Journal Name

COMMUNICATION

Modelling the binding constant for potential drugs targeting malarial protein plasmepsin II

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

C. P. Ross[a]

In a journal article by K. Ersmark et al[1] 25 compounds were described as being potential anti-malarial drugs. The aim for this report was to create a model to predict binding affinity to the target protein plasmepsin (Plm) II, using data collected from a variety of online resources.

As one of the most significant infectious diseases the planet currently faces, malaria attracts vital research to be carried out[1]. The affecting parasite is becoming increasingly resistant to existing drugs, resulting in new targets being identified. The aspartic protease (Plm) II has been extensively studied and is involved in haemoglobin catabolism, essential for parasitic survival. Similar in structure to other plasmepsins, they take part in the initial cleavage during the catabolism. This causes the structure to unfold, thus increasing its susceptibility to further proteolysis[2]; an inhibitor would block this degradation.

Crystal structures and simulations of the protease indicate a flexible binding cleft that can accommodate ligands of varying sizes. Inter-domain flexibility is believed to be important for Plm II to recognise and bind specific sequences on the native haemoglobin molecule[3]. A single chain of 329 amino acids folded into two similar domains, the structure has a bilobal shape and topology, typical to that of eukaryotic aspartic proteases. The Protein Data Bank (PDB) represents the structure of Plm II with the unique code 1LF4 and biological assembly image in Fig. 1.

Using a series of inhibitors that have been reported with their chemical structure and binding constants (Ki)[1], identifiers were collected, including SMILES strings and InChI codes. An example set for these are shown in Tab. 1. Some descriptors for the set of compounds were found using a variety of online databases, these included basic information such as molar mass, H-bond donors/acceptors (HBD/A), as well as more complex predicted values for the polar surface area and log P. Many of the descriptors were chosen because they represented properties that were related to Lipinski's rule of five[4]. Where an orally active drug should abide by: no more than 5 HBD and 10 HBA, weigh less than 500 $gmol^{-1}$ and have a log P value of less than 5. Though there are exceptions to these rules, they remain a good indication of which descriptors may be useful for modelling the binding constant. Other descriptors were chosen because they were related to polarity of the molecule, and thus likely to represent the aptitude to binding experienced with the enzyme.

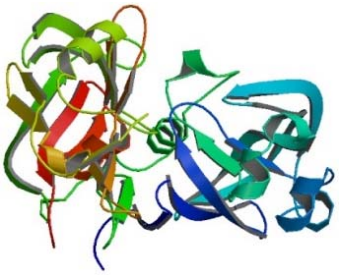
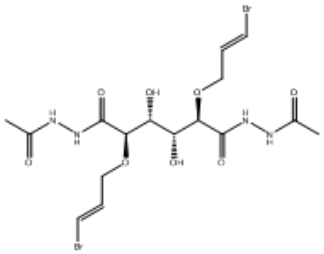
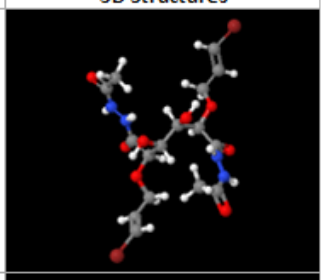
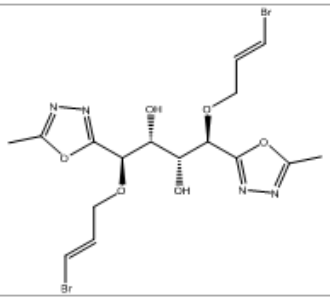
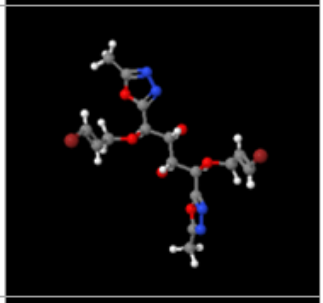
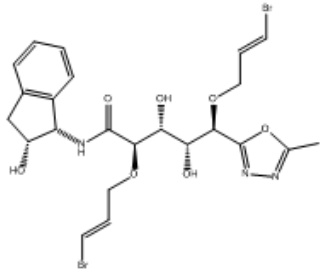
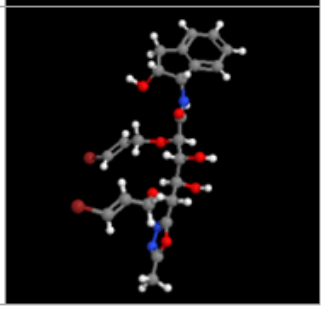


Fig. 1: Biological assembly of the malarial protein Plm II, showing the bilobal nature.

