# 3D-QSAR

Three-dimensional quantitative structure-activity relationships (3D-QSAR) involve the analysis of the quantitative relationship between the biological activity of a set of compounds and their three-dimensional properties using statistical correlation methods. 3D-QSAR uses probe-based sampling within a molecular lattice to determine three-dimensional properties of molecules (particularly steric and electrostatic values) and can then correlate these 3D descriptors with biological activity.

1. **Molecular shape analysis (MSA)**

Molecular shape analysis wherein matrices which include common overlap steric volume and potential energy fields between pairs of superimposed molecules were successfully correlated to the activity of series of compounds. The MSA using common volumes also provide some insight regarding the receptor-binding site shape and size.

2. **Molecular topological difference (MTD)**

Simons and his coworkers developed a quantitative 3D-approach, the minimal steric (topologic) difference approach. Minimal topological difference use a 'hypermolecule' concept for molecular alignment which correlated vertices (atoms) in the hypermolecule (a superposed set of molecules having common vertices) to activity differences in the series.

3. **Comparative molecular movement analysis (COMMA)**

COMMA – a unique alignment independent approach.The 3D QSAR analysis utilizes a succinct set of descriptors that would simply characterize the three dimensional information contained in the movement descriptors of molecular mass and charge up to and inclusive of second order.

4. **Hypothetical Active Site Lattice (HASL)**

Inverse grid based methodology developed in 1986-88, that allow the mathematical construction of a hypothetical active site lattice which can model enzyme-inhibitor interaction and provides predictive structure-activity relationship for a set of competitive inhibitors. Computer-assisted molecule to molecule match which makes the use of multidimensional representation of inhibitor molecules. The result of such matching are used to construct a hypothetical active site by means of a lattice of points which is capable of modeling enzyme-inhibitor interactions.

**5. Self Organizing Molecular Field Analysis (SOMFA)**

SOMFA – utilizing a self-centered activity, i.e., dividing the molecule set into actives (+) and inactives (-), and a grid probe process that penetrates the overlaid molecules, the resulting steric and electrostatic potentials are mapped onto the grid points and are correlated with activity using linear regression.

**6. Comparative Molecular Field Analysis (COMFA)**

The comparative molecular field analysis a grid based technique, most widely used tools for three dimensional structure-activity relationship studies was introduced in 1988, is based on the assumption that since, in most cases, the drug-receptor interactions are noncovalent, the changes in biological activities or binding affinities of sample compound correlate with changes in the steric and electrostatic fields of these molecules. These field values are correlated with biological activities by partial least square (PLS) analysis.

**7. Comparative Molecular Similarity Indices (COMSIA)**

COMSIA is an extension of COMFA methodology where molecular similarity indices can serve as a set of field descriptors in a novel application of 3d QSAR referred to as COMSIA.

**Scientific Roots of 3D QSAR**

Even before computers, medicinal chemists knew that a set of molecules will typically display an understandable structure–activity relationship. Usually this is manifest in the observation that the smaller the change in the structure of the molecule, the less likely is there to be a change in its biological properties. The similarity principle is another way to say the same thing: compounds with similar chemical and physical properties also have similar biological properties. In QSAR the similarity principle is considered to apply within a series or structural class only, although the pharmacophore hypothesis generalizes the similarity to 3D properties independent of the underlying structure diagrams of the compounds. Another important observation is that the effect on biological activity of changing a substituent at one position of a molecule is often independent of the effect of changing a substituent at a second position, quantified in the early Free–Wilson QSAR method. This has been discussed the previous lecture. Supplanting these qualitative insights by 3D quantitative structure–activity relationships was accomplished by the conscious or unconscious incorporation of insights from many different disciplines.

