www.3d-qsar.com. A Web Applications that Brings 3-D QSAR to all Electronic Devices. 1. CoMFA Models from pre-Aligned Datasets.

Rino Ragno[§,*]

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

*Rino Ragno: rino.ragno@uniroma1.it

Abstract

Comparative molecular field analysis (CoMFA), introduced in 1988, was the first 3-D QSAR method ever published and sold. Since then thousands of application, articles and citation have proved 3-D QSAR as a valuable method to be used in drug design. Several other 3-D QSAR methods have appeared, but still CoMFA remains the most used and cited. Nevertheless from a survey on the Certara web site it seems that CoMFA is no more available.

Herein is presented a python implementation of the CoMFA (Py-CoMFA). Py-CoMFA is usable through the www.3d-qsar.com web applications suites portal by mean of any electronic device that can run a web browser. A benchmark using 30 different publicly available datasets were used to assess the Py-CoMFA usability and robustness. Comparisons with published results proved Py-COMFA to highly overlap those obtained with the original CoMFA. The used datasets were pre-aligned, in a future report the www.3d-qsar.com will be proved also as a tool to develop 3-D QSAR models from scratch. In conclusion Py-CoMFA is a valuable tools for non informatics skilled user and also as a possible support to teach 3-D QSAR.

Introduction

Quantitative structure-activity relationships (QSARs) is a general term to indicate computational mathematical methods aimed to build models which attempts to find statistically significant correlations between a series of molecular structures and their associated biological property.[1] In term of drug design (DD), molecular structures refers to molecules' properties, descriptors and/or their substituents or interaction energy fields, biological property corresponds to an experimental biological/biochemical endpoint such as binding affinity, activity, toxicity or rate constants. In QSARs, the structure-activity correlation is carried out by means of chemometric method including multiple linear regression (MLR),[2] principal component analysis (PCA),[3] principal component regression (PCR)[4, 5] and partial least square (PLS).[6] Various QSAR approaches have been developed and reported[7] since its first broad formalisms,[8, 9] particularly focusing in drug design and agrochemicals sciences. QSAR methods suffer from serious limitations due to only one/two-dimensional (1/2-D) structures description inclusion. Stereochemistry of compounds and their spatial arrangement are usually neglected; thus providing inadequate features that might be important to describing potential drug-receptor interactions, also QSAR application is limited only within congeneric or scaffold related series.[10] Moreover the lack of practically no graphical output, makes interpretation of results in chemical terms, difficult or almost impossible.[11] Given these limitations, three-dimensional (3-D) QSAR methodologies[12, 13] (3-D QSAR) emerged as an evolution of Hansch[9] and Free-Wilson [8] QSAR approaches.

For many years, comparative molecular field analysis (CoMFA) has been used as a synonym for 3-D QSAR as it was the first method developed by Cramer[14] who joined the

Wold's projection of latent variables (also partial least square, PLS)[6] and the Goodford's GRID[15] technologies. CoMFA worked in a fashioned smooth multistep procedure to build 3-D QSAR model usable to predict biological activity of a ligand from its 3-D structure without the use of any experimental and/or calculated physical-chemical features.

The underlying idea of CoMFA, similarly to any QSAR and 3-D QSARs, is that differences in a target property, e.g. biological activity, are often closely related to equivalent changes in shapes and intensities of non-covalent calculated interaction surrounding the molecules (the molecular interaction fields, MIFs[16-18]), these are the basis of the so called field based 3-D QSAR (FB 3-D QSAR)[19] methods. During the procedure (see below), the molecules virtually merged in a cuboid grid are described by MIFs, calculated at each grid point by means of a predefined probe atom or group of atoms. In the standard CoMFA only two potentials are used, namely steric (STE) and electrostatic (ELE), calculated by means of the Lennard-Jones[17] and Coulomb law definition. Although other statistical techniques can be used, the MIF are linearly correlated with the training set biological activity data by means of PLS,[6] which identifies and extracts the quantitative influence of specific chemical features (latent variables or principal components) of molecules on their biological activity. For a visual interpretation, (greatest innovation introduced by the CoMFA method) the 3-D QSAR model is presented by freely rotatable 3-D pictures consisting of colored contour plots representing the values for the corresponding field variables.

Many books and reviews detailing CoMFA and hence 3-D QSAR methods have been published and extensively discussed,[6, 20-31] therefore herein herein are just pointed the most important objectives of a 3-D QSAR:

- Explanation of biological data in relationship with three-dimensional and quantitative properties of molecules;
- Optimization of existing compounds by analyzing the 3-D contour maps;
- Prediction of biological activities of untested and yet unavailable compounds. Predicting the biological activity of a candidate drug, as well as its pharmacokinetic properties and toxicity, early in the drug discovery process, has the ability to reduce cost and time in early stage of drugs design and developing.

The procedure to build a 3-D QSAR model involves the following steps (Figure 1):

1. *Training-set selection* – although not strictly necessary, for greater statistical significance, it is advisable to consider a number of molecules of at least 15-20, in the same biological potency unit of measurement (homogeneity of data) and with the same mechanism of action. The range of biological activity should be as large (at least 2 log units) and as continuous as possible (with no clusters).[32]
2. *Definition of Alignment Rules* – by using one of the many available approaches to superimpose molecular structures[26] (atom by atom, pharmacophore based, receptor based, etc.). This allow molecular comparison based on chemical and structural differences. The alignment rules should be as much as possible reproducible, thus any manual or arbitrary setting should be avoided.[28] In this step are also included the generation of 3-D conformations and eventually their selection from a conformational search.[31]
3. *MIF calculation* – Superimposed molecules are placed in the center of a grid box, to calculate MIFs between ligands and probe atoms located at each grid intersection

(nodes). These data represent the molecular descriptors, i.e. the independent variable matrix for the subsequent statistical analysis. Different grid extension, steps and/or probes can be set and data pretreatment is needed to reduce redundancy and to optimized the subsequent statistical model definition.

