

# Linking ordinal log-linear models with Correspondence Analysis: an application to estimating drug-likeness in the drug discovery process

S. Zafar<sup>a</sup>, S. A. Cheema<sup>a</sup>, E. J. Beh<sup>a</sup>, I. L. Hudson<sup>a</sup>, S. A. Hudson<sup>b</sup>, A. D. Abell<sup>c</sup>

<sup>a</sup> School of Mathematical and Physical Sciences, University of Newcastle, Newcastle, Australia.

<sup>b</sup> Department of Pharmaceutical Chemistry and Cellular & Molecular Pharmacology, University of California San Francisco (UCSF), California, USA. <sup>c</sup> School of Chemistry and Physics, Adelaide University, Adelaide, South Australia

Email: [Sidra.Zafar@uon.edu.au](mailto:Sidra.Zafar@uon.edu.au)

**Abstract:** Ordinal log-linear models (OLLM's) are amid the most widely used and powerful techniques to model association among ordinal variables in categorical data analysis. The parameters of such models are traditionally estimated using iterative algorithms, such as the Newton-Raphson method and iterative proportional fitting. More recent advances involve a non-iterative estimation method that performs equally well for estimation of the linear-by-linear association in OLLM's for a two-way table. This paper establishes a link between the Beh-Davy non-iterative estimation method (BDNI) (Beh & Davy, 2003) and the well-known ordinal correspondence analysis (CA) technique for two dimensional tables. The BDNI estimator of association relies on orthogonal polynomials (OP's), an approach dating from Lancaster (1953) to Beh and Davy (2003). OP's provide insight into the origin and development of non-iterative estimation in OLLM's, as an alternative to popular iterative procedures. The main advantage of OP's is that the resultant parameters enable estimation of the linear, and also quadratic and higher order association structures amongst the ordered categories. Ordinal CA was first introduced by Beh (1997). We compare the linear-by-linear BDNI association procedure with the linear-by-linear association method depicted via graphical representation in ordinal CA. To demonstrate this link and theory we analyzed the relationships between predictors of drug-likeness used in drug discovery to filter out small molecule (drugs) that may fail clinical trials. In vitro absorption, distribution, metabolism and elimination (ADME) assays are now being conducted throughout the drug discovery process, from hit to lead optimization (Kerns & Li, 2008). The analytical community needs still to develop faster and better analytic methods to enhance the 'developability' of drug leads, and to formalize strategies for ADME assessment of candidates in the discovery and pre-clinical stages (Kassel 2004). Assessing drug-likeness depends on the nature of relationships between surrogate measures of drug-likeness (aqueous solubility, permeability) and physicochemical properties (lipophilicity, molecular weight (MW)). To date, lipophilicity is expressed quantitatively as logP, the most popular predictor for permeation. We apply our methods to test rules of druggability (Lipinski 2000). In this study 1,279 small molecules from Hudson *et al.* (2012), based on the DrugBank3.0 database (Knox *et al.*, 2011), a unique chem-informatics resource are analysed. The pair-wise association between categorised variants of 2 of the 4 traditional parameters of Lipinski's rule of five (Ro5), namely MW and logP, and an additional parameter, polar surface area (PSA), introduced by Veber *et al.*, (2002), are shown to differ in magnitude or swap sign across strata, where strata are defined by a molecule's druggable (Ro5 compliant) versus non-druggable (Ro5 violation) status. Log P's association with MW, assumed to be positive, is shown to: [1] change sign from significantly negative to positive for non-druggable vs druggable strata, when data is tertiled within the stratum and the first level category (0) satisfies the new cutpoints for violation developed by Hudson *et al.*, (2012), i.e.  $\log P \leq 1.9$  and  $MW \leq 305$ , in contrast to Ro5's cutpoints of  $\log P \leq 5$  and  $MW \leq 500$ ; or [2] be lower (positive) for non-druggable vs druggable, for data stratified within quartiles. These findings support recent criticisms about using log P (Bhal *et al.*, 2007) in ADME assessment. Also PSA's association with MW, traditionally assumed to be positive, is shown to change sign from significantly negative to significantly positive for non-druggable vs druggable molecules, for data stratified within quartiles; with the first level category (0) satisfying the cutpoints for violation of Hudson *et al.*, (2012), i.e.  $PSA \leq 65$ ,  $MW \leq 305$ , in contrast to conventional cutpoints of 140 and 500, respectively. This study shows that assumed relationships between predictors need to be questioned. Log D, as a distribution coefficient (Bhal *et al.*, 2007), may be a better surrogate than log P.

