

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# CoMFA/CoMSIA and Pharmacophore Modelling as a Powerful Tools for Efficient Virtual Screening: Application to Anti-Leishmanial Betulin Derivatives

Leo Ghemtio, Yuezhou Zhang and Henri Xhaard  
Centre for Drug Research, Faculty of Pharmacy, University of Helsinki,  
Finland

## 1. Introduction

Improvement of efficiency as well as speed and accuracy in the step of identification of chemicals that excerpt *in vitro* or *in vivo* activity would help reduce the huge investments made by pharmaceutical companies in drug development projects. Traditionally, *in silico* high-throughput screening techniques, either based on protein binding site fitting (docking) or ligand similarity, are used to select the most promising molecules from large chemical libraries. (Ling & Xuefeng, 2008; Stahura & Bajorath, 2004; Tuccinardi, 2009; Villoutreix et al., 2009) Nonetheless, these computational techniques are hampered by high rates of false positives and high demand in computational resources. (Ghemtio et al., 2010) To avoid these shortcomings, predictive three-dimensional (3D) quantitative structure–activity relationship (3D-QSAR) combined with pharmacophore computational models coupled to ligand-based three-dimensional virtual screening (3D-VS) are becoming increasingly popular. (Clark, 2009; Ekins et al., 2007; Ghemtio et al., 2010; Kirchmair et al., 2008; Kirchmair et al., 2008; Langer & Hoffmann, 2001; Lengauer, 2004; Rognan, 2010; Sippl, 2002; Spitzer et al., 2010; Tropsha & Golbraikh, 2007). For example, previously, combination of 3D-QSAR studies and 3D-VS have been successfully applied for screening large collection of natural products and synthetic chemicals. (Clark, 2009; Liu et al., Nagarajan et al., 2010; Sippl, 2002; Spitzer et al., 2010)

The general idea is that after building a 3D-QSAR model that predicts usually the binding constant or *in vitro* biological activities of compounds from their 3D chemical properties, the 3D pharmacophoric representation of the shared chemical features that are most important towards activity can be used as a constraint for 3D-VS. Generally, a binding constant is accurately measured experimentally, relates to a single type of molecular event, and therefore is a suitable source of data for 3D-QSAR modeling. There is however a large gap between the binding constant to given protein and any therapeutic effect that may be provided by a compound. *In vitro* activities relate in many cases to a compound binding to several target proteins or to other cellular effects but can still be useful for 3D-QSAR. *In vivo* activities on the other hand are the sum of so many complex and mechanistically different processes that they are not a reasonable source of data for 3D-QSAR modeling.

Here, we aim to interpret the *in vitro* anti-leishmanial activities of a set of 24 betulin derivatives (BDIs), i.e. compounds derived from a betulin scaffold, as well as to screen for

novel potentially interesting chemicals. Leishmaniasis are diseases caused by protozoan parasites that affect millions of people in more than 88 countries worldwide.(Alakurtti et al., 2010) Several drugs are available for the treatment of these diseases, for example pentavalent antimony compounds derived from the heavy metal antimony (Sb), pentamidine or amphotericin B, and miltefosine compounds. However, these drugs present severe side effects, parasite resistance, are too expensive for use in less-developed countries, and for some are dangerous to use in pregnant women.(Pink et al., 2005) There is therefore an urgent need for the development of safe chemicals for the treatment of all clinical forms of leishmaniasis. For this purpose, betulin derivatives are one of the most investigated classes of compounds. While the molecular mechanism of the inhibitory action of betulin derivatives on *Leishmania donovani* growth is to date unknown, several protein targets have been suggested including the Topoisomerase 2 enzyme. Betulins present several advantages that make them a very suitable class of compounds to run quantitative structure–activity relationship (SAR) studies and, despite their large size and hydrophobicity, to be investigated as a therapeutic class of compounds: a five-ring chemical scaffold allows a straightforward three dimensional superimposition, while the parent molecule can be extracted easily and in large quantities from the bark of birch tree, and is easily chemically modifiable at three sites. In addition to anti-Leishmania activity, betulin derivatives have shown anti-inflammatory, antimalarial and especially cytotoxic activity against several tumor cell lines by inducing apoptosis in cells.(Alakurtti et al., 2010; Alakurtti et al., 2006) Structure–activity relationship studies and pharmacological properties of betulin and its derivatives have been reviewed recently.(Alakurtti et al., 2006)

In this chapter, we have developed predictive 3D-QSAR models that help to interpret the *in vitro* anti-leishmanial activities of a small but consistent set of 24 betulin derivatives (the chemical structures are shown in Table 1). (Alakurtti et al., 2010). We first use two popular and well-studied 3D-QSAR methods; comparative molecular field analysis (CoMFA)(Cramer et al., 1988), and comparative molecular similarity indices analysis (CoMSIA)(Klebe et al., 1994) implemented in Sybyl-X to construct predictive 3D-QSAR models that predict the activities of betulin derivatives from a small but consistent dataset. In these methods the proper alignment of molecular structures across the series and the selection of the bioactive conformation are critical yet often problematic. The 3D-QSAR models developed here should serve as a useful tool to predict the inhibitory properties of untested compounds and therefore help to guide synthesis for the further development of more potent anti-leishmanial inhibitors. Secondly, the 3D-QSAR models together with compound 3D structures were used to develop 3D pharmacophore models that describe the chemical features most important for activity, using the GALAHAD (Genetic Algorithm with Linear Assignment of Hypermolecular Alignment of Database)(Richmond et al., 2006) implemented in Sybyl-X. Only the five most active molecules well predicted by the 3D-QSAR models were used for pharmacophore development. Thirdly, these pharmacophore models were used as 3D constraints to query two libraries for new chemical structures, one containing 120.000 compounds and the other containing 240.000 compounds.

## 2. Materials and methods

### 2.1 Compounds and biological data

The molecular structures and biological data used in this study were retrieved from a series of 24 betulin derivatives developed by Alakurtti et al. The chemical structures and

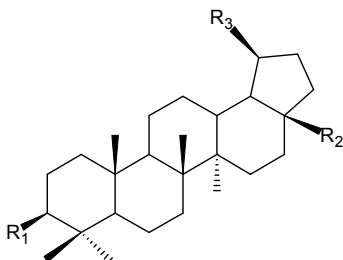
experimental activities are shown in Table 1.(Alakurtti et al., 2010) The biological activities are reported as the percent inhibition of *Leishmania donovani* axenic amastigotes growth at 50  $\mu$ M of betulin derivatives and were used as dependent variables in this study. These represent the percentage of growth reduction of *Leishmania donovani* axenic amastigotes associated with adding 50  $\mu$ M of betulin derivatives to the cells.

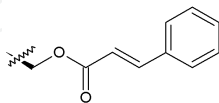
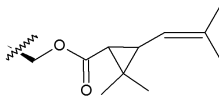
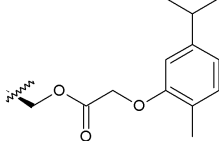
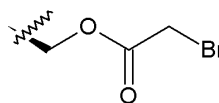
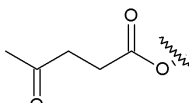
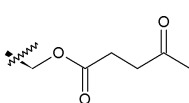
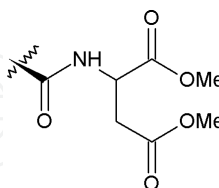
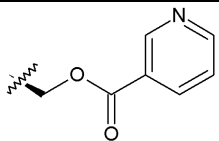
Other type of data, such as growth rate constants, would have been perhaps been more reliable for 3D-QSAR modeling but were not accessible to us at time of this study.

All pharmacological data were obtained from the same laboratory, eliminating the potential noise that might have been introduced by the pooling of data sets from different sources. The inhibition ( $I_{50}$ ) percentage values were converted in negative logarithmic units ( $pI_{50}$ , M) using Sybyl-X. The CoMFA/CoMSIA models were developed using 16 compounds as training set, and externally validated using 8 compounds as test set (see Table 1). The compound set was randomly divided into a training set and a test set (distributed using a 2/3 and 1/3 rule). After this division, we checked that both sets represent equally well the chemical and biological properties of the whole data set. The range of  $pI_{50}$  values for both the training and test set spans at least three orders of magnitude (2.30-5.91), and in addition the biological activity values are well distributed over the entire range. The compound 18 is later used as a reference, since it is the most active.

Accounting for outliers, either activity outliers, i.e. similar compounds for which different activities have been recorded, or leverage outliers, i.e. compounds chemically dissimilar from the rest of the set, is an important step in any type of QSAR modelling. One of the main deficiencies of some chemical datasets is that they do not satisfy the hypothesis that similar compounds share similar biological activities or properties. Outliers may originate from genuine effects, i.e. activity cliffs, may be due to artifacts and errors in structure representation, may result from a poor identification of chemical similarity, or may come from a poor annotation of biological activity. In addition, outliers may originate from different molecular mechanisms of action that may involve seemingly similar compounds. Outlier detection and removal before proceeding to model development is the best way to avoid model instability with significant differences in external predictive power of models. (Tropsha, 2010).

In this study, the compounds in the dataset are based on the same betulin chemical scaffold and therefore should be chemically similar one to another. As we will show below, there is a chemical reason to suspect that a few compounds can use a mechanism of action based on formation of covalent adducts, a mechanism that is quite different from the other molecules in the set. Removing these outliers from training and test set clearly led to improved models. Interestingly enough, we first noticed these compounds being outliers based on CoMFA/CoMSIA modeling and only afterwards identified a reasonable molecular explanation for this behaviour.