4. *Statistical model definition* – By means of PLS (or other statistical techniques) MIFs are correlated with the 'dependent variable' (biological activity, Y space), to find any possible relationships between them. Since independent variables (X space) are numerically much larger than the number of tested compounds, PLS is used to extract principal components (PCs) to reduce X space dimensionality and establish a valid statistical correlation.[33, 34]
5. *Model validation* – The defined model need to be checked for robustness, chance correlation and predictive ability.[33, 34] To this a q^2 and SDEP coefficients are used at any validation stage. Model statistical robustness is evaluated by cross-validation (CV),[35-37] while lack of chance correlation can be assessed by Y-scrambling procedure.[33-35] Although Clark et al. report,[35] CV is often improperly used to indicate the model's predictive ability, the use of external test sets should be considered the only outmost method to evaluate any model predictive ability.[32]
6. *Graphical interpretation* – This is the radical innovation introduced by CoMFA and the most valuable and informative step of any FB 3-D QSAR model. Differently from Hansch type QSARs, the 3-D QSAR model can be actually visualized through a number of polyhedrons obtained by connecting similar grid points. The visualization is normally obtained by selecting iso-value nodes that can represent different chemical features determined by the used probe. At each grid point can be associated different values: the

actual field (MIF); the PLS coefficients at a given number of principal components to obtain the PLS Coefficients plots; the standard CoMFA contour maps associating at each grid point the product of PLS coefficient by the corresponding MIF standard deviation calculated on all the training set molecules. By means of a deep and time consuming analysis of the graphical output it is possible to describe how training set molecular structural features can positively or negatively affect the endowed bioactivities and thus design new molecular entities as potential new and more active derivatives, the soul of any 3-D QSAR model.

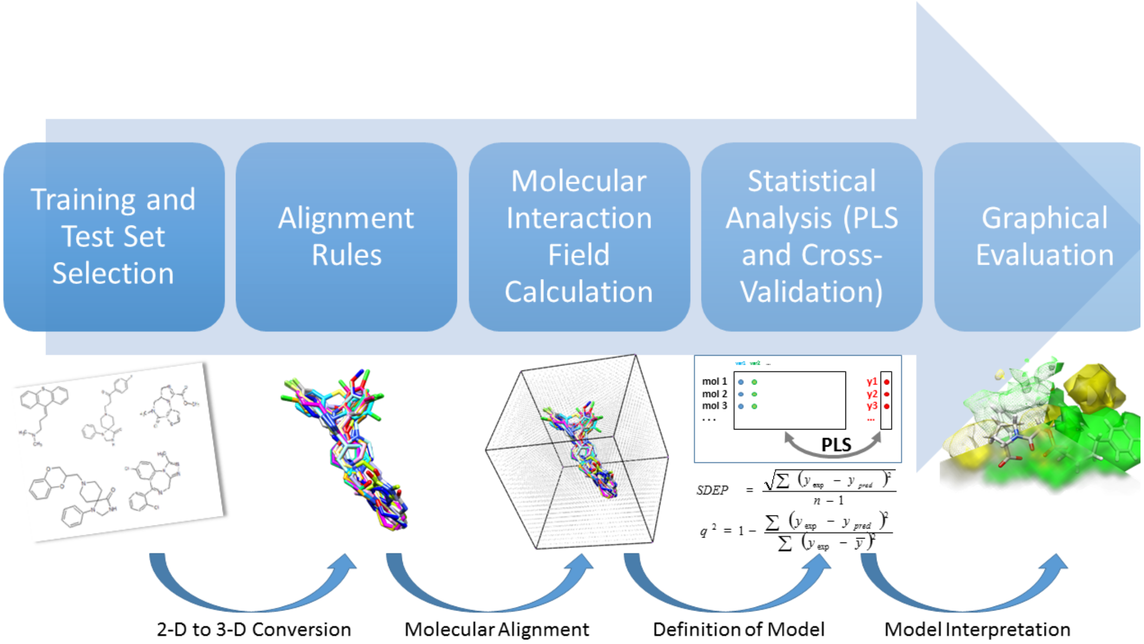
Training and Test Set Selection
Alignment Rules
Molecular Interaction Field Calculation
Statistical Analysis (PLS and Cross-Validation)
Graphical Evaluation
var1 var2 ...
mol 1
mol 2
mol 3
...
y1
y2
y3
...
PLS
$SDEP = \frac{\sqrt{\sum (y_{exp} - y_{pred})^2}}{n-1}$
$q^2 = 1 - \frac{\sum (y_{exp} - y_{pred})^2}{\sum (y_{exp} - \bar{y})^2}$
2-D to 3-D Conversion
Molecular Alignment
Definition of Model
Model Interpretation

Figure 1. 3-D QSAR model building flowchart

In general 3-D QSAR methodology is of big demand to the scientific community as evidenced by a literature survey (data from scopus.com accessed 2019 July 26th) since the original CoMFA article in 1988 as many as 4802 articles were retrieved using several combination of “3D” and “QSAR” keywords with the “CoMFA” one with a global Hirsch index of 106 and a total number of citation of 95744 (Table 1). Since 1988, a general positive trend demonstrate the high appealing of 3-D QSAR during the years (Figure 2). Only a little flexion was observed in the last few years, mainly due likely to the fact that Certara (the owner of the Sybyl CoMFA containing suites package initially belonging to Tripos) seems not anymore selling the CoMFA package. Due to its supremacy, until year 2000, CoMFA was almost the only applied 3-D QSAR method (Figure 2), some other procedures or methodologies, like CoMSIA[38] or the GRID/GOLPE pair,[39, 40] and the recently Open3DQSAR[41] appeared on the scenario.

