JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

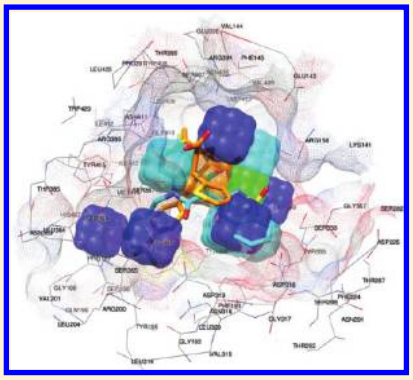
3-D QSAutogrid/R: An Alternative Procedure To Build 3-D QSAR Models. Methodologies and Applications

Flavio Ballante[†] and Rino Ragno*[,†]

[†]Rome Center for Molecular Design, Dipartimento di Chimica e Tecnologie del Farmaco, Sapienza Università di Roma, P. le A. Moro 5, 00185, Rome, Italy

Supporting Information

ABSTRACT: Since it first appeared in 1988 3-D QSAR has proved its potential in the field of drug design and activity prediction. Although thousands of citations now exist in 3-D QSAR, its development was rather slow with the majority of new 3-D QSAR applications just extensions of CoMFA. An alternative way to build 3-D QSAR models, based on an evolution of software, has been named 3-D QSAutogrid/R and has been developed to use only software freely available to academics. 3-D QSAutogrid/R covers all the main features of CoMFA and GRID/GOLPE with implementation by multiprobe/multiregion variable selection (MPGRS) that improves the simplification of interpretation of the 3-D QSAR map. The methodology is based on the integration of the molecular interaction fields as calculated by AutoGrid and the R statistical environment that can be easily coupled with many free graphical molecular interfaces such as UCSF-Chimera, AutoDock Tools, JMol, and others. The description of each R package is reported in detail, and, to assess its validity, 3-D QSAutogrid/R has been applied to three molecular data sets of which either CoMFA or GRID/GOLPE models were reported in order to compare the results. 3-D QSAutogrid/R has been used as the core engine to prepare more that 240 3-D QSAR models forming the very first 3-D QSAR server (www.3d-qsar.com) with its code freely available through R-Cran distribution.



INTRODUCTION

The main requirements of molecular analyses today are as follows: speed, automation, optimization, and economy. Three-Dimensional Quantitative Structure−Activity Relationships (3-D QSARs) approaches are widely used and represent a viable medicinal chemistry tool whose application domain range from rationalizing a structure−activity relationship quantitatively and retrospectively to prioritizing the synthesis of molecules for synthesis and testing; its development considering the actual technology and insight sought becomes important. Till recently,[1] the well-known CoMFA[2] technique and the GRID[3]/GOLPE[4,5] approaches were the 3-D QSAR tools most widely used in the last two decades; although successful, both these methods utilize proprietary software and require significant user interaction. The classical flowchart of a 3-D QSAR can be summarized as reported in Figure 1, and it lists the following:

1) Selection and alignment of Training and Test Sets
2) Calculation of Molecular Interaction Fields (MIFs)
3) Importing of Bioactivities and linking to the MIFs
4) Statistical evaluation
5) Interpretation of results by means of 2-D and 3-D plots

Excellent reviews[6,7] of 3-D QSAR methods have been recently published; for any further details, the reader is referred to them.

Herein is described an alternative approach based on the use of open-source software to perform 3-D QSAR studies fully optimized to minimize costs, calculation time, user/computer

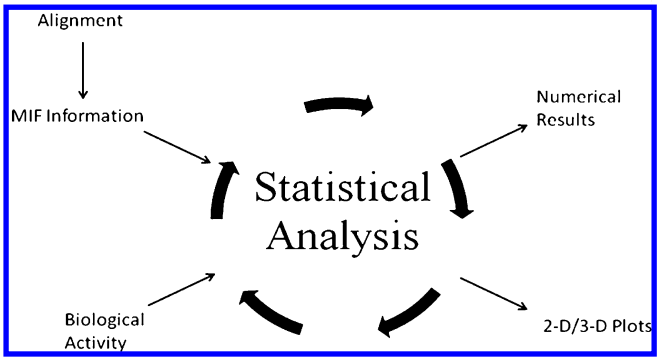


Figure 1. Overview of a classical 3-D QSAR.

interactions, and random and systematic errors. As conceived, the procedure needs only the prealigned training and test sets molecules. The protocol allows the iterative generation of hundreds/thousands 3-D QSAR models and selection of the best on the basis of conventional squared correlation (r^2), predictive cross-validation squared correlation (q^2), and standard deviation error of prediction (SDEP or root mean squared error of prediction, RMSEP) statistical coefficients.

After this project was started, another open-source method, namely OPEN3DQSAR, was reported by Tosco and Balle.[1]

Received: March 7, 2012

   dx.doi.org/10.1021/ci300123x | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

OVERVIEW OF THE NEW PROCEDURE

The new procedure is characterized by a set of new R[8]-based packages that make it possible to carry out, automatically and in parallel, 3-D QSAR studies (like CoMFA and GRID/GOLPE). Different from the previously mentioned methods, the united atom force field (FF) implemented in the AutoGrid program (AutoDock Suite[9]) was used to generate the Molecular Interaction Fields (MIFs). Upon statistic treatment of the merged biological activity and MIFs by means of the R environment and molecular graphic softwares (UCSF Chimera,[10] Python Molecular Viewer[11] (PMV), Autodock Tools[12] (ADT)) useful 2-D (actual vs recalculated (fitting), actual vs predicted (cross validation), principal component analysis (PCA scores, loadings and scores/loadings), partial least-squares (PLS *t-u* and weights), and 3-D plots (actual fields, PLS-Coefficients, Activity contribution, CoMFA-Like (PLS-Coeff*StDev), and various related PLS parameters) are generated to graphically inspect, analyze, and interpret the 3-D QSAR models. Each R based package was conceived to perform specific steps ensuring high specificity, versatility, and, compared to other methods, deep optimization of the models.

Worthy of note is the included ability to determine, through a combinatorial calculation, the most appropriate pretreatment values to get preoptimized 3-D QSAR models; therefore, particular effort was given to data pretreatment and variable selection. To this aim, heavy use of the cross-validation (CV) techniques such as leave-one-out (LOO), leave-some-out (LSO), k-fold (KF), and Monte Carlo (MC) based CVs were applied either in standalone or in conjunction with a genetic algorithm (GA) as implemented in the genalg R package.[13] Guided Region Selection (GRS) using just one probe, or a compilation of different probes (Multi Probe Guided Region Selection (MPGRS)), is a further available variable-selection method as previously reported.[14−16]

The whole approach is described in detail below with its applications, either on ligand-based[17] or structure-based[18] prealigned, to molecular data sets previously reported using CoMFA and GRID/GOLPE, respectively.

At the time all the models were completed, comparison with OPEN3DQSAR software[1] was not possible due to patent restrictions that prevented CoMFA's free release in Italy.

COMPUTATIONAL METHODS

All calculations used a 6 blades (8 Intel-Xeon E5520 2.27 GHz CPU and 24 GB DDR3 RAM each) cluster (48 CPU total) running Debian GNU/Linux 5.03 64 bit operating system. The entire sequence was automated; to obtain 3-D QSAR models, the user needed only to input the prealigned data set and the values of the corresponding experimental parameters (i.e., biological activity).

Alignment Rules. The described methodology does not include an alignment engine; therefore, all the molecules contained in the data sets were used prealigned. Alignment procedures using several molecular superimposition programs are currently under investigation at the Rome Center for Molecular Design (RCMD, www.rcmd.it).

MIF Calculation. MIFs were generated using AutoGrid Software (based on the AMBER united-atom Force Field), although almost any probe can be used, in the current implementation 8 different probes (Table 1) were used in agreement with the most common residue atomic composition.

Table 1. List of the AutoGrid Probes Employed for MIF Calculation

probe type	description
A	aromatic carbon
C	aliphatic (sp^3) carbon
OA	hydrogen-bond-accepting oxygen
HD	hydrogen bonded to heteroatom
NA	hydrogen-bond-accepting amine nitrogen
N	amide nitrogen
e	electrostatic
d	desolvation

The sulfur probe (SA) was eliminated due to its close similarity to the OA (hydrogen-bond accepting oxygen) probe.

The calculated AutoGrid MIFs were imported in the R environment by means of the D2M R package (see below).

Statistical Analysis. The actual construction of statistical models was performed by a series of dedicated R packages as listed in Table 2 and arranged in Figure 2.

Table 2. List of the R Compiled Packages

R package	description
D2M	data to model
CAPP	combinatorial analysis of pretreatment parameters
MDP	model data pretreatment
CV	cross validation
VS	variable selection
GRS	guided region selection
MPGRS	multi probe guided region selection
ESP	external set prediction
YS	Y-scrambling

For a given Training Set, 3-D QSAR PLS[19] models were generated according to the MIF calculations described above. Different MIFs calculated with other softwares[3] could easily be imported as well.

While running the 3-D QSAR procedure, each package (named using the acronym build on the particular stage performed) was designed to achieve a statistical objective while saving the statistical information, workspace, and logs.