Compound	R1	R2	R3	pI <sub>50exp</sub>	Prediction	
					CoMFA pI <sub>50pred</sub>	CoMSIA pI <sub>50pred</sub>
Training set						
1*	OH	CH <sub>2</sub> OH	CH <sub>3</sub> -C=CH <sub>2</sub>	4.03	4.11	3.94
2*	OH	CO <sub>2</sub> H	CH <sub>3</sub> -C=CH <sub>2</sub>	4.12	4.03	4.11
4*	OH	CHO	CH <sub>3</sub> -C=CH <sub>2</sub>	4.55	4.60	4.49
6	OH		CH <sub>3</sub> -C=CH <sub>2</sub>	2.30	-	-
9	OH		CH <sub>3</sub> -C=CH <sub>2</sub>	3.49	3.44	3.49
10	OH		CH <sub>3</sub> -C=CH <sub>2</sub>	3.60	3.69	3.60
12	OH		CH <sub>3</sub> -C=CH <sub>2</sub>	5.08	-	-
14	OAc	CH <sub>2</sub> OAc	CH <sub>3</sub> -C=CH <sub>2</sub>	2.30	2.30	2.20
15			CH <sub>3</sub> -C=CH <sub>2</sub>	2.30	2.19	2.42
17*	O=	CHO	CH <sub>3</sub> -C=CH <sub>2</sub>	4.23	4.72	4.16
18*	O=	CO <sub>2</sub> H	CH <sub>3</sub> -C=CH <sub>2</sub>	5.91	4.72	5.79
20*	O=	CO <sub>2</sub> Me	CH <sub>3</sub> -C=CH <sub>2</sub>	4.12	4.67	4.28
21	O=		CH <sub>3</sub> -C=CH <sub>2</sub>	4.65	4.81	4.67
23*	-	CH <sub>2</sub> OH	CH <sub>3</sub> -C=CH <sub>2</sub>	3.48	3.82	3.41
24*	OH	CH=NOH	CH <sub>3</sub> -C=CH <sub>2</sub>	4.65	4.59	4.77
25*	=NOH	CH=NOH	CH <sub>3</sub> -C=CH <sub>2</sub>	4.73	4.44	4.80
Testing set						
7	OH		CH <sub>3</sub> -C=CH <sub>2</sub>	3.28	3.91	3.75

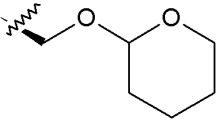
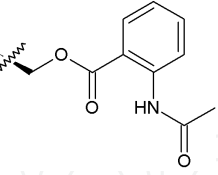
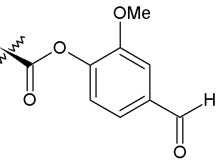
8	OH		CH3-C=CH2	3.37	3.62	3.83
11	OH		CH3-C=CH2	4.46	4.63	4.04
13*	OAc	OH	CH3-C=CH2	4.07	4.05	3.95
16*	O=	CH2OAc	CH3-C=CH2	4.13	4.28	3.85
19*	O=	CO2H	CH3CHCH2	4.71	4.34	5.01
22	O=		CH3-C=CH2	2.30	2.64	2.34
26*	OAc	CN	CH3-C=CH2	4.52	4.18	4.05

Table 1. Experimental and predicted Leishmanial growth inhibitory activities of Betulin derivatives used in the study. The first group is the training set and the second group is the test set. Compounds used for pharmacophore modelling are shown in bold. When the set was used as control (see section 3.3) in pharmacophore based virtual screening 3D search, the compounds marked with (\*) were retrieved.

## 2.2 Generating the molecular structures and conformational analysis

The molecular structures of betulin derivatives were sketched using Sybyl-X v1.2 software(Tripos International, St. Louis). The fragment libraries in Sybyl-X database were used as building blocks to build three-dimensional structures of functional groups added to the betulin scaffold. A single conformation for each chemical was randomly picked, but it should be noted that conformational space is available for the R groups while the betulin scaffold is rigid. All the structures were assigned Gasteiger-Huckel charges and energy minimized using the standard Tripos force field (Powell method and 0.05 kcal/(mol.Å) energy gradient convergence criteria) (Tripos International, St. Louis). These conformations were used as starting conformations to perform the following 3D-QSAR and pharmacophore studies.

## 2.3 3D-QSAR models

3D-QSAR methodologies such as CoMFA/CoMSIA, aim to correlate biological activities with the three dimensional structures of compounds. Among these, comparative molecular field analysis (CoMFA) is widely used and historically the first, and has been improved since. CoMFA is restricted to electrostatic fields and therefore accounts only for the enthalpic contribution of binding(Klebe et al., 1994), with the underlying idea that if the aligned molecules share global shape and location in the 3D lattice, the entropic contributions to the free energy of binding to a molecular target are expected to be similar.(Perkins et al., 2003) The CoMFA Lennard-Jones and Coulomb potentials are sharp



and may introduce errors in scaling, alignment sensitivity, and interpretation of contours. (Bostrom et al., 2003) In order to improve these shortcomings, the comparative molecular similarity indices (CoMSIA) methods have been developed that make usage in addition to the electrostatic fields of hydrophobic fields, supposed to account better for differences in the entropic contribution to binding free energy, hydrogen bonding fields, as well as use smoother potentials based on Gaussian functions, which are less sensible to variation in alignment and lead to more interpretable contours. (Buolamwini & Assefa, 2003)

### **CoMFA/CoMSIA alignment rules**

The three-dimensional alignment of chemical structures is one of the most important steps in 3D-QSAR methodologies. For a set of congeneric chemicals, an optimal alignment of a set of molecules can be defined as the alignment that achieves the maximum superposition of steric and electrostatic fields. In CoMFA/CoMSIA modeling, significant and relevant results should be expected only for valid alignments. There are multiple strategies available in the literature depending on the specificity of each dataset for compound alignment as well as resources. Commonly used among commercial solutions are Sybyl-X 1.2 database alignment, Sybyl-X 1.2 atom fit alignment, SYBYL-X 1.2 Surflex-Sim(Jain, 2004), BRUTUS(Rönkkö et al., 2006; Tervo et al., 2005), or for freely available softwares ShaEP(Vainio et al., 2009). These tools can be used separately or together to identify the effect of the alignment on the final prediction. In reality, the alignment that we aim to recreate should reflect the superimposition that the set of compounds adopt when binding to a given molecular target; however a given set of molecules may bind in different ways when confronted to another binding site. In the present study, no binding site is available to guide the molecular alignment of betulin derivatives. We can however take advantage from the fact that the compounds used in our study are very similar and share a common five-member ring scaffold while they vary with the attached functional groups.

In this study, the alignment of training set was made with database alignment algorithms (Sybyl-X 1.2 database alignment) by using template compound (compound 18) as the basis for the alignment. Database alignment corresponds to the superposition of the common substructure shared by all molecules (Fig. 1). For superposition, compound 18 with the highest pI<sub>50</sub> (5.91) was used as template molecule. All the five rings of the betulin scaffold were selected for superimposition and a rigid body superposition performed. The substituent R<sub>3</sub> is highly conformational flexible, however we did not select the individual conformations that would lead to an optimal superimposition of R<sub>3</sub>. Optimizing this region would have required further work but could have lead to more predictive models.

### **CoMFA/CoMSIA fields calculation**

The aligned training sets of molecules were positioned inside grid boxes with grid spacing value of 2 Å (default distance) in all Cartesian directions and CoMFA fields were calculated using the QSAR module of Sybyl-X(Tripos International, St. Louis). The interaction energies for each molecule were calculated at each grid point using probe atom: an sp<sup>3</sup> hybridized carbon atom with a van der Waals (vdW) radius of 1 Å and a +1 charge (default probe). The steric (vdW interaction) and electrostatic (Coulombic values) fields were calculated at each intersection on the regularly spaced grid. The cutoff value for both steric and electrostatic interaction was set to 30 kcal/mol. CoMSIA similarity index descriptors was derived using the same lattice boxes as those used in CoMFA calculations. Five physicochemical properties steric, electrostatic, hydrophobic, hydrogen

bond donor and acceptor were evaluated using a common probe atom of 1 Å radius. In CoMSIA, the steric indices are related to the third power of the atomic radii, the electrostatic descriptors are derived from atomic partial charges, the hydrophobic fields are derived from atom-based parameters developed by Viswanadhan and co-workers (Viswanadhan et al., 1989), and the hydrogen bond donor and acceptor indices are obtained from a rule-based method derived from experimental values. Similarity indices were calculated using Gaussian-type distance dependence between the probe and the atoms of the molecules of the data set. This functional form requires no arbitrary definition of cutoff limits, and the similarity indices can be calculated at all grid points inside and outside the molecule. The value of the attenuation factor was set to 0.30.

### 3D-QSAR models calculation, internal and external validation

In order to generate statistically significant 3D-QSAR models, Partial Least Square (PLS) regression was used to analyse the training set by correlating the variation in the pI<sub>50</sub> values (the dependent variable) with variations in their CoMFA/CoMSIA interaction fields (the independent variables). The grid was chosen with resolution of 2 Å and extended beyond the molecular dimensions by 4 Å in all directions. Column filtering was set to 4 kcal/mol. CoMFA and CoMSIA models were developed using the conventional stepwise procedure. The leave-one-out cross validation (LOO-CV) was performed to determine optimum number of component leading to the highest cross-validated coefficient  $q^2$  (equation 1) and the lowest standard error of prediction (SEP) that indicates the consistency and predictive ability of models. After that, non-cross-validation was performed to derive the final PLS regressions models with the explained variance  $r^2$ , standard error of estimate (S) and F ratio. S represents the measure of the target property uncertainty still unexplained after the model has been derived, and F the ratio of  $r^2$  to  $1 - r^2$  weighted so that the fewer the explanatory properties and more the values of the target property, the higher the F-ratio.

$$q^2 = 1 - \frac{\sum_y (Y_{\text{predict}} - Y_{\text{experimental}})^2}{\sum_y (Y_{\text{experimental}} - Y_{\text{mean}})^2} \quad (1)$$

Where:

- $Y_{\text{predict}}$  = a predicted pI<sub>50</sub>
- $Y_{\text{experimental}}$  = an experimental pI<sub>50</sub>
- $Y_{\text{mean}}$  = the best estimate of the mean of all values that might be predicted
- The numerator is the sum of the squared deviations between predicted and experimental pI<sub>50</sub> values for the training set compounds.
- The denominator is the sum of the squared deviation between the experimental pI<sub>50</sub> values and the mean pI<sub>50</sub> predicted value of the training set

As the name suggests, leave-one-out cross-validation involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. The coefficients of the independent variables of the original PLS model are calculated, excluding one compound (i.e., activity values and calculated properties) from the original training set at once, and this “new” model is used to predict the activity of the excluded compound. This procedure is repeated through the whole data set, until all compounds have been excluded once, and then,  $q^2$  values and SEP are calculated.



