# The Application of COMBINE Analysis to Generate Target-Specific Scoring Functions

Stefan Henrich[1], Isabella Feierberg[2], Ting Wang[1], Niklas Blomberg[2], Rebecca Wade[1]

[1] EML Research gGmbH, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany

[2] DECS Chem Comp Dep, SC 2, AstraZeneca R&D Mölndal, S-43183 Mölndal, Sweden

## Introduction

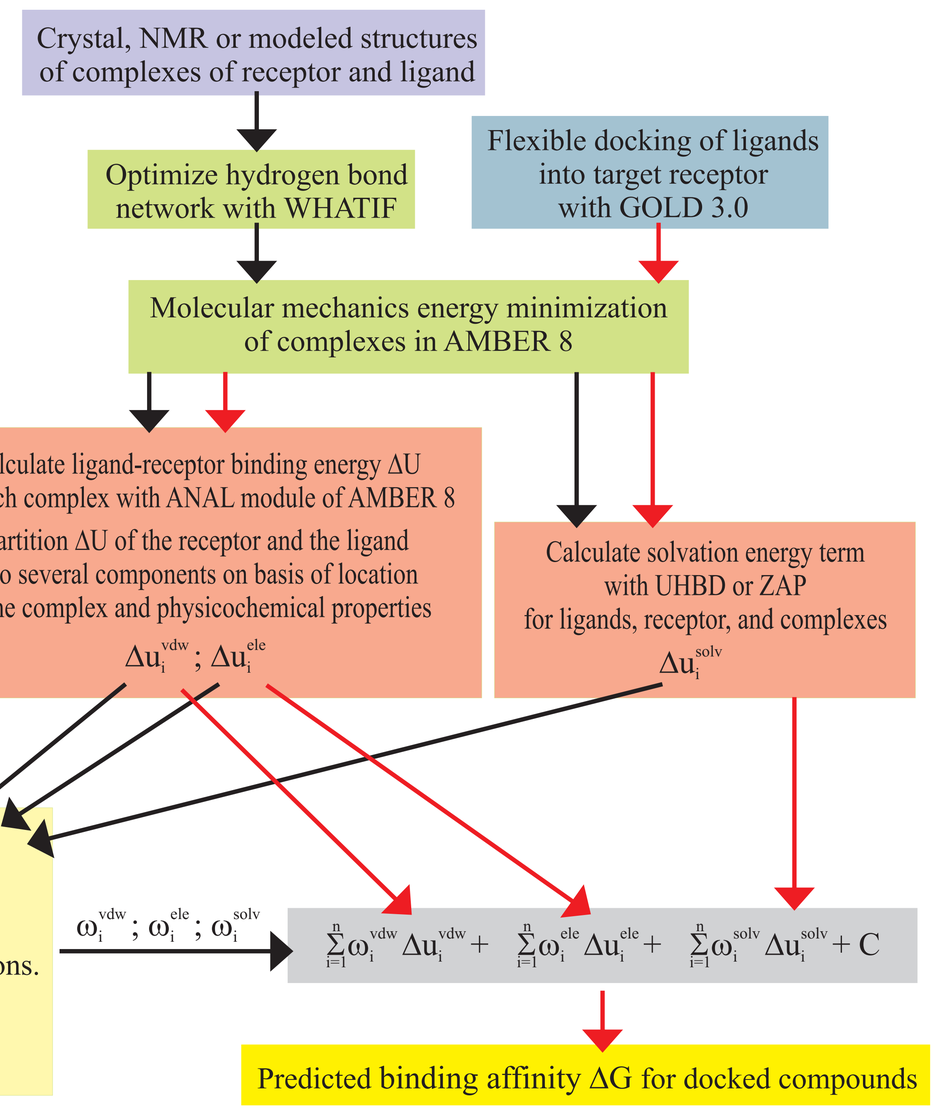
Quantitative structure-activity relationship (QSAR) analysis is an essential method to correlate the properties of a series of molecules with their biological activities and to predict the activities of new compounds. Tailor-made scoring functions can be constructed by using the macromolecular structure-based COMparative BINding Energy (COMBINE) analysis (ref. 1-3). This method provides the possibility to derive 3D QSARs for a set of receptor-ligand complexes whose 3D structures can be modeled.

## Method

The principal idea of COMBINE analysis is the assumption that the binding free energy ΔG is correlated with a subset of weighted interaction energy terms determined by the structures of receptor and ligand. COMBINE analysis starts with an energy minimized model of a receptor-ligand complex that is divided for energy calculations into parts according to their spatial location, normally its amino acid residues and the bound ligand. These parts are used for calculating electrostatic and Lennard-Jones interaction energies as well as solvation energy terms between parts of the ligand and of the receptor.

