Mol Divers
DOI 10.1007/s11030-013-9462-x

FULL-LENGTH PAPER

Exploring the chemical space of aromatase inhibitors

Chanin Nantasenamat · Hao Li · Prasit Mandi · Apilak Worachartcheewan · Teerawat Monnor · Chartchalerm Isarankura-Na-Ayudhya · Virapong Prachayasittikul

Received: 12 April 2013 / Accepted: 4 July 2013

Abstract Aromatase, a rate-limiting enzyme catalyzing the conversion of androgen to estrogen, is overexpressed in human breast cancer tissue. Aromatase inhibitors (AIs) have been used for the treatment of estrogen-dependent breast cancer in post-menopausal women by blocking the biosynthesis of estrogen. The undesirable side effects in current AIs have called for continued pursuit for novel candidates with aromatase inhibitory properties. This study explores the chemical space of all known AIs as a function of their physicochemical properties by means of univariate (i.e., statistical and histogram analysis) and multivariate (i.e., decision tree and principal component analysis) approaches in order to understand the origins of aromatase inhibitory activity. Such a non-redundant set of AIs spans a total of 973 compounds encompassing both steroidal and non-steroidal inhibitors. Substructure analysis of the molecular fragments provided pertinent information on the structural features important for ligands providing high and low aromatase inhibition. Analyses were performed on data sets stratified according to their structural scaffolds (i.e., steroids and non-steroids) and bioactivities (i.e., actives and inactives). These analyses have uncover a set of rules characteristic to active and inactive AIs as well as revealing the constituents giving rise to potent aromatase inhibition.

Keywords Aromatase · Aromatase inhibitor · Breast cancer · Chemical space · Principal component analysis · Data mining

Introduction

Cancer is the leading cause of death worldwide, accounting for an estimated 7.6 million deaths in 2008 [1]. Breast cancer is the most common type of cancer in women for both developed and developing countries. The effectiveness of anti-hormonal treatment of early breast cancer can be attributed to the fact that approximately two-thirds of breast tumors are hormone-dependent and require growth factors like estrogen to grow [2]. Tamoxifen, a selective estrogen receptor modulator, is an adjuvant endocrine therapy that had for the past 30 years been the gold standard of care in the treatment of women with estrogen receptor-positive breast cancer [3]. Although being an effective agent for breast cancer treatment, it possesses some inherent adverse estrogenic properties in the uterus and vascular system that ultimately culminates in an increased risk of endometrial cancer and thromboembolism [4].

Human aromatase is the expression product of the *CYP19A1* gene located on chromosome 15q21.1 and a sole member of family 19 of the cytochrome P450 superfamily, which is one of the largest superfamily. The protein is composed of 503 amino acids with a heme co-factor at the central cavity coordinated by Cys437. The crystal structure of aromatase has first been determined by Ghosh et al. [5] at a resolution of 2.9 Å and recently refined by the same group to 2.75 Å [6]. The major source of estrogen production in

Electronic supplementary material The online version of this article (doi:10.1007/s11030-013-9462-x) contains supplementary material, which is available to authorized users.

C. Nantasenamat (✉) · H. Li · P. Mandi · A. Worachartcheewan · T. Monnor
Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
e-mail: chanin.nan@mahidol.ac.th

C. Nantasenamat · P. Mandi · A. Worachartcheewan · C. Isarankura-Na-Ayudhya · V. Prachayasittikul
Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

Published online: 16 July 2013

post-menopausal women originates from the biosynthetic conversion of androgens to estrogens by the aromatase enzyme. The enzyme catalyzes the biosynthesis of estrogens from androgens in a three-step process whereby the C19 methyl group of the androgenic substrate is oxidized to formic acid in concomitant with aromatization of the A ring to the characteristic phenolic A ring of estrogen [7]. Therefore, the inhibition of aromatase would greatly reduce the level of estrogen and compounds achieving such effects are known as aromatase inhibitors (AIs).

AIs have emerged to become the standard of care for the treatment of post-menopausal women with estrogen receptor-positive breast cancer [8]. AIs can be categorized into two major types according to their structural properties: (1) steroids (type I inhibitor) and non-steroids (type II inhibitor) [9]. The former class is known as mechanism-based inhibitors as they are converted into a chemically reactive species that bind covalently and irreversibly to aromatase; they are thus termed enzyme inactivators or suicide inhibitors. The latter class interacts reversibly with the heme co-factor by employing its azole moiety. Over the years, three generations of AIs have emerged [7]. The sole member of the first-generation of inhibitor is the non-steroidal aminoglutethimide. Owing to its poor specificity, aminoglutethimide inhibits other cytochrome P450 enzymes involved in cortisol aldosterone biosynthesis, which led to its toxicity and ultimately led to its withdrawal from clinical use. Second-generation inhibitors are composed of the non-steroidal imidazole derivative fadrozole and the steroidal analog formestane. Although fadrozole was more selective and potent than aminoglutethimide but its inhibitory properties toward aldosterone, corticosterone, and progesterone biosynthesis was undesirable. Formestane is a steroidal and the first selective AI to be used clinically. In spite of its high potency, the requirement that it be administered intramuscularly had limited its usage. Third-generation inhibitors are composed of triazole derivatives anastrozole and letrozole as well as the steroidal exemestane. In comparison to the first and second generations, the third-generation of inhibitors provided greater clinical benefits while also displaying robust aromatase inhibition of 98 % or more, which is in contrast to the 80–90 % inhibition of the former two generations [10].

In spite of good therapeutic potentials afforded by existing AIs, there is still ample opportunities for improving the bioactivities of these inhibitors. Therefore, this calls for the search for novel highly potent chemotypes possessing robust aromatase inhibition and drug-like properties. The analysis by Lipinski et al. [11] on oral drugs in their formulation of the well-known "rule of five" has immense impact for drug discovery efforts. In light of this, we report for the first time an exploration of the chemical space of AIs as well as the formulation of a set of simple and intuitive rules defining the preferred physicochemical properties for aromatase inhibition. The derived rules may be applicable for future screening of putative AIs. In order to achieve our proposed goals, we performed an exhaustive compilation of a large data set of AIs to serve as a knowledge base for further mining of useful information. Compounds were derived from the primary literature and this encompasses both steroidal and non-steroidal scaffolds. This comprehensive data set offers great opportunity for investigating the underlying profiles governing aromatase inhibition and in this study this was achieved with univariate and multivariate approaches.

Materials and methods

Data set compilation and curation

An exhaustive search for all compounds with reported aromatase inhibitory activities have been compiled from the primary literature. Such bioactivities were available in several formats: IC_{50} (and pIC_{50}), EC_{50}, and K_i. As IC_{50} values were the most abundant, therefore they were subsequently used for further investigations. Redundant compounds were identified as those with duplicate compound names, SMILES and IC_{50} values and were subjected to removal from the data set. This resulted in the final set of 973 non-redundant compounds, which is composed of 280 steroidal and 693 non-steroidal AIs. Compounds were further classified by their biological activity into active AIs for compounds having pIC_{50} values above 6 (corresponding to IC_{50} value less than or equal to 1 μM) and as inactive AIs for those having values below 5 (corresponding to IC_{50} value greater than or equal to 10 μM) while those with intermediate biological activity with pIC_{50} values in the range of 5 and 6 were not considered. This resulted in the removal of 319 compounds (comprising of 100 steroids and 219 non-steroids) to yield a new total of 654 compounds (comprising of 180 steroids and 474 non-steroids). The full data set of 973 compounds and subset of 654 compounds after intermediates removed are available in Supplementary Tables S1 and S2, respectively.

Molecular descriptors

Molecular descriptors are numerical description of the physicochemical properties of compounds as a function of their molecular structures. Chemical structures of investigated compounds were drawn using VIDA [12] and converted to the suitable file format using Babel [13]. A set of 13 easy-to-interpret molecular descriptors (i.e., six quantum chemical descriptors and seven molecular descriptors) was selected to account for the physicochemical properties of AIs. Descriptors were generated by means of computational chemistry and these descriptors have been widely used for elucidating the physicochemical property and reactivity of

compounds [14–21] as well as used in modeling their biological activities [22–29] and chemical properties [30–34].

The first set of six quantum chemical descriptors were calculated by semi-empirical AM1 method using Gaussian 09 [35]: (1) mean absolute charge (Q_m), (2) energy, (3) dipole moment (μ), (4) highest occupied molecular orbital (HOMO), (5) lowest unoccupied molecular orbital (LUMO), (6) energy gap of the HOMO and LUMO state (HOMO–LUMO gap). An additional set of seven molecular descriptors were calculated using DRAGON 5.5 Professional [36]: (7) molecular weight (MW), (8) rotatable bond number (RBN), (9) number of rings (nCIC), (10) number of hydrogen bond donors (nHDon), (11) number of hydrogen bond acceptors (nHAcc), (12) Ghose–Crippen octanol–water partition coefficient (ALogP), and (13) topological polar surface area (TPSA).

Univariate and multivariate analysis

Univariate and multivariate analysis were carried out to investigate patterns, features, and trends that were inherently present in the calculated molecular descriptors. The former was performed by constructing histogram plots of descriptors using Python scripts implemented with the matplotlib module while the latter was performed by performing principal component analysis (PCA) using The Unscrambler software package [37] and decision tree analysis using the J48 algorithm of the Weka software package, version 3.4.12 [38]. Confidence factor of 25 % [39] was employed in the decision tree analysis.

