How the energy evaluation method used in the geometry optimization step affect the quality of the subsequent QSAR/QSPR models

Åsmund Rinnan · Niels Johan Christensen · Søren Balling Engelsen

Abstract The quantitative influence of the choice of energy evaluation method used in the geometry optimization step prior to the calculation of molecular descriptors in QSAR and QSPR models was investigated. A total of 11 energy evaluation methods on three molecular datasets (toxicological compounds, aromatic compounds and PPARγ agonists) were studied. The methods employed were: MMFF94 s, MM3* with ε_r (relative dielectric constant) $= 1$, MM3* with $\varepsilon_r = 80$, AM1, PM3, HF/STO-3G, HF/6-31G, HF/6-31G(d,p), B3LYP/STO-3G, B3LYP/6-31G, and B3LYP/6-31G(d,p). The 3D-descriptors used in the QSAR/QSPR models were calculated with commercially available molecular descriptor programs primarily directed toward pharmaceutical research. In order to evaluate the uncertainties involved in the QSAR/QSPR predictions bootstrapping was used to validate all models using 1,000 drawings for each data set. The scale free error-term, q^2, was used to compare the relative quality of the models resulting from different optimization methods on the same set of molecules. Depending on the dataset, the average 0.632 bootstrap estimated q^2 varies from 0.55 to 0.57 for the toxicological compounds, from 0.58 to 0.62 for the aromatic compounds, and from 0.69 to 0.75 for the PPARγ agonists. The B3LYP/6-31G(d,p) provided the best overall results, albeit the increase in q^2 was small in all cases. The results clearly indicate that the choice of the energy evaluation method has very limited impact. This study suggests that QSAR or QSPR studies might benefit from the choice of a rapid optimization method with little or no loss in model accuracy.

Keywords QSAR · QSPR · Energy evaluation · PLS regression · Quantum mechanics · Semi-empirical · Molecular mechanics

Abbreviations

QSPR	Quantitative structure property relationship
QSAR	Quantitative structure activity relationship
MM3*	Allinger's molecular mechanics
AM1	Austin model 1
PM3	Parameterized model 3, HF, Hartree–Fock
B3LYP	The hybrid exchange–correlation functional based on work from Becke, Lee, Yang and Par
PLS	Partial least squares
RMSD	Root mean square distance
MCMM	Monte Carlo multiple minimum

Å. Rinnan (✉) · N. J. Christensen · S. B. Engelsen
Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark
e-mail: aar@life.ku.dk

Introduction

QSAR and QSPR methods are becoming more and more attractive for screening purposes in pharmaceutical [1], toxicological [2], nutritional sciences [3] and health sciences [4]. In QSPR and QSAR a wide variety of computational methods for calculating the potential energy surface of the molecules, and thus affecting the geometry optimization of the molecular structures, may be chosen prior to the calculation of molecular descriptors and the subsequent regression step. The computational time increases manifold when

going from the more pragmatic to the more sophisticated energy evaluation methods. For typical small drug-sized molecules, molecular mechanics methods can generate an optimized structure in seconds while quantum mechanical methods may require hours or days on present-day computers. Considering that QSPR and QSAR applications often involve datasets with tens, hundreds or even thousands of molecules, the choice of energy evaluation method becomes a real concern when contrasted with the demands for rapid development. While accurate results in computational chemistry often necessitate calculations at a high level of theory, it has also been demonstrated that the geometry obtained using energy evaluations at the highest level of theory does not always lead to the best results in subsequent calculations of molecular properties [5]. This suggests that any presumption about the relationship of the energy evaluation method and the quality of a particular type of results should be carefully examined. Nevertheless, only few attempts have been made to elucidate how the choice of energy evaluation method affects QSAR and QSPR models [6–8]. The most commonly employed energy evaluation methods in QSAR/QSPR studies are AM1 [9], PM3 [10], and HF [11], and B3LYP [12–15] with the 6-31G(d) basis set [16–20]. However, rarely is any rationale given for using a particular method and almost never are several methods compared within the same study. In a surprisingly high percentage (about 25%) of the published works in literature, any description of the energy evaluation method used during geometry optimization is absent. Such widespread omission of information suggests that the matter in question is either firmly established or irrelevant. In this study the influence of the choice of energy evaluation method used in the geometry optimization step upon the final QSAR/QSPR prediction model is investigated for three diverse datasets selected from the literature. A series of QSAR/QSPR models for the prediction of the measured quantity were built using structures from each of the energy evaluation schemes and the relative quality of these models was compared using q^2.

