A 3D-QSAR Study on DPP-4 inhibitors

Giovanna Tedesco[†]

[†] Cresset, New Cambridge House, Bassingbourn Road, Litlington, Cambridgeshire, SG8 0SS, UK

Abstract

3D-QSAR (Quantitative Structure Activity Relationship) models can be built in Forge[1], Cresset's powerful ligand-focused workbench for understanding SAR and design. These models can be created for any available dataset, consisting of a significant number of compounds, which are believed to share a common binding mode and with a reasonable range of binding strength or activity. 3D-QSAR models can explain the currently observed SAR and aid in the design of new molecules where this is called for. In this case study, a data set of 73 dipeptidyl peptidase IV (DPP-4, a serine protease) inhibitors were used to develop a robust 3D-QSAR model within Forge. Ad hoc Forge 3D display capabilities were used to visualize and interpret the model.

Introduction

Many 3D-QSAR methods determine descriptors by calculating molecular properties at the intersection points of a 3D lattice or grid, which covers the entire volume of the aligned molecules. This is necessary because these methods have no way of knowing which region of space around the molecules is likely to be relevant to molecular recognition.

However, Cresset's field point description of molecules provides information about the regions of space around a molecule relevant to molecular recognition. As summarized in Figure 1, the 3D-QSAR method within Forge uses probe positions that are determined directly from the field points of the aligned molecules training set. These positions are used to sample the electrostatic potential or the volume taken up by molecules. The advantage of this method

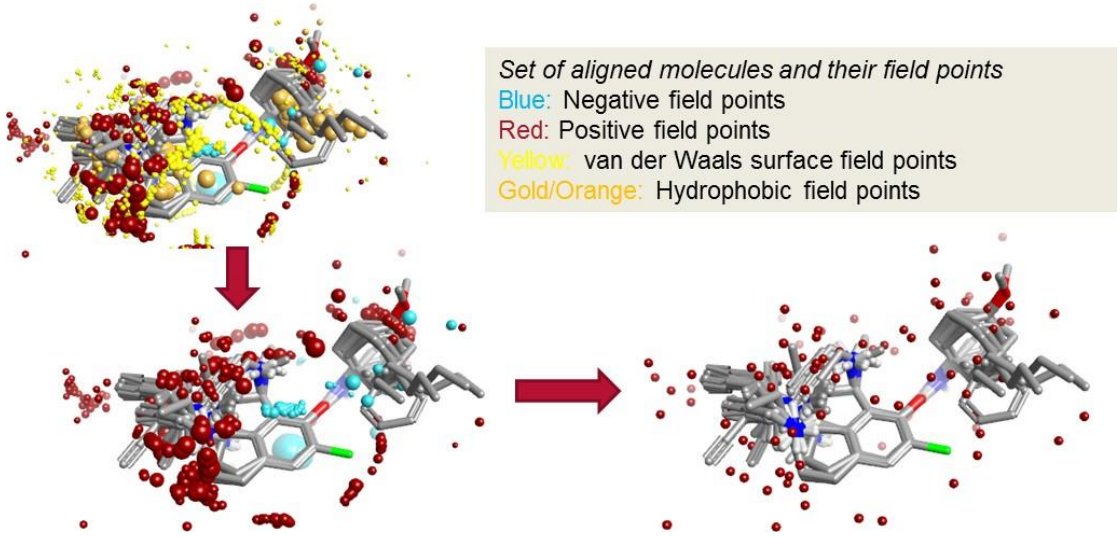


Figure 1. Forge probe location process.

over lattice based methods is that far fewer sample positions are used. Additionally the sample values do not change when the molecules are rotated in cartesian space.

The sample values are combined using Partial Least Squares (PLS) to derive an equation that describes activity. This 3D-QSAR model can help to explain SAR data and, with the best models, used to predict an activity value for newly designed molecules.

However, getting a good 3D-QSAR is challenging. This is due to the requirements of getting good, and consistent, biological data and then generating the correct alignments for all compounds with the lowest degree of noise. Visual inspection of alignments is recommended. This ensures that there are no anomalies present and enables Forge to use the best possible alignment in the model building. Where the calculated alignment is sub-optimal manual intervention can be used to improve them. However, caution must be exercised not to manually create a model of activity that is dependent on the alignment (e.g. all the actives access a different space to all the inactives).

