QSAR modeling of aromatase inhibitory activity of 1-substituted 1,2,3-triazole analogs of letrozole. Eur J Med Chem

ARTICLE *in* EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY · AUGUST 2013

Impact Factor: 3.45 · DOI: 10.1016/j.ejmech.2013.08.015 · Source: PubMed

5 AUTHORS, INCLUDING:

Chanin Nantasenamat
Mahidol University
82 PUBLICATIONS **985** CITATIONS
SEE PROFILE

Apilak Worachartcheewan
Mahidol University
56 PUBLICATIONS **471** CITATIONS
SEE PROFILE

Supaluk Prachayasittikul
Mahidol University
70 PUBLICATIONS **673** CITATIONS
SEE PROFILE

Virapong Prachayasittikul
Mahidol University
200 PUBLICATIONS **1,863** CITATIONS
SEE PROFILE

European Journal of Medicinal Chemistry 69 (2013) 99–114

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

QSAR modeling of aromatase inhibitory activity of 1-substituted 1,2,3-triazole analogs of letrozole

Chanin Nantasenamat [a,b,*], Apilak Worachartcheewan [a], Supaluk Prachayasittikul [a], Chartchalerm Isarankura-Na-Ayudhya [b], Virapong Prachayasittikul [b]

[a] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[b] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

ARTICLE INFO

Article history:
Received 22 February 2013
Received in revised form
28 July 2013
Accepted 7 August 2013
Available online 19 August 2013

Keywords:
Aromatase
Breast cancer
Triazole
Letrozole
QSAR
Structure–activity relationship
Chemical space

ABSTRACT

Aromatase is an estrogen biosynthesis enzyme belonging to the cytochrome P450 family that catalyzes the rate-limiting step of converting androgens to estrogens. As it is pertinent toward tumor cell growth promotion, aromatase is a lucrative therapeutic target for breast cancer. In the pursuit of robust aromatase inhibitors, a set of fifty-four 1-substituted mono- and bis-benzonitrile or phenyl analogs of 1,2,3-triazole letrozole were employed in quantitative structure–activity relationship (QSAR) study using multiple linear regression (MLR), artificial neural network (ANN) and support vector machine (SVM). Such QSAR models were developed using a set of descriptors providing coverage of the general characteristics of a molecule encompassing molecular size, flexibility, polarity, solubility, charge and electronic properties. Important physicochemical properties giving rise to good aromatase inhibition were obtained by means of exploring its chemical space as a function of the calculated molecular descriptors. The optimal subset of 3 descriptors (i.e. number of rings, ALogP and HOMO–LUMO) was further used for QSAR model construction. The predicted pIC_{50} values were in strong correlation with their experimental values displaying correlation coefficient values in the range of 0.72–0.83 for the cross-validated set (Q_{CV}) while the external test set (Q_{Ext}) afforded values in the range of 0.65–0.66. Insights gained from the present study are anticipated to provide pertinent information contributing to the origins of aromatase inhibitory activity and therefore aid in our on-going quest for aromatase inhibitors with robust properties.

1. Introduction

Breast cancer is considered to be the most common type of cancer and is the leading cause of cancer-related death in women accounting for an estimated 23% of new cases and 14% of all cancer deaths in 2008 [1]. It is widely accepted that the majority of breast cancers are hormone-dependent and that estrogen is a key mediator in the progression and metastasis of breast tumors. Particularly, for postmenopausal women it has been reported that the concentration of 17β-estradiol (E2) in breast tumor can be ten-fold higher than those in plasma [2]. The high concentration of E2 in breast tumors could be attributed to increased uptake from plasma or in situ aromatization of androgens to estrogens [3]. The latter is afforded by aromatase, an enzyme involved in the rate-limiting step of estrogen biosynthesis by catalyzing three consecutive hydroxylation reactions that aromatizes C19 androgens to C18 estrogens. Thus, blockade of the synthesis of estrogens by inhibiting aromatase represents a lucrative therapeutic approach for the treatment of hormone-sensitive breast cancer [4,5].

Aromatase inhibitors (AIs) are comprised of 2 major classes of compounds according to their molecular structure and mechanism of action. Type I or steroidal inhibitors typically contain an androgen core structure and snugly binds to the substrate-binding site while type II or non-steroidal inhibitors are typically comprised of azole moieties that bind to the heme co-factor of the enzyme. The former class usually binds irreversibly to the substrate-binding site and is thus termed suicidal inhibitors as they inactivate the enzyme while the latter class binds reversibly by positioning the nitrogen atom from its azole moiety to coordinate with the iron atom of the heme co-factor.

* Corresponding author. Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand. Tel.: +66 2 441 4371; fax: +66 2 441 4380.
E-mail address: chanin.nan@mahidol.ac.th (C. Nantasenamat).

0223-5234/$ – see front matter © 2013 Elsevier Masson SAS. All rights reserved.
http://dx.doi.org/10.1016/j.ejmech.2013.08.015

Three generations of AIs had been introduced over the past three decades. The third generation AIs demonstrated good tolerability profiles, higher selectivity and potency when compared with the first and second generation AIs [6,7]. Particularly, the first and second generation AIs had been shown to inhibit *in vivo* estrogen synthesis by up to 90% whereas the third generation AIs provided greater than 98% inhibition [8]. Letrozole (also known as CGS 20267) is a third generation AI that was first introduced by Novartis (then Ciba-Geigy) as a non-steroidal AI possessing potent pharmacological profile [9]. In 1996 it was approved in Europe and this is followed by its subsequent approval in the United States in 1997 [10]. The compound is marketed as Femara and has an IUPAC name of 4-[(4-cyanophenyl)(1H-1,2,4-triazol-1-yl)methyl]benzonitrile. Letrozole has been particularly demonstrated to be a highly potent AI both *in vitro* and *in vivo*. Particularly, letrozole exhibited higher potency than several AIs such as aminoglutethimide, anastrozole, exemestane and formestane [9,11,12].

The seminal work by Hansch et al. [13] in correlating the biological activity of plant growth regulators with Hammett constants and partition coefficients popularized the concept of quantitative structure–activity relationship (QSAR). Over the years, the QSAR paradigm has been employed in modeling a wide range of biological activities (i.e. spectral properties of green fluorescent protein [14,15], antioxidant activity [16,17], quorum quenching of *N*-acyl homoserine lactone lactonase [18], lipopolysaccharide neutralization activity of anti-endotoxins [19], furin inhibition [20], anti-prion activity [21] and anti-cancer activity [22]), chemical properties (i.e. imprinting factor of molecularly imprinted polymers [23,24]) and medical conditions (i.e. prediction of ischemic heart disease [25] and identification of metabolic syndrome [26,27]). The concepts of QSAR modeling had been reviewed previously [28,29]. Briefly, QSAR constructs statistically validated models that are capable of quantitatively predicting the bioactivity of the explored compounds. This is performed by employing multivariate analysis methods to discern linear or non-linear relationships between the molecular structures (i.e. physicochemical descriptors) and their respective biological activities (i.e. aromatase inhibitory activity).

Herein, a series of 1-substituted mono- and bis-benzonitrile or phenyl analogs of 1,2,3-triazole letrozole are employed for QSAR modeling of the aromatase inhibitory activity. A diverse set of quantum chemical and molecular descriptors were employed to provide numerical description of the investigated compounds. Descriptors accounting for the unique structural features of the compounds were correlated to their respective aromatase inhibitory activity using multiple linear regression (MLR), artificial neural network (ANN) and support vector machine (SVM).

2. Results and discussion

2.1. Structure–activity relationship of 1,2,3-triazole letrozole analogs

Letrozole is a non-steroidal AI known by its IUPAC name of 4-[(4-cyanophenyl)(1H-1,2,4-triazol-1-yl)methyl]benzonitrile. As can be seen from Fig. 1 the chemical structure of letrozole **1** is comprised of a 1,2,4-triazole and 1-substituted bis-benzonitrile. Mechanistically, the azole functional moiety coordinate its N4-atom to interact with Fe^{2+} of the heme prosthetic group while its two phenyl rings provide tighter fit inside the binding cavity by mimicking the steroidal backbone of the natural substrate, androstenedione. Furthermore, the two nitrile groups at the para positions of the phenyl ring mimics the carbonyl group of androstenedione and functions as hydrogen bond acceptors. Previous investigation on structure–activity relationship signifies the importance of electron withdrawing groups at the para position of the phenyl ring while demonstrating that nitrile groups afforded the best activity [30]. This is in concomitant with the findings by Schuster et al. [31] that two aromatic rings along with two hydrogen bond donors are important pharmacophores for strong aromatase inhibition. Moreover, the importance of hydrogen bonding in aromatase inhibition was suggested by Neves et al. [32] from their computational analysis of potent AIs. In a comparative analysis of the structures of letrozole and androstenedione, Doiron et al. [33] observed that distances between nitrogens of the nitrile group and those at positions 3 or 4 of the triazole heterocycle are similar to distances between oxygens in the structure of androstenedione.

Several findings have suggested that letrozole is a highly potent AI both *in vitro* and *in vivo* as well as exhibiting higher potency than several AIs such as aminoglutethimide, anastrozole, exemestane and formestane [9,11,12]. Owing to the success of letrozole, intense efforts have been invested in deriving novel AIs from this structural scaffold. In a series of investigations, Le Borgne et al. [34–36] synthesized several letrozole analogs possessing arylindole moiety with imidazole or 1,2,4-triazole. Furthermore, Farag et al. [37,38] synthesized several pyrazole-based letrozole and celecoxib analogs affording aromatase inhibitory activities. Moreover, Potter et al. [39–43] synthesized an array of sulfamate-containing letrozole, anastrozole and vorozole analogs, which provided interesting dual inhibitory activities toward aromatase and steroid sulfatase. Recently, Doiron et al. [33] sought out to investigate the aromatase inhibitory activity of 1,2,3-triazole instead of the 1,2,4-triazole present in letrozole. In their study, a library of substituents at the N1 position of the triazole heterocycle was examined by observing the influence of one or two phenyl/benzonitrile on the aromatase inhibitory activity. The authors concluded that nitrogen atom at positions 3 or 4 were both important for aromatase inhibition. A subset of compounds from their investigation comprising of fifty-four 1-substituted mono- and bis-benzonitrile or phenyl analogs of 1,2,3-triazole letrozole were compiled as a data set for QSAR investigation in this study (Table 1). A schematic representation of the computational methodology used herein is depicted in Fig. 2.

2.2. Exploring the chemical space of 1,2,3-triazole letrozole analogs

The linkage between the molecular structures of compounds with its respective biological activities is central to the QSAR paradigm. Molecular descriptors play a crucial role in providing numerical description of the physicochemical properties of molecules. In order to properly account for these structural features, it is essential that suitable descriptors be chosen for QSAR investigation. A handbook providing comprehensive coverage of molecular descriptors has been summarized by Todeschini and Consonni [44] while an in-depth review of quantum chemical descriptors was provided by Karelson et al. [45] In spite of the wide availability and relatively large number of molecular descriptors to choose from, this study selected a small subset of descriptors representing the general characteristics of a molecule (i.e. molecular size, flexibility, polarity, solubility, charge and electronic properties as well as chemical reactivity) and whose physicochemical properties can be easily understood. Such descriptors were recently used in exploring the chemical space of all known aromatase inhibitors [46].