Statistical assessment

Normality of each data subset (i.e., molecular descriptors for steroids and non-steroids as well as actives and inactives) was assessed using Kolmogorov–Smirnov test. Since descriptors exhibited non-normal distribution therefore Mann–Whitney U test was employed to measure the statistical significance of the investigated pairs (i.e., steroids vs. non-steroids and actives vs. inactives).

The predictive power of classification models was statistically evaluated using sensitivity, specificity, and accuracy as described by the following equations:

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{2}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}, \tag{3}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives or over-predictions, and FN is the number of false negatives or missed predictions.

Molecular fragment analysis

In addition to analyzing the AI properties as expressed by molecular descriptors, substructure analysis of the underlying molecular fragments was performed using Fragmenter and FragmentStatistics, which are the molecular fragment creation and analysis components, respectively, of ChemAxon's JChem technology [40]. Briefly, chemical structures in SDF format were generated for all AIs using ChemAxon's MarvinSketch. Drug activity was appended to these SDF files with an in-house developed text processing program coded in C++. Fragmenter was used to process the activity-tagged SDF files containing the structural information of all AIs. Substructures were generated according to the FragmenterAll protocol. Resulting fragments were then subjected to further processing by a component of JChem known as the FragmentStatistics toolkit, which allows the division of molecular fragments according to the activity value of their parent molecules thereby enabling identification of fragments that are possibly associated with therapeutic activity. An activity cut-off value of 6 and 5 was used to distinguish active and inactive fragments, respectively. Fragments were ranked according to their molecular score as defined by ChemAxon FragmentStatistics as summarized in the equation below:

$$\text{Score} = \text{Atom count} \times (n_{\text{occurence in active set}} - n_{\text{occurence in inactive set}}). \tag{4}$$

Results and discussion

Univariate analysis of steroidal and non-steroidal aromatase inhibitors

From the total of 973 AIs collected in the data set herein it was observed that 280 were steroids and 693 were non-steroids. Simple histogram (Figs. 1 and 2) and statistical analysis (Tables 1 and 2) for each descriptor were performed to provide an overview on the relative distribution of the data values. As AIs are composed of two major classes (i.e., steroids and non-steroids) belonging to different structural scaffolds, it was therefore deemed appropriate that separate analysis should be performed as to optimally shed light on their unique physicochemical properties as well as gain insights on their mechanisms of therapeutic activity. Simple statistical analysis of bioactivity values revealed that the mean pIC_{50} value of AIs was 5.4 (about 3.98 μM) and 6.19 (about 0.64 μM) for steroids and

Table 1 Summary of statistical analysis of steroids and non-steroids

	Steroids	Non-steroids	*P* value
MW	331.950 ± 39.960	298.319 ± 62.365	<0.001
RBN	1.718 ± 1.832	3.609 ± 2.098	<0.001
nCIC	4.207 ± 0.406	3.203 ± 0.776	<0.001
nHDon	0.389 ± 0.646	0.685 ± 1.123	0.003
nHAcc	2.493 ± 0.924	3.535 ± 1.767	<0.001
ALogP	4.045 ± 1.241	3.267 ± 1.122	<0.001
TPSA	40.916 ± 14.468	53.698 ± 30.830	<0.001
Q_m	0.207 ± 0.012	0.211 ± 0.036	0.412
Energy	−1,097.594 ± 374.352	−1,127.723 ± 621.275	<0.001
Dipole moment	3.601 ± 1.150	4.215 ± 1.782	<0.001
HOMO	−0.228 ± 0.011	−0.224 ± 0.019	<0.001
LUMO	−0.042 ± 0.028	−0.051 ± 0.026	<0.001
HOMO–LUMO gap	0.186 ± 0.026	0.173 ± 0.026	<0.001
pIC_{50}	5.399 ± 1.035	6.187 ± 1.307	<0.001

non-steroids, respectively. It can be seen that non-steroids on average had slightly higher therapeutic activity than steroids ($P < 0.001$).

Several aspects of the chemical properties of AIs will be discussed either by referencing to individual descriptors or several descriptors in conjunction. As shall be seen, the AIs exhibited good agreement with the properties of known drugs as outlined by Lipinski's rule of 5 for drug-like molecules [11]. This rule stated that known therapeutic drugs generally exhibit the following molecular features: (1) MW < 500 Da, (2) LogP < 5, (3) nHDon < 5, and (4) nHAcc < 10. In order to facilitate the elucidation of structure–activity relationships, compounds were further classified as actives and inactives using pIC_{50} cut-offs of ≥6 (i.e., $IC_{50} \leq 1$ μM) and ≤5 (i.e., $IC_{50} \geq 10$ μM), respectively. Statistical analyses were then performed on these subsets of compounds stratified by their bioactivities (Table 2). Intermediate compounds with pIC_{50} in the range of 5 and 6 were not considered.

MW is a general measure of the molecular size for the investigated compounds. The histogram plot of MW of non-steroidal AIs (Fig. 1) was more normally distributed than that of steroidal AIs (Fig. 2) owing to the larger size of the data set for the former class. Furthermore, it was observed that MW values for steroidal and non-steroidal AIs were well within the range of Lipinski's rule. Moreover, steroidal AIs were generally larger in size than the non-steroidal AIs bearing values of 331.950 ± 39.960 and 298.319 ± 62.365 Da, respectively ($P < 0.001$). Non-steroidal AIs had larger size variability than steroidal AIs with ranges of 425.040 and 174.180, respectively. A majority of steroidal (172 or 61.42 %) and non-steroidal (494 or 71.28 %) AIs had MW values within one standard deviation of the mean. Moreover, more than half of the steroidal (169 or 60 %) and non-steroidal (370 or 53.39 %) AIs had values less than the mean. Statistical analysis revealed that active steroids were slightly larger in size (but not statistically significant) than their inactive counterparts as deduced from values of 334.952 ± 42.013 and 332.888 ± 42.672, respectively ($P = 0.587$), while active non-steroids were significantly larger in size than their inactive counterparts as observed from values of 309.823 ± 58.062 and 292.012 ± 81.903, respectively ($P < 0.001$).

Q_m, or the mean absolute charge, is a global measure of the molecular charge. As shown in Figs. 1 and 2, the distribution of Q_m values was not strikingly different between non-steroids and steroids as was the case for MW. Steroidal AIs were found to exhibit a more jagged distribution indicating that more steroids Q_m were distributed in narrow ranges. Both inhibitor classes had similar mean for the charge descriptor with values of 0.207 ± 0.012 and 0.211 ± 0.036 for steroids and non-steroids, respectively ($P = 0.412$). Particularly, 189 (67.5 %) steroidal and 491 (70.8 %) non-steroidal AIs had Q_m within one standard deviation from the mean while 136 (48.6 %) steroidal and 394 (56.9 %) non-steroidal AIs had Q_m below the mean. Statistical analysis suggested that there was no difference in Q_m of steroidal actives and inactives as deduced from values of 0.207 ± 0.011 and 0.206 ± 0.012, respectively ($P = 0.948$), while non-steroids showed lower values in actives than inactives with values of 0.209 ± 0.032 and 0.221 ± 0.041, respectively ($P = 0.005$).

ALogP is a computational estimation of the logarithm of 1-octanol/water partition coefficient (LogP) and it is a well-known measure of molecular hydrophobicity. The mean and standard deviation of ALogP was 4.045 ± 1.241 and 3.267 ± 1.122 for steroidal and non-steroidal AIs, respectively ($P < 0.001$). The distribution range of ALogP was higher for non-steroids at 7.8 as compared to that of steroids at 6.8. It was observed that 184 (67.5 %) steroidal and 497 (71.7 %) non-steroidal AIs had ALogP within one standard deviation from the mean. The distribution of ALogP