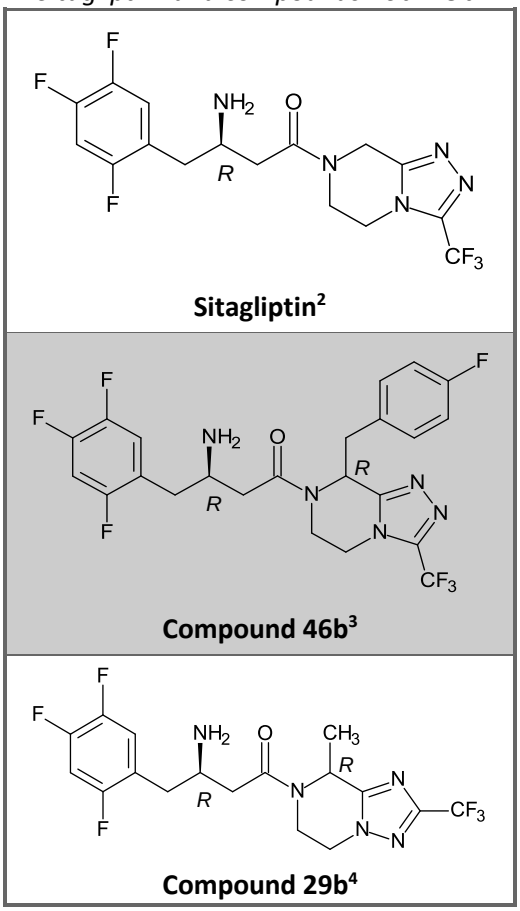
In this case study, the published structures of 73 inhibitors of DPP-4, and related biological activity data (*in vitro* inhibition of DPP-4), were used as a training set for 3D-QSAR model.

Inhibitors of DPP-4 are a class of oral hypoglycemics that can be used to treat diabetes mellitus type 2. Sitagliptin (Table 1), a potent, selective, and orally active DPP-4 inhibitor, was approved by the U.S. FDA in 2006. An extensive chemical exploration of the phenyl ring (left), and of the heterocyclic ring (right), have recently been reported in the literature[2,5].

The absolute stereochemistry of Sitagliptine[2], and compounds 46b, 29b, as determined experimentally by X-ray crystal structure[3,4], is shown in Table 1.

For analogues of compounds 46b and 29b the stereochemistry of the most potent diastereoisomer (where not experimentally determined) was also assumed to be *[R, R]*. Only the most potent diastereoisomer of each compound (corresponding to the bioactive conformation) was included in this study.

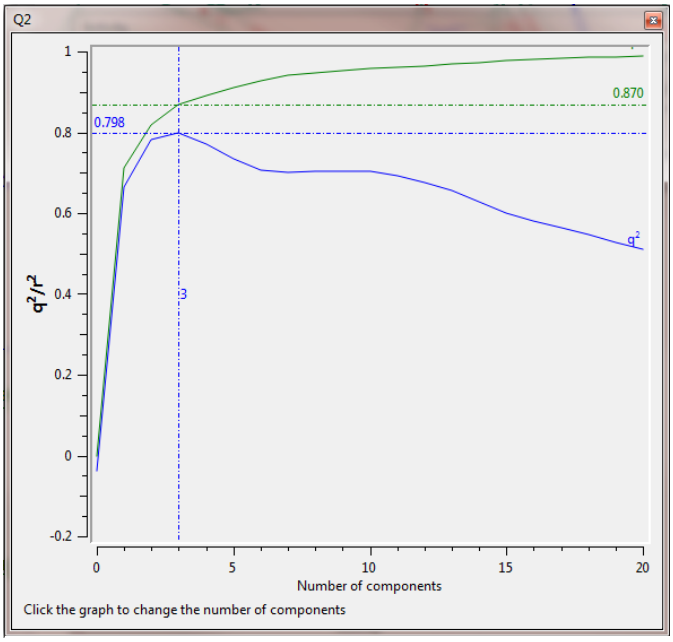
Table 1. Absolute stereochemistry of Sitagliptin[2] and compounds 46b[3], 29b[4].



Conformation hunt and alignment of compounds

Compound 46b (pIC50 = 9.74, the most potent in the training set) was used as the reference compound to align the training set.

The alignment of a few compounds was manually adjusted by flipping the phenyl ring. This aligned the ortho substituents in a manner consistent with the whole dataset.