**Keywords:** Ordinal log-linear model, Non-iterative estimation, linear-by-linear association parameter, ordinal correspondence analysis, Lipinski's rule of 5 (Ro5)

## 1. INTRODUCTION

Ordinal log-linear models (OLLM's) are one of the most widely used techniques for studying the association between ordinal categorical variables that are cross-classified to form a contingency table. The estimation of the parameters of these models has, traditionally, involved iterative procedures, such as Newton-Raphson algorithm and iterative proportional fitting (Agresti, 2010). A more direct method built on a non-iterative approach, in contrast to iterative algorithms, as proposed by Beh and Davy (2003) and developed by Beh and Farver (2009, 2012) can be adopted. These authors showed that their non-iterative method provides exceptionally stable and reliable estimates of the linear-by-linear association parameter, compared with Newton's algorithm. Recently Zafar *et al.* (2013) also showed that two non-iterative procedures, the Beh-Davy non-iterative estimate (BDNI) (Beh and Davy, 2003) and the Log non-iterative estimate (LogNI) give unbiased estimates. In this paper we shall confine our attention to the key non-iterative estimate, the BDNI estimate. The BDNI procedure relies on orthogonal polynomials, which allow the study of the linear, quadratic and higher order association structures among ordered categories. Rather than considering model based non-iterative techniques for quantifying the association, one may consider instead graphically summarising the association using ordinal CA (Beh, 1997). Ordinal CA applies to situations when one of the variables is ordered, or both row and column variables are ordered (single *vs* doubly ordered). Orthogonal polynomials underpin the development of both the non-iterative procedures and the ordinal CA approach. As such both methods elucidate the structure of one or more ordinal variables, and can identify any linear, quadratic and higher order association present in a two-way contingency table.

This paper defines the practical link between the BDNI procedure for OLLM's and ordinal CA, depicted by ordinal CA's graphical representation. We demonstrate the application of non-iterative methods and ordinal CA in establishing the nature of relationships between predictors of drug-likeness used by the drug discovery industry (Leesom, 2012). High throughput docking of small molecules (candidate drugs) into high resolution protein structures is now standard in computational drug discovery (Ursu *et al.*, 2011). Druggability predictions help the drug industry to avoid intractable targets and to identify superior sites. Predicting druggability and prioritising certain disease modifying targets is still a high priority in pharmaceutical research. Assessing drug-likeness depends on the nature of relationships between surrogate measures (aqueous solubility, permeability) and physicochemical properties (lipophilicity, molecular weight (MW)). Lipophilicity, expressed quantitatively as logP, to date, is the most popular predictor for permeation (across cell membranes). In this study the data of Hudson *et al.* (2012), namely, 1,279 small molecules from the DrugBank3.0 database (Knox *et al.*, 2011) are analysed. DrugBank3.0 is a unique chem-informatics resource with detailed drug (i.e. chemical, pharmacological and pharmaceutical) and drug target data (i.e. sequence, structure, and pathway). We evaluate the association relationships between bivariate pairs of categorised variants of 2 of the 4 traditional parameters of Lipinski's rule of five (Ro5) (Lipinski & Hopkins, 2004), namely MW and logP, and an additional parameter, polar surface area (PSA), introduced by Veber *et al.*, (2002). The aim of the study is to investigate differences in magnitude or swaps in sign of the pairwise associations across strata, where strata are defined by a molecule's druggable (Ro5 compliant) versus non-druggable (Ro5 violation) status. See Hudson *et al.* (2012) for a new druggability rule and scores.

## 2. MATHEMATICAL METHODS

### 2.1 ORDINAL LOG-LINEAR MODELS & BEH-DAVY NON-ITERATIVE ESTIMATION

For a doubly ordered  $I \times J$  contingency table,  $N$ , denote the proportion of individuals/units in the  $(i, j)$ th cell as  $p_{ij} = n_{ij}/n$  where  $n_{ij}$  is the  $(i, j)$ th cell value of  $N$ , for  $i = 1, 2, \dots, I$ , and  $j = 1, 2, \dots, J$ . Therefore  $\sum_{i=1}^I \sum_{j=1}^J p_{ij} = 1$ . Denote  $p_i$  and  $p_j$  as the marginal proportion of the  $i$ th row and  $j$ th column categories, respectively, such that  $\sum_{i=1}^I p_i = \sum_{j=1}^J p_j = 1$ . Moreover, let  $m_{ij}$  be the expected cell frequency of the  $(i, j)$ th cell. The OLLM for a doubly ordered contingency table is then defined as