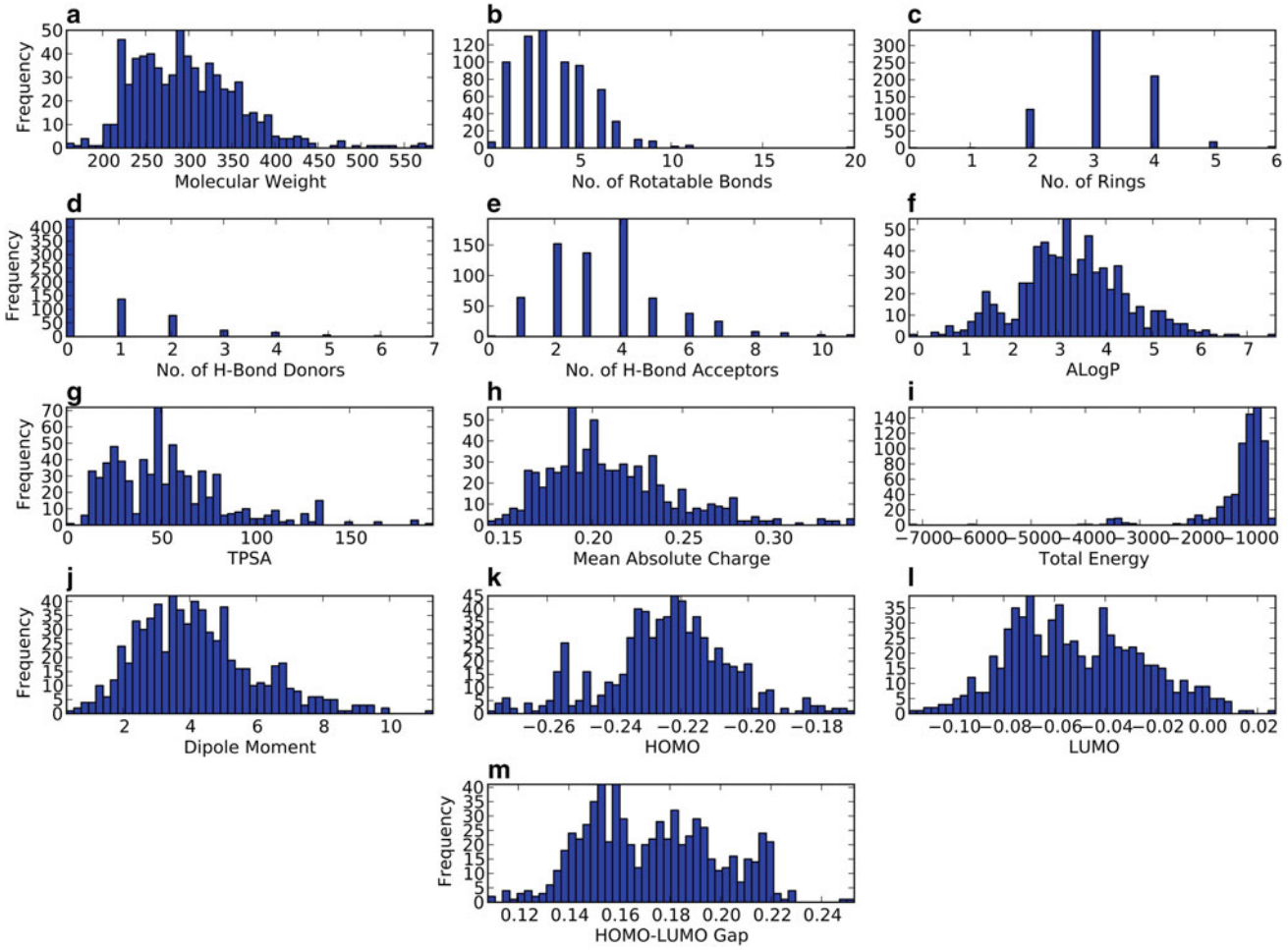


Fig. 1 Histogram of molecular descriptors of non-steroidal aromatase inhibitors

for steroids was shown to be jagged in relation to that of non-steroids indicating that steroids tended to have ALogP distributed in narrow ranges. The results indicated that 146 (52.1 %) steroids and 362 (52.2 %) non-steroids had ALogP below the mean. Statistical analysis revealed that steroidal actives had less (but not statistically significant) ALogP than steroidal inactives with corresponding values of 3.905 ± 0.992 and 4.220 ± 1.534, respectively ($P = 0.337$). Furthermore, non-steroidal actives had higher ALogP than their inactive counterparts with corresponding values of 3.323 ± 1.016 and 2.896 ± 1.209, respectively ($P < 0.001$).

nHDon and nHAcc are essentially the number of hydrogen bond donors and acceptors, respectively, present in a molecule. Non-steroidal AIs on average had higher counts for both nHDon (mean value of 0.685) and nHAcc (mean value of 3.535) when compared to mean values of 0.389 and 2.493, respectively, for steroids ($P = 0.003$ and <0.001, respectively). Hydrogen bond acceptors were more prevalent than donors for both inhibitor classes. The histogram plots clearly showed that both types of inhibitors had nHDon values close to zero. In fact, 195 (69.6 %) steroids and 431 (62.2 %) non-steroids do not possess any hydrogen bond donors at all whereas all steroids had at least one hydrogen bond acceptor and all but one non-steroidal AI had nHAcc of zero. It was observed that nHAcc were not statistically different for both steroids and non-steroids with values of 2.568 ± 0.836 and 2.485 ± 0.919, respectively, for steroids ($P = 0.511$) and values of 3.645 ± 1.678 and 4.056 ± 2.262, respectively, for non-steroids ($P = 0.277$). It was found that the values of nHDon were lower in active compounds as compared to their inactive counterpart for both steroids and non-steroids with values of 0.222 ± 0.500 and 0.687 ± 0.791, respectively, for steroids ($P < 0.001$) and values of 0.501 ± 0.889 and 1.008 ± 1.406, respectively, for non-steroids ($P < 0.001$).

Dipole moment is a measure of the asymmetric distribution of charge in a molecule where low value suggests minimal charge distribution and vice versa. The dipole moment of actives from both steroids and non-steroids was found to be higher than their inactive counterparts and a notable difference was seen in non-steroids. The corresponding

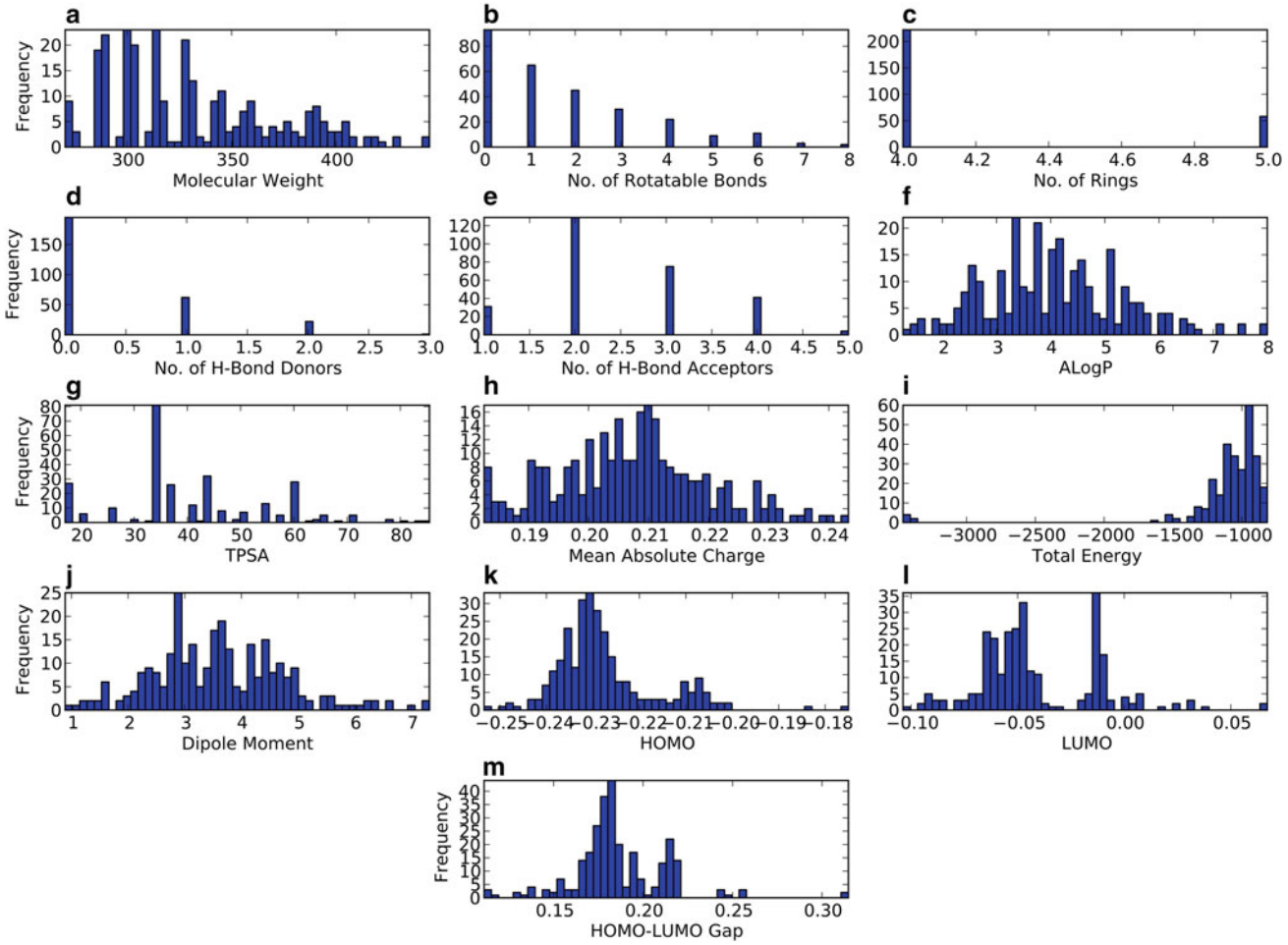


Fig. 2 Histogram of molecular descriptors of steroidal aromatase inhibitors

values of dipole moment for actives and inactives of steroids were 3.853 ± 1.083 and 3.533 ± 1.245 ($P = 0.032$), respectively, while values of 4.589 ± 1.894 and 4.016 ± 1.642 ($P = 0.002$), respectively, for non-steroids.

Energy is essentially the sum of the atomic energy and it was observed that non-steroids had higher values than their steroidal counterpart with values of $-1{,}127.723 \pm 621.275$ and $-1{,}097.594 \pm 374.352$ ($P < 0.001$), respectively. The histogram plot of the energy was found to be negatively skewed for both steroids and non-steroids with peaks at approximately $-1{,}000$, which corresponded to their mean values. It was found that active non-steroids had higher energy values than steroids and their inactive counterpart. Particularly, the corresponding values of energy for actives and inactives were $-1{,}070.557 \pm 179.896$ and $-1{,}090.429 \pm 358.407$ ($P = 0.742$), respectively, for steroids and $-1{,}244.244 \pm 744.858$ and -978.791 ± 316.172 ($P < 0.001$), respectively, for non-steroids.