Table 1. Aggregate results from a scopus.com literature survey on 3-D QSAR and CoMFA from 1988 to 2018.

#	Keywords and logical connection	Arts[1]	Hs[2]	Cits[3]
1	"CoMFA"	2766	88	59277
2	"3d qsar" OR "3-d qsar" OR "3d-qsar" OR "3dqsar"	3925	91	68063
3	1 OR 2	4802	106	95744

1: Numebr of publishe articles; 2: Hirsh index; 3: Number of citations

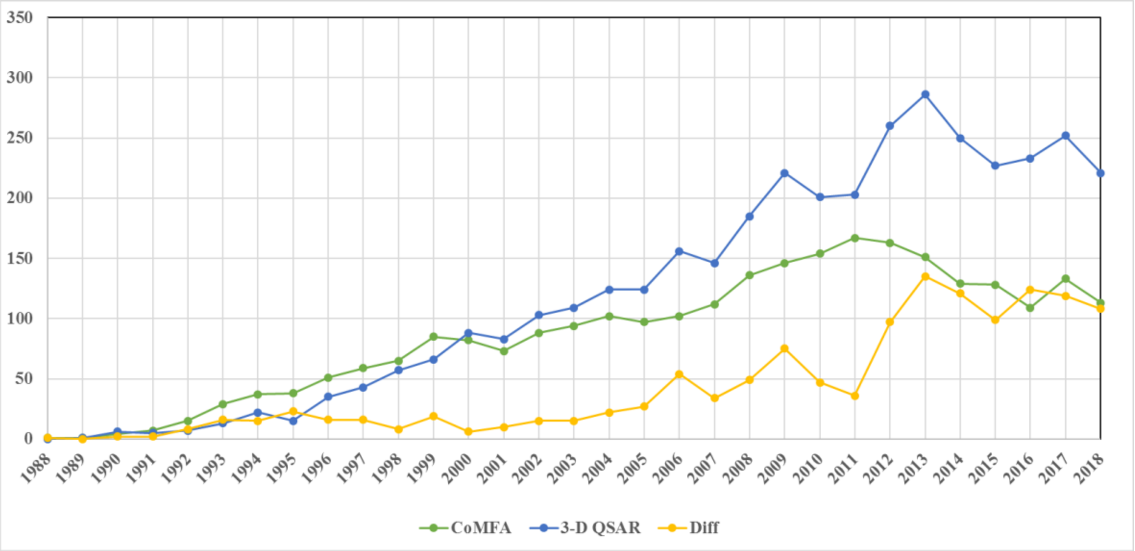

Figure 2. Number of scientific publication retrieved with scopus.com (accession date 2019 July 26th) from 1988 to 2018. In the ordinates are reported the number of

occurrences for the for the “CoMFA”, “3-D QSAR” occurrences and their difference “Diff”. The search cutoff was set to 2018 to avoid uncomplete indexed references in the recent period.

The above listed 3-D QSAR model building steps (Figure 2) have been deeply investigated and several protocols have been reported.[42-44] Nearly all the steps can be performed using different software[14, 39, 41, 45] so that using a single data set different 3-D QSAR models can be built obtaining similar and overlapping results. Nevertheless, to perform the steps of flowchart in Figure 1 any user is asked to install specialized software either costly or even open source.[41, 45]. Here, it is presented Py-CoMFA the 3-D QSAR engine of the very first 3-D QSAR web application accessible to anyone and by which a model can be easily build and graphically analyzed by means of any electronic device able to run a web browser (personal computer, tablet and smartphones). Along with Py-CoMFA, three other web applications (Was) are also included (Py-MolEdit, Py-ConfSerch and Py-Align, and many other will be opened and inserted soon) to perform all the above described 3-D QSAR steps and all available through www.3d-qsar.com electronic address. Py-MolEdit, Py-ConfSerch and Py-Align web applications will be detailed elsewhere, while herein will be discussed and detailed Py_CoMFA, the embedded python based implementation of the original CoMFA method.

Computational methods.

The above cited WAs are hosted in the Rome Center for Molecular Design (RCMD) computer server running linux operating system. All code was developed in python 3.5 language with combination of javascript for graphical integration.

Py-CoMFA. The procedure used in this study, is a python[46] flexible implementation of the original CoMFA developed by TRIPOS This procedure exploits implementation of