The CoMFA/CoMSIA results were graphically represented by field contour maps, where the coefficients were generated using the field type “Stdev\*Coeff”. Favored and disfavored levels were fixed at 80% and 20%, respectively.

The models generated have to be validated. Even if the model is validated as high quality by internal cross validation, uncertainty will remain regarding its ability to predict chemicals not in the training set. To address this question, external validation data sets are used.

### External validation

In order to assess the actual predictive ability of the best models generated by the CoMFA/CoMSIA approaches, the  $pI_{50}$  values of the external validation set (i.e., test set compounds not included in the training set) were calculated using the same CoMFA/CoMSIA parameters as those used to generate the models. The non cross validated analyses were used to make predictions of the percent inhibitions of the betulin derivatives compounds from the test set and to display the coefficient contour maps. The actual versus predicted percent inhibitions of the test betulin derivatives compounds were fitted by linear regression, and the “predictive”  $r^2$ , S, and F ratio were determined. The quality of the external prediction is documented using the standard deviation of error prediction ( $r^2_{pred}$ ).

$$r^2_{pred} = 1 - \frac{PRESS}{SD} \quad (2)$$

In Equation (2), PRESS is the sum of the squared deviations between predicted and actual  $pI_{50}$  values for the test set compounds and SD is the sum of the squared deviation between the actual  $pI_{50}$  values of the compounds from the test set and the mean  $pI_{50}$  value of the training set compounds.

Establishing an applicability domain is a major step in QSAR analysis since an applicability domain allows to avoid trying to predict irrelevant molecules, i.e. molecules that differ too much from those included in the training set. (Tropsha, 2010) We have verified that several statistical criteria from activity/property of training and test set prediction defined by Tropsha and al are satisfied by our predictive model, including a correlation coefficient > 0.50, coefficient of determination > 0.60. (Tropsha & Golbraikh, 2007)

## 2.4 Pharmacophore models

In this study, five compounds from the training set, whose functional R groups are matching well the molecular fields suggested by the CoMFA or CoMSIA models to be important for activity, were selected to generate pharmacophore models: compounds 4, 18, 21, 24 and 25. In the case of the present study, these compounds are also the most active compounds and therefore could have been chosen without the help of the CoMFA/CoMSIA model, but this is not always the case. GALAHAD was run with default values to generate a set of pharmacophore models, used for screening, and molecular alignments, that we do not further use.

GALAHAD is a proprietary pharmacophore module from Tripos Ltd, which generates pharmacophore models and alignments from sets of compounds (Tripos International, St. Louis). A pharmacophore model consist of a group of features located relatively close one to each other in 3D space, surrounded by a sphere of tolerance, which encode location-

dependent chemical characteristics that account for activity. The sphere represents the 3D area that should be occupied by specific chemical functional groups for optimal activity. GALAHAD identifies a set of molecular conformations with an optimal combination of low strain energy, steric overlap, and pharmacophoric similarity. The search of conformations is performed in two steps. First, the ligands are aligned one onto each other in internal coordinate space. In this stage a genetic algorithm is used to identify a set of ligand conformations that both minimizes energy and maximizes pharmacosteric similarity. Simultaneously, pharmacophore multiplet similarity between ligands is maximized. This stage is fully flexible. The second step is a rigid-body hypermolecular alignment process in Cartesian space. (Richmond et al., 2004) GALAHAD uses a multi-objective (MO) function in which each term is considered independently for three different purposes: to assess reproductive fitness, to pick the candidates that survive to the next generation, and to rank models after Cartesian alignment of their constituent ligand conformers. (Gillet et al., 2002) The three MO functions (multi-objective triage approach) make use of Pareto ranking for each individual model, which is defined as the number of alternative candidates that are better than the model being assessed by all criteria. (Clark & Abrahamian, ). Among the selected models, the ones with the best energy, steric and pharmacophoric concordance values based on Pareto ranking were selected as the best model.

## 2.5 Database searching and compound selecting

Database searching and compounds selection from the resulting 3D pharmacophore models was carried out on a Linux Pentium (2 CPUs). The private FIMM library (119027 compounds) of Institute for Molecular Medicine Finland and the public NCI library (234054 compounds) of National Cancer Institute were chosen for virtual screening; compounds from these libraries are easily accessible to us for experimental testing. All compounds in these databases were first converted from 2D (sdf format) to 3D using the Concord module in Sybyl-X, and one representative 3D conformation, (lowest energy) selected. Then, for all of these representative conformations, Gasteiger-Huckel charges were assigned and the compounds energy minimized using the standard Tripos force field (Powell method and 0.05 kcal/(mol.Å) energy gradient convergence criteria).

A 3D query was defined based on the best 3D pharmacophore model derived by GALAHAD in Sybyl-X. The query was used to perform virtual screening experiment, by using the Unity 3D database search protocol with all options set to default. In the default option, in order to save screening and hit selection time, the oral bioavailability drug-likeness rule (Lipinski's rule of five with one violation) was applied as a pre-filter with following criterion: molecular weight < 500,  $-4 < \log p < 5$ , number of donor/acceptor < 10, the numbers of rotatable bond < 10 and the number of rotatable bonds < 10. This should be also beneficial for the quality of selected compounds at the end of screening.

During the Unity search procedure, the conformations of the compounds in the screened database were generated on the fly by means of the Directed Tweak method. (Hurst, 1994) This procedure attempts to determine if a candidate structure can reasonably flex into a conformation that matches the query. In this way, the data storage problem and the search time are minimized, since only the relevant conformations are dynamically generated. If the query uses spatial constraints, then the query is aligned to the target values for the constraints. If the query contains normal constraints, then the query is aligned to the target

for the features. As a result, two hit lists were generated, one for the FIMM and one for the NCI library, which contain compounds with chemical functionalities and spatial properties similar to those of the 3D pharmacophore query. For each compound, a fit value was returned that represented how well the compound fit into the pharmacophore model. The molecular conformations that have been identified by UNITY as hits are not necessarily the lowest possible energy conformations. In some cases, UNITY returns highly strained conformations that energetically cannot exist. The post-processing ranking, relaxing, and tightening functionality allows to rank the hits after the search. In addition, predicted values of the inhibition ( $I_{50}$ ) percentage and the water-octanol partition coefficients LogP for each compound in the hit lists are computed with CoMSIA predictive model and the Sybyl-X software respectively. LogP is linked to solubility and we may encounter experimental problems to solubilize these compounds. In addition, betulin derivatives are extremely hydrophobic and not too much drug-like, where a logP <5 or often < 3 are recommended. Nonetheless, previous studies on betulin have shown it has other advantages that make this scaffold attractive to medicinal chemists.

### 3. Results and discussion

#### 3.1 3D-QSAR results

##### CoMFA/CoMSIA modeling and outlier removal

The Fig. 1 shows how the training set molecules are aligned within the grid box (grid spacing 2 Å). The summary of results from CoMFA and CoMSIA models using LOO-CV is presented in Table 2. The predictability of the models is one of the most important parameters for appreciation of 3D-QSAR methods. The first CoMFA/CoMSIA model generated with all compounds in the training set has a  $Q^2$  value of 0.27 and 0.30 respectively. The analysis of correlations between the calculated and experimental values of  $pI_{50}$  (Fig. 2) show the presence of two compounds, 6 and 12, poorly predicted. Compounds 6 and 12 are at extreme of experimental property/activity range from all others values: compound 6 is inactive and compound 12 is highly active. In addition, these compounds carry strong electrophilic centers that are good leaving groups (compounds 12) or Michael acceptors (compound 6). It is likely that the bioactivities of these compounds are due to their high reactivity, i.e. their tendencies to covalently attach to nucleophilic centers. In order to avoid mixing up different mechanistic effects, the CoMFA/CoMSIA models were rebuilt omitting these compounds, considering them as outliers. Compound 16, which present similarly to compound 12 a good leaving group, was accordingly deleted from the test set. On the other hand, compounds 4 and 22, which contain potentially reactive aldehydes in their substituent, were kept in the dataset since aldehydes are not very reactive groups and can be found in known drugs.

##### CoMFA model, predictivity

As a result, the CoMFA model without compounds 6 and 12 describing Betulin derivatives inhibition used both steric and electrostatic fields and had a  $Q^2$  value of 0.58, and using two components. This CoMFA model indicated different contributions of both steric and electrostatic fields of 0.33 and 0.66, respectively. The model had cross-validated  $r^2 = 0.58$ , non cross-validated  $r^2 = 0.81$  and Fischer ratio ( $F = 24.11$ ). The predictions of  $pI_{50}$  values for the 14 BDIs in the training set using CoMFA models are shown in Table 1. The correlations

between the calculated and experimental values of  $pI_{50}$  (from training and LOO cross-validation) are shown in Fig. 2.

This model was validated by an external test set of eight compounds not included in the model construction. We found that this model was able to describe the test set variance with predictive  $r^2 = 0.78$ . The predicted activity values of test set are listed in Table 1, and the correlations between the predictions and experimental values are represented in Fig. 2. This analysis revealed that the proposed model is able to predict successfully compounds that were not used in the training process.

