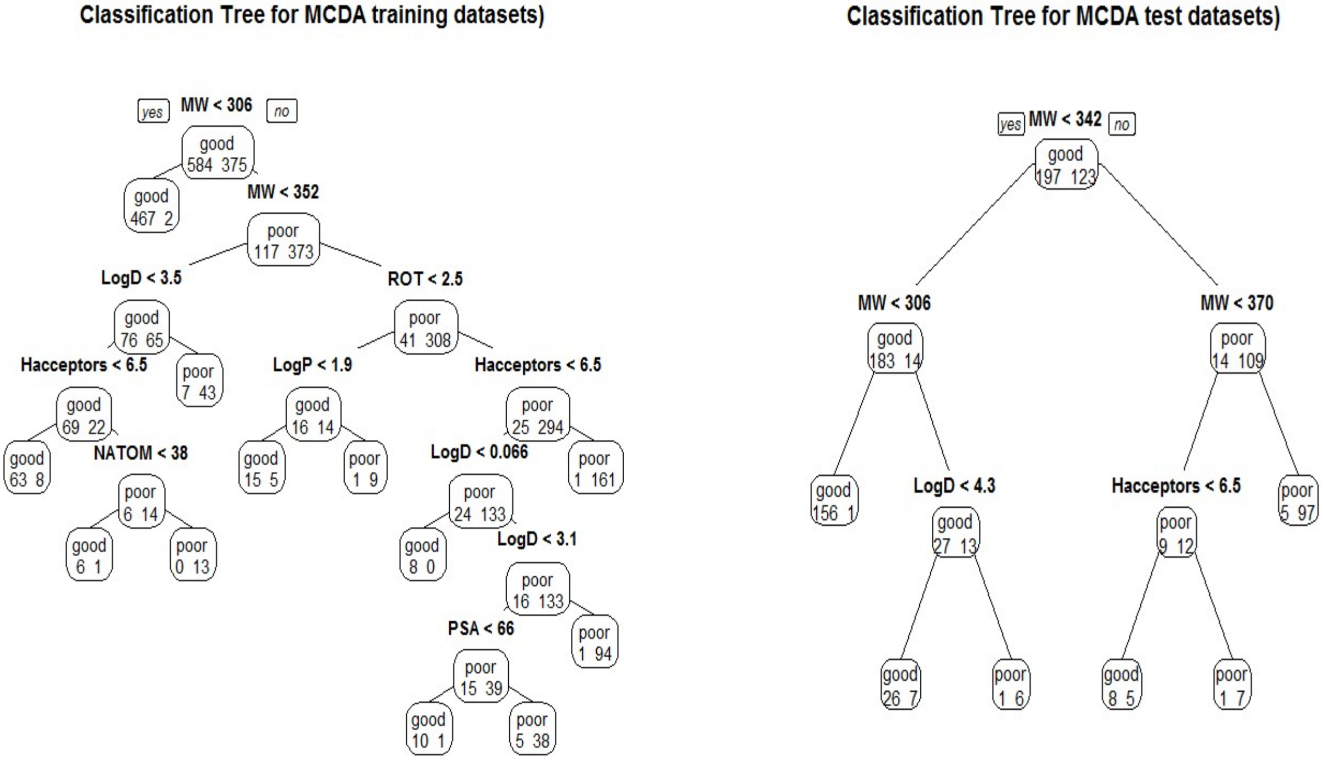


Figure 2. RP Classification Tree for partitions based on Score 10 with cutpoint 5.

4. DISCUSSION AND CONCLUSIONS

Our work illustrates that a simple scoring function of counts of violations can partition chemospace and help identify both good and poor druggable molecules, and associated targets. Moreover, ligands with 5 or more violations, based on adding subcomponents of score10, were shown to be associated with specific disease targets. The Anti-Bacterial target drugs (median score10 = 6) were found to have high values for all 10 molecular parameters, consistent with the results of Giordanetto et al., (2014), who reported that compounds that fall into the bRo5 space (higher MW and PSA) include the Anti-bacterial target. In contrast Dietary target drugs had low values for the 10 descriptors, lowest median score10 of 1, lowest median MW and also negative medians for both log P = -2.45 and log D = -2.81. Further work that aims to evaluate which of the Log D and Log P best reflects permeability will test different score functions, based on 9 parameters, which respectively omit Log P (score9D) and log D (score9P), and find associated best score cutpoints (Hudson et al., *in prep*). Recently log P's association with MW and PSA was shown to change magnitude/sign according to the molecule's Lipinski's Ro5, not so for log D (Zafar et al., 2016; 2013). Consequently future work will test for

Log P and Log D's association with the remaining ligand parameters, according to the new strata based on our score10 partition (< 5 vs ≥ 5) in this paper and in regard to new partitions based on 9 parameters. Ongoing work using mixture clustering of the target-specific medians of the 10 molecular parameters aims to help identify so-called poor and good targets. How these targets correlate with new partitions based on 9 molecular parameters is a future research topic (Hudson, Shafi et al., *in prep*).

REFERENCES

Bhal, S. K., Kassam, K., Peirson, I. G. and 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., Zheng, J., Dobritzsch, 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.

Finan, C., Gaulton, A., Kruger, F. A., Lumbers, R. T., Shah, T., Engmann, J., Galver, L., Kelley, R., Karlsson, A., Santos, R., Overington, J. P., Hingorani, A. D. and Casas, J. P. (2017). The druggable genome and support for target identification and validation in drug development. *Sci Transl Med,* 9(383), eaag1166.

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: Dept Statistics, University of Washington.

Giordanetto, F. and Kihlberg, J. (2014). Macrocyclic drugs and clinical candidates: What can medicinal chemists learn from their properties? *Journal of Medicinal Chemistry,* 57(2), 278-295.

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. Shanmu-ganathan & 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, Feb 2014. University of Wollongong, Australia.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C. and Wishart, D. S. (2011). Drugbank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research,* 39, D1035-D1041. doi: 10.1093/nar/gkq1126. Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today.* 20(3):318-31.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y. F., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z. T., Han, B. S., Zhou, Y. and Wishart, D. S. (2014). Drugbank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Research,* 42(D1), D1091-D1097.

Lazo, J. S. and Sharlow, E. R. (2016). Drugging undruggable molecular cancer targets. *Annual Review of Pharmacology and Toxicology, Vol 56,* 56, 23-40. doi: 10.1146/annurev-pharmtox-010715-103440.

Lee, S. and McLachlan, G. J. (2014). Finite mixtures of multivariate skew t-distributions: Some recent and new results. *Statistics and Computing,* 24(2), 181-202.

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, 3441.

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.

Rask-Andersen, M., Masuram, S. and Schioth, H. B. (2014). The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annual Review of Pharmacology and Toxicology, Vol 54,* 54, 9-26. doi: 10.1146/annurev-pharmtox-011613-135943.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., et al. (2011). pROC: An open-source package for R and S plus to analyze and compare roc curves. *BMC Bioinformatics,* 12. doi: Artn 7710.1186/1471-2105-12-77.

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 Res,* 45(D1), D932-D939. doi: 10.1093/nar/gkw993.

Veber, D.F., Johnson, S.R., *et al.* (2002). Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem*. 45(12): 2615–2623.

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.