TRIPOS 5.2 force field, to calculate the molecular interaction fields (MIFs), and the scikit-learn module python implementation[47] for the statistical analysis (such as PLS regression and internal and external validation). For each data set, xyz coordinates (in angstroms) of the cuboid grid box used for the MIFs computation are automatically set to embrace all the training and test sets aligned compounds spanning a user desired value with the default set to 5 Å in all six directions. For each considered dataset (see table 1, in Benchmark datasets paragraph) was developed a CoMFA like standard pretreated model (Energy cutoff of ±30 Kcal/mol). Two level of grid steps were used at 1.0 and 2.0 Å as common values used in GRID/GOLPE and CoMFA, respectively. All models robustness were evaluated by cross-validation (CV) using both leave one out (LOO) and leave 20% of the experiments out (5 groups, L5O) using the k-fold method with 100 iterations. During CV the minimum standard deviation error, in analogy to the CoMFA's minimum sigma (min_sigma), was set to 2.0 to speed-up the calculations, reduce memory usage and data redundancy. The code allow also to run CV with lower min_sigma, as suggested in the GOLPE manual (http://www.miasrl.com/software/golpe/manual/), to evaluate final model and crossvalidation with the same set of data. Although easily computable, neither optimization of the pretreatment setting or grid steps were performed (further investigations are ongoing on systematic variation of variable pretreatment and selection). Y-scrambling,[33, 48] (100 permutations) was performed to check absence of chance correlation and finally available external test sets were used to evaluate models' goodness of prediction, the last validation step before real application to the ultimate goal use of any QSAR model: prediction of new untested compounds. For the graphical output a Gaussian cube file output format writing routine was implemented. This file can be easily downloaded and used as it is readable from most of the molecular graphic programs (pymol,[49] UCSF Chimera,[50] Jmol,[51-53] JSmol,[54] etc). In the present version graphical output for the MIFs, the PLS coefficients and CoMFA like plots (i.e. PLS coefficients × Standard Deviation) are embedded in the python code. All the calculation are be saved into the database for future reprocess of the built 3-D QSAR model and output of all possible data. Differently from the original CoMFA application of minimum sigma affected both fitting and crossvalidation runs. Differently from how suggested in the literature r^2_{pred} was calculated using the average of the test set experimental values instead of the average of the training set (Cramer suggestion[14]). This choice was preferred, as the original CoMFA r^2_{pred} would indicate good values only in the case the used test set would have an experimental values average comparable to that of the training set.

Benchmark datasets. As reported by Coats[27, 55, 56] the original CoMFA steroid dataset (ID 21 in table 1) is normally used as benchmark for 3-D QSAR procedures, therefore the dataset was incorporated in the list of all datasets here used as the first list of molecules to explore Py-CoMFA features. Furthermore, to investigate on the Py-CoMFA potentialities a series of 30 publicly available pre-aligned molecular datasets and associated bioactivities (Table 1) were retrieved from literature and used to build Py-CoMFA based FB 3-D QSAR models.

Table 1. List of datasets used in this study.

Dataset ID	**Dataset Name**	**Numbers of Molecules**			**Activity Range[a]**
		Dataset	***Training***	***Test***	
1	**ACE**[57,58]	114	76	38	7.8
2	**AchE**[58, 59]	111	74	37	5.2
3	**BZR**[58, 60]	147	98	49	3.9
4	**GPB**[58, 61]	66	44	22	5.5
5	**COX2**[58, 62]	282	188	94	5

6	**DHFR**[58, 63]	361	237	124	6.5
7	**THERM**[58, 61, 64]	76	51	25	9.7
8	**THR-1**[58]	88	59	29	4.1
9	**ATA**[65]	94	72	22	4.9
10	**AT2**[20]	28	28	NA	3.9
11	**CCR5**[66]	75	63	12	3.4
12	**YOPH**[67]	39	35	4	4.3
13	**KOA**[68]	39	31	8	2.9
14	**MX**[69]	29	29	NA	5.3
15	**DAT**[70]	42	36	6	4
16	**TP2A**[71]	25	25	NA	3.5
17	**CBRA**[72]	32	32	NA	4
18	**AI**[73,74]	78	78	NA	4.5
19	**HIVPR**[28, 75]	113	93	20	5.9
20	**GSK3B**[21,75, 76]	42	34	8	3.7
21	**STEROIDS**[77,14]	21	21	NA	2.9
22	**GHS**[77]	31	31	NA	3.5
23	**D2R**[24,78]	38	32	6	4.6
24	**D4R**[24]	38	32	6	4
25	**DIAZEPAM DI/DS**[79]	42	42	NA	4.1
26	**DIAZEPAM DI**[79]	42	42	NA	4.1
27	**DIAZEPAM DS**[79]	42	42	NA	4.1
28	**THR-2**[80]	88	72	16	4.0
29	**TRY**[80]	88	72	16	4.7
30	**FXA**[80]	88	72	16	3
	Min[b]	21	21	4	2.9
	Max[c]	361	237	124	9.7
	Total[d]	2051	1554	497	30[e]

NA: Not Available; a) bioactivity range; b) minimum number of compounds in data, training and test sets or activity range; c) maximum number of compounds in data, training and test sets or activity range; d) comprehensive number of compounds in data, training and test sets; e) number of biological activities.

Py-CoMFA at Work.

A detailed tutorial on the use Py-CoMFA to build 3-D QSAR models from pre-aligned dataset is reported in the corresponding blog implemented in www.3d-qsar.com.

Results and Discussion

Py-CoMFA applied to the 30 datasets returned r^2s, q^2s and r^2_{pred} values of good level (Table 2 and Supplementary Material Table 1S) showing the implemented python code as an effective tool to develop 3-D QSAR models using pre-aligned datasets. The only poor model was GHS displaying an r^2s of 0.463. Excluding GHS the r^2s and q^2s were in the ranges of 0.656-0.997 and 0.191-0.792, respectively. For the dataset with available external test sets the r^2_{pred} values ranged from -4.323 (DAT dataset) to 0.933 (YOPH dataset). Only 5 models returned negative r^2_{pred} values.

Table 2. Py CoMFA models' r^2s, q^2s and r^2_{pred} data. The reported models were built with C.3 atom probe with a +1 charge using the combination of steric and electrostatic fields (STE+ELE) and 2Å grid spacing.