Package "D2M" (Data to Model). Two main steps are achieved by D2M: (1) merging of MIF(s) data with biological activities; (2) building as many raw 3-D QSAR models as the number of user-defined principal components (PC), saving the reloadable workspace and spreadsheet files containing the conventional correlation coefficients (r^2), the standard deviation errors of recalculation (SDEC), and the PLS recalculations for each PC.

Package "CV" (Cross-Validation). While the application of PLS, or other statistical techniques, to the training set are necessary to obtain a set of 3-D QSAR models, internal validation by CV is essential to assess chance correlation,[20] select the optimal model dimensionality (number of PCs), and measure the internal predictive ability by means of statistical coefficients such as cross-validated correlation coefficient (q^2) and standard deviation error of prediction (SDEP).

Different validation methods were included: LOO (leave-one-out), LSO (leave-some-out), KF (*k*-fold), and MC (Monte Carlo). For any implemented CV, several PLS calculations are performed as outlined in Table 3. All partial and final statistical

dx.doi.org/10.1021/ci300123x | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

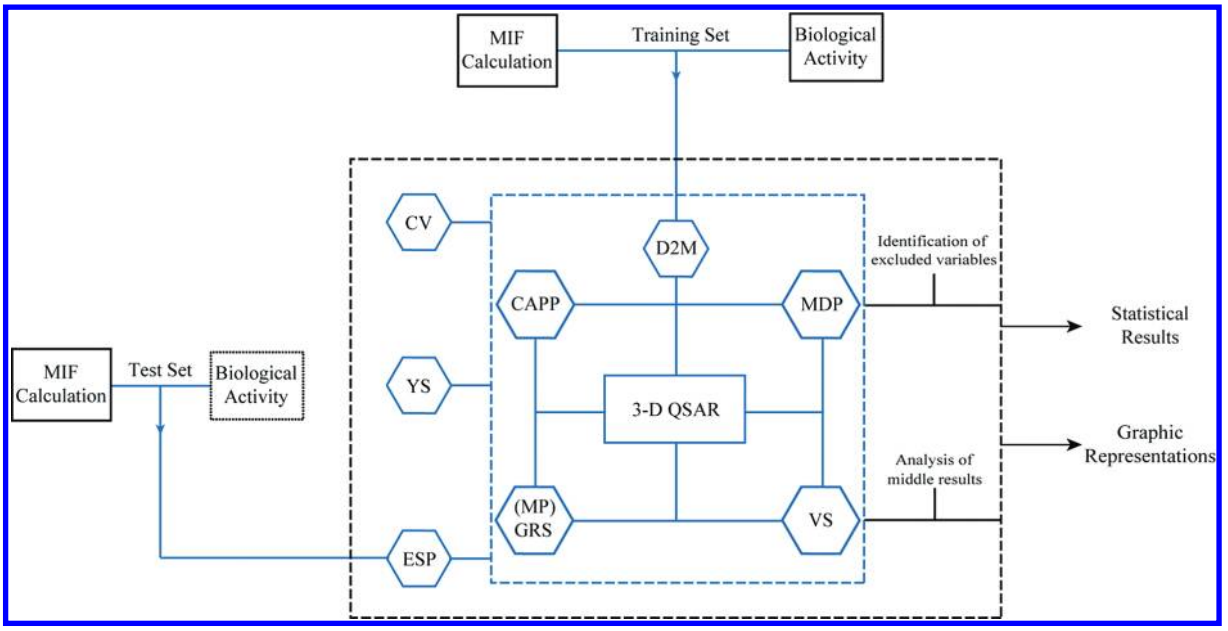


Figure 2. 3-D QSAutogrid/R process workflow. Acronyms inside the blue meshed square refer to the packages (steps) that effectively build or optimize the statistical PLS 3-DQSAR models; others are related to packages that perform analysis on these created models (see below for the description of each package). D2M: "Data to Model"; CAPP: "Combinatorial Analysis of Pretreatment Parameters"; MDP: "Model Data Pretreatment"; CV: "Cross-Validation"; VS: "Variable Selection"; (MP)GRS: "(Multi Probe) Guided Region Selection"; ESP: "External Set Prediction"; YS: Y-scrambling.

coefficients to exhaustively analyze the cross-validation process for each method are stored and can be inspected.

Table 3. Cross-Validation (CV) Methods Implemented in the "CV" Package, with the Relative Number of PLS Calculations[a]

CV method	LOO	LTO	KF	MC
CV combinations	N	$\frac{N!}{2!(N-2)!}$	K * iterations	iterations

[a]LOO: Leave One Out; LTO: Leave Two Out, KF: *k*-Fold, MC: Monte Carlo; *N*: no. of molecules in the Training Set, K: no. of *k* subsamples (folds).

Package "CAPP" (Combinatorial Analysis of Pretreatment Parameters). Raw data usually need to be pretreated to minimize redundancy,[21] even though pretreatment parameters' values are generally arbitrarily assigned without any systematic approach, thus ignoring a possible refinement based on the specific statistical model under development. To face this issue, a methodology was developed to systematically seek the more efficient data pretreatment values (energy cutoff, zeroing of very low data points, and minimum standard deviation cutoff). The CAPP package, through combinatorial analysis for each combination of parameters' values, builds the relative 3-D QSAR PLS model readily evaluated for each PC by different cross-validations (choice between LOO, LSO or KF). The optimal pretreatment combination is then selected according to the maximum q^2 while considering the percentage decrement value of sPRESS for each PC (Figure 3) as suggested by Gillet.[22]

Package "MDP" (Model Data Pretreatment). By the means of MDP, the user filters the data set values (MIFs) either setting the pretreatment parameters in an arbitrary way or as supplied from the CAPP procedure. A further parameter not included in the CAPP package is recognition of 1N kind of 2-level variable elimination (variables which take only 2 values in all of the data file, one of which appears only in one object). In the data pretreatment, the user can freely set which pretreatment to switch on or off. Although the logical sequence should be Field Cut-Off → Zeroing → SD Cut-Off →1N kind of 2-level variable elimination, no restriction is set to the chosen data pretreatment sequence. At each chosen pretreatment stage, the PLS is applied while saving the r^2, SDEC, and all the recalculated vs experimental responses for each extracted PC.

Package "VS" (Variable Selection). In order to improve the predictability of the statistical model, different variable-selection procedures like D-optimal design[23] (DOD), Fractional Factorial Design[24] (FFD), simulated annealing[25] (SA), and Genetic Algorithm[26] (GA) are currently used in QSARs[27−29] and 3-D QSARs.[30] In the VS package only the GA was actually implemented, while DOD, FFD, and SA are currently under development.

In the VS package, the GA-selection variable was implemented by the means of the R-binary genetic algorithm (genalg R package)[13] in combination with an *ad hoc* fitness evaluation R script in which SDEP/RMSEP was used as the discriminator.

Package "GRS" (Q2-Guided Region Selection). Along with the above variable selection procedures, other approaches to improve both the robustness and goodness of the models, such as q^2-guided region selection (q^2-GRS) and smart-region definition (SRD), were developed.[14,31] The q^2-GRS procedure was implemented in this approach and, for each separate probe field, was performed by the following steps: (1) the box is divided into a user-defined number of subregions leading to many PLS submodels; (2) each generated submodel is automatically validated through one the above-described cross-validation methods (LOO, LSO, MC, or KF); (3) for each PC, only those regions displaying a q^2 value higher than a user-defined threshold value are selected, and a new 3-D QSAR

dx.doi.org/10.1021/ci300123x | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

