4D-QSAR Models in Cheminformatics

# Introduction

Quantitative Structure–Activity Relationship (QSAR) models are statistical models that predict bioactivity of ligands given numerical measures of their chemical structure, or chemical descriptors. These descriptors often have different levels of organization, which can be expressed in terms of dimensionality - starting at 1D and reaching as high as 6D according to some authors.

For descriptors at dimensions lower than 3D, the conformation of the chemical is irrelevant. However, for 3D descriptors, a single “active” conformation of each ligand is assumed and descriptors of the ligands three dimensional structure are computed (e.g., size, shape, protein-ligand interactions). Computing predictive 3D descriptors can be challenging as these are often highly sensitive to the conformation of the ligand (e.g, COMFA descriptors1) and an incorrect binding pose of the ligand will lead to an uninformative QSAR model. Identifying the correct orientation of each ligand is often challenging, as docking can often result in distinct, energetically favorable binding poses2. Furthermore, ligands are not static objects. Both ligands and protein structures have considerable conformational flexibility. For example, kinase binding pockets are known to be especially flexible3,4.

In response to the limitations of 3D descriptors, 4D QSAR methods have been developed that consider variation in ligand conformation (as well as other variations that may affect 3D descriptor values, such as orientation, protonation states and stereoisomers). The central hypotheses of these approaches is that by considering the variation in the spatial features of protein-ligand complexes, the non-covalent interactions between a protein and ligand may be better characterized, leading to a more highly predictive QSAR model. The main advantage of these methods is that they can result in models highly predictive of ligand binding affinity, while the main disadvantage is that they are computationally expensive. The 4D QSAR level of accuracy may not be necessary during the high throughput virtual screening stages of drug discovery, but it can be essential at the lead optimization stage, where a few hundred leads have been identified and the task is to identify the small structural modifications among analogs that result in the highest target affinity.

# Hopfinger’s 4D-QSAR Formalism

In his seminal paper, Hopfinger et al5 presented a concept of 4D-QSAR that has served as a framework for the field ever since. The key idea in this paper was that 3D-QSAR predictions could be improved by representing the ligand as an ensemble of conformations, foregoing the assumption that the active conformation of a ligand was known. These models are capable of learning the true active conformation from the ensemble. Two main variations of these methods exist: Receptor Independent (RI) 4D-QSAR and Receptor Dependent (RD) QSAR6.

In RI 4D-QSAR, information about structure of each protein-ligand complex is either unknown or not utilized. Short run (100 ps) molecular dynamics (MD) simulations are performed for each ligand with the intent of approximating the Boltzman sampling distribution for their conformations7. Conformations are sampled along the MD simulation and aligned on a grid. Then, the presence/absence of interaction pharamacophore elements (IPEs) are computed at each grid cell. The IPEs are essentially ligand atom types, representing their polarity, hydrophobicity, capacity to form a hydrogen bond, etc. Since the number descriptor variables this generates is large and the sample size is small (i.e., p >> n), data reduction is performed using partial least squares (PLS). A genetic algorithm is used to perform variable selection for a linear model relating the PLS components and the response. The end result is a highly interpretable multiple linear regression model relating a few grid cell IPEs and ligand binding affinity. These IPEs can be interpreted as a simple pharmacophore hypothesis.

RD 4D-QSAR performs the exact same protocol, except that information about the constraints that are placed on the ligand conformation by its target’s binding pocket is utilized. Each ligand is docked in the target crystal structure before MD simulation, and MD is performed on the pruned protein-ligand complex.

Several applications of these models have illustrated their predictive power for a variety of targets, such as p38-MAPK8, HIV-1 protease9, and Glyocogen Phosphorlyase10. However, the data sets analyzed in these applications were small (under 50 compounds in the training set and less than 10 in the test set). As a representative example, the Glycogen Phosphorlyase study10 produced a RI 4D-QSAR model with excellent prediction performance on held out data. A virtual screen on 225 new compounds in the same analog series resulted in an astonishing R2 of .82. While this virtual screen was by no means high throughput, the results do demonstrate that these models can retain their predictive performance for a sizeable held out data set, albeit for only compounds in the same chemical series. A co-crystalized protein-ligand complex was available, but was not used in the model building process, affording an opportunity to validate the pharmacophore hypothesis. The pharmacophore hypothesis recapitulated some of the binding mode observed in the crystal structure.

Another notable application of Hopfinger’s 4D-QSAR formalism was in the prediction of pharmodynamics. A 4D-QSAR model has been utilized to predict a set of ligands’ ability to cross the blood brain barrier.

The number of successful applications of Hopfinger’s 4D-QSAR protocol is due to a few of its important advantages. First, they generate a large number of variables that contain information about the conformational variation of each ligand within its target’s binding pocket. This provides an extra dimension of structural information about a ligand that can be related to its activity. Next, PLS data reduction and a genetic algorithm intelligently searches for the minimum number of variables necessary to predict ligand activity, negating overfitting. Finally, the resulting models are highly interpretable as a pharmacophore hypothesis.

