JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

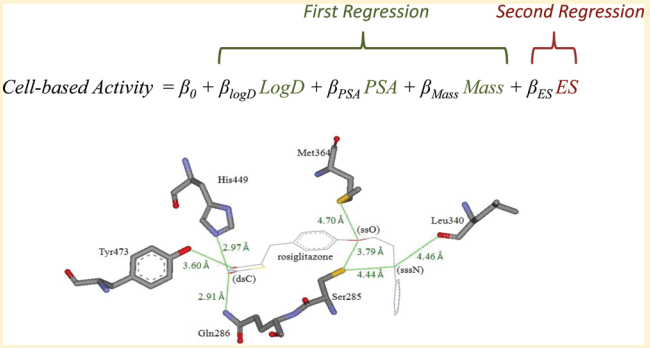
# A Scaffold-Independent Subcellular Event-Based Analysis: Characterization of Significant Structural Modifications

Ying-Ting Lin*[,†,‡] and Guan-Yu Chen[†]

[†]Department of Biotechnology, College of Life Sciences and [‡]Center of Excellence for Environmental Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

Supporting Information

**ABSTRACT:** General and singular subcellular events within the ligand-dependent receptor-mediated cellular response were separated by using the Jurs and the electrotopological state (ES) descriptors, allowing characterization of the significant structural modifications in a given set of collected peroxisome proliferator-activated receptor $\gamma$ (PPAR$\gamma$) agonists. The identified Jurs descriptor is the integrated function of all the general events but is scaffold-dependent. The top captured ES descriptors stand for significant structural modifications, i.e., singular events. To further elucidate the descriptor-event relationship, three biological data sets show that the Jurs descriptor can be further divided into three important descriptors, the log *D*, polar surface area, and shape-like descriptor. The identification of the essential descriptors for general events is the first regression, and the prioritization of all the possible structural modifications of the 46 collected thiazolidinedione PPAR$\gamma$ agonists is the second regression. As results, the top captured ES symbols can correspond to the singular ligand−receptor interactions as highlighted in the X-ray crystallographic image of rosiglitazone−PPAR$\gamma$ complex.




## 1. INTRODUCTION

Ligand-dependent receptor-mediated cellular data (ligand cellular data) have long been considered an inadequate source for the analysis of ligand−receptor interactions, as they contain many confounding factors. To analyze the ligand cellular data, general and singular subcellular events in the ligand-dependent receptor-mediated cellular response were separated by using the Jurs[1] and the electrotopological state (ES) descriptors,[2−4] allowing characterization of the significant structural modifications in given collected peroxisome proliferator-activated receptor gamma (PPAR$\gamma$) agonists; the proposed framework of molecular description is illustrated in Figure 1. In cellular response, some alterations of molecular recognition, e.g., hydrogen-bond formation or deformation, can cause drastic alteration in activity. These particular structural modifications as singular events are considered to be statistical breakdown points for many correct analyses, i.e., outliers of statistic regression, also known as activity cliffs.[5,6] An activity cliff comes from a singular subcellular event. The possible structural modifications in a given set of agonists are predefined here by ES descriptors, and all the ES descriptors can be statistically prioritized through a "second regression". The first regression is to identify the descriptor or descriptors suitable for the representation of general subcellular events, and the second regression is to prioritize the ES descriptors for singular events.

A two-stage regression has been formulated to analyze the collected thiazolidinedione (TZD) PPAR$\gamma$ agonists.[7] One of the resulting outcomes is shown in Figure 2: The descriptor, Jurs_RNCG, was selected from first regression, and the ES descriptor, ES_Count_ssO, was prioritized in order of potency. The former indicates an integrated description of all the possible general subcellular events of the agonists. The latter, acting as the outlier of the first regression, was captured through the statistical prioritization of the second regression. This top-captured ES descriptor also indicates that the significant structural modification of ether linkage (ssO) corresponds to the singular ligand−receptor interaction, which can be shown in the X-ray crystallographic image of the potent ligand−PPAR$\gamma$ complex,[8] also in Figure 2. In this mapping of general and singular subcellular events with Jurs and ES descriptors, one can directly prioritize the potency order of "informative outlier", as a so-called activity cliff,[5,6] from ligand-dependent receptor-mediated cellular data.

Furthermore, in order to elucidate the descriptor-event relationship of general subcellular events, we used three biological data sets to demonstrate that the Jurs descriptor can further be divided into more subtle descriptors. The first data set is a collection of 110 topoisomerase I (TopI) inhibitors,[9,10] which shows log *D* as a most important descriptor in the inhibitor-dependent cellular response. The second data set is composed of 72 analogs[11,12] of raloxifene with both cell-based and cell-free

**Received:** September 23, 2011
**Published:** January 30, 2012

| Overall Molecular Description | | | | Approach | Event-Based |
|---|---|---|---|---|---|
| Significant Structural Modification (ES) | Significant Structural Modification (ES) | Significant Structural Modification (ES) | Significant Structural Modification (ES) | *Second* | Singular Events |
| Integrated Function (Jurs) | | | | *First* | Combination of General Events, except singular event |
| Structure of agonists in a set | | | | Regression | Subcellular Event |

**Figure 1.** Illustrations of overall molecular description in the subcellular event-based analysis of singular events, significant structural modifications (ES descriptors), for a set of agonists. A two-stage regression is employed. The descriptor, Jurs, is the integrated function representing the combination of general subcellular events.