The resultant energy terms of many receptor-ligand complexes are analysed by Principial Component Analysis (PCA) and are correlated to activity values by Partial Least Squares (PLS) coupled with suitable variable selection and data pretreatment. With this correlation, important residues of the target can be identified for describing the binding affinity between receptor and ligand.

New ligands, which are docked into the active site, can be ranked by this target-specific scoring function and activity can be predicted.



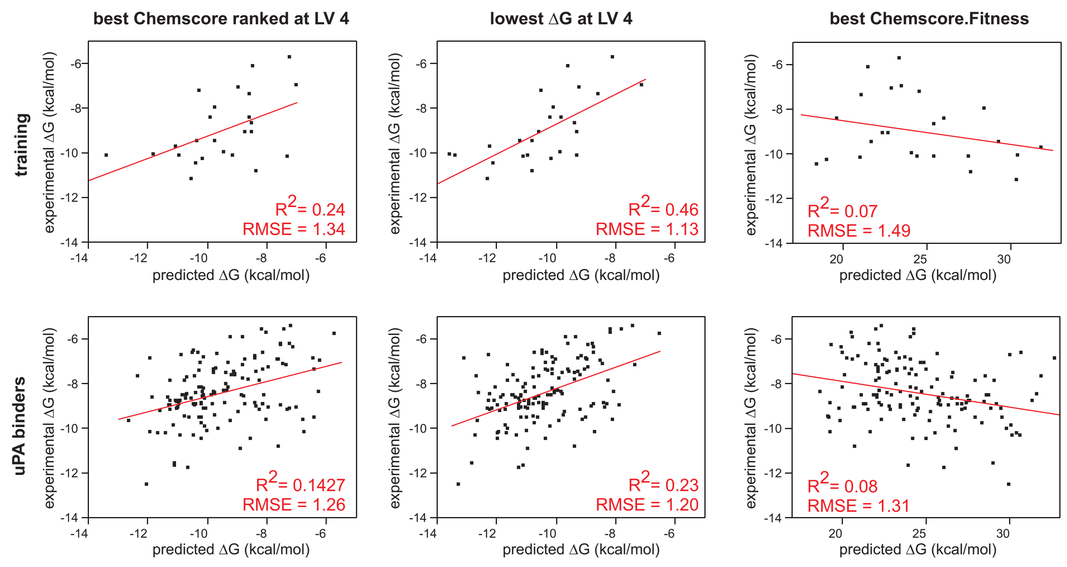

## Model building

For generating target-specific scoring functions by COMBINE analysis, we are focusing on different trypsin-like serine proteases of the blood coagulation cascade, because of the large amount of published data as well as their importance in diseases.

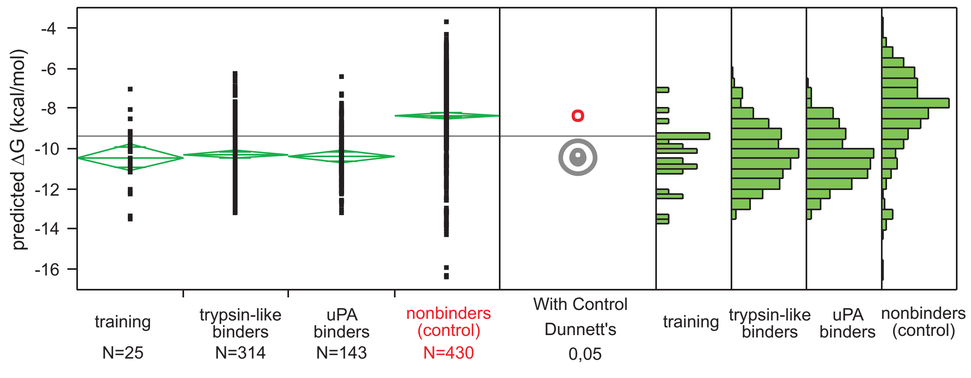
The COMBINE model for urokinase based on 25 X-ray structures of the PDB and their published inhibitor constants. The structures were superimposed and a representative receptor model was build. This model was minimized in complex with the experimental determined conformations of the ligands. After calculating the interaction and solvation energy terms the initial model resulted in a coefficient of determination $R^2$ of 0.75 and a predictive correlation coefficient $Q^2$ of 0.36 at latent variable (LV) 4. A variable selection improved the model to a $R^2$ of 0.75 (0.83) and a $Q^2$ of 0.59 (0.62) at LV 4 (LV 5).