Potential issues were acknowledged with these descriptors, the limited range in values for HBD/A and the many shared values for polar surface area. Molecular weight was unlikely to have a noteworthy effect; instead descriptors based on the functional groups present were more likely to be useful when constructing a model. Predicted log P values were obtained from multiple online databases, unlike the other descriptors a greater variation was observed.

[a] *University of Southampton, Highfield Campus, Southampton, SO17 1BJ*
Electronic Supplementary Information (ESI) available: [More detailed description of the work undertaken, including the models construction]. See DOI: 10.1039/x0xx00000x

Tab. 1: InChI codes and 3D structures for an example set of compounds.

No.	Structure	InChI	3D Structures
3		InChI=1S/C16H24Br2N4O8/c1-9(23)19-21-15(27)13(29-7-3-5-17)11(25)12(26)14(30-8-4-6-18)16(28)22-20-10(2)24/h3-6,11-14,25-26H,7-8H2,1-2H3,(H,19,23)(H,20,24)(H,21,27)(H,22,28)/b5-3+,6-4+/t11-,12-,13-,14-/m1/s1	
17		InChI=1S/C16H20Br2N4O6/c1-9-19-21-15(27-9)13(25-7-3-5-17)11(23)12(24)14(26-8-4-6-18)16-22-20-10(2)28-16/h3-6,11-14,23-24H,7-8H2,1-2H3/b5-3+,6-4+/t11-,12-,13-,14-/m1/s1	
19		InChI=1S/C23H27Br2N3O7/c1-13-27-28-23(35-13)21(34-11-5-9-25)19(31)18(30)20(33-10-4-8-24)22(32)26-17-15-7-3-2-6-14(15)12-16(17)29/h2-9,16-21,29-31H,10-12H2,1H3,(H,26,32)/b8-4+,9-5+/t16-,17+,18-,19-,20-,21-/m1/s1	

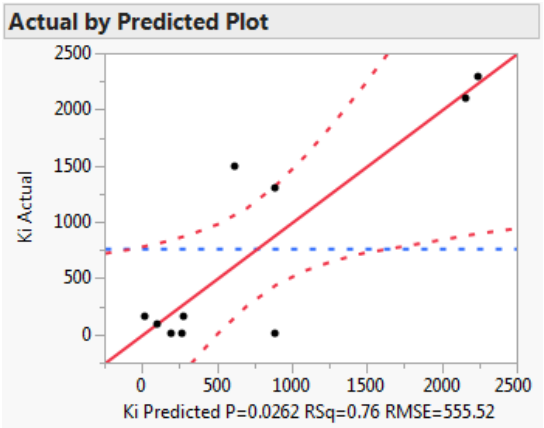
Each descriptor was plotted against Ki, with no significant correlation observed, even with logarithmic and inverse values. However, when used in combination, the data could still be useful. To begin building a QSAR statistical model, the number of descriptors to use was chosen based on the following guidelines:

$$\text{Observations} > 5 \times (\text{no. of descriptors})$$

$$\text{Observations} > 3^{(\text{no. of descriptors})}$$

As the training set will use 10 compounds, the number of descriptors used should be two. A multivariate method was then used to determine the correlation between descriptors, if the relationship was too close, only one would be included in model building. Testing all the descriptors gives an R^2 value of 1, this result is due to the model fitting to noise arising from the numerous descriptors. Removing a descriptor based on the multivariance results, followed by how high the Prob>[t] value was, resulted in four descriptors being leftover. These were modelled as pairs or trios, the three combinations giving the highest residual squared values (R^2) were then retained to use on the testing set.

Model 1: Using HBD, molar volume and log P.