$$\log m_{ij} = \mu + \alpha_i + \beta_j + \varphi(u_i - \bar{u})(v_j - \bar{v}), \quad (2.1)$$

where  $i = 1, 2, \dots, I$  and  $j = 1, 2, \dots, J$ ; refer to, for example, Agresti (2010). Here,  $u_i$  and  $v_j$  represent row and column monotonic scores for the categories of the row and column variable, respectively. Choose the scores as  $u_i = i$  and  $v_j = j$ . For the OLLM model (2.1),  $\mu$  is the grand mean,  $\alpha_i$  and  $\beta_j$  the main effects of  $i$ th row and  $j$ th column, respectively. The parameter of interest in this model is the linear-by-linear association parameter,  $\varphi$ . This parameter,  $\varphi$ , can be interpreted in terms of the local log-odds ratio as follows,

$$\log\left(\frac{m_{ij}m_{i+1,j+1}}{m_{i,j+1}m_{i+1,j}}\right) = \varphi(u_i - u_{i+1})(v_j - v_{j+1}). \tag{2.2}$$

Therefore, when natural scores are used to reflect the ordered structure of the two categorical variables, so that  $u_i - u_{i+1} = 1$  and  $v_j - v_{j+1} = 1$ , the local log-odds ratio is equal to  $\varphi$ . The estimation of the parameter of interest,  $\varphi$ , is conventionally performed using the iterative procedures, such as Newton’s uni-dimensional algorithm and iterative proportional fitting. However, the BDNI estimation method introduced by Beh and Davy (2003) and further studied by Beh and Farver (2009) can be used alternatively to estimate the linear-by-linear association parameter. The BDNI estimate relies on recurrence formulae to generate the OP’s; see Emerson (1968). These formulae produce polynomials that are akin to the Gram-Schmidt orthogonalisation procedure where the row, or column, scores are the basis vector and have been extensively considered in the analysis of association for contingency tables, see, for example Lancaster (1953). Therefore, by using OP’s, a more general OLLM than that of (2.1) is

$$\ln m_{ij} \approx \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \sum_{u=1}^{I-1} \sum_{v=1}^{J-1} a_u(i) \frac{z_{uv}}{\sqrt{n}} b_v(j), \tag{2.3}$$

being the saturated form of the OLLM (Beh and Davy, 2003). Here,  $a_u(i)$  is the  $u$ th order OP for the  $i$ th row category. Similarly,  $b_v(j)$  is the  $v$ th order OP for  $j$ th column category. Also

$$Z_{uv} = \sqrt{n} \sum_{i=1}^I \sum_{j=1}^J a_u(i) b_v(j) p_{ij}, \tag{2.4}$$

is the  $(u,v)$ th generalised correlation and the  $Z_{uv}$  are independently and identically standard normals, refer to Rayner and Beh (2009) for more details. The most commonly considered correlation, and the one that we shall confine our attention to, is the  $(1, 1)$ th correlation. The grand mean effect,  $\mu$ , row effect,  $\alpha_i$ , and the column effect,  $\beta_j$ , parameters in (2.3), are estimated as follows

$$\hat{\mu} = \ln n + \frac{1}{I} \sum_{i=1}^I \ln p_{i.} + \frac{1}{J} \sum_{j=1}^J \ln p_{.j}, \quad \hat{\alpha}_i = \ln p_{i.} - \frac{1}{I} \sum_{i=1}^I \ln p_{i.}, \quad \hat{\beta}_j = \ln p_{.j} - \frac{1}{J} \sum_{j=1}^J \ln p_{.j}. \tag{2.5}$$

Model (2.3) is dynamic in the sense that, unlike the usual OLLM, it can account for various parameters which reflect effects other than the linear-by-linear association parameter. Now setting  $\bar{u} = \sum_{i=1}^I p_{i.} u_i$ ,  $\bar{v} = \sum_{j=1}^J p_{.j} v_j$ ,  $\sigma_I = (\sum_{i=1}^I p_{i.} u_i^2 - (\sum_{i=1}^I p_{i.} u_i)^2)^{1/2}$ , and  $\sigma_J = (\sum_{j=1}^J p_{.j} v_j^2 - (\sum_{j=1}^J p_{.j} v_j)^2)^{1/2}$  is an unsaturated version of model (2.3), and with only the linear-by-linear association ( $u = 1, v = 1$ ), is then

$$\ln m_{ij} \approx \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \frac{z_{11}}{\sqrt{n}} \frac{(u_i - \bar{u})(v_j - \bar{v})}{\sigma_I \sigma_J}. \tag{2.6}$$