#	**Dataset**	$\boldsymbol{r^2}$	**ONC**	$\boldsymbol{q^2}$	$\boldsymbol{r^2_{\text{pred}}}$
1	**ACE**	0.965	8	0.709	0.523
2	**AchE**	0.879	6	0.525	0.525
3	**BZR**	0.656	4	0.404	0.046
4	**GPB**	0.967	8	0.466	0.246
5	**COX2**	0.710	7	0.432	0.072
6	**DHFR**	0.757	4	0.656	0.569
7	**THERM**	0.951	7	0.552	0.565
8	**THR**	0.846	4	0.574	0.662
9	**ATA**	0.771	4	0.306	-1.402
10	**AT2**	0.859	3	0.191	NA
11	**CCR5**	0.932	4	0.792	-0.302
12	**YOPH**	0.979	4	0.772	0.933
13	**KOA**	0.967	7	0.753	0.660
14	**MX**	0.949	6	0.772	NA
15	**DAT**	0.997	8	0.290	-4.323
16	**TP2A**	0.856	2	0.619	NA
17	**CBRA**	0.920	2	0.615	NA
18	**AI**	0.761	3	0.497	NA
19	**HIVPR**	0.975	8	0.523	0.497
20	**GSK3B**	0.952	8	0.736	0.266
21	**STEROIDS**	0.961	3	0.704	NA
22	**GHS**	0.463	1	0.323	NA
23	**D2R**	0.977	7	0.759	0.420
24	**D4R**	0.756	3	0.522	-0.134
25	**DIAZEPAM_DS_DI**	0.833	3	0.576	NA

26	**DIAZEPAM_DI**	0.967	8	0.421	NA
27	**DIAZEPAM_DS**	0.967	8	0.421	NA
28	**THR**	0.888	5	0.696	0.416
29	**TRY**	0.747	3	0.548	0.655
30	**FXA**	0.874	6	0.437	-0.160

The models' q^2s values were then compared with those reported in the datasets' corresponding original articles (Table 3). Interestingly, although expected, a good overlap between the orginal CoMFA and herein Py-CoMFA code q^2s were observed. In general, q^2s discrepancies reported in Table 3 can be mainly ascribed to differences in grid definitions and in numerical approximations. The dimension of the grid was not possible to be replicated due to lack of information from many original articles. Whereas for the numerical approximation the original CoMFA was written in C language and mainly run on SGI IRIX running workstations using RISC CPUs, while Py-CoMFA rely on Python code using partly C encoded libraries and was run on a CISC CPU. Nevertheless an absolute average discrepancy of 12.8% ± 15.2 was recorded for all 30 datasets. Eleven dataset showed an absolute difference higher than 10% while all the others displayed an absolute difference of only 3.3%. Fourteen out of 30 Py-CoMFA model reported q^2 values higher than those reported in the original articles.

Table 3. Comparison of Py CoMFA models' q^2s data with those reported in the literature for the benchmark datasets. The reported models were obtained with C.3 atom probe with a +1 charge using the combination of steric and electrostatic fields (STE+ELE).

#	Dataset	Published CoMFA		Py-CoMFA[a]	
		q^2	ONC[d]	q^2	ONC
1	**ACE**	0.68[a]	3[58]	0.71	8
2	**AchE**	0.52[a]	5[58]	0.53	6
3	**BZR**	0.32[a]	3[58]	0.40	4
4	**GPB**	0.42[a]	4[58]	0.47	8
5	**COX2**	0.49[a]	5[58]	0.43	7

6	**DHFR**	0.65[a]	5[58]	0.66	4
7	**THERM**	0.52[a]	4[58]	0.55	7
8	**THR-1**	0.59[a]	4[58]	0.57	4
9	**ATA**	0.49[b]	8[65]	0.31	4
10	**AT2**	0.48[a]	5[20]	0.19	3
11	**CCR5**	0.79[a]	3[66]	0.79	4
12	**YOPH**	0.73[a]	3[67]	0.77	4
13	**KOA**	0.69[a]	4[68]	0.75	7
14	**MX**	0.78[a]	5[69]	0.77	6
15	**DAT**	0.29[a]	6[70]	0.29	8
16	**TP2A**	0.61[a]	3[71]	0.62	2
17	**CBRA**	0.57[a]	2[72]	0.62	2
18	**AI**	0.61[a]	3[73]	0.50	3
19	**HIVPR**	0.52[a]	6[28]	0.52	8
20	**GSK3B**	0.78[a]	7[21]	0.74	8
21	**STEROIDS**	0.68[a]	4[14]	0.70	3
22	**GHS**	0.41[c]	NA[75]	0.32	1
23	**D2R**	0.75[a]	3[24]	0.76	7
24	**D4R**	0.49[a]	2[24]	0.52	3
25	**DIAZEPAM DI/DS**	0.79[a]	6[79]	0.58	3
26	**DIAZEPAM DI**	0.70[a]	7[79]	0.42	8
27	**DIAZEPAM DS**	0.73[a]	11[79]	0.42	8
28	**THR**	0.69[a]	4[80]	0.70	5
29	**TRY**	0.63[a]	5[80]	0.55	3
30	**FXA**	0.38[a]	3[80]	0.44	6

NA: Not Available; [a)] q^2 obtained with LOO;[b)] q^2 obtained with L10%O; [c)] q^2 obtained with L30%O q^2 value as a mean of reported from Wang et al. [77], [d)] ONC: Optimal Numeber of Components.