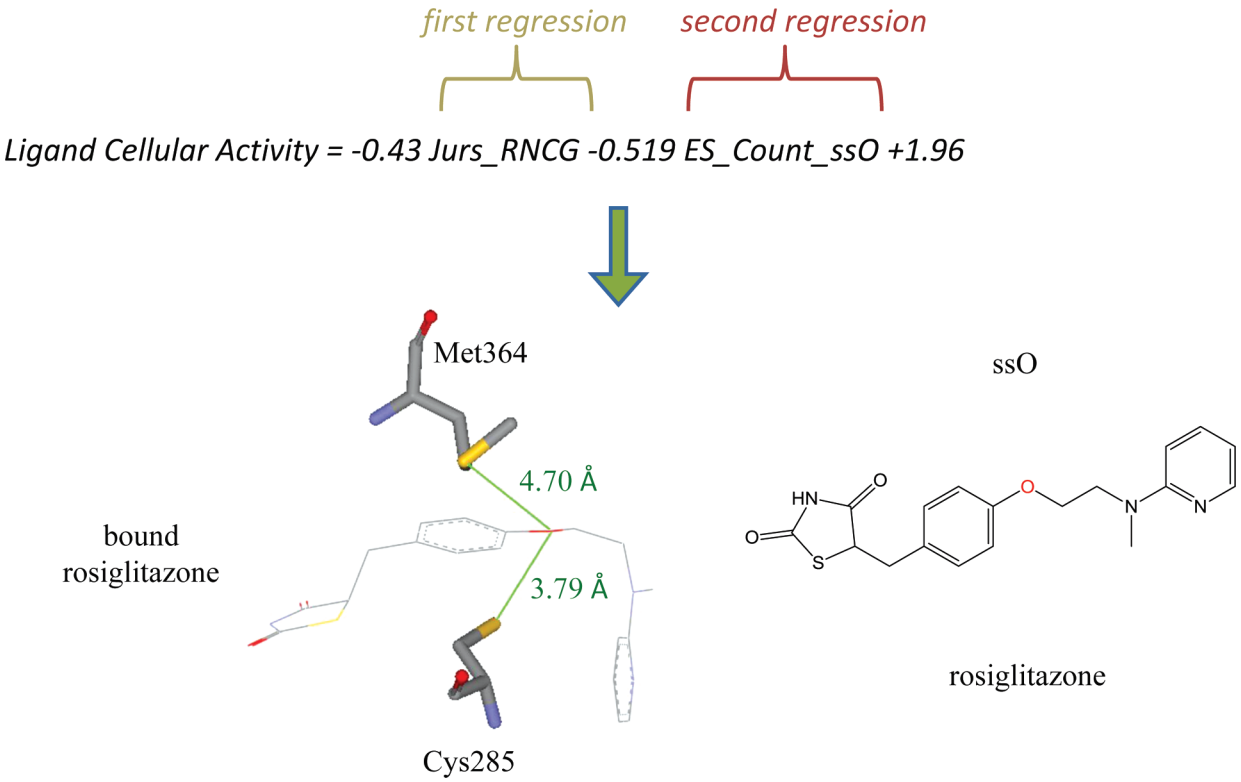



**Figure 2.** One of the resulting equations results from a two-stage regression to the TZD PPARγ agonists. The Jurs_RNCG is from first regression and the descriptor ES_Count_ssO from second regression. The ES symbol, ssO, indicates singular ligand−receptor interactions, as shown in the X-ray crystallographic image (PDB code: 2PRG), in which the singular interactions are between the rosiglitazone tyrosine oxygen (ssO) and the sulfurs of the PPARγ residues Cys285 and Met364; the respective distances (in green) are 3.79 and 4.70 Å. The chemical structure of rosiglitazone is at the bottom right, and the oxygen is in red.

data, which shows polar surface area (PSA) the most important factor counting for the difference between cell-based and cell-free responses. The third data set employs 46 TZD PPARγ agonists,[7,13,14] which shows that the Jurs_RNCG descriptor[1] can be further divided into the three descriptors, log *D*, PSA, and a shape-like descriptor. These identifications of the three former subtle descriptors for all the general subcellular events comprised of the first regression. The significant structural modifications in the 46 collected TZD PPARγ agonists[7,13,14] as informative outliers are the singular subcellular events. Through second regression, all the ES descriptors of the agonists were statistically prioritized. As a result, the top captured ES descriptors can also find their corresponding singular interactions of molecular recognition, which were highlighted in the X-ray crystallographic image of the rosiglitazone−PPARγ complex.

## 2. THE ROLE OF LOG *D*

**2.1. TopI Inhibitors.** A total of 110 nonredundant TopI inhibitors had been created by a single lab in a decade-long synthesis endeavor[9,10] spanning from 1999 to 2008. The ligand (inhibitor) cellular phenomenon is the cell growth inhibition caused by the inhibition of TopI (a protein involved in DNA topology modification). These inhibitors were tested in 55 cell lines obtained from the National Cancer Institute (NCI) that represented a range of cancer cell types, including lung (HOP-62), colon (HCT-116), central nervous system (SF-539), melanoma (UACC-62), ovarian (OVCAR-3), renal (SN12C),


dx.doi.org/10.1021/ci200540y | J. Chem. Inf. Model. 2012, 52, 506−514

prostate (DU-145), and breast (MDA-MB-435) cancers. The 110 TopI inhibitors used were selected for having definite experimental values for GI50, the concentration producing 50% growth inhibition, for the above 8 cell lines explicitly listed in the literature. The inhibitors which had indefinite experimental values, i.e., with the mathematical greater than or less than symbols, were excluded from analysis. The mean graph midpoint (MGM) values were averaged from the GI50 values over all 55 cell lines, and the details were described in the experimental sections of the related papers.[9,10] Overall, the MGM value represents overall cell growth inhibition by a TopI inhibitor across the NCI cell lines. The ligand-dependent cellular phenomenon is the cell growth inhibition resulting from inhibition of the TopI enzyme. All 110 molecular structures and the experimental inhibition values expressed as $-\log_{10}(\mathrm{MGM})$, i.e., pMGM, are listed in Table S2.1, Supporting Information. All inhibitor structures were geometrically optimized with respect to energy by the molecular mechanism optimization program MMF94 in the ChemBio3D suite from the ChemBioOffice package.[15]

**2.2. Most Important Descriptor in Ligand Cellular Phenomenon.** The selecting one descriptor out mechanism (SODO mechanism) aims to use a suitable working equation to measure any descriptor in order of potency. For the purpose of selection, a simple linear equation is used. A measure of the correlation between descriptor and biological activity in a working equation is the least-squares fit. When Pearson's correlation coefficient is used as the fit, $r^2$ is utilized to prioritize the potential descriptors in order of rank. Note that in the SODO mechanism, we use statistical quantity as a measure of descriptor potency during this stage of descriptor-event mapping. For the inhibitor cellular phenomenon of TopI inhibitors, we will choose the single most important descriptor in this inhibitor cellular data. Mathematically, the dependent variable, $y$, is the ligand-dependent enzyme-mediated cellular phenomenon. We seek the most representative descriptor, $x_{ch}$, as the primary factor for the ligand cellular activity. The resulting equation is as follows:

$$y = \beta_0 + \beta_{ch} x_{ch} \quad (1)$$

where $y$ is the dependent variable standing for the ligand-dependent enzyme-mediated cellular phenomenon, $x_{ch}$ is the descriptor to be chosen, and $\beta_0$ and $\beta_{ch}$ are the regression coefficients after the least-squares fit. Once the activity values for the ligand cellular phenomena against a given set of molecules are available, the correlation fit $r^2$ can be obtained for each descriptor from a specific descriptor pool. Here, more than 500 descriptors of eminent classes were employed. All descriptors in the working equation with correlation fits are prioritized in rank order of the regression fit. All calculation of descriptors was performed using the Discovery Studio 2.1 QSAR module.[16] The regression fitting for each descriptor in the working equation and the Pearson's coefficient were performed using R 2.11.0.[17]

**2.3. Observations and Results.** The MGM values were the average inhibitions for all 55 NCI cancer cell lines, representing lung, colon, central nervous system (CNS), melanoma, ovarian, renal, prostate, breast, and other cancers. The average value means that no cancer type-dependent subcellular event dominates and that the MGM is the combination of the common subcellular biochemical events of each TopI inhibitor. Therefore, these inhibitor cellular data are the perfect outcome of a combination of all possible *general* subcellular events, which may include solubility, membrane transport, cytosol mobility, general degree of agonism, and general molecular recognition. The first question in the framework of a subcellular event-based approach is which of these general subcellular events can be dominant. To answer this question, we used the SODO mechanism to pick the single most representative descriptor from a large data set containing 110 collective molecular structures of TopI inhibitors.