#### CoMFA model, contour plots

The contour plots of the CoMFA steric and electrostatic fields are presented in Fig. 3 for the modeled BDI activities. For simplicity, only the most active compound (compound 18) contour map is shown. In this figure, green and yellow contours indicate regions where steric bulk groups favored and disfavored the activity, respectively. A green contour, shown in Fig. 3, indicates that large substituents near C2 of the compound 18 (substituent R1) are important for a high BD inhibitory activity. The red contours at the same position indicate regions where an increase of positive charge decreases the activity. The R1 substituents are therefore preferably large and negatively charged. Green contours are also located near the R2 substituent at C16 (Fig. 3) that therefore prefers large substituents too. Overall, polar groups are favored at R2. Red contours indicate that functional groups containing with an electronegative character, such as carboxylic acid ( $COO^-$ ) and OH, are beneficial at R2. Positively charged groups, such as bases and OH groups, are also beneficial at R2 for gaining BDI activity, as seen by the blue contours. It seems that the R3 substitution does not have any effect on compound activity according the steric and electrostatic field contours, but as mentioned before the conformational alignment of this substituent was not optimized.

#### CoMSIA model, predictivity

In comparison to CoMFA, CoMSIA is less affected by changes in molecular alignment and provides smoother and interpretable contour maps as a result of employing Gaussian type distance dependence with the molecular similarity indices it uses. Furthermore, in addition to the steric and electrostatic fields, CoMSIA defines explicit hydrophobic and HBD and HBA descriptor fields. A more statistically robust model was obtained from the CoMSIA study. The CoMSIA model has a better cross-validated  $r^2$  value of 0.662 using five components, non cross-validated  $r^2$  value of 0.991 and a Fischer ratio ( $F = 179.83$ ). CoMSIA model indicated contributions of steric, electrostatic, hydrophobic, H bonds donor and acceptor field contributions of 0.03, 0.31, 0.07, 0.34 and 0.22 respectively. Thus, in contrast to CoMFA, the steric contribution to the CoMSIA model is almost negligible. The predictions of  $pI_{50}$  values for the 14 BDIs in the training set using CoMSIA model are shown in Table 1. The correlations between the calculated and experimental values of  $pI_{50}$  (from training and LOO-CV are shown in Fig. 4. The CoMSIA model was also used to predict the inhibitory activities of the external test set compounds, and this model was able to describe the test set with predictive  $r^2 = 0.91$ . The external test set predicted values are listed in Table 1, and the correlations between the predicted activity values and experimental values are represented in Fig. 4. As for CoMFA, the model is therefore able to predict successfully compounds that were not used in the training process.



### CoMSIA model, contour plots

The contour plots of the CoMSIA steric, electrostatic, hydrophobic, HB acceptor and HB donor fields are presented in Fig. 5. Generally, the steric and electrostatic field contributions respectively are similar to those one in CoMFA analysis (Fig. 5a). They were interpreted in the same manner as in the above-mentioned CoMFA model and therefore not described here.

CoMSIA models present, in addition to CoMFA models, hydrophobic and hydrogen-bond fields. The hydrophobic contributions are presented in Fig. 5b. An orange contour covering the area near substituent R2 indicates that hydrophobic groups are favoured at R2. Instead, the presence of black contours at R1 substitution, near C3, suggests that hydrophilic groups are useful to increase activity. The hydrogen bond donor and acceptors contours are generally in agreement with the contours based on negative/positive charges, as seen on Fig. 5c and 4d, however giving more precise information. A large magenta contour near R2 shows that HB donor groups are favourable to activity at R2. A significant volume red contour is present near to C16 indicates a detrimental effect of HB acceptor groups at R1, an information that was not given by the CoMFA model. Red contour near the R1 substituent, that hydrogen bond acceptor groups at R1 decrease the activity.

### Performance comparison

The superior performance of CoMSIA relative to CoMFA with this dataset may be attributed to the smoother potentials or to the higher contributions from the HBD and HBA fields to the CoMSIA models (Table 2). Unlike CoMSIA, CoMFA does not have explicit hydrogen-bonding descriptors, which are assumed to be implicitly treated in the CoMFA steric and electrostatic fields, respectively. The CoMSIA steric and electrostatic PLS contours were similarly placed as those of the CoMFA model. The HBD fields made the highest contribution to the CoMSIA models (Table 2), which suggest that among the descriptors considered, the HBD is the most important factor influencing the activity of the betulin derivatives in the training set.

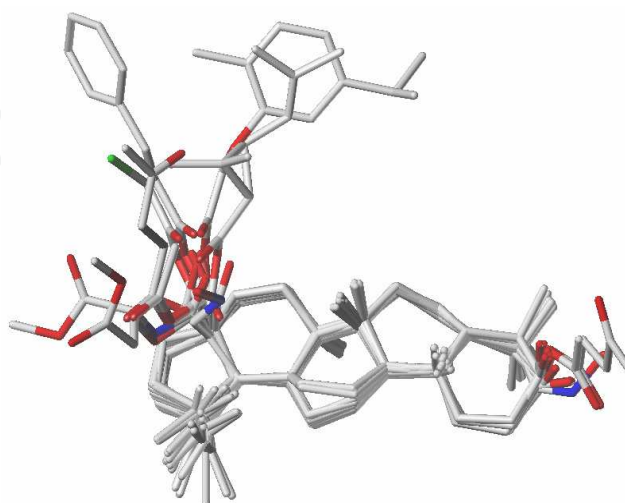


Fig. 1. Database alignment superposition of training set compounds used for 3D-QSAR analysis



	Parameter						Contributions				
	$r^2_{cv}$	NC	$r^2_{ncv}$	SEE	F-value	$r^2_{pred}$	S	E	H	D	A
CoMFA	0.27	1	-	-	-	-	-	-	-	-	-
CoMSIA	0.30	2	-	-	-	-	-	-	-	-	-
CoMFA*	0.58	2	0.81	0.44	24.11	0.78	0.33	0.66	-	-	-
CoMSIA*	0.66	5	0.99	0.11	179.83	0.91	0.03	0.31	0.07	0.34	0.22

Table 2. Summary of Analysis Results of the CoMFA and CoMSIA Models. NC is the number of components from PLS analysis,  $r^2_{cv}$  are the correlation coefficients of the leave-one-out (LOO) cross-validation,  $r^2_{ncv}$  are the correlation coefficients for training set without cross-validation analysis. S = Steric, E = Electrostatic, H = Hydrophobic, D = H bond donor, A = H bond acceptor

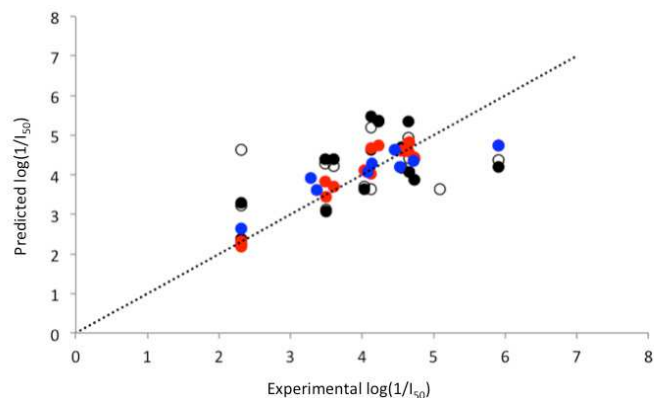


Fig. 2. Scatter plot of the experimental activities versus predicted activities for the CoMFA model. Empty circles: LOO cross-validated predictions on the full training set. Black circles: LOO cross-validated predictions on training set predictions without compound 6 and 12. Red circles: training set without cross validation, Blue circle: test-set predictions.

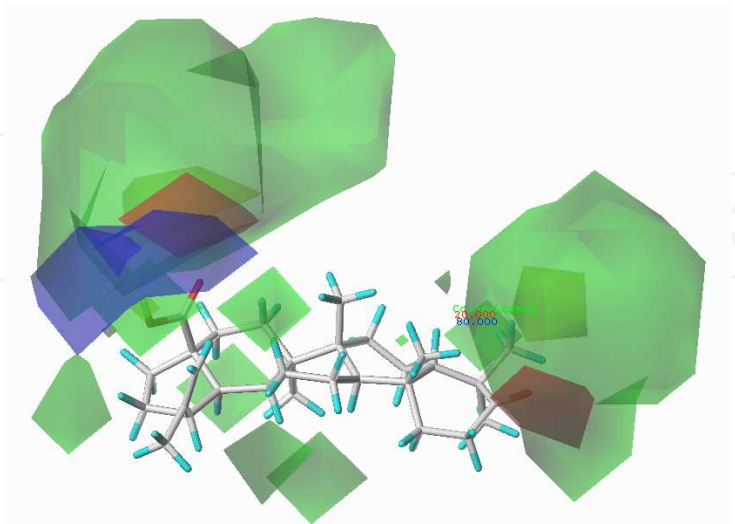


Fig. 3. Contour maps for CoMFA with compound 18 shown as a representative structure. Green contours indicate regions where bulky groups enhance the activity. Blue contours indicate regions where an increase of positive charge enhances the activity, and red contours indicate regions where more negative charges are favourable for activity.