Structural chemistry provides valuable insights into why changing a substituent on a molecule might change its biological activity. For decades scientists have realized that the three-dimensional arrangement of dispersion, electrostatic and hydrophobic interactions, as well as hydrogen-bonds, determines the strength of intermolecular interactions. Small-molecule crystallography has contributed greatly to our knowledge of the structural aspects of intermolecular interactions. However, only recently have we had the requisite macromolecular structural information, theoretical models and computer power to attempt to forecast macromolecular structure and binding affinity. 3D QSAR capitalizes on these developments and insights of structural and physical biochemistry. Quantum chemistry changes focus from the nuclei of the atoms, the traditional structure, to the electrons of molecule. Today's computers have changed this discipline from one practiced by only devoted experts to one that laboratory chemists can practice or at least set up on their desk-top computer. Although *ab initio* methods remain the benchmark method, semiempirical quantum mechanical methods allow one to calculate fairly accurately the molecular structure and electronic properties of almost any organic molecule — one doesn't need numerous parameters to do so. Recently developed solvation models expand the scope of problems that one can tackle. Although physical organic chemistry traditionally focuses on the rate and equilibrium constants of organic reactions, it has provided both a precedent and an understanding that has been critical to the development of 3D methods. First, it has provided methods for the quantitation of the electronic, steric and hydrophobic effects of substituents on the reaction centre. Second, it demonstrated that multivariate statistical analysis can suggest the physical basis of biological structure–activity relationships, QSAR. It provided the jump-start to combine molecular modelling and statics into 3D QSAR. Molecular modelling in the form of molecular mechanics of small molecules grew from the early hand-held molecular models so useful in conformation analysis. The computer allows the incorporation of electrostatic effects as well as steric ones; the generation and comparison of many conformers of the same molecule; and comparison of the 3D structures of different molecules. Kier pioneered comparing the 3D structures of bioactive molecules to discovering the pharmacophore, the 3D requirements, for a particular biological activity which Marshall later developed into the active analog approach. Lastly, the development of computer graphics provided the platform with which scientists would interact with their structure–activity data. Molecular graphics provides visual insight into 3D structures with colour used to distinguish atoms types and color-coded dot surfaces showing the surface distribution of molecular properties such as electrostatic or hydrophobic potential. It also allows one to

easily compare, by superimposing, different molecules. Most 3D QSAR methods provide some 3D graphics as part of their output. Since 3D QSAR uses insights from so many scientific disciplines, different implementations differ in the concepts and strategies employed. In a perfect world, we would have the requisite understanding to develop a perfect method. In the current world, our scientific understanding is primitive and often qualitative and we continually strive to approximate the truth more closely. Part of the enthusiasm for continued development of 3D QSAR methods is that researchers recognize that each approach has deficiencies in either theoretical background or implementation. This recognition provides the incentive for continuing attempts to improve the methods.

**3D QSAR versus Traditional 2D QSAR**

As noted, computer analysis in the form of linear free energy relationships allowed scientists for the first time to quantitate the relationship between the change in structure of molecules with the change in their biological activity. Traditional QSAR, also known as Hansch-Fujita or 2D QSAR, accurately forecasts the potency of additional compounds and has led to the development of several commercial drugs and pesticides. Statistical analysis distinguishes between steric, hydrophobic and electrostatic effects of substituents on biological activity. This strategy identifies which few of these are the dominant features behind the change in biological properties. When only the statistically important features are considered, a larger number of substituents will be predicted to have the same effect on biological activity. For example, if the QSAR indicates that increasing hydrophobicity leads to increased potency, then both electron-donating and electron-withdrawing substituents can increase potency if they are hydrophobic, and neither will if they are hydrophilic. This is true provided, of course, that the original QSAR was derived from a dataset that included both electron-donating and electron-withdrawing substituents. 3D QSAR methods generalize further to hypothesize that the critical factor is the 3D spatial arrangement of these chemical and physical properties. There are those who conjecture that its structure diagram encodes all the information about the chemical, physical and biological properties of a molecule. In fact, our own studies demonstrated that simple substructure keys are more successful in grouping diverse active compounds together than are more elaborate keys based on 3D structures. Indeed, we found the same trend for the prediction of octanol–water and cyclohexane–water logP, pKa, surface area and a number of other physical properties. Although we have found more sophisticated 3D descriptors that separate actives from inactives more effectively, the impressive performance of simple descriptors must not be ignored. A key difference between traditional and 3D QSAR is the form of the output. Although both provide statistical evidence