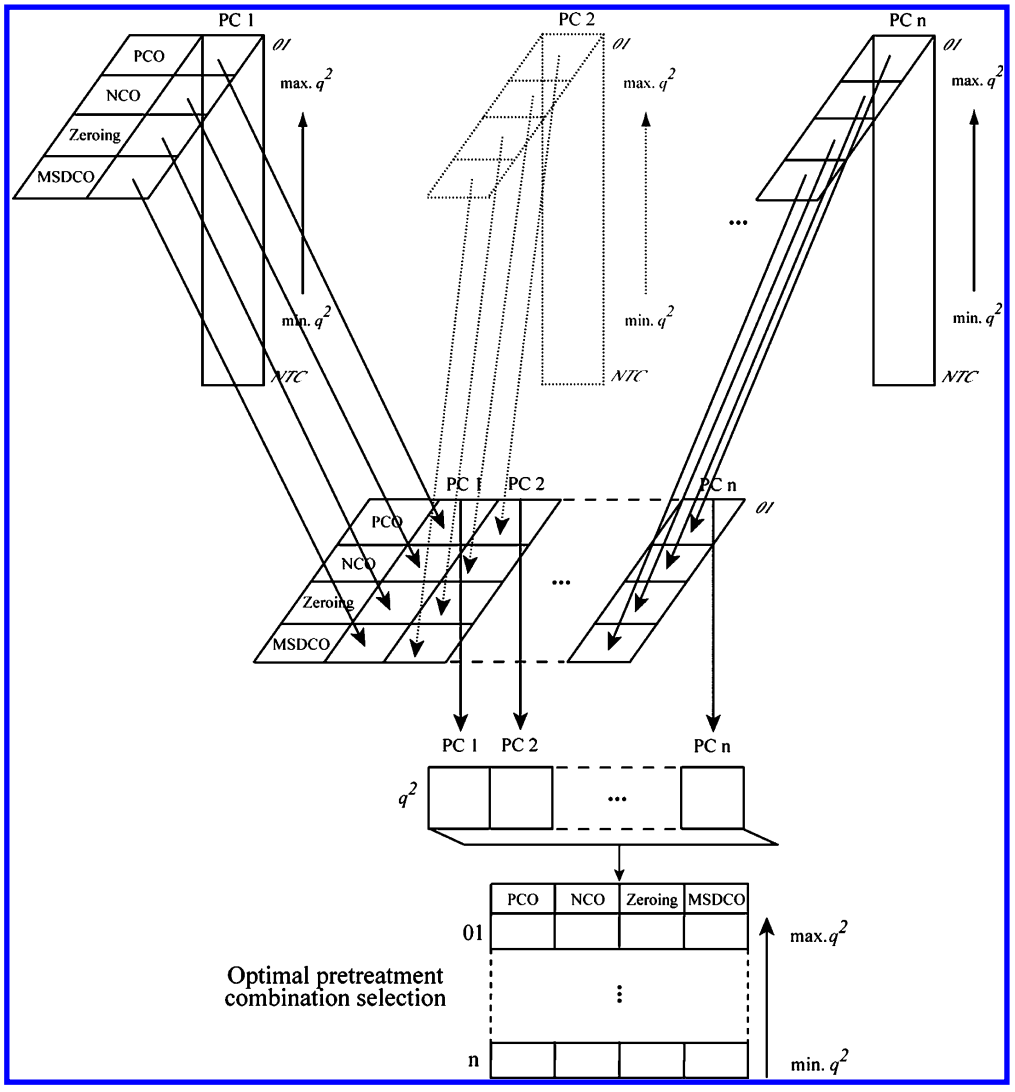


Figure 3. The CAPP Process. Through a combinatorial analysis according to predetermined setting values (as listed in Table 4), a certain number of combinations is investigated; after evaluation for each PC, the best pretreatment setting, the optimal overall-pretreatment combination is selected considering both the maximum q^2 and the percentage decrement sPRESS values.[22]

model is built and cross-validated; (4) the final model is chosen at the optimal number of PC.[22]

The q^2-GRS method implemented differs from the previously reported[32] as step 3 uses a bidimensional approach to select the more informative regions, so that for each dimension (PC) a different number of boxes may be retained to define the best model.

The maximum number of regions is only limited by the available amount of computer memory. Work is in progress to reduce this limitation.

Package "MPGRS" (Multi Probe Guided Region Selection). This package, an extension of the above single-probe q^2-GRS variable selection, considers as a further dimension probe interexchange, resembling what was previously described.[15] In particular, the following steps were considered: (1) starting from a series of monoprobe MIFs, for each subregion, a PLS model was built and cross-validated for i numbers of PCs (Figure 4, Step 1); (2) for the considered probe list, the obtained q^2 values (herein referred as first level q^2, q^2_{FL}) belonging to the same region and PC were sorted. The q^2_{FL} maximum values indicated the optimal probe for each PC (called "first level" PC, PC_{FL}) and region; all values constituted the PC_{FL} spreadsheet (Figure 4, Step 2); (3) as indicated in the PC_{FL} spreadsheet for each PC, the relative final multiprobe MIF and 3-D QSAR model were generated for q numbers of "second level" PCs (PC_{SL}) with associated q^2_{SL}; thus, each multiprobe model was indicated by two indexes recalling both PC_{FL} and PC_{SL} ($PC_{FL:SL}$) to which corresponded a $q^2_{FL:SL}$ value (Figure 4, Step 3); (4) the optimal multiprobe 3-D QSAR model was selected according to the $q^2_{FL:SL}$ values applying the percentage decrement value of sPRESS in a bidimensional way. First were selected models at fixed first-level PCs ($PC_{FL:q}$) (first dimension), then the PC_{SL} index (second dimension) was directly retrieved from the relative q^2_{SL}values. (Figure 4, Step 4). As a result, the optimal MPGRS model was characterized by two determined values of PCs ($PC_{FL:SL}$) that implicitly contained the most informative probe for each subregion and their best combination. Notably, the final model obtained by merging the selected subregions back into a single multiprobe MIF represented a very useful tool to derive advanced 3-D QSAR studies. The same *FL:SL* notation can also be applied to the simple GRS procedure described above.

Package "YS" (Y-Scrambling). Elimination of chance correlations of generated models was checked via the

dx.doi.org/10.1021/ci300123x | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 4. List of the Pretreatment Parameters (With Relative Editable Values and Number of Combinations) Analyzed by CAPP[a]

parameter	max value	min value	step value	no. comb
PCO	Max_{PCO}	Min_{PCO}	$Step_{PCO}$	$PCO_{CB} = \left[\left(\frac{Max_{MCO} - Min_{MCO}}{Step_{MCO}}\right) + 1\right]$
NCO	Max_{NCO}	Min_{NCO}	$Step_{NCO}$	$NCO_{CB} = \left[\left(\frac{\lvert Max_{NCO}\rvert - \lvert Min_{NCO}\rvert}{Step_{NCO}}\right) + 1\right]$
Zeroing	Max_Z	Min_Z	$Step_Z$	$Z_{CB} = \left[\left(\frac{Max_Z - Min_Z}{Step_Z}\right) + 1\right]$
MSDCO	Max_{MSDCO}	Min_{MSDCO}	$Step_{MSDCO}$	$MSDCO_{CB} = \left[\left(\frac{Max_{MSDCO} - Min_{MSDCO}}{Step_{MSDCO}}\right) + 1\right]$

no. total combinations = $PCO_{CB} \times NCO_{CB} \times Z_{CB} \times MSDCO_{CB}$

[a]PCO (Positive Cut Off); Max_{PCO} (maximum PCO value); Min_{PCO} (minimum PCO value); $Step_{PCO}$ (incremental PCO value); PCO_{CB} (number of combinations that origins only for the PCO analysis); NCO (Negative Cut Off); Max_{NCO} (maximum NCO value); Min_{NCO} (minimum NCO value); $Step_{NCO}$ (incremental NCO value); NCO_{CB} (number of combinations that origins only for the NCO analysis); Zeroing: zeroing of very low data points; Max_Z (maximum zeroing value); Min_Z (minimum zeroing value); $Step_Z$ (incremental zeroing value); Z_{CB} (number of combinations that origins only for the zeroing analysis); MSDCO (Minimum SD cut-off); Max_{MSDCO} (maximum MSDCO value); Min_{MSDCO} (minimum MSDCO value); $Step_{MSDCO}$ (incremental MSDCO value); $MSDCO_{CB}$ (number of combinations that origins only for the MSDCO analysis).

experimental response scrambling approach.[33] The YS package allowed a user-defined number of iterations randomly coupled property/activity values to evaluate the risk of chance correlation.[20]

Package "ESP" (External Set Prediction). As predictions are the main purpose for any QSAR-related model, validation through external test sets is mandatory. Furthermore, ESP was compiled as an independent program in place of internal validation (CV) to select the optimal number of PCs and as an extension for CAPP, GRS, and MPGRS variable selections. Such an approach allowed model optimization for external prediction.

In the current version, the ESP applied the same training-set pretreatment. In the case of (MP)GRS models, the training set selected/merged regions were retained and applied to the test set.

Graphical Results. Besides the essential role of PLS, a successful 3-D QSAR is also due to the number of graphical insights that can be generated to help interpretation of numerical results. Without graphical analyses, 3-D QSAR would be reduced to QSAR with a great number of parameters. CoMFA success measurable in more than 3340 papers citing it (SciFinder accessed February 2012) is surely due to the fact that SYBYL allowed depiction of user-friendly 3-D plots correlating structure with activity. Regarding the current method, the gnuplot style implemented in R[34] through the ggplot library[35] allowed creation of 2-D graphics score plots, loading plots, regression plots, inner-correlation plots, biplots, and many others in a straightforward way. Through an ad-hoc "in house" utility, the MIF, PLS coefficient, activity contribution 3-D plots were written in a format to be used by molecular viewers such as UCSF Chimera,[10] Python Molecular Viewer (PMV),[11] Autodock Tools (ADT),[12] and Jmol[36] to generate high-quality colored molecular maps. To better interpret the 3-D QSAR model, more than one map can be overlapped to generate a complete scenario (see Figures in the application section).

RESULTS AND DISCUSSION

The application of the new 3-D QSAR procedure to a data set of aligned opioid-receptor antagonists[17] (LB data set) and two data sets of HCV NS5B allosteric inhibitors[18] (SB data sets) is reported.

Ligand-Based Case Studies: Opioid-Receptor Antagonists. To test the new 3-D QSAR procedure, a series of opioid-receptor antagonists, previously described by Peng et al.[17] in a CoMFA application, were used to build several 3-D QSAR models. The data set was comprised of prealigned 74 compounds, separated into training and test sets with associated δ, μ, κ opioid-binding affinities.

