From Hansch-Fujita Analysis to AFMoC: A Road to Structure-Based QSAR

Christoph G. W. Gertzen and Holger Gohlke

Dedicated to the Memory of Corwin Hansch and Toshio Fujita and Their Outstanding Contributions to the Field of QSAR
Abstract: Since the pioneering effort of Hansch and Fujita, quantitative structure–activity relationships (QSAR) have proved valuable in optimizing lead structures. Enriching classical 3D-QSAR analysis, which exploits the three-dimensional structure of ligands, with structural information of the target has helped to improve the interpretability of the derived models and to increase their predictive power. One such method is the Adaption of Fields for Molecular Comparison (AFMoC) approach where protein-specifically adapted knowledge-based pair-potentials are tailored to one particular protein by considering additional structural and energetic information about ligands. Here, we summarize applications of AFMoC, describe recent developments, and provide an outlook on how to improve the method.

Keywords: Molecular bioinformatics • Structure–activity relationship • Lead optimization • Scoring function • Protein flexibility

1 Introduction

Correlating the structure of ligands to their activity has been an important goal since the very early days of modern chemistry.\[1\] In a pioneering effort, Hansch and Fujita were the first to succeed in establishing quantitative structure–activity relationships (QSAR) with the p-o-μ-analysis.\[2\] They correlated the biological activity of chemical compounds using a Hammett-type relationship\[3\] with the compounds’ lipophilicity (expressed as the octanol-water partition coefficient) and electronic and steric properties, that is physicochemical properties, by means of a regression analysis.\[4,5\] The method is still used to date due to its vast applicability.\[6\] In the same year, Free and Wilson published a method for deriving structure–activity relationships that does not model activity contributions with physicochemical terms but rather attributes incremental activity values to compounds’ groups and substituents.\[6\] In many instances, both the Hansch and Free Wilson analyses have been combined to a mixed approach, which has turned out to be a powerful tool of classical QSAR.\[6a\] Moving from the level of physicochemical properties (1D descriptors) and structural formulas (2D descriptors), as in the case of Hansch and Free Wilson analyses, to considering the three-dimensional structure of ligands then provided another major breakthrough that established the field of 3D-QSAR methods.\[7\]

2 3D-QSAR Methods

2.1 Ligand-Based Methods

3D-QSAR methods can be grouped into structure- and ligand-based methods.\[8\] The first 3D-QSAR method, Comparative Molecular Field Analysis (CoMFA),\[7\] introduced by Cramer and coworkers, belongs to the latter group. Here, 3D structural information of a compound is correlated to activity data by means of exploiting molecular interaction fields.

To start with, an alignment of 3D structures of a set of ligands onto the “bioactive conformation” of one of the ligands is required. In the initial application of CoMFA,\[7\] the bioactive conformation was obtained by generating a low-energy conformation of one of the more rigid compounds in the dataset. Alternatively, if a crystal structure of the target with a co-crystallized ligand is available, the compounds of interest can be aligned onto this conformation.\[9\] Finally, if an apo structure of the target or a homology model is available, docking poses of the ligands can be generated and used for the structural alignment.\[9\] In the latter cases, it is reasonable to do an energy minimization for each of the aligned ligands in the presence of the target structure. That way receptor information is implicitly included in the ligand alignment. This very likely leads to a decrease in the structural overlap of the ligands resulting in models that show a lower correlation between predicted and actual pKi values. Nonetheless, CoMFA models derived that way have been shown to have a high predictive power.\[10\]

After the structural alignment molecular interaction fields based on van der Waals and Coulomb potentials are generated outside each ligand at fixed intersections of an orthorhombic grid; these fields represent the ligand’s steric and electrostatic properties. Field values are subsequently correlated to activity data by means of partial least-squares (PLS) analysis. This takes into account that usually the number of field values exceeds by far the number of activity data. As a result of the PLS analysis, weights (“coefficients”) are obtained for each intersection point and field value that specify to what extent this combination contributes to the activity of a compound. For new compounds, activity data can then be predicted based on the new compounds’ interaction fields and the weights.