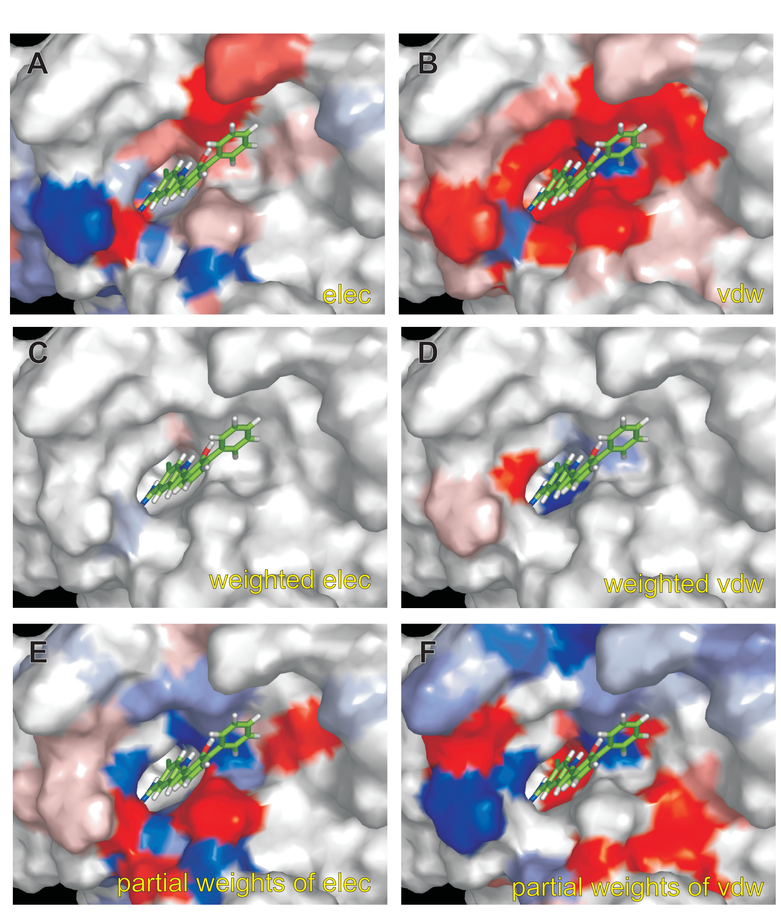
In a subsequent step, the used ligands ('training') as well as further ligands with known binding affinity to trypsin-like proteases ('trypsin-like binders'), especially to urokinase ('uPA binders') and randomly selected ligands from the NCI database were docked ten times into the receptor model. For all of the docking solutions ΔG were predicted and were statistically analyzed.

For each of the ten docking solutions of the test sets 'training' and 'uPA binders' ΔG values were predicted. These were plotted (left and middle) together with the best Chemscore values (right) against the experimental ΔG values. On the left side, predicted ΔG values of the best Chemscore ranked docking solution and in the middle of the docking solution with the highest predicted binding affinity at LV 4 were used. The corresponding $R^2$ and RMSE values are given in the plots. A re-ranking by results of COMBINE analysis gave a much better correlation than taking the best Chemscore ranked docking solution or Chemscore values alone.




Means of lowest predicted ΔG values at LV 4 of the test sets 'training', 'trypsin-like binders', 'uPA binders' are significant distinguishable from randomly selected ligands from the NCI database 'nonbinders'. The figure shows the 95% confidence interval of the means, the distribution of predicted ΔG values and results of Dunnett's test.




Visualization of interaction energy terms. Electrostatic (A, C) and van der Waals (B, D) interaction values between a receptor model and one of the compounds in the training set were mapped onto the receptor binding site surface (in red stabilizing and in blue destabilizing regions for complex formation). A and B show the unweighted interaction energy terms and C as well as D the corresponding ones, weighted by the partial weights shown in E and F (in red unfavored and in blue favored parts). It can be seen that not all of the interaction energies have the same impact to the binding affinity.

## Conclusion

In the present work we applied the COMBINE analysis approach to the problem of predicting the selectivity against trypsin-like serine proteases. With target-specific scoring functions generated by COMBINE analysis for urokinase, we performed virtual screening. Because of problems like overfitting, a too diverse test set and probable incorrect experimental inhibitor constants, we got a weak correlation between experimental and predicted ΔG values. However, the correlation was suitable to discriminate probable nonbinders from proved binders.

**References**

1. Wade, R.C., Henrich, S., Wang, T. (2004) Using 3D protein structures to derive 3D-QSARs. *Drug Discovery Today: Technologies* 1:241-246.
2. Ortiz, A.R., Pisabarro M.T., Gago F., Wade R.C. (1995) Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* 38:2681-2691.
3. Wade, R.C., Ortiz, A.R., Gago F. (1998) Comparative Binding Energy Analysis. *Persp. Drug Disc. and Des.* 9:19-34.

We gratefully acknowledge support of AstraZeneca, Sweden, and of the Klaus Tschira Foundation.