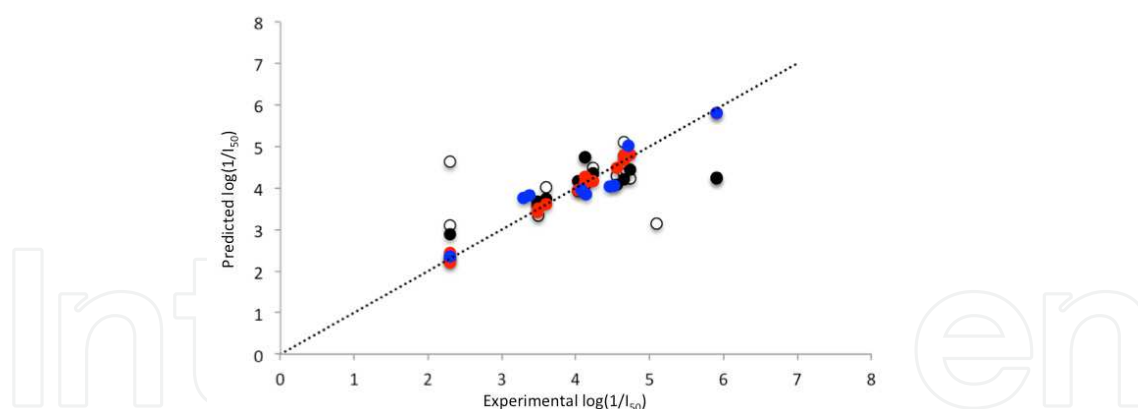
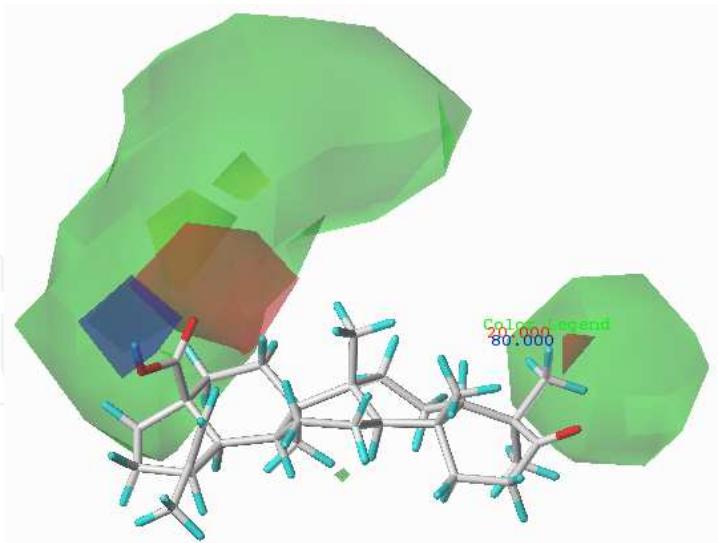


Fig. 4. Scatter plot of the experimental activities versus predicted activities for CoMSIA model. Empty circles: LOO cross-validated predictions on full training set, Black circles: LOO cross-validated predictions on training set predictions without compound 6 and 12, Red circles: training set without cross validation, Blue circles: test-set predictions.

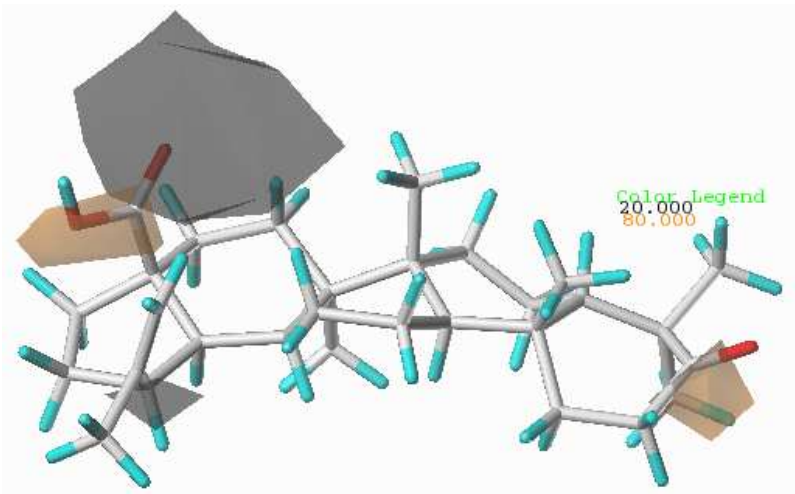
### 3.2 Pharmacophoric representations

GALAHAD pharmacophore models were derived, by using the 5 most active ligands in the training set (these 5 compounds are shown in bold in Table 1). Ten pharmacophore models were retained after the GALAHAD runs. All these models present eight to nine pharmacophoric features. Seven hydrophobic moieties of the pharmacophore reflect the presence for a large hydrophobic structure as the skeleton of the BDIs. It would be possible for us to reduce the number of these pharmacophoric points if we wished to retrieve chemical compounds more distant from the betulin scaffold. The remaining 2 to 3 pharmacophoric points corresponds to the three R groups. In Fig. 7, the pharmacophore for model 3 is represented. It includes 8 pharmacophore features: 7 hydrophobes (HY\_2, HY\_3, HY\_4, HY\_5, HY\_6, HY\_7 and HY\_8) and 1 HD donors (DA\_1). The HB donor moieties reflect the importance of OH groups at these positions of the betulin for BD inhibitory activity. In Fig. 7, cyan and magenta spheres represent indicate hydrophobes and HB donors, respectively.

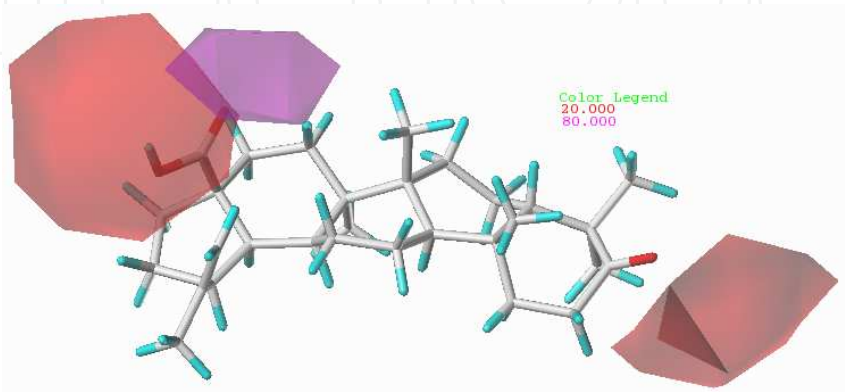
Each of the obtained models represents a different trade-off among the conflicting demands of maximizing steric consensus, maximizing pharmacophore consensus, and minimizing energy. They had Pareto rank 0; this means no one model is superior to any other. During GALAHAD runs, it is recognized that high-energy values are due to steric clashes. (Dorfman et al., 2008) The algorithm retains these models to keep good characteristics to be passed on to less strained offspring during genetic algorithm process. All the GALAHAD models are derived from at least 4 ligands of the training set and were compared according to Pareto ranking. Table 3 shows energy, steric and pharmacophoric concordance values for models with all the 5 ligands. Minimum and maximum values for each characteristic between all the obtained twenty models are also reported in this table. The model ten had energy very high than the other nine models and is not included in the statistic. Small value of energy and high values of steric and pharmacophoric concordance are desired for the best model. Now, the higher energy value between all the models is 6149; the models containing all the five ligands had values between 41.79 (the minimum) and 6149.39 (the maximum), in this sense energy value varies widely distributed among the considered models. Steric had a small



(A)



(B)



(C)

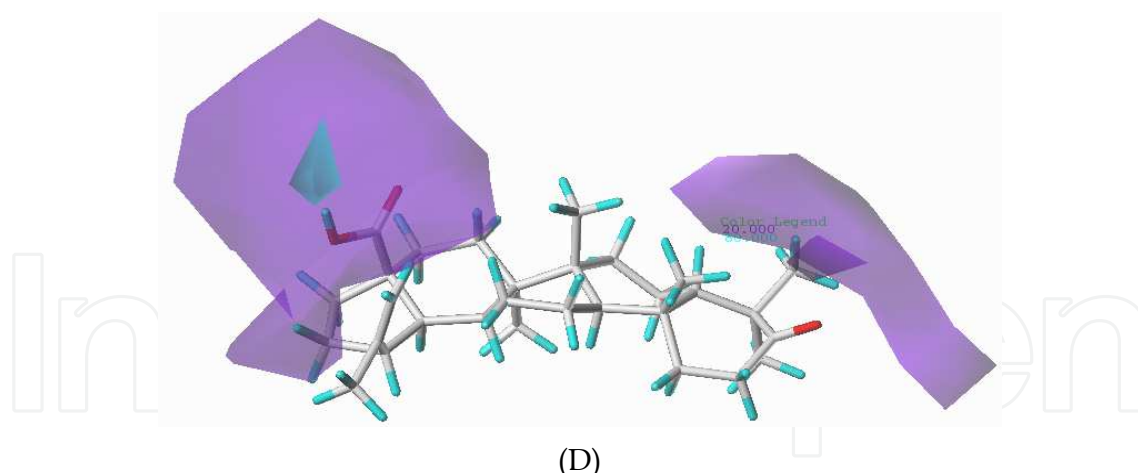
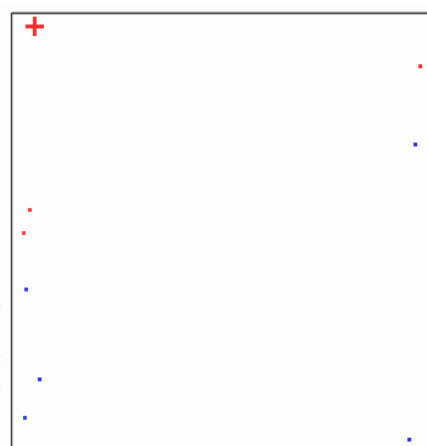


Fig. 5. Contour maps for the CoMSIA model, shown with compound 18 as a representative structure. (A) Steric field: green contours indicates region where bulky groups enhance the activity. Electrostatic field: Blue indicates regions where positive charge is favoured and enhances activity and red the regions where it decrease it. (B) Hydrophobic field: orange contours indicate regions where hydrophobic/hydrophilic groups enhance/decrease the activity, and black contours indicate regions where hydrophobic/hydrophilic groups decrease/enhance the activity. (C) HB acceptor field: Magenta represents areas where HB acceptors favor the activity and red, area where it disfavour it. (D) HB donor field: Cyan represents areas where HB donors favor the activity and purple the area where it decrease it.

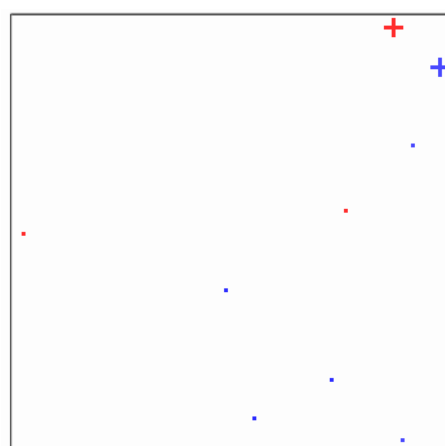
variation between the minimum (20978.80) and the maximum (21538.59) considering all the models. Finally, pharmacophoric concordance had a small variation between the minimum (325) and the maximum (392.50). With the intention to select the best model, we constructed a 3D plot to visualize the Pareto surface (Fig. 6).