The new method was conducted maintaining the molecular alignment used in the original paper.[17] Applying a spacing grid of 1 Å and considering the binding-affinity sets of data (δ, μ, κ) with 8-molecular probes, 24 3-D QSAR models were obtained.

During the model definitions the CV was conducted via (1) Leave-One-Out (LOO), (2) Leave-Two-Out (LTO), (3) k-Fold (KF), and (4) Monte Carlo (MC) methodologies.

Initially, the raw models (Tables S1, S15, S28) were optimized through the CAPP package setting the pretreatment intervals as listed in Table 5 using the *k*-fold cross-validation with 5-random groups and 100 iterations and monitoring the q^2 and SDEP values. For the e probe, several trials (data not shown) led to set a fixed Negative Cut Off (NCO) value, equal to −0.5.

A total of 12,221 combinations for each 3-D QSAR model were processed using 5% sPRESS reduction as suggested by Gillet[22] to select the best combination and derive the pretreated PLS model; this led to an average q^2_{KSFCV} value increment ranging from 26% to 55% (Tables S7, S21, S34).

The best pretreated models were then optimized through the GA-variable selection (VS package), setting the number of chromosomes, number of generation, percentage of mutation chance, and % of best individuals that are kept into the next generation to 50, 100, 0.005, and 20, respectively (Tables S6, S20, and S33). As reported in Table 6, the statistical quality of the models is similar to those obtained by the original CoMFA[17] (Table 7), although the 3-D QSAutogrid/R models

dx.doi.org/10.1021/ci300123x | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

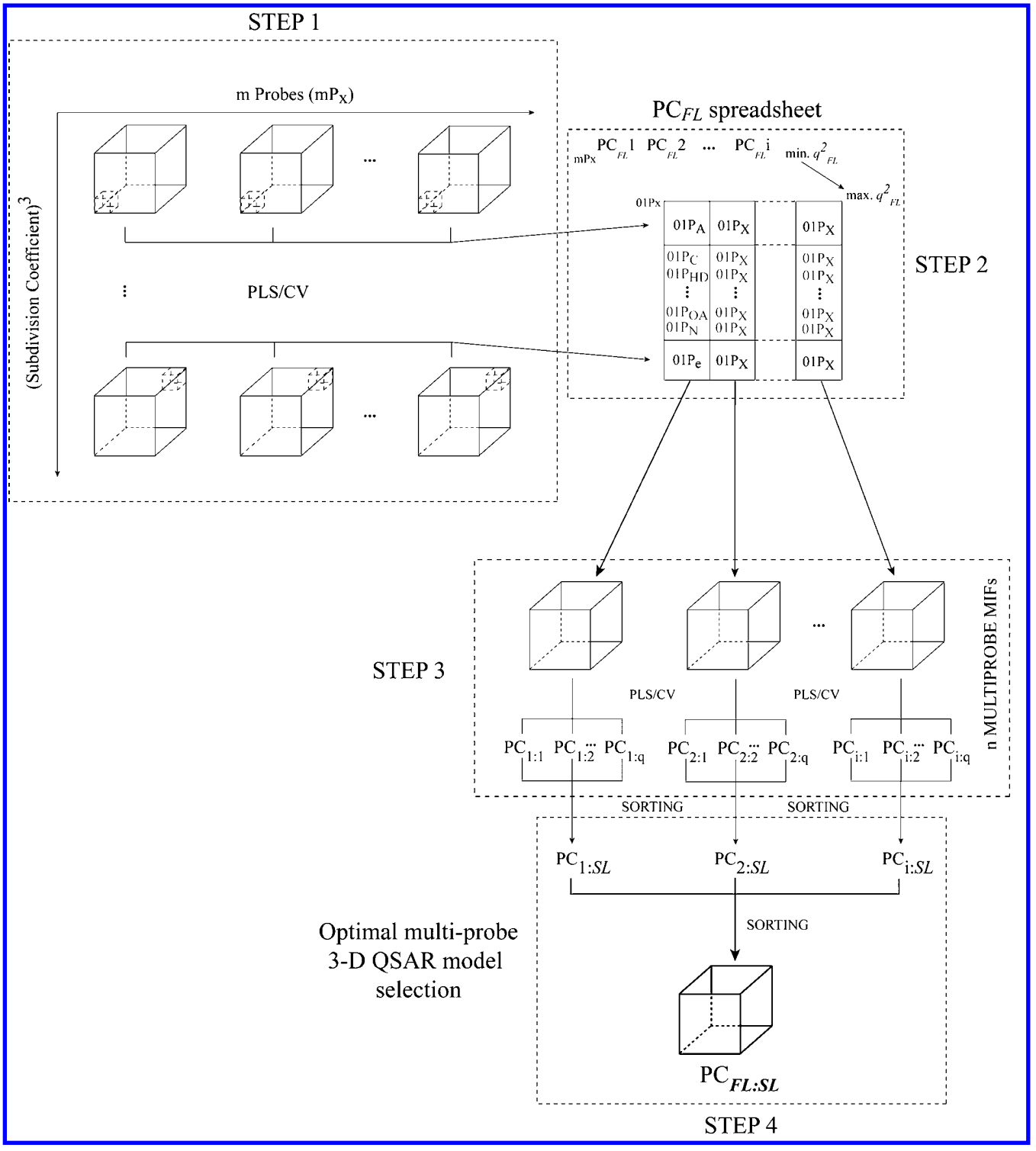


Figure 4. The MPGRS Process. For each subregion a PLS model was built and cross-validated (Step 1) in order to obtain, for each "first level" PC (PC_{FL}), the relative optimal multiprobe list (Step 2) and 3-D QSAR model characterized by an established number of "second level" PCs (PC_{SL}, Step3). Applying a bidimensional criterion (Step 4), the optimal multiprobe 3-D QSAR model was then selected.

seem slightly more robust being described by a fewer number of PCs. For direct comparison, a further 3-D QSAR model was built using only aliphatic carbon-atom and the electrostatic probes in a single model applying only the CAPP procedure. As reported in Table 7 and Table 8, the CoMFA and the Autogrid/R double-probe models (DP) were statistically similar. Furthermore, the YS package was applied leading to low r^2_{YS} and negative q^2_{YS} values thus supporting the lack of chance correlation in reported models.

Along with numerical output, a series of plots (scores, loadings, actual field, PLS coefficients, activity contribution and CoMFA-like) were generated to allow interpretation of the 3-D QSAutogrid/R models. Analyses of the 3-D QSAR models were conducted using all the graphical plots. Regarding the δ-opioid receptors, the CoMFA-like plots (the default TRIPOS

dx.doi.org/10.1021/ci300123x | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 5. CAPP Settings Adopted for the δ-, μ-, κ-Opioid Receptor Antagonists 3-D QSAR Models[a]

min value	parameter	max value	step
0	PCO	50	5.0
0	Zeroing	0.1	0.01
0	MSDCO	5	0.05

[a]PCO: Positive Cut Off; Zeroing: zeroing of very low data points; MSDCO: Minimum SD Cut Off.

Table 6. Opioid-Receptor Antagonists: Autogrid/R PLS Models Statistical Results (CAPP and GA Processes Were Applied)[b]

model	OR	P	PC	r^2	q^2_{LOO}	q^2_{KSFCV}	r^2_{YS}	q^2_{YS}
1	δ	A	2	0.81	0.73	0.70	0.27	−0.37
2	δ	C	2	0.82	0.74	0.71	0.32	−0.35
3	δ	HD	2	0.83	0.75	0.72	0.33	−0.34
4	δ	NA	2	0.83	0.75	0.73	0.31	−0.34
5	δ	N	2	0.83	0.76	0.72	0.29	−0.32
6	δ	OA	2	0.83	0.74	0.71	0.32	−0.37
7	δ	e	3	0.69	0.58	0.56	0.22	−0.19
8	δ	d	3	0.70	0.59	0.55	0.24	−0.30
9	μ	A	3	0.91	0.82	0.76	0.57	−0.50
10	μ	C	3	0.90	0.81	0.78	0.59	−0.50
11	μ	HD	3	0.90	0.81	0.75	0.47	−0.49
12	μ	NA	3	0.91	0.81	0.78	0.59	−0.50
13	μ	N	3	0.91	0.83	0.78	0.52	−0.61
14	μ	OA	3	0.91	0.83	0.77	0.51	−0.61
15	μ	e[a]	1	0.31	0.21	0.20	0.06	−0.10
16	μ	d	3	0.72	0.60	0.52	0.27	−0.39
17	κ	A	2	0.78	0.58	0.49	0.42	−0.37
18	κ	C	3	0.81	0.62	0.55	0.54	−0.53
19	κ	HD	3	0.82	0.72	0.65	0.34	−0.41
20	κ	NA	3	0.80	0.62	0.54	0.55	−0.47
21	κ	N	3	0.80	0.61	0.52	0.54	−0.48
22	κ	OA	3	0.82	0.65	0.59	0.54	−0.44
23	κ	e[a]	2	0.35	0.20	0.18	0.13	−0.18
24	κ	d	3	0.58	0.38	0.34	0.29	−0.36

[a]The e models 15 and 23 reported were only pretreated due to too few variable after GA selection. [b]OR: Opioid-receptor data, P: Autogrid Probe, PC: optimal number of principal components/latent variables, r^2: conventional square-correlation coefficient; q^2_{LOO}: cross-validation correlation coefficient using the leave-one-out method; q^2_{KSFCV}: cross-validation correlation coefficient using the *k*-fold cross-validation with 5 random groups and 100 iterations; r^2_{YS}: average square correlation coefficient obtained after Y-scrambling process using 100 iterations; q^2_{YS}: average cross-validation correlation coefficient using the leave-one-out method obtained after Y-scrambling process using 100 iterations.