TPSA is an empirical measure of the polar surface area of a molecule and it describes the contribution of polar atoms to the molecular charge. It is frequently used in the study of drug transport properties such as intestinal absorption [41] and blood–brain barrier permeability [42]. High TPSA, in addition to showing that the molecule has a complex surface charge environment, also indicates that the molecule inherently had poor membrane permeability and would need to rely on active transportation such as membrane-bound receptors. It was observed that steroids displayed quite jagged distribution for the TPSA descriptor where a large number of steroidal AIs had TPSA of about 35. In fact, 108 (64.3 %) steroids had TPSA between 32.0 and 38.0 while exhibiting a range of 68.4 with mean and standard deviation of 40.916 ± 14.468. Notably, 185 (66.1 %) steroids had TPSA within one deviation from the mean. By contrast, the TPSA values of non-steroids were more spread out and had less jagged distribution. Analysis revealed the range for TPSA in non-steroids to be 185.0, which is more than two times wider than that of the steroids. The majority appeared to have TPSA between 20 and 70 with a mean of 53.698 ± 30.830. Results indicated that 501 (72.2 %) were within one standard deviation. It should be noted that 164 (58.6 %) steroids and 389 (56.1 %) non-steroids had TPSA

below the mean. Statistical analysis revealed that actives provided higher TPSA than their inactive counterparts for both steroids and non-steroids with non-steroids affording a more significant different. The corresponding TPSA values for actives and inactives of steroidal AIs were 42.903 ± 14.998 and 41.344 ± 13.777 ($P = 0.638$), respectively, while 56.789 ± 29.403 and 60.205 ± 36.489 ($P = 0.943$), respectively, were observed for non-steroids. It appears that steroidal AIs had higher hydrophobicity and lower TPSA than their non-steroidal counterpart, both in terms of mean value and number of inhibitors with TPSA below the mean, suggesting that these compounds possessed greater membrane-crossing
capability.

HOMO–LUMO gap is the energetic difference between the HOMO and LUMO states. It is a measure of kinetic stability and chemical reactivity as HOMO and LUMO descriptors play fundamental roles in electron donation and acceptance [43]. A large gap implies high kinetic stability and low chemical reactivity because it is energetically unfavorable to add electrons to a high-lying LUMO or to extract electrons from a low-lying HOMO and so to form the activated complex of any potential reaction. Conversely, a molecule with small or no HOMO–LUMO gap is chemically reactive [44]. As can be seen in Figs. 1 and 2, the majority of steroidal AIs had HOMO–LUMO gap of approximately 0.18. HOMO–LUMO gap for steroids had mean and standard deviation of 0.186 ± 0.026 with 201 (71.8 %) being within one standard deviation from the mean. Therefore, it can be deduced from the HOMO–LUMO values that the majority of steroids had fairly similar chemical reactivity. As for non-steroidal AIs, the mean and standard deviation was 0.173 ± 0.026 and 453 (65.4 %) non-steroids were within one standard deviation. The lower percentage of non-steroidal AIs with HOMO–LUMO gap within one standard deviation indicated greater difference in terms of chemical reactivity. This is further supported by Fig. 1m in which there is greater spread of HOMO–LUMO gap values for non-steroids. However, the range of HOMO–LUMO gap for steroids was 0.204 as compared to 0.145 for non-steroids indicating that even though the majority of steroidal AIs did possess similar reactivity but there were a few atypical ones that were very different in terms of reactivity. Figure 2m shows a few steroids had unusually high HOMO–LUMO gap and thus were highly inert. Particularly, 180 (64.3 %) steroids and 355 (51.2 %) non-steroids had HOMO–LUMO gap below the mean indicating a tendency to be chemically inert. Statistical analysis indicated that the HOMO–LUMO gap for both steroids and non-steroids were smaller in the active set as compared to the inactive ones further supporting the fact that the chemical reactivity was higher in active AIs. Steroids had HOMO–LUMO gap values of 0.181 ± 0.024 and 0.185 ± 0.026 for actives and inactives ($P = 0.734$), respectively, while non-steroids provided values of 0.171 ± 0.024 and 0.174 ± 0.029 ($P = 0.541$), respectively.

PCA analysis of steroidal and non-steroidal aromatase inhibitors

In addition to simple univariate analysis of calculated molecular descriptors, PCA was also employed to provide a more detailed account of the information presented in the molecular descriptor data. PCA is a powerful multivariate data analysis technique that is extremely useful for revealing hidden data structures, trends, and correlations that are otherwise difficult to discern. PCA results in mutually orthogonal axis are called principal components (PCs) and they lie along directions of decreasing variance of the data matrix. The first PC, PC1, lies along the direction of maximal data variance, which in the vast majority of cases represents the property that distinguishes different samples of the data matrix the most. For a more detailed introduction to the principles of PCA, excellent book and review articles [45,46] are recommended. Two of the most useful features of PCA are that it offers the ability to reveal correlations between all variables simultaneously (via the loadings plot) as well as visualizing similarities and differences among samples (via the scores plot). In this study, PCA was performed on a set of 13 molecular descriptors as described in our previous navigation of the chemical space of molecularly imprinted template molecules [19]. Prior to PCA analysis, all data were standardized to comparable scale by transforming variables to zero mean and unit variance.

The purpose of PCA is to construct an easy to interpret model of the original data by separating useful information in the data from the noise. The fundamental assumption of PCA is that large variances and trends possessed by many samples are systemic variances and constitutes information while small variances present in few samples are the data noise. Therefore, lower order PCs capturing a majority of the data variances represent structures and information whereas higher order PCs tend to represent the data noise. It is therefore up to the analyst to decide on the number of PCs to adequately represent the information present in the data. The inclusion of higher order PCs usually implies model overfitting in which the model, while representative of the data used to build it, offers poor generalization. Hence, five PCs were deemed sufficient to provide meaningful information on the chemical space of AIs and for subsequent model interpretation.

In order to provide a general account of the chemical space spanned by constituting compounds in the investigated sets of non-steroidal and steroidal AIs, PCA analysis was performed. As AIs can be classified primarily as steroids and non-steroids therefore separate PCA analysis for these two types of inhibitors is deemed to be suitable in

Table 2 Summary of statistical analysis of molecular descriptors stratified by their bioactivity and structural scaffold

	Active		Inactive		*P* value[a]			
	Steroid ($n = 81$)	Non-steroid ($n = 349$)	Steroid ($n = 99$)	Non-steroid ($n = 125$)	1	2	3	4
MW	334.952 ± 42.013	309.823 ± 58.062	332.888 ± 42.672	292.012 ± 81.903	<0.001	<0.001	0.587	<0.001
RBN	1.802 ± 2.009	3.564 ± 1.708	1.838 ± 1.718	3.600 ± 2.304	<0.001	<0.001	0.454	0.328
nCIC	4.198 ± 0.401	3.407 ± 0.716	4.253 ± 0.437	3.024 ± 0.893	<0.001	<0.001	0.383	<0.001
nHDon	0.222 ± 0.500	0.501 ± 0.889	0.687 ± 0.791	1.008 ± 1.406	<0.001	0.366	<0.001	<0.001
nHAcc	2.568 ± 0.836	3.645 ± 1.678	2.485 ± 0.919	4.056 ± 2.262	<0.001	<0.001	0.511	0.277
ALogP	3.905 ± 0.992	3.323 ± 1.016	4.220 ± 1.534	2.896 ± 1.209	<0.001	<0.001	0.337	<0.001
TPSA	42.903 ± 14.998	56.789 ± 29.403	41.344 ± 13.777	60.205 ± 36.489	<0.001	<0.001	0.638	0.943
Q_m	0.207 ± 0.011	0.209 ± 0.032	0.206 ± 0.012	0.221 ± 0.041	0.401	0.010	0.948	0.005
Energy	−1,070.557 ± 179.896	−1,244.244 ± 744.858	−1,090.429 ± 358.407	−978.791 ± 316.172	0.925	<0.001	0.742	<0.001
Dipole moment	3.853 ± 1.083	4.589 ± 1.894	3.533 ± 1.245	4.016 ± 1.642	0.001	0.058	0.032	0.002
HOMO	−0.230 ± 0.008	−0.226 ± 0.019	−0.244 ± 0.013	−0.226 ± 0.018	0.001	0.373	0.004	0.621
LUMO	−0.048 ± 0.022	−0.055 ± 0.025	−0.039 ± 0.030	−0.052 ± 0.027	0.003	0.004	0.173	0.210
HOMO–LUMO	0.181 ± 0.024	0.171 ± 0.024	0.185 ± 0.026	0.174 ± 0.029	0.001	0.005	0.734	0.541
pIC_{50}	6.561 ± 0.049	7.243 ± 0.047	4.257 ± 0.059	4.504 ± 0.051	<0.001	<0.001	<0.001	<0.001

[a] Statistical significance test was performed using Mann–Whitney *U* test. *1*, Active steroid versus Active non-steroid; *2*, inactive steroid versus inactive non-steroid; *3*, active steroid versus inactive steroid; *4*, active non-steroid versus inactive non-steroid

shedding light on their properties. For that purpose, the PCA scores and loadings plot for steroids and non-steroids were examined separately and comparisons between them were drawn. Scores and loadings plot for the first three PCs are shown in panels a and b of Supplementary Figs. S1 and S2 for non-steroidal and steroidal AIs, respectively. These PCs accounted for roughly half of the data variance with values of 53.8 and 55.1 % for non-steroids (Supplementary Fig. S1c) and steroids (Supplementary Fig. S2c), respectively. The relative distribution of active and inactive compounds for non-steroids (Supplementary Fig. S1a), as shown in blue and red color, respectively, were found to be more dispersed in the active set as compared to the inactive set whereas the corresponding distribution in steroids (Supplementary Fig. S2a) displayed no significant difference. Loadings plot for non-steroids and steroids are provided in Supplementary Figs. S1b and S2b, respectively, and were shown to exhibit different rearrangement of descriptors in their respective chemical spaces.