for the validity of the proposed relationships, the result of a 3D QSAR analysis is typically supplied as a 3D graphics image superimposed on a molecule of the dataset. This visualization of the results increases the fidelity of the communication between the QSAR modeler and collaborators, such as the synthetic chemists who are interested to see why or if certain molecules are suggested by the model. Another key difference between traditional and 3D QSAR lies in the source of the numerical descriptors of the molecules. In traditional QSAR, one relies on the observed correlation between the effect of a particular substituent on the rate or equilibrium constant for one reaction with the effect of the same substituent on the rate or equilibrium constant for another reaction. Since substituents affect the electronic, steric and hydrophobic properties of molecules, independent parameters are used for each of these properties. The substituent constants themselves are derived from measured effects in model reactions or equilibria. Accordingly, to derive a traditional QSAR equation the scientist or the computer looks up in a table the values of such parameters for each substituent. In contrast, in 3D QSAR one calculates the properties of the molecules of interest. Usually, these properties are calculated in such a way that their 3D distribution is retained in the final model. Although they are appealing because they are measured and not estimated by calculation, a fundamental problem with using measured substituent constants is that the model reactions used to define substituent constants are often themselves only postulated to represent the named feature. This is particularly true of the long-standing argument whether Taft Es values are purely steric, as originally proposed, or whether the measured rate is also influenced by electronic effects. Moreover, recent studies of solvation properties of molecules emphasize that the relative octanol–water partition coefficients of molecules depend on their hydrogen-bonding character, as well as their ‘innate’ hydrophobicities. Thus, the traditional logP is a composite measure of the hydrophobic and hydrogen-bonding properties of the compounds. A practical handicap to using traditional QSAR can be the unavailability of substituent constants for the compounds of interest. Should one then omit those compounds, or guess at the values? Another problem arises when the molecules do not represent a series that can be described by substituent constants. In some cases, overall molecular properties, such as octanol–water logP and calculated pKa, will provide a useful equation. However, this is not always true. Of course, the solution to the difficulty of finding tabulated parameters is to use calculated properties since the definitions are clear and usually all the compounds can be included. However, since this usually involves calculations on the 3D structures of the molecules, why not move directly to 3D QSAR? One must also ask if the calculations are accurate enough to represent such measured properties, a

question answered affirmatively by several workers. A final limitation of traditional QSAR, and a reason why 3D QSAR is considered so attractive by contrast, is that the equations discovered by traditional QSAR do not directly suggest new compounds to synthesize. Rather, one must be experienced with the values of the substituent constants in order to imagine which molecules will have the desired properties. In spite of these limitations, traditional QSAR has contributed greatly to computer assisted molecular design. Many other types of descriptors have been suggested: often these can be directly calculated from the structure diagram of the compounds. Equally important, workers in this field have introduced a wide variety of methods for the quantitative analysis of structure–property relationships. This supplement or replace the traditional multiple regression analysis with statistically based methods such as discriminant analysis, principal components and partial least squares; neural networks; genetic algorithms; and artificial intelligence strategies. Important also is the early recognition that, in order to derive a satisfactory QSAR, one must design the set of compounds carefully, this presages the current interest in diversity analysis and selection of subsets of compound collections. Two early 3D QSAR methods used traditional QSAR descriptors for electronic and hydrophobic effects of substituents, but generate a single steric descriptor by comparing the 3D structures of the molecules with references. Although these methods include 3D properties, they suffer from difficulties in choosing the appropriate reference for the calculation and from ambiguities in how to handle both positive and negative steric influences on potency. An alternative early 3D QSAR method describes the properties of the molecules by their calculated interaction energies with a model of the binding site. Although this method has led to interesting results and enhancements, it was too complex and ambiguous to be adapted for general use. 3D QSAR, as we know it, started with CoMFA. It was invented when Cramer and colleagues recognized that

(i) they could describe, as had others before or simultaneously with them, the 3D distribution of electrostatic and steric properties of molecules by calculating interaction energies on a 3D lattice surrounding the molecules
(ii) they could use partial least squares to extract the relationships between biological potency and these fields; and
(iii) they could produce a visual summary of the QSAR by contouring of the influence of each lattice point to potency.