Geometry optimization of the molecular structures was performed in a two-step fashion starting from an initial optimization with the semi-empirical AM1 method as to afford reasonably good starting structures followed by a subsequent and more refined optimization at the DFT level. The resulting low-energy conformers served as the basis for the extraction of 6 quantum chemical descriptors followed by a subsequent calculation using the Dragon

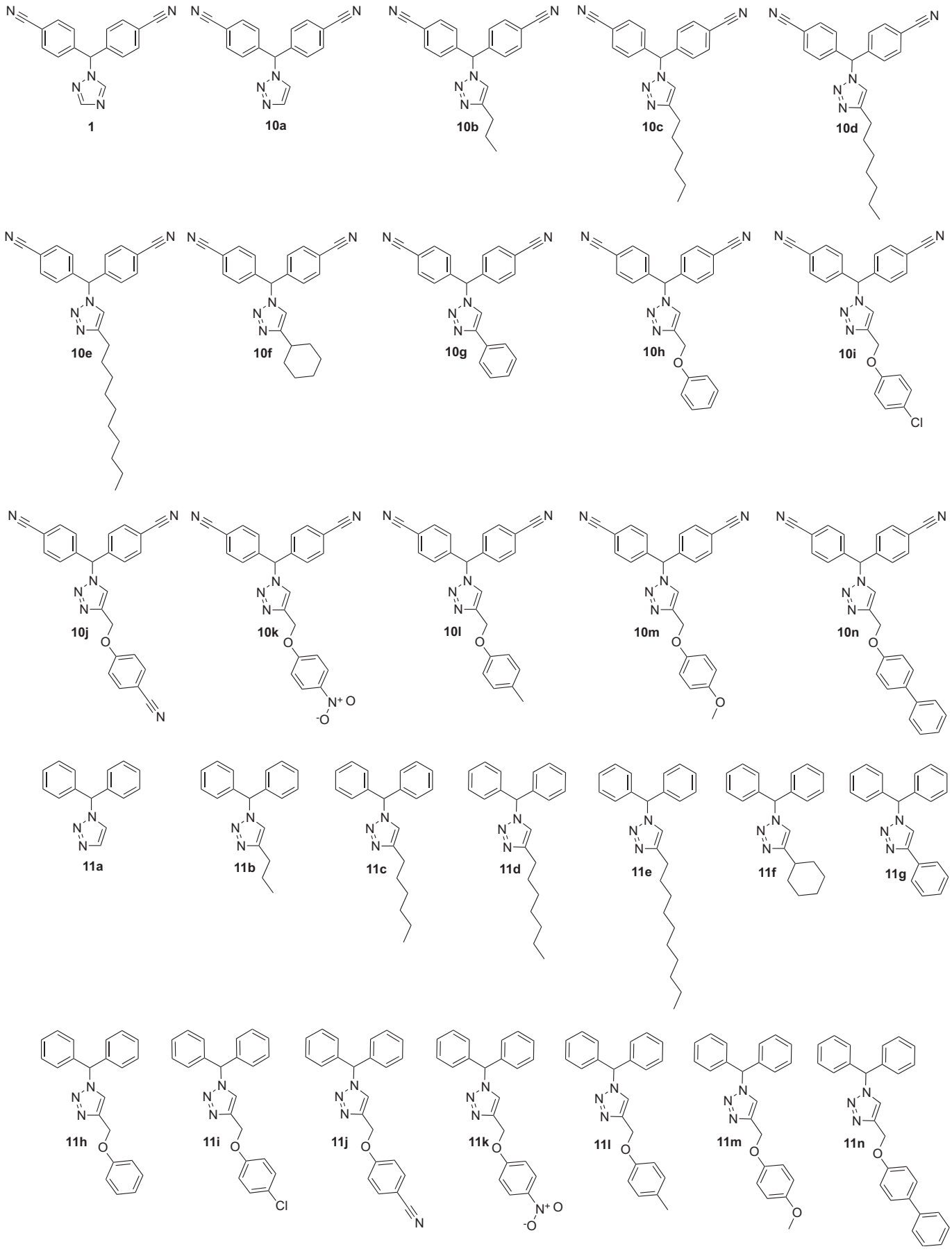

Fig. 1. Chemical structures of fifty-four 1-substituted 1,2,3-triazole analogs of letrozole.

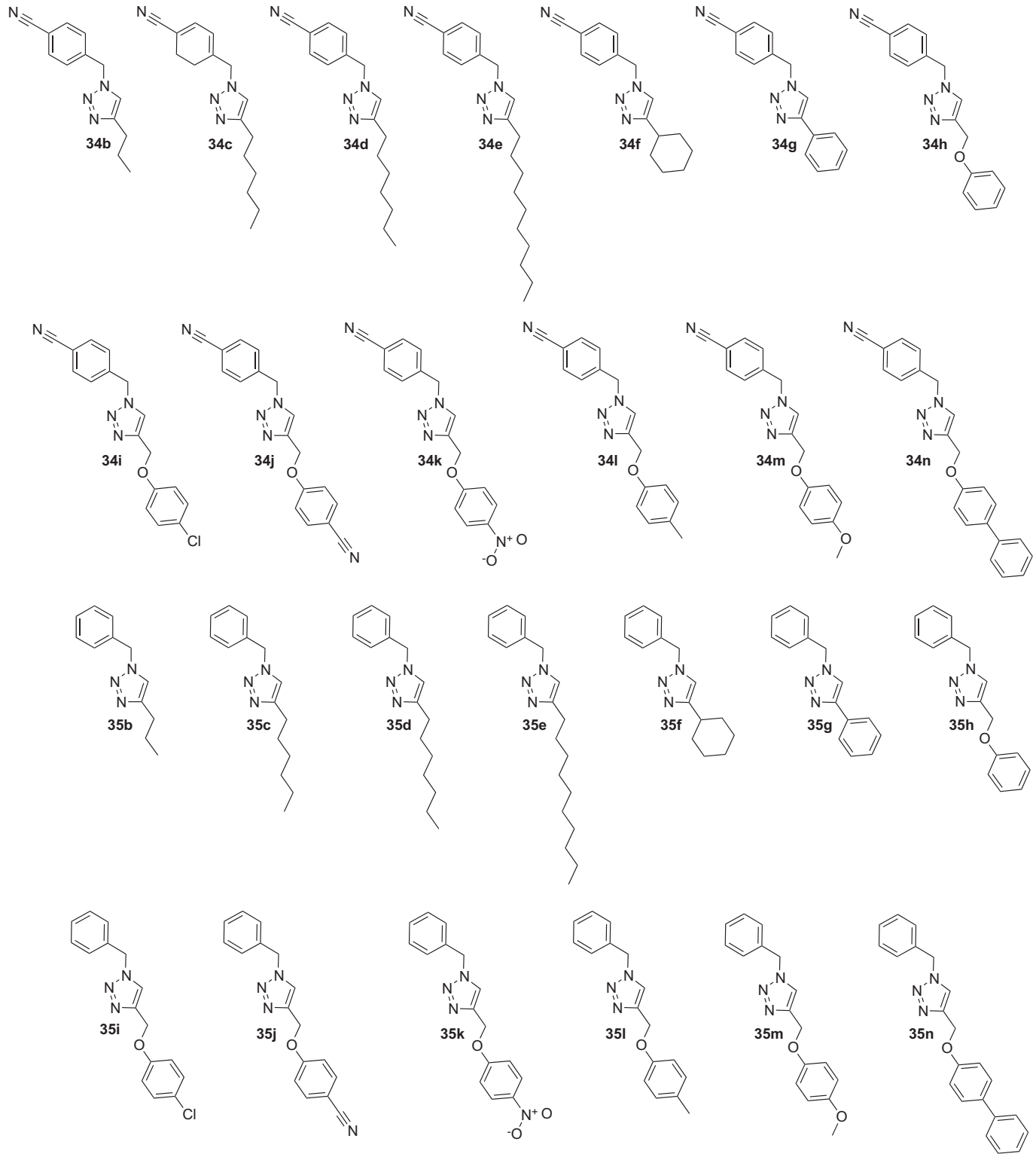

Fig. 1. (*continued*).

software package to produce an additional set of 6 molecular descriptors.

In order to visualize the relative distribution of the bioactivity and descriptor values, radar plots were created as shown in Figs. 3 and 4, respectively. It can be seen from the IC_{50} radar plot that compounds **11** occupied the most area on the plot indicating poor IC_{50} values while compounds **34** occupied the least area thereby suggesting potent IC_{50} values. The same but inverse trend is also deduced from the pIC_{50} radar plot where the most potent set of compounds occupied the most area of the plot while the least active set of compounds occupied the least area of the plot. Furthermore, it is observed from the IC_{50} values that the most

Table 1
Summary of the aromatase inhibitory activity of the letrozole analogs.

Name	IC_{50} (μM)	pIC_{50}	Name	IC_{50} (μM)	pIC_{50}
10a	0.008	8.10	**11n**	15.48	4.81
10b	4.7	5.33	**34b**	4.37	5.36
10c	5.07	5.29	**34c**	5.26	5.28
10d	10.6	4.97	**34d**	3.07	5.51
10e	16.02	4.80	**34e**	7.88	5.10
10f	19.32	4.71	**34f**	10.7	4.97
10g	10.33	4.99	**34g**	3.38	5.47
10h	10.97	4.96	**34h**	6.09	5.22
10i	16.11	4.79	**34i**	3.91	5.41
10j	4.64	5.33	**34j**	9.16	5.04
10k	7.73	5.11	**34k**	5.62	5.25
10l	10.95	4.96	**34l**	4.17	5.38
10m	12.16	4.92	**34m**	2.34	5.63
10n	13.35	4.87	**34n**	11.57	4.94
11a	4.58	5.34	**35b**	12.63	4.90
11b	10.96	4.96	**35c**	7.47	5.13
11c	14.83	4.83	**35d**	9.53	5.02
11d	19.36	4.71	**35e**	6.4	5.19
11e	13.54	4.87	**35f**	7.98	5.10
11f	13.72	4.86	**35g**	9.77	5.01
11g	12.94	4.89	**35h**	10.6	4.97
11h	31.26	4.51	**35i**	5.04	5.30
11i	13.37	4.87	**35j**	1.36	5.87
11j	16.41	4.78	**35k**	2.78	5.56
11k	11.61	4.94	**35l**	2.27	5.64
11l	26.77	4.57	**35m**	17.34	4.76
11m	15.21	4.82	**35n**	10.07	5.00

active set of compounds (in order of potency) were **34** > **35** > **10** > **11**, which had corresponding mean values of 5.963 ± 2.980, 7.942 ± 4.481, 10.919 ± 4.600 and 16.574 ± 5.990 μM, respectively. In parallel, an inverse trend is also observed from the pIC_{50} values where the most active set of compounds (in order of potency) had corresponding values of 5.273 ± 0.215, 5.188 ± 0.321, 5.003 ± 0.206 and 4.802 ± 0.133 M, respectively. A closer analysis of the molecular structures of compounds **10** revealed that **10b**, **10c**, **10j** and **10k** afforded the best aromatase inhibitory activity with IC_{50} values of 4.7, 5.07, 4.64 and 7.73 μM, respectively. A notable feature of the former 2 is that the substituents are small aliphatic moieties whereas the latter 2 are electrophilic moieties 4-methoxybenzonitrile and 4-nitroanisole, respectively. Class **11** generally had poor aromatase inhibitory activities and is the class with the worst activity. Of particular note is the good activity afforded by compounds **10a** and **11a**, which afforded IC_{50} values of 0.008 and 4.58 μM, respectively. The absence of substituents at the C4 position of the triazole ring of **10a** and **11a** is reminiscent of the letrozole drug with the exception that the 1,2,3-triazole ring of the former 2 compounds is in place of the 1,2,4-triazole of letrozole. In general, class **34** exhibited the best aromatase inhibitory activity. All compounds except for **34f**, **34j** and **34n** provided inhibitory activity less than 8 μM. Notably, compounds **34d**, **34g**, **34i** and **34m** exhibited the best activity in this class with IC_{50} values of 3.07, 3.38, 3.91 and 2.34 μM, respectively. Although class **35** provided the second best aromatase inhibitory activity there are several highly potent compounds namely **35j**, **35k** and **35l** with IC_{50} values of 1.36, 2.78 and 2.27 μM, respectively.