PCA results of non-steroidal and steroidal AIs were analyzed for informative PCs and screened for outliers. Plots of their cumulative explained variance are provided in Supplementary Figs. S1c and S2c for non-steroids and steroids, respectively. Steroids were found to require more PCs than their non-steroidal counterpart in providing cumulative data variance of 70 %. Particularly, six PCs accounted for 70.8 % of the data variance in steroids whereas five PCs accounting for 71.6 % of the data variance were deemed sufficient to model the properties of non-steroidal AIs. The PC coefficient values of the loadings plot are provided in Supplementary Figs. S3 and S4 for non-steroids and steroids, respectively.

PC1 accounted for 24.4 and 13.2 % of the data variance for non-steroidal and steroidal AIs, respectively. It should be noted that PC1 was the most informative PC for non-steroids as it provided the highest explained variance of all the PCs obtained. In PCA, it matters not where the data dots are on the scores or loadings plots but only on their relative position in relation to other data dots in which information on their similarity with other samples or correlation between variables can be deduced. It was observed that descriptors with the highest influence on PC1 for non-steroids were nHAcc and TPSA that dominated one end of the PC while energy and LUMO occupied the other end. On the other hand, descriptor having the highest influence on PC1 for steroids was Q_m followed closely by TPSA and nHAcc (as also observed for non-steroids) with ALogP predominating the other side of the PC. A noteworthy difference between both classes of inhibitors can be observed from the ALogP descriptor. For non-steroidal AIs, it was shown that ALogP had low loading on PC1 indicating that it provided small data variances and that ALogP did not cause non-steroids to differ much from one another. However, for steroidal AIs ALogP had the highest loading on the negative end of PC1 thereby suggesting the importance of hydrophobicity in describing the data variance of steroids. Furthermore, another difference that can be seen in both inhibitor classes are nCIC, RBN, and MW, which provided similar level of loadings for steroids whereas in non-steroids the following trends were discerned in which MW afforded the highest loadings followed by RBN and then nCIC. Moreover, the set of quantum chemical descriptors comprising of HOMO, LUMO, and HOMO–LUMO gap had higher loadings for non-steroids than that of steroids indicating that these descriptors were less important in explaining the data variance in steroids. It can be observed that the trend in which nHAcc provided higher loadings than that of nHDon can be found in both inhibitor classes.

PC2 accounted for 16.3 and 32.9 % of data variance in non-steroids and steroids, respectively. Such high value of

loadings for steroids suggests that this PC was the most informative as it was also the PC with the highest value for all PCs obtained for steroids. Descriptor providing the highest loadings for non-steroids in PC2 was ALogP whereas the same descriptor provided the lowest loadings in PC1. Aside from the ALogP descriptor, nCIC and MW afforded the next highest loadings on the positive end of PC2 in non-steroids while Q_m provided high loadings on the negative end of the PC, which is followed by nHDon and energy descriptors. Interestingly, it was found that the set of quantum chemical descriptors (i.e., HOMO, LUMO, and HOMO–LUMO gap) afforded the highest loadings in PC2 for steroids while the same descriptor gave the lowest loadings in the preceding PC1. On the negative end of the PC of steroids, dipole moment and MW provided the highest loadings, which is followed by nHAcc and TPSA.

PC3 accounted for 13.1 and 9.0 % of data variance in non-steroids and steroids, respectively. The loadings of PC3 for non-steroids stems from LUMO and HOMO–LUMO gap on the positive end of the PC whereas the dominating descriptor on the negative end was dipole moment. In steroids, nHDon provided the highest loadings on the positive side of the PC while energy dominated the other end. The scores plot of both classes of inhibitors exhibited atypical samples on PC3, which is an indication that important trends for the data variance has become less obvious. It should be noted that atypical samples in this context refer to samples differing from the majority by properties measured by PC3 only. PC3 represented a rather small proportion of the data variance. Therefore, such atypical samples do not have leverage on the overall PCA model that is strong enough to warrant removal. The higher order PCs increasingly capture weaker trends as determined by the decreasing number of samples rather than common properties present amongst the majority. Nevertheless, the overall spread of samples, aside from the outlying ones, are still quite large, hence PC3 can still be considered to be an informative PC for both class of inhibitors, although no noteworthy features were observed from the scores plot of either inhibitor classes.

PC4 accounted for 12.2 and 9.3 % of the data variance in non-steroids and steroids, respectively. PC4 captured the variance of a small proportion of inhibitors in which there were inversed correlation of HOMO–LUMO gap and HOMO. Interestingly, these descriptors provided the highest loadings in both steroids and non-steroids.

PC5 accounted for 5.6 and 2.1 % of the data variance in non-steroids and steroids. Dipole moment was found to be the dominating descriptor on the positive end of the PC for non-steroids, which is followed by energy and RBN. On the other side of the PC, nCIC and nHDon provided the highest loadings on the negative end of the PC although to a lesser extent than the positive end. In steroids, structural descriptors, comprising of RBN on the positive of the PC and nCIC on the negative end of the PC, afforded the highest loadings in PC5. Finally, PC6 provided 4.4 % of the explained variance in steroids and descriptors giving the highest loadings were energy and nCIC.

Decision tree analysis

Decision tree is a popular machine learning technique for elucidating the underlying rules governing the inherent relationship of independent variables with a dependent variable of interest [39,47,48], which in this study is the structure–activity relationship of a set of AIs where molecular descriptors represent the former and the bioactivity represents the latter. Decision tree is a supervised approach that generates a set of if-then rules to classify compounds from the data set as actives and inactives through the use of interconnected nodes (i.e., independent variables for internal nodes and dependent variable for the terminal node) and branches (i.e., the cut-off value in which compounds are classified by). The tree finds the most informative attribute or important node (i.e., the topmost root node) followed by subsequent essential variables until terminal branches are reached for classifying the data.

Several decision tree models were constructed using the following data sets: (i) steroidal AIs and (ii) non-steroidal AIs. The predictive performance of the resulting decision tree models were assessed and compared through the use of statistical parameters (i.e., accuracy, sensitivity, and specificity). As shown in Table 3, the results indicated that the use of a set of 13 descriptors provided accuracies of 71.67 and 77.85 % for the ten-fold cross-validation set of steroidal and non-steroidal AIs, respectively. Interestingly, it was found that LUMO and HOMO–LUMO gap were the root nodes of steroidal and non-steroidal AIs, respectively, and are thus deemed as important descriptors.

Feature selection was performed by constructing an intercorrelation matrix of molecular descriptors (shown in Fig. 3) and descriptors having intercorrelation greater than 0.7 was subjected to removal. Particularly, this led to the removal of four descriptors comprising of LUMO, TPSA, Q_m, and RBN. Results indicated that the decision tree model for non-steroids provided better level of accuracy affording values of approximately 77 % for the former and 71.67 % for the latter. It should be noted that feature selection did not increase the accuracy for both data sets as the same level of accuracy was observed at 71.67 % for steroids before and after feature selection whereas a slight drop in accuracy was seen for non-steroids from 77.85 to 76.79 %. Furthermore, after feature selection it was found that nHDon and HOMO–LUMO gap were root nodes of decision tree models for steroidal and non-steroidal AIs, respectively, and are therefore deemed the most informative attributes.

A closer analysis of the prediction results revealed that although non-steroids provided better overall accuracy and

Table 3 Summary of predictive performance of decision tree models

Model	Before feature selection					After feature selection				
	Active	Inactive	Accuracy (%)	Sensitivity (%)	Specificity (%)	Active	Inactive	Accuracy (%)	Sensitivity (%)	Specificity (%)
Steroidal AIs										
Training set			93.89	95.83	91.67			92.22	93.81	90.36
Active	92	7				91	8			
Inactive	4	77				6	75			
Ten-fold CV			71.67	74.49	68.29			71.67	76.09	67.05
Active	73	26				70	29			
Inactive	25	56				22	59			
Non-steroidal AIs										
Training set			94.09	95.33	93.73			93.88	94.44	93.72
Active	102	23				102	23			
Inactive	5	344				6	343			
Ten-fold CV			77.85	60.64	82.11			76.79	57.58	81.87
Active	57	68				57	68			
Inactive	37	312				42	307			

higher specificity but this comes at the cost of a lower level of sensitivity. Particularly, non-steroids afforded accuracy, sensitivity, and specificity of approximately 77, 60, and 82 %, respectively, as compared to 72, 75, and 67 % of steroids, respectively.

