Electronic Supplementary Material (ESI) for MedChemComm.

In silico human volume of distribution at steady state model

This document describes the building of an *in silico* human volume of distribution at steady state (V_{ss}) model. The model was built using the process described below and the key technical details summarised in Table 1:

Table 1 Summary of technical details associated to the building of the *in silico* human V_{ss} model.

Model	V_{ss}
Experimental units	L kg^{-1}
Number of unrefined measurements from ChEMBL	1293
Number of associated SMILES strings needing desalting	11
Number of results removed as the associated compound contained P	27
Number of results removed as the associated compound contained B	1
Number of results removed as the associated compound was charged	24
Number of results removed as the associated compound was inorganic	1
Number of results removed as the associated compound was a dimer	2
Data transformation for modelling	$\log_{10}(V_{ss})$
Number of unique compounds with an experimentally derived value	659
Compounds removed for associated measurement's having standard deviation >0.4	3
Compounds removed for containing sugar functional groups	34
Compounds removed for having steroid based structures	27
Compounds removed for having macrocyclic based structures	36
Compounds removed for having 8-fused ring system structures	2
Compounds removed for containing ≥4 SO_3H groups	2
Compounds removed for containing a long aliphatic chain (C ≥13)	1
Compounds removed for being predominantly peptide based	2
Compounds removed for having a molecular weight >850 Da	1
Compounds removed for having an incorrect structure	2
Model's lower dynamic range	-1.3

Model	V_{ss}
Model's upper dynamic range	2.0
Compounds removed for having a measurement < lower dynamic range	1
Compounds removed for having a measurement > upper dynamic range	3
Size of the model's data set	545
Training set size	463
External test set size	82
Machine learning methods used	RF
Number of trees	500
Number of tries	55
Model predictive lower dynamic range (on its original scale)	0.05
Model predictive upper dynamic range (on its original scale)	100.00

Data

Unrefined chemical and experimental data sets was sourced from the ChEMBL database (version 20, January 2015, https://www.ebi.ac.uk/chembl/) using data reported in specific publications (see Reference section below). Extracted SMILES strings were checked for salts and desalted where necessary. Results associated with compounds containing P, B, an overall charge, or being inorganic or dimeric were removed. Where necessary the experimental data was transformed to normalise it. Repeat measurements for compounds were averaged and those with a standard deviation >0.4 removed. Compounds that contained a sugar group, were steroid or macrocyclic based or had an 8-fused ring system were removed. Compounds that contained ≥4 sulphonic acid groups, a long aliphatic chain (C ≥13), were predominantly peptide based, had a molecular weight >850 Da or had incorrect SMILES strings were removed. Compounds with measurements less than or greater than the specified lower and upper dynamic ranges for the *in silico* human V_{ss} model were also removed. The subsequent refined data set were randomly split into

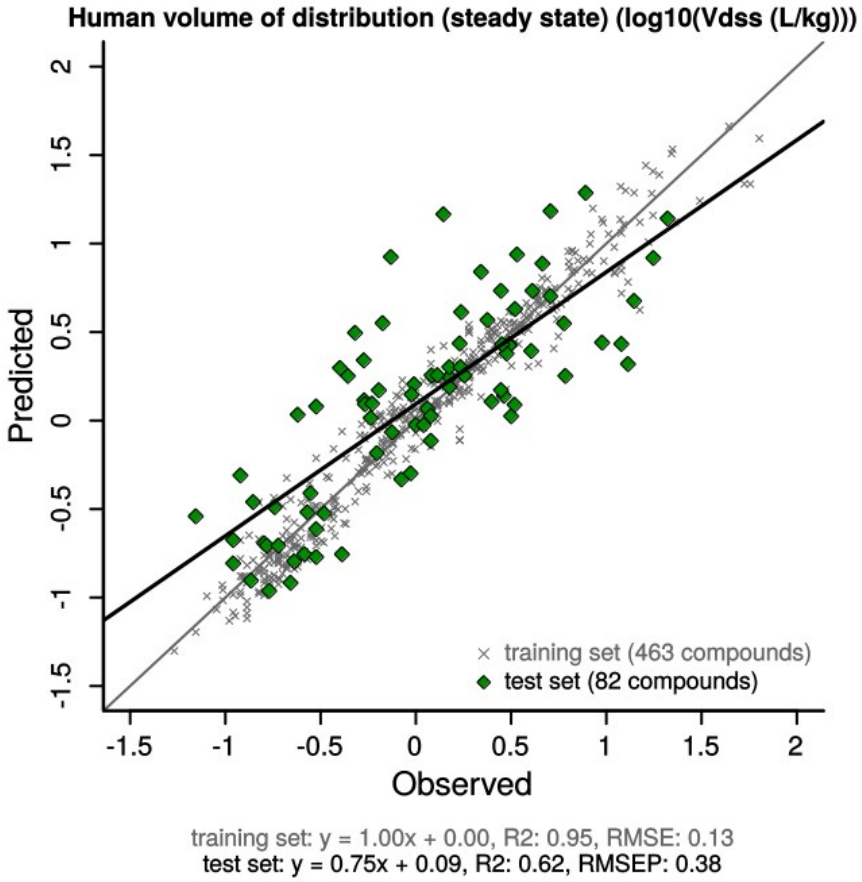
training and external test sets as specified in Table 1.

Descriptors

RDKit descriptors (http://www.rdkit.org) calculated from a molecule's smiles string.

Modelling

Performed within R (http://www.R-project.org/) with the use of the randomForest (RF) library (http://CRAN.R-project.org/package=randomForest). Zero variance and collinear properties were removed prior to modelling. The regression model used k-fold cross-validation to optimise the number of trees and the number of variables tried at each split (*i.e.*, tries), see Table 1 for their details. A correction factor, that transforms the training set model about unity, was applied to all predictions. The best-fit linear regression line for the external test set: predicted $\log_{10}(V_{ss})$ = 0.75 * observed $\log_{10}(V_{ss})$ + 0.09, r^2: 0.62, RMSEP: 0.38.



References

1. *J. Med. Chem.*, 2002, **45**, 2867–2876.
2. *J. Med. Chem.*, 2004, **47**, 1242–1250.
3. *Drug Metab. Dispos.*, 2006, **34**, 1255–1265.
4. *Drug Metab. Dispos.*, 2008, **36**, 1385–1405.
5. *Eur. J. Med. Chem.*, 2009, **44**, 4455–4460.
6. *J. Med. Chem.*, 2010, **53**, 1098–1108.