Supplementary Material for:

Predicting pKa Using a Combination of Semi-Empirical Quantum Mechanics and Radial Basis Function Methods

Peter Hunt[1], Layla Hosseini-Gerami[2], Tomas Chrien[1], Jeffrey Plante[3], David J. Ponting[3], Matthew Segall[1]

1 Optibrium Ltd, F5-6 Blenheim House, Cambridge Innovation Park, Denny End Road, Cambridge, CB25 9PB, UK Tel. +44 (0) 1223 815 900

2 Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW, UK Tel. +44 (0) 1223 336 300

3 Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, UK Tel. +44 (0)113 394 6020

Figure S1

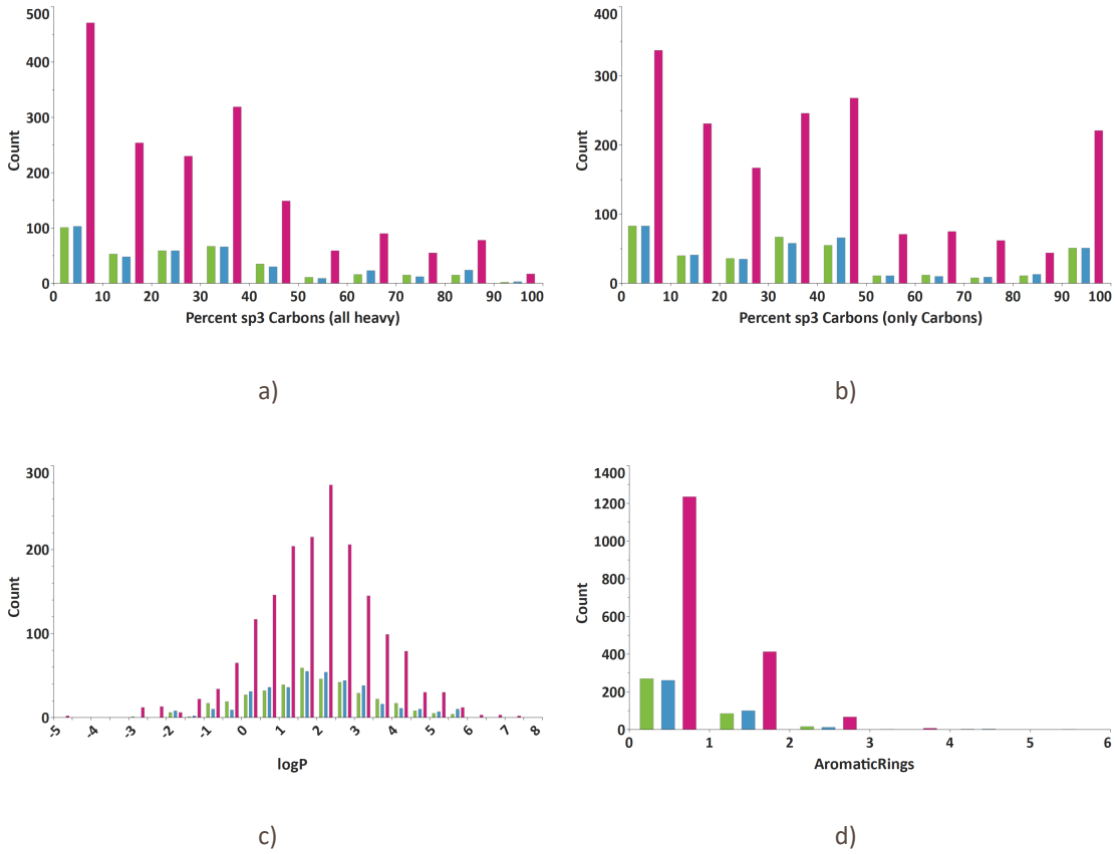
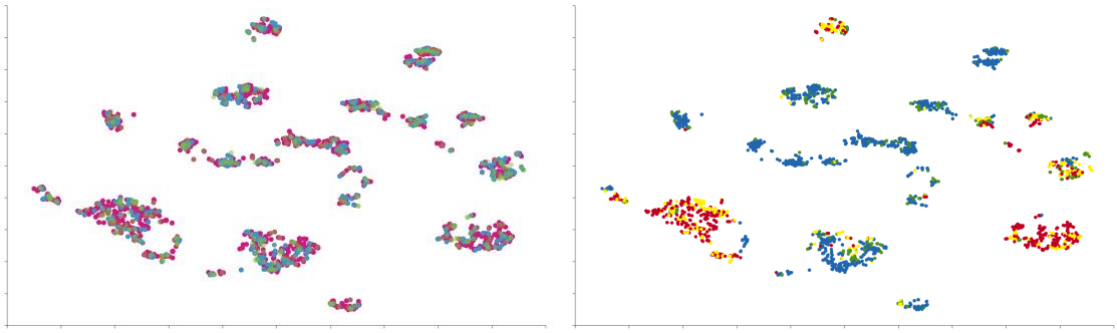


Figure S1. Plots of the distributions of simple properties across the training (pink bars), validation (blue bars) and test (green bars) sets. a) percentage of sp3 carbon atoms taking all heavy atoms into account, b) percentage of sp3 carbon atoms taking only carbon atoms into account, c) calculated LogP, d) Number of aromatic rings.