Training set: 58 compounds
pIC_{50} range: 5.8 - 9.7
Test set (activity stratified): 15 compounds
pIC_{50} range: 6.1 - 9.4

Optimal number of components: 3
r^2 = 0.870
q^2 = 0.798
r^2 test set = 0.623
RMSE = 0.339
RMSEpred = 0.422

Figure 2. 3D-QSAR model statistics.

The conformation of the scaffold of 46b was derived from the X-ray conformation of Sitagliptin bound to DPP-4 (PDB 1X70). The orientation of the $-CH_2$(4-fluorophenyl) group was adjusted according to published data[3] .

The other compounds in the training set were aligned to compound 46b by Maximum Common Substructure using a customized 'accurate but slow' set-up for the conformation hunt:

- Max number of conformations: 500
- RMS cut-off for duplicate conformers: 0.2
- Gradient cut-off for conformer minimization: 0.1 kcal/mol
- Energy window: 3 kcal/mol.

Statistical analysis and results

The regression method used in Forge is PLS[6]. Specifically, the SIMPLS algorithm[7] was used. The initial training set of 73 compounds was partitioned into 80% training set (58 compounds) and 20% test-set (15 compounds). The activity stratified method was used.

The following conditions were used to calculate the field QSAR model:

- Maximum 20 PLS components
- Leave-many-out cross-validation (20% of training set, 1000 repeats)
- 50 Y scrambles
- Sample point minimum distance threshold: 1 Å.

The results are shown in Figures 2 and 3.

The 3-components model shows both good descriptive and predictive ability. This is shown by the good r^2 and q^2 values for the training and the cross-validated training set (Figure 2).

The 'model coefficient' view shows the regions where the QSAR model suggests that the local fields have a strong effect on activity. Large points indicate that the model has found a strong correlation between the electrostatic/steric field in that position and

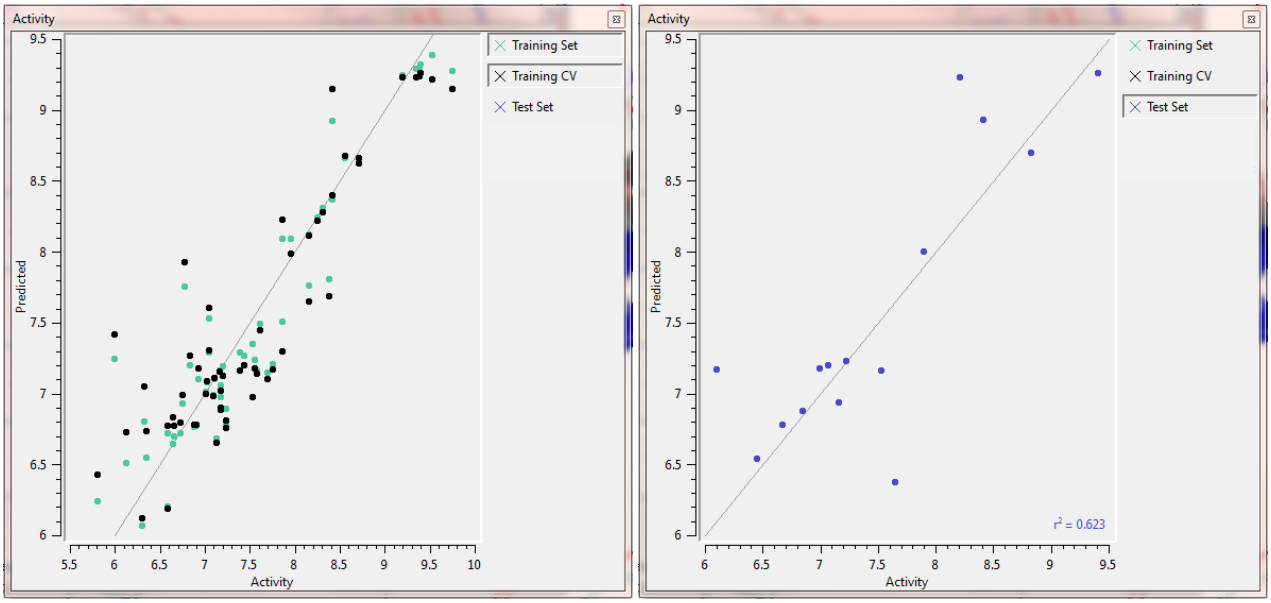


Figure 3. Three components DPP-4 3D-QSAR model - experimental vs. predicted activity of the compounds in the training set (left) and the test set (right).