Table 7. Opioid-Receptor Antagonists: Original CoMFA Models Statistical Results[a]

model	OR	P	PC	r^2	q^2_{LOO}	q^2_{KSFCV}
25	δ	CoMFA	4	0.91	0.69	-
26	μ	CoMFA	4	0.92	0.67	-
27	κ	CoMFA	6	0.96	0.60	-

[a]OR: Opioid-receptor data; P: standard CoMFA Probe C_{sp3}^+, PC: optimal number of principal components/latent variables, r^2: conventional square-correlation coefficient; q^2_{LOO}: cross-validation correlation coefficient using the leave-one-out method; q^2_{KSFCV}: cross-validation correlation coefficient using the *k*-fold cross-validation with 5 random groups and 100 iterations.

Table 8. Opioid-Receptor Antagonists: Autogrid Double-Probe (DP) PLS Models Statistical Results (Only the CAPP Process Was Applied)[a]

model	OR	P	PC	r^2	q^2_{LOO}	q^2_{KSFCV}	r^2_{YS}	q^2_{YS}
28	δ	Autogrid DP	3	0.83	0.70	0.67	0.41	−0.50
29	μ	Autogrid DP	4	0.85	0.65	0.63	0.52	−0.53
30	κ	Autogrid DP	3	0.84	0.67	0.63	0.50	−0.53

[a]OR: Opioid-receptor data; P: Autogrid double probe (DP, C, and e probes); PC: optimal number of principal components/latent variables; r^2: conventional square-correlation coefficient; q^2_{LOO}: cross-validation correlation coefficient using the leave-one-out method; q^2_{KSFCV}: cross-validation correlation coefficient using the *k*-fold cross-validation with 5 random groups and 100 iterations; r^2_{YS}: average square correlation coefficient obtained after Y-scrambling process using 100 iterations; q^2_{YS}: average cross-validation correlation coefficient using the leave-one-out method obtained after Y-scrambling process using 100 iterations.

StdDev*PLS Coeff contour plots) for the C probe that allowed highlighting the molecular features indicating where sterically bulky groups were favorable (green) or unfavorable (yellow) are reported in Figure 5. A CoMFA model of the data was reproduced with a recent SYBYL version, and the related contour plots confirmed that the new procedure generated similar graphical information (compare Figures 5 and S2 with Figure S4).

Model 2 (Table 6) CoMFA-like maps (Figure 5) were in good agreement with those reported by Peng[17] and displayed two green areas indicating that around bulky groups in positions 5′, 6′, 7′ (R5 substituents of Core 1 as in ref 17) of **18**, **20**, **22**, and **50** (Figure 5) were well tolerated. On the other hand, two yellow contours (unfavorable steric interactions) were present and give some hints to explain the reduced activities of **30** and **67** that bear bulky groups in R1 (Core 1, **30**) and R3 (Core 4, **67**) and that of **68** (one of the least active) which to some extent occupies both regions. The latter region was not viewable in Figure S4, likely due to differences in force field and molecular formats.

Slightly less agreement was observed between the electrostatic probe-derived plots (compare Figure 6 and Figure S3 with Figure S5 in the Supporting Information) and its CoMFA counterparts; differences were mainly located on a supplemental region that was found on the NTI indole group. Likely, these differences were surely due to the force-field differences; AutoGrid uses a united-atom force field, while CoMFA uses the all-atom TRIPOS force-field.

Similar results were obtained analyzing the models for the μ- and κ-opioid receptors; therefore, to avoid redundancy, the analyses are reported as Supporting Information (pages S13−S32).

As in the original CoMFA paper,[17] the 3-D QSAutogrid/R models were externally validated using test sets (TS1) compiled from the original data sets and external test sets (TS2) compiled from different literature sources (Table 9). The predictions were similar to the original CoMFA paper and, therefore, are not commented on in detail (see pages S51−S52 in the Supporting Information).

Structure-Based Case Studies: Hepatitis C Virus NS5B-Polymerase Inhibitors. A detailed GRID/GOLPE application was reported on HCV NS5B non-nucleoside inhibitors (NNI)[18] binding at two distinct allosteric sites (thumb and

dx.doi.org/10.1021/ci300123x | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

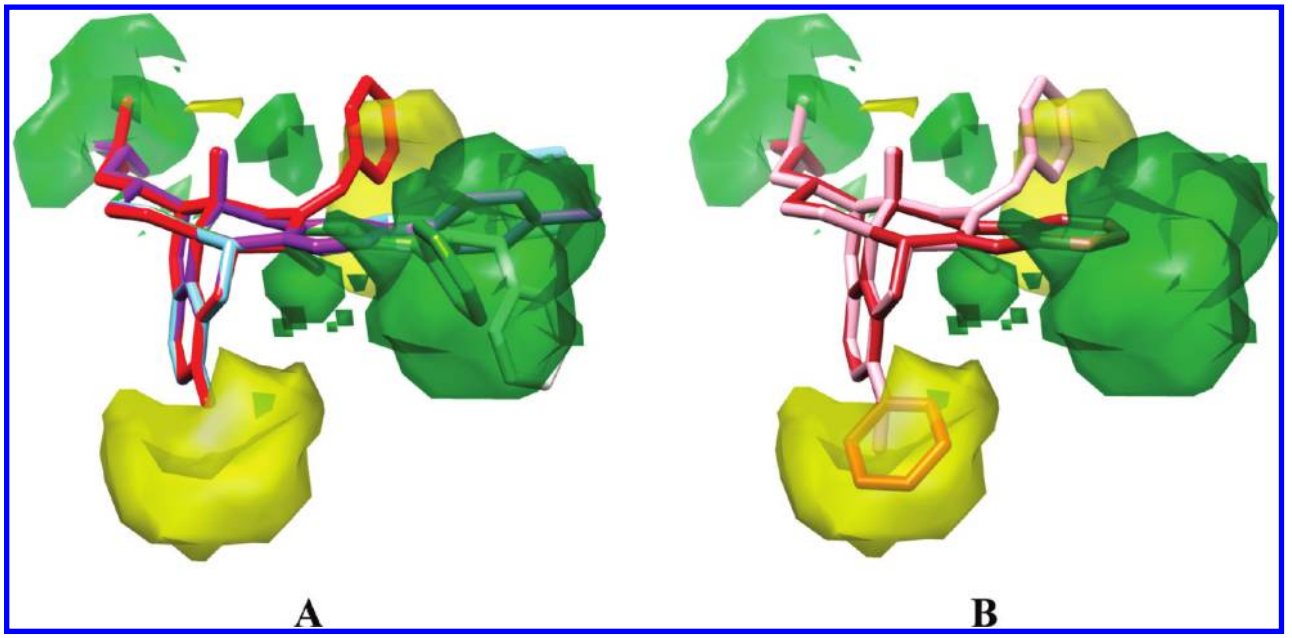


Figure 5. CoMFA-like steric-contour map derived from the C probe for the δ-opioid receptors. A: compounds **18** (sky blue), **20** (white), **22** (green), **50** (purple), and **67** (red). B: compounds **30** (brown) and **68** (pink). Contour levels: 85% (positive green, negative yellow). Hydrogen atoms are omitted for the sake of clarity. In A and B are reported the same contour maps..