The main problem with Hopfinger’s 4D-QSAR is that the methods employed are at least two decades old and have limitations that could be addressed by more current technology. For example, the force field employed for MD is MM25,11 which is simple fore field parameterized for hydrocarbons. MD simulations are only run for 100 ps without explicit solvent molecules and when protein-ligand complexes are utilized they are pruned. No information about computational time required to train these models. However, the fact that no applications found involved training set larger than 50 compounds suggests the protocol is computationally expensive. In particular, the genetic algorithms that are used for variable selection can be computationally costly.

# LQTA 4D-QSAR

More recently, Martins et al.12 developed LQTA-QSAR, which combined Hopfinger’s 4D-QSAR formalism with comparative molecular field analysis (CoMFA13), a widely used 3D descriptor computation method. LQTA-QSAR generates conformational ensembles and superimposes on a grid as Hopfinger et al5 do, but instead of grid occupancy descriptors, LQTA QSAR uses the CoMFA14 to compute interaction energies for each grid cell.  Different functional groups (e.g carboxyl groups, cations, and anions) are used as probes at each grid point where a sum of interaction energies is computed. 500 ps receptor independent MD simulations are performed to create conformational ensembles. A more recent MD simulation method (GROMACS with explicit solvent and ff43a1 force field parametrization) is utilized. QSAR models are constructed using the grid cell interaction energies and a PLS regression model with an in house ordered predictor selection algorithm15 for variable selection.

There are two main advantages of this approach over Hopfinger et al’s 4D-QSAR. The MD simulations are conducted with greater biophysical realism with a more appropriate force field parametrization and explicit solvent. Also, the use of CoMFA to compute grid cell descriptors carries more information about interaction energies than Hopfinger’s grid cell occupancy descriptors. However, this change in methodology provides its own disadvantages. CoMFA is very sensitive to ligand conformations, suggesting a tendency in this method to over fit the data and sensitivity to the subjective decision of the method used to align ligand conformations. Also, this method has only been developed for receptor interdependent 4D-QSAR so receptor ligand interactions are not considered. Only 500 ps MD simulations are performed and, again, as far as we are aware thorough analysis of computation time was not conducted.

# Quasar 4D to 6d-Qsar

The Quasar software developed by Vedani et al.16 approaches 4D-QSAR from a different perspective than Hopfinger’s original formalism. Quasar uses a "quasi-atomic" representation of the receptor with atom like properties (hydrogen bonds, hydrophobic particles and virtual solvent) to compute interaction energies between the ligand and the receptor.  The traditional 4D-formalism is also employed to some extent, where an ensemble of conformations, orientations and protonation states are considered for each ligand.  Quasar is unique in that it simulates a local induced fit by adapting the van der Waal's surface to each ligand. Relative free energies of ligand binding are estimated by as a function of several terms including the force-field energy of ligand binding and ligand desolvation energy. Then, a genetic algorithm is use to find the optimal model among the family of receptor ligand models. Vidani et al extended this model further by considering an ensemble of conformations for the quasi-atomic receptor surrogate and termed this approach 5D-QSAR17. The advantage of this approach is that large changes in the conformation of the protein are modeled such as the Apo (unbound) and Holo (Bound) conformations occurring during induced fit binding.  These large conformational changes would be difficult to model with MD simulations alone. They extended this approach even further and incorporated an ensemble of solvation models which they call 6D-QSAR, which was being successful at predicting SARs in a few applications18.

# ConClusions

4D QSAR models can better forecast ligand activity because they consider another level of ligand structural organization. As Tseng et al19 point out, many different descriptor sets should be considered when constructing a QSAR model. It is often unclear which descriptor set is going to perform the best for a particular data set, QSAR models should be fit to many different combinations of descriptor sets and some validation on held out data used to assess model performance. Often 1D and 2D descriptors are not combined with 3D and 4D in QSAR models, while there is no particular reason to avoid these combinations. Avoiding combining descriptors at many different levels of organization may miss important structural features related to ligand activity.

Several other 4D-QSAR methods have recently been developed20–22. Collectively, the main limitation for these approaches is their computational time required to train such models. This is due to the fact that molecular dynamics simulations need to run for each ligand. To compensate for this unrealistic parameterizations of MD simulations are selected for ease of simulation. For example, explicit solvent is not included in the simulation or a simple water model is used. Also, MD simulations are run for short periods of time (usually on the order of picoseconds), which may limit the amount of conformational variability in the ligand that can be sampled. Even after making these concessions, these 4D-QSAR approaches typically only have their predictive power validated on a data set with less than 50 compounds6,12,17. There is a need for more thorough assessment of how well these methods will perform when analyzing large, structurally diverse data sets.

However, this methodology does present some exciting new opportunities for drug discovery. It has been repeatedly demonstrated that these models can be highly predictive models of pki or IC50 in applications relevant to lead optimization. 4D-QSAR can better characterize dynamic noncovalent protein–ligand interactions and thus build target-specific QSAR models with enhanced prediction performances. 4D-QSAR can also reduce the bias of selecting a single conformation out of a set energetically similar ones by using an conformational ensemble to identify binding modes. This reduces the sharp dependency on single ligand conformation. We also have shown in a recent paper23 that MD descriptors are useful in resolving activity cliffs that are present in other descriptor spaces.