Additionally, the derived CoMFA model can be interpreted in that so-called Stdev*Coeff fields show where a favorable or disfavorable contribution to the activity is to be expected in the presence of sterically demanding or polar substituents. This information can be exploited to design novel ligands in the process of lead optimization. Especially when considered in the binding region of a crystal structure or a homology model, the Stdev*Coeff fields can be interpreted in view of the interactions between the target structure and the ligands, which simplifies the interpretation of the fields and, hence, the optimization of lead structures.

\[a\] C. G. W. Gertzen, H. Gohlke
Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf
fax: +49-211-8113847
e-mail: Gohlke@hhu.de

One should note, however, that as the protein environment is not considered during calculation of the fields, the CoMFA analysis examines interactions to all parts of the ligands, even if the parts are solvent-exposed and so should not contribute to activity. Furthermore, as a drawback of the generic (steric, electrostatic) interaction fields used, only suggestions such as “place a more bulky group with more negative electrostatic potential in this binding pocket site” can be made, which may be difficult to translate into actual chemical modifications.\textsuperscript{[11]} To some extent, the latter disadvantage is reduced in the 3D-QSAR approach Comparative Molecular Similarity Index Analysis (CoMSIA).\textsuperscript{[12]} Here, ligands are compared in terms of relative similarities in their steric, electrostatic, hydrophobic, and H–bond donor/acceptor properties, which provides more detailed information for interpretation of the QSAR model.

2.2 AFMoC: Protein-Based Tailoring of Knowledge-Based Potentials

Both of the above disadvantages are overcome in the AFMoC approach,\textsuperscript{[11]} a “reverse” protein-based CoMFA method, where protein-specifically adapted knowledge-based pair-potentials are tailored to one particular protein by considering additional structural and energetic information about ligands. For this, a regular-spaced grid is placed into the binding site of a target, and pair-potentials between protein atoms and ligand atom probes are mapped onto the grid intersections resulting in “potential fields” (Figure 1). By multiplying distance-dependent atom-type properties of actual ligands docked into the binding site with the neighboring grid values, “interaction fields” are produced from the original “potential fields”. In a PLS analysis, these atom-type specific interaction fields are correlated to the actual binding affinities of the embedded ligands, resulting in individual weights for each field value. As in CoMFA, the results of the analysis can be interpreted in graphical terms by Stdev*Coeff maps, and binding affinities of novel ligands are predicted by applying the derived 3D-QSAR equation. As only protein-ligand interactions up to a distance of 6 Å are mapped onto neighboring grid points, parts of the ligands that are solvent exposed will not be taken into account in the model derivation, in contrast to CoMFA and CoMSIA.
AFMoC requires the ligands to be structurally aligned, yet not only with respect to a bioactive conformation but also inside a binding pocket. The structural alignment of the ligands can be achieved by the same means as in the CoMFA analysis. In the case of the AFMoC analysis, each ligand must be energy minimized in the presence of the target structure to avoid a compound clashing with the target, which would have a strong influence on the analysis.

For incorporating the information about the structural environment of the ligands, DrugScore pair-potentials have been applied so far for calculating the potential fields. These potentials have proven valuable as scoring and docking functions. They have been derived by converting structural database information of experimentally determined protein-ligand complexes, which implicitly includes entropy-driven effects arising from (de-)solvation. This may explain as to why convincing results were obtained in AFMoC analyses even if (structural) water molecules were not included. Furthermore, although these potentials explicitly depend on the distance between two atoms, they implicitly contain information about directional features of an interaction, e.g. the angular dependence of the strength of a hydrogen bond. This arises from the superimposition of multiple potentials at one point in space. Finally, as the pair-potentials are atomtype-specific, so are the resulting interaction fields. This enables one to propose structural modifications during lead optimization in terms of favorable ligand atom types, in contrast to generic properties as in CoMFA or CoMSIA. We note that by now several variants of the DrugScore potentials have been derived from structural information on RNA-ligand complexes (DrugScore[RNA],[17]) protein-protein complexes (DrugScore[PP]), and nonbonded interactions in small organic molecule crystal packings (DrugScore[CSP],[19]) Each of these potentials can be used in connection with AFMoC depending on the application area, as could be any other distance-dependent pair-potential such as the Astex Scoring Potential (ASP).[20] The use of knowledge-based atom type-specific pair-potentials distinguishes AFMoC from a related approach, COMBINE.[21]

Here, first, a ligand-macromolecule interaction energy is computed for a set of ligands using molecular mechanics calculations. Then, by selecting and scaling components of the ligand-macromolecule interaction energy that show good predictive ability, a regression equation is obtained in which activity is correlated with the interaction energies of parts of the ligands and key regions of the macromolecule. Consequently, COMBINE highlights general interaction types that are (dis)-favorable for ligand binding between ligand parts and subregions of the macromolecule; in contrast, AFMoC emphasizes atom type-specific contributions at the location of ligand atoms, clearly denoting regions where it is either more favorable or disfavorable to place ligand atoms of a given type with respect to binding affinity. We believe that this atom type-specific information is easier to interpret in the light of ligand design than information showing generally (dis)-favorable interactions between a ligand and a target structure. These differences are summarized in Table 1.