Fragment-based analysis

The search for lead structures is a vital step in the design of novel drugs. It is noted that amongst the vast number of known drugs only a comparably limited number of common structures, termed *privileged structures*, were shared. Previous evidence [49] supports the view that such privileged structures have an inherent tendency toward biological activity and that the backbone of these structures could be further modified to provide novel therapeutic activities. The discovery of such privileged fragments requires splitting the molecules of interest into their constitutional fragments and then identifying what fragments correlate with drug activity. High-throughput screening for identification of lead compounds is a well-established concept [49]. Computational fragment-based drug design offers a fast, efficient, and low cost alternative to high-throughput screening. This approach is an algorithmic methodology that utilizes molecular fragments from existing databases or generates them from established fragmentation rules and subsequently relate the fragments to their respective drug activities [50].

This study employed ChemAxon's Molecular Fragmenter [40] in decomposing AIs into their constitutional fragments. Empirical analysis of the generated fragments revealed that multiple structures were prevalent in inhibitors designated as active ($IC_{50} \leq 1$ μM corresponding to $pIC_{50} \geq 6$) as well as fragments that preferentially occur in inhibitors designated as inactive ($IC_{50} \geq 10$ μM corresponding to $pIC_{50} \leq 5$). Hence, these discriminating fragments provided tentative clues as to the substructures that are essential for the observed aromatase inhibitory activity.

Computed fragments

Molecular fragmentation was performed according to the FragmenterAll protocol to yield a set of 239 and 1,958 unique fragments for steroids and non-steroids, respectively, with corresponding fragment occurrences of 360 and 2,963, respectively. Strikingly, while a large number of unique fragments were generated from the inhibitors, fragments shared amongst the different inhibitors were surprisingly low.

Particularly, of the total of 1,958 unique fragments that were generated for non-steroidal AIs, quite a high number of these fragments or 1,668 (85.2 %) had an occurrence of one. Furthermore, 290 unique fragments with occurrences of two or more had a frequency of 1,295. This is only 14.8 % of the unique fragments that accounted for 43.7 % of the fragment occurences, which are almost half of the 2,963 fragment occurrences from non-steroids. Moreover, it is was found that only 12 unique fragments (or 0.006 % from the total number of fragments generated) possessed occurrences of more than ten, which accounted for a total of 561 occurrences (18.9 %) or roughly a fifth of the total occurrences.

It was found that molecular fragmentation gave rise to 239 unique fragments for the steroidal AIs where 202 of the fragments had an occurrence of one thereby accounting for 84.5 % of the unique fragments generated. Furthermore, fragments occurring for two or more times accounted for the remaining 15.5 % or 158 of the fragment occurrence. Surprisingly, only two fragments had occurrences of more than ten, which is a number that is significantly less than their non-steroidal counterpart. Partial explanation for such observation may be due to the inherently rigid structural scaffold of steroids that consequently limits the breadth of the chemi-

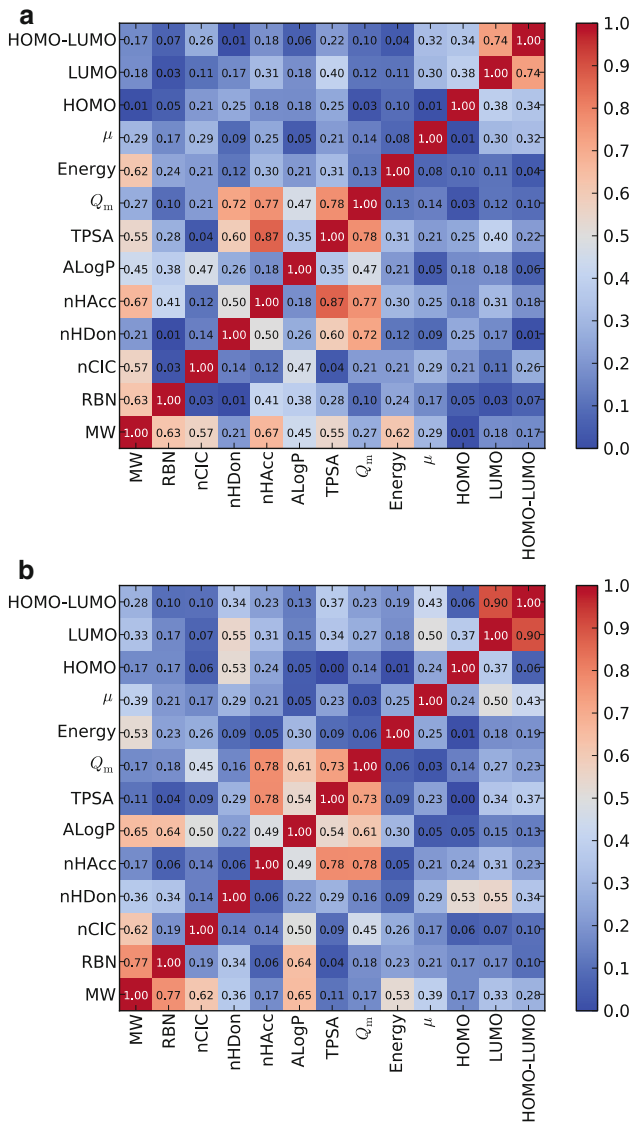


Fig. 3 Heat map representation of the intercorrelation matrix of molecular descriptors for non-steroidal (**a**) and steroidal (**b**) aromatase inhibitors. Pearson's correlation coefficient of descriptor pairs are indicated in the respective *box* and *color-coded* according to their degree of correlation (Color figure online)

cal space. Furthermore, the rather limited size of compounds present in the steroidal data set is another factor contributing to the few number of fragments generated in this class.

These figures indicated that the set of AIs investigated in this study possessed a very limited number of common structures that were, however, highly prevalent. The privileged structures shared amongst many of the inhibitors are likely to be significantly associated with their activity as it was observed that many of the fragments occurred just once, which is an indication that many of the drug development attempts were made by modifying the few privileged structures. Thus, they appear to be necessary for aromatase inhibitory activity.

Fragments associated with activity and inactivity

The concept of searching for promising lead structures is a well-established concept in drug design. It should be noted that although the goal is to search for structures associated with high activity but those that preferentially occur in less successful drugs should also deserve attention. Such fragments occurring in the inactive set could provide useful knowledge on which structures to avoid as they give rise to low therapeutic activity. Therefore, theoretical understanding of why such structures tend to occur in drugs with low activity would further benefit the rational design of novel therapeutics. Hence, fragments from both the active and inactive set of AIs were surveyed in this study.

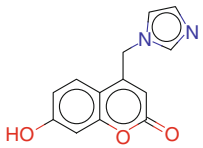
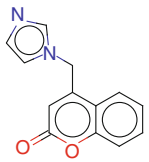
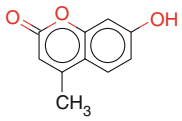
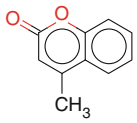
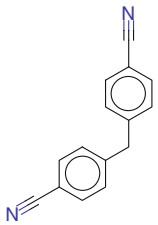
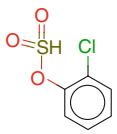
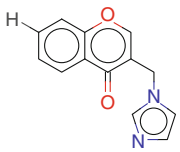
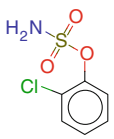
Analysis of non-steroidal AIs using the Fragment Statistics toolkit revealed that from the total of 2,963 fragment occurrences, 2,292 (77.4 %) fragments were found in the active set of compound. Such fragment occurrences having at least one occurrence in the active set belonging to 1,561 (79.7 %) unique fragments. Results indicated that steroidal AIs had a total fragment occurrence of 360 and 200 (55.6 %) of these fragments were associated with the active set and that these occurrences belong to 122 unique fragments.

Analysis of non-steroidal AIs indicated that from the total of 2,963 fragment occurrences, 671 (22.6 %) were found in the inactive set of compounds. Such fragment occurrences having at least one occurrence in the inactive set belongs to 478 (24.4 %) unique fragments for the inactive set of compounds. Of the 360 fragment occurrence in steroidal AIs, 160 were associated with the inactive set and that these occurrences belong to 127 unique fragments. It should be noted that unique fragments could be associated with either the active or inactive set as well as belong to both sets at the same time.

The ten top-ranking fragments for both active and inactive set of fragments are shown in Tables 4 and 5, respectively, for non-steroidal AIs and in Tables 6 and 7, respectively, for steroidal AIs. Thus, the ten fragments with the highest or lowest molecular score are likewise considered common structures shared amongst inhibitors having high or low activity.

It is interesting to note that five of the top ten fragments occurring in the active set of non-steroids (Table 4) contain the azole ring. Of particular note is that the three top-ranking fragments (as well as the ninth rank) all contain 1,3-imidazole ring while the sixth rank contain the 1,2,4-triazole ring. This is in agreement with the literature in which the azole ring is known to effectively coordinate the heme iron in its inhibition of aromatase [2]. Another interesting fragment found in the active set is 4-[(4-cyanophenyl)methyl]benzonitrile, which is a substructural component of the third-generation of AIs namely letrozole. Analysis of the inactive set of fragments from non-steroids (Table 5) pointed out that all ten are composed of 5-hydroxy-2-phenyl-4*H*,10*H*-pyrano[2,3-

Table 4 Summary of the top ten active fragments from non-steroidal AIs sorted according to their molecular score

Rank	Structure	Atom count	Fragment occurrence		Molecular score
			Active	Inactive	
1		5	154	20	670
2		18	24	0	432
3		17	24	0	408
4		13	24	0	312
5		12	24	0	288
6		5	60	3	285
7		17	8	0	136
8		11	8	0	88
9		17	5	0	85
10		12	7	0	84

f]chromen-4-one, which is a fusion of the flavone ring with the six-membered heterocyclic 4*H*-pyran.