Conclusion

A python implementation of CoMFA as embedded in the www.3d-qsar.com portal proved to be effective in building 3-D QSAR models and in predicting external test sets molecules' activity, as well as the original commercial software. Aside the great advantages to have a free implementation of a 3-D QSAR available for anyone, www.3d-qsar.com can be run from any electronic device able to run a web browser. Herein it has been focused on the computational aspect of the Py-CoMFA module demonstrating that having prealigned datasets it feasible to build 3-D QSAR models. In a future report the portal will be assessed for giving the possibility to build 3-D QSAR models from scratch and also to run feature selection to optimize goodness of fit, robustness and predictive ability of initial models.

In conclusion www.3d-qsar.com represent a valid service to help not informatics skilled researcher in the design of new compounds to prioritize the ones that will be most likely biologically active, enabling significant cost benefits and time savings. Furthermore, www.3d-qsar.com can also be used as a didactic tool to teach 3-D QSAR at any school level, from high school to PhD students.

Supplemetary Material

A zip compressed file containing all the original datasets with biological activities. These data will enable the user to reproduce through www.3d-qsar.com the results herein presented.

Table 1S reporting the r^2, q^2 and r^2_{pred} for the 30 datasets at grid steps of 1 and 2 Å.

References

1. Reker, D.; Schneider, G., Active-learning strategies in computer-assisted drug discovery. *Drug Discov Today* **2015**, 20, 458-65.

2. Cohen, J., *Applied Multiple Regression/correlation Analysis for the Behavioral Sciences*. Taylor & Francis: 2003.

3. Pearson, K., LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6* **1901**, 2, 559-572.

4. H, M. J. R., *OR* **1958**, 9, 63-65.

5. Hotelling, H., THE RELATIONS OF THE NEWER MULTIVARIATE STATISTICAL METHODS TO FACTOR ANALYSIS. *British Journal of Statistical Psychology* **1957**, 10, 69-79.

6. 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.

7. Dearden John, C., The History and Development of Quantitative Structure-Activity Relationships (QSARs). *International Journal of Quantitative Structure-Property Relationships (IJQSPR)* **2016**, 1, 1-44.

8. Free, S. M., Jr.; Wilson, J. W., A Mathematical Contribution to Structure-Activity Studies. *Journal of medicinal chemistry* **1964**, 7, 395-9.

9. Hansch, C.; Fujita, T., p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J Am Chem Soc* **1964**, 86, 1616-1626.

10. Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, II; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A., QSAR modeling: where have you been? Where are you going to? *Journal of medicinal chemistry* **2014**, 57, 4977-5010.

11. Stanton, D. T., On the physical interpretation of QSAR models. *J Chem Inf Comput Sci* **2003**, 43, 1423-33.

12. Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C., Applications of artificial intelligence for chemical inference. 37. GENOA: a computer program for structure elucidation utilizing overlapping and alternative substructures. *The Journal of Organic Chemistry* **1981**, 46, 1708-1718.

13. M, W.; RD, C.; D, S.; I, E. Progress in three-dimensional drug design: the use of real time color graphics and computer postulation of bioactive molecules in DYLOMMS. In *Quantitative approaches to drug design*, JC, D., Ed.; Elsevier: Amsterdam, 1983, pp 145–146.

14. 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.

15. Goodford, P. J., A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of medicinal chemistry* **1985**, 28, 849-57.

16. Artese, A.; Cross, S.; Costa, G.; Distinto, S.; Parrotta, L.; Alcaro, S.; Ortuso, F.; Cruciani, G., Molecular interaction fields in drug discovery: recent advances and future perspectives. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, 3, 594-613.

17. Jones, J. E.; Chapman, S., On the determination of molecular fields. —II. From the equation of state of a gas. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **1924**, 106, 463-477.

18. Cruciani, G., *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*. 2006; Vol. 27, p 1-303.

19. Merz, K. M.; Ringe, D.; Reynolds, C. H., *Drug Design: Structure- and Ligand-Based Approaches*. Cambridge University Press: 2010.

20. Belvisi, L.; Bravi, G.; Catalano, G.; Mabilia, M.; Salimbeni, A.; Scolastico, C., A 3D QSAR CoMFA study of non-peptide angiotensin II receptor antagonists. *J Comput Aided Mol Des* **1996**, 10, 567-82.

21. Zhang, N.; Jiang, Y.; Zou, J.; Zhang, B.; Jin, H.; Wang, Y.; Yu, Q., 3D QSAR for GSK-3beta inhibition by indirubin analogues. *Eur J Med Chem* **2006**, 41, 373-8.

22. Kubinyi, H.; Folkers, G.; Martin, Y. C., *3D QSAR in drug design*. Kluwer Academic: Dordrecht ; Boston, Mass, 1998; p v. < 2- >.

23. Kellogg, G. E.; Semus, S. F., 3D QSAR in modern drug design. *EXS* **2003**, 223-41.

24. Bostrom, J.; Bohm, M.; Gundertofte, K.; Klebe, G., A 3D QSAR study on a set of dopamine D4 receptor antagonists. *J Chem Inf Comput Sci* **2003**, 43, 1020-7.

25. Martin, Y. C., 3D QSAR: Current state, scope, and limitations. *Perspect Drug Discov* **1998**, 12, 3-23.

26. Jewell, N. E.; Turner, D. B.; Willett, P.; Sexton, G. J., Automatic generation of alignments for 3D QSAR analyses. *Journal of Molecular Graphics and Modelling* **2001**, 20, 111-121.

27. Coats, E. A. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. In *3D QSAR in Drug Design: Recent Advances*, Kubinyi, H.; Folkers, G.; Martin, Y. C., Eds.; Springer Netherlands: Dordrecht, 1998, pp 199-213.