The results, shown in Table S2.3, Supporting Information, ranked the descriptors Molecular_Solubility and log $D$ as most important. Solubility is the solubility in water, and log $D$ is the partition coefficient. In fact, these 2 are mutually correlated, and log $D$ dominates in a smaller data set. As a general illustration of the analysis, we will discuss the log $D$. Log $D$ is the ratio of concentrations of a small molecule in the two phases of a mixture of two immiscible solvents (octanol and water) at equilibrium. When an inhibitor is tested in cells, some of the subcellular events are related to the log $D$. The first general event related to log $D$ is the solvent solubility, i.e., the ability of a molecule to cross the gas–solvent interface. The second and third general events are the solvent–membrane and membrane–cytosol interfaces. However, these two latter events can be more precisely represented by the PSA, as discussed later. The last event is the solvation–desolvation prior to inhibitor binding.[18] In order to emphasize the general description of many related subcellular interface events, we used log $D$ as the most crucial factor in an inhibitor cellular data set.

In addition, when the agonist set becomes larger and more diversified, the structural modifications tend not to dominate under these conditions. Here in Table 1, the dominant descriptors detected by the SODO mechanism against the increasing data set size are listed. Inhibitors of each data set size were randomly selected 500 times. Each set of the same size was put into the SODO mechanism, and the best of the 500 first-selected descriptors was listed. We can clearly see that with data set sizes of 10 and 20 inhibitors, the topological shape indices (CHI_V_2, CHI_V_1)[19,20] are dominant and the correlation coefficients ($r^2$ values) are 0.99 and 0.68. With data set sizes of 30 and 40 inhibitors, the ES descriptor (ES_Count_sNH2) is dominant, and the correlation coefficients ($r^2$ values) are 0.58 and 0.44. With data set sizes of more than 50 inhibitors, the partition coefficient (log $D$) is dominant, although the correlation coefficients ($r^2$ values) decreased. When the data set size is 110 inhibitors, the Molecular_Solubility is dominant.

**Table 1. Dominant Descriptors Using Different Data Set Sizes from the TopI Inhibitor Molecular System**

| data size | dominant descriptor | $r^2$ |
|---|---|---|
| 10 | CHI_V_2 | 0.99 |
| 20 | CHI_V_1 | 0.68 |
| 30 | ES_Count_sNH2 | 0.58 |
| 40 | ES_Count_sNH2 | 0.44 |
| 50 | log $D$ | 0.39 |
| 60 | log $D$ | 0.31 |
| 70 | log $D$ | 0.32 |
| 80 | log $D$ | 0.30 |
| 90 | log $D$ | 0.25 |
| 100 | log $D$ | 0.18 |
| 110 | Molecular_Solubility | 0.11 |

As mentioned above, Molecular_Solubility and log $D$ are mutually correlated, and the effect of solvation–desolvation


dx.doi.org/10.1021/ci200540y | J. Chem. Inf. Model. 2012, 52, 506–514

becomes dominant, although with decreasing $r^2$, when the size of the data set is larger than 50 inhibitors. Note that all of the dominant descriptors for the various data set sizes listed in Table 1 are ranked by the SODO mechanism in first place out of more than 500 various descriptors. The dominant descriptor of the inhibitors is the log $D$ (or Molecular_Solubility), revealing the most crucial subcellular biochemical event in the 55 NCI cell lines. Thus, in other words, the property log $D$ is the most important and necessary descriptor for consideration underlying the ligand-dependent cellular data.

## 3. THE ROLE OF PSA

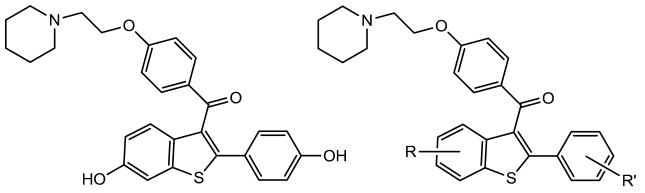
**3.1. Estrogen Receptor (ER) Antagonists.** To determine the most crucial factor between cell-free and cell-based systems within a ligand cellular data, we compared 72 synthetic analogs[11,12] of raloxifene in two types of in vitro biological assays, the cell-free radioligand binding assay and the cell-based antiproliferation assay. The cell-free assay measured ER$\alpha$-binding affinities determined by displacement of bound radiolabeled [$^3$H]-17$\beta$-estradiol from MCF-7 cell lysates. The cell-based assay assessed the antagonism of tested antagonists in the MCF-7 cell line by measuring the inhibition of cell proliferation, which cell growth is induced by $10^{-11}$ M 17$\beta$-estradiol. The results of both experiments were reported in nanomolar units as $IC_{50}$ and $EC_{50}$, respectively. The relative binding affinity (RBA) refers to the comparison of binding of a given raloxifene analog to that of 2-$\beta$-estradiol. The RIA refers to the relative inhibitory activity compared that of raloxifene. Furthermore, the abbreviation LRBA represents $\log_{10}$ of the RBA value, and the LRIA is the $\log_{10}$ of the RIA. The final mathematical expressions of LRBA and LRIA from cell-free and cell-based assays are shown in Figure 3. The chemical structures of the raloxifene analogs

$$LRBA = \log \frac{IC_{50}(E2)}{IC_{50}(ana\log)} \quad \text{(data from cell-free assay)}$$

$$LRIA = \log \frac{IC_{50}(Ral)}{IC_{50}(ana\log)} \quad \text{(data from cell-based assay)}$$

**Figure 3.** Formulas for the LRBA and the LRIA. LRBA was determined from the cell-free assay and LRIA from the cell-based assay.

were modified around the benzothiophene core as shown in Figure 4. Analogs that were listed in the literature without




Raloxifene Analogs modified to benzothiophene core

**Figure 4.** Raloxifene analogs are from the modifications of benzothiophene core of raloxifene.

definite assay quantities were not used in the analysis in order to avoid introducing uncertainties. Twelve analogs were enantiomeric, and we assumed that only one enantiomer possessed binding activity, thus causing the active enantiomer to have half of the $IC_{50}$ value. This would then require that the RBA be multiplied by 2 due to the chirality. All of the chemical structures, LRIAs, and LRBAs of the 72 raloxifene analogs are listed in Table S3.1, Supporting Information.