Considering only the energy and steric criteria, the best of all models lies in the upper left hand corner of the graph in Fig. 6a, where the energy (x axis) is low and the steric (y axis) score is high. In terms of pharmacophoric concordance and steric criteria, the best of all models lies in the upper right hand corner of the graph in Fig. 6b, where the HBond (x axis) score is high and steric (y axis) is high. Finally, in terms of pharmacophoric concordance and energy scores, the best of all models now lies at the lower right corner, where HBOND (x axis) are high and energy (y axis) are low both (Fig. 6c). According Fig. 6, there is only one model (Model 3), which filled all the three requirements described above and was selected for the next of study. This model is represented in Fig. 7. Model 3 has low energy, the higher steric but with high pharmacophoric concordance values. All conformers aligned represent low-energy conformations of the molecules, and it can be seen that the final alignment shows a satisfactory superimposition of the pharmacophoric points.

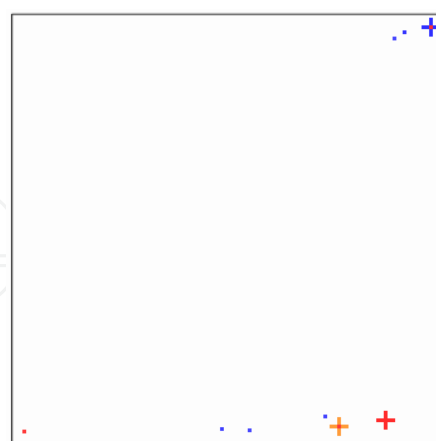
We evaluated how well the model identifies active compounds in virtual screening experiment of a larger database. For this, the model was used to screen a large database constituted by FIMM database, NCI database and the set of 24 compounds from the previous 3D-QSAR studies. This indicates how these models can be used as a theoretical screening tool and how they were able to discriminate between active and inactive molecules, and consequently, to predict whether a new molecule inhibits BD.



(A)



(B)



(C)

Fig. 6. Plot of the strain energy, steric overlap and pharmacophoric concordance values for GALAHAD models with all the 5 ligands with contribution to the consensus feature.

Plot of steric overlap vs. energy. (B) Plot of steric overlap vs. pharmacophoric concordance (HBOND). (C) Plot of HBOND vs. steric overlap.



MODEL	Features	ENERGY	STERICS	HBOND
MODEL 1	9	6074.6299	21377.9004	388.1
MODEL 2	8	41.79	21258.1992	325
<b>MODEL 3</b>	<b>8</b>	<b>210.97</b>	<b>21538.5996</b>	<b>385.1</b>
MODEL 4	9	75.67	21181.9004	357.9
MODEL 5	8	126.38	21289.6992	377.3
MODEL 6	9	5978.9502	20978.8008	386.4
MODEL 7	9	279.49	21061.1992	375
MODEL 8	9	56.97	21009	362.5
MODEL 9	8	6149.3901	21483.5	392.5
MODEL 10	9	132872792	21596.6992	366
Min <sup>a</sup>		41.79	20978.8008	325
Max <sup>a</sup>		6149.3901	21538.5996	392.5

Table 3. Summary of Analysis Results of the CoMFA and CoMSIA Models. The selected model (MODEL 3) is indicated in boldface. <sup>a</sup>Minimum and maximum values between all the obtained 21 models.

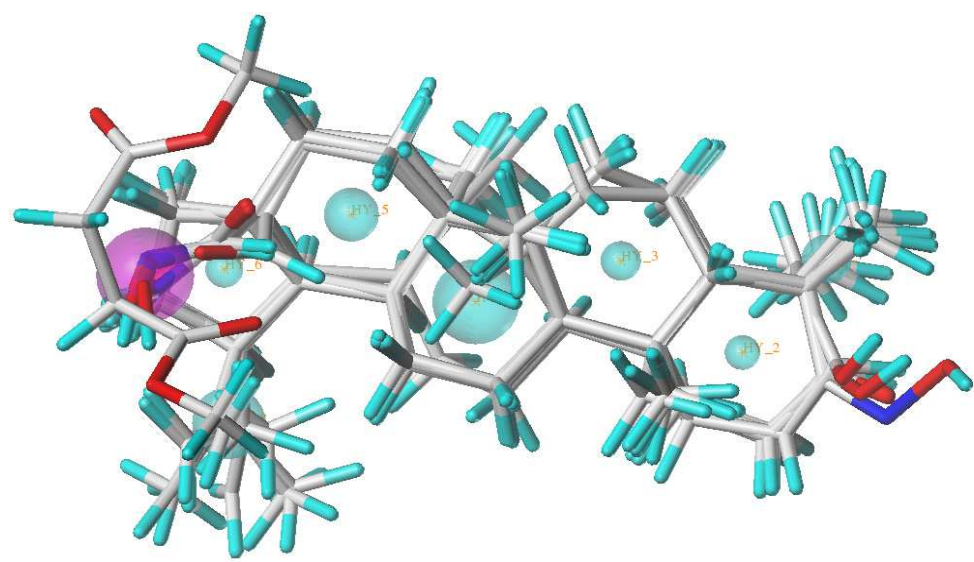


Fig. 7. Selected pharmacophore model 3 and molecular alignment of the compounds used to elaborate the models. Cyan, and magenta spheres are represented for hydrophobes, and HB donors, respectively.

3.3 Virtual screening results

Before screening a virtual database it is important to verify the (dis)similarity of compounds present in the database. This enables us to visualize the chemical space covered by the compounds inside the database and the level of diversity of these compounds. The diversity of a library of compounds denotes the degree of heterogeneity, structural range within the set of compounds. Such exploration of yet unknown chemical space might help to solve the problem of the high attrition rates in drug development by giving more diverse compounds to choose, a broader range of structures at the hit prioritization level, which should increase the chances of success at later stages and also may allow us to avoid target promiscuity that

is apparent in many drugs and allow the design of safer drugs. The chemical space of chemical (diversity/similarity) of chemical structure libraries can be characterized by the distribution of Tanimoto coefficients (equation 3). The Tanimoto coefficient is the most commonly used coefficient in chemical similarity/diversity work, following a study of the performance of a range of similarity coefficients by Willet and Winterman.(Willett & Winterman, 1986) It firstly requires that the molecules are represented by appropriate structural descriptors. Many different structural descriptors have been developed for similarity searching in chemical databases including 2D fragment based descriptors, 3D descriptors, and descriptors that are based on the physical properties of molecules.(Downs et al., 1994) The Tanimoto coefficient is usually calculated from the 2D structure fingerprint, or the 3D shape/feature similarity. A fingerprint is an ordered list of bits. Each bit represents a Boolean determination of, or test for, the presence of, for example, an element count, a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure.

The Tanimoto score equation:

$$T(A,B) = \frac{AB}{\|A\|^2 + \|B\|^2 - AB} \quad (3)$$

Where:

T(A,B) is the similarity score, a fraction between 0 and 1.

A is the count of bits set in fingerprint A

B is the count of bits set in fingerprint B

AB is the count of bits set in common in fingerprints A and B

The Unity module generates a binary substructure fingerprint for chemical structures of our screening libraries. These fingerprints are used by Unity for similarity neighboring and similarity searching. As shown in Fig. 8 and Fig. 9, when we calculated the matrix of pairwise similarity based on 2D structure for the FIMM and NCI libraries. This 2D fingerprints is based on a combination of a hashing function (which represents connected molecular fragments in a efficient but unintelligible manner) and an explicit count of specific fragments, such as rings. The histograms of the Tanimoto indices show diverse distribution of the compounds in the databases and relative distribution of type of compounds inside. A 2D Tanimoto mean of 0.88 and 0.87 are indicative of a suitable range of diversity within each of the two databases.

The Model shown in Fig. 7 was used to generate the query for 3D search virtual screening via the 3D search method implemented in UNITY module encoded in Tripos. Compounds had to map at least 6 features in the pharmacophore model.

<<

```
DONOR_ATOM[NAME=DA_1;TARGET=(-6.205,0.423,2.026)]
HYDROPHOBIC[NAME=HY_2;TARGET=(3.520,-0.863,-0.432)]
HYDROPHOBIC[NAME=HY_3;TARGET=(1.540,0.665,0.009)]
HYDROPHOBIC[NAME=HY_4;TARGET=(-0.820,-0.017,-0.450)]
HYDROPHOBIC[NAME=HY_5;TARGET=(-2.846,1.434,-0.588)]
HYDROPHOBIC[NAME=HY_6;TARGET=(-4.916,0.526,-1.145)]
HYDROPHOBIC[NAME=HY_7;TARGET=(4.890,0.689,-0.736)]
```

```
HYDROPHOBIC[NAME=HY_8;TARGET=(-4.841,-2.039,-0.846)]
spatial_point[name=SPAT_DA_1;feature=DA_1;point=(-6.205,0.423,2.026);tolerance=0.800]
spatial_point[name=SPAT_HY_2;feature=HY_2;point=(3.520,-0.863,-0.432);tolerance=0.290]
spatial_point[name=SPAT_HY_3;feature=HY_3;point=(1.540,0.665,0.009);tolerance=0.330]
spatial_point[name=SPAT_HY_4;feature=HY_4;point=(-0.820,-0.017,-0.450);tolerance=0.710]
spatial_point[name=SPAT_HY_5;feature=HY_5;point=(-2.846,1.434,-0.588);tolerance=0.450]
spatial_point[name=SPAT_HY_6;feature=HY_6;point=(-4.916,0.526,-1.145);tolerance=0.310]
spatial_point[name=SPAT_HY_7;feature=HY_7;point=(4.890,0.689,-0.736);tolerance=0.410]
spatial_point[name=SPAT_HY_8;feature=HY_8;point=(-4.841,-2.039,-0.846);tolerance=0.600]
partial_match[min=6;max=6;features=DA_1,HY_2,HY_3,HY_4,HY_5,HY_6,HY_7,HY_8]
>>
```

As a result (see Table 4), pharmacophore based virtual screening yielded 13 hits (out of 24) from Table 1 compounds, 16 hits from FIMM library (out of 120k compounds) and 76 hits from NCI library (out of 240k compounds) that meet the specified requirements. The hits list selected from the Table 1 compounds confirm the selectivity of our query. Among the set of 24 betulin derivatives used as controls, about half (13) were retrieved by the procedure, mostly the highly active ones. Finally, as a result of this study the best 20 hits from FIMM and NCI were selected for further pharmacological assay (Table 5 and 6). The inhibition ( $I_{50}$ ) percentage of the selected compounds are in the range of the values of active compounds present in the Table 1 dataset and their values of LogP are near the value (8.07) of the most active compound (compound 18) of Table 1. These predicted values of  $I_{50}$  and LogP should be used to prioritize compounds to send in experimental test.