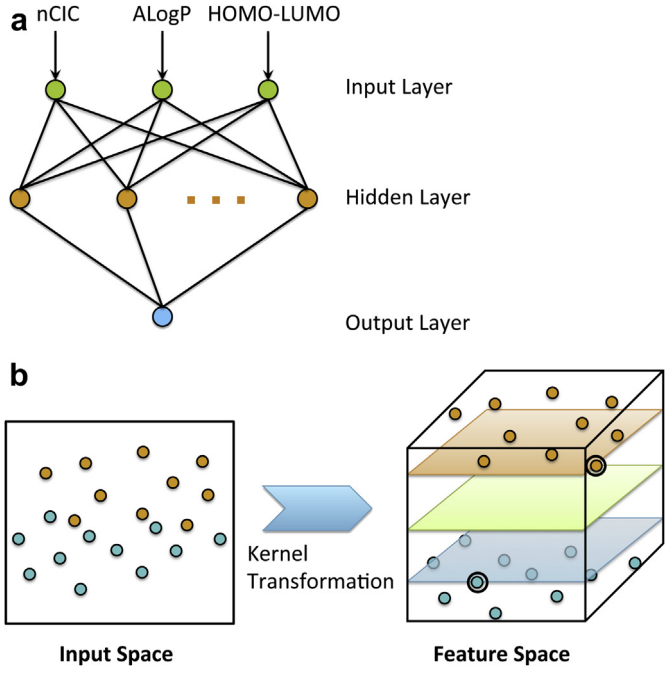

Fig. 2. General overview of the computational methodology employed in this study.

A more thorough analysis of highly potent compounds was performed in order to discern the essential physicochemical properties for good aromatase inhibitory activity. As previously mentioned, compound **10a** provided the best aromatase inhibitory activity in the investigated set of molecules. The compound has small molecular size with an MW of 285.33 Da (mean value of 328.070 Da), low flexibility with RBN of 3 (mean of 6.074), moderate hydrogen bond accepting capacity with nHAcc of 4 (mean of 3.574), high polarity as indicated by ALogP of 2.966 (mean of 4.767), TPSA of 78.290 (mean of 59.399) and Q_m of 0.482 (mean of 0.345). Furthermore, **10a** displayed the lowest μ value indicating low asymmetric distribution of charge in the molecule. Of all the investigated compounds, **10a** had the lowest HOMO value of −0.284 (mean of −0.247) and among one of the lowest LUMO value of −0.093 (minimum of −0.106) thereby indicating low propensity for electron donation and high capacity for electron acceptance, respectively. The compound had moderative chemical reactivity as indicated by a HOMO−LUMO value of 0.191 (mean of 0.182, minimum of 0.129 and maximum of 0.221).

The first Dragon descriptor MW considers the molecular size of compounds in the data set and it was observed that the largest molecule was **10** > **11** > **34** > **35**, which had corresponding values of 399.988 ± 37.717, 349.968 ± 37.717, 298.878 ± 37.717 and 273.868 ± 37.717, respectively. Such sequential ordering is not surprising as it is a direct function of the molecular substituents of the 4 sets of molecules (in order of decreasing bulkiness and size) namely the 2 benzonitrile, 2 phenyl, 1 benzonitrile and 1 phenyl moieties attached to the N1 position of the triazole ring of letrozole.

It should be noted that the standard deviation for all sets of molecules was the same. This can be attributed to the fact that all 4 sets of molecules each contained 13 compounds with the same group of 13 functional moieties at the C4 position of the 1,2,3-triazole ring. Such observation on the same variability of descriptor values is also true for the other 5 Dragon descriptors, which is to be expected, as the descriptors are either functional group counts or derived by summing descriptor values from their molecular fragments. In contrast, as will be seen in the following paragraphs that there are descriptor variability of quantum chemical descriptors for the 4 sets of molecules. This can be attributed to the fact that such descriptors were derived from quantum mechanical calculations in which the electronic structure is considered and perturbation to the molecular structure as a result of changes in functional moieties would also lead to alterations in the calculations physicochemical properties.

RBN accounted for the degree of flexibility in a molecule as obtained from the number of rotatable bonds and it was shown

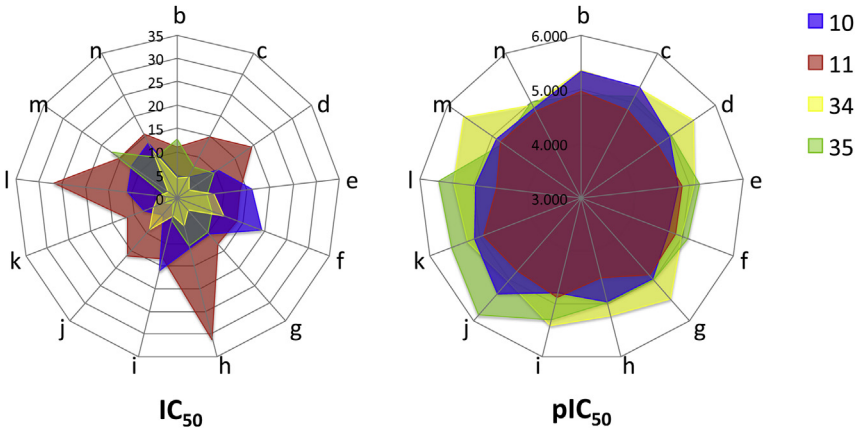

Fig. 3. Radar plots of aromatase inhibitory activity of 1-substituted 1,2,3-triazole analogs of letrozole shown in IC_{50} (a) and pIC_{50} (b) format.

that the most flexible set of molecules were (**10** = **11**) > (**34** = **35**) with corresponding values of 6.692 ± 2.136, 6.692 ± 2.136, 5.692 ± 2.136 and 5.692 ± 2.136, respectively. nCIC is a count of the number of rings present in a molecule and sets of molecule with the most rings were (**10** = **11**) > (**34** = **35**) with corresponding values of 3.769 ± 0.599, 3.769 ± 0.599, 2.769 ± 0.599 and 2.769 ± 0.599, respectively. The two former sets of molecules displayed higher RBN and nCIC than the latter two namely because there are two rings connected to the N1 position of the triazole ring as compared to one for the latter two.

nHAcc is essentially the number of hydrogen bond acceptors and sets of molecules with the highest number were compounds **10** > **34** > (**11** = **35**) having corresponding values of 4.846 ± 0.987, 2.846 ± 0.987, 3.846 ± 0.987 and 2.846 ± 0.987, respectively. It can

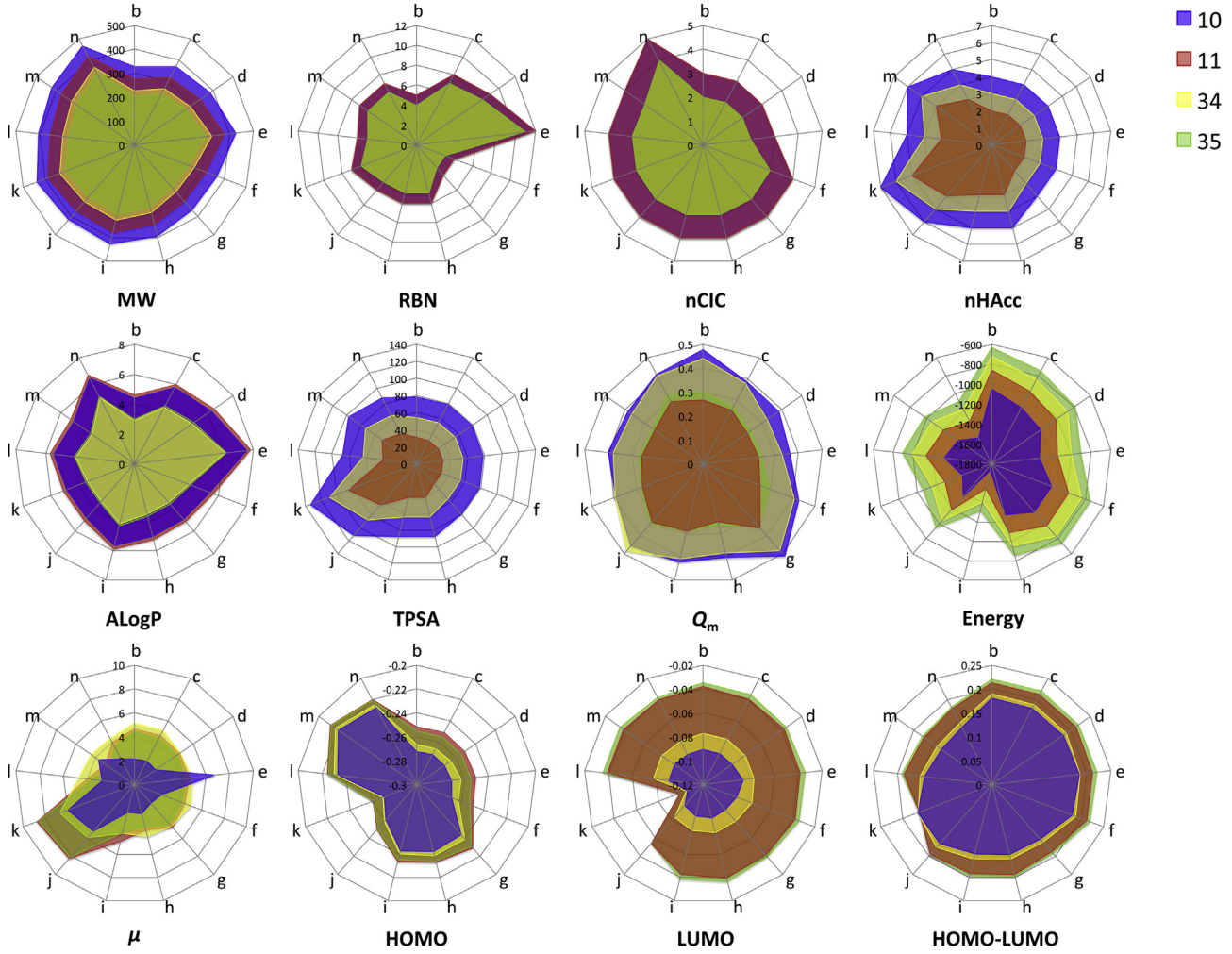

Fig. 4. Radar plots of all calculated molecular and quantum chemical descriptors.

be seen that molecules **11** and **35** displayed the same low mean value and this can be attributed to the fact that there is an absence of nitrile groups on the phenyl rings attached to the N1 position of the triazole ring. Furthermore, the fact that molecules **10** possessed the highest nHAcc value can be explained by the 2 additional nitrile groups of benzonitrile at the N1 position of triazole. Taking a closer look into the molecular structures, it is observed that compounds b–g accounted for 2 hydrogen bond acceptors arising from two basic nitrogens of triazole. Furthermore, 1 and 2 additional hydrogen bond acceptors are provided by the nitrile group of benzonitrile attached to the N1 position of the triazole ring. The remaining compounds h–n accounted for the remaining hydrogen bond acceptors.

ALogP provides a measure of a molecule's lipophilicity where high ALogP value indicates high lipophilicity while low value suggests low lipophilicity. The results indicated that compounds having the highest lipophilicity were **11** > **10** > **35** > **34** with corresponding values of 5.728 ± 0.859, 5.486 ± 0.859, 4.116 ± 0.859 and 3.995 ± 0.859, respectively. The former two sets of molecules possessed the highest lipophilicity namely because of the presence of two rings at the N1 position of triazole. The higher lipophilicity of **35** than **34** can be explained by the absence of the nitrile group from the rings at N1 position of triazole.