Figure S2



a) b)

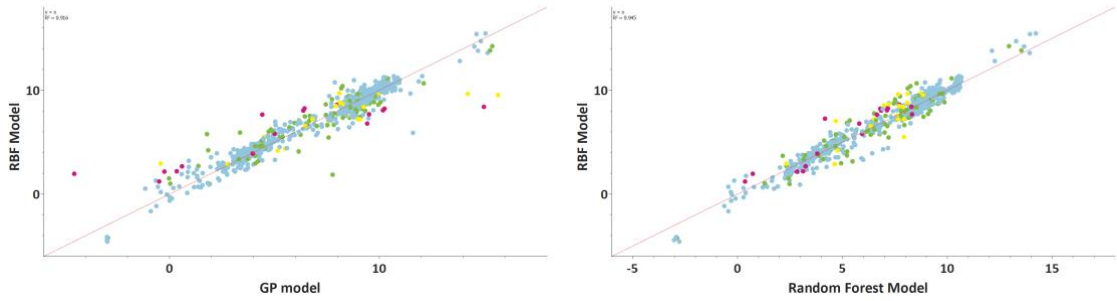
Figure S2. A pair of property space representations derived from the descriptor matrix used in the model generation. Plots a) and b) show the whole pk_a data set and the points are coloured in a) by their set membership either training (pink), validation (blue) and test (green) and in b) by the experimental pk_a values associated with the ionisable site(s) within the molecule pk_a <=4 (red), pk_a >4 and <= 6 (yellow), pk_a >6 and <=8 (green), and pk_a >8 (blue).

Table T1

Torsion of central bond should be 155 degrees:	[#7,#8,#16,#15,F,Cl,Br,I]-,=[#6,#7,#8,#16,#15]-!@[#6,#7,#8,#16,#15]-,=[#7,#8,#16,#15,F,Cl,Br,I]
Torsion should be 25 degrees:	[!#1]~[!#1;!#6]-[#6,#7](:[!#1]):[!#1]
	[#1]-[#7H2]-[#6,#7](:[!#1]):[!#1]
	[#1]-[#8H]-[#6,#7](:[!#1]):[!#1]
	[#1]-[#16H]-[#6,#7](:[!#1]):[!#1]
	[#7,#6,#8,#16]=[#6]-[#6,#7](:[!#1]):[!#1]

Table T1. SMARTS patterns used to assign the torsion angles shown prior to geometry optimisation with AM1 in order to minimise failed descriptor calculations due to proton transfer during optimisation.

Figure S3

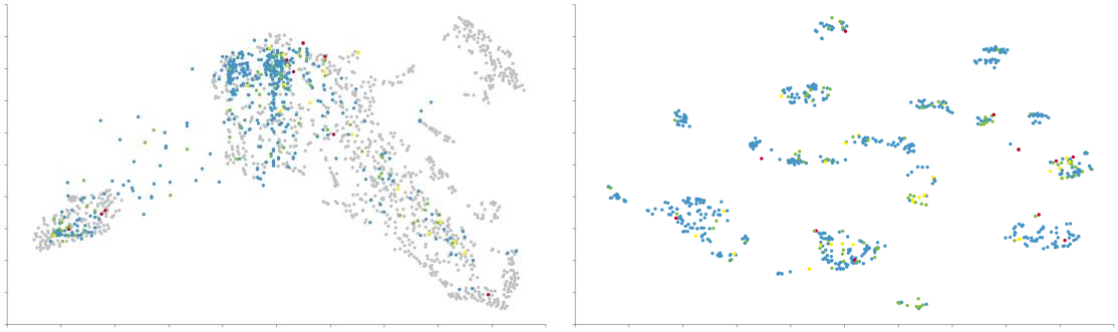


a) b)

Figure S3. Plots of the correlation between predictions for the test and validation sets by different methods. The points are coloured by the absolute deviation from the experimental value for the RBF model prediction

(deviation <= 1 log unit blue, 1 < deviation <=2 log units green, 2 < deviation <=3 log units yellow, >3 log units pink). The models are generally well correlated but the poorly predicted sites in the RBF model are not always poorly predicted by the models from the other methods.

Figure S4



a) b)

Figure S4 A set of property space representations of the data set. In a) the test and validation set compounds are displayed in a launched drug chemical space along with the approximately 1,300 launched small molecule drugs that defined the space, whilst in b) only the test set and validation set compounds are displayed in a property space derived from the descriptor matrix used in the model generation. The colour scheme is consistent across both spaces with the grey points being the launched drugs whilst the pk_a data set points are coloured by the absolute error of the RBF model prediction compared to the experimental pk_a values associated with the ionisable site(s) within the molecule (deviation <= 1 log unit blue, 1 < deviation <=2 log units green, 2 < deviation <=3 log units yellow, >3 log units pink). The plots show that there are no consistently poorly predicted areas in either chemical or descriptor property space.