**3.2. Most Important Descriptor between Cell-Based and Cell-Free Phenomenon.** The SODO mechanism prioritizes given descriptors in order of potency with a suitable working equation. To identify and better understand the most crucial factor in cell-drug interaction responsible for the discrepancies between cognate cell-based data and cell-free data for ligand binding, a simple linear three-variable equation was used. The working equation in the context of "cell-based phenomena = $a$(cell-free phenomena) + $b$(descriptor to be chosen) + $c$" is as follows:

$$y_{\text{cell-based}} = \beta_0 + \beta_{\text{cell-free}} x_{\text{cell-free}} + \beta_{\text{ch}} x_{\text{ch}} \quad (2)$$

where $y_{\text{cell-based}}$ is the cell-based antiproliferation data, $x_{\text{cell-free}}$ is the cell-free radioligand binding data, $x_{\text{ch}}$ is the descriptor to be chosen, and $\beta_0$, $\beta_{\text{cell-free}}$, and $\beta_{\text{ch}}$ are the regression coefficients after the least-squares fit. Thus, LRIA is $y_{\text{cell-based}}$, and LRBA is $x_{\text{cell-free}}$. Once both the cognate cell-based and cell-free ligand-binding data for a given set of inhibitors are available, the correlation fit ($r^2$) for each descriptor from a descriptor pool[16] can be calculated. The SODO mechanism was thus applied to more than 500 descriptor equations against the data obtained for the 72 raloxifene analogs. All compounds were geometrically optimized with respect to energy by the molecular mechanism MMF97 program from the ChemBio3D suite of the ChemBioOffice software package.[15]

**3.3. Observations and Results.** The four descriptors with best fits to the working equation are shown in Table 2. The

**Table 2. First Four Ranking Descriptors and Their Fitting Coefficients from the TopI Inhibitor Molecular System**

| rank | descriptors | $r^2$ |
|---|---|---|
| 1 | Molecular_PolarSurfaceArea | 0.55 |
| 2 | Molecular_FractionalPolarSurfaceArea | 0.54 |
| 3 | S_Count | 0.52 |
| 4 | Molecular_PolarSASA | 0.52 |

largest correlation coefficient for the working equation was $r^2$ = 0.55 for the molecular descriptor Molecular_PolarSurfaceArea (PSA).[21] Molecular polar surface area is the surface area covered by polar atoms and was interpreted as representative of passive molecular transport through membranes. Since PSA emerged as the most significant descriptor from a total of approximately 500 different descriptors, we considered the membrane transport of an antagonist the most crucial subcellular event accounting for the discrepancy between cell-based and cell-free data. We further analyzed the following three top descriptors in the working equation with highest correlation fits and assessed their mutual correlations. These descriptors were Molecular_FractionalPolarSurfaceArea (FPSA),[21] S_Count, and Molecular_PolarSASA (PSASA).[21] FPSA, fractional polar surface area, is the ratio of the polar surface area to the total surface area. PSASA, polar solvent accessible surface area, is the total polar solvent accessible surface area for a molecule. S_Count is the number of sulfur atoms in a molecule. We neglect S_Count here as it is considered to be a structural modification rather than an effect of the general subcellular events we are attempting to identify.

In Table 3, a correlation-coefficient (**r**, not $r^2$) matrix is calculated in order to examine whether the four indicated descriptors are mutually independent and to determine their


dx.doi.org/10.1021/ci200540y | J. Chem. Inf. Model. 2012, 52, 506–514

Table 3. Correlation Coefficient Matrix[a]

| | LRIA | LRBA | PSA | FPSA | SC | PSASA |
|---|---|---|---|---|---|---|
| LRIA | 1.000 | | | | | |
| LRBA | 0.687 | 1.000 | | | | |
| PSA | −0.284 | 0.004 | 1.000 | | | |
| FPSA | −0.126 | 0.196 | 0.910 | 1.000 | | |
| SC | −0.225 | 0.005 | 0.313 | 0.321 | 1.000 | |
| PSASA | −0.237 | −0.013 | 0.937 | 0.884 | 0.262 | 1.000 |

[a]LRIA: cell-based Data; LRBA: cell-free Data; PSA: Molecular_PolarSurfaceArea; FPSA: Molecular_FractionalPolarSurfaceArea; SC: S_Count; and PSASA: Molecular_PolarSASA.

contributions to the LRIA. First, we infer from the first coefficient column in Table 3 that LRBA is highly correlated with LRIA. It is reasonable to interpret this as an indication that the binding affinity of the raloxifene analogs plays the dominant role in the growth inhibition of MCF-7 cells. On the other hand, we also infer that the best four descriptors shown in the first column of Table 3 display relatively small negative correlations with LRIA. The minor contribution to LRIA for each single descriptor indicates that LRIA is a phenomenon of combined effects to which ligand binding is one of the most important contributors. Therefore, the goal we have attempted to achieve in this portion of the analysis is to determine which parameter, besides ligand binding, of the cell-based phenomenon can be considered to be the crucial subcellular factor. The four descriptors in the second column of Table 3 exhibit almost no correlations with LRBA (the highest value being 0.196), with three of the correlation coefficients falling within ±0.013. This means that the four best descriptors can be regarded as factors that are independent of molecular recognition. In the remaining four columns of Table 3, PSA, FPSA, and PSASA are highly mutually correlated. The correlation coefficients are 0.910 between PSA and FPSA, 0.937 between PSA and PSASA, and 0.884 between FPSA and PSASA. The fact that three of the four best descriptors are closely related indicates a common mechanism of drug−cell interaction. Obviously, this points to the passive transport through cell membrane. This membrane event acts as the most important subcellular factor of these raloxifene analogs in the MCF-7 cell line when counting for the difference between cell-based and cell-free data. Thus, the PSA is the most representative descriptor for the crucial subcellular event of molecular membrane transport.

## 4. DIVISION OF THE JURS DESCRIPTOR INTO LOG *D*, PSA, AND THE SHAPE-LIKE DESCRIPTOR

**4.1. TZD PPARγ Agonists.** A total of 46 compounds were collected with the TZD moiety,[7,13,14] an important class of synthetic PPARγ agonists. PPARγ[7] is a ligand-activated transcription factor belonging to the nuclear hormone family. Its biological functions are involved in the regulation of lipid and glucose storage and catabolism, and it is an established biological target for drug discovery. The PPARγ agonists bind to the receptor as parts of the transactivation machinery that activates the biological response, and this activity is usually detected on a cellular basis. The indeterminate and uncertain efficacy concentration ($EC_{50}$) values were excluded. The $EC_{50}$ values were expressed as negative $\log_{10}$ values before modeling. The sources of the agonists and their log values are listed in Table S4.1, Supporting Information. The most important descriptor of the 46 PPARγ agonists with the TZD moiety is the Jurs_RNCG. RNCG means the relative negative charge, the quantity of the charge of the most negative atom divided by the summation of the total negative charge in a molecule. Note that the Jurs_RNCG thus can be an integrated function of all the general subcellular events for the 46 collected TZD PPARγ agonists.