TPSA is essentially the sum of surfaces of polar atoms in a molecule and molecules with the highest value were **10** > **34** > (**11** = **35**) with corresponding values of 89.325 ± 16.379, 65.535 ± 16.379, 41.745 ± 16.379 and 41.745 ± 16.379, respectively. Molecules **10** and **34** displayed higher TPSA value than those of **11** and **35** namely because of the presence of benzonitrile moieties at the N1 position of the triazole ring of the former 2 classes of compound. The lack of polar functional moieties at the C4 position of the triazole ring of all 4 classes of compounds led to the same set of values for compounds b–g. However, the presence of polar functional groups at the C4 position of triazole in compounds h–n contributed to increases in the TPSA values.

As for the quantum chemical descriptor Q_m or the mean absolute charge of a molecule, it was observed that molecules with the highest Q_m were **10** > **34** > **35** > **11** having values of 0.418 ± 0.043, 0.399 ± 0.049, 0.285 ± 0.034 and 0.271 ± 0.037, respectively. The former 2 classes of compounds exhibited greater Q_m values than the latter two particularly owing to the presence of benzonitrile, which are strong electron withdrawing groups.

Energy is the total electronic energy present in a molecule and it was found that molecules with the highest values were **10** > **11** > **34** > **35** having values of −1316.771 ± 183.135, −1132.242 ± 183.135, −993.405 ± 183.135 and −901.140 ± 183.135, respectively. The former 2 classes possessed higher energy than the latter two, which can be explained by the fact that there are 2 rings in the former 2 classes as opposed to 1 in the former 2 classes. The presence of more cyclic rings translates to more atoms and as the total energy of a molecule is essentially a summation of the atomic energy of its constituents, therefore molecules having more atoms will also have higher molecular energy.

μ refers to the dipole moment and it essentially provides a measure of the asymmetric distribution of charges in a molecule. It can be seen that compounds with the highest dipole moment were compounds **11** > **34** > **35** > **10** with values of 4.960 ± 1.652, 4.943 ± 0.688, 4.629 ± 1.742 and 3.283 ± 1.633, respectively. It was observed that compounds bearing the benzonitrile moiety also had high dipole moment when compared to their phenyl counterpart. Such high value is associated with the asymmetric distribution of electrons as afforded by the strong electron withdrawing nature of benzonitrile.

HOMO can be defined as the electronic energy of the highest occupied molecular orbital and it provides a measure of the electron donating ability of a molecule. It was observed that values with the highest HOMO were **11** > **35** > **34** > **10** with corresponding values of −0.240 ± 0.015, −0.242 ± 0.016, −0.249 ± 0.018 and −0.254 ± 0.019, respectively. High HOMO values are indicative of a high propensity of the molecule to donate electrons to acceptor molecules. It can be deduced that the former 2 classes bearing phenyl rings at the N1 position of triazole afforded higher HOMO values than the latter 2 classes. This suggested that the lack of the nitrile group from the phenyl ring increases the HOMO values and thereby augmenting its electron donating ability. Substituents at the C4 position of compounds b–e and f are aliphatic and alicyclic hydrocarbons, respectively. Such structural features led to a decrease in the HOMO values, which subsequently leads to a reduction in electron donating propensity. The same is also observed for compounds j and k in which the substituents at the C4 position of the triazole ring are the electrophilic moieties 4-methoxybenzonitrile and 4-nitroanisole, respectively. The substituents of remaining compounds g–i and l–n are comprised of substituted aromatic rings and it follows that there are increases in the HOMO values, which suggests high propensity for electron donation to acceptor molecules.

LUMO refers to the electronic energy of the lowest unoccupied molecular orbital and it measures the electron accepting ability of a molecule. Low LUMO values suggest that the molecule has a high tendency to accept electrons from donor molecules. It can be seen that molecules with the lowest LUMO were **10** < **34** < **11** < **35** having corresponding values of −0.092 ± 0.005, −0.081 ± 0.008, −0.045 ± 0.017 and −0.042 ± 0.018, respectively. It is observed that the former 2 classes containing benzonitrile moieties at the N1 position of triazole contributed to lower LUMO values than the latter 2 classes. Such observation suggests that the presence of nitrile group on the phenyl ring reduces the LUMO values and consequently leads to enhanced capacity for electron acceptance. A closer observation into the substituent effects revealed that compounds b–f and l–n exhibited similar level of LUMO values while compounds g–l displayed slight reduction in the LUMO values. Notably, compounds j and k had low LUMO values; particularly the latter compound possessed the lowest LUMO values. Combining these observations together leads to the summary that compounds having the lowest LUMO values were in the following order (k < j) < (g–i) < (b–f, l–n).

HOMO–LUMO refers to the energetic difference of HOMO and LUMO. Such energetic gap can be used as an indicator of chemical reactivity where low energy gap correlates to high chemical reactivity and vice versa. It was observed that the most reactive set of molecules were **10** < **34b**< **11** < **35** with corresponding values of 0.161 ± 0.019, 0.169 ± 0.018, 0.195 ± 0.017 and 0.200 ± 0.019, respectively. Furthermore, it can be seen that the former 2 classes had the least HOMO–LUMO values while the latter 2 classes displayed the largest values, which corresponds to high and low chemical reactivity, respectively. A closer look into the molecular structure as to discern the substituent effects revealed that compounds b–f displayed the largest HOMO–LUMO gap, which suggests that these molecules are chemically stable and relatively inert. An observation of compounds **10** and **34** revealed that compounds with the lowest LUMO values were (in decreasing order) (l–n) < (g–i) < (j–k) < (b–f). For these set of compounds, it can be deduced that molecules l–n were the best electron acceptors. As for compounds **11** and **35** it was observed that compounds with the lowest LUMO values were k < (l–n) < (g–i) < j < (b–f). This revealed that compound k was the best electron acceptors.

2.3. Prediction of aromatase inhibition using multiple linear regression

Feature selection was performed on this set of 12 independent variables by constructing an intercorrelation matrix of Pearson's correlation coefficient values (Table 2). Redundant variables were identified as those having correlation coefficient values greater than 0.7 and one of the descriptor from the pair was subjected to removal. This resulted in the removal of 5 descriptors (i.e. MW, RBN, nHAcc, TPSA and LUMO) and therefore the retention of 7 - descriptors (i.e. nCIC, ALogP, Q_m, Energy, μ, HOMO and HOMO–LUMO) that were subsequently used hereafter for QSAR model building. Prior to performing multivariate analysis, the data set was pre-processed by converting the dependent y variables (i.e. IC_{50} values) to negative logarithmic scales (i.e. pIC_{50}).

As shown in Table 3, the first MLR model provided rather low predictive performance as observed from the R_{Tr} and Q_{CV} values of 0.6108 and 0.2692. An additional statistical parameter was employed to assess the predictive performance of the QSAR model. This statistical parameter is the F ratio, which is a statistical measure represented by the ratio of the explained variance to that of the unexplained variance. This first MLR model provided an F ratio of 0.5134 also suggesting low predictive performance. Outlying compounds were identified from the data set as those having maximal standardized residual and were subsequently removed. This was performed until no further improvements in F ratio were observed (data not shown), particularly this resulted in the construction of seven MLR models. It should be noted that a total of 7 compounds were identified as outliers and removed from the data set to produce a data set of 47 compounds. This resulted in a gradual increase of the predictive performance from Q_{CV} of 0.2692 in model 1 to 0.5579, 0.6019, 0.6370, 0.6779, 0.6709 and 0.6965 for models 2–7, respectively, which corresponded to F ratio values of 0.5134, 2.9051, 3.4897, 4.0971, 4.9804, 4.6773 and 5.2492. The MLR equation in model 7 is expressed as follows:

$$\begin{aligned} pIC_{50} = {} & -0.201(nCIC) - 0.0982(ALogP) - 0.3825(Q_m) \\ & - 0.0001(Energy) - 0.0034(\mu) - 0.9179(HOMO) \\ & - 4.1662(HOMO{-}LUMO) + 6.7649 \end{aligned} \quad (1)$$

$N = 47$, $R_{Tr} = 0.8054$, $RMSE_{Tr} = 0.1452$, $Q_{CV} = 0.6965$, $RMSE_{CV} = 0.1778$, F ratio $= 5.2492$.

To further enhance the model's performance, a further round of feature selection using the M5 algorithm led to the removal of 4 additional descriptors (i.e. Q_m, Energy, μ and HOMO) resulting in final set of 3 descriptors (i.e nCIC, ALogP and HOMO–LUMO). The MLR equation for model 8 is shown below:

$$\begin{aligned} pIC_{50} = {} & -0.1973(nCIC) - 0.0823(ALogP) \\ & - 3.7849(HOMO{-}LUMO) + 6.7933 \end{aligned} \quad (2)$$

$N = 47$, $R_{Tr} = 0.7998$, $RMSE_{Tr} = 0.1470$, $Q_{CV} = 0.7552$, $RMSE_{CV} = 0.1612$, F ratio $= 7.3952$.

Interpretation of the descriptor constituents of this equation suggested that the number of rings, aqueous solubility and chemical reactivity as deduced from the HOMO–LUMO gap are important descriptors for predicting the aromatase inhibitory activity.

A second trial of calculation was performed to assess the predictive power of MLR on an external test set comprising of 15% (i.e., 8 compounds from the original data set of 54 compounds). Thus, 46 compounds were used in the development of a predictive model and later tested on the 8 compound external test set. In the development of the 46 compound training and LOO-CV set, six compounds were identified as outlying compounds and were discarded from the data set. The MLR equation is described below:

$$\begin{aligned} pIC_{50} = {} & -0.2008(nCIC) - 0.0651(ALogP) \\ & - 6.0052(HOMO{-}LUMO) + 7.0829 \end{aligned} \quad (3)$$

$N = 40$, $R_{Tr} = 0.8145$, $RMSE_{Tr} = 0.1447$, $Q_{CV} = 0.7628$, $RMSE_{CV} = 0.1623$, F ratio $= 16.6988$, $Q_{Ext} = 0.6515$, $RMSE_{Ext} = 0.2212$.

As seen above, the MLR model performed moderately on the external test set. Statistical quality of the model is also summarized in Table 4 and plots of the predicted versus experimental pIC_{50} values are shown in Fig. 5a.

Table 3
Summary of predictive performance of MLR models.

Model	N	Training set		LOO-CV set		
		R_{Tr}	$RMSE_{Tr}$	Q_{CV}	$RMSE_{CV}$	F ratio
1	54	0.6108	0.3927	0.2692	0.5071	0.5134
2	53	0.7007	0.2032	0.5579	0.2403	2.9051
3	51	0.7449	0.1699	0.6019	0.2070	3.4897
4	50	0.7640	0.1638	0.6370	0.1986	4.0971
5	49	0.7897	0.1568	0.6779	0.1902	4.9804
6	48	0.7879	0.1512	0.6709	0.1844	4.6773
7	47	0.8054	0.1452	0.6965	0.1778	5.2492
8[a]	47	0.7998	0.1470	0.7552	0.1612	7.3952

[a] Feature selection using the M5 method was applied to reduce the number of descriptors from 7 to 3.

Table 2
Intercorrelation matrix of the molecular descriptors. Descriptors having high intercorrelation greater than 0.7 (shown as bold italic text) were subjected to removal.