Comparing (2.1) and (2.6) the BDNI estimate is  $\hat{\varphi}_{BDNI} = \frac{1}{\sigma_I \sigma_J} \sum_{i=1}^I \sum_{j=1}^J p_{ij} (u_i - \bar{u})(v_j - \bar{v})$ . This estimate has the same interpretation as the OLLM parameter,  $\varphi$ , and is thus equivalent to the common log-odds ratio.

## 2.2 ORDINAL CORRESPONDENCE ANALYSIS

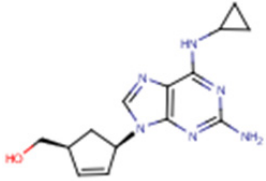
The ordinal CA was first introduced by Beh (1997). Unlike traditional CA of a two-way contingency table, ordinal CA accommodates the ordinal nature of the variables. In order to obtain a visual interpretation of the association between the ordered row and ordered column categories, the CA plot may be constructed using a variety of coordinate systems involving orthogonal polynomials and generalised correlations. The commonly adopted system involves the calculation of *profile coordinates* which, for the row and column categories along the  $m$ th dimension of this plot, may be obtained from the  $m$ ’th column,  $= \mathbf{A} \times \frac{\mathbf{Z}}{\sqrt{n}}$ ,  $\mathbf{G} = \mathbf{B} \times \frac{\mathbf{Z}^T}{\sqrt{n}}$ , respectively. Here,  $\mathbf{A}$  and  $\mathbf{B}$  are  $I \times (I-1)$  and  $J \times (J-1)$  column matrices consisting of row and column orthogonal polynomials.  $\mathbf{Z}$  is the  $(I-1) \times (J-1)$  matrix whose entries are (2.4). These profile coordinates can be used to reflect the total association that is present in the contingency table and is quantified by the chi-squared statistic divided by  $n$ , and is referred to as the total inertia of the table,  $\frac{\chi^2}{n} = \frac{\text{trace}(\mathbf{Z}^T \mathbf{Z})}{n} = \text{trace}(\mathbf{F}^T \mathbf{D}_I \mathbf{F}) = \text{trace}(\mathbf{G}^T \mathbf{D}_J \mathbf{G})$ , where,  $\mathbf{D}_I = \text{diag}(p_{i.})$  and  $\mathbf{D}_J = \text{diag}(p_{.j})$ . Therefore, like the BDNI estimate, the coordinates of the row and column categories along the first axis of a plot, obtained by performing an ordinal CA on the contingency table, rely on the magnitude and sign of  $Z_{uv}$ . Hence, as we shall demonstrate in section 4’s application, the BDNI estimation procedure and ordinal CA are interrelated. The difference

between the two techniques is as follows – the BDNI estimate, as are all estimates of the linear-by-linear parameter of an OLLM quantifies the magnitude and direction of the association between two categorical variables. Ordinal CA provides a visual inspection of this magnitude and direction and also provides the analyst with a means of identifying those row and column categories that impact on the estimate. For example, the magnitude of the BDNI estimate is visually reflected in an ordinal CA plot by identifying the distance of the row and column profile coordinates along the first axis from the origin; the origin is located where all the categories would be positioned if there was complete independence between the two ordinal categorical variables. If the second axis (dispersion) is also dominant, it would suggest that there are non-linear association terms contributing to the association that may be important. We shall not explore this aspect of the analysis here, but invite the interested reader to consider the findings in Beh (1997) and work subsequently done on this topic.

### 3. DATA AND DESIGN APPROACH FOR STRATA, CATEGORIZATION & CUTPOINTS

We evaluate the bivariate association between categorised variants of 2 of the 4 traditional parameters of Lipinski's rule of five (Ro5) (Lipinski & Hopkins, 2004), namely MW and logP, and also Veber *et al.*'s (2002) additional parameter, polar surface area (PSA). The aim of the study is to investigate differences in magnitude or swaps in sign of the pairwise BDNI estimates across strata, where strata are defined by a molecule's druggable (Ro5 compliant) versus non-druggable (Ro5 violation) status. We study 3 of the 9 drug-likeness surrogate variables investigated by Hudson *et al.* (2012) in an examination of 1,279 small molecules from the DrugBank3.0 (Knox *et al.*, 2011) database, which contains 6,711 drug entries, including 1441 FDA-approved small molecule drugs (one candidate molecule example is shown in Table 1).