**4.2. Three Contributors to Jurs_RNCG.** Log *D* and PSA, the molecular descriptors identified from the above observations and results, can correspond to two crucial subcellular events: molecular solvation−desolvation and membrane transport. The Jurs_RNCG is the integrated description of all the possible general subcellular events for these TZD PPARγ agonists. To identify the third contributor to Jurs_RNCG, a simple linear four-variable equation was used. The working equation is as follows:

$$\text{JursRNCG} = \beta_0 + \beta_{\log D}\log D + \beta_{\text{PSA}}\text{PSA} + \beta_{\text{ch}}x_{\text{ch}} \quad (3)$$

where Jurs_RNCG is the calculated descriptor of 46 TZD PPARγ agonists. RNCG means the relative negative charge, the quantity of the charge of the most negative atom divided by the sum of the total negative charge of a molecule, log *D* is the calculated partition coefficient, PSA is the calculated polar surface area, $x_{\text{ch}}$ is the descriptor to be chosen, and $\beta_0$, $\beta_{\log D}$, $\beta_{\text{PSA}}$, and $\beta_{\text{ch}}$ are the regression coefficients after the least-squares fit. The chosen descriptor is selected from a set of more than 500 molecular descriptors. The least-squares fitting was performed using R 2.11.0[17] and was applied to the descriptor equations against the calculated data on the TZD PPARγ agonists. Prior to descriptor calculation, the geometries of the molecular structures were optimized using the molecular mechanics program MMF97.[15]

**4.3. Observations and Results.** Table 4 lists the top 16 equations after the SODO selection. There, the 16 top selected descriptors are indices of topological shape (Kappa_1, Kappa_1_AM, CHI_V_0, CHI_0), electronic energy (Electronic_Energy, Total_Energy), electric multipole moment (Dipole_mag, Mean_Polarizability, Apol), molecular weight (MOLWEIGHT Molecular_Mass, Molecular_Weight, Organic_Count), and the molecular surface (Molecular_SASA, Molecular_SurfaceArea). In eq 3, log *D* represents the solvation−desolvation effect and PSA represents membrane transport in the subcellular event-based approach. Among them, the solubility and cytosol mobility are apparently related to the log *D* and the membrane transport to PSA. The general degree of agonism and general ligand−receptor interaction remains as the subcellular events possibly contributing to Jurs_RNCG. One observation is that the Dipole_mag in Table 4 has a significant impact on the log *D* when the correlation coefficient of log *D* changes from the average to the lowest, −0.006, and on the PSA when the correlation coefficient of PSA changes from the average to the lowest, −0.0002. Therefore, except for the electric moment or multipole moment, through observations of Table 4 we infer that the molecular surface, molecular weight, or even electronic energy is the descriptor for the description of the contact shape of molecular recognition. Therefore, based on the above observations of the descriptor-event relationships, the Jurs_RNCG of TZD PPARγ agonists can be or is postulated to be divided into three important descriptors: log *D*, PSA, and the shape-like descriptor.

Table 4. The 16 Top Equations from the SODO Mechanism in the Molecular System of the 46 TZD PPARγ Agonists

| 16 top equations | $r^2$ |
|---|---|
| Jurs_RNCG = +0.020log*D* + 0.0012PSA − 0.012_Kappa1 + 0.21 | 0.75 |
| Jurs_RNCG = +0.021log*D* + 0.0014PSA − 0.013Kappa_1_AM + 0.19 | 0.75 |
| Jurs_RNCG = +0.019log*D* + 0.0012PSA + 0.000005Electronic_Energy + 0.12[a] | 0.74 |
| Jurs_RNCG = +0.024log*D* + 0.0012PSA − 0.018CHI_V_0 + 0.23 | 0.73 |
| Jurs_RNCG = −0.006log*D* − 0.0002PSA + 0.004Dipole_mag + 0.18 | 0.72 |
| Jurs_RNCG = +0.022log*D* + 0.0013PSA − 0.0007MOLWEIGHT + 0.22[a] | 0.71 |
| Jurs_RNCG = +0.029log*D* + 0.0014PSA − 0.008Mean_Polarizability + 0.23[a] | 0.71 |
| Jurs_RNCG = +0.022log*D* + 0.0013PSA − 0.0007Molecular_Mass + 0.22 | 0.71 |
| Jurs_RNCG = +0.022log*D* + 0.0013PSA − 0.0007Molecular_Weight + 0.22 | 0.71 |
| Jurs_RNCG = +0.024log*D* + 0.0013PSA − 0.0007Molecular_SAVol + 0.28 | 0.71 |
| Jurs_RNCG = +0.020log*D* + 0.0011PSA − 0.014CHI_0 + 0.24 | 0.71 |
| Jurs_RNCG = +0.026log*D* + 0.0013PSA − 0.0006Molecular_SASA + 0.29 | 0.71 |
| Jurs_RNCG = +0.029log*D* + 0.0019PSA − 0.001Molecular_SurfaceArea + 0.21 | 0.70 |
| Jurs_RNCG = +0.019log*D* + 0.0009PSA − 0.000014Apol + 0.21 | 0.67 |
| Jurs_RNCG = +0.015log*D* + 0.0009PSA + 0.00005Total_Energy + 0.23[a] | 0.66 |
| Jurs_RNCG = +0.021log*D* + 0.0010PSA − 0.010Organic_Count + 0.23 | 0.66 |

[a]These descriptors, Total_Energy, Polarizability, MOLWEIGHT, and Electronic_Energy, were calculated and derived from the semiempirical VAMP/AM1 quantum-chemical wave function.,.

## 5. CHARACTERIZATION OF SIGNIFICANT STRUCTURAL MODIFICATIONS THROUGH SECOND REGRESSION

**5.1. Indentification of Log *D*, PSA, and Mass as First Regression.** In order to understand the descriptor-event relationship of the subcellular event-based QSAR analysis, we have demonstrated that the Jurs_RNCG can be divided into the log *D*, PSA, and shape-like descriptors. Among the subcellular events, log *D* is the general description of solvation and desolvation, PSA is the description of membrane transport, and the shape-like descriptor is the description of general ligand−receptor interaction. In other words, the Jurs_RNCG for the molecular system of PPARγ agonists is dependent on the log *D*, PSA, and the shape-like descriptor. This means that the Jurs_RNCG is dependent on the general molecular scaffold, i.e., "general shape" of ligand−receptor binding. In this analysis, the shape-like descriptors for the general binding description can be many, as demonstrated in Section 4.3. However, to avoid fitting with singular events, we employ the descriptor of molecular mass, mass, as the description of general molecular recognition. Put together, the log *D*, PSA, and mass can suitably represent the effect of all the general subcellular events in cells.

**5.2. Statistical Prioritization of ES Descriptors as Second Regression.** With the use of the descriptor of molecular mass, a scaffold-independent event-based working equation is designed here as follows:

$$y_{\text{cellular}} = \beta_0 + \beta_{\log D}\log D + \beta_{\text{PSA}}PSA + \beta_{\text{Mass}}\text{Mass} + \beta_{\text{ES}}\text{ES} \quad (4A)$$

where $y_{\text{cellular}}$ is the ligand cellular data, log *D* is the calculated molecular partition coefficient, PSA is the calculated molecular polar surface area, mass is the molecular mass representing the general ligand−receptor interaction, ES is a ES descriptor representing a predefined structural modification. $\beta_0$, $\beta_{\log D}$, $\beta_{\text{PSA}}$, $\beta_{\text{mass}}$, and $\beta_{ES}$ are the regression coefficients after the least-squares fit. Log *D*, PSA, and mass were identified for general events, and all the ES descriptors are going to be prioritized for singular events.