	MW	**RBN**	nCIC	**nHAcc**	ALogP	**TPSA**	Q_m	Energy	Dipole	HOMO	**LUMO**	HOMO–LUMO
MW	1.000											
RBN	0.438	1.000										
nCIC	***0.741***	−0.149	1.000									
nHAcc	0.633	0.025	0.412	1.000								
ALogP	0.677	***0.741***	0.368	−0.087	1.000							
TPSA	0.602	0.041	0.320	***0.959***	−0.054	1.000						
Q_m	0.217	−0.333	0.151	0.546	−0.251	0.665	1.000					
Energy	***−0.911***	−0.288	−0.690	−0.649	−0.513	−0.588	−0.238	1.000				
Dipole	−0.111	0.093	−0.143	0.089	−0.142	0.077	−0.226	0.091	1.000			
HOMO	0.135	−0.128	0.446	−0.049	0.047	−0.287	−0.257	−0.180	−0.233	1.000		
LUMO	−0.461	−0.014	−0.159	***−0.807***	0.073	***−0.911***	***−0.771***	0.450	−0.039	0.416	1.000	
HOMO–LUMO	−0.582	0.080	−0.500	***−0.806***	0.041	***−0.737***	−0.614	0.604	0.133	−0.312	***0.734***	1.000

Table 4
Summary of predictive performance using MLR, ANN and SVM methods.

Method	Training set		LOO-CV			External test set	
	R_{Tr}	$RMSE_{Tr}$	Q_{CV}	$RMSE_{CV}$	F ratio	Q_{Ext}	$RMSE_{Ext}$
MLR	0.8145	0.1447	0.7628	0.1623	16.6988	0.6515	0.2212
ANN[a]	0.8242	0.1526	0.7594	0.1639	16.3479	0.6511	0.2228
SVM[b]	0.8786	0.1200	0.8326	0.1385	27.1163	0.6603	0.2210

[a] Optimal ANN parameters: nodes in the hidden layer, number of learning epochs, learning rate and momentum were 1, 40, 0.1 and 0.4, respectively.
[b] Optimal SVM parameters: C and γ were 2^3 and 2^{-1} for global grid search and C and γ were $2^{3.7}$ and $2^{-1.5}$ for local grid search.

2.4. Prediction of aromatase inhibition using artificial neural network

Data set of the final MLR model (comprising of 3 descriptors) was subjected to further model development using the back-propagation algorithm of ANN. A cartoon illustration of the network topology of ANN is shown in Fig. 6a. Before performing the ANN calculations, the data set was pre-processed by subjecting the independent x variables (i.e. molecular descriptors) to standardization as described by Eq. (2). Optimization of the following parameters was investigated in a sequential manner: number of nodes in the hidden layer, number of learning epochs, learning rate and momentum. The results indicated that the optimal number of nodes in the hidden layer is 1 (Fig. 7a) providing R_{Tr} and Q_{CV} values of 0.8122 and 0.6894, respectively, with corresponding $RMSE_{Tr}$ and $RMSE_{CV}$ values of 0.1658 and 0.1835, respectively. This resulted in a network topology of 3–1–1 corresponding to 3 nodes in the input layer, 1 node in the hidden layer and 1 node in the output layer. It is observed that the optimal number of learning epochs was 20 (Fig. 7b), which gave R_{Tr} and Q_{CV} values of 0.8105 and 0.7034, respectively, with corresponding $RMSE_{Tr}$ and $RMSE_{CV}$ values of 0.1666 and 0.1787, respectively. The results indicated that the optimal values for the learning rate and momentum were 0.1 and 0.6 (Fig. 7c), respectively. The use of these set of values further increased the predictive performance as deduced from R_{Tr} and Q_{CV} values of 0.8063 and 0.7218, respectively, with corresponding $RMSE_{Tr}$ and $RMSE_{CV}$ values of 0.1601 and 0.1719, respectively.

As shown in Table 5, optimization of the three ANN parameters progressively improved the predictive performance as deduced from the gradual reduction in the $RMSE_{CV}$ values from 0.1928 to 0.1835 then 0.1787 and finally to 0.1719. Furthermore, a similar but inverse trend was also observed for the Q_{CV} values, which demonstrated gradual increase from 0.6816 to 0.6894 then 0.7034 and finally to 0.7218. Moreover, it was observed that the F ratio improved from 4.8343 to 5.0463 then 5.4561 and finally to 6.0598.

The 40 compound data set from Eq. (5) using 3 descriptors was further used for model development using ANN. Optimization of ANN parameters afforded the following parameters: number of nodes in the hidden layer of 1, number of learning epochs of 40, learning rate of 0.1 and momentum of 0.4. As shown in Table 4, ANN afforded similar level of performance as that of MLR with R_{Tr} = 0.8242 and $RMSE_{Tr}$ = 0.1526 for the training set, Q_{CV} = 0.7594, $RMSE_{CV}$ = 0.1639 and F ratio = 16.3479 for the LOO-CV set whereas the external test set provided Q_{Ext} = 0.6511 and $RMSE_{Ext}$ = 0.2228. A plot depicting experimental versus predicted activities of compounds as modeled by ANN is shown in Fig. 5b.

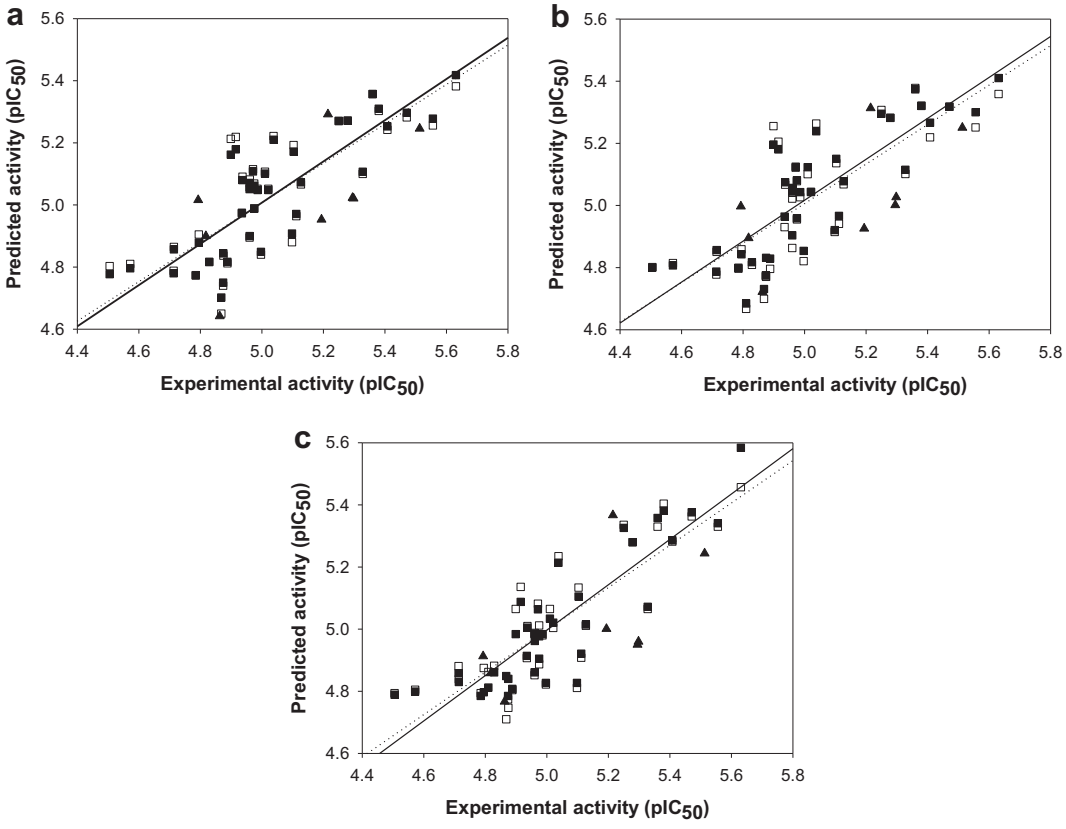

Fig. 5. Plot of experimental versus predicted pIC_{50} values of aromatase inhibition for QSAR models generated by MLR (a), ANN (b) and SVM (c) for the training set (■; regression line is represented as dotted line), the leave-one-out cross-validated testing set (□; regression line is represented as solid line) and external testing set (▲).

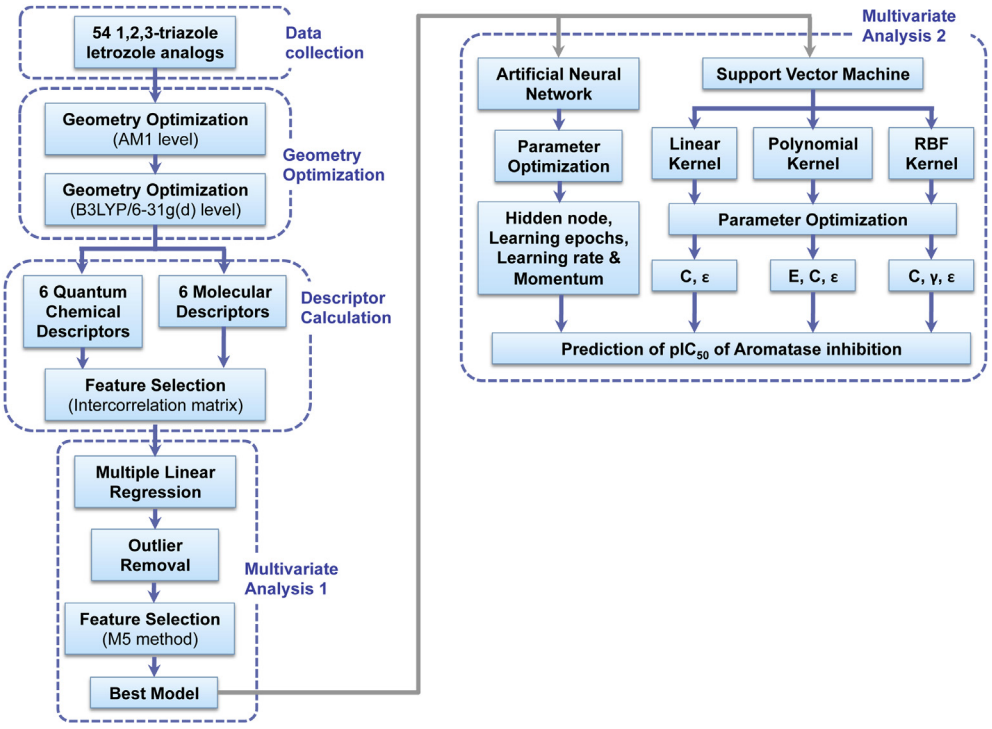

Fig. 6. Schematic illustration of the general concept of artificial neural network (a) and support vector machine (b).

2.5. Prediction of aromatase inhibition using support vector machine

The final MLR model was also utilized for QSAR model building using SVM by making use of three kernel functions (i.e. linear, polynomial and RBF). The best SVM parameters were obtained from parameter optimization using in-house developed Python scripts. Such optimization was performed by means of a two-level grid search comprising of an initial global search and subsequently by a more refined local search.

Results from the global grid search of SVM parameters are summarized in Table 6. It can be seen that the worst kernel function for predicting aromatase inhibition was the polynomial kernel (Fig. 8b), which provided R_{Tr}, Q_{CV} and F ratio values of 0.6715, 0.1892 and 4.5753, respectively. This was achieved using the best exponential value of 2 (from among the tested values of 2–10), which is used in conjunction with a C value of 2^{-1}. Furthermore, it is observed that the best prediction results were afforded by the RBF kernel (Fig. 8c), which is deduced from R_{Tr}, Q_{CV} and F ratio values of 0.7755, 0.1608 and 8.4061, respectively. The optimal C and γ values for the RBF kernel were 2^{-3} and 2^{1}, respectively. Finally, the linear kernel (Fig. 8a) afforded the second best predictive performance as shown by the R_{Tr}, Q_{CV} and F ratio values of 0.7527, 0.1652 and 7.2825, respectively. This was obtained using the optimal C value of 2^{-1}. The RBF kernel was selected for further model building by performing a refined parameter search.