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 <sub>14</sub> H <sub>18</sub> N <sub>6</sub> O		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.

In this study a molecule is categorized as druggable if Lipinski's rule of 5 (Ro5) is satisfied. In total there are 105 violators/non-druggables in the data set. Table 2 gives the cross tabulation of PSA by MW for data stratified according to Ro5 status (druggable or not) within quartiles created from the whole data set. Numbers in brackets in Table 2 correspond to the 105 non-druggable molecules as judged by the Ro5. Two methods of creating 4 level categorized variables are used – tertile or quartile methods. Data is: [1] stratified within quartiles (bottom of Table 3); or [2] the data is tertiled within the given druggability stratum (levels 1, 2, 3) (see top of Table 3), and the first level category (0) defined to satisfy the new cutpoints determined by Hudson *et al.*, (2012). These cutpoints are PSA ≤ 65, and MW ≤ 305, in contrast to conventional cutpoints for PSA and MW of 140 and 500, respectively, and log P ≤ 1.9 (top part of Table 3), in contrast to Lipinski's cutpoints of log P ≤ 5.

Table 2: PSA by MW cross table (for data stratified by Ro5 status within quartiles)

Druggable (non-druggable)		MW quartiles				Total
		0	1	2	3	
PSA quartiles	0	116 (0)	104 (0)	77 (0)	21 (2)	318 (2)
	1	102 (0)	96 (0)	91 (0)	36 (2)	325 (2)
	2	82 (0)	68 (0)	85 (0)	70 (7)	305 (7)
	3	20 (0)	52 (0)	64 (3)	90 (91)	226 (94)
Total		320 (0)	320 (0)	317 (3)	217 (102)	1174 (105)

Numbers in brackets correspond to the 105 non-druggable molecules as judged by Ro5.

#### 4. RESULTS

Table 3 gives BDNI estimates (95% CIs) and corresponding linear-by-linear CA estimates for the Ro5 based druggability strata, for data tertiled within the stratum, with the first level category (0) satisfying the cutpoints for violation of Hudson *et al.*, (2012) (top part of Table 3); or for data stratified within quartiles (bottom of Table 3). All BDNI estimates are highly significant ( $P < 0.0001$ ). For a given variable pairing and stratum, the sign of the BDNI and the linear by linear CA based association estimates agree (Table 3).

The pairwise associations are traditionally assumed to be significant and positive given the notion of the Ro5 scoring of violations. This is clearly violated by the association between logP and MW, and between logP and PSA for nondruggable molecules. Specifically from Table 3, log P's association with MW changes sign from significantly negative (-0.1127, 95% CI [-0.1135, -0.1119],  $P < 0.0001$ ) for the nondruggables, to significantly positive (0.3025, 95% CI [0.3016, 0.3034],  $P < 0.0001$ ) for the Ro5 based druggable strata. Here data is tertiled within the stratum and the first level category satisfies new cutpoints for violation developed by Hudson *et al.*, (2012), i.e.  $\log P \leq 1.9$  and  $MW \leq 305$  (top part of Table 3), in contrast to classic Lipinski's cutpoints of  $\log P \leq 5$  and  $MW \leq 500$ . Table 3 also shows that for data stratified within quartiles (bottom part of Table 3) Log P's association with MW, whilst positive (0.1101, 95% CI [0.1085, -0.1117],  $P < 0.0001$ ) for