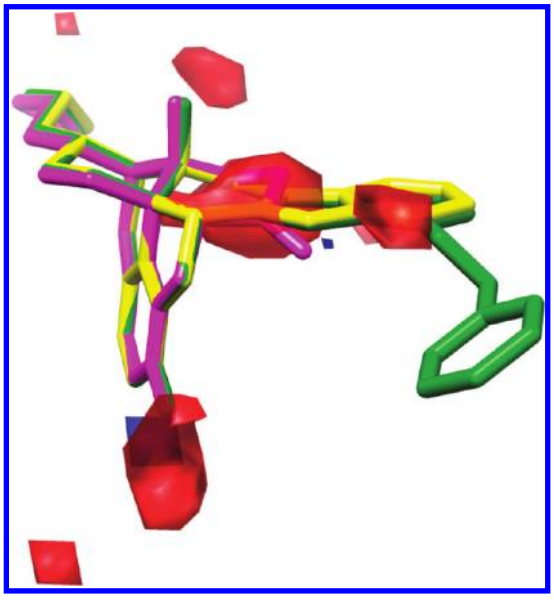


Figure 6. CoMFA-like electrostatic contour map derived from e probe for the δ-opioid receptors. Compounds: **Naltrexone** in magenta, **NTI** in yellow, **24** in green. Contour levels: 85% (positive blue, negative red). Hydrogen atoms are omitted for clarity.

palm). Thumb and palm NNIs data sets were chosen, as training sets to derive 3-D QSAR models using the new procedure with the purpose of comparing the results with those obtained by the well-established GRID/GOLPE method.[18]

Interaction energies between the eight probes and each molecule were computed using a grid spacing of 1 Å; thus a total of 16 3-D QSAR initial raw models were built. To build optimized models, CAPP analyses were conducted analyzing 27,775 combinations (Table S47) with the best ones chosen applying the same criteria for the previous opioid case. The CAPP procedure for both thumb- and palm-training sets

Table 9. δ Test Set Predictions Indicated by SDEP Values[a]

OR model	P	PC	$SDEP_{TS1}$	$SDEP_{TS2}$
1	A	2	0.66	0.80
2	C	2	0.64	0.77
3	HD	2	0.62	0.74
4	NA	2	0.64	0.82
5	N	2	0.64	0.76
6	OA	2	0.65	0.75
7	e	3	0.81	1.20
8	d	3	0.90	1.12

[a]OR model: Opioid-receptor model of Table 6; P: Autogrid probe; PC: optimal number of principal components/latent variables; $SDEP_{TS1}$: standard deviation error of prediction for the original test set; $SDEP_{TS2}$: standard deviation error of prediction for the external test set.

(Table 10) led to models comparable to those previously reported.[18]

Furthermore, similarly as for the LB case study, the YS package was applied leading to low r^2_{YS} and negative q^2_{YS} values thus supporting the lack of chance correlation in reported models.

Activity contribution, PLS coefficients, and CoMFA-like maps were generated, and their interpretation was in full agreement with those reported for the previous GRID/GOLPE models.

For comparison purposes, the PLS-coefficients plots obtained are shown in Figure 7 with the two methodologies, and their similarity and information content are clearly visible.

Furthermore for each SB model, the reduced test set of 21 (thumb) and 23 (palm) compounds in the original paper[18] were employed to compare the predictive ability of AutoGrid/R and GRID/GOLPE. As for the statistical values, predictions were also very similar with no further comments (Table 11).

Application of Multi Probe Guided Region Variable Selection. Variable selection is an important task in 3-D QSAR in order to achieve models with an enriched data/noise ratio

dx.doi.org/10.1021/ci300123x | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 10. PLS Analysis Results for the Thumb- and the Palm-Structure Based Autogrid/R and Original GRID/GOLPE C1= 3-D QSAR Models[a]

data set	P	PC	r^2	q^2_{LOO}	q^2_{KSFCV}	r^2_{YS}	q^2_{YS}
thumb	A	2	0.90	0.67	0.64	0.70	−0.63
thumb	C	2	0.90	0.68	0.65	0.70	−0.60
thumb	HD	2	0.92	0.75	0.73	0.68	−0.69
thumb	NA	3	0.95	0.75	0.73	0.79	−0.66
thumb	N	3	0.95	0.76	0.73	0.78	−0.67
thumb	OA	3	0.95	0.77	0.73	0.77	−0.54
thumb	e	3	0.98	0.58	0.52	0.92	−0.55
thumb	d	1	0.58	0.36	0.36	0.27	−0.38
thumb	GRID/GOLPE/C1=	3	0.99	-	0.69	-	-
palm	A	3	0.96	0.73	0.62	0.68	−1.62
palm	C	3	0.96	0.73	0.62	0.69	−1.59
palm	HD	1	0.90	0.75	0.71	0.44	−0.76
palm	NA	2	0.97	0.62	0.52	0.84	−0.76
palm	N	2	0.97	0.62	0.55	0.85	−0.87
palm	OA	1	0.86	0.67	0.64	0.32	−0.66
palm	e	3	0.96	0.85	0.82	0.73	−1.01
palm	d	3	0.93	0.62	0.39	0.73	−1.80
palm	GRID/GOLPE/C1=	3	0.99	-	0.55	-	-

[a]P: Autogrid Probe or GRID C1= probe; PC: optimal number of principal components/latent variables; r^2: conventional square-correlation coefficient; q^2_{LOO}: cross-validation correlation coefficient using the leave-one-out method; q^2_{KSFCV}: cross-validation correlation coefficient using the *k*-fold cross-validation with 5 random groups and 100 iterations; r^2_{YS}: average square correlation coefficient obtained after Y-scrambling process using 100 iterations; q^2_{YS}: average cross-validation correlation coefficient using the leave-one-out method obtained after Y-scrambling process using 100 iterations.

Table 11. Thumb- and Palm-External Test Set Prediction Obtained from Structure Based AutoGrid/R and Original GRID/GOLPE C1= 3-D QSAR Models[a]

data set	P	PC	$SDEP_{ext}$
thumb	A	2	0.69
thumb	C	2	0.69
thumb	HD	2	0.76
thumb	NA	3	0.66
thumb	N	3	0.66
thumb	OA	3	0.67
thumb	e	3	0.63
thumb	d	1	0.67
thumb	GRID/GOLPE/C1=	3	0.59
palm	A	3	1.14
palm	C	3	1.11
palm	HD	1	1.29
palm	NA	2	1.04
palm	N	2	1.04
palm	OA	1	1.03
palm	e	3	1.18
palm	d	3	1.18
palm	GRID/GOLPE/C1=	3	1.08

[a]P: Autogrid Probe or GRID C1= probe; PC: optimal number of principal components/latent variables; $SDEP_{ext}$: standard deviation error of prediction for the external test set.

and predictability.[16,37] The default 3-D QSAR approaches use one or more probes distributed on regularly spatial grids without the possibility of mixing probe information into one single grid leading to a multiprobe (MP) grid. This was achieved by selecting the most informative subregions (guided region selection, q^2-GRS package) for each considered probe so that the whole grid was reconstituted with pieces from several MIFs as described above (MPGRS package). This approach was initially reported by Tropsha (modified q^2 GRS)[15] and implemented in the new approach designated multi probe guided region selection (MPGRS). The MPGRS as conceived, if correctly applied, should result in a powerful modified 3-D QSAR technique; therefore, with the aim of optimizing the 3-D QSAR models, the MPGRS procedure was implemented and applied to the above case studies to test its validity and potentiality.

The MPGRS was thus applied to the three case studies [Table S12, S14, S25, S27, S38, S40, S54, S55, and S56], and as reported in Table 12, the mixed-probes models maintained a comparable level of statistical coefficients (compare Table 12 data with those of the above monoprobe 3-D QSARs: Tables S42 and S45 and Table 11).

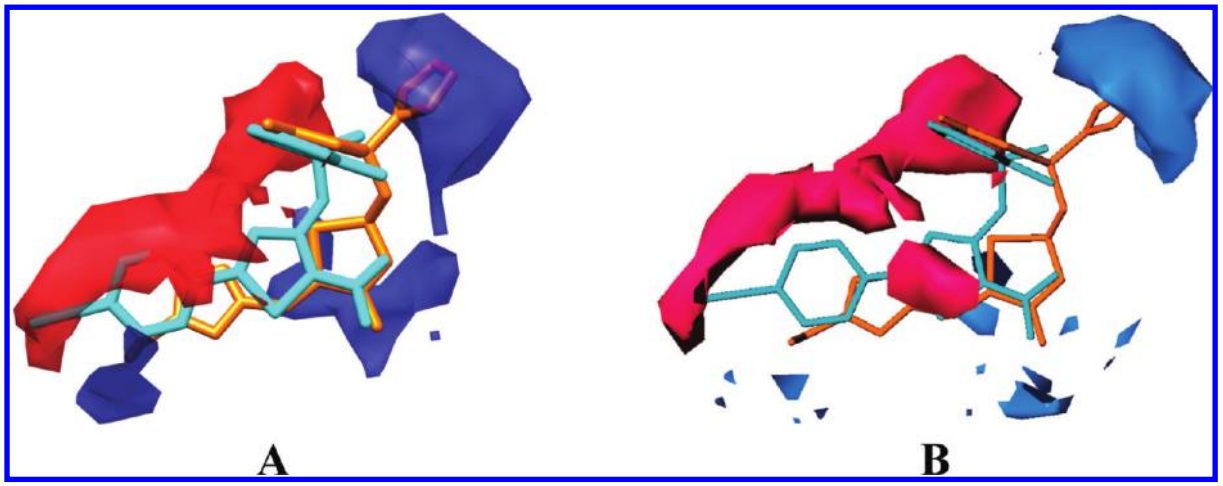


Figure 7. PLS-coefficients contour maps using the thumb-training set; only the highest active (**6** in cyan) and one of the lowest active (**11** in orange) compounds are shown. A: AutoGrid/R PLS coefficients contour maps derived from A probe analysis (contour levels: 60%, positive red, negative blue). B: GRID/GOLPE PLS coefficients contour maps derived from C1= GRID probe analysis (contour levels: 0.0008 red, −0.0008 blue).

dx.doi.org/10.1021/ci300123x | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 12. Statistical Results Obtained from MPGRS Analysis for the Thumb- and the Palm-HCV Training Sets[a]

	MPGRS 3-D QSAR					
data set	$PC_{FL:SL}$	r^2	q^2_{K5FCV}	r^2_{YS}	q^2_{YS}	$SDEP_{ext}$
thumb	2:2	0.95	0.90	0.50	−0.67	0.74
palm	1:2	0.99	0.91	0.61	−0.93	1.06

[a] $PC_{FL:SL}$: optimal number of principal first level (FL) and second level (SL) components/latent variables for the MPGRS model; r^2: conventional square-correlation coefficient; q^2_{LOO}: cross-validation correlation coefficient using the leave-one-out method; q^2_{K5FCV}: cross-validation correlation coefficient using the k-fold cross-validation with 5 random groups and 100 iterations; r^2_{YS}: average square correlation coefficient obtained after Y-scrambling process using 100 iterations; q^2_{YS}: average cross-validation correlation coefficient using the leave-one-out method obtained after Y-scrambling process using 100 iterations.