28. Tervo, A. J.; Nyronen, T. H.; Ronkko, T.; Poso, A., Comparing the quality and predictiveness between 3D QSAR models obtained from manual and automated alignment. *J Chem Inf Comput Sci* **2004**, 44, 807-16.

29. Kubinyi, H., QSAR and 3D QSAR in drug design .1. methodology. *Drug Discovery Today* **1997**, 2, 457-467.

30. Kubinyi, H., QSAR and 3D QSAR in drug design .2. Applications and problems. *Drug Discovery Today* **1997**, 2, 538-546.

31. Wildman, S. A.; Crippen, G. M., Validation of DAPPER for 3D QSAR: conformational search and chirality metric. *J Chem Inf Comput Sci* **2003**, 43, 629-36.

32. Tropsha, A., Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, 29, 476-488.

33. Topliss, J. G.; Costello, R. J., Chance correlations in structure-activity studies using multiple regression analysis. *Journal of medicinal chemistry* **1972**, 15, 1066-1068.

34. Topliss, J. G.; Edwards, R. P., Chance factors in studies of quantitative structure-activity relationships. *Journal of medicinal chemistry* **1979**, 22, 1238-1244.

35. Clark, M.; Cramer, R. D., The Probability of Chance Correlation Using Partial Least-Squares (Pls). *Quant Struct-Act Rel* **1993**, 12, 137-145.

36. Kohavi, R., A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. **2001**, 14.

37. Xu, Y.; Goodacre, R., On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J Anal Test* **2018**, 2, 249-262.

38. Klebe, G.; Abraham, U.; Mietzner, T., Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *Journal of medicinal chemistry* **1994**, 37, 4130-4146.

39. 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. *Journal of medicinal chemistry* **1994**, 37, 2589-2601.

40. Ragno, R.; Simeoni, S.; Valente, S.; Massa, S.; Mai, A., 3-D QSAR Studies on Histone Deacetylase Inhibitors. A GOLPE/GRID Approach on Different Series of Compounds. *Journal of chemical information and modeling* **2006**, 46, 1420-1430.

41. Tosco, P.; Balle, T., Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields. *Journal of molecular modeling* **2011**, 17, 201-8.

42. Akamatsu, M., Current state and perspectives of 3D-QSAR. *Current topics in medicinal chemistry* **2002**, 2, 1381-94.

43. Mor, M.; Rivara, S.; Lodola, A.; Lorenzi, S.; Bordi, F.; Plazzi, P. V.; Spadoni, G.; Bedini, A.; Duranti, A.; Tontini, A.; Tarzia, G., Application of 3D-QSAR in the rational design of receptor ligands and enzyme inhibitors. *Chem Biodivers* **2005**, 2, 1438-51.

44. Verma, J.; Khedkar, V. M.; Coutinho, E. C., 3D-QSAR in drug design--a review. *Current topics in medicinal chemistry* **2010**, 10, 95-115.

45. Ballante, F.; Ragno, R., 3-D QSAutogrid/R: an alternative procedure to build 3-D QSAR models. Methodologies and applications. *Journal of chemical information and modeling* **2012**, 52, 1674-85.

46. Perkel, J. M., Programming: Pick up Python. *Nature* **2015**, 518, 125-6.

47. Pedregosa, F.; Ga; #235; Varoquaux, l.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; #201; Duchesnay, d., Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, 12, 2825-2830.

48. Rucker, C.; Rucker, G.; Meringer, M., y-Randomization and its variants in QSPR/QSAR. *Journal of chemical information and modeling* **2007**, 47, 2345-57.

49. Schrodinger, LLC, In; 2010.

50. Chimera UCSF homepage. http://www.cgl.ucsf.edu/chimera

51. Murray-Rust, P.; Rzepa, H. S.; Williamson, M. J.; Willighagen, E. L., Chemical markup, XML, and the World Wide Web. 5. Applications of chemical metadata in RSS aggregators. *J Chem Inf Comput Sci* **2004**, 44, 462-9.

52. Herráez, A., Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education* **2006**, 34, 255-261.

53. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

54. Hanson, R. M.; Prilusky, J.; Renjian, Z.; Nakane, T.; Sussman, J. L., JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. *Israel Journal of Chemistry* **2013**, 53, 207-216.

55. Cherkasov, A.; Ban, F.; Santos-Filho, O.; Thorsteinson, N.; Fallahi, M.; Hammond, G. L., An updated steroid benchmark set and its application in the discovery of novel nanomolar ligands of sex hormone-binding globulin. *Journal of medicinal chemistry* **2008**, 51, 2047-56.

56. Polanski, J.; Gieleciak, R.; Magdziarz, T.; Bak, A., GRID formalism for the comparative molecular surface analysis: application to the CoMFA benchmark steroids, azo dyes, and HEPT derivatives. *J Chem Inf Comput Sci* **2004**, 44, 1423-35.

57. Depriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R., 3d-Qsar of Angiotensin-Converting Enzyme and Thermolysin Inhibitors - a Comparison of Comfa Models Based on Deduced and Experimentally Determined Active-Site Geometries. *J Am Chem Soc* **1993**, 115, 5372-5384.

58. Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F., A comparison of methods for modeling quantitative structure-activity relationships. *Journal of medicinal chemistry* **2004**, 47, 5541-54.

59. Golbraikh, A.; Bernard, P.; Chretien, J. R., Validation of protein-based alignment in 3D quantitative structure-activity relationships with CoMFA models. *Eur J Med Chem* **2000**, 35, 123-136.

60. Maddalena, D. J.; Johnston, G. A. R., Prediction of Receptor Properties and Binding-Affinity of Ligands to Benzodiazepine/Gaba(a) Receptors Using Artificial Neural Networks. *Journal of medicinal chemistry* **1995**, 38, 715-724.