Table 3: BDNI estimates and CA (linear-by-linear) estimates

With cut points: BDNI and CA estimates for Ro5 based Druggability (Quartile within Strata and for the un-quartiled data (All))				
Variable pair	<b>Non druggable</b> $\hat{\Phi}_{BDNI}^{ND}$ (P-value) [Confidence Interval] CA(Linear-by-linear)	<b>Druggable</b> $\hat{\Phi}_{BDNI}^D$ (P-value) [Confidence Interval] CA(Linear-by-linear)	<b>ALL data</b> $\hat{\Phi}_{BDNI}^D$ (P-value) [Confidence Interval]	<b>ALL data</b> Traditional assumed correlation and estimate (r)
LogP , MW	<b>-0.1127</b> (< 0.0001) [-0.1135, -0.1119] -1.16766	<b>0.3025</b> (< 0.0001) [0.3016, 0.3034] 10.36308	<b>0.2448</b> (<0.0001) [0.2430, 0.2466]	Positive $r = 0.08$ NS
PSA, MW	<b>0.4352</b> (< 0.0001) [0.4334, 0.4371] 4.40905	<b>0.3410</b> (< 0.0001) [0.3390, 0.3430] 11.68384	<b>0.4802</b> (<0.0001) [0.4781, 0.4823]	Positive $r = 0.70^{***}$
LogP, PSA	<b>-0.7163</b> (< 0.0001) [-0.7186, -0.7139] -7.34034	<b>-0.3597</b> (< 0.0001) [-0.3607, -0.3586] -12.32382	<b>-0.3528</b> (<0.0001) [-0.3544, -0.3512]	Negative not positive as assumed $r = -0.51^{**}$
Without Cut points: BDNI estimates and CAestimates for Ro5 Druggability (Strata Within Quartile)				
Variable pair	<b>Non druggable</b> $\hat{\Phi}_{BDNI}^{ND}$ (P-value) [Confidence Interval] CA(Linear-by-linear)	<b>Druggable</b> $\hat{\Phi}_{BDNI}^D$ (P-value) [Confidence Interval] CA(Linear-by-linear)	<b>ALL data</b> $\hat{\Phi}_{BDNI}^D$ (P-value) [Confidence Interval]	<b>ALL data</b> Traditional assumed correlation and estimate (r)
LogP vs. MW	<b>0.1101</b> (< 0.0001) [0.1085, 0.1117] 1.0591	<b>0.3980</b> (0.0000) [0.3962, 0.3998] 13.6359	<b>0.2448</b> (<0.0001) [0.2430, 0.2466]	Positive $r = 0.08$ NS
PSA vs. MW	<b>-0.0517</b> (< 0.0001) [-0.0527, -0.0507] -8.3528	<b>0.3151</b> (< 0.0001) [0.3134, 0.3168] 10.7950	<b>0.4802</b> (<0.0001) [0.4781, 0.4823]	Positive $r = 0.70^{***}$
LogP vs. PSA	<b>-0.4444</b> (< 0.0001) [-0.4497, -0.4391] -4.5537	<b>-0.4029</b> (< 0.0001) [-0.4047, -0.4011] -13.8041	<b>-0.3528</b> (<0.0001) [-0.3544, -0.3512]	Negative not positive as assumed $r = -0.51^{**}$

the Ro5 violators, is still less than one third of the positive association for the druggable stratum (0.3980, 95% CI [0.3962, 0.3998],  $P < 0.00001$ ). Likewise the association between PSA and MW, traditionally assumed to be positive, changes sign from significantly negative (-0.0517, 95% CI [-0.0527, -0.0507],  $P < 0.0001$ ) to significantly positive (0.3151, 95% CI [0.3134, 0.3168],  $P < 0.0001$ ) for nondruggable vs druggables, respectively, for data stratified within quartiles (bottom of Table 3). This is also reflected in the ordinal CA plot for the PSA and MW pairing (Figure 1).

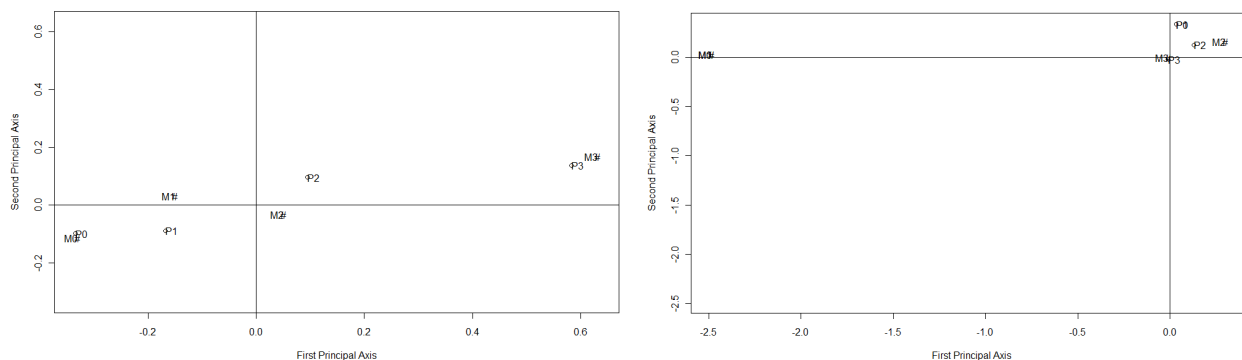


Figure 1 Ordinal CA plot for PSA, MW for data stratified within quartile. Druggable stratum on the left.