All the MPGRS models were analyzed; of particular interest were the SB-derived alignments that checked for the MPGRS ability to propose a pseudoreceptor. Therefore, the detailed analyses is reported for the HCV palm-training set.

Applying q^2 threshold value of 0.4, 11-MIFs subregions were selected (Figure 8) to build the multiprobe MIF, and were color coded according to that reported in Table S53. In particular, 5 regions were taken from the N MIF, 4 from the NA, and the last two from HD and e probes, respectively. High agreement between the selected regions and the HCV NS5B-palm binding pocket surface was observed. These 11 subregions were highly informative to allow a very detailed interpretation of the final MPGRS 3-D QSAR model (Figure 8).

Furthermore, by analyzing the selected subregions' PLS coefficients, a series of pharmacophoric-like points were extrapolated (Figures 9 and S20). According with the relative

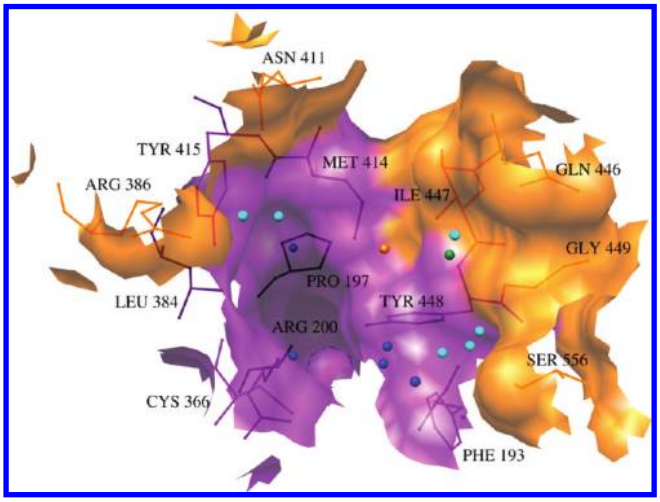


Figure 9. MPGRS 3-D QSAR palm model key points. The points are color coded: in blue N (amidic nitrogen) probe key points, in cyan those from NA (hydrogen acceptor nitrogen) probe, in green and orange those from HD (hydrogen donator) and e (electrostatic) probes, respectively.

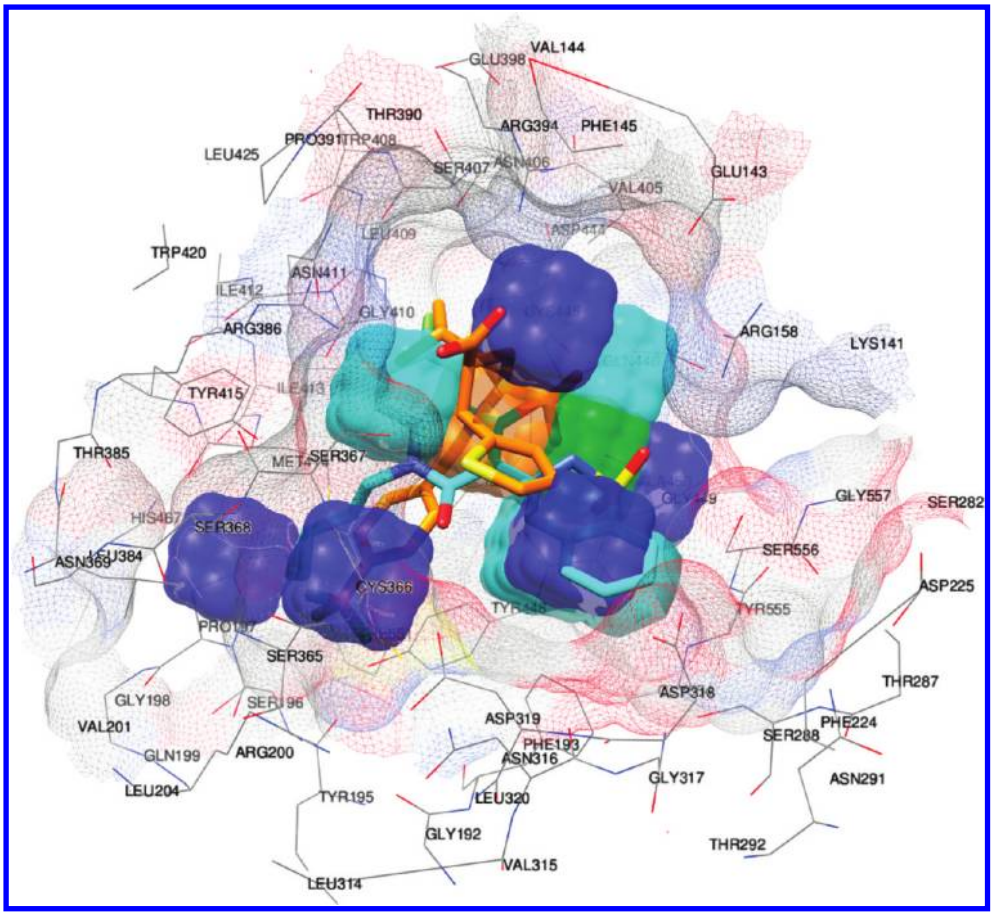


Figure 8. Most informative subregions derived from the final MPGRS 3-D QSAR palm model. The regions are color coded: in blue N, in cyan NA, in green HD, and in orange e.

dx.doi.org/10.1021/ci300123x | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

probe subregion, similar signs, as in the GRID/GOLPE approach, could have a different meaning, and correct interpretation remains problematic.

The five N (amidic nitrogen) probe subregions in Figure 8 indicate the localization of hydrophobic interactions, and their positions in the palm allosteric-binding site were indeed characterized by the nonpolar residues Phe193, Pro197, Arg200, Cys366, Leu384, Met414, Tyr415, and Tyr448, forming a hydrophobic pocket (violet surface in Figure 9) in agreement with that previously reported.[38−40]

The 4 NA (hydrogen acceptor nitrogen)-associated key points overlapped with residues bearing hydrogen-acceptor groups[38−40] (Arg386, Asn411, Gln446, Ile447, Gly449, Tyr448, Ser555, orange surfaces in Figure 9), while the single green HD (hydrogen donator) region was in proximity with the Tyr448 and Gly449 main-chain nitrogens.[39,40] The last electrostatic subregion presented difficulties of interpretation, and no specific role was assigned.

Regarding the internal predictive ability of the MPGRS models, the multiprobe approach in general was not improved; nevertheless, the interpretation of the model was greatly enhanced. MPGRS allowed focusing on the most informative regions around the ligands and used all the probes together to reduce the chances of missing important correlations when using single probe 3-D QSARs.

CONCLUSION

The use of the AutoGrid software coupled with ad-hoc R-based scripts allows an alternative procedure (3-D QAutogrid/R) to generate 3-D QSAR analyses, similarly as the well-established CoMFA and GRID/GOLPE techniques, improving both the use of chemical data and minimizing time and human-machine interactions. The procedure was validated with three data sets, covering both ligand-based and structure-based alignment methodologies. The main features of the new procedure are automation and flexibility that permit the iterative generation of hundreds/thousands of 3-D QSAR models selecting the best one in a completely independent way and improving the amount of important information generated from detailed 3-D QSARs analyses. Furthermore, the possibility to extrapolate/merge the more informative interactions from different probe fields into a single multiprobe MIF lead to more comprehensive interpretations. Case studies results and comparisons with the other mentioned methods show how the new procedure should be a useful tool, based on free software, to conduct advanced 3-D QSAR analyses. The implementation of the MPGRS, although not improving the models overall predictive abilities, greatly enhanced their interpretation. To the best of the author's knowledge, this is the very first free MPGRS implementation. Furthermore, 3-D QSAutogrid/R has been recently used as core engine to prepare more that 240 3-D QSAR models used to generate the first 3-D QSAR server[41−43] (www.3d-qsar.com).

All the described R packages are available through the CRAN-package repository; tutorials with example files are also available through the www.3d-qsar.com and www.rcmd.it Web sites.