Through first regression, log *D*, PSA, and mass are identified, and through second regression, the structural modifications of ES descriptors in the 46 collected TZD PPARγ agonists were statistically prioritized by using eq 4A. Figure 5 illustrated the overall concept of this approach. The descriptors, log *D*, PSA, and a shape-like descriptor, are scaffold-independent, representing the combination of general subcellular events. The significant structural modifications, indicated by the top captured ES descriptors, can directly correspond to the singular interactions of molecular recognition.

**5.3. Ordering of ES Descriptors with Different Representation of General Events.** Through the separation of general and singular events as illustrated in Figure 1, Jurs_RNCG is taken as the integrated function of all possible subcellular general events. The corresponding working equation is as follows:

$$y_{\text{cellular}} = \beta_0 + \beta_{\text{Jurs\_RNCG}}\,\text{Jurs\_RNCG} + \beta_{\text{ES}}x_{\text{ES}} \quad (4B)$$

where $y_{\text{cellular}}$ is the ligand cellular data, Jurs_RNCG is the calculated Jurs descriptor, and $x_{\text{ES}}$ is the descriptor to be chosen from the ES descriptors for the subcellular structural modifications.

For examination of the ordering of ES descriptors with different representation of all possible general events, all ES descriptors of the 46 TZD PPARγ agonists were prioritized with eq 4B. The resulting 14 top ES descriptors are listed in order in the first column of Table 5. The top ES symbol, ssO, which indicates the singular ligand−receptor interaction, has been correlated with the crystallographic image, as shown in

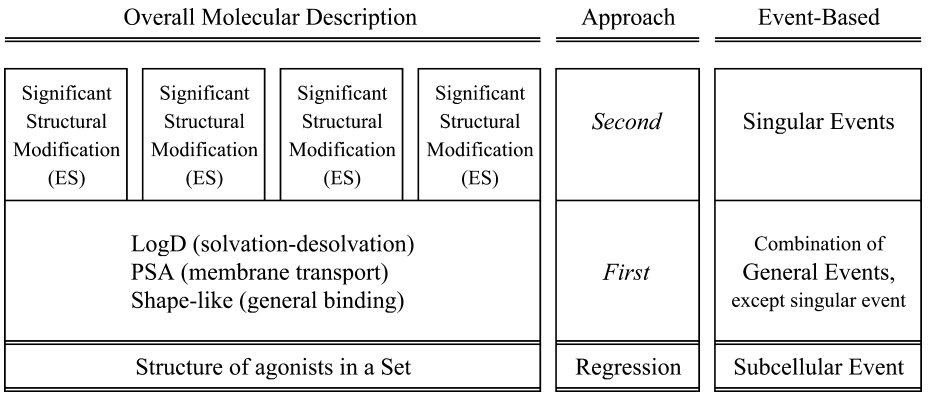



**Figure 5.** Illustrations of overall molecular description in the subcellular event-based analysis of singular events, significant structural modifications (ES descriptors), for a set of agonists. A two-stage regression is employed. The descriptors, log *D*, PSA, and a shape-like descriptor are scaffold-independent, representing for the combination of general subcellular events.

**Table 5. Top Captured ES Descriptors[a] using Eqs 4B and 4A**

| equation 4B | equation 4A |
|---|---|
| ES_Count_ssO | ES_Count_ssO |
| ES_Sum_ssO | ES_Count_dssC(−)[b] |
| ES_Sum_sssN | ES_Sum_ssO |
| ES_Count_sssN | ES_Count_sssN |
| ES_Count_ssCH2 | ES_Sum_sssN |
| ES_Sum_aaO | ES_Sum_aaO |
| ES_Count_aaO | ES_Count_aaO |
| ES_Count_sCH3 | ES_Sum_aaN |
| ES_Sum_aaN | ES_Count_aaN |
| ES_Count_aaN | ES_Count_ssCH2 |
| ES_Count_dsCH | ES_Count_dsCH |
| ES_Sum_dsCH | ES_Count_sCH3 |
| ES_Sum_ssS | ES_Sum_dsCH |
| ES_Count_ssS | ES_Sum_ssS |
| | ES_Count_ssS |

[a]Prioritized through second regression using eqs 4B and 4A, respectively. [b]ES_Count_dssC is a negative structural modification.

Figure 2. Next, with eq 4A, the resulting 14 top ES descriptors are in the second column of Table 5. Comparison of the two columns of the table shows that one additional, important ES symbol, dssC, shows up in the second column, at the same time the rest of ES symbols are in the same order as those in the first column. In ES terminology, dssC is the symbol of a carbonyl carbon, with two single-bond linkages and one double bond. (The "d" in dssC stands for "double bond" and the "s" for "single bond".) Every TZD moiety has two carbonyl groups, and each TZD agonist has at least two counts for this value. The minus sign of the dssC indicates that the addition of more carbonyl carbons to the TZD PPARγ agonists can drastically decrease activity. In other words, this tendency of the ES_Count_dssC means that the two carbonyl carbons of TZD indicated by dssC are significant; the ordering is just after the ES symbol, ssO.

Notably, the ES_Count_dssC ranks in the 45th place when eq 4B is used but ranks in the second place when eq 4A is used. This means that the ES_Count_dssC originally was absorbed into the Jurs descriptor as a common molecular scaffold when using eq 4B but becomes a significant structural modification when using eq 4A. An important observation is thus that the lack-of-modification TZD moiety of the 46 PPARγ agonists tends to be absorbed into the Jurs_RNCG as part of the general ligand−receptor interaction. Furthermore, when the general "shape" of the ligand−receptor interaction is represented by molecular mass in the division of the Jurs descriptor into log *D*, PSA, and mass, the significant structural moiety (TZD) previously absorbed in the Jurs_RNCG re-emerges. This shows an advantage of using molecular mass to represent the general ligand−receptor interaction when the Jurs_RNCG divided into log *D*, PSA, and mass. As resulting outcomes, the top three ES symbols (ssO, dssC, and sssN) form a pharmacophore of three-point features with a general shape for this molecular recognition, which were directly extracted from the ligand-dependent receptor-mediated cellular data of the 46 TZD PPARγ agonists, through second regression.

## 6. REAL CORRESPONDENCE AS VALIDATION

There is no regression method currently used for capturing activity cliffs in a given set because activity cliff itself acts as the outlier of regression.[5] Consequently, there is no currently available regression validation method for the captured activity cliffs as well. We have conducted the second regression to capture the ES descriptors, which stands for the outliers of first regression. The second regression for capturing the informative outliers may reflect the real physical situation. Ranking order can serve as a measure of confidence level for each ES descriptor, and the top captured ES symbols reflect the singular ligand−receptor interaction. Therefore, one can examine their actual physical correspondences as validation.