A notable difference between fragments from the active (Table 6) and inactive (Table 7) set of steroids is the composition of carbonyl and hydroxyl moieties at the C3 position in the former and latter, respectively. It should be noted that the presence of hydroxyl group at the C3 position is characteristic of sterols. It is also worthy to mention that the three top-ranking fragments for the inactive set lack a methyl group at the C10 position whereas this moiety is present in all active

Table 5 Summary of the top ten inactive fragments from non-steroidal AIs sorted according to their molecular score

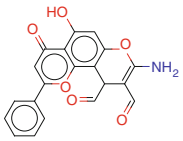
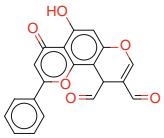
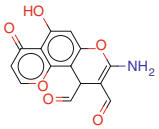
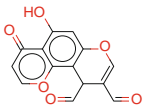
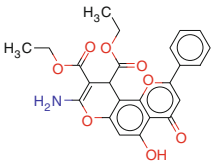
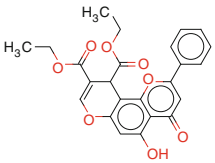
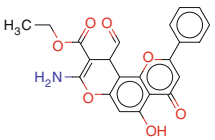
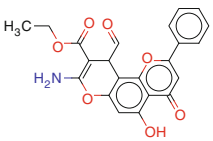
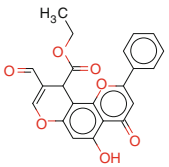
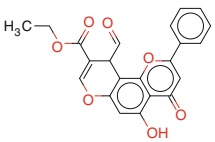
Rank	Structure	Atom count	Fragment occurrence		Molecular score
			Active	Inactive	
1		27	0	5	−135
2		26	0	5	−130
3		21	0	5	−105
4		20	0	5	−100
5		33	0	3	−99
6		32	0	3	−96
7		30	0	3	−90
8		30	0	3	−90
9		29	0	3	−87
10		29	0	3	−87

Table 6 Summary of the top ten active fragments from steroidal AIs sorted according to their molecular score

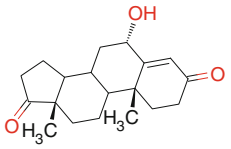
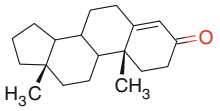
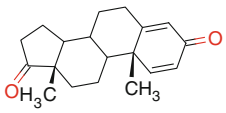
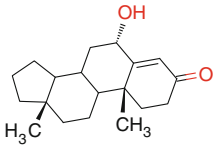
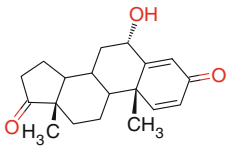
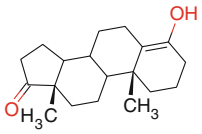
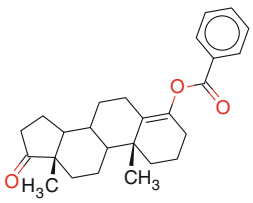
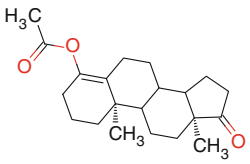
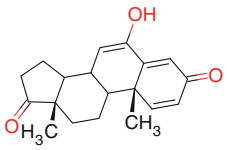
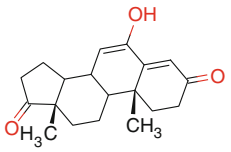
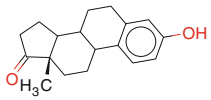
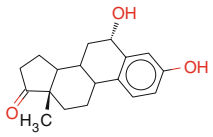
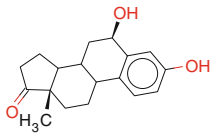
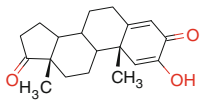
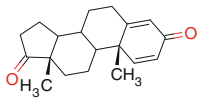
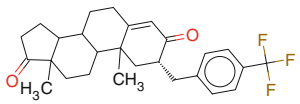
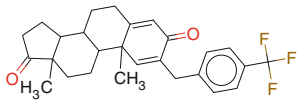
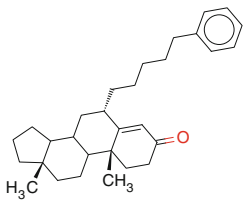
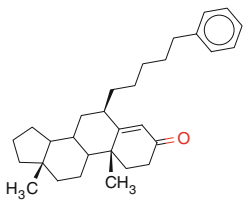
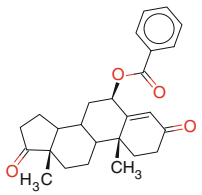
Rank	Structure	Atom count	Fragment occurrence		Molecular score
			Active	Inactive	
1		22	8	0	176
2		20	7	0	140
3		21	6	0	126
4		21	5	0	105
5		22	4	0	88
6		21	4	0	84
7		29	2	0	58
8		24	2	0	48
9		22	2	0	44
10		22	2	0	44

Table 7 Summary of the top ten inactive fragments from steroidal AIs sorted according to their molecular score

Rank	Structure	Atom count	Fragment occurrence		Molecular score
			Active	Inactive	
1		20	0	6	−120
2		21	0	3	−63
3		21	0	3	−63
4		22	0	2	−44
5		21	0	2	−42
6		32	0	1	−32
7		32	0	1	−32
8		31	0	1	−31
9		31	0	1	−31
10		30	0	1	−30

fragments. The absence of methyl group at the C10 position is characteristic of estrogens whereas its presence is reminiscent of androgens.

Substructural analysis performed herein indicated that these fragments are likely to be important structures that are associated with aromatase inhibitory activity. Furthermore,

results indicated that non-steroidal AIs, when compared to their steroidal counterpart, had higher mean activity as well as a greater percentage of molecular fragments occurring in the active set. Hence, it appears that non-steroidal AIs were generally more effective. Particularly, fragments having high molecular score are empirically indicative of structural components necessary for effective aromatase inhibitory activity. Thus, these types of fragments are worthy to be considered in further drug design. Of note is the fact that fragments from non-steroidal AIs had much higher molecular score than their steroidal counterpart. This may be indicative that they have higher tolerance to modifications and thus were shared amongst many of the non-steroidal AIs. Although detailed study of fragment–activity contribution was not carried out in this study, such observed differences provided some clues as to the relationship between fragment structures and therapeutic activity.

Conclusion

This study explored the chemical space of AIs that have been exhaustively collected from the literature. To provide insights into the origins of aromatase inhibitory activity, the data set was stratified according to their structural scaffolds (i.e., steroids and non-steroids) and bioactivities (i.e., actives and inactives) where they were subjected to chemical space navigation by means of univariate analysis, decision tree, PCA analysis, and molecular fragment analysis. Substructural analysis of molecular fragments revealed substructures that were preferentially shared by the active set, which were different from those shared by the inactive set. It is possible that highly prevalent fragments from the active set were responsible for their therapeutic effectiveness and a further refined study is highly desirable. Systematic revelation of major trends in the physicochemical property of investigated compounds served as an informational basis that is useful for further rational design of novel AIs. Thus, the methodology described herein may be applied to other molecular systems of interest.