SVM parameters C and γ were optimized by investigating the ranges of 2^{1} to 2^{5} and 2^{-1} to 2^{3}, respectively, using interval steps of $2^{0.1}$. Results from this local grid search are shown in Table 7 and Fig. 8a. It is observed that the optimal C value was $2^{-2.2}$ while the best γ value was 2^{3}. This set of SVM parameter augmented the prediction performance as deduced from the increase of Q_{CV} value from 0.7755 to 0.7896 along with a corresponding increase of the F ratio from 8.4061 to 9.2253.

The default ε value used for all SVM calculations thus far was 0.001. An empirical search for the optimal ε value was performed in the range of 0.00001–0.01 initially using an incremental step size of 0.00001 (for the range of 0.00001–0.0001) and later using an incremental step size of 0.0001 (for the range from 0.0001 to 0.01). Results indicated that there were no further improvements in the predictive performance when compared to the default value. It can be seen from Fig. 9b that the performance deteriorates as the ε value increases from the default value whereas no significant changes were observed as the ε value decreases from the default value.

The data set from MLR model as described in Eq. (2) comprising of 40 compounds and using 3 descriptors were also employed for QSAR model building using SVM with the RBF kernel function, which was found to be the best kernel function as previously described. Optimization of SVM parameters (i.e., C and γ) was performed in two sequential steps in which the initial global grid search indicated the preliminary optimal C and γ values to be 2^{3} and 2^{-1}, respectively. A subsequent local grid search identified the optimal C and γ values to be $2^{3.7}$ and $2^{-1.5}$, respectively. This set of parameters was then used in the construction of SVM model. As shown in Table 4, it was found that SVM models provided better predictive performance than those of MLR and ANN as deduced from the following statistical parameters: R_{Tr} = 0.8786 and $RMSE_{Tr}$ = 0.1200 for the training set, Q_{CV} = 0.8326, $RMSE_{CV}$ = 0.1385 and F ratio = 27.1163 for the LOO-CV set while the external test set afforded Q_{Ext} = 0.6603 and $RMSE_{Ext}$ = 0.2210. A plot of experimental versus predicted activities of compounds as modeled by SVM is shown in Fig. 5c.

3. Conclusion

The present study explores the structure–activity relationship for a library of fifty-four 1-substituted 1,2,3-triazole analogs of letrozole with known aromatase inhibitory activity. The molecular

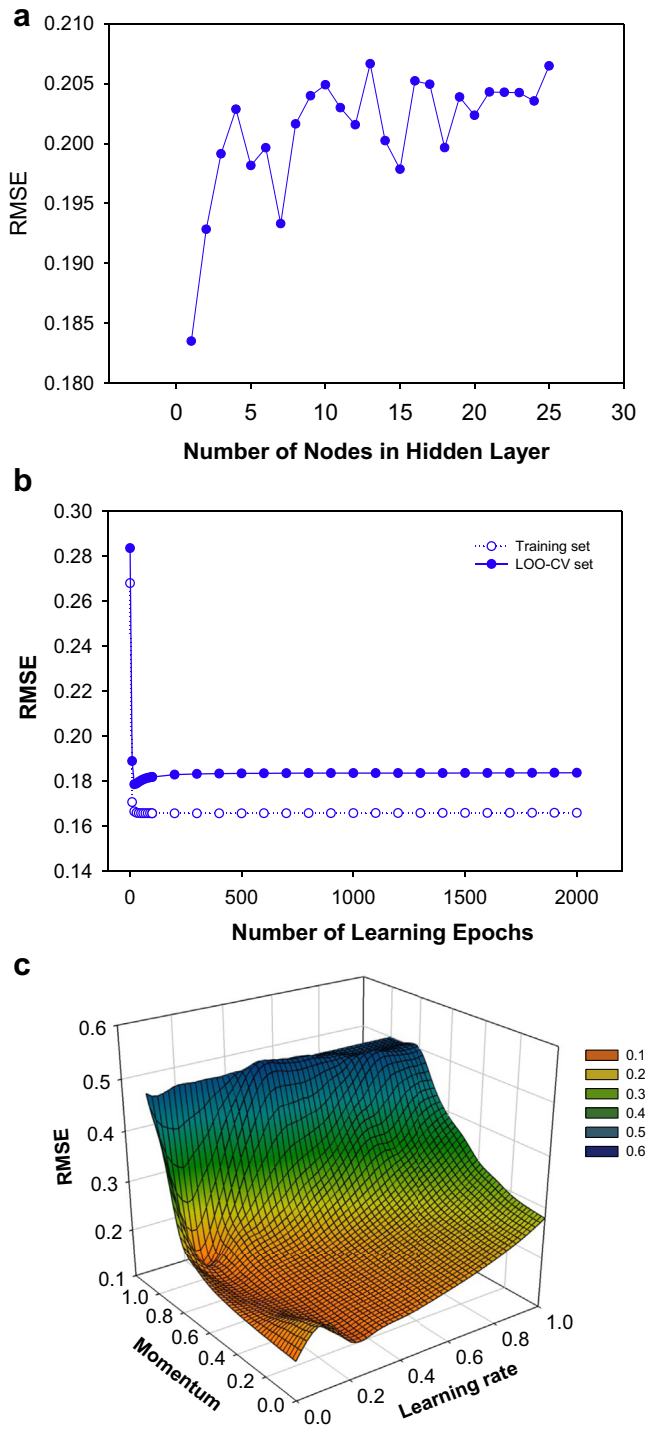

Fig. 7. Optimization of ANN parameters comprising of the number of nodes in the hidden layer (a), number of learning epochs (b) as well as the learning rate and momentum (c).

structures of the compounds were geometrically optimized initially using the semi-empirical AM1 method followed by further optimization at DFT level. A diverse set of quantum chemical and molecular descriptors were utilized to provide numerical description of the investigated compounds. Exploration of the chemical space of letrozole analogs as performed by analyzing the calculated molecular descriptors provided pertinent insights into the molecular features of the compounds. Feature selection by means of intercorrelation matrix and M5 method afforded a subset of important descriptors. Particularly, this set of descriptors is comprised of nCIC, ALogP and HOMO–LUMO suggesting that the number of rings, aqueous solubility and chemical reactivity may be critical in predicting aromatase inhibitory activity. Of all the multivariate analysis methods used, the results indicated the following order of accuracy SVM > MLR > ANN, where SVM employing the RBF kernel provided the highest predictive performance for modeling the activity of aromatase inhibition. Although SVM provided the highest accuracy it does not provide discernible rules as it is inherently a black-box approach. In contrast, regression coefficients as provided by MLR equation provided an easy-to-interpret metric that may be used to explain the relative importance of descriptors. In spite of this, both methods may be used in combination for predictive modeling as to complement the inherent shortcomings of the other method. It is anticipated that the constructed QSAR models proposed herein may be used to guide the further design of novel aromatase inhibitors with robust inhibitory properties.

4. Materials and methods

4.1. Data set

The experimental IC_{50} data of fifty-four 1-substituted mono- and bis-benzonitrile or phenyl analogs of 1,2,3-triazole letrozole (Table 1 and Fig. 1) were obtained from the work of Doiron et al. [33] The compounds were synthesized by click chemistry via Cu(I)-catalyzed Huisgen 1,3-dipolar cycloaddition onto azides bearing aliphatic alkynes or substituted propargyl phenol ethers. Nucleophilic substitution onto brominated precursors gave rise to mono-benzonitrile and phenyl analogs while base-induced condensation with 4-fluorobenzonitrile onto mono-benzonitrile precursors gave rise to bis-benzonitrile analogs.

The reported IC_{50} values of the compounds were in the range of 0.008 and 31.26 μM. In order to afford a more uniform distribution of the dependent variable, negative logarithmic transformation to the base of 10 was performed on the IC_{50} values. Such transformation is described by the following equation:

$$pIC_{50} = -\log(IC_{50}) \quad (4)$$

and it is worthy to note that the IC_{50} values were first converted from μM to M unit prior to the transformation. This resulted in pIC_{50} values in the range of 4.505 and 8.097 M.

4.2. Geometry optimization

Molecular structures of letrozole analogs were drawn into VIDA [47] and converted to suitable file format using Babel [48]. Low-energy conformers were obtained by geometrically optimizing the structures in Gaussian 09 [49] preliminarily at the semi-empirical AM1 method as to afford reasonably good starting structures. This is followed by full optimization with no symmetry constraint at the density functional theory (DFT) level using Becke's three-parameter Lee–Yang–Parr functional [50,51] with the 6–31g(d) basis set.

4.3. Descriptor calculation and feature selection

Quantum chemistry considers the electron density in a molecule by solving the Schrödinger equation. Such quantum chemical descriptors are essentially numerical descriptions of the molecular structures and have been demonstrated to be useful in explaining the origins of biological activities [52–56] and chemical properties

Table 5
Summary of ANN parameter optimization and their predictive performance.

Model	Parameters				Training set		LOO-CV set		
	No. of hidden nodes	No. of learning epochs	Learning rate	Momentum	R_{Tr}	$RMSE_{Tr}$	Q_{CV}	$RMSE_{CV}$	*F* ratio
Default parameters	2	500	0.3	0.2	0.8246	0.1642	0.6816	0.1928	4.8343
Optimal hidden nodes	1	500	0.3	0.2	0.8122	0.1658	0.6894	0.1835	5.0463
Optimal learning epochs	1	20	0.3	0.2	0.8105	0.1666	0.7034	0.1787	5.4561
Optimal parameters	1	20	0.1	0.6	0.8063	0.1601	0.7218	0.1719	6.0598

[57–59]. The quantum chemical descriptors were extracted from the resulting low-energy conformers as obtained from Gaussian 09 using an in-house developed script written in Python. These descriptors included mean absolute charge (Q_m), energy, dipole moment (μ), highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), energy gap of the HOMO and LUMO state (HOMO–LUMO). Furthermore, an additional set of molecular descriptors were obtained from Dragon version 5.5 [60] using low-energy conformers as input files and are comprised of molecular weight (MW), rotatable bond number (RBN), number of rings (nCIC), number of hydrogen bond acceptors (nHAcc), Ghose-Crippen octanol–water partition coefficient (ALogP) and topological polar surface area (TPSA). Full details of the data set comprising of the name of the compounds, chemical structure in SMILES notation and values of all 12 descriptors employed in this study are provided as Supplementary Data.

Both sets of descriptors comprising of 6 quantum chemical descriptors and 6 molecular descriptors were then analyzed for redundancy. This was performed by first constructing an inter-correlation matrix of Pearson's correlation coefficient values for all descriptors (Table 2). Subsequently, collinear and redundant variables were identified using correlation coefficient cut-off value of 0.7. Particularly, for any given pair of descriptors exhibiting a correlation coefficient value exceeding 0.7 one of the variables was subjected to removal.

4.4. Data scaling

As to allow comparability among the independent variables, the descriptors were scaled to zero mean and unit variance by performing standardization as described by the following equation:

$$x_i^{stdn} = \frac{x_i - \overline{x}_i}{\sigma_i} \quad (5)$$

where x_i^{stdn} represents the standardized *i*th descriptor, x_i represents the *i*th descriptor value of interest, $\overline{x}_i$ represents the mean value of the *i*th descriptor and σ_i represents the standard deviation of the *i*th descriptor.

4.5. Internal validation of predictive models

The data set was separated as training and testing sets by means of leave-one-out cross-validation (LOO-CV). This process was performed by leaving out one data sample as the testing set and uses the remaining $N - 1$ samples as the training set. Such process is carried out for *N* times by leaving out a different sample as the testing set so that all samples had a chance to be left out. The benefit of this data sampling approach is two-folds. First, this approach makes the most use of all available data by employing all samples from the data set. Second, as there is no random sampling involved the results obtained from this sampling approach will always be the same thus there will be no need for repeating the calculations iteratively. However, the inherent drawback of this approach is the high computational cost of performing calculations for *N* times meaning that it is not suitable for big data sets. In spite of this, LOO-CV is an attractive approach for making the most economical use of small data sets [61].