The authors have chosen to use the phrase "energy evaluation" regarding the methods investigated in this manuscript as we do not investigate the effect of energy minimization algorithms such as conjugate gradients [21], BFGS, DFP [22] or quasi-Newton [23] in conjunction with an energy evaluation method such as B3LYP [12–15], HF [11] or PM3 [10]. The conjunction between the two could be referred to as geometry optimization.

Methods

Figure 1 shows how the work of this paper was designed. The grey circle with the molecule depicts the geometry optimization step wherein the energy evaluation method used was changed. These optimizations are then used for two purposes: to calculate the difference in the geometries (RMSD) and to make prediction models. The black squares to the right of the molecules are the data-matrices used for the subsequent prediction models. A total of 1,000 bootstrap drawings were performed for each energy evaluation method and the samples left out by the bootstrap drawing were used as the validation set. The q^2-values of these validation sets where subsequently used in order to compare the importance of the choice of the energy evaluation method. This whole process was repeated for each of the energy evaluation methods, i.e. 11 times in total. The number and names of both the compounds and the descriptors were kept constant for all repetitions in order to be able to compare the effect of the energy evaluation methods on the subsequent regression model. Further description of each of the steps follows below.

Datasets

This study investigates three datasets already published and discussed in earlier papers, all showing good predictive power. Although the datasets used in this study are somewhat reduced from these original studies, the subsets of molecules should still lead to acceptable models.

The three sample sets are diverse: 290 compounds exhibiting acute aquatic toxicity in fathead minnow [24]—named toxicological compounds, 80 compounds with various degrees of penetration of a polydimethylsiloxane membrane [25]—named aromatic compounds, and 12 PPARγ agonists [26] with different pK_i-values for PPARγ. Further details about these datasets are given in Supplementary Material—Part I.

The last dataset contains 12 molecules, but this low number is justified by the goal of this project: investigating the perturbation effect the energy evaluation method has on the subsequent regression analysis,

Geometry optimization

All structures were built with GaussView [27] and subjected to 5,000 steps of MCMM conformational searching in MacroModel [28] using the MMFF94 s force-field [29] and the PRCG optimization algorithm [30]. Only the lowest conformer was selected for further use in the study. The conformational search was included to remove human bias from the initially generated structures.

The lowest conformer of each molecule was geometry optimized further using 10 energy evaluation methods: MM3* [31] with $\varepsilon_r = 1$, MM3* with $\varepsilon_r = 80$, AM1, PM3, HF with STO-3G [32], HF/6-31G, HF/6-31G(d,p), B3LYP/ STO-3G, B3LYP/6-31G, and B3LYP/6-31G(d,p). For the