ASSOCIATED CONTENT

Supporting Information

Full statistical results for all the data sets with example of GRS, 3-D QSAutogrid/R double-probe, and bidimensional plots. Graphical comparisons between the new method and the original CoMFA analyses for the opioid data sets. MPGRS statistical values for all the data sets and related plots. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rino.ragno@uniroma1.it.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Many thanks to Prof. Garland R. Marshall (Department of Biochemistry and Molecular Biophysics, Washington University in St. Louis) for critical reading of the manuscript, helpful discussions, and encouragement and also for the use of TRIPOS software in his laboratory. Many thanks are also due to Youyi Y. Peng for having provided the opioid-receptor antagonists data sets. This study was supported by grants from Italian Ministry of University and Research (MIUR Grant 2008F8T894_002 and 2008ZTN724_003).

REFERENCES

(1) Tosco, P.; Balle, T. Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields. *J. Mol. Model.* **2011**, *17*, 201−208.

(2) 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.

(3) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(4) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear Pls Estimations (Golpe) - an Advanced Chemometric Tool for Handling 3d-Qsar Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.

(5) Cruciani, G.; Watson, K. A. Comparative Molecular-Field Analysis Using Grid Force-Field and Golpe Variable Selection Methods in a Study of Inhibitors of Glycogen-Phosphorylase-B. *J. Med. Chem.* **1994**, *37*, 2589−2601.

(6) Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in drug design--a review. *Curr. Top. Med. Chem. 10*, 95−115.

(7) Sippl, W. 3D-QSAR - Applications, Recent Advances, and Limitations. In *Recent Advances in QSAR Studies*; Puzyn, T., Leszczynski, J., Cronin, M. T., Eds.; Springer: Netherlands: Vol. *8*, pp 103−125.

(8) Team, R. D. C. *The R Foundation for Statistical Computing*. http://www.r-project.org/ (accessed month day, year).

(9) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791.

(10) The University of California, S. F. U. UCSF chimera. http://www.cgl.ucsf.edu/chimera/ (accessed month day, year).

(11) Python Molecular Viewer (PMV). Molecular Graphics Laboratory, T. S. R. I. http://mgltools.scripps.edu/ (accessed month day, year).

(12) AutoDockTools (ADT). Molecular Graphics Laboratory, T. S. R. I. http://mgltools.scripps.edu/ (accessed month day, year).

(13) Willighagen, E. *genalg: R Based Genetic Algorithm*, 0.1.1.; 2005.

(14) Cho, S. J.; Tropsha, A. Cross-validated R2-guided region selection for comparative molecular field analysis: a simple method to achieve consistent results. *J. Med. Chem.* **1995**, *38*, 1060−1066.

(15) Cho, S. J.; Tropsha, A.; Suffness, M.; Cheng, Y. C.; Lee, K. H. Antitumor agents. 163. Three-dimensional quantitative structure-activity relationship study of 4′-O-demethylepipodophyllotoxin analogs

using the modified CoMFA/q2-GRS approach. *J. Med. Chem.* **1996**, *39*, 1383−1395.
(16) Cruciani, G.; Clementi, S.; Pastor, M. GOLPE-guided region selection. *Perspect. Drug Discovery Des.* **1998**, *12*−*14*, 71−86.
(17) Peng, Y.; Keenan, S. M.; Zhang, Q.; Kholodovych, V.; Welsh, W. J. 3D-QSAR comparative molecular field analysis on opioid receptor antagonists: pooling data from different studies. *J. Med. Chem.* **2005**, *48*, 1620−1629.
(18) Musmuca, I.; Caroli, A.; Mai, A.; Kaushik-Basu, N.; Arora, P.; Ragno, R. Combining 3-D quantitative structure-activity relationship with ligand based and structure based alignment procedures for in silico screening of new hepatitis C virus NS5B polymerase inhibitors. *J. Chem. Inf. Model.* **2010**, *50*, 662−676.
(19) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−743.
(20) Clark, M.; Cramer, R. D. The Probability of Chance Correlation Using Partial Least-Squares (Pls). *Quant. Struct.-Act. Relat.* **1993**, *12*, 137−145.
(21) Cruciani, G. *Molecular interaction fields: applications in drug discovery and ADME prediction*; Wiley-VCH: Weinheim, 2006; p xviii, 307 p.
(22) Wold, S.; Johansson, E.; Cocchi, M. *PLS: Partial Least Squares Projections to Latent Structures in 3D QSAR in Drug Design: Theory Methods and Applications*; ESCOM Science Publishers: 1993.
(23) Mitchell, T. J. An algorithm for the construction of "D-optimal" experimental designs. *Technometrics* **2000**, *42*, 48−54.
(24) Box, G. E. P.; Hunter, W. G.; Hunter, J. S.; *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*; John Wiley & Sons: 1978; p 653.
(25) Kirkpatrick, S.; Gelatt, C. D., Jr.; Vecchi, M. P. Optimization by simulated annealing. *Science* **1983**, *220*, 671−680.
(26) Holland, J. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*; A Bradford Book: 1992.
(27) Baroni, M.; Clementi, S.; Cruciani, G.; Kettanehwold, N.; Wold, S. D-Optimal Designs in Qsar. *Quant. Struct.-Act. Relat.* **1993**, *12*, 225−231.
(28) Puzyn, T.; Leszczynski, J.; Cronin, M. T. D. *Recent advances in QSAR studies: methods and applications*; Springer: Dordrecht; New York, p xiv, 423 p.
(29) Shen, M.; LeTiran, A.; Xiao, Y. D.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant-agents using k nearest neighbor and simulated annealing PLS methods. *J. Med. Chem.* **2002**, *45*, 2811−2823.
(30) Kubinyi, H.; Folkers, G.; Martin, Y. C. *3D QSAR in drug design*; Kluwer Academic: Dordrecht; Boston, MA, 1998; p v. < 2- >.
(31) Pastor, M.; Cruciani, G.; Clementi, S. Smart region definition: a new way to improve the predictive ability and interpretability of three-dimensional quantitative structure-activity relationships. *J. Med. Chem.* **1997**, *40*, 1455−1464.
(32) Cho, S. J.; Tropsha, A. Cross-validated R2-guided region selection for comparative molecular field analysis: a simple method to achieve consistent results. *J. Med. Chem.* **1995**, *38*, 1060−1066.
(33) Wold, S.; Eriksson, L. *Chemometrics Methods in Molecular Design*; VCH: Weinheim: 1995.
(34) Crawley, M. J. *The R book*; Wiley: Chichester, England; Hoboken, NJ, 2007; p viii, 942 p.
(35) Wickham, H. *ggplot2: elegant graphics for data analysis*; Springer: New York, 2009.
(36) *Jmol: an open-source Java viewer for chemical structures in 3D.*
(37) Cruciani, G.; Clementi, S.; Baroni, M. Variable Selection in PLS Analysis. In *3D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 551−564.
(38) Tedesco, R.; Shaw, A. N.; Bambal, R.; Chai, D.; Concha, N. O.; Darcy, M. G.; Dhanak, D.; Fitch, D. M.; Gates, A.; Gerhardt, W. G.; Halegoua, D. L.; Han, C.; Hofmann, G. A.; Johnston, V. K.; Kaura, A. C.; Liu, N.; Keenan, R. M.; Lin-Goerke, J.; Sarisky, R. T.; Wiggall, K. J.; Zimmerman, M. N.; Duffy, K. J. 3-(1,1-dioxo-2H-(1,2,4)-benzothiadiazin-3-yl)-4-hydroxy-2(1H)-quinolinones, potent inhibitors of hepatitis C virus RNA-dependent RNA polymerase. *J. Med. Chem.* **2006**, *49*, 971−983.
(39) Li, T.; Froeyen, M.; Herdewijn, P. Insight into ligand selectivity in HCV NS5B polymerase: molecular dynamics simulations, free energy decomposition and docking. *J. Mol. Model.* **2010**, *16*, 49−59.
(40) Ryu, K.; Kim, N. D.; Il Choi, S.; Han, C. K.; Yoon, J. H.; No, K. T.; Kim, K. H.; Seong, B. L. Identification of novel inhibitors of HCV RNA-dependent RNA polymerase by pharmacophore-based virtual screening and in vitro evaluation. *Bioorg. Med. Chem.* **2009**, *17*, 2975−2982.
(41) Ballante, F.; Musmuca, I.; Patsilinakos, A.; Ragno, R. An Alternative Method for Generating 3-D QSAR Models using Free Software. In *5th Joint Sheffield Conference on Chemoinformatics*; Sheffield, UK, 2010.
(42) Patsilinakos, A.; Ballante, F.; Musmuca, I.; Ragno, R. 3-D QSAR SERVER − A 3-D QSAR Models Database for Virtual Screening. In *14th Hellenic Symposium on Medicinal Chemistry*; Thessaloniki, Greece, 2010.
(43) Musmuca, I.; Ballante, F.; Ragno, R. R/AUTOGRID/ADT Combination As An Alternative To Build 3-D QSAR Models. Methodologies And Applications. In *18th European Symposium on Quantitative Structure-Activity Relationships*; Rhodes, Greece, 2010.