# References

(1) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J. C.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz, V. E.; Cramer, R. D.; Benigni, R.; Yang, C.; Rathman, J. F.; Terfloth, L.; Gasteiger, J.; Richard, A. M.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**.

(2) Lill, M. A. Multi-Dimensional QSAR in Drug Discovery. *Drug Discov. Today* **2007**, *12* (23).

(3) Mazanetz, M.; Laughton, C.; Fischer, P. Investigation of the Flexibility of Protein Kinases Implicated in the Pathology of Alzheimer’s Disease. *Molecules* **2014**, *19* (7), 9134–9159.

(4) McClendon, C. L.; Kornev, A. P.; Gilson, M. K.; Taylor, S. S. Dynamic Architecture of a Protein Kinase. *Proc. Natl. Acad. Sci.* **2014**, *111* (43), E4623–E4631.

(5) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism.

(6) Andrade, C. H.; Pasqualoto, K. F. M.; Ferreira, E. I.; Hopfinger, A. J. 4D-QSAR: Perspectives in Drug Design. *Molecules* **2010**, *15* (5), 3281–3294.

(7) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119* (43), 10509–10524.

(8) Correia Romeiro, N.; Albuquerque, M. G.; Bicca De Alencastro, R.; Ravi, M.; Hopfinger, A. J. Construction of 4D-QSAR Models for Use in the Design of Novel P38-MAPK Inhibitors.

(9) Santos-Filho, O. A.; Hopfinger, A. J. Structure-Based QSAR Analysis of a Set of 4-Hydroxy-5,6-Dihydropyrones as Inhibitors of HIV-1 Protease:  An Application of the Receptor-Dependent (RD) 4D-QSAR Formalism. *J. Chem. Inf. Model.* **2006**, *46* (1), 345–354.

(10) Venkatarangan, P.; Hopfinger, A. J. Prediction of Ligand-Receptor Binding Free Energy by 4D-QSAR Analysis: Application to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase.

(11) Allinger, N. L. Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. *J. Am. Chem. Soc* **1977**, *99* (25), 8127–8134.

(12) Paulo, J.; Martins, A.; Barbosa, E. G.; Pasqualoto, K. F. M.; Ferreira, M. M. C. LQTA-QSAR: A New 4D-QSAR Methodology.

(13) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc* **1988**, *110*, 5959–5967.

(14) Damale, M. G.; Harke, S. N.; Kalam Khan, F. A.; Shinde, D. B.; Sangshetti, J. N. Recent Advances in Multidimensional QSAR (4D-6D): A Critical Review.

(15) Teófilo, R. F.; Martins, J. P. A.; Ferreira, M. M. C. Sorting Variables by Using Informative Vectors as a Strategy for Feature Selection in Multivariate Regression. *J. Chemom.* **2009**, *23* (1), 32–48.

(16) Vedani, A.; Dobler, M.; Zbinden, P. Quasi-Atomistic Receptor Surface Models: A Bridge between 3-D QSAR and Receptor Modeling.

(17) Vedani, A.; Dobler, M. 5D-QSAR: The Key for Simulating Induced Fit?

(18) Vedani, A.; Dobler, M.; Lill, M. A. Combining Protein Modeling and 6D-QSAR. Simulating the Binding of Structurally Diverse Ligands to the Estrogen Receptor.

(19) Tseng, Y. J.; Hopfinger, A. J.; Esposito, E. X. The Great Descriptor Melting Pot: Mixing Descriptors for the Common Good of QSAR Models. *J. Comput. Aided. Mol. Des.* **2012**, *26* (1), 39–43.

(20) Wieder, M.; Garon, A.; Perricone, U.; Boresch, S.; Seidel, T.; Almerico, A. M.; Langer, T. Common Hits Approach: Combining Pharmacophore Modeling and Molecular Dynamics Simulations.

(21) Dreher, J.; Scheiber, J.; Stiefl, N.; Baumann, K. XMaP—An Interpretable Alignment-Free Four-Dimensional Quantitative Structure–Activity Relationship Technique Based on Molecular Surface Properties and Conformer Ensembles. *J. Chem. Inf. Model.* **2018**, *58* (1), 165–181.

(22) Bak, A.; Kozik, V.; Smolinski, A.; Jampilek, J. Multidimensional (3D/4D-QSAR) Probability-Guided Pharmacophore Mapping: Investigation of Activity Profile for a Series of Drug Absorption Promoters. *RSC Adv.* **2016**, *6* (80), 76183–76205.

(23) Ash, J.; Fourches, D. Characterizing the Chemical Space of ERK2 Kinase Characterizing the Chemical Space of ERK2 Kinase Inhibitors Using Descriptors Computed from Molecular Dynamics Trajectories. *J. Chem. Inf. Model.* **2017**.