Fig. 1 Design of the analysis
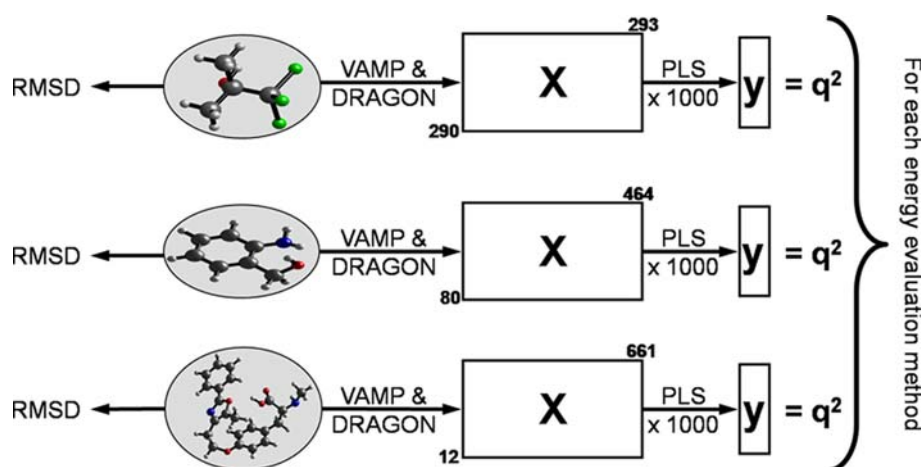
MM3* method the PRCG optimization algorithm were used, while the Berny optimization algorithm [33] was used for the remaining methods. It is assumed that the choice of the optimization algorithm is insignificant compared to the choice of energy evaluation method. Thus a total of 11 energy evaluation methods were evaluated. In the remainder of the text we use the shorthand notations MM3*-1 and MM3*-80 for MM3* with $\varepsilon_r = 1$ and MM3* with $\varepsilon_r = 80$, respectively.

Prediction models

VAMP [34] was employed to calculate the total energy, electronic energy, nuclear energy, surface area, mean polarizability, heat of formation, ionization potential, LUMO/HOMO energies, total dipole, and partial charges for all molecules at the AM1 level. DRAGON [35] provided additional 3D descriptors in the categories: Randic molecular profiles, geometrical descriptors, RDF descriptors, 3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction) descriptors [36], WHIM (Weighted Holistic Invariant Molecular descriptors) descriptors [37], GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptors [38], and charge descriptors (using VAMP charges). Thus all the descriptors used in this study are geometry sensitive descriptors, capturing variations due to the choice of energy evaluation method. Descriptors with no variance in at least one of the molecules of the data set across the optimization methods studied were excluded, in order to focus the study on the geometry sensitive descriptors (the toxicological compounds thus included 293 variables, the aromatic compounds 464 variables and PPARγ agonists included 661 variables, as also can be seen from Fig. 1). This was furthermore done because the prediction models including all 681 descriptors are in general worse and less consistent. Furthermore, our focus is not on how to make the best PLS-model, but rather on the differences between comparable

PLS-models. Auto-scaling was used in order to give all descriptors the same chance of influencing the model. The prediction models were all PLS-models [39]. In order to estimate the uncertainty of the models, 1,000 bootstrap [40] drawings were performed for each dataset (the number of samples in each bootstrap drawing equals the number of samples in the original dataset under investigation). The same set of bootstrap drawings were performed on all energy evaluation methods, ensuring that the selection of which samples goes in the calibration or the validation set does not influence the quality of the final model. The resulting RMSE-values were used to calculate the 0.632 bootstrap estimates of the mean error and of the standard error of the error. The number 0.632 is approximately the probability that any one sample is in a bootstrap drawing [40]. Bootstrapping was performed instead of simple cross-validation, as bootstrapping also provides an estimate of the uncertainty of the prediction error [40, 41], which leads to options to test for significance between two sets of datasets (two different energy evaluation methods in this particular case). The 0.632 bootstrap RMSE estimates were transformed into q^2-values by the formula shown in Equation 1.

$$q^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} = 1 - \left(\frac{\text{RMSE}}{s_y}\right)^2 \frac{n-1}{n} \qquad (1)$$

Here y is the reference value, $\hat{y}$ is the predicted value, $\bar{y}$ is the mean reference value, s_y is the standard deviation of the reference values and n is the number of samples in the dataset. All the numbers in the equation refers to the samples in the validation set, i.e. different for each bootstrap drawing, similar to formula 2 in the work by Consonni et al. [42].

The difference in the 0.632 bootstrap estimate of the q^2 was tested using a Games-Howell comparison of mean test [43]. This test was performed on the basis of the 0.632 bootstrap estimates of the q^2 values and the uncertainty of these, i.e. using the results from the 1,000 bootstrap

drawings. The *p*-values were corrected according the Bonferroni correction, stating that $p_{corr} = 1 - p^{1/k}$, where p is the normal significance value (e.g. 0.05), k is the number of comparisons and p_{corr} is the corrected significance value. For a comparison of n different groups $k = \frac{n(n-1)}{2}$.

Variations in geometry

To quantify geometrical differences, the optimized structures of all methods were superimposed on the structure obtained at the highest level of theory (B3LYP/6-31G(d,p)) using all heavy atoms. The RMS distance (RMSD) was calculated for the superposition between equivalent atoms excluding hydrogens, as several of the hydrogens have rotational freedom, for example in a methyl group. The RMSD is defined as shown in Equation 2.

$$RMSD = \sqrt{\sum_{i=1}^{N} \frac{(r_{i,1} - r_{i,2})^2}{N}}. \tag{2}$$

Here $r_{i,1}$ and $r_{i,2}$ are the positions of atom i in structure 1 and 2, respectively and N is the total number of atoms in the molecule. Furthermore, the number of close contacts were counted; defined as two atoms joined through at least three covalent bonds and separated by less than 75% of the sum of their van der Waals radii. Hydrogen bonds were counted using the following criteria: The H—A-R angle had to be larger than 90 °, the H-A distance had to be smaller than 2.5 Å, and the D-H—A angle had to be larger than 120 ° (H = Hydrogen, A = Acceptor, D = Donor, R = Remainder of molecule).

Software

Molecular structures were created with GaussView [27]. The MM3* optimizations were carried out with Macro-Model [28]. All other optimizations were performed with Gaussian 03 [44]. RMSD values were calculated with Maestro [28]. The prediction models were made in Matlab 7.6 [45] with in-house functions.

Results and discussion

All molecules in this study were geometry optimized in the isolated state neglecting intermolecular interactions which may induce geometry changes important for QSAR/QSPR models.

Influence of energy evaluation method on geometry

For the toxicological compounds, 42 (about 1%) of the optimized structures had close contacts. These structures were all produced with PM3, HF/STO-3G, or B3LYP/STO-3G. PM3 and B3LYP/STO-3G gave several bad contacts for the same compound. Ten compounds in this dataset allow for internal hydrogen bonds. All methods reproduced the hydrogen bonds, except in case of MM3*-80 for four structures. The reluctant capability of creating hydrogen bonds in the latter case is a natural consequence of large separation between hydrogen bond donors and acceptors for the five compounds and high relative dielectric constant used for MM3*.

For the aromatic compounds, a close contact was found in 13 structures (11 different compounds) optimized with B3LYP/STO-3G, HF/STO-3G or PM3. All optimization methods reproduced intramolecular hydrogen bonds in the four structures where they were expected. B3LYP/STO-3G and HF/STO-3G, and MM3*-1 produced a false hydrogen bond in one structure.

In the PPARγ agonists a single close contact was found for 13 optimized structures, five structures had two close contacts, and 3 close contacts were found for one compound. The close contacts were in all cases produced with PM3 or B3LYP/STO-3G. One structure is expected to have two hydrogen bonds which were reproduced with most methods, except AM1 and the force field methods, which produce only one hydrogen bond.

A simple comparison of the three datasets indicates that the toxicological compounds and the aromatic compounds generally have low (max median 0.06 Å) and comparable RMSD values, see Fig. 2. The RMSDs (max median 1.3 Å) for the PPARγ agonists are larger reflecting the increased flexibility of the larger molecules. A more detailed examination reveals that the median RMSD values for the toxicological compounds and the aromatic compounds follow the same behavior for DFT and Hartree–Fock methods, although median RMSD values are slightly but consistently lower for the toxicological compounds. For both datasets the lowest median RMSDs are produced with HF/6-31G(d,p), but the difference between this method and B3LYP/6-31G or HF/STO-3G is marginal. B3LYP/STO-3G gives relatively large RMSDs, as evident from medians and 95-percentile values. Visual inspection of the optimized geometries reveals that the energy evaluation methods do not in general introduce any prominent structural distortions when applied to the toxicological compounds or the aromatic compounds. Comparison of the 95-percentiles for these two datasets (Fig. 2a, b) shows that the evaluation methods give roughly the same pattern of relative deviations from B3LYP/6-31G(d,p). The significant flexibility of the molecules of the PPARγ agonists is reflected in the generally high values of RMSD medians (max 1.26 Å) and other statistics. B3LYP/6-31G gives the lowest median, but not the smallest spread in RMSD.
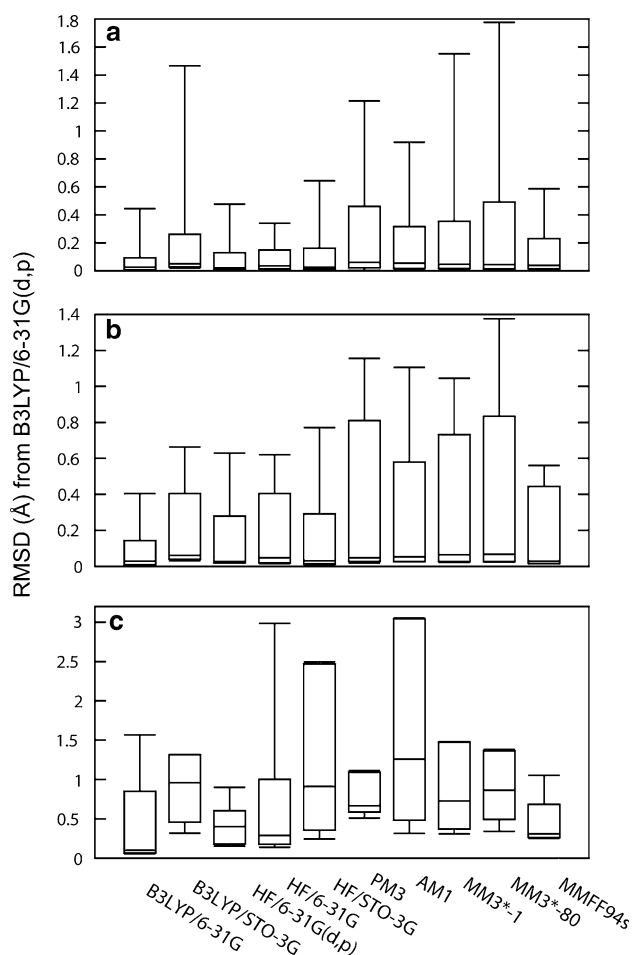
Fig. 2 RMSD values for the three QSAR/QSPR datasets. **a** Toxicological compounds, **b** aromatic compounds, and **c** PPARγ agonists. Minimum and maximum values are indicated by *whiskers*. The *boxes* are vertically limited by the 5 and 95 percentile in (**a**) and (**b**) and by the second lowest and second highest value in (**c**). Median values are indicated with *black bars* in the interior of the *boxes*

Prediction results

The datasets in this study give mediocre models for the toxicological compounds and aromatic compounds ($q^2 \in [0.55, 0.57]$ and $q^2 \in [0.58, 0.62]$ respectively), and satisfactory models for the PPARγ agonists ($q^2 \in [0.69, 0.75]$). The variation in the q^2-values is largest for the PPARγ agonists, and smallest for the toxicological compounds, which is consistent with the RMSD values calculated above, indicating that the differences in the prediction models are due to differences in 3D structure of the molecules. Figure 3 shows the % deviation in the mean q^2 of the 11 energy evaluation methods compared to the best method (see Eq. 3), including the standard error of the q^2. The smallest average deviation is found for the most expensive method—B3LYP/6-31G(d,p). Furthermore, in two of the three datasets this method either has the highest

q^2 or there is no significant difference ($p > 0.05$) between B3LYP/6-31G(d,p) and the best model (see Tables 4–6 in Supplementary Material—Part II). For the last dataset (the aromatic compounds) it is only significantly worse than B3LYP/6-31G (best model) and B3LYP/STO-3G ($p < 0.001$), but only by a decrease in q^2 by 0.03 and 0.02 respectively. However, more pragmatic methods such as B3LYP/6-31G, B3LYP/STO-3G, HF/6-31G and even PM3 show similar results. From these observations it becomes evident that the choice of basis set is not so important for the B3LYP method. Taking into account the discussion of the close contacts above, it becomes clear that B3LYP with 6-31G or 6-31G(d,p) are the best choices. From a time perspective B3LYP/6-31G(d,p) is approximately 3, 10, and 100 times more expensive than B3LYP/6-31G, HF/6-31G and PM3 respectively. This study indicates that if a larger QSAR/QSPR screening is to be performed, cheaper methods such as HF/6-31G or PM3 may be employed with success. This is in accordance to the conclusion in the paper by Puzyn et al. [46], where they state that it is better to use the semi-empirical methods PM6 (a new version of PM3) [47] or RM1 (a new version of AM1) [48] instead of the more expensive DFT methods.

$$\text{Deviation} = \frac{q^2_{\text{best}} - q^2_{\text{current}}}{q^2_{\text{best}}} \times 100\%. \tag{3}$$

Conclusions

The influence of the choice of energy evaluation method in the geometry optimization step on the predictive quality of QSAR/QSPR models for three different molecular datasets has been investigated. The lowest energy conformer in each dataset was optimized with 11 different methods and
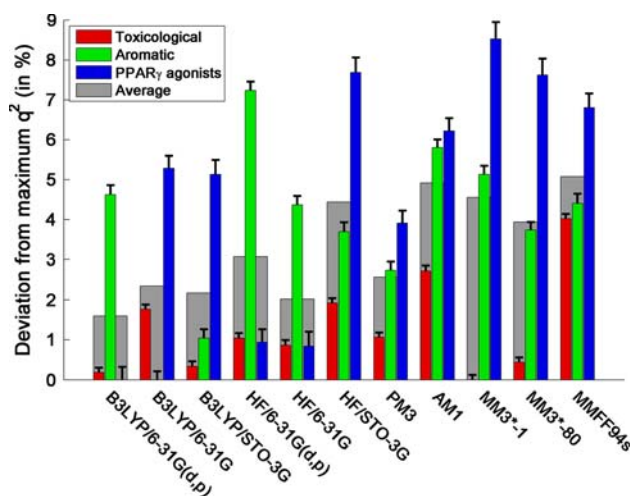

Fig. 3 Deviation from the maximum 0.632 bootstrap estimate of the mean q^2 given in %, with the standard error showed by *whiskers*

subsequently 3D molecular descriptors were calculated with VAMP and DRAGON. The results show that the energy evaluation methods only to a small extent influence the QSAR/QSPR prediction model. The most time consuming method—B3LYP/6-31G(d,p)—is the method which in general gives the best prediction models, albeit the increase in q^2 is rather small. This further suggests that the usage of more pragmatic methods such as HF/6-31G and PM3 can be used, especially in larger screening analyses with little or no loss in model accuracy.

References

1. Agüero-Chapin G, Varona-Santos J, de la Riva GA, Antunes A, Gonzáles-Villa T, Uriarte E, González-Díaz H (2009) J Proteome Res 8:2122–2128
2. Pasha FA, Nez MM, Cho SJ, Ansari M, Mishra SK, Tiwari S (2009) Chem Biol Drug Des 73:537–544
3. Jensen BF, Sørensen MD, Kissmeyer A-M, Björkling F, Sonne K, Engelsen SB, Nørgaard L (2003) J Comput Aid Mol Des 17:849–859
4. Thorsteinson N, Ban F, Santos-Filho O, Tabaei SMH, Miguel-Queralt S, Underhill C, Cherkasov A, Hammond GL (2009) Toxicol Appl Pharm 234:47–57
5. Foresman JB, Frisch A (1996) Exploring chemistry with electronic structure methods, 2nd edn. Gaussian, Pittsburgh, pp 157–158
6. Hudáky I, Hudáky I, Perczel A (2004) J Comput Chem 25:1522–1531
7. Swart M, Snijders JG (2003) Theor Chem Acc 110:34–41
8. Bayari S, Saglam S, Ustundag HF (2005) J Mol Struct 726:225–232
9. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902–3909
10. Stewart JJP (1989) J Comput Chem 10:209–220
11. Roothaan CCJ (1951) Rev Mod Phys 23:69–89
12. Becke AD (1993) J Chem Phys 98:5648–5652
13. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785–789
14. Vosko SH, Wilk L, Nusair M (1980) Can J Phys 58:1200–1211
15. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ (1994) J Phys Chem 98:11623–11627
16. Ditchfield R, Hehre WJ, Pople JA (1971) J Chem Phys 54:724–728
17. Hehre WJ, Ditchfield R, Pople JA (1972) J Chem Phys 56:2257–2261
18. Hariharan PC, Pople JA (1974) Mol Phys 27:209–214
19. Gordon MS (1980) Chem Phys Lett 76:163–168
20. Hariharan PC, Pople JA (1973) Theor Chim Acta 28:213–222
21. Fletcher R, Reeves CM (1964) Computer J 7:149–154
22. Stewart JPP (1989) Comput Chem 13:157–158
23. Dennis JE, More JJ (1977) SIAM Rev 19:46–89
24. He L, Jurs PC (2005) J Mol Graphics Modell 23:503–523
25. Agantovic-Kustrin S, Beresford R, Pauzi A, Yusof M (2001) J Pharm Biomed Anal 26:241–254
26. Rücker C, Scarsi M, Meringer M (2006) Bioorg Med Chem 14:5178–5195
27. Dennington II R, Keith T, Milliam J, Eppinnett K, Hovell WL, Gilliland R (2003) GaussView, Version 3.09, Semichem, Inc., Shawnee Mission. http://www.gaussian.com
28. Schrödinger LLC (2007) http://www.schrodinger.com
29. Halgren TA (1999) J Comput Chem 20:720–729
30. Polak E, Ribiere G (1969) Revenue Francaise Inform Rech Operationelle 16:35
31. Allinger NL, Yuh YH, Lii J-H (1989) J Am Chem Soc 111:8551–8566
32. Hehre WJ, Stewart RF, Pople JA (1969) J Chem Phys 51:2657–2664
33. Peng CY, Ayala PY, Schlegel HB, Frisch MJ (1996) J Comp Chem 17:49–56
34. Vamp 6.5 (1997) Oxford Software Limited
35. Talete srl (2009) DRAGON for Windows (Software for Molecular Descriptor Calculations) http://www.talete.mi.it
36. Schuur JH, Selzer P, Gasteiger J (1996) J Am Chem Soc 36:334–344
37. Todeschini R, Lasagni M (1994) J Chemom 8:263–272
38. Consonni V, Todeschini R, Pavan M, Gramatica P (2002) J Chem Inf Comput Sci 42:693–705
39. Martens H, Martens M, Næs T (2001) Multivariate analysis of quality. John Wiley, Chichester, pp 111–125
40. Wehrens R, Putter H, Buydens LMC (2000) Chemom Intell Lab Syst 54:35–52
41. Efron B, Gong G (1983) Am Stat 37:36–48
42. Consonni V, Ballabio D, Todeschini R (2009) J Chem Inf Model 49:1669–1678
43. Howell JF, Games PA (1974) Brit J Math Stat Psy 27:72–81
44. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Jr VrevenT, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2009) Gaussian 03 revision C02. Gaussian Inc, Wallingford
45. The Mathworks ™ (2009) USA, http://www.mathworks.com
46. Puzyn T, Suzuki N, Haranczyk M, Rak J (2008) J Chem Inf Model 48:1174–1180
47. Rocha GB, Freire RO, Simas AM, Stewart JJP (2006) J Comput Chem 27:1101–1111
48. Stewart JJP (2007) J Mol Model 13:1173–1213