For example, the ES symbols ssO, dssC, and sssN prioritized in the leading places using eq 4A (Table 5, second column). In particular, dssC represents the carbonyl carbon in TZD as discussed above. In Figure 6, the bound crystallographic structure of rosiglitazone (PDB code: 2PRG)[8] is used to examine the correspondences. Rosiglitazone is a PPARγ drug, a very important full agonist. And the significant interactions between the agonist oxygen atom, indicated by ssO, with the sulfurs of the two specific residues Cys285 and Met364 are at the distances 3.79 and 4.70 Å, respectively. The distances of the significant interactions between the two oxygen atoms of the two agonist carbonyl groups (dssC) with the His449 nitrogen atom, Gln286 nitrogen atom, and Tyr473 oxygen atom are 2.97, 2.91, and 3.60 Å, respectively. The nitrogen atom of the rosiglitazone trialkylamine (sssN) extends by 4.44 Å to reach the Cys285 sulfur atom and by 4.46 Å to reach the backbone oxygen of Leu340. Therefore, the top-ranked ES symbols

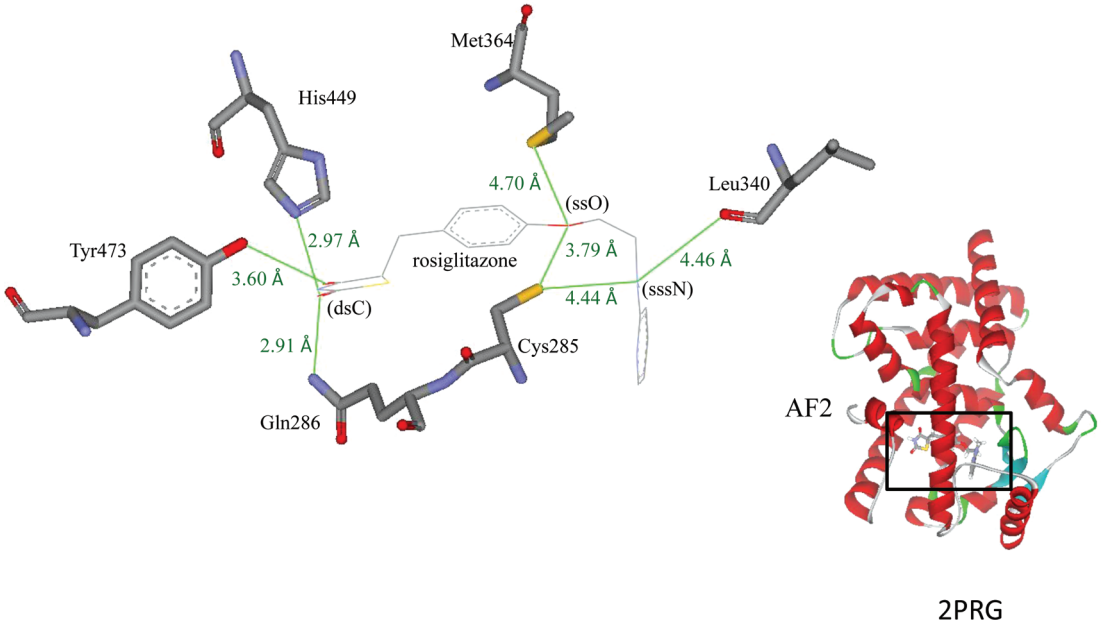



**Figure 6.** Illustration of the significant structural modifications of the ES symbols ssO, dssC, and sssN. The X-ray crystallographic image of bound rosiglitazone (PDB code: 2PRG) shows the singular interactions of the rosiglitazone tyrosine oxygen, indicated by ssO, with the sulfurs of the two specific residues Cys285 and Met364 at distances of 3.79 and 4.70 Å, respectively. The distances of the singular interactions between the two TZD carbonyl oxygen, dssC, with the His449 nitrogen, Gln286 nitrogen, and Tyr473 oxygen are 2.97, 2.91, and 3.60 Å, respectively. The distances of the singular interactions between the TZD trialkylamine nitrogen, sssN, with the sulfur of Cys285 and the sulfur atom of Leu340 are 4.44 and 4.46 Å, respectively. The overall picture of 2PRG with bound rosiglitazone in black box is shown at lower right.

prioritized through second regression have their corresponding singularity interactions.

## 7. DISCUSSION

Across the 55 NCI human cancer cell lines, log *D* is the most crucial factor for the inhibitors. This is not surprising as the inhibitors need to be dissolved before the in vitro experiment is performed. However, the partition coefficient, log *D*, is by its nature potentially related to another subcellular event, the solvation−desolvation before inhibitor binding. Recently, the molecular solvation and desolvation in the presence of the solvent and receptor have been shown to make a major thermodynamic contribution to molecular binding, and the fundamental issue of solvation−desolvation is thus attracting much attention.[18] Here, the log *D* is postulated to be the general description of both solvent solubility and ligand solvation−desolvation before binding.

The identification of PSA as the most crucial descriptor of the discrepancy between the two types of in vitro biological assays of the raloxifene analogs,[11,12] the cell-free radioligand binding assay and the cell-based antiproliferation assay, indicates that the membrane transport of a molecule is the most crucial general subcellular event accounting for the discrepancy between these two assays, only one of which involves cell membranes. PSA has been correlated with various types[21] of membrane transport, but the correlation with the membrane as a subcellular event within a single cell shown by these ER antagonists is the first such observation. Therefore, based on these two observations, the descriptor-event correspondence is postulated to mean that log *D* and PSA are the two necessary and essential molecular descriptors for the molecular description of the subcellular events.

As shown in the above section, the Jurs descriptor can be further divided into log *D*, PSA, and a shape-like descriptor. A shape-like descriptor appears to be a description of the general molecular recognition, given that the overall ligand−receptor contact area is usually proportional to the energy of the van der Waals contacts. The molecular mass, a shape-like descriptor, is here postulated to be a suitable description of general ligand−receptor interaction that avoids incorporation of significant structural modifications, singular events. When the Jurs descriptor was divided into three important descriptors, log *D*, PSA, and mass, these three essential descriptors are scaffold-independent as well as independent of cellular system. The ligand-dependent cellular system of the same target family may have the similar scaffolds of ligands.[22,23]

The functionality of present study is to characterize the significant structural modifications of a given set of agonists with ligand-dependent receptor-mediated data. In this subcellular event-based molecular description, the view of molecular recognition can have two parts: the binding shape of a molecule can be described by its molecular mass, as part of general subcellular events, and the significant structural modifications can be described by the ES descriptors, as singular subcellular events. Therefore, this molecular recognition can be abstracted by a pharmacophore of three-point features, ssO, dssC, and sssN, with a general shape, mass. Taken together, this characterization of singular events of the 46 TZD PPARγ agonists provides a heuristic approach that through second regression, with the three essential descriptors, log *D*, PSA, and mass for all the possible general subcellular events, the structural modifications, predefined by ES descriptor, of agonists can be statistically prioritized. Here, the number of descriptors has been kept to a minimum to avoid possible chance correlations and top captured ES symbols can find their correspondences.