In the literature up to 1993, CoMFA models reported from 90 biological datasets show the range of R2 fit to be 0.73–1.00, of RMSE fit to be 0.034–0.91 and of RMSE cv to be

0.32–1.52. Although CoMFA overcomes some of the deficiencies of traditional QSAR, new difficulties arise; these will be discussed below. We showed that CoMFA reproduces traditional QSAR descriptors; that is, that a traditional QSAR and a CoMFA analysis provide the same information. Whether traditional or 3D QSAR, only the structure–activity relationships of the ligands contribute to the statistical comparisons. They require no knowledge or hypothesis of the 3D structure or chemical nature of the complementary macromolecule. The comparisons may imply something about this macromolecule, but the implication is by correlation and not direct structural evidence. Although it is not necessary for deriving models, both traditional and 3D QSAR models are usually interpreted as if the common portions of all molecules interact in the same way with the target biomolecule.

**3D QSAR versus Protein-based Affinity Prediction Methods**

The revolution in structural biology means that today the computational chemist often has the 3D structure of the macromolecular binding site with which the ligands of interest interact. Increasing numbers of protein and nucleic acid structures are being solved. As well as being directly useful, these structures supply the basis for homology models of related proteins. Does this make 3D QSAR useless, or do the two approaches complement each other? Knowing the 3D structure of the target makes it easier to perform a 3D QSAR analysis. Many 3D QSAR methods base their property calculation on some absolute orientation of the molecules in space. Usually this means that either the user or the computer program selects the conformation of each molecule to use and how to compare each molecule to the others. Obviously if one has the 3D structure of the macromolecular target, particularly if one also has the structure of at least one ligand of each series bound to the protein, then it will be easier to propose a bioactive conformation and superposition rule. The location of key binding sites should help suggest an orientation for the other molecules of interest. One could also directly observe the structure of the complex crystallographically, or optimize a model to provide a bioactive conformation. Is 3D QSAR necessary if one has a 3D structure of the protein on which to base predictions? Much attention has been paid recently to perturbation free energy method of predicting protein–ligand affinity. Although this method is based on solid theoretical foundations, in practice such calculations involve days to weeks of computer time per pair of ligands and are limited to calculating affinity differences resulting from rather modest differences in structure. Their accuracy is probably limited by the approximations used in the force fields and electrostatic calculations: greater computer power and deeper insight

into the biophysics of macromolecular structure may result in improved precision of calculations. A more recent method, Linear Interaction Energy calculations, combines features of perturbation free energy calculations and QSAR to produce simple equations in steric and electronic energy using only three to four compounds. The calculation on each ligand requires less than a day of computer time. In some research, four compounds were used to determine a regression equation that predicted the affinity of seven structurally different compounds with a mean error of 0.55 kcal/mol. Clearly, this method deserves watching: it currently would be useful for predicting the potency of a handful of compounds, more if several computers were available and as computer speeds increase. However, its limitations are also becoming known: both errors in prediction and correct predictions of affinity based on the wrong structure of the complex. Another approach to using protein structures to predict binding affinity involves deriving generalized QSAR equations that predict the strength of any protein–ligand complex. They are used mainly in the computer *de novo* design and docking of ligands. The descriptors for each ligand are calculated from an experimental 3D structure of a complex. Typically, they include features such as the number and quality of the intermolecular hydrogen-bonds, as well as electrostatic, dispersion and hydrophobic interactions and an estimate of the ligand entropy lost on binding. A universal model is derived by regression or PLS analysis of dissociation constants of a variety of protein–ligand complexes using many different proteins. Once a model is derived, it can be used quickly to predict the affinities of any ligand interacting with any protein. Forecasts from these empirical equations are less precise than from perturbation or linear interaction energy analysis, typically of the order of 1.3 log units. A problem with these approaches is that steric misfit is not explicitly included since such molecules will bind in another configuration. In contrast, all QSAR methods include explicit terms that reflect steric misfit. In yet another approach to using the structure of a protein–ligand complex as a basis of a QSAR analysis, several groups have used molecular descriptors derived from energy minimization of docked ligands with a target protein. Either the calculated interaction energy or separated components of the interaction energy are correlated with affinity. Sometimes other properties, such as estimates of the relative entropy cost of binding the ligand, are added to the prediction equation. Interestingly, the cross-validation statistics suggest that these equations are approximately of the same precision as typical equations derived without knowledge of the protein structure. One problem with this approach may be that since the force fields are parameterized to reproduce the structure and dynamics of a single compound, they