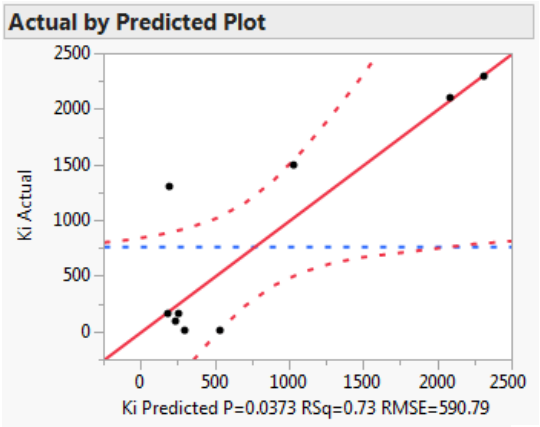
Summary of Fit

RSquare	0.763713
RSquare Adj	0.64557
Root Mean Square Error	555.5168
Mean of Response	766
Observations (or Sum Wgts)	10

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3201.331	2018.541	1.59	0.1638
HBD	-1173.692	294.7978	-3.98	0.0073*
Molar Volume	10.226659	6.344488	1.61	0.1581
log P	-282.2079	244.7308	-1.15	0.2927

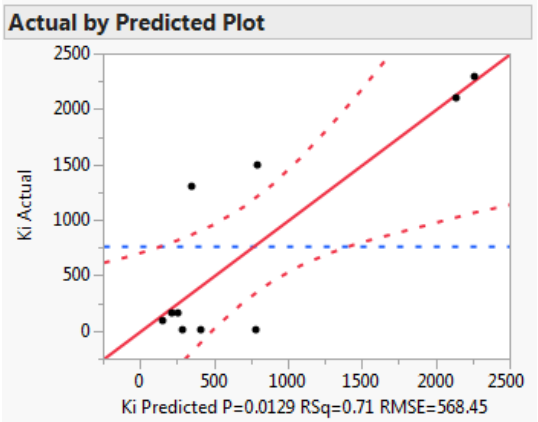
Model 2: Using HBD, molar volume and polar surface area.



Summary of Fit	
RSquare	0.732756
RSquare Adj	0.599134
Root Mean Square Error	590.7879
Mean of Response	766
Observations (or Sum Wgts)	10

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4701.428	2060.446	2.28	0.0626
HBD	-776.8737	349.9184	-2.22	0.0682
Molar Volume	7.128899	5.720282	1.25	0.2591
Polar Surface Area	-17.29956	24.9528	-0.69	0.5141

Model 3: Using HBD and molar volume.



Summary of Fit	
RSquare	0.711347
RSquare Adj	0.628875
Root Mean Square Error	568.4495
Mean of Response	766
Observations (or Sum Wgts)	10

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4209.9904	1861.524	2.26	0.0582
HBD	-953.937	230.1544	-4.14	0.0043*
Molar Volume	4.0840388	3.526512	1.16	0.2848

Each model had rather mediocre performance when it came to predicting the binding constants for the test set; the results are shown in Tab. 2. Whilst there were some strong estimates, the variation showed how weak the models were.

Tab. 2: Predicted values for Ki using the constructed models.

No.	Binding Constant Ki / nM			
	Actual	Model 1	Model 2	Model 3
2	47	54.0	391	249
7	142	-933	-134	-473
9	7	229	415	351
13	586	1150	416	415
23	2900	2200	2200	2200

Conclusions

Using data for the potential anti-malarial drugs, QSAR models were produced. They were not able to accurately predict binding energy; this is likely to be due to having such a small data set. The lack of diversity also contributes to not being able to obtain a reliable model.

To judge whether these compounds would be appropriate as anti-malarial drugs requires further assessment of the binding ability and delivery to the target. Whilst some compounds do achieve a low Ki value, suitable for Plm II inhibition, none of the compounds followed all of Lipinski's rules of five. The lack of adherence to these rues suggests few, if any, are likely to be suitable anti-malarial drugs.

Notes and references

[1] K. Ersmark et al, *J.Med.Chem.,* 2005, **48**, 6090-6106.
[2] *AsparticAcid Proteases as Therapeutic Targets,* ed. A. K. Ghosh, John Wildey & Sons, 2011.
[3] A. Asojo et al, *J.Mol.Biol.*, 2003, **327**, 173-181.
[4] P. Duchowicz et al, *Bioorg.Med.Chem.,* 2007, **15**, 3711-3719.