dx.doi.org/10.1021/ci200540y | *J. Chem. Inf. Model.* 2012, 52, 506−514

## 8. CONCLUSION

In the present study, the Jurs descriptor was divided into three important descriptors, log *D*, PSA, and mass (shape-like descriptor), revealed by the three selected biological data sets. Log *D* is the general description of solvation and desolvation events, PSA is the general description of membrane transport events, and mass is the description of general ligand−receptor interaction event. Figure 5 depicts the overall concept of this overall event-based description of given agonists with ligand-dependent receptor-mediated cellular data. Through first regression, the three descriptors log *D*, PSA, and mass are well-characterized as the representative descriptions of overall subcellular events of ligand cellular data. These findings improve our understanding of the descriptor-event relationship of the subcellular events in cells. Through second regression, all the ES descriptors of the 46 TZD PPAR$\gamma$ agonists were prioritized. As shown, the three top ES symbols (ssO, dssC, and sssN) form a pharmacophore of three-point features with a general shape that can account for the molecular recognition. In the end, this two-stage regression suggests a possible analysis of real ligand−receptor interactions directly from cellular data with use of minimal, but essential, descriptors

## ■ ASSOCIATED CONTENT

### Supporting Information

Tables S2.1, S3.1 and S4.1 list the structures of 110 TopI inhibitors, 46 TZD PPARs, and 72 raloxifene analogs and their experimental values. Table S2.3 lists the prioritized ES descriptors. The information is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION


### Corresponding Author

*E-mail: ytlin@kmu.edu.tw. Telephone: 886-7-3221514.


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

These works were gratefully supported by the Taiwan National Science Council (grant no. NSC100-2113-M-037-010) and Kaohsiung Medical University (grant no. KMU-EM-99-2-4). The authors are also grateful to the helpful comments of the anonymous reviewers.

## ■ REFERENCES


(1) Stanton, D. T.; Jurs, P. C. Development and use of charged partial surface area structure descriptors in computer-assisted quantitative structure-property relationship. *Anal. Chem.* **1990**, *62*, 2323−2329.

(2) Kier, L. B.; Hall, L. H. An electrotopological-state index for atom in molecules. *Pharm. Res.* **1990**, *7*, 801−807.

(3) Hall, L. H.; Mohney, B.; Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.

(4) Hall, L. H.; Kier, L. B. The E-state as the basis for molecular structure space definition and structure similarity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 784−791.

(5) Maggiora, G. M. On outliers and activity cliffs--why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

(6) Johnson, S. R. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **2008**, *48*, 25−26.

(7) Willson, T. M.; Brown, P. J.; Sternbach, D. D.; Henke, B. R. The PPARs: from orphan receptors to drug discovery. *J. Med. Chem.* **2000**, *43*, 527−550.

(8) Nolte, R. T.; Wisely, G. B.; Westin, S.; Cobb, J. E.; Lambert, M. H.; Kurokawa, R.; Rosenfeld, M. G.; Willson, T. M.; Glass, C. K.; Milburn, M. V. Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor-gamma. *Nature* **1998**, *395*, 137−143.

(9) Morrell, A.; Placzek, M.; Parmley, S.; Grella, B.; Antony, S.; Pommier, Y.; Cushman, M.; Cinelli, M. A.; Dexheimer, T. S.; Scher, E. S. Optimization of the indenone ring of indenoisoquinoline topoisomerase I inhibitors: Design, synthesis, and biological evaluation of 14-substituted aromathecins as topoisomerase I inhibitors. *J. Med. Chem.* **2007**, *50*, 4388−4404.

(10) Morrell, A.; Placzek, M. S.; Steffen, J. D.; Antony, S.; Agama, K.; Pommier, Y.; Cushman, M. Investigation of the lactam side chain length necessary for optimal indenoisoquinoline topoisomerase I inhibition and cytotoxicity in human cancer cell cultures. *J. Med. Chem.* **2007**, *50*, 2040−2048.

(11) Grese, T. A.; Cho, S.; Finley, D. R.; Godfrey, A. G.; Jones, C. D.; Lugar, C. W. 3rd; Martin, M. J.; Matsumoto, K.; Pennington, L. D.; Winter, M. A.; Adrian, M. D.; Cole, H. W.; Magee, D. E.; Phillips, D. L.; Rowley, E. R.; Short, L. L.; Glasebrook, A. L.; Bryant, H. U. Structure-activity relationships of selective estrogen receptor modulators: modifications to the 2-arylbenzothiophene core of raloxifene. *J. Med. Chem.* **1997**, *40*, 146−167.

(12) Grese, T. A.; Pennington, L. D.; Sluka, J. P.; Adrian, M. D.; Cole, H. W.; Fuson, T. R.; Magee, D. E.; Phillips, D. L.; Rowley, E. R.; Shetler, P. K.; Short, L. L.; Venugopalan, M.; Yang, N. N.; Sato, M.; Glasebrook, A. L.; Bryant, H. U. Synthesis and pharmacology of conformationally restricted raloxifene analogues: highly potent selective estrogen receptor modulators. *J. Med. Chem.* **1998**, *41*, 1272−1283.

(13) Lehmann, J. M.; Moore, L. B.; Smith-Oliver, T. A.; Wilkison, W. O.; Willson, T. M.; Kliewer, S. A. An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor gamma (PPAR gamma). *J. Biol. Chem.* **1995**, *270*, 12953−12956.

(14) Willson, T. M.; Lambert, M. H.; Kliewer, S. A. Peroxisome proliferator-activated receptor gamma and metabolic disease. *Annu. Rev. Biochem.* **2001**, *70*, 341−367.

(15) Kerwin, S. M. ChemBioOffice Ultra 2010 suite. *J. Am. Chem. Soc.* **2010**, *132*, 2466−2467.

(16) *The QSAR module in Discovery Studio*, 2.1; Accelrys Software Inc.: San Diego, CA, 2008.

(17) R, 2.11.0; R Foundation for Statistical Computing: Vienna, Austria, 2005.

(18) Syme, N. R.; Dennis, C.; Bronowska, A.; Paesen, G. C.; Homans, S. W. Comparison of entropic contributions to binding in a "hydrophilic" versus "hydrophobic" ligand-protein interaction. *J. Am. Chem. Soc.* **2010**, *132*, 8682−8689.

(19) Kier, L. B.; Hall, L. H. Derivation and significance of valence molecular connectivity. *J. Pharm. Sci.* **1981**, *70*, 583−589.

(20) Kier, L. B.; Hall, L. H. General definition of valence delta-values for molecular connectivity. *J. Pharm. Sci.* **1983**, *72*, 1170−1173.

(21) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(22) Hu, Y.; Bajorath, J. Molecular scaffolds with high propensity to form multi-target activity cliffs. *J. Chem. Inf. Model* **2010**, *50*, 500−510.

(23) Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons learned from molecular scaffold analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1742−1753.

(24) Goller, A. H.; Hennemann, M.; Keldenich, J.; Clark, T. In silico prediction of buffer solubility based on quantum-mechanical and HQSAR- and topology-based descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 648−658.