may be deficient in the treatment of solvation energy. This varies more dramatically between compounds than between different conformations of the same compound. Additionally, the parameter values for the types of atoms of the ligands may not have been as carefully established: it appears that especially assigning values for the partial atomic charges may present a problem. An emerging method to predict binding energy is based on the observed preferences of certain types of atoms to be near each other in macromolecular complexes. The accuracy appears to be approximately the same as the generalized QSAR equations. The main limitation of this approach, at the moment, is the limited numbers of better than 2.0Å resolution protein–ligand complexes available compared to the number of atom types present in drug molecules and the number of examples of each that would be needed to derive a preference score. This survey suggests that 3D QSAR methods are an important complement to structure-based affinity prediction methods. If one already has a series of molecules and their corresponding binding affinities, then a 3D QSAR equation may provide a valuable method to forecast affinity of further analogs. Knowledge of the structure of the binding site would guide the molecular modelling and should prevent unwarranted extrapolation of such equations. At the moment, the observed structure–activity relationships of ligands provide a more sensitive measure of ligand–receptor affinity than do computational methods. On the other hand, structure-based calculations of affinity can be done, even if one has no or limited structure–activity and if the suggested compounds are very different from any known ligands.

**References**

1. Kim, K.H., Greco, G. and Novellino, E., *A critical review of recent CoMFA applications*, In Kubinyi, H., Folkers, G., and Martin, Y.C., (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 257–316.

2. Dunn III, W.J. and Hopfinger, A.J., *3D QSAR of flexible molecules using tensor representation*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 167–182.

3. Hahn, M. and Rogers, D., *Receptor surface models*, in Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 117–134.

4. Heritage, T.W., Ferguson, A.M., Turner, D.B. and Willett, P., *EVA — a novel theoretical descriptor for QSAR studies*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 381–398.

5. Klebe, G., *Comparative molecular similarity indices analysis — CoMSIA*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 87–104.

6. Walters, D.E., *Genetically evolved receptor models (GERM) as a 3D QSAR tool*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 159–166.

7. Wade, R.C., Ortiz, A.R. and Gago, F., *Comparative binding energy analysis*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 19–34.

8. Holloway, M.K., *A priori predicition of ligand affinity by energy minimization*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 63–84.

9. Todeschini, R. and Gramatica, P., *New 3D molecular descriptors: The WHIM theory and QSAR applications*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 355–380.

10. Silverman, B.D., Platt, D.E., Pitman, M. and Rigoutsos, I., *Comparative molecular moment analysis (COMMA)*, in Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 183–196.

11. Jain, A.N., Koile, K. and Chapman, D., *Compass: Predicting biological activities from molecular surface properties — performance comparisons on a steroid benchmark*, J. Med. Chem., 37 (1994) 2315–2327.

12. Martin, Y.C., Kim, K.-H. and Lin, C.T., *Comparative molecular field analysis: CoMFA*, In Charton, M. (Ed.) Advances in quantitative structure property relationships, JAI Press, Greenwich, CT, 1996, pp. 1–52.

13. Greco, G., Novellino, E. and Martin, Y.C., *Approaches to 3D-QSAR*, In Martin, Y.C. and Willett, P. (Eds.) Designing bioactive molecules: Three-dimensional techniques and applications, America Chemical Society, Washington, DC, 1997 (in press).

14. Ajay and Murcko, M.A., *Computational methods to predict binding free-energy in ligand–receptor complexes*, J. Med. Chem., 38 (1995) 4953–4967.

15. Kollman, P.A., *Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules*, Acc. Chem. Res., 29 (1996) 461–469.

16. Bush, B.L. and Nachbar Jr., R.B., *Sample-distance partial least-squares — PLS optimized for many variables, with application to CoMFA*, J. Comput.-Aided Mol. Design, 7 (1993) 587–619.

17. Burger, A., *Medical chemistry — the first century*, Med. Chem. Res., 4 (1994) 3–15.

18. Willett, P., *Similarity and clustering techniques in chemical information systems*, Research Studies Press, Letchworth, 1987.

19. Hodgkin, E.E. and Richards, W.G., *Molecular similarity based on electrostatic potential and electric field*, Int. J. Quantum Chem., 14 (1987) 105–110.

20. Kier, L.B., *Molecular orbital theory in drug research*, Academic Press, New York, 1971, p. 258.