Acknowledgments The Goal-Oriented Research Grant of Mahidol University to C.N. is gratefully acknowledged for financial support of this research. H.L., P.M., and T.M. are grateful for research assistantship supported by Grants from Mahidol University. This project was also supported in part by the Office of the Higher Education Commission and Mahidol University under the National Research Universities Initiative.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011) Global cancer statistics. CA Cancer J Clin 61:69–90. doi:10.3322/caac.20107
2. Miller WR (2003) Aromatase inhibitors: mechanism of action and role in the treatment of breast cancer. Semin Oncol 30:3–11
3. Jordan VC (2004) Selective estrogen receptor modulation: concept and consequences in cancer. Cancer Cell 5:207–213. doi:10.1016/S1535-6108(04)00059-5
4. Fisher B, Costantino JP, Redmond CK, Fisher ER, Wickerham DL, Cronin WM (1994) Endometrial cancer in tamoxifen-treated breast cancer patients: findings from the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-14. J Natl Cancer Inst 86:527–537. doi:10.1093/jnci/86.7.527
5. Ghosh D, Griswold J, Erman M, Pangborn W (2009) Structural basis for androgen specificity and oestrogen synthesis in human aromatase. Nature 457:219–223. doi:10.1038/nature07614
6. Ghosh D, Lo J, Morton D, Valette D, Xi J, Griswold J, Hubbell S, Egbuta C, Jiang W, An J, Davies HM (2012) Novel aromatase inhibitors by structure-guided design. J Med Chem 55:8464–8476. doi:10.1021/jm300930n
7. Simpson ER, Clyne C, Rubin G, Boon WC, Robertson K, Britt K, Speed C, Jones M (2002) Aromatase—a brief overview. Annu Rev Physiol 64:93–127. doi:10.1146/annurev.physiol.64.081601.142703
8. Burstein HJ, Prestrud AA, Seidenfeld J, Anderson H, Buchholz TA, Davidson NE, Gelmon KE, Giordano SH, Hudis CA, Malin J, Mamounas EP, Rowden D, Solky AJ, Sowers MR, Stearns V, Winer EP, Somerfield MR, Griggs JJ (2010) American Society of Clinical Oncology clinical practice guideline: update on adjuvant endocrine therapy for women with hormone receptor-positive breast cancer. J Clin Oncol 28:3784–3796. doi:10.1200/jco.2009.26.3756
9. Ponzone R, Mininanni P, Cassina E, Pastorino F, Sismondi P (2008) Aromatase inhibitors for breast cancer: different structures, same effects? Endocr Relat Cancer 15:27–36. doi:10.1677/erc-07-0249
10. Lønning PE (2004) Aromatase inhibitors in breast cancer. Endocr Relat Cancer 11:179–189
11. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46:3–26
12. VIDA (2013) Version 4.2.1. OpenEye Scientific Software, Santa Fe, NM
13. Babel (2013) Version 3.3. OpenEye Scientific Software, Santa Fe, NM
14. Isarankura-Na-Ayudhya C, Nantasenamat C, Buraparuangsang P, Piacham T, Ye L, Bülow L, Prachayasittikul V (2008) Computational insights on sulfonamide imprinted polymers. Molecules 13:3077–3091. doi:10.3390/molecules13123077
15. Suksrichavalit T, Prachayasittikul S, Piacham T, Isarankura-Na-Ayudhya C, Nantasenamat C, Prachayasittikul V (2008) Copper complexes of nicotinic–aromatic carboxylic acids as superoxide dismutase mimetics. Molecules 13:3040–3056. doi:10.3390/molecules13123040
16. Suksrichavalit T, Prachayasittikul S, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V (2009) Copper complexes of pyridine derivatives with superoxide scavenging and antimicrobial activities. Eur J Med Chem 44:3259–3265. doi:10.1016/j.ejmech.2009.03.033
17. Prachayasittikul V, Isarankura-Na-Ayudhya C, Tantimongcolwat T, Nantasenamat C, Galla HJ (2007) EDTA-induced membrane fluidization and destabilization: biophysical studies on artificial lipid membranes. Acta Biochim Biophys Sin 39:901–913. doi:10.1111/j.1745-7270.2007.00350.x
18. Prachayasittikul S, Wongsawatkul O, Worachartcheewan A, Nantasenamat C, Ruchirawat S, Prachayasittikul V (2010) Elucidating the structure–activity relationships of the vasorelaxation and antioxidation properties of thionicotinic acid derivatives. Molecules 15:198–214. doi:10.3390/molecules15010198

19. Nantasenamat C, Li H, Isarankura-Na-Ayudhya C, Prachayasittikul V (2012) Exploring the physicochemical properties of templates from molecular imprinting literature using interactive text mining approach. Chemometr Intell Lab Syst 116:128–136. doi:10.1016/j.chemolab.2012.05.006
20. Piacham T, Isarankura-Na-Ayudhya C, Nantasenamat C, Yainoy S, Ye L, Bülow L, Prachayasittikul V (2006) Metalloantibiotic Mn(II)–bacitracin complex mimicking manganese superoxide dismutase. Biochem Biophys Res Commun 341:925–930. doi:10.1016/j.bbrc.2006.01.045
21. Piacham T, Nantasenamat C, Suksrichavalit T, Puttipanyalears C, Pissawong T, Maneewas S, Isarankura-Na-Ayudhya C, Prachayasittikul V (2009) Synthesis and theoretical study of molecularly imprinted nanospheres for recognition of tocopherols. Molecules 14:2985–3002. doi:10.3390/molecules14082985
22. Mandi P, Nantasenamat C, Srungboonmee K, Isarankura-Na-Ayudhya C, Prachayasittikul V (2012) QSAR study of anti-prion activity of 2-aminothiazoles. EXCLI J 11:453–467
23. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V (2008) Prediction of bond dissociation enthalpy of antioxidant phenols by support vector machine. J Mol Graph Model 27:188–196. doi:10.1016/j.jmgm.2008.04.005
24. Nantasenamat C, Piacham T, Tantimongcolwat T, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V (2008) QSAR model of the quorum-quenching *N*-acyl-homoserine lactone lactonase activity. J Biol Syst 16:279–293. doi:10.1142/S021833900800254X
25. Pingaew R, Tongraung P, Worachartcheewan A, Nantasenamat C, Prachayasittikul S, Ruchirawat S, Prachayasittikul V (2012) Cytotoxicity and QSAR study of (thio)ureas derived from phenylalkylamines and pyridylalkylamines. Med Chem Res. doi:10.1007/s00044-012-0402-6
26. Thippakorn C, Suksrichavalit T, Nantasenamat C, Tantimongcolwat T, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V (2009) Modeling the LPS neutralization activity of anti-endotoxins. Molecules 14:1869–1888. doi:10.3390/molecules14051869
27. Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul S, Prachayasittikul V (2011) Predicting the free radical scavenging activity of curcumin derivatives. Chemometr Intell Lab Syst 109:207–216. doi:10.1016/j.chemolab.2011.09.010
28. Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V (2013) Predicting antimicrobial activities of benzimidazole derivatives. Med Chem Res. doi:10.1007/s00044-013-0539-y
29. Worachartcheewan A, Nantasenamat C, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V (2009) Modeling the activity of furin inhibitors using artificial neural network. Eur J Med Chem 44:1664–1673. doi:10.1016/j.ejmech.2008.09.028
30. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V (2007) Quantitative structure–imprinting factor relationship of molecularly imprinted polymers. Biosens Bioelectron 22:3309–3317. doi:10.1016/j.bios.2007.01.017
31. Nantasenamat C, Isarankura-Na-Ayudhya C, Tansila N, Naenna T, Prachayasittikul V (2007) Prediction of GFP spectral properties using artificial neural network. J Comput Chem 28:1275–1289. doi:10.1002/jcc.20656
32. Nantasenamat C, Naenna T, Prachayasittikul V (2005) Quantitative prediction of imprinting factor of molecularly imprinted polymers by artificial neural network. J Comput Aided Mol Des 19:509–524. doi:10.1016/S1535-6108(04)00059-5
33. Nantasenamat C, Srungboonmee K, Jamsak S, Tansila N, Isarankura-Na-Ayudhya C, Prachayasittikul V (2013) Quantitative structure–property relationship study of spectral properties of green fluorescent protein with support vector machine. Chemometr Intell Lab Syst 120:42–52. doi:10.1016/j.chemolab.2012.11.003
34. Worachartcheewan A, Dansethakul P, Nantasenamat C, Pidetcha P, Prachayasittikul V (2012) Determining the optimal cutoff points for waist circumference and body mass index for identification of metabolic abnormalities and metabolic syndrome in urban Thai population. Diabetes Res Clin Pract 98:e16–e21. doi:10.1016/j.diabres.2012.09.018
35. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas Ö, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian 09, Revision A.02. Gaussian, Inc., Wallingford
36. DRAGON for Windows (Software for Molecular Descriptor Calculations) (2007) Version 5.5. Talete srl, Milano, Italy
37. The Unscrambler (2005) Version 9.5. Camo Process AS, Oslo, Norway
38. Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington
39. Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P, Prachayasittikul V (2010) Identification of metabolic syndrome using decision tree analysis. Diabetes Res Clin Pract 90:e15–18. doi:10.1016/j.diabres.2010.06.009
40. JChem (2012) Version 5.10. ChemAxon Ltd., Hungary
41. Li SF, He HD, Parthiban LJ, Yin HQ, Serajuddin ATM (2005) IV–IVC considerations in the development of immediate-release oral dosage form. J Pharm Sci 21. doi:10.1002/jps.20378
42. Strazielle N, Ghersi-Egea JF (2005) Factors affecting deivery of antiviral drugs to the brain. Rev Med Virol 15:105–133. doi:10.1002/rmv.454
43. Bulat FA, Chamorro E, Fuentalba P, Toro-Labbe A (2004) Condensation of frontier molecular orbital fukui functions. J Phys Chem A 108:342–349. doi:10.1021/jp036416r
44. Aihara J (1999) Reduced HOMO–LUMO gap as an index of kinetic stability for polycyclic aromatic hydrocarbons. J Phys Chem A 103:7487–7495. doi:10.1021/jp990092i
45. Esbensen KH, Guyot D, Westad F, Houmoller LP (2004) Multivariate data analysis—in practice. CAMO Process AS, Esbjerg
46. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometr Intell Lab Syst 2:37–52. doi:10.1016/0169-7439(87)80084-9
47. Suenderhauf C, Hammann F, Huwyler J (2012) Computational prediction of blood–brain barrier permeability using decision tree induction. Molecules 17:10429–10445. doi:10.3390/molecules170910429
48. Yang XG, Chen D, Wang M, Xue Y, Chen YZ (2009) Prediction of antibacterial compounds by machine learning approaches. J Comput Chem 30:1202–1211. doi:10.1002/jcc.21148
49. DeSimone RW, Currie KS, Mitchell SA, Darrow JW, Pippin DA (2007) Privileged structures: applications in drug discovery. Comb Chem High Throughput Screen 7:473–493. doi:10.2174/1386207043328544
50. McInnes C (2007) Virtual screening strategies in drug discovery. Curr Opin Chem Biol 11:494–5025. doi:10.1016/j.cbpa.2007.08.033