The plot of experimental *vs.* predicted activity for the compounds, in the training set and the cross-validated training set (Figure 3, left), shows a good distribution of the values with only a few outliers. The plot of experimental *vs.* predicted activity for the compounds in the test set (Figure 3, right) is still reasonably good with only three outliers and a cross-validated r^2 = 0.623.

Model visualization and interpretation

A number of different views are available in Forge to help the visualization and interpretation of the 3D-QSAR model.

higher activity values.

Electrostatic and steric model coefficients for the three components DPP-4 3D-QSAR model are shown in Figure 4. This 3D-QSAR model is clearly dominated by the steric effects of substituents, as indicated by the large size of green and purple polyhedra. The electrostatic effects seem to play a minor role, as indicated by the very small size of red and cyan polyhedra.

In Figure 4, the 3D-QSAR model coefficients are superposed to the structures of the two most potent (46b[3] pIC50 9.74; 44b[3] pIC50 9.51), and two less potent (29[5] pIC50 5.99, 36[5] pIC50 5.8) compounds in the training set. It can clearly be seen that substitution in position 8 of the

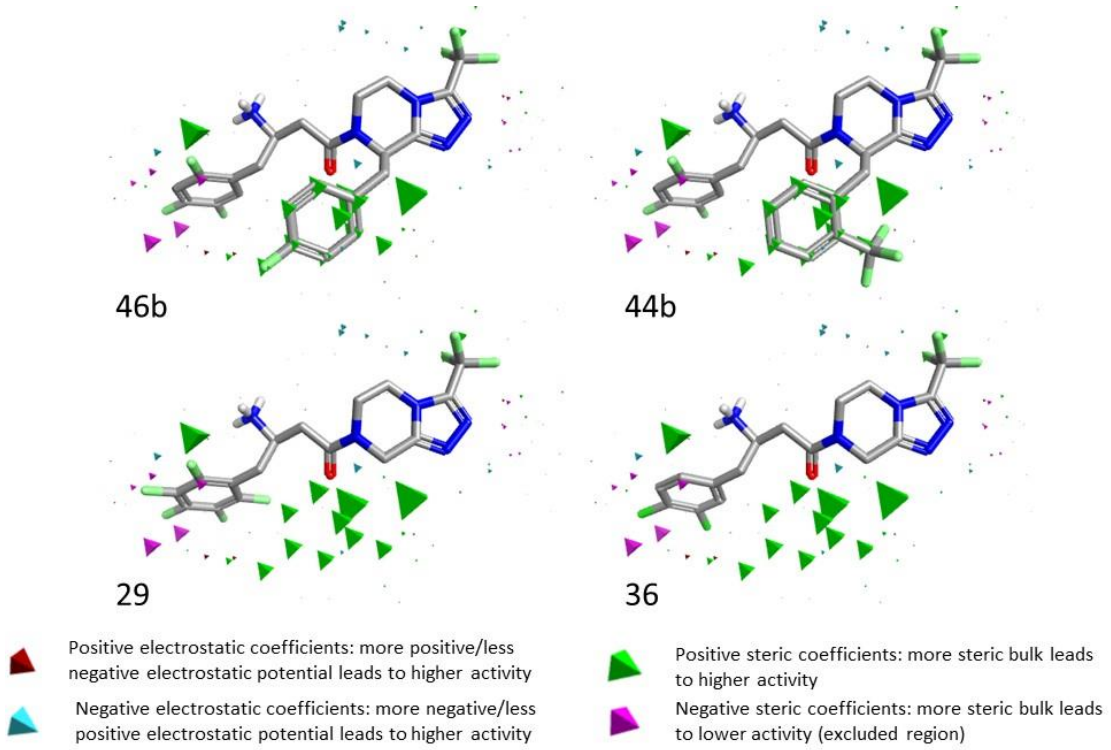


Figure 4. DPP-4 3D-QSAR electrostatic and steric model coefficients superposed to the two most potent (46b[3], 44b[3]) and two less potent (29[5], 36[5]) compounds in the training set.

heteroaromatic ring (as in compound 46b and 44b, top row) improves DPP-4 enzyme activity. Ortho substitution on the phenyl ring on the left is also beneficial, while an increase in steric bulk in the para position is detrimental to activity.