Table 6
Summary of global search of optimal SVM parameters and their predictive performance.

Kernel	Optimal parameters[a]			Training set		LOO-CV set		
	E	*C*	γ	R_{Tr}	$RMSE_{Tr}$	Q_{CV}	$RMSE_{CV}$	*F* ratio
Linear		−1		0.7859	0.1575	0.7527	0.1652	7.2825
Polynomial	2	−1		0.7417	0.1694	0.6715	0.1892	4.5753
RBF		−3	1	0.8304	0.1484	0.7755	0.1608	8.4061

[a] Default ε value of 0.001 was used for all computations.

4.6. External validation of predictive models

Aside from internal validation of predictive models, a second round of model building was performed by dividing the data set into 2 partition: (i) subset for training and LOO-CV comprising of 46 compounds as well as (ii) subset for external validation comprising of 8 compounds (i.e., **10c**, **10i**, **11f**, **11m**, **34d**, **34k**, **35e** and **35i**) or 15% of the data set. Particularly, the former was used for internal validation and the latter for external validation.

4.7. Multivariate analysis

Three multivariate analysis methods were employed in this study namely multiple linear regression, artificial neural network and support vector machine. These approaches seek to reveal the correlation that exists between the structure (i.e. the descriptors) and activity (i.e. the aromatase inhibitory activity). All calculations were performed using the Weka software package [62].

4.7.1. Multiple linear regression

MLR discerns such structure–activity relationship according to the following linear equation:

$$y = m_1x_1 + m_2x_2 + \cdots + m_nx_n + b \quad (6)$$

where *y* refers to the pIC_{50} values, *m* are the regression coefficients for molecular descriptors *x* (encompassing descriptor 1 to *n*) and *b* denotes the *y*-intercept value. MLR essentially represents the dependent variable *y* as a linear combination of the independent *x* variables.

Each compound from the data set is represented by a set of independent *x* variables and its corresponding dependent variable *y*. The predicted value of *y* for the first compound (denoted by the value of 1 in superscript) from the data set can be expressed as:

$$y^{(1)}_{predicted} = \sum_{i=1}^{n} m_i x_i^{(1)} = m_1x_1^{(1)} + m_1x_2^{(1)} + \cdots + m_nx_n^{(1)} \quad (7)$$

MLR seeks to minimize the sum of squares of the residual values (i.e. the difference between the predicted and actual values) over all

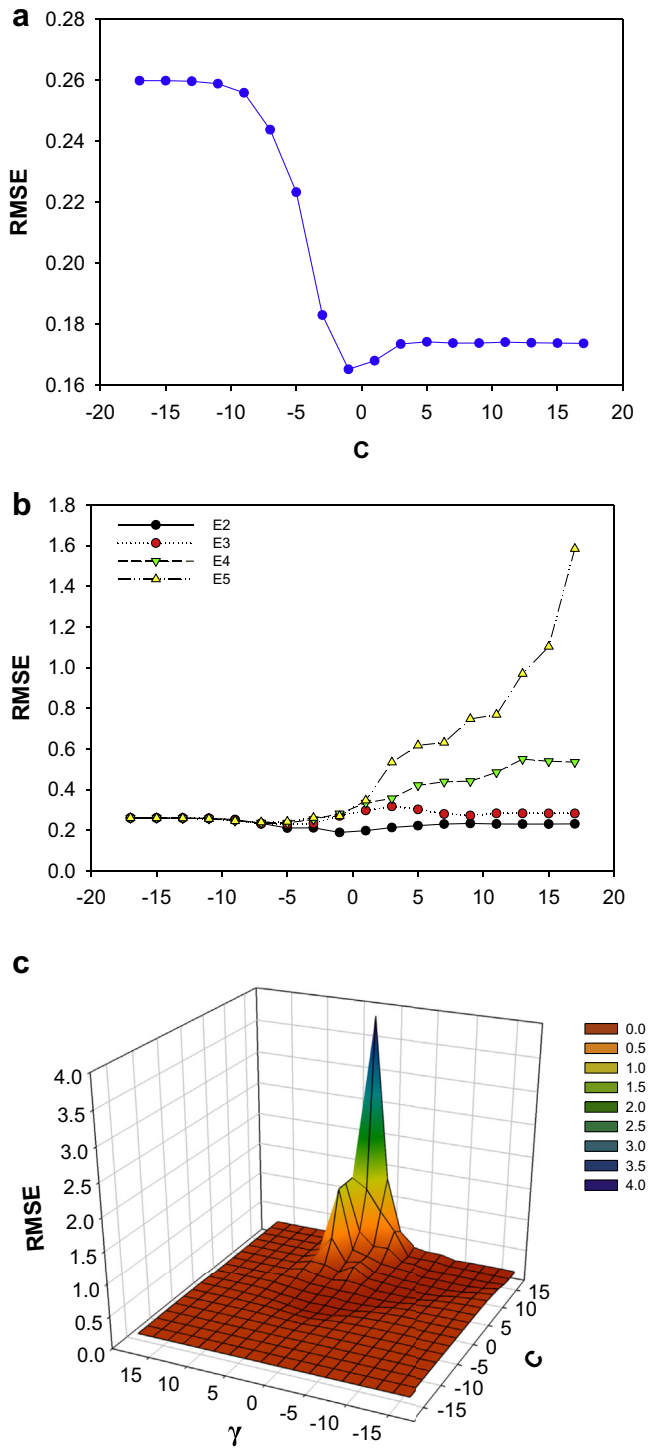

Fig. 8. Global grid search of parameters for SVM using linear (a), polynomial and radial basis function (c) kernels.

Table 7
Summary of local search of optimal SVM parameters using RBF kernel and their predictive performance.

Optimal parameters			Training set		LOO-CV set		
C	γ	ε	R_{Tr}	$RMSE_{Tr}$	Q_{CV}	$RMSE_{CV}$	F ratio
−2.2	3	0.001[a]	0.9131	0.1149	0.7896	0.1554	9.2253

[a] Default ε value of 0.001 was used.

training samples. Therefore, the sum of squares of the residuals of each jth sample for sample size N can be represented as:

$$\sum_{j=1}^{N}\left(y_{\text{actual}}^{(1)} - y_{\text{predicted}}^{(1)}\right)^2 \tag{8}$$

4.7.2. Artificial neural network

ANN is a supervised machine learning method and a simplistic representation of the human brain. The neuronal nodes in an ANN system are comparable to the human neuron while the dendrites and axons that interconnect the neurons are computationally represented by synaptic weights. Such weights of the ANN system are adjusted in an on-going manner as the learning process proceeds. A typical ANN system is comprised of three layers: input, hidden and output. The values of the independent variables are relayed directly to the nodes of the input layer. Signals are subsequently sent in a feed-forward manner to the hidden layer and finally onto the output layer via synaptic weights (i.e. the connection between nodes of each adjacent layer). Neurons of the hidden layer contain a sigmoidal transfer function that essentially limits the output signal to 0 and 1 as described by the following equation:

$$\text{sf}(\text{input}) = \frac{1}{1+e^{-\text{input}}} \tag{9}$$

Output neuron having a numerical class is an unthresholded linear unit. The inner workings of this network can be summarized by the equation:

$$\hat{y} = f(x) = \sum_{h}\left[\text{sf}\left(\sum_{i} x_i w_{ih} - \theta_i\right)\right] w_h - \theta_h \tag{10}$$

where w_{ih} represents the synaptic weight between input node i and hidden node h, w_h denotes the synaptic weight between the hidden node and output node y. θ_i and θ_h refer to biases of the input and hidden layers. The target error is calculated as the difference between $\hat{y}$ and y, which is back-propagated from the output layer through the hidden layer onto the input layer followed by readjustment of the synaptic weights in order to obtain good prediction. This is performed in an iterative manner until the designated number of learning epochs is reached. As the initial values of the synaptic weights are randomly assigned at the onset of the learning process and this may result in slightly varying prediction. Therefore, ten calculations were performed and the average value from these runs was used.

The optimal set of values for the ANN parameters (i.e. number of hidden nodes, number of learning epochs, learning rate and momentum) were obtained through an empirical systematic search using RMSE as a measure of predictive performance.

4.7.3. Support vector machine

SVM represents a supervised machine learning method based on the statistical learning theory proposed by Vapnik and coworkers [63,64]. SVM calculations performed herein are based on John Platt's Sequential Minimal Optimization (SMO) algorithm [65]. Several review articles [66–68] and books [69,70] are available for in-depth description of the SVM theory. Briefly, SVM was traditionally a linear learning classifier but has been adapted to solve non-linear regression problems by considering the ε-insensitive loss function:

$$L_\varepsilon(y, f(x,w)) = \begin{cases} |y - f(x,w)| - \varepsilon & \text{for } |y - f(x,w)| \geq \varepsilon \\ 0 \quad \text{otherwise} \end{cases} \tag{11}$$

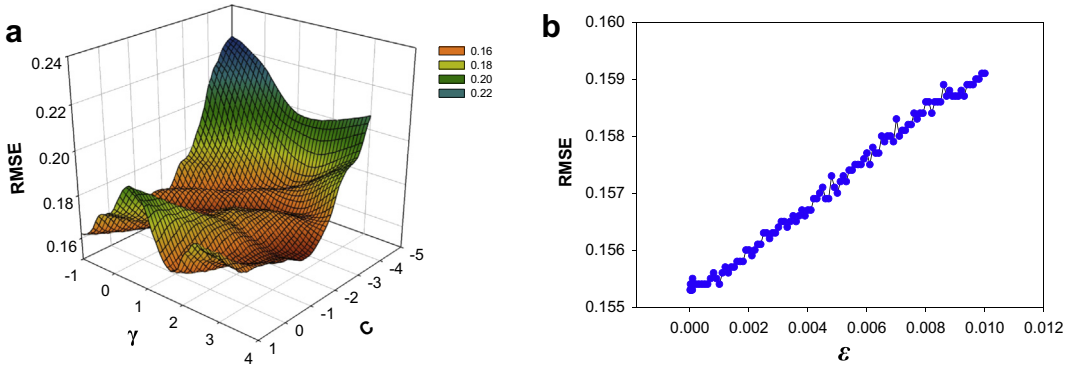

Fig. 9. Local grid search of parameters for SVM using radial basis function kernel (a) followed by optimization of the ε value (b).

where ε denotes the tube size that accounts for the approximation accuracy of training data samples. Support vector regression seeks to uncover for all training data samples an $f(x)$ function where there is a maximum of ε deviation from experimental values of y_i while trying to minimize the flatness. This means that the loss function places negligible consideration on errors as long as the value is less than ε. In any cases, it forbids significant deviation from the value.

Linear SVM are suitable for samples that are separable by linear mapping of their feature vectors. In a non-linear SVM the input vector X are projected into a higher dimensional feature space using kernel function:

$$K(x,y) = \langle \varphi(x) \cdot \varphi(y) \rangle \tag{12}$$

where K represents the kernel function and φ is a mapping function from the original input space into the feature space.

Typical kernel functions include linear, polynomial and radial basis function (RBF). The polynomial kernel can be described by the following equation:

$$K(x,y) = (\langle x,y \rangle + 1)^E \tag{13}$$

where E designates the exponent value while a linear kernel will have a value of 1.