61. Gohlke, H.; Klebe, G., DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. *Journal of medicinal chemistry* **2002**, 45, 4153-4170.

62. Chavatte, P.; Yous, S.; Marot, C.; Baurin, N.; Lesieur, D., Three-dimensional quantitative structure-activity relationships of cyclo-oxygenase-2 (COX-2) inhibitors: a comparative molecular field analysis. *Journal of medicinal chemistry* **2001**, 44, 3223-30.

63. Sutherland, J. J.; Weaver, D. F., Three-dimensional quantitative structure-activity and structure-selectivity relationships of dihydrofolate reductase inhibitors. *J Comput Aid Mol Des* **2004**, 18, 309-331.

64. Klebe, G.; Abraham, U.; Mietzner, T., Molecular Similarity Indexes in a Comparative-Analysis (Comsia) of Drug Molecules to Correlate and Predict Their Biological-Activity. *Journal of medicinal chemistry* **1994**, 37, 4130-4146.

65. Nayyar, A.; Malde, A.; Jain, R.; Coutinho, E., 3D-QSAR study of ring-substituted quinoline class of anti-tuberculosis agents. *Bioorganic & medicinal chemistry* **2006**, 14, 847-56.

66. Aher, Y. D.; Agrawal, A.; Bharatam, P. V.; Garg, P., 3D-QSAR studies of substituted 1-(3, 3-diphenylpropyl)-piperidinyl amides and ureas as CCR5 receptor antagonists. *Journal of molecular modeling* **2007**, 13, 519-29.

67. Hu, X.; Stebbins, C. E., Molecular docking and 3D-QSAR studies of Yersinia protein tyrosine phosphatase YopH inhibitors. *Bioorganic & medicinal chemistry* **2005**, 13, 1101-9.

68. Li, W.; Tang, Y.; Zheng, Y. L.; Qiu, Z. B., Molecular modeling and 3D-QSAR studies of indolomorphinan derivatives as kappa opioid antagonists. *Bioorganic & medicinal chemistry* **2006**, 14, 601-10.

69. Bang, S. J.; Cho, S. J., Comparative molecular field analysis (CoMFA) and comparative molecular similarity index analysis (CoMSIA) study of mutagen X. *B Kor Chem Soc* **2004**, 25, 1525-1530.

70. Yuan, H. B.; Kozikowski, A. P.; Petukhov, P. A., CoMFA study of piperidine analogues of cocaine at the dopamine transporter: Exploring the binding mode of the 3 alpha-substituent of the piperidine ring using pharmacophore-based flexible alignment. *Journal of medicinal chemistry* **2004**, 47, 6137-6143.

71. Jensen, L. H.; Liang, H.; Shoemaker, R.; Grauslund, M.; Sehested, M.; Hasinoff, B. B., A three-dimensional quantitative structure-activity relationship study of the inhibition of the ATPase activity and the strand passing catalytic activity of topoisomerase II alpha by substituted purine analogs. *Mol Pharmacol* **2006**, 70, 1503-1513.

72. Salo, O. M.; Savinainen, J. R.; Parkkari, T.; Nevalainen, T.; Lahtela-Kakkonen, M.; Gynther, J.; Laitinen, J. T.; Jarvinen, T.; Poso, A., 3D-QSAR studies on cannabinoid CB1 receptor agonists: G-protein activation as biological data. *Journal of medicinal chemistry* **2006**, 49, 554-66.

73. Sulea, T.; Oprea, T. I.; Muresan, S.; Chan, S. L., A different method for steric field evaluation in CoMFA improves model robustness. *J Chem Inf Comp Sci* **1997**, 37, 1162-1170.

74. Oprea, T. I.; Garcia, A. E., Three-dimensional quantitative structure-activity relationships of steroid aromatase inhibitors. *J Comput Aided Mol Des* **1996**, 10, 186-200.

75. Mittal, R. R.; Harris, L.; McKinnon, R. A.; Sorich, M. J., Partial charge calculation method affects CoMFA QSAR prediction accuracy. *Journal of chemical information and modeling* **2009**, 49, 704-9.

76. Polychronopoulos, P.; Magiatis, P.; Skaltsounis, A. L.; Myrianthopoulos, V.; Mikros, E.; Tarricone, A.; Musacchio, A.; Roe, S. M.; Pearl, L.; Leost, M.; Greengard, P.; Meijer, L., Structural basis for the synthesis of indirubins as potent and selective inhibitors of glycogen synthase kinase-3 and cyclin-dependent kinases. *Journal of medicinal chemistry* **2004**, 47, 935-46.

77. Wang, R. X.; Gao, Y.; Liu, L.; Lai, L. H., All-orientation search and all-placement search in comparative molecular field analysis. *Journal of molecular modeling* **1998**, 4, 276-283.

78. Melville, J. L.; Hirst, J. D., On the stability of CoMFA models. *J Chem Inf Comput Sci* **2004**, 44, 1294-300.

79. Wong, G.; Koehler, K. F.; Skolnick, P.; Gu, Z. Q.; Ananthan, S.; Schonholzer, P.; Hunkeler, W.; Zhang, W.; Cook, J. M., Synthetic and computer-assisted analysis of the structural requirements for selective, high-affinity ligand binding to diazepam-insensitive benzodiazepine receptors. *Journal of medicinal chemistry* **1993**, 36, 1820-30.

80. Bohm, M.; Sturzebecher, J.; Klebe, G., Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *Journal of medicinal chemistry* **1999**, 42, 458-477.