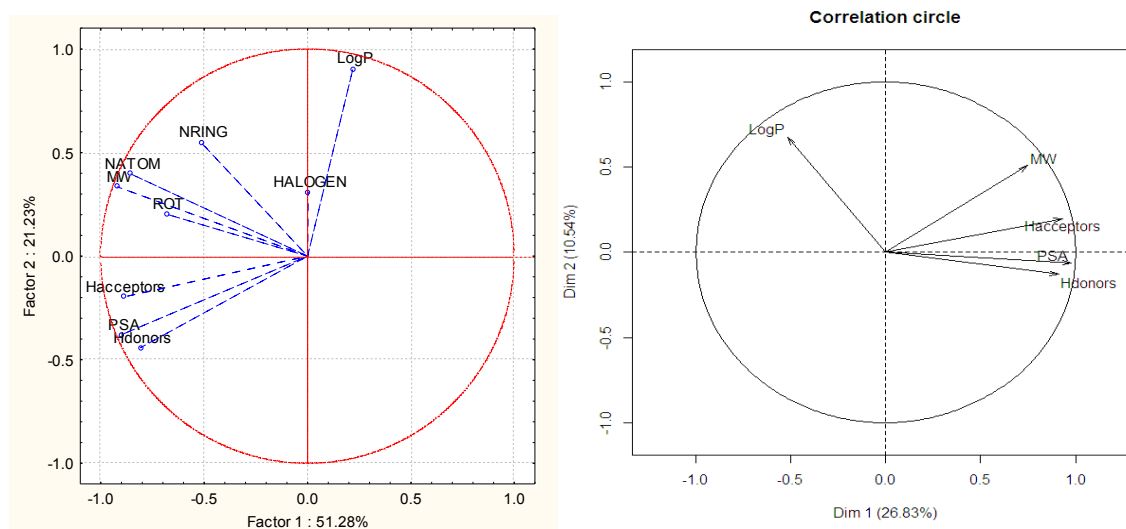


Figure 2 Factor 2 vs Factor 1 plot (Hudson *et al.*, 2012) (LHS), and of the 4 Ro5 variables and  $\log P$  (RHS).

## 5. DISCUSSION AND CONCLUSIONS

Drug-likeness is not a precisely defined, concept in drug discovery. Opinions on so-called drug-likeness are governed primarily by the relationships observed between surrogate measures of drug-likeness and fundamental physicochemical properties. Kenny and Montanari (2013) recently advised that in spite of the industry's conventional thinking that control of the fundamental physicochemical properties of molecules is essential in pharmaceutical design, correlations between these and ADMET properties may not be as strong as is traditionally assumed. We postulate here that strengths of association, as well as directionality of changes in trends overall and across, say, oral or target status etc, should be understood if data-driven decisions are to be accurate (Kenny & Montanari, 2013).

Log P seems the odd variable out (as implied in the factor plot, Figure 2). In this study log P's association with MW and PSA is shown to change magnitude/sign according to the molecule's Ro5 or its oral status (not reported here). This supports recent criticisms about using log P (Bhal *et al* 2007). Our results and recent literature suggest that Log D, as a distribution coefficient (Bhal *et al.*, 2007), may be preferable to log P. Log D needs to be tested formally via these BDNI and CA association methods, and also be tested via self organising maps (SOMs) and a mixture discriminant approach (MC/DA) (Fralely *et al.*, 2013), as developed by Hudson *et al.* (2012).