The RBF kernel can be described by the following equation:

$$K(x,y) = e^{-\left(\gamma \cdot \langle x-y, x-y \rangle^2\right)} \tag{14}$$

Subsequently, a linear model is constructed in this higher dimensional feature space. Essentially, a data set comprising of m compounds with known aromatase inhibitory activity y_i (i.e. pIC_{50}) and physicochemical descriptors x_i can be described as $\{(x_i, y_i)\}_{i=1}^m$. Structure–activity correlations can be represented as $y_i = f(x_i)$ where the $f(x_i)$ term is defined by a linear function in the form of:

$$f(x_i) = \langle w_i, x_i \rangle + b \tag{15}$$

where w_i is the weight vector of the linear function of x_i and b is the threshold coefficient. Data in the high dimensional feature space is then subjected to the following linear function:

$$y = \sum_{i=1}^{m} w_i \varphi(x_i) + b \tag{16}$$

where $\{\varphi(x_i)\}_{i=1}^m$ are features of the input variables after kernel transformation whereas $\{w_i\}_{i=1}^m$ and b are coefficients. Data in such higher dimensional feature space are now linearly separable by hyperplane. The *maximal-margin hyperplane* maximizes distance between classes by being centrally located in between the hyperplanes of each data classes. *Margins* can be defined as the distance between these two hyperplanes while *support vectors* refer to samples lying on the hyperplanes. A schematic illustration on the concepts of SVM is depicted in Fig. 6b. Particularly, the illustration shows the non-linear mapping of data samples from the input space into the higher dimensional feature space followed by subsequent linear learning on the transformed space.

Empirical search for the optimal set of SVM parameters was performed as to achieve good generalization performance. Parameters that were investigated are comprised of: (i) C, or the complexity parameter, searches for a balance between misclassification and simplicity of the decision surface, (ii) γ determines the extent at which one training sample has on the model and (iii) E refers to the exponential value in a polynomial kernel where a value of 1 is essentially a linear kernel. Parameter optimizations were performed utilizing a two-level grid search. This is comprised of an initial coarse grid search that systematically adjusts the exponential n values in the form of 2^n for the parameters (i.e. C, γ). Subsequently, a more refined local grid search was then carried out on regions from the coarse grid search affording good performance. A step size of 2 was used for the coarse grid search whereas a step size of 0.1 was used for the local grid search. The RMSE value was used as a measure of the prediction performance.

4.8. Statistical assessment of QSAR models

Predictive performance of QSAR models was assessed from three statistical parameters: root mean squared error (RMSE), Pearson's correlation coefficient for the training set (R) and cross-validation testing set (Q) as well as the Fisher (F) ratio.

The RMSE values were calculated according to the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{\mathrm{exp}} - y_{\mathrm{pred}}\right)^2} \tag{17}$$

where N, y_{exp} and y_{pred} represents the sample size, experimental values and predicted values, respectively.

Pearson's correlation coefficient were calculated as described by the following equation:

$$r = \frac{\sum_{i=1}^{N}\left(y_{\mathrm{exp}} - \overline{y_{\mathrm{exp}}}\right)\left(y_{\mathrm{pred}} - \overline{y_{\mathrm{pred}}}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_{\mathrm{exp}} - \overline{y_{\mathrm{exp}}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_{\mathrm{pred}} - \overline{y_{\mathrm{pred}}}\right)^2}} \tag{18}$$

The F ratio between the explained (R^2) and unexplained ($1 - R^2$)

variance using m and $n - m - 1$ degrees of freedom can be calculated according to the following equation:

$$F \text{ ratio} = \frac{R^2/m}{(1-R^2)/(n-m-1)} \quad (19)$$

where m is the number of independent variables and n is the number of compounds in the data set. The critical F value for each set of m and $n - m - 1$ degrees of freedom was obtained from the F distribution table.

Compounds having maximal standardized residual were identified as outliers and were therefore removed from the QSAR models until no further improvement in F ratio was observed.

Acknowledgments

C.N. gratefully acknowledges financial support from the Goal-Oriented Research Grant of Mahidol University. This project is also supported in part by the Office of the Higher Education Commission and Mahidol University under the National Research Universities Initiative.

Appendix A. Supplementary material

Supplementary material associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ejmech.2013.08.015.

References

[1] A. Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, D. Forman, Global cancer statistics, CA Cancer J. Clin. 61 (2011) 69–90.
[2] A.A. van Landeghem, J. Poortman, M. Nabuurs, J.H. Thijssen, Endogenous concentration and subcellular distribution of androgens in normal and malignant human breast tissue, Cancer Res. 45 (1985) 2907–2912.
[3] W. Yue, J.P. Wang, C.J. Hamilton, L.M. Demers, R.J. Santen, In situ aromatization enhances breast tumor estradiol levels and cellular proliferation, Cancer Res. 58 (1998) 927–932.
[4] A.M.H. Brodie, Aromatase inhibitors in the treatment of breast cancer, J. Steroid Biochem. Mol. Biol. 49 (1994) 281–287.
[5] S. Chumsri, T. Howes, T. Bao, G. Sabnis, A. Brodie, Aromatase, aromatase inhibitors, and breast cancer, J. Steroid Biochem. Mol. Biol. 125 (2011) 13–22.
[6] J.M. Nabholtz, A. Buzdar, M. Pollak, W. Harwin, G. Burton, A. Mangalik, M. Steinberg, A. Webster, M. von Euler, Anastrozole is superior to tamoxifen as first-line therapy for advanced breast cancer in postmenopausal women: results of a North American multicenter randomized trial. Arimidex Study Group, J. Clin. Oncol. 18 (2000) 3758–3767.
[7] H. Mouridsen, M. Gershanovich, Y. Sun, R. Perez-Carrion, C. Boni, A. Monnier, J. Apffelstaedt, R. Smith, H.P. Sleeboom, F. Janicke, A. Pluzanska, M. Dank, D. Becquart, P.P. Bapsy, E. Salminen, R. Snyder, M. Lassus, J.A. Verbeek, B. Staffler, H.A. Chaudri-Ross, M. Dugan, Superior efficacy of letrozole versus tamoxifen as first-line therapy for postmenopausal women with advanced breast cancer: results of a phase III study of the International Letrozole Breast Cancer Group, J. Clin. Oncol. 19 (2001) 2596–2606.
[8] J. Geisler, P.E. Lonning, Aromatase inhibition: translation into a successful therapeutic approach, Clin. Cancer Res. 11 (2005) 2809–2821.
[9] A.S. Bhatnagar, A. Häusler, K. Schieweck, M. Lang, R. Bowman, Highly selective inhibition of estrogen biosynthesis by CGS 20267, a new non-steroidal aromatase inhibitor, J. Steroid Biochem. Mol. Biol. 37 (1990) 1021–1027.
[10] A.S. Bhatnagar, The early days of letrozole, Breast Cancer Res. Treat. 105 (2007) 3–5.
[11] A.S. Bhatnagar, The discovery and mechanism of action of letrozole, Breast Cancer Res. Treat. 105 (Suppl. 1) (2007) 7–17.
[12] A.S. Bhatnagar, A.M. Brodie, B.J. Long, D.B. Evans, W.R. Miller, Intracellular aromatase and its relevance to the pharmacological efficacy of aromatase inhibitors, J. Steroid Biochem. Mol. Biol. 76 (2001) 199–202.
[13] C. Hansch, R. Muir, T. Fujita, P. Maloney, F. Geiger, M. Streich, The correlation of biological activity of plant growth regulators and chloromycetin derivatives with hammett constants and partition coefficients, J. Am. Chem. Soc. 85 (1963) 2817–2824.
[14] C. Nantasenamat, C. Isarankura-Na-Ayudhya, N. Tansila, T. Naenna, V. Prachayasittikul, Prediction of GFP spectral properties using artificial neural network, J. Comput. Chem. 28 (2007) 1275–1289.
[15] C. Nantasenamat, K. Srungboonmee, S. Jamsak, N. Tansila, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Quantitative structure–property relationship study of spectral properties of green fluorescent protein with support vector machine, Chemometr. Intell. Lab. Syst. 120 (2013) 42–52.
[16] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, Prediction of bond dissociation enthalpy of antioxidant phenols by support vector machine, J. Mol. Graph. Model. 27 (2008) 188–196.
[17] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, S. Prachayasittikul, V. Prachayasittikul, Predicting the free radical scavenging activity of curcumin derivatives, Chemometr. Intell. Lab. Syst. 109 (2011) 207–216.
[18] C. Nantasenamat, T. Piacham, T. Tantimongcolwat, T. Naenna, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, QSAR model of the quorum-quenching *N*-acyl-homoserine lactone lactonase activity, J. Biol. Syst. 16 (2008) 279–293.
[19] C. Thippakorn, T. Suksrichavalit, C. Nantasenamat, T. Tantimongcolwat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, Modeling the LPS neutralization activity of anti-endotoxins, Molecules 14 (2009) 1869–1888.
[20] A. Worachartcheewan, C. Nantasenamat, T. Naenna, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Modeling the activity of furin inhibitors using artificial neural network, Eur. J. Med. Chem. 44 (2009) 1664–1673.
[21] P. Mandi, C. Nantasenamat, K. Srungboonmee, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, QSAR study of anti-prion activity of 2-aminothiazoles, EXCLI J. 11 (2012) 453–467.
[22] R. Pingaew, P. Tongraung, A. Worachartcheewan, C. Nantasenamat, S. Prachayasittikul, S. Ruchirawat, V. Prachayasittikul, Cytotoxicity and QSAR study of (thio)ureas derived from phenylalkylamines and pyridylalkylamines, Med. Chem. Res. (2012) 1–14.
[23] C. Nantasenamat, T. Naenna, C. Isarankura Na-Ayudhya, V. Prachayasittikul, Quantitative prediction of imprinting factor of molecularly imprinted polymers by artificial neural network, J. Comput. Aid. Mol. Des. 19 (2005) 509–524.
[24] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, Quantitative structure-imprinting factor relationship of molecularly imprinted polymers, Biosens. Bioelectron. 22 (2007) 3309–3317.
[25] Y. Kangwanariyakul, C. Nantasenamat, T. Tantimongcolwat, T. Naenna, Data mining of magnetocardiograms for prediction of ischemic heart disease, EXCLI J. 9 (2010) 82–95.
[26] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, P. Pidetcha, V. Prachayasittikul, Identification of metabolic syndrome using decision tree analysis, Diab. Res. Clin. Pract. 90 (2010) e15–e18.
[27] A. Worachartcheewan, P. Dansethakul, C. Nantasenamat, P. Pidetcha, V. Prachayasittikul, Determining the optimal cutoff points for waist circumference and body mass index for identification of metabolic abnormalities and metabolic syndrome in urban Thai population, Diab. Res. Clin. Pract. 98 (2012) e16–e21.
[28] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, A practical overview of quantitative structure–activity relationship, EXCLI J. 8 (2009) 74–88.
[29] C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Advances in computational methods to predict the biological activity of compounds, Exp. Opin. Drug Discov. 5 (2010) 633–654.
[30] L.J. Browne, C. Gude, H. Rodriguez, R.E. Steele, A. Bhatnager, Fadrozole hydrochloride: a potent, selective, nonsteroidal inhibitor of aromatase for the treatment of estrogen-dependent disease, J. Med. Chem. 34 (1991) 725–736.
[31] D. Schuster, C. Laggner, T.M. Steindl, A. Palusczak, R.W. Hartmann, T. Langer, Pharmacophore modeling and in silico screening for new P450 19 (aromatase) inhibitors, J. Chem. Inf. Model. 46 (2006) 1301–1311.
[32] M.A. Neves, T.C. Dinis, G. Colombo, M.L. Sa e Melo, Fast three dimensional pharmacophore virtual screening of new potent non-steroid aromatase inhibitors, J. Med. Chem. 52 (2009) 143–150.
[33] J. Doiron, A.H. Soultan, R. Richard, M.M. Touré, N. Picot, R. Richard, M. Cuperlović-Culf, G.A. Robichaud, M. Touaibia, Synthesis and structure–activity relationship of 1- and 2-substituted-1,2,3-triazole letrozole-based analogues as aromatase inhibitors, Eur. J. Med. Chem. 46 (2011) 4010–4024.
[34] P. Marchand, M. Le Borgne, M. Palzer, G. Le Baