JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

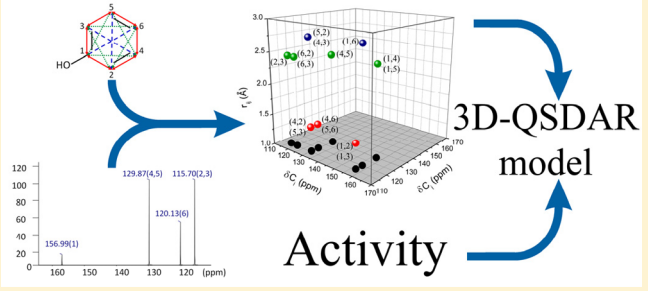
pubs.acs.org/jcim

# $^{13}$C NMR–Distance Matrix Descriptors: Optimal Abstract 3D Space Granularity for Predicting Estrogen Binding

Svetoslav H. Slavov, Elizabeth L. Geesaman, Bruce A. Pearce, Laura K. Schnackenberg, Dan A. Buzatu, Jon G. Wilkes,* and Richard D. Beger*

Division of Systems Biology, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Rd., Jefferson, Arkansas 72079, United States

**ABSTRACT:** An improved three-dimensional quantitative spectral data–activity relationship (3D-QSDAR) methodology was used to build and validate models relating the activity of 130 estrogen receptor binders to specific structural features. In 3D-QSDAR, each compound is represented by a unique fingerprint constructed from $^{13}$C chemical shift pairs and associated interatomic distances. Grids of different granularity can be used to partition the abstract fingerprint space into congruent "bins" for which the optimal size was previously unexplored. For this purpose, the endocrine disruptor knowledge base data were used to generate 50 3D-QSDAR models with bins ranging in size from 2 ppm × 2 ppm × 0.5 Å to 20 ppm × 20 ppm × 2.5 Å, each of which was validated using 100 training/test set partitions. Best average predictivity in terms of $R^2_{test}$ was achieved at 10 ppm ×10 ppm × $Z$ Å ($Z$ = 0.5, ..., 2.5 Å). It was hypothesized that this optimum depends on the chemical shifts' estimation error (±4.13 ppm) and the precision of the calculated interatomic distances. The highest ranked bins from partial least-squares weights were found to be associated with structural features known to be essential for binding to the estrogen receptor.




## INTRODUCTION

Beginning in the 1980s, attempts to reduce the complexity and mitigate the interpretation challenges in QSAR were undertaken by defining restricted sets of universal and easy to interpret descriptors.[1−6] Some of these efforts favored experimentally determined parameters such as aqueous solvation energy, octanol–water partition coefficient, boiling point, molar refractivity, volume and vaporization enthalpy, molar refraction, polarity/polarizability, hydrogen bond acidity and basicity, and infrared and UV spectra.[1,2,7,8]

In a recent review, Verma and Hansch emphasized the importance of the $^{13}$C NMR chemical shifts as parameters in the development of QSARs/QSPRs.[9] Two major advantages were underlined: (i) the spectra are taken in a liquid solution that mimics the conditions in biological systems and (ii) the high sensitivity of the technique allows the observation of minor changes in the molecular conformation. Moreover, it has been demonstrated[10−13] that the $^{13}$C chemical shifts correlate exceptionally well ($R^2$ above 0.99) with various partial atomic charges and the Hammett substituent constants. Hence, it can serve as a measure of the electron density distribution within the molecules. On the basis of the above foundations, many successful quantitative spectrometric data–activity relationships (QSDARs) were reported.[14−25]

An important distinction between the NMR and the infrared (which encodes bond strength related information) or the ultraviolet (which encodes electron conjugation related information) spectral characteristics is that the NMR chemical shifts can be assigned to specific atoms. Thus, they allow for relatively easy interpretation of the models in terms of structural changes. Because the chemical shift ($\delta$) range for $^{13}$C is much larger than that for $^{1}$H NMR, there is much less overlap in peaks, and $^{13}$C NMR, therefore, can reflect slight structural variations with less ambiguity.[26] Furthermore, when compared to $^{1}$H, the $^{13}$C NMR peaks are less affected by solvent effects and hence are often preferred when building QSDAR models.[11,12,16−20] Very few QSDARs utilizing atoms other than C or H have been published.[27,28] However, the addition of spectral data for other biologically important heteroatoms such as $^{15}$N, $^{17}$O, $^{19}$F, and $^{31}$P may also prove beneficial.

One limitation of using only NMR chemical shifts in establishing QSDAR relationships is that they encode information related primarily to the electron density distribution. This may prove sufficient to explain the variance in biological data in cases in which the electrostatic (Coulombic) interactions are dominant. On the other hand, in cases in which the steric (Lennard–Jones) effects are substantial, the addition of interatomic distance related information would be desirable. Therefore, to complete the picture of the intermolecular interactions within the framework of our QSDAR approach, we combined NMR spectral data with interatomic distances.

**Received:** March 30, 2012
**Published:** June 9, 2012

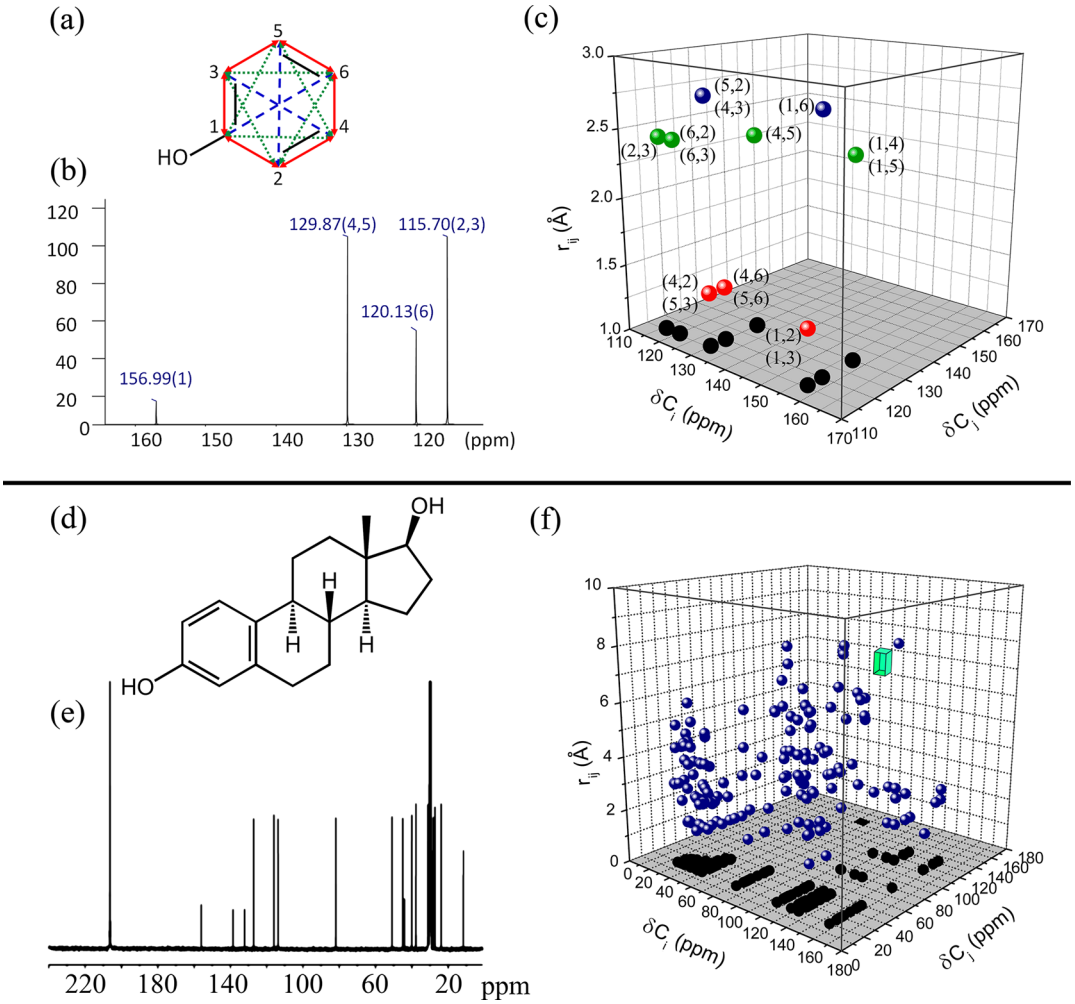



**Figure 1.** (a) Phenol molecule, (b) $^{13}$C NMR spectra of phenol, (c) 3D fingerprint of phenol, (d) structure of 17$\beta$-estradiol, (e) $^{13}$C NMR spectra of 17$\beta$-estradiol, and (f) 3D fingerprint of 17$\beta$-estradiol. The black circles in (c) and (f) represent the shadows of the fingerprint elements in the *XY*-plane. The green square box element in (f) represents a single nonoccupied 10 ppm × 10 ppm × 1.0 Å bin.

Our current work extends on previous efforts[21] to utilize $^{13}$C NMR and interatomic distance data by (i) using a regular 3D grid to partition the 3D-QSDAR space with no assumptions about the importance of specific interatomic ($r_{ij}$) distances; (ii) a partial least-squares (PLS) method, rather than multiple linear regression (MLR), principal component regression (PCR), or artificial neural networks (ANN) was used to establish a correlation with the property; (iii) the ranked weights of the original variables were used to extract relevant information explaining the variance in biological data, and (iv) the frequency of occurrence of the original variables was used to identify structural fragments responsible for the observed biological effect.

## METHODOLOGY

**General 3D-QSDAR Approach.** The theory behind the 1D-QSDAR approach, which uses only NMR spectral data, is described elsewhere.[19,20,28−31] The 3D-QSDAR[21−23,25,32] approach extends 1D-QSDAR by incorporating an atom-to-atom distance matrix that accounts for the shape and size of the molecule (i.e., steric configuration). Within the framework of the 3D-QSDAR approach, all molecules are represented by their 3D spectral distance fingerprints. For a given molecule with a total of *N* carbon atoms, these 3D fingerprints are constructed using the chemical shifts of all nonordered ($C_iC_j \equiv C_jC_i$; $i,j = 1...N$) carbon atom pairs in conjunction with the $\delta C_i \geq \delta C_j$ condition, in which $\delta$ denotes the corresponding chemical shifts. Under these conditions a three-dimensional abstract space with the following orthogonal axes is created: (i) the chemical shift of atom $C_i$ is placed on the *X*-axis; (ii) the chemical shift of atom $C_j$ is placed on the *Y*-axis, and (iii) the distance ($r_{ij}$) between atoms $C_i$ and $C_j$ forms the *Z*-axis. According to the above definition ($C_iC_j \equiv C_jC_i$ and $\delta C_i \geq \delta C_j$), all fingerprints are characterized by a single plane of symmetry $C_s$ intersecting the *XY*-plane through its main diagonal and are invariant to rotation and/or translation of the atomic coordinates.

Such "fingerprint" representations are shown in Figure 1c and f. As the fingerprint of the phenol molecule in Figure 1c consists of very few elements, it will be used as an example on how these fingerprints are constructed. For clarity, the Euclidean distances of similar length between all carbon atom pairs in the phenol molecule are shown in red (~1.4 Å), green (~2.4 Å), or blue (~2.8 Å) (Figure 1a). The same colors (and respective atom identifiers) are used in the fingerprint representation of phenol (Figure 1c), so that the distances $r_{ij}$ between carbon atoms $C_i$ and $C_j$ can be easily distinguished. As shown in Figure 1a, there is a total of 15 carbon atom pairs.


dx.doi.org/10.1021/ci3001698 | J. Chem. Inf. Model. 2012, 52, 1854−1864

**Table 1. Experimental and Averaged Predicted log(RBA) Values According to the Model Using 10 ppm × 10 ppm × 0.5 Å Square Box Elements**

| compound name | class | CAS | log(RBA) | predicted log (RBA) | standard deviation | number of times predicted |
|---|---|---|---|---|---|---|
| Diethylstilbestrol (DES) | DES | 56−53−1 | 2.60 | 1.62 | 1.00 | 19 |
| Hexestrol | DES | 84−16−2 | 2.48 | 2.41 | 0.53 | 21 |
| Ethynylestradiol | Steroid | 57−63−6 | 2.28 | 3.28 | 0.60 | 21 |
| 4-OH-Tamoxifen | DES | 68047−06−3 | 2.24 | 2.88 | 0.60 | 15 |
| 17$\beta$-Estradiol (E2) | Steroid | 50−28−2 | 2.00 | 1.80 | 0.40 | 13 |
| 4-OH-Estradiol | Steroid | 5976−61−4 | 1.82 | 1.45 | 0.54 | 19 |
| Zearalenol | Phyto | 36455−72−8 | 1.63 | 0.69 | 0.70 | 20 |
| ICI 182780 | Steroid | 129453−61−8 | 1.57 | 1.82 | 0.84 | 25 |
| Dienestrol | DES | 84−17−3 | 1.57 | −0.86 | 0.44 | 15 |
| $\alpha$-Zearalanol | Phyto | 55331−29−8 | 1.48 | 2.32 | 0.98 | 18 |
| 2-OH-Estradiol | Steroid | 362−05−0 | 1.47 | 1.94 | 0.79 | 24 |
| Monomethyl ether diethylstilbestrol | DES | 7773−60−6 | 1.31 | 1.93 | 1.24 | 18 |
| 3,3′-Dihydroxyhexestrol | DES | 79199−51−2 | 1.19 | 0.37 | 0.44 | 28 |
| Droloxifene | DES | 82413−20−5 | 1.18 | 1.86 | 0.61 | 24 |
| ICI 164384 | Steroid | 98007−99−9 | 1.16 | 4.29 | 1.35 | 23 |
| Dimethylstilbestrol | DES | 552−80−7 | 1.16 | −0.23 | 0.35 | 14 |
| Moxestrol | Steroid | 34816−55−2 | 1.14 | 1.60 | 0.47 | 25 |
| 17-Deoxyestradiol | Steroid | 53−63−4 | 1.14 | 0.45 | 0.22 | 21 |
| 2,6-Dimethyl hexestrol | DES | | 1.11 | 2.07 | 0.68 | 19 |
| Estriol | Steroid | 50−27−1 | 0.99 | 0.50 | 0.51 | 18 |
| Monomethyl ether hexestrol | DES | 13026−26−1 | 0.97 | 3.58 | 0.90 | 20 |
| Estrone | Steroid | 53−16−7 | 0.86 | 0.26 | 0.38 | 20 |
| 3-(p-Phenol)-4-(p-tolyl)-hexane | DES | | 0.60 | 1.90 | 0.54 | 19 |
| 17$\alpha$-Estradiol | Steroid | 57−91−0 | 0.49 | −0.81 | 0.30 | 27 |
| Dihydroxymethoxychlor olefin | DDT | 14868−03−2 | 0.42 | −1.54 | 0.34 | 21 |
| Mestranol | Steroid | 72−33−3 | 0.35 | 2.70 | 0.59 | 20 |
| Zearalanone | Phyto | 5975−78−0 | 0.32 | 1.38 | 0.93 | 26 |
| Tamoxifen | DES | 10540−29−1 | 0.21 | 3.79 | 0.81 | 24 |
| Toremifene | DES | 89778−26−7 | 0.14 | 1.23 | 0.56 | 17 |
| $\alpha,\alpha$-Dimethyl-$\beta$-ethyl allenolic acid | Phenol | 65118−81−2 | −0.02 | −1.79 | 0.32 | 18 |
| Coumestrol | Phyto | 479−13−0 | −0.05 | −2.76 | 0.23 | 17 |
| 4-Ethyl-7-OH-3-(p-methoxyphenyl)coumarin | Phyto | 5219−17−0 | −0.05 | −0.54 | 0.39 | 17 |
| Clomiphene | DES | 911−45−5 | −0.14 | −1.05 | 0.44 | 18 |
| Nafoxidine | DES | 1845−11−0 | −0.14 | 2.61 | 0.81 | 12 |
| 6$\alpha$-OH-Estradiol | Steroid | 1229−24−9 | −0.15 | 0.48 | 0.28 | 24 |
| $\beta$-Zearalanol | Phyto | 42422−68−4 | −0.19 | 1.62 | 0.85 | 28 |
| 3-OH-Estra-1,3,5(10)-trien-16-one | Steroid | 3601−97−6 | −0.29 | −0.18 | 0.28 | 15 |
| 3-Deoxyestradiol | Steroid | 2529−64−8 | −0.30 | −0.40 | 0.73 | 25 |
| 3,6,4′-Trihydroxyflavone | Phyto | | −0.35 | −2.61 | 0.26 | 23 |
| Genistein | Phyto | 446−72−0 | −0.36 | −2.36 | 0.26 | 24 |
| 4,4′-Dihydroxystilbene | DES | 659−22−3 | −0.55 | −1.95 | 0.48 | 22 |
| HPTE | DDT | 2971−36−0 | −0.60 | −1.87 | 0.37 | 20 |
| Monohydroxymethoxychlor olefin | DDT | 75938−34−0 | −0.63 | −1.24 | 0.41 | 21 |
| 2,3,4,5-Tetrachloro-4′-biphenylol | PCB | 67651−34−7 | −0.64 | −3.21 | 0.28 | 23 |
| Norethynodrel | Steroid | 68−23−5 | −0.67 | −0.17 | 0.49 | 21 |
| 2,2′,4,4′-Tetrahydroxybenzil | Other | 5394−98−9 | −0.68 | −2.24 | 0.46 | 23 |
| B-Zearalenol | Phyto | 71030−11−0 | −0.69 | 3.08 | 0.98 | 16 |
| Equol | Phyto | 531−95−3 | −0.82 | −3.30 | 0.24 | 21 |
| 4′,6-Dihydroxyflavone | Phyto | 63046−09−3 | −0.82 | −3.12 | 0.33 | 13 |
| Monohydroxymethoxychlor | DDT | 28463−03−8 | −0.89 | −2.10 | 0.49 | 13 |
| 3$\beta$-Androstanediol | Steroid | 571−20−0 | −0.92 | −4.56 | 1.45 | 18 |
| Bisphenol B | DDT | 77−40−7 | −1.07 | −2.73 | 0.62 | 15 |
| Phloretin | Phyto | 60−82−2 | −1.16 | −2.42 | 0.24 | 15 |
| Diethylstilbestrol dimethyl ether | DES | 7773−34−4 | −1.25 | 3.72 | 0.59 | 15 |
| 2′,4,4′-Trihydroxychalcone | Phyto | 961−29−5 | −1.26 | −1.94 | 0.29 | 23 |
| 2,5-Dichloro-4′-biphenylol | PCB | 53905−28−5 | −1.44 | −3.12 | 0.25 | 25 |
| 4,4′-(1,2-Ethanediyl)bisphenol | DES | 6052−84−2 | −1.44 | −2.84 | 0.38 | 25 |
| 16$\beta$−OH-16-Methyl-3-methyl-estradiol | Steroid | 5108−94−1 | −1.48 | 0.36 | 0.45 | 23 |
| Aurin | DDT | 603−45−2 | −1.50 | −1.52 | 0.66 | 21 |


dx.doi.org/10.1021/ci3001698 | J. Chem. Inf. Model. 2012, 52, 1854−1864

**Table 1. continued**

| compound name | class | CAS | log(RBA) | predicted log (RBA) | standard deviation | number of times predicted |
|---|---|---|---|---|---|---|
| Nordihydroguaiaretic acid | Other | 500−38−9 | −1.51 | 0.40 | 1.17 | 16 |
| Nonylphenol | Phenol | 25154−52−3 | −1.53 | −2.72 | 0.31 | 26 |
| Apigenin | Phyto | 520−36−5 | −1.55 | −2.57 | 0.36 | 21 |
| Kaempferol | Phyto | 520−18−3 | −1.61 | −2.66 | 0.38 | 16 |
| Daidzein | Phyto | 486−66−8 | −1.65 | −1.63 | 0.22 | 21 |
| 3-Methyl-estriol | Steroid | 3434−79−5 | −1.65 | 0.90 | 0.47 | 26 |
| 4-Dodecylphenol | Phenol | 104−43−8 | −1.73 | −1.93 | 0.45 | 24 |
| 2-Ethylhexyl-4-hydroxybenzoate | Phenol | 5153−25−3 | −1.74 | −1.95 | 0.24 | 23 |
| 4-t-Octylphenol | Phenol | 140−66−9 | −1.82 | −3.08 | 0.44 | 24 |
| Phenolphthalein | DDT | 77−09−8 | −1.87 | −2.29 | 0.61 | 23 |
| Kepone | Pesticide | 143−50−0 | −1.89 | −3.97 | 0.28 | 34 |
| Heptyl p-hydroxybenzoate | Phenol | 1085−12−7 | −2.09 | −2.68 | 0.26 | 16 |
| Bisphenol A | DDT | 80−05−7 | −2.11 | −4.06 | 0.39 | 21 |
| Naringenin | Phyto | 480−41−1 | −2.13 | −2.98 | 0.29 | 18 |
| 4-Chloro-4′-biphenylol | PCB | 28034−99−3 | −2.18 | −3.89 | 0.16 | 19 |
| 3-Deoxy-estrone | Steroid | 53−45−2 | −2.20 | 0.04 | 0.23 | 26 |
| p-Cumyl phenol | DDT | 599−64−4 | −2.30 | −4.06 | 0.22 | 18 |
| 4-n-Octylphenol | Phenol | 1806−26−4 | −2.31 | −2.48 | 0.21 | 12 |
| Fisetin | Phyto | 528−48−3 | −2.35 | −2.08 | 0.24 | 17 |
| 3′,4′,7-Trihydroxy isoflavone | Phyto | 485−63−2 | −2.35 | −1.96 | 0.23 | 18 |
| Biochanin A | Phyto | 491−80−5 | −2.37 | −1.71 | 0.55 | 31 |
| 4′-Hydroxychalcone | Phyto | 2657−25−2 | −2.43 | −2.86 | 0.17 | 23 |
| 2,2′-Methylenebis(4-Chlorophenol) | DDT | 97−23−4 | −2.45 | −3.06 | 0.24 | 16 |
| 4,4′-Dihydoxy-benzophenone | DDT | 611−99−4 | −2.46 | −2.70 | 0.30 | 16 |
| Benzyl 4-hydroxybenzoate | Phenol | 94−18−8 | −2.54 | −3.38 | 0.20 | 19 |
| 4-Hydroxychalcone | Phyto | 20426−12−4 | −2.55 | −2.87 | 0.20 | 27 |
| 2,4-Hydroxybenzophenone | DDT | 131−56−6 | −2.61 | −3.08 | 0.22 | 20 |
| 4′-hydroxyflavanone | Phyto | 6515−37−3 | −2.65 | −3.66 | 0.19 | 25 |
| 3$\alpha$-Androstanediol | Steroid | 1852−53−5 | −2.67 | −2.13 | 0.62 | 12 |
| 4-Phenethylphenol | Phenol | 6335−83−7 | −2.69 | −3.67 | 0.20 | 18 |
| Doisynoestrol | Other | 15372−34−6 | −2.74 | 0.02 | 0.38 | 26 |
| Prunetin | Phyto | 552−59−0 | −2.74 | −0.83 | 0.81 | 22 |
| Myricetin | Phyto | 529−44−2 | −2.75 | −2.95 | 0.34 | 23 |
| 2-Chloro-4-biphenylol | PCB | 92−04−6 | −2.77 | −2.91 | 0.23 | 15 |
| Triphenylethylene | DES | 58−72−0 | −2.78 | −2.84 | 0.55 | 24 |
| 3′-Hydroxyflavanone | Phyto | 1621−55−2 | −2.78 | −3.12 | 0.17 | 22 |
| Chalcone | Phyto | 94−41−7 | −2.82 | −4.91 | 0.19 | 20 |
| o,p′-DDT | DDT | 789−02−6 | −2.85 | −3.66 | 0.20 | 24 |
| 4-Heptyloxyphenol | Phenol | 13037−86−0 | −2.88 | −2.39 | 0.27 | 18 |
| Dihydrotestosterone | Steroid | 521−18−6 | −2.89 | −1.72 | 0.42 | 19 |
| Formononetin | Phyto | 485−72−3 | −2.98 | −1.36 | 0.31 | 18 |
| Bis(4-hydroxyphenyl)methane | DDT | 620−92−8 | −3.02 | −2.64 | 0.40 | 18 |
| 4-Hydroxybiphenyl | PCB | 92−69−3 | −3.04 | −3.68 | 0.17 | 12 |
| Baicalein | Phyto | 491−67−8 | −3.05 | −3.10 | 0.32 | 23 |
| 6-Hydroxyflavanone | Phyto | 4250−77−5 | −3.05 | −3.98 | 0.33 | 16 |
| 4,4′-Sulfonyldiphenol | DDT | 80−09−1 | −3.07 | −3.27 | 0.30 | 22 |
| *n*-Butyl 4-hydroxybenzoate | Phenol | 94−26−8 | −3.07 | −3.06 | 0.20 | 18 |
| Morin | Phyto | 480−16−0 | −3.09 | −1.73 | 0.31 | 20 |
| Diphenolic acid | DDT | 126−00−1 | −3.13 | −2.18 | 0.27 | 21 |
| 1,3-Diphenyltetramethyldisiloxane | Siloxane | 56−33−7 | −3.16 | −4.26 | 0.30 | 15 |
| *n*-Propyl 4-hydroxybenzoate | Phenol | 94−13−3 | −3.22 | −3.03 | 0.13 | 23 |
| Ethyl 4-hydroxybenzoate | Phenol | 120−47−8 | −3.22 | −3.27 | 0.17 | 21 |
| 3,3′,5,5′-Tetrachloro-4,4′-biphenyldiol | PCB | 13049−13−3 | −3.25 | −3.98 | 0.20 | 17 |
| 4-tert-Amylphenol | Phenol | 80−46−6 | −3.26 | −3.01 | 0.33 | 16 |
| 4-s-Butylphenol | Phenol | 99−71−8 | −3.37 | −3.30 | 0.21 | 18 |
| 4-Chloro-3-methylphenol | Phenol | 59−50−7 | −3.38 | −2.88 | 0.16 | 20 |
| 6-Hydroxyflavone | Phyto | 6665−83−4 | −3.41 | −3.28 | 0.30 | 23 |
| 3-Phenylphenol | PCB | 580−51−8 | −3.44 | −3.47 | 0.12 | 19 |
| 4-(Benzyloxyl)phenol | Phenol | 103−16−2 | −3.44 | −3.66 | 0.33 | 15 |
| Methyl 4-hydroxybenzoate | Phenol | 99−76−3 | −3.44 | −3.35 | 0.17 | 19 |


dx.doi.org/10.1021/ci3001698 | *J. Chem. Inf. Model.* 2012, 52, 1854−1864

Table 1. continued

| compound name | class | CAS | log(RBA) | predicted log (RBA) | standard deviation | number of times predicted |
|---|---|---|---|---|---|---|
| 2-s-Butylphenol | Phenol | 89−72−5 | −3.54 | −2.79 | 0.17 | 18 |
| 4-tert-Butylphenol | Phenol | 98−54−4 | −3.61 | −4.06 | 0.28 | 17 |
| 2,4′-Dichlorobiphenyl | PCB | 34883−43−7 | −3.61 | −3.90 | 0.19 | 25 |
| 2-Cholor-4-methyl phenol | Phenol | 6640−27−3 | −3.66 | −3.36 | 0.18 | 15 |
| Phenolphthalin | DDT | 81−90−3 | −3.67 | −0.51 | 0.49 | 12 |
| 4-Chloro-2-methyl phenol | Phenol | 1570−64−5 | −3.67 | −3.52 | 0.19 | 25 |
| 7-Hydroxyflavanone | Phyto | 6515−36−2 | −3.73 | −2.63 | 0.19 | 16 |
| 3-Ethylphenol | Phenol | 620−17−7 | −3.87 | −2.74 | 0.22 | 20 |
| Rutin | Phyto | 153−18−4 | −4.09 | −0.03 | 0.50 | 17 |
| 4-Ethylphenol | Phenol | 123−07−9 | −4.17 | −2.71 | 0.19 | 19 |
| 4-Cresol | Phenol | 106−44−5 | −4.50 | −3.15 | 0.26 | 15 |

Note that due to molecular symmetry, the $^{13}$C NMR spectrum of phenol shown in Figure 1b has only four unique peaks. In this spectrum, the corresponding carbon atom identifiers are shown in parentheses next to $\delta$.

Because all elements of a fingerprint are constructed in a similar manner, the "1,4" pair (and its symmetry partner "1,5") will be used as an illustration. In Figure 1a, carbon atoms "1" and "4" (and respectively "1" and "5") are 2.41 Å apart. Atom "1" has a $\delta$ of 156.99 ppm, while atoms "4" and "5" have $\delta$ of 129.87 ppm (Figure 1b). These three values ($\delta_1\delta_4 r_{1,4} \equiv \delta_1\delta_5 r_{1,5}$) define the positions of fingerprint elements "1,4" and "1,5", which coincide due to the symmetry of the phenol molecule (Figure 1c). All remaining fingerprint elements are constructed in a similar manner. Because there are six symmetry partners, the fingerprint of phenol consists of only nine distinct elements. The black circles representing the shadows of these elements in the *XY*-plane are shown for the sole purpose of visualizing the elements' positions (to simulate depth in fingerprint abstract space) and are not used further in calculations.

As obvious from Figure 1, the complexity of the 3D spectral fingerprints depends on the symmetry (point group) and size of the molecule. In the case of simple highly symmetric molecules, the fingerprints consist of very few elements (Figure 1c), while for larger asymmetric molecules (Figure 1d) the fingerprints are much more complex (Figure 1f). For any molecule with more than five carbon atoms, the number of the fingerprint elements $[N(N-1)/2]$, exceeds the degrees of freedom $(3N-6)$ resulting in a system overdetermined by $(N^2 - 7N + 12)/2$.

Because the fingerprints' elements exist along continua, to capture more efficiently the chemical information contained in them, the concept of a 3D grid dividing the space into smaller compartments ("bins") was introduced; one such bin is shown in green in Figure 1f. In this regard the "bins" played the role of a measure of similarity putting together carbon atom pairs placed in similar chemical environments. Depending on their congruency, these "bins" can form either regular or irregular grids. As there is a simple rule (setting a step size in each dimension) that allows an objective generation of regular grids, this type was preferred and used further.

As multiple linear regression is inefficient in cases in which the number of independent variables greatly exceeds the number of cases (compounds),[33] it would be preferable to handle the data processing using projection methods such as principal component analysis (PCA) or PLS. A major advantage of PLS over PCA is that it uses the response variable ($\mathbf{Y}$) to transform the original descriptor space ($\mathbf{X}$) into an orthogonal set of predictors that explain variance in both $\mathbf{X}$ and $\mathbf{Y}$ (PCA operates only on $\mathbf{X}$).[34] Similarly to CoMFA, the PLS weights of the original variables can be mapped back to the original fingerprint space allowing the identification of positively and negatively contributing 3D spatial regions (clusters of bins) with specific chemical shifts and atom-to-atom distances.

Being a relatively new approach, the 3D-QSDAR parametric space has not yet been systematically studied. This work is intended to explore the characteristics of the parametric space defined by the chemical shifts of the $^{13}$C atoms and their corresponding interatomic distances and to provide preliminary estimates of the optimal granularity of the 3D grid. Furthermore, the PLS bin weights (i.e., the associated chemical shifts and interatomic distances) were used to identify key substructural features related to activity, thus illustrating the interpretability of the 3D-SDAR models in terms of structural characteristics.

**Case Study: 3D-QSDAR with Congruent Bins.** Because of their enormous impact on the health and behavior of a wide range of species,[35−38] a diverse data set of 130 chemicals found in the endocrine disruptors knowledge base (EDKB)[39] (Table 1) was selected as a subject of the present study. As data sets derived from EDKB were studied extensively,[40−59−42−61] the structural features responsible for binding are well-known; this permitted a direct comparison of our results to earlier findings for the same data set.

Because the applied PLS technique is prone to over-parameterization, the ability to identify potentially overfitted models would be critical in assessing reliability of the derived QSDARs. Hence, it would be desirable to set an upper limit for $R^2$ on the basis of the standard deviation ($\sigma$) of the original measurements. Applying the methodology described by Doweyko et al.,[43] 100 simulated models using data with $\sigma$ equal to that of the EDKB data ($IC_{50}$ $\sigma = 2.73 \times 10^{-5}$) were generated. The average $R^2$ was 0.89 (ranging from 0.55 to 0.98) with a standard deviation of 0.073. However, this upper limit would be achievable only if the utilized descriptor set is able to explain the total variance in the biological data, thus an unlikely outcome.

To determine the global minimum of the potential energy surface, each of the 130 molecules was subjected to a conformational search analysis. A random walks search method with an acceptance energy criterion of 6 kcal/mol, as implemented in HyperChem 8.0,[44] was used. The lowest energy conformer was further optimized employing a semi-

Table 2. Average Statistical Parameters for the 50 QSDAR Models Using a 2−20 ppm Granularity in the Chemical Shifts Plane (*XY*) and a 0.5−2.5 Å Resolution on the Distance Axis (*Z*)[a]

| (*Z*)/(*XY*)[b] | | 2 ppm | 4 ppm | 6 ppm | 8 ppm | 10 ppm | 12 ppm | 14 ppm | 16 ppm | 18 ppm | 20 ppm | average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.5 Å | $R^2_{trn}$ | 0.910 | 0.890 | 0.826 | 0.835 | 0.803 | 0.772 | 0.713 | 0.708 | 0.641 | 0.683 | 0.778 |
| | $R^2_{test}$ | 0.450 | 0.519 | 0.496 | 0.507 | 0.546 | 0.561 | 0.531 | 0.501 | 0.494 | 0.546 | 0.515 |
| | $R^2_{scr}$ | 0.062 | 0.058 | 0.058 | 0.066 | 0.064 | 0.065 | 0.080 | 0.075 | 0.074 | 0.084 | 0.069 |
| 1.0 Å | $R^2_{trn}$ | 0.885 | 0.843 | 0.775 | 0.768 | 0.735 | 0.711 | 0.665 | 0.652 | 0.599 | 0.646 | 0.728 |
| | $R^2_{test}$ | 0.448 | 0.502 | 0.488 | 0.485 | 0.529 | 0.541 | 0.513 | 0.477 | 0.480 | 0.534 | 0.500 |
| | $R^2_{scr}$ | 0.067 | 0.064 | 0.062 | 0.071 | 0.072 | 0.071 | 0.083 | 0.081 | 0.082 | 0.088 | 0.074 |
| 1.5 Å | $R^2_{trn}$ | 0.868 | 0.819 | 0.735 | 0.733 | 0.705 | 0.679 | 0.645 | 0.617 | 0.579 | 0.622 | 0.700 |
| | $R^2_{test}$ | 0.449 | 0.507 | 0.484 | 0.497 | 0.537 | 0.525 | 0.512 | 0.477 | 0.465 | 0.519 | 0.497 |
| | $R^2_{scr}$ | 0.069 | 0.068 | 0.061 | 0.070 | 0.075 | 0.073 | 0.085 | 0.081 | 0.078 | 0.088 | 0.075 |
| 2.0 Å | $R^2_{trn}$ | 0.850 | 0.790 | 0.719 | 0.702 | 0.690 | 0.664 | 0.624 | 0.619 | 0.573 | 0.620 | 0.685 |
| | $R^2_{test}$ | 0.450 | 0.504 | 0.499 | 0.491 | 0.542 | 0.521 | 0.502 | 0.474 | 0.464 | 0.516 | 0.496 |
| | $R^2_{scr}$ | 0.071 | 0.072 | 0.066 | 0.074 | 0.080 | 0.075 | 0.088 | 0.083 | 0.081 | 0.098 | 0.079 |
| 2.5 Å | $R^2_{trn}$ | 0.839 | 0.771 | 0.696 | 0.672 | 0.674 | 0.649 | 0.611 | 0.590 | 0.561 | 0.613 | 0.668 |
| | $R^2_{test}$ | 0.449 | 0.493 | 0.495 | 0.506 | 0.547 | 0.511 | 0.491 | 0.461 | 0.458 | 0.503 | 0.491 |
| | $R^2_{scr}$ | 0.073 | 0.072 | 0.068 | 0.074 | 0.084 | 0.081 | 0.089 | 0.083 | 0.081 | 0.098 | 0.080 |
| average | $R^2_{trn}$ | 0.870 | 0.823 | 0.750 | 0.742 | 0.721 | 0.695 | 0.652 | 0.637 | 0.591 | 0.637 | 0.870 |
| | $R^2_{test}$ | 0.449 | 0.505 | 0.492 | 0.497 | 0.540 | 0.532 | 0.510 | 0.478 | 0.472 | 0.524 | 0.449 |
| | $R^2_{scr}$ | 0.068 | 0.067 | 0.063 | 0.071 | 0.075 | 0.073 | 0.085 | 0.081 | 0.079 | 0.091 | 0.068 |

[a]The three "average" rows show the change in the average statistical parameters as a function of the granularity in the *XY*-plane. The "average" column shows the change in the average statistical parameters as a function of the resolution on the *Z*-axis. [b]$R^2_{trn}$ denotes $R^2$ for the training set; $R^2_{test}$ represents the $R^2$ for the test set; $R^2_{scr}$ denotes $R^2$ from scrambling.

empirical AM1 Hamiltonian with a root-mean-square gradient of 0.01 kcal/Å × mol and saved as a mol file. These mol files were then imported to the ACD/NMR $^{13}$C Predictor version 12.0,[45] and the NMR spectra of the corresponding compounds were generated. In case of unknowns, the ACD/NMR $^{13}$C Predictor relies on a spectral library and an algorithm correlating similar structures with similar NMR chemical shifts.[46] These simulated spectra were preferred to those calculated by ab initio approaches because of their high accuracy and speed of generation. Using the ACD/NMR $^{13}$C Predictor output files, an average error per carbon shift of ±4.13 ppm for the current data set was obtained. The NMR spectral data and the distances between the carbon atoms were used to generate a unique 3D-fingerprint for each molecule from the data set. The fingerprints of all 130 compound were binned using an in-house program with a 2 ppm step in the chemical shifts plane *XY* (e.g., 2, 4, 6, ..., 20 ppm) and 0.5 Å step for the interatomic distances on the *Z*-axis (e.g., 0.5, 1.0, 1.5, 2.0, and 2.5 Å). That is a total of 50 regular grids with bins ranging in size from 2 ppm × 2 ppm × 0.5 Å to 20 ppm × 20 ppm × 2.5 Å were created. The number of fingerprint elements in each bin (bin occupancy) was then counted and stored in columns.

## RESULTS AND DISCUSSION

The bin occupancy for each of the 50 generated 3D grids was further used to explain a portion of the variance of the ER binding data. For this purpose, each of the 50 data matrices formed by rows representing the bin occupancy for every individual compound was randomly split into training (4/5 of the total or 104 compounds) and "hold-out" test (1/5 of the total or 26 compounds) subsets. To minimize the probability of incorrectly estimating the performance of a particular QSDAR architecture due to arbitrarily choosing a training and test set pair, 100 such combinations were randomly generated. After 100 runs, the random number generator was reinitialized with the same seed in order to recreate the same training/test set sequence for each of the 50 QSDARs.

A PLS code written in R v.2.13.2[47] was used to process the binned fingerprint data; depending on the bin size, somewhere between 283 (20 ppm × 20 ppm × 2.5 Å) and 6792 (2 ppm × 2 pmm × 0.5 Å) original variables were processed. A 10-fold cross-validation procedure running within the training set was aimed to estimate the optimal number of latent variables (LVs). In more than 70% of the cases, 3LVs resulted in models with best predictive accuracy. Hence, to study the performance of the QSDAR models as a function of the 3D grid granularity, the number of the LVs was fixed to three. At the end, the PLS models obtained for the training subsets were used to predict the logarithm of the relative binding affinities (log(RBA)) for the complementary "hold-out" test subsets (Table 2). As described above, 100 training/test set combinations were generated for each of the 50 QSDAR models. The statistical parameters in Table 2 ($R^2_{trn}$, $R^2_{test}$, and $R^2_{scr}$) are reported as averages of these 100 run cycles.

As shown in Table 2, the coefficients of determination ($R^2$) for the training subsets decrease from left to right and from top to bottom. The model, using bins of 2 ppm × 2 ppm × 0.5 Å (the upper left corner of Table 2), slightly exceeds the upper limit of $R^2$ (0.89) estimated on the basis of the standard deviation of the experimental data. However, this was an expected statistical artifact of the constant number of LVs used throughout the calculations and the strong PLS fitting ability.

The predictive power of the models estimated on the basis of the "hold-out" test sets reveals a much more interesting pattern. As a function of the granularity of the chemical shifts plane *XY*, $R^2_{test}$ reaches a maximum at 10 (±1) ppm (i.e., 10 ppm ×10 ppm × *Z* Å, *Z* = 0.5...2.5 Å; average $R^2_{test}$ = 0.54, $\sigma$ = 0.007). Such an outcome may seem counterintuitive to the expectation that the high resolution QSDAR models should perform better than the more coarse-grained ones because of their potentially higher information content. However, this expectation does not take into account the chemical shift estimation error (±4.13


dx.doi.org/10.1021/ci3001698 | J. Chem. Inf. Model. 2012, 52, 1854−1864

ppm) and the discretization nature of the bin occupancy count procedure. As each bin is defined by a given range on each of the three axes (e.g., *X*, 130−140 ppm; *Y*, 120−130 ppm; and *Z*, 2.5−3.0 Å), a negative error at the lower range end combined with a positive error at the higher range end would effectively double the error. Hence, at high resolutions in the *XY* chemical shifts plane (2−8 ppm), the NMR estimation error would be a significant factor that may cause a random assignment of fingerprint elements to neighboring bins. Such uncertainty introduces noise into the model and hence lowers $R^2_{test}$.

Similar considerations should also apply to the distance axis (*Z*-axis). However, the AM1 method utilized for geometry optimization is known to provide bond length estimates deviating from the experiment by an average of 0.02 Å per bond.[48] Therefore, in most cases, the inaccuracy of the calculated $atom_i$-to-$atom_j$ distances will be much lower than the step of 0.5 Å used for binning the *Z*-axis. Incorrect assignment of fingerprint elements to neighboring *Z*-bins can be expected either due to cumulative effects (significant for carbon atoms separated by several bonds) or the presence of flexible residues with a high degree of conformational freedom. As most of the structures in our data set are relatively rigid, this uncertainty plays a negligible role and the best models should be those at high resolutions on the *Z*-axis.

The existence of an optimum at 20 ppm in the $R^2_{test}$ vs chemical shift granularity plot (Figure 2) suggests that more

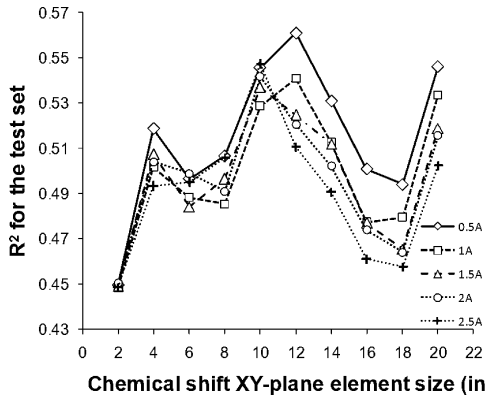



**Figure 2.** Average $R^2$ for the test set as a function of the bin size.

coarse-grained and less computationally demanding models may perform as well as those obtained at higher spectral binning resolutions. However, the low-resolution models would be of a limited utility for decoding the underlying structure−activity relationship due to the significant overlap of the chemical shifts of chemically/biologically nonequivalent carbon atoms.

To ensure the quality of the carried out models and to avoid the probability of generating correlations by chance, a scrambling procedure was applied to all 50 QSDAR models. The biological activities characterizing the training set were randomly assigned to any of the 104 compounds, then the respective PLS models were built, and the test set compounds were predicted. The average $R^2_{scr}$ of 100 scrambling cycles per QSDAR model are reported in Table 2. As shown, the significant difference between the actual $R^2_{test}$ and the $R^2_{scr}$ confirms that none of the reported models was obtained as a result of chance correlations.

**Interpretation.** As discussed above, the models with a 10 ppm granularity in the chemical shifts plane *XY* result in consistently high predictions (having the highest average $R^2_{test}$ in Table 2). Among these the 10 ppm × 10 ppm × 0.5 Å performs best. For this model, the averaged predicted log(RBA) values for all 130 compounds when part of the test set are given in Table 1. Because the 100 test sets were randomly generated and the size of the hold-out test set was set at 20% of the total, each of the 130 compounds had a probability of 1/5 to be a selected as a part of the test set. As for relatively small samples the random number generators are unable to produce uniformly distributed random numbers there are slight deviations from the above probability. The actual numbers showing how many out of a hundred times each compound was predicted (i.e., was part of the test set) are given in the last column of Table 1.

Using a standard outlier detection technique, seven $2\sigma$ outliers were detected (Figure 3), namely rutin, ICI 164384,

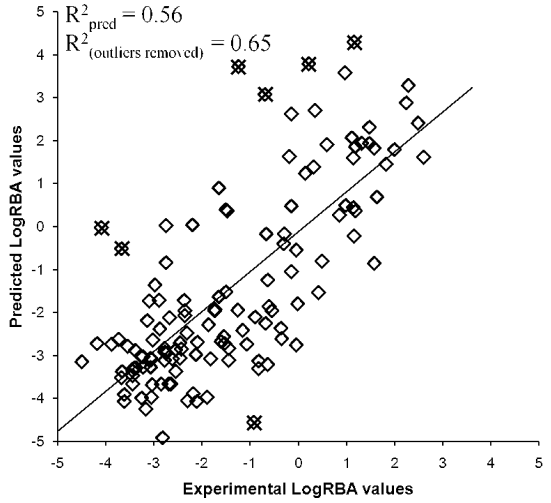



**Figure 3.** Plot of predicted vs experimental log(RBA) values. Predicted values are an average of 100 predictions for all 130 compounds.

3$\beta$-androstanediol, tamoxifen, phenolphthalin, diethylstilbestrol dimethyl ether, and $\beta$-zearalenol. The binding affinity of six of these outliers (except for 3$\beta$-androstanediol) was significantly overestimated. However, for regulatory purposes, overestimation of toxicity is preferable to underestimation. Except for the 3$\beta$-androstanediol, which is rigid, the remaining six outliers (all of them overestimated) were characterized by a varying degree of structural flexibility. Hence, the lowest energy conformations in vacuum for these six molecules might have been considerably different from the biologically relevant ones. This is particularly true for the highly flexible structure of rutin, which has large conformational entropy in its unbound state that has been associated with its weak ER binding.[49] The removal of the above identified outliers improved significantly the predictive $R^2_{pred}$ (calculated on the basis of the averaged predictions for all 130 compounds) from 0.56 to 0.65.

Depending on the distribution of the outliers between the 100 training and test sets and the probability that some of the compounds in the randomly formed test sets may have been out of the applicability domain of the model, $R^2_{test}$ was found to vary from 0.11 to 0.75 (average $R^2_{test}$ = 0.55, $\sigma$ = 0.12). The variations of $R^2_{test}$ for the 10 ppm × 10 ppm × 0.5 Å models are shown in Figure 4. The 10 least accurate submodels had $R^2_{test}$ in the range of 0.11−0.40 (indicated using diamonds), while the 10 most accurate submodels had $R^2_{test}$ between 0.71 and


dx.doi.org/10.1021/ci3001698 | J. Chem. Inf. Model. 2012, 52, 1854−1864

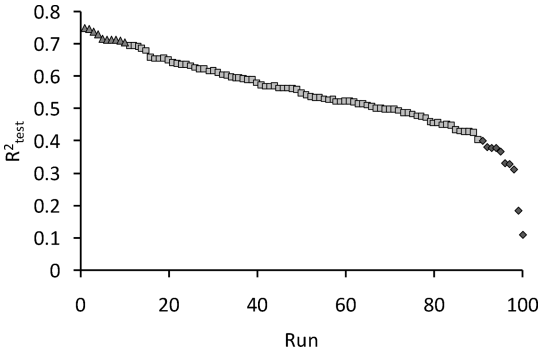



**Figure 4.** Variation of $R^2_{test}$ for the 100 submodels using 10 ppm × 10 ppm × 0.5 Å bin size. The 10 most accurate models are indicated by triangles; the 10 least accurate by diamonds; all the rest are shown using squares.

0.75 (indicated by triangles). The former had between 1 and 4 outliers, while the latter had either 1 or no outliers presented in the test set. On average, the 10 worst performing models had 3.2 times more outliers that the 10 best ones.

In order to identify the bins explaining most of the variance of the log(RBA) data, their corresponding weights were extracted and ranked. The topmost 10 positive weights for each of the three LVs generated as a result of 100 training/test set combinations resulted in a total of 3000 significant bins; among these, 88 were unique. Table 3 shows the ranges of the chemical shifts of the carbon atoms, their proximity range, and the frequency of occurrence of the unique bins. The bins with a frequency of occurrence of more than 1% were then transferred to Table 4, in which each of the chemical shift ranges (shown in the first two columns of Table 3) were associated with carbon atoms having $\delta$ in this range. The distances from column 3 and the frequencies from column 4 in Table 3 are shown in the corresponding intersecting cells of Table 4. For example, the bin with the highest frequency of occurrence (6.03%) shown in the first row of Table 3 was transferred to Table 4 in which the given chemical shift ranges (130−140 ppm and 120−130 ppm) were associated with the presence of specific carbon atoms; the two values in the intersecting cell show the interatomic distance and the frequency of occurrence of the bin (1−1.5 Å; 6.03%). This bin is consistent with the presence of an aromatic ring, chlorine substituted aromatic ring, or tetrahydronaphthalene moiety.

As shown in Table 4, the highest frequency of occurrence can be attributed to the presence of three structural fragments: an aromatic ring (120−130 ppm), a monosubstituted tetrahydronaphthalene (130−140 ppm), and dodecahydro−cyclopentanaphthalene systems (40−50 ppm). These three fragments are the building blocks of the steroid skeleton—a confirmation that structure−activity inferences based on 3D-QSDAR descriptors are correct because the strongest estrogens are steroids.

**Table 3. Frequency of Occurrence (in % of the total) of the Bins Found Significant in All 100 10 ppm × 10 ppm × 0.5 Å Models (100 models × 3 LVs × 10 bins = 3000 total bins)**

| $C_1$ chemical shift | $C_2$ chemical shift | distance in Å | frequency of occurrence | $C_1$ chemical shift | $C_2$ chemical shift | distance in Å | Frequency of occurrence | $C_1$ chemical shift | $C_2$ chemical shift | distance in Å | frequency of occurrence |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 130−140 | 120−130 | 1−1.5 | 6.03 | 30−40 | 20−30 | 6.5−7 | 0.97 | 130−140 | 130−140 | 2−2.5 | 0.13 |
| 120−130 | 120−130 | 2.5−3 | 5.87 | 190−200 | 120−130 | 3.5−4 | 0.90 | 30−40 | 20−30 | 3−3.5 | 0.13 |
| 40−50 | 30−40 | 2.5−3 | 5.80 | 120−130 | 120−130 | 5.5−6 | 0.73 | 110−120 | 30−40 | 4.5−5 | 0.10 |
| 120−130 | 120−130 | 1−1.5 | 5.57 | 120−130 | 110−120 | 6.5−7 | 0.70 | 130−140 | 120−130 | 3−3.5 | 0.10 |
| 120−130 | 120−130 | 2−2.5 | 5.43 | 160−170 | 110−120 | 3.5−4 | 0.63 | 140−150 | 110−120 | 2−2.5 | 0.10 |
| 130−140 | 120−130 | 2−2.5 | 4.90 | 130−140 | 120−130 | 8.5−9 | 0.53 | 120−130 | 110−120 | 1−1.5 | 0.07 |
| 30−40 | 30−40 | 1.5−2 | 4.23 | 150−160 | 110−120 | 1−1.5 | 0.53 | 120−130 | 120−130 | 4.5−5 | 0.07 |
| 130−140 | 130−140 | 5.5−6 | 3.67 | 120−130 | 110−120 | 7−7.5 | 0.47 | 130−140 | 120−130 | 3.5−4 | 0.07 |
| 160−170 | 130−140 | 2−2.5 | 3.37 | 120−130 | 120−130 | 7−7.5 | 0.47 | 150−160 | 130−140 | 7.5−8 | 0.07 |
| 40−50 | 20−30 | 2.5−3 | 3.10 | 40−50 | 20−30 | 6.5−7 | 0.47 | 150−160 | 50−60 | 2−2.5 | 0.07 |
| 130−140 | 120−130 | 2.5−3 | 3.00 | 110−120 | 40−50 | 3.5−4 | 0.43 | 160−170 | 120−130 | 2.5−3 | 0.07 |
| 120−130 | 30−40 | 3.5−4 | 2.83 | 130−140 | 130−140 | 1−1.5 | 0.43 | 30−40 | 20−30 | 1.5−2 | 0.07 |
| 30−40 | 20−30 | 4.5−5 | 2.80 | 130−140 | 110−120 | 7−7.5 | 0.40 | 30−40 | 20−30 | 7.5−8 | 0.07 |
| 40−50 | 30−40 | 1.5−2 | 2.63 | 120−130 | 110−120 | 7.5−8 | 0.37 | 120−130 | 100−110 | 5.5−6 | 0.03 |
| 120−130 | 110−120 | 2−2.5 | 2.33 | 120−130 | 110−120 | 8−8.5 | 0.33 | 120−130 | 100−110 | 6.5−7 | 0.03 |
| 30−40 | 30−40 | 2.5−3 | 2.33 | 170−180 | 130−140 | 4.5−5 | 0.33 | 120−130 | 120−130 | 6.5−7 | 0.03 |
| 30−40 | 20−30 | 2.5−3 | 2.23 | 120−130 | 120−130 | 8−8.5 | 0.30 | 120−130 | 20−30 | 3.5−4 | 0.03 |
| 40−50 | 30−40 | 3.5−4 | 2.20 | 130−140 | 30−40 | 2.5−3 | 0.30 | 120−130 | 40−50 | 3.5−4 | 0.03 |
| 160−170 | 90−100 | 1−1.5 | 2.17 | 140−150 | 130−140 | 2−2.5 | 0.27 | 130−140 | 110−120 | 3.5−4 | 0.03 |
| 70−80 | 70−80 | 1.5−2 | 2.10 | 130−140 | 110−120 | 2.5−3 | 0.23 | 130−140 | 120−130 | 5.5−6 | 0.03 |
| 110−120 | 30−40 | 3.5−4 | 1.83 | 130−140 | 120−130 | 4.5−5 | 0.23 | 130−140 | 120−130 | 7−7.5 | 0.03 |
| 130−140 | 110−120 | 2−2.5 | 1.63 | 140−150 | 120−130 | 1−1.5 | 0.23 | 130−140 | 120−130 | 7.5−8 | 0.03 |
| 70−80 | 70−80 | 2.5−3 | 1.60 | 150−160 | 120−130 | 7−7.5 | 0.23 | 130−140 | 20−30 | 2.5−3 | 0.03 |
| 120−130 | 30−40 | 2.5−3 | 1.57 | 160−170 | 150−160 | 2−2.5 | 0.23 | 130−140 | 40−50 | 2.5−3 | 0.03 |
| 150−160 | 110−120 | 2−2.5 | 1.53 | 50−60 | 20−30 | 8.5−9 | 0.23 | 140−150 | 110−120 | 3.5−4 | 0.03 |
| 40−50 | 30−40 | 8−8.5 | 1.37 | 120−130 | 0−10 | 6.5−7 | 0.20 | 140−150 | 130−140 | 2.5−3 | 0.03 |
| 130−140 | 110−120 | 5.5−6 | 1.30 | 20−30 | 20−30 | 3.5−4 | 0.20 | 160−170 | 110−120 | 1−1.5 | 0.03 |
| 150−160 | 130−140 | 2.5−3 | 1.13 | 120−130 | 120−130 | 8.5−9 | 0.17 | 30−40 | 20−30 | 3.5−4 | 0.03 |
| 120−130 | 110−120 | 2.5−3 | 1.00 | 120−130 | 50−60 | 5.5−6 | 0.17 | | | | |
| 150−160 | 120−130 | 2−2.5 | 0.97 | 40−50 | 30−40 | 6.5−7 | 0.17 | | | | |


dx.doi.org/10.1021/ci3001698 | J. Chem. Inf. Model. 2012, 52, 1854−1864

**Table 4. Bins and Associated Substructural Fragments with a Frequency of Occurrence of More Than 1% (first 29 bins from Table 3)**

| Chemical shifts | | 30-40ppm | 40-50ppm | 70-80ppm | 110-120 ppm | 120-130 ppm | 130-140ppm | 150-160ppm | 160-170 ppm |
|---|---|---|---|---|---|---|---|---|---|
| | | 5-membered ring<br>aliphatic chain | aliphatic chain | -COR | | aromatic ring | | | -COOR |
| 20-30 ppm | aliphatic chain or 6-membered ring | 2.5-3Å:2.23%<br>4.5-5Å:2.80% | 2.5-3Å:3.10% | | | | | | |
| 30-40 ppm | 5-membered ring aliphatic chain | 1.5-2Å:4.23%<br>2.5-3Å:2.33% | 1.5-2Å:2.63%<br>2.5-3Å:5.80%<br>3.5-4Å:2.20%<br>8-8.5Å:1.37% | | 3.5-4Å:1.83% | 2.5-3Å:1.57%<br>3.5-4Å:2.83% | | | |
| 70-80 ppm | | | | 1.5-2Å:2.10%<br>2.5-3Å:1.60% | | | | | |
| 90-100 ppm | | | | | | | | | 1-1.5Å:2.17% |
| 110-120 ppm | | | | | | 2-2.5Å:2.33%<br>2.5-3Å:1.00% | 2-2.5Å:1.63%<br>5.5-6Å:1.30% | 2-2.5Å:1.53% | |
| 120-130 ppm | Aromatic ring | | | | | 1-1.5Å:5.57%<br>2-2.5Å:5.43%<br>2.5-3Å:5.87% | 1-1.5Å:6.03%<br>2-2.5Å:4.90%<br>2.5-3Å:3.00% | | |
| 130-140 ppm | | | | | | | 5.5-6Å:3.67% | 2.5-3Å:1.13% | 2-2.5Å:3.37% |

The ability of the 3D-QSDAR approach not only to identify important structural fragments but also to determine the optimal distances between the carbon atoms belonging to such fragments is illustrated in Figure 5. These optimal distances and their associated substructural units can be regarded as two-center toxicophores/pharmacophores.

**Comparison with Previously Reported Models.** The EDKB utilized in this study is often used as a benchmark data set to compare the performance of various models. Because of the promiscuity of the ER receptor ($\alpha$ and/or $\beta$) and hence the diversity of the ligands binding it, the ability to build high quality predictive QSAR models has proven a challenging task.

An early attempt to utilize a similar set of compounds from the NCTR data was reported by Shi et al.[40−59] CoMFA and HQSAR models with $R^2$ of 0.91 and 0.76, respectively, were proposed. A leave-many-out (LMO) cross-validation with an average 5-fold $Q^2$ of 0.62 (CoMFA) and 0.57 (HQSAR) was carried out. On average, more than five principal components were used by both modeling techniques. The addition of a phenol indicator improved the predictive power of the models, which when applied to external test sets produced an $R^2_{test}$ ranging from 0.15 (HQSAR) to 0.71 (CoMFA). In a later work, Ghafourian et al.[41−60] reported multiple MLR, PLS, and formal inference-based recursive modeling (FIRM) models utilizing 157 quantum mechanical, graph theoretical, indicator variables and logP descriptors calculated by the TSAR package. Models with $R^2$ for the training set ranging from 0.52 for the PLS model to 0.73 for the MLR were reported. On the basis of EDKB data and CODESSA derived descriptor sets of different size, Marini et al. proposed various PLS counter- and back-propagation neural network models.[42−61] Three component PLS models employing alternatively 281 ($R^2$ = 0.55 and $Q^2_{LOO}$ = 0.41) or 68 ($R^2$ = 0.64 and $Q^2_{LOO}$ = 0.62) descriptors were

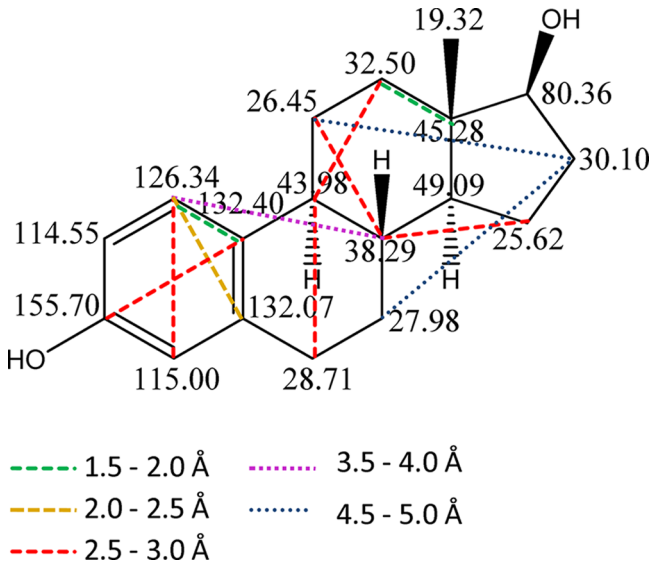



**Figure 5.** An illustration of 3D-QSDAR results superimposed over the structure of 17$\beta$-estradiol. The chemical shifts of the $^{13}C$ atoms are shown. For simplicity, only a few of the carbon-to-carbon distances are visualized.


dx.doi.org/10.1021/ci3001698 | J. Chem. Inf. Model. 2012, 52, 1854−1864

obtained. However, only leave-one-out (LOO) cross-validation was performed, which in itself is an insufficient criterion for quality evaluation.[50] Counter-propagation and back-propagation ANN models with $R^2$ as high as 0.96 and $Q^2_{LOO}$ = 0.84 were also reported. As these are significantly higher than the estimated upper bound on the basis of the $\sigma$ of the EDKB data (see the discussion above), they likely result from overtraining.

Although all of the above models utilized the EDKB data set, direct comparison of their performance is impossible either due to the lack of a test set, different training-to-test set ratios, incompatible cross-validation procedures (LMO vs LOO), or statistical artifacts such as overtraining. Assuming similarity between the hold out test set procedure and the LMO cross-validation, it seems that the unmodified CoMFA[59] provides a slightly higher 5-fold $Q^2$ when compared to $R^2_{test}$ for our best model. However, unlike CoMFA, the 3D-QSDAR approach using energy minimized unbound structures does not necessarily require alignment. As the utilized 3D-QSDAR descriptors are atom specific, the method allows the identification of structural fragments, which when combined with the distances form two-center pharmacophores/toxicophores. Similarly to CoMFA, it provides a way to map back to the original 3D fingerprint space regions with a high contribution to the observed biological effect.

**Outlooks.** Several ways to improve the performance of the 3D-QSDAR approach are currently being investigated. One of the most obvious enhancements would be the addition to the fingerprints of chemical shifts and distances involving heteroatoms such as $^{15}N$, $^{17}O$, $^{19}F$, $^{33}S$, and $^{35}Cl$. Beyond $^{1}H$ and $^{13}C$, the only nucleus with a large base of accurate experimental spectral data is $^{15}N$. The same does not hold true for other biologically important atoms. For example, because of its low sensitivity and natural abundance of only 0.037%, $^{17}O$ is less frequently used[51], and the NMR data for it are still scarce. However, even the addition of only $^{15}N$ spectral data would benefit significantly the applicability of QSDAR to areas where nitrogen atoms in compounds are known to play a role in biological activity such as drug efficacy or toxicity.[25,52]

Currently, the 3D-QSDAR methodology relies on a conformational search analysis aimed to identify the lowest energy conformations in the unbound state, which are further subject to geometry optimization. There are several limitations to this approach: (i) the conformational search is performed in a vacuum, (ii) all structures are optimized independently and molecules with similar binding profiles may assume different conformations, (iii) the obtained conformation may not coincide with the biologically relevant one in the bound state, and (iv) a single conformation does not take into account entropy contributions to the ligand−receptor interaction. Unlike the 2D fingerprint or the topological descriptor-based approaches, the 3D-QSDAR performance depends strongly on the conformation chosen to represent each molecule. Hence, an alignment procedure based on X-ray data or an objective computational approach may substantially improve the quality of the models especially for sets including compounds with multiple degrees of freedom.

Further generalization and extension of the QSDAR technique to four dimensions (4D-QSDAR) would allow the consideration of multiple conformers for each individual compound. A combination of systematic conformational search analysis and energy ranking may be used to select conformers on the basis of their probability of existence as per the Boltzmann distribution. As each compound will be represented by more than one conformer, the binning procedure would require the introduction of a probability factor in the atom-to-atom distances (Z-axis) allowing partial occupancy.

## CONCLUSIONS

The incorporation of $^{13}C$ NMR spectral data and structural distance information has proven to be of value for the modeling of various relationships between biological endpoints/properties and the structural characteristics of compounds. As such, it was incorporated into our 3D-QSDAR approach, which combines NMR and structural data to form unique 3D fingerprints that can be used to discriminate between slightly different compounds. These chemical shift−distance-based fingerprints encode physical and structural information that is related to the electrostatic and steric effects important for the correct description of the interactions in modeling the biological effect of compounds. Our study suggests that for this particular data set the optimal granularity of the 3D grid (10 ppm × 10 ppm × 0.5 Å) depends on the accuracy of the predicted chemical shifts (±4.13 ppm) and the atom-to-atom distances (0.02 Å per bond).

## AUTHOR INFORMATION

### Corresponding Author

* Telephone: (870) 543-7080(R.D.B); (870) 543-7108 (J.G.W.). E-mail: Richard.Beger@fda.hhs.gov (R.D.B.); Jon.Wilkes@fda.hhs.gov (J.G.W.). Fax: (870) 543-7686(R.D.B.; J.G.W.).

### Notes

The views presented in this article are those of the authors and do not necessarily reflect those of the U.S. Food and Drug Administration. No official endorsement is intended nor should be inferred.

The authors declare no competing financial interest.

## REFERENCES

(1) Cramer, R. D. BC(DEF) Parameters. 1. The intrinsic dimensionality of intermolecular interactions in the liquid state. *J. Am. Chem. Soc.* **1980**, *102*, 1837−1849.

(2) Abraham, M. H. Scales of solute hydrogen-bonding: Their construction and application to physicochemical and biochemical processes. *Chem. Soc. Rev.* **1993**, *22*, 73−83.

(3) Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464−477.

(4) Cherkasov, A. Inductive QSAR descriptors. Distinguishing compounds with antibacterial activity by artificial neural networks. *Int. J. Mol. Sci.* **2005**, *6*, 63−86.

(5) Bersuker, I. QSAR without arbitrary descriptors: The electron-conformational method. *J. Comput. Aided Mol. Des.* **2008**, *22*, 423−430.

(6) Burden, F. R.; Polley, M. J.; Winkler, D. A. Toward novel universal descriptors: Charge fingerprints. *J. Chem. Inf. Model.* **2009**, *49*, 710−715.

(7) Andersson, P.; Haglund, P.; Rappe, C.; Tysklind, M. Ultraviolet absorption characteristics and calculated semi-empirical parameters as chemical descriptors in multivariate modelling of polychlorinated biphenyls. *J. Chemometrics* **1996**, *10*, 171−185.

(8) Benigni, R.; Giuliani, A.; Passerini, L. Infrared spectra as chemical descriptors for QSAR models. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 727−730.

(9) Verma, R.; Hansch, C. Use of $^{13}C$ NMR chemical shift as QSAR/QSPR descriptor. *Chem. Rev.* **2011**, *111*, 2865−2899.

(10) Levy, J. B. A quantum chemical and electrostatic study of alkyl and substituted benzenonium cations and related molecules: The

effect of atomic charge distribution on carbocation energy and geometry. *Struct. Chem.* **1999**, *10*, 121−127.
(11) Lin, S.-T.; Lee, C.-C.; Liang, D. W. Analysis of substituent effects on C-13 NMR parameters of substituted arylacetylene derivatives. Linear free energy relationships and PM3 semiempirical calculations. *Tetrahedron* **2000**, *56*, 9619−9623.
(12) Neuvonen, H.; Neuvonen, K.; Koch, A.; Kleinpeter, E.; Pasanen, P. Electron-withdrawing substituents decrease the electrophilicity of the carbonyl carbon. An investigation with the aid of $^{13}C$ NMR chemical shifts, $\nu$(CO) frequency values, charge densities, and isodesmic reactions to interprete substituent effects on reactivity. *J. Org. Chem.* **2002**, *67*, 6995−7003.
(13) Thirunarayanan, G. IR and NMR spectral studies in substituted styryl 1-naphthyl ketones. *Acta Cienc. Indica* **2005**, *31*, 299−304.
(14) Xu, R.; Sim, M. K.; Go, M. L. Synthesis, Antimuscarinic activity and quantitative structure-activity relationship (QSAR) of tropinyl and piperidinyl esters. *Chem. Pharm. Bull.* **1998**, *46*, 231−241.
(15) Bursi, R.; Dao, T.; van Wijk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861−867.
(16) Beger, R.; Freeman, J. P.; Lay, J., Jr.; Wilkes, J.; Miller, D. $^{13}C$ NMR and EI Mass spectrometric data−activity relationship model of estrogen receptor binding. *Toxicol. Appl. Pharmacol.* **2000**, *169*, 17−25.
(17) Beger, R. D.; Freeman, J. P.; Lay, J. O., Jr.; Wilkes, J. G.; Miller, D. W. Use of $^{13}C$ NMR spectrometric data to produce a predictive model of estrogen receptor binding activity. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 219−224.
(18) Beger, R. D.; Wilkes, J. G. Developing $^{13}C$ NMR quantitative spectrometric data−activity relationship (QSDAR) models of steroid binding to the corticosteroid binding globulin. *J. Comput.-Aided Mol. Design.* **2001**, *15*, 659−669.
(19) Beger, R. D.; Buzatu, D. A.; Wilkes, J. G. $^{13}C$ NMR quantitative spectrometric data−activity relationship (QSDAR) models of steroids binding the aromatase enzyme. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1360−1366.
(20) Beger, R. D.; Wilkes, J. G. Models of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls binding affinity to the aryl hydrocarbon receptor developed using $^{13}C$ NMR data. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1322−1329.
(21) Beger, R. D.; Buzatu, D. A.; Wilkes, J. G.; Lay, J. O., Jr. Comparative structural connectivity spectra analysis (CoSCoSA) models of steroid binding to the corticosteroid binding globulin. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1123−1131.
(22) Beger, R. D.; Buzatu, D. A.; Wilkes, J. G. Combining NMR spectral and structural data to form models of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls binding to the AhR. *J. Comput. Aided Mol. Des.* **2002**, *16*, 727−740.
(23) Beger, R. D.; Buzatu, D. A.; Wilkes, J. G. In *Drug Discovery Handbook Vol. 1: Pharmaceutical Development and Research Handbook*, 2nd ed.; Shayne, C. G., Ed.; John Wiley & Sons: Hoboken, NJ, 2005; pp 227−286.
(24) Thirunarayanan, G. IR and NMR spectral studies in substituted styryl 1-naphthyl ketones. *Acta Cienc. Indica,Chem.* **2005**, *31*, 299−304.
(25) Beger, R. D. Computational modeling of biologically active molecules using NMR spectra. *Drug Discovery Today* **2006**, *11*, 429−435.
(26) Ning, Y.-C. *Interpretation of Organic Spectra*; John Wiley and Sons (Asia) Pte Ltd: Singapore, 2011; p 41.
(27) Wilman, D. E.V.; Palmer, B. D.; Denny, W. A. Application of 15N nuclear magnetic resonance spectroscopy to the determination of the stability of aryl nitrogen mustards. *J. Med. Chem.* **1995**, *38*, 2256−2258.
(28) Matter, H.; Schudok, M.; Elshorst, B.; Jacobs, D. M.; Saxena, K.; Kogler, H. QSAR-by-NMR: Quantitative insights into structural determinants for binding affinity by analysis of $^{1}H/^{15}N$ chemical shift differences in MMP-3 ligands. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1779−1783.
(29) Shade, L.; Beger, R. D.; Wilkes, J. G. The use of carbon thirteen nuclear magnetic resonance spectra to predict dioxin and furan binding affinities to the aryl hydrocarbon receptor. *Environ. Toxicol. Chem.* **2003**, *22*, 501−509.
(30) Miller, D. W.; Beger, R.; Lay, J. O., Jr.; Wilkes, J. G.; Freeman, J. P. WO 2001057495 A2 20010809, 2001.
(31) Miller, D. W.; Beger, R.; Lay, J. O., Jr.; Wilkes, J. G.; Freeman, J. P. US 6898533 B1 20050524, 2005.
(32) Beger, R. D.; Wilkes, J. G. US 20030229456 A1 20031211, 2003.
(33) Wang, Y.; Li, Y.; Ding, J.; Wang, Y.; Chang, Y. Prediction of binding affinity for estrogen receptor $\alpha$ modulators using statistical learning approaches. *Mol. Divers.* **2008**, *12*, 93−102.
(34) Abdi, H. Partial least squares regression and projection on latent structure regression (PLS regression). *WIREs Comp Stat.* **2010**, *2*, 97−106.
(35) Swan, S. H.; Elkin, E. P.; Fenster, L. The question of declining sperm density revisited: An analysis of 101 studies published 1934−1996. *Environ. Health Perspect.* **2000**, *108*, 961−966.
(36) Joensen, U. N.; Jorgensen, N.; Rajpert-De Meyts, E.; Skakkebaek, N. E. Testicular dysgenesis syndrome and leydig cell function. *Basic Clin. Pharmacol. Toxicol.* **2008**, *102*, 155−161.
(37) Cleary, M. P.; Grossmann, M. E. Obesity and breast cancer: The estrogen connection. *Endocrinology* **2009**, *150*, 2537−2542.
(38) Diamanti-Kandarakis, E.; Palioura, E.; Kandarakis, S. A.; Koutsilieris, M. The impact of endocrine disruptors on endocrine targets. *Horm. Metab. Res.* **2010**, *42*, 543−552.
(39) *Endocrine Disruptors Knowledge Base*. http://www.fda.gov/ScienceResearch/BioinformaticsTools/EndocrineDisruptorKnowledgebase/ucm136091.htm (accessed 02/14/2012).
(40) (40−59) Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186−195.
(41) (41−60) Ghafourian, T.; Cronin, M. T. The impact of variable selection on the modelling of oestrogenicity. *SAR QSAR Environ. Res.* **2005**, *16*, 171−190.
(42) (42−61) Marini, F.; Roncaglioni, A.; Nović, M. Variable selection and interpretation in structure−affinity correlation modeling of estrogen receptor binders. *J. Chem. Inf. Model.* **2005**, *45*, 1507−1519.
(43) Doweyko, A. M.; Bell, A. R.; Minatelli, J. A.; Relyea, D. I. Quantitative structure−activity relationships for 2-[(phenylmethyl)-sulfonyl]pyridine 1-oxide herbicides. *J. Med. Chem.* **1983**, *26*, 475−478.
(44) *HyperChem 8 Tools for Molecular Modeling*, version 8.0; HyperCube Inc.: Gainesville, FL, 2007.
(45) *ACD/NMR Predictor Release 12.00*, version 12.5; Advanced Chemistry Development: Toronto, Canada, 2011.
(46) Bremser, W. HOSE: A novel substructure code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(47) *R Statistical Package* [Online], version v.2.13.2. www.r-project.org (accessed 02/14/2012).
(48) Lewars, E. G. *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*, 2nd ed.; Kluwer: Boston, MA, 2003; p 413.
(49) Beger, R. D.; Buzatu, D. A.; Wilkes, J. G. Quantitative spectrometric data−activity relationships (QSDAR) models of endocrine disruptor binding activities; CRC Press:Boca Raton, FL, 2009, pp 237−260.
(50) Golbraikh, A.; Tropsha, A. Beware of $q^2$! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.
(51) Boykin, D. W. *$^{17}O$ NMR Spectroscopy in Organic Chemistry*; CRC Press: Boca Raton, FL, 1991; pp 2−22.
(52) Sugimura, T.; Wakabayashi, K.; Nakagama, H.; Nagao, M. Heterocyclic amines: Mutagens/carcinogens produced during cooking of meat and fish. *Cancer Sci.* **2004**, *95*, 290−299.