Whilst in this study a molecule is categorized as druggable if Ro5 is satisfied, future work will involve BDNI and CA methods to test rules of druggability that go beyond Lipinski's Ro5 (Walters 2012). These will involve 9 parameters, namely, the number of halogens, rotatable bonds, rings and N and O atoms, along with

PSA and Lipinski's 4 parameters (MW, logP, number of hydrogen bond acceptors and number of hydrogen bond donors (see Hudson *et al.*, 2012). Recently an alternative score for violations based on Lipinski's 4 variables, but using different cutpoints (Hudson *et al.*, 2012) categorized molecules as druggable, if they satisfied less or equal to 2 violations. This MC/DA based scoring was able to correctly classify 92.7% of the Ro5 based druggables studied here, and 81.9% of the Ro5 nondruggables. In contrast the classical Ro5 rule, whilst identifying 89.5% of the MC/DA score based druggable contenders, could only correctly classify 51.6% of the new MC/DA score based violators. Future research will also test our association methods with respect to targeted disease categories.

## REFERENCES

- Beh, E.J. (1997). Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials, *Biometrical Journal*, 39, 589-613.
- Bhal, S.K., Kassam, K., Peirson, I.G., Pearl, GM. (2007). The rule of five revisited: applying log D in place of log P in drug-likeness filters. *Mol. Pharm.* 4(4):556–560.
- Beh, E.J. & Davy, P.J. (2003). A non-iterative alternative to ordinal log-linear models, *Journal of Applied Mathematics & Decision Sciences*, 7(2), 1-20.
- Beh, E.J. & Farver, T.B. (2009). an evaluation of non-iterative methods for estimating the linear-by-linear parameter for ordinal log-linear models, *Australian & New Zealand Journal of Statistics*, 51(3), 335-352.
- Beh, E.J. & Farver, T.B. (2012). A computational study assessing maximum likelihood and non-iterative methods for estimating the linear-by-linear parameter for ordinal log-linear models, *ISRN Comp. Math.* doi:10.5402/2012/396831.
- Fraley, C., Raftery, A., Scrucca, L. (2013). Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. MCLUST Version 4.1 in R.
- Emerson, P.L. (1968). Numerical construction of orthogonal polynomials from a general recurrence formula, *Biometrics*, 24(3), 695-701.
- Hudson, I., Shafi, S., Lee, S., Hudson, S., Abell, A.D. (2012). Druggability in drug discovery: SOMs with a mixture discriminant approach. *Aust Stats Conference, ASC 2012*, July 2012, Adelaide, Australia, 105.
- Kassel, D.B. (2004). Applications of high-throughput ADME in drug discovery. *Curr Opin Chem Biol.* 8(3):339-45.
- Kenny, P.W., Montanari, C.A. (2013). Inflation of correlation in the pursuit of drug-likeness. *Journal of Computer-Aided Molecular Design* 27, 1-13.
- Kerns, E.H., & Di, L. (2008). Drug-like properties: Concepts, structure, design and methods, from ADME to toxicity optimization. ISBN 978-0-1236-9520-8, Academic Press, Burlington, MA 01803, USA, pp64.
- Knox C., Law V., Jewison T., *et al.* (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* 2011 Jan;39 (Database issue):D1035-41.
- Lancaster, H.O. (1953). A reconciliation of  $\chi^2$  from metrical and enumerative aspects, *Sankhya*, 13, 1-10.
- Leesom, P. (2012). Drug discovery: Chemical beauty contest. *Nature* 481, 455–456.
- Lipinski, C. & Hopkins, A. (2004). Navigating chemical space for biology and medicine. *Nature* 432, 855–861.
- Lipinski, C.A. (2000). Drug-like properties and the cause of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* 44, 235–249.
- Rayner, J. C. W. & Beh, E. J. (2009). Towards a better understanding of correlation, *Statistica Neerlandica*, 63, 324-333.
- Sirois, S., Hatzakis, G., Wei, D. Q., Du, Q. S., Chou, K. C. (2005). Assessment of chemical libraries for their druggability. *Computational Biology and Chemistry* 29, 55–67.
- Ursu, O., Rayan, A., Goldblum, A., Oprea, T. I. (2011). Understanding drug-likeness. *WIREs Comput Mol Sci*, 1: 760–781.
- 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.
- Walters, P. (2012). Going further than Lipinski's rule in drug design. *Expert Opinion on Drug Discovery*, 7, 2, 99-107.
- Wishart D.S., Knox C., Guo A.C., *et al.* (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* 36 (Database issue): D901–6 ADMET
- Zafar, S., Cheema, S.A., Beh, E.J. and Hudson, I.L. (2013), A Study of Bias for Non-Iterative Estimates of the Linear-by-Linear Association Parameter from the Ordinal Log-Linear Model. *Italian Statistical Society: Statistical Conference*, University of Brescia.