Figure 5 shows the model field contributions to predicted activity. This view displays how well each particular molecule fits the model. The two most potent compounds in the top row (46b[3], 44b[3]) have all the 'good' features (substitution in position 8, ortho substituent on the phenyl ring of the left, no bulky para substituent on the same ring), while the two less potent compounds on the bottom row (29[5], 36[5]) both miss the substituent in position 8 and 36 has a bulky para-Cl substituent on the left phenyl ring.

Conclusion

Forge was used to build a statistically robust 3D-QSAR model for a set of 73 DPP-4 enzyme inhibitors. Forge visualization capabilities made model visualization and interpretation straightforward.

However, while powerful statistical techniques can give seemingly miraculous results, one needs to be constantly aware of the potential issues that can arise. These include over-parameterized models, statistically questionable predictions, robustness to extrapolation rather than interpolation and many more. Further complications arise due to the complexities of conformational searching, molecular alignment, variations in binding modes across series, dealing with inconsistent biological data, etc.

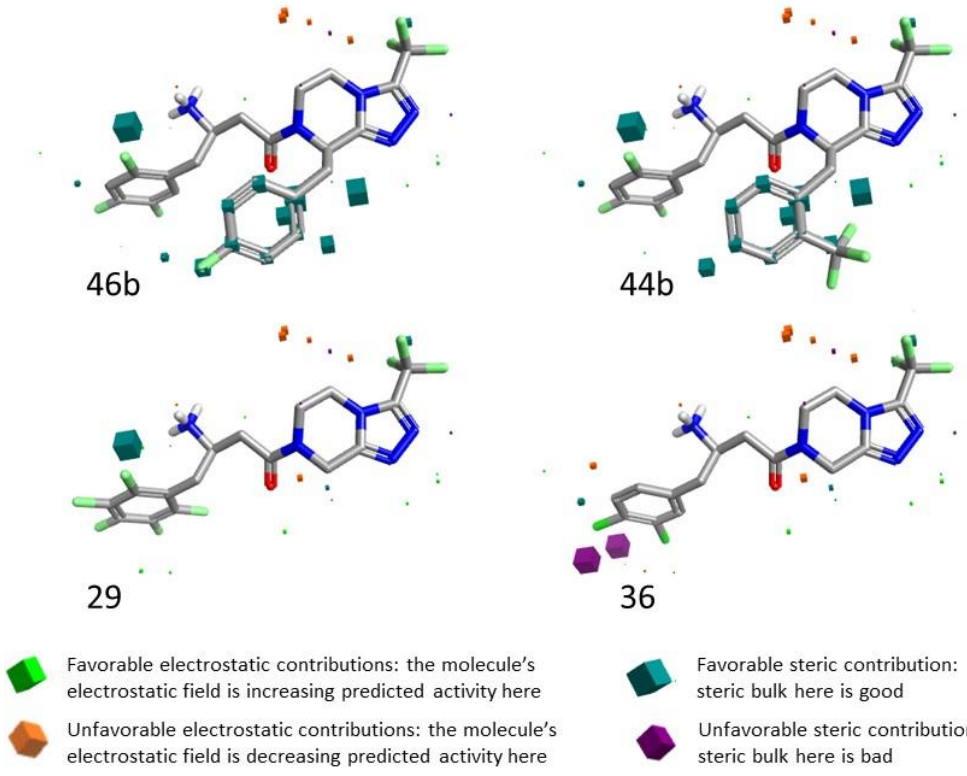


Figure 5. DPP-4 3D QSAR field contributions to predicted activity for the two most potent (46b[3], 44b[3]) and two less potent (29[5], 36[5]) compounds in the training set.

References and Links

1. http://www.cresset-group.com/products/forge/
2. Kim, D. et al., *J. Med. Chem.* **2005**, 48, 141-151
3. Kim, D. et al., *J. Med. Chem.* **2008**, 51, 589–602
4. Kowalchick, J. E. et al., *Bioorg. Med. Chem. Lett.* 17 (**2007**) 5934–5939
5. Kim, D. et al., *Bioorg. Med. Chem. Lett.* 17 (**2007**) 3373–3377
6. Wold, S. et al., *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109-130
7. de Jong, S. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251-263.