So far 3D-QSAR models have been derived by AFMoC analyses for ligands of the DOXP-reductoisomerase,[22] the carbonic anhydrase isoenzymes,[23] and factor Xa.[24] In the first case, a predictive AFMoC model was obtained despite a small set of ligands and a heterogeneous set of crystal structures to work with: The crystal structures either had different loop conformations or missing metal ions or co-substrates resulting in different orientations of co-crystalized antagonists. Still, fusing parts of the structures to form a complete enzyme representing a near-native state yielded the structural information necessary to conduct the AFMoC analysis. Compared to CoMFA and CoMSIA studies on this set of ligands the AFMoC model showed superior predictive power. In particular, AFMoC’s ability to gradually transform between generally applicable unadapted interaction fields to case-specifically adapted ones proved to be of major importance. In line with the small training set, using 50% tailored fields was found to permit the accurate prediction of binding affinities for related ligands without losing the capability to estimate the affinities of structurally distinct inhibitors.[22]

In the study on carbonic anhydrase, the task was to identify ligand features that determine isoenzyme selectivity of 140 ligands. For this, classical 3D-QSAR techniques (CoMFA, CoMSIA), protein-based consensus principal component analysis (CPCA), and AFMoC was applied. Encouragingly, the AFMoC approach showed regions for enhancing ligand selectivity that purely ligand-based methods were unable to detect; this was attributed to the fact that AFMoC, in ad-

### Table 1. Comparison of QSAR approaches.

<table>
<thead>
<tr>
<th>Approach</th>
<th>2D/3D</th>
<th>Target structure</th>
<th>Parameters</th>
<th>Interpretation[a]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hansch-Fujita</td>
<td>2D</td>
<td>Not considered</td>
<td>Lipophilicity, electronic and steric properties</td>
<td>Increase lipophilicity</td>
</tr>
<tr>
<td>analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoMFA</td>
<td>3D</td>
<td>Not considered</td>
<td>Coulomb and van der Waals interactions</td>
<td>Add a bulky, electropositive group at specific position</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Increase van der Waals interaction energy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bonded &amp; non-bonded force-field energies</td>
<td>interactions at specific position</td>
</tr>
<tr>
<td>COMBINE</td>
<td>3D</td>
<td>Considered</td>
<td>Bonded &amp; non-bonded force-field energies</td>
<td>Add amine group at specific position</td>
</tr>
<tr>
<td>AFMoC</td>
<td>3D</td>
<td>Considered</td>
<td>Atom-type specific pair interactions</td>
<td></td>
</tr>
</tbody>
</table>

[a] Example for an interpretation of the results leading to an improved binding affinity of a ligand.
dition to ligand information, also exploits information on structural differences between the isoenzymes.\[23\]

In the case of factor \(\text{Xa}\), a combination of ligand- and structure-based methods including AFMoC was used for guiding ligand design. While both the ligand- and structure-based methods were used to identify effects that play a dominant role in ligand-receptor interactions, the predictive power of only the latter ones was exploited to identify new synthesis candidates with improved affinities towards factor \(\text{Xa}\). As a side goal, the potency towards thrombin was to remain unchanged.\[24\]

2.3 Recent Developments of AFMoC

Recently, the standard AFMoC approach has been extended to take into account dynamical behavior and/or structural inaccuracies of receptor-ligand systems. This resulted in the consensus AFMoC (AFMoC\textsuperscript{con}) approach that considers multiple ligand conformations in an ensemble of protein conformations.\[25\] The ensemble can either be generated by molecular dynamics simulations or result from multiple X-ray structures or homology models of the target. This alleviates the need to decide a priori which target structure to take for deriving an AFMoC model. Rather, as shown for a data set of 79 thrombin inhibitors and three structurally diverse protein models, the AFMoC\textsuperscript{con} approach led to a QSAR model the internal and external predictivity of which was comparable to the best model derived from a standard AFMoC run.\[23\] As to the regression technique, AFMoC\textsuperscript{con} applies partial least-squares regression considering multimode binding (MMB)\[26\] and a variable influence on the model(VINFM)-based region selection\[27\] to an extended descriptor matrix (Figure 2). After an initial PLS analysis considering MMB and a VINFM-based variable selection, a consensus descriptor matrix is generated and subjected to another PLS analysis considering MMB. Note that there is no principal limitation regarding the number of target structure-ligand alignments (TSLA) in this procedure. Furthermore, the approach allows determining the influence of a single TSLA on the overall AFMoC\textsuperscript{con} model, and the AFMoC\textsuperscript{con} model can be interpreted in terms of contour plots that aid in proposing variations of the ligand structure to improve binding.