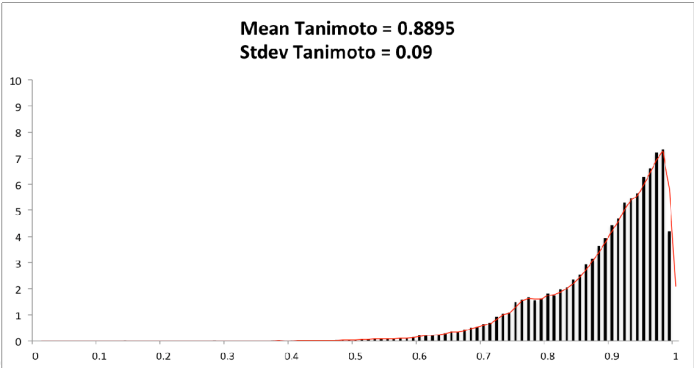


Fig. 8. FIMM library distribution

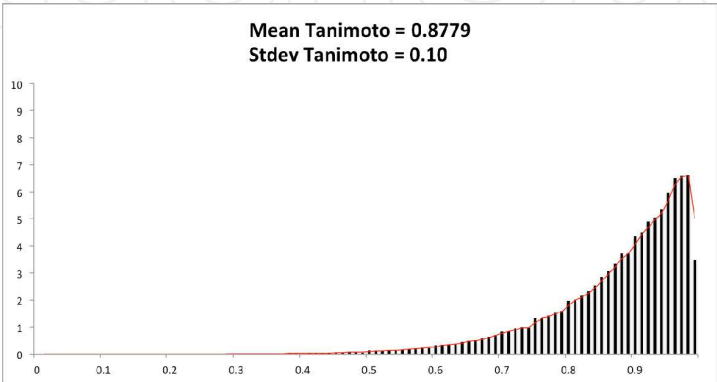
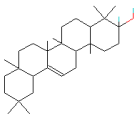
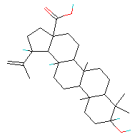
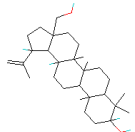
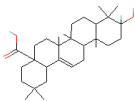
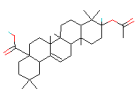
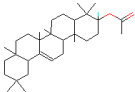
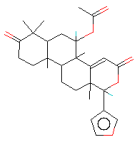


Fig. 9. NCI library distribution

	Table 1	FIMM	NCI
Number hits	13	16	76

Table 4. Summary of hits found by pharmacophore based 3D search virtual screening on FIMM, NCI and table 1 compound.

FIMM ID	Image	QFIT	RANK	LOGBIO	CLOGP
538990053		34.33	1	3.63	10.66
538990110		22.94	2	3.66	7.95
538990111		27.42	3	3.67	8.52
538990154		34.33	4	3.69	8.62
538990271		34.33	5	3.52	9.53
538990190		34.33	6	3.39	11.60
538990189		46.28	7	4.37	3.88

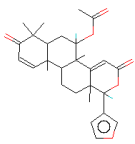
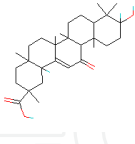
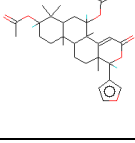
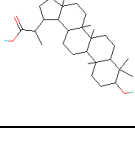
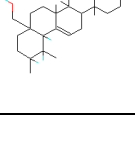
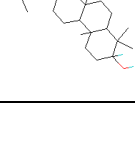
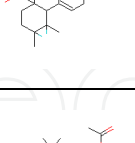
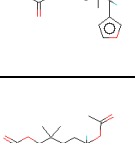

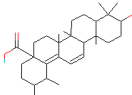
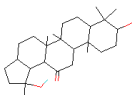
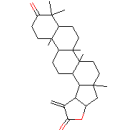
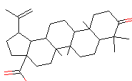
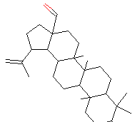
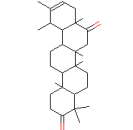
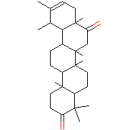
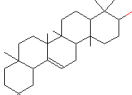
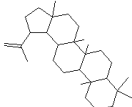
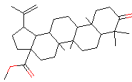
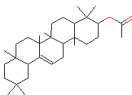
538990181		35.60	8	4.25	3.74
538990200		39.75	9	3.67	6.47
538990295		45.39	10	4.00	5.23
AE-641/00404032		56.28	11	3.76	6.97
538990112		47.15	12	3.66	8.67
538988558		56.28	13	3.55	10.51
538990155		47.15	14	3.73	8.62
538990368		42.53	15	3.68	4.67
538990294		45.39	16	4.07	5.23

Table 5. List of compounds selected with pharmacophore based 3D search virtual screening on FIMM compounds.



NCI ID	Image	QFIT	RANK	LOGBIO	CLOGP
661747		80.42	1	3.55	8.40
144946		64.10	2	3.70	6.14
680072		30.81	3	3.68	7.07
152534		30.60	4	3.81	8.075
250423		60.03	5	3.70	8.44
277277		30.62	6	4.06	7.77
119118		30.62	7	4.06	7.77
527971		53.78	8	3.70	10.66
90487		60.03	9	3.68	10.51
152535		30.60	10	3.83	8.45
403166		53.78	11	3.59	11.60

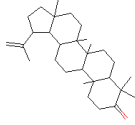
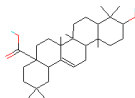
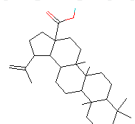
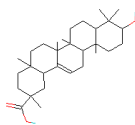
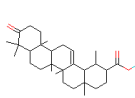
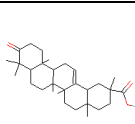
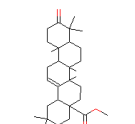
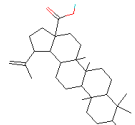
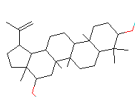
281807		20.53	12	3.82	10.11
114945		53.78	13	3.70	8.62
113090		60.03	14	3.70	8.47
655415		53.78	15	3.67	8.62
94656		50.21	16	3.93	8.22
655414		50.21	17	3.81	8.22
125854		51.71	18	3.91	6.86
677578		60.03	19	3.70	8.47
133914		29.32	20	3.63	8.42

Table 6. List of compounds selected with pharmacophore based 3D search virtual screening on NCI compounds.

4. Conclusions

In the current study, 3D-QSAR and pharmacophore models were derived for betulin derivatives as inhibitors of Leishmaniasis, which should be useful for assisting the design of active compounds. Such models correlate well structural features with inhibitory activities and bring valuable information about the relevant characteristics of inhibitors. CoMFA and CoMSIA approaches were developed to derive structure–activity relationships. CoMFA and

CoMSIA modeling were efficient tools to suggest outliers that we could link to a specific molecular mechanism, in that case covalent crosslinking. The models are reliable and were obtained by using steric and electrostatic CoMFA fields, and by using steric, electrostatic, hydrophobic, HB acceptor and donor CoMSIA fields. In this study, CoMSIA outperforms CoMFA, but this is not always the case. Moreover, contour plots may help identify relevant regions where any change can affect binding preference. According to the obtained statistics, prediction of betulin derivatives activities with sufficient accuracy should be possible by using these models. In a second phase, pharmacophore models were derived with GALAHAD. Models derived from 5 active compounds that all match best the CoMSIA predictions were obtained. These models include hydrophobes, and HB donors. The obtained pharmacophore models were used as queries for 3D flexible search engine to search for the FIMM, NCI and QSAR dataset collection. Without the verification of the predictive characteristic of the compounds in our dataset with 3D-QSAR model, it would have been much more speculative to do a pharmacophore-based screening. The process of screening takes less than two hours (standard 2 CPUs workstation). In comparison, a molecular docking study involving the same two libraries in the same conditions, counting 2-5 seconds (2 CPU) for each compound, would take 20 to 50 hours. The search was really efficient, allowing us to retrieve among the hit lists 9 out of the 14 molecules that had been used to build the model and had been put in the library as a control, as well as 4 out of 8 molecules in the test set also used as a control. As a result of this study, 20 first molecules were selected from FIMM and NCI hit list for further biological binding assay.

While this study is conducted for a small number of compounds, for which biological activity was easily obtainable and testing conducted in a single laboratory, it could easily be generalized to larger sets and databases. The results described in this paper indicate that this method is very efficient in the study of hit identification and lead optimization.

## 5. Acknowledgment

Leo GHEMTIO thanks The Drug Discovery and Chemical Biology Consortium and CDR of University of Helsinki for financial support through a postdoctoral fellowship. Yuezhou Zhang would like to thank the Chinese Scholarship Council for financial support and the Informational and Structural Biology doctoral programme (ISB) for organizing graduate studies. The Finnish IT Center for Science (CSC) is thanked for computational resources.