![Figure 2](https://www.molinf.com)

**Figure 2.** Variation influence on the model (VINFM)-based variable selection for the generation of a consensus descriptor matrix from which an AFMoC\textsuperscript{con} model is derived. Figure taken with permission from *J. Med. Chem.*\[19\]
A second line of development was followed to allow for the use of protein-specifically adapted DrugScore potential fields in docking. This resulted in AFMoC\textsuperscript{(26)}\textsuperscript{(28)} which implicitly takes into account effects due to protein flexibility and information about multiple solvation schemes within a binding pocket. Compared to the application of AFMoC for binding affinity predictions, a Shannon entropy based column filtering of the descriptor matrix and the capping of adapted repulsive potentials within the binding site turned out to be crucial for the success of this method. When applied to a dataset of 66 HIV-1 protease inhibitors, a significant improvement in the accuracy of binding mode prediction was found.

In addition, a significantly higher predictive power of binding affinities for the AFMoC\textsuperscript{(26)} function was observed in this case, compared to if unmodified DrugScore potentials were used. Overall, AFMoC\textsuperscript{(26)} should be a valuable tool for similarity-driven correct binding mode identification, which is a prerequisite for accurate binding affinity prediction and successful virtual screening.

### 3 Outlook

Although the AFMoC approach has proven successful in several validation and application studies, further developments can be envisaged. In particular, considering protein mobility could be improved. So far, AFMoC\textsuperscript{(26)} uses multiple TSLA, which models protein mobility in terms of discrete target conformations to which ligand conformations are adapted. However, the approach does not allow using individually adapted protein-ligand complexes, i.e., those where both the protein and ligand conformations have been mutually fit to another. A potential way to overcome this drawback is to use irregular, deformable 3D potential fields instead of the regular, static ones currently being applied in CoMFA, CoMSIA, or AFMoC. The underlying idea is to adapt a 3D grid with pre-calculated potential field values, which were derived from an initial protein conformation, to another conformation by moving intersection points in space, but keeping the potential field values constant. As before, interaction fields could then be generated as described above (Figure 1). A representation of an irregular, deformable 3D potential grid has been introduced by us recently in terms of modeling the grid as a homogeneous linear elastic body that deforms according to displacements of the surrounding protein atoms.\textsuperscript{(26)} Such a grid representation has been successfully applied for protein- and RNA-ligand docking already.\textsuperscript{(29–30)}

Thinking in the opposite direction, information about protein-ligand interactions that come out from an AFMoC analysis could also be used to improve the quality of homology models of targets. So far, the Modeling Binding Sites using Ligand Information Explicitly (MOBILE) approach\textsuperscript{(31)} already uses such information from an alignment of ligands in a model binding site as restraints in the next cycle of homology modeling. That is, structural information about ligands together with generic DrugScore potentials are applied to adapt the binding pocket region. However, energetic information in terms of experimentally determined binding affinities or activities of the ligands has been neglected so far. This could be overcome by a self-consistent, iterative procedure that combines adaptation of interaction fields for an (initial) alignment of ligands in the model binding site with a subsequent usage of such fields in the MOBILE approach.

Thus, even after almost 50 years of Hansch’ groundbreaking introduction of the concept of QSAR,\textsuperscript{(2)} and after almost 20 years of Hansch thinking to relate polarization effects of an enzyme binding pocket to the rate of ester hydrolysis,\textsuperscript{(32)} and, hence, to consider the effect of a binding pocket for QSAR, the roads that originate from there are still directed towards exciting new horizons.

### Acknowledgements

This study was supported by the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center SFB 974 (“Communication and Systems Relevance during Liver Damage and Regeneration”, Düsseldorf).

### References


Received: February 9, 2012
Accepted: July 10, 2012
Published online: August 7, 2012