## 6. References

- Alakurtti, S., Bergström, P., Sacerdoti-Sierra, N., Jaffe, C. L., & Yli-Kauhaluoma, J. (2010) Anti-leishmanial activity of betulin derivatives. *J Antibiot (Tokyo)* 63, 123-6.
- Alakurtti, S., Mäkelä, T., Koskimies, S., & Yli-Kauhaluoma, J. (2006) Pharmacological properties of the ubiquitous natural product betulin. *Eur J Pharm Sci* 29, 1-13.
- Bostrom, J., Bohm, M., Gundertofte, K., & Klebe, G. (2003) A 3D QSAR Study on a Set of Dopamine D<sub>4</sub> Receptor Antagonists. *J. Chem. Inf. Comput. Sci.* 43, 1020--1027.
- Buolamwini, J. K. & Assefa, H. (2003) Overview of Novel Anticancer Drug Targets. 85, .
- Clark, R. (2009) Prospective ligand- and target-based 3D QSAR: state of the art 2008.. *Current topics in medicinal chemistry* 9, 791--810.

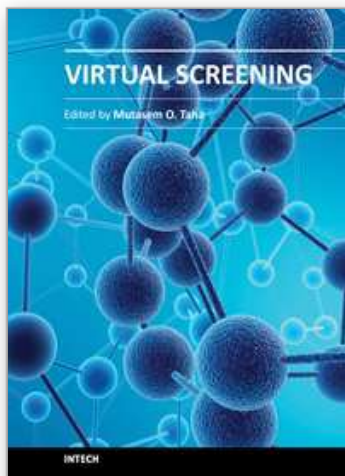
- Clark, R. D. & Abrahamian, E. () Using a staged multi-objective optimization approach to find selective pharmacophore models. *Journal of Computer-Aided Molecular Design* 23, 765--771.
- Cramer, R., Patterson, D., & Bunce, J. (1988) Comparative molecular field analysis (CoMFA). Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* 110, 5959--5967.
- Dorfman, R., Smith, K., Masek, B., & Clark, R. (2008) A knowledge-based approach to generating diverse but energetically representative ensembles of ligand conformers. *Journal of Computer-Aided Molecular Design* 22, 681-691.
- Downs, G. M., Willett, P., & Fisanick, W. (1994) Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *Journal of Chemical Information and Computer Sciences* 34, 1094-1102.
- Ekins, S., Mestres, J., & Testa, B. (2007) In silico pharmacology for drug discovery: applications to targets and beyond.. *British journal of pharmacology* 152, 21--37.
- Ghemtio, L., Devignes, M.-D., Smail-Tabbone, M., Souchet, M., Leroux, V., & Maigret, B. (2010) Comparison of Three Preprocessing Filters Efficiency in Virtual Screening: Identification of New Putative LXR $\beta$  Regulators As a Test Case. *Journal of Chemical Information and Modeling* 50, 701-715.
- Gillet, V. J., Khatib, W., Willett, P., Fleming, P. J., & Green, D. V. S. (2002) Combinatorial Library Design Using a Multiobjective Genetic Algorithm. *Journal of Chemical Information and Computer Sciences* 42, 375-385.
- Hurst, T. (1994) Flexible 3D searching: The directed tweak technique. *Journal of Chemical Information and Computer Sciences* 34, 190--196.
- Jain, A. N. (2004) Ligand-Based Structural Hypotheses for Virtual Screening. *Journal of Medicinal Chemistry* 47, 947--961.
- Kirchmair, J., Distinto, S., Schuster, D., Spitzer, G., Langer, T., & Wolber, G. (2008) Enhancing Drug Discovery Through In Silico Screening: Strategies to Increase True Positives Retrieval Rates. *Current Medicinal Chemistry* 15, 2040--2053.
- Kirchmair, J., Markt, P., Distinto, S., Wolber, G., & Langer, T. (2008) Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection---What can we learn from earlier mistakes?. *Journal of Computer-Aided Molecular Design* 22, 213--228.
- Klebe, G., Abraham, U., & Mietzner, T. (1994) Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *Journal of Medicinal Chemistry* 37, 4130--4146.
- Langer, T. & Hoffmann, R. (2001) Virtual screening: an effective tool for lead structure discovery?. *Current pharmaceutical design*. 7, 509--527.
- Lengauer, T. (2004) Novel technologies for virtual screening. *Drug Discovery Today* 9, 27--34.
- Ling & Xuefeng, B. (2008) High Throughput Screening Informatics. *Combinatorial Chemistry & High Throughput Screening* 11, 249--257.
- Liu, H.-Y., Liu, S.-S., Qin, L.-T., & Mo, L.-Y. () CoMFA and CoMSIA analysis of 2,4-thiazolidinediones derivatives as aldose reductase inhibitors. *Journal of Molecular Modeling*, .

- Nagarajan, S., Ahmed, A., Choo, H., Cho, Y., Oh, K.-S., Lee, B., Shin, K., & Pae, A. (2010) 3D QSAR pharmacophore model based on diverse IKK $\beta$  inhibitors. *Journal of Molecular Modeling*, .
- Perkins, R., Fang, H., Tong, W., & Welsh, W. J. (2003) Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology. *Environ Toxicol Chem* 22, 1666-79.
- Pink, R., Hudson, A., Mouriès, M.-A., & Bendig, M. (2005) Opportunities and challenges in antiparasitic drug discovery. *Nat Rev Drug Discov* 4, 727-40.
- Richmond, N., Abrams, C., Wolohan, P., Abrahamian, E., Willett, P., & Clark, R. (2006) GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. *Journal of Computer-Aided Molecular Design* V20, 567--587.
- Richmond, N., Willett, P., & Clark, R. (2004) Alignment of three-dimensional molecules using an image recognition algorithm. *Journal of Molecular Graphics and Modelling* 23, 199--209.
- Rognan, D. (2010) Structure-Based Approaches to Target Fishing and Ligand Profiling. *Molecular Informatics* 29, 176--187.
- Rönkkö, T., Tervo, A., Parkkinen, J., & Poso, A. (2006) BRUTUS: Optimization of a grid-based similarity function for rigid-body molecular superposition. II. Description and characterization. *Journal of Computer-Aided Molecular Design* 20, 227-236.
- Sippl, W. (2002) Development of biologically active compounds by combining 3D QSAR and structure-based design methods.. *J Comput Aided Mol Des* 16, 825--830.
- Spitzer, G., Heiss, M., Mangold, M., Markt, P., Kirchmair, J., Wolber, G., & Liedl, K. (2010) One Concept, Three Implementations of 3D Pharmacophore-Based Virtual Screening: Distinct Coverage of Chemical Search Space.. *Journal of chemical information and modeling* 0, .
- Stahura, F. & Bajorath, J. (2004) Virtual Screening Methods that Complement HTS. *Combinatorial Chemistry & High Throughput Screening* 7, 259--269.
- Tervo, A. J., Rönkkö, T., Nyrönen, T. H., & Poso, A. (2005) BRUTUS: Optimization of a Grid-Based Similarity Function for Rigid-Body Molecular Superposition. 1. Alignment and Virtual Screening Applications. *Journal of Medicinal Chemistry* 48, 4076--4086.
- SYBYL-X 1.2, Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.
- Tropsha, A. (2010) Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* 29, 476--488.
- Tropsha, A. & Golbraikh, A. (2007) Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Current pharmaceutical design* 13, 3494--3504.
- Tuccinardi, T. (2009) Docking-based virtual screening: recent developments.. *Combinatorial chemistry & high throughput screening* 12, 303--314.
- Vainio, M. J., Puranen, J. S., & Johnson, M. S. (2009) ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential. *Journal of Chemical Information and Modeling* 49, 492--502.
- Villoutreix, B., Eudes, R., & Miteva, M. (2009) Structure-based virtual ligand screening: recent success stories.. *Combinatorial chemistry & high throughput screening* 12, 1000--1016.
- Viswanadhan, V. N., Ghose, A. K., Revankar, G. R., & Robins, R. K. (1989) Atomic physicochemical parameters for three dimensional structure directed quantitative



structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *Journal of Chemical Information and Computer Sciences* 29, 163-172.

Willett, P. & Winterman, V. (1986) A Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity Measures of Inter-Molecular Structural Similarity. *Quantitative Structure-Activity Relationships* 5, 18--25.



## **Virtual Screening**

Edited by Prof. Mutasem Taha

ISBN 978-953-51-0308-0

Hard cover, 100 pages

**Publisher** InTech

**Published online** 14, March, 2012

**Published in print edition** March, 2012

Pharmacophore modeling, QSAR analysis, CoMFA, CoMSIA, docking and molecular dynamics simulations, are currently implemented to varying degrees in virtual screening towards discovery of new bioactive hits. Implementation of such techniques requires multidisciplinary knowledge and experience. This volume discusses established methodologies as well as new trends in virtual screening with aim of facilitating their use in drug discovery.

### **How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Leo Ghemtio, Yuezhou Zhang and Henri Xhaard (2012). CoMFA/CoMSIA and Pharmacophore Modelling as a Powerful Tools for Efficient Virtual Screening: Application to Anti-Leishmanial Betulin Derivatives, Virtual Screening, Prof. Mutasem Taha (Ed.), ISBN: 978-953-51-0308-0, InTech, Available from: <http://www.intechopen.com/books/virtual-screening/comfa-comsia-and-pharmacophore-modelling-as-a-powerful-tools-for-efficient-virtual-screening-applica>

**INTECH**  
open science | open minds

### **InTech Europe**

University Campus STeP Ri  
Slavka Krautzeka 83/A  
51000 Rijeka, Croatia  
Phone: +385 (51) 770 447  
Fax: +385 (51) 686 166  
[www.intechopen.com](http://www.intechopen.com)

### **InTech China**

Unit 405, Office Block, Hotel Equatorial Shanghai  
No.65, Yan An Road (West), Shanghai, 200040, China  
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元  
Phone: +86-21-62489820  
Fax: +86-21-62489821

© 2012 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the [Creative Commons Attribution 3.0 License](https://creativecommons.org/licenses/by/3.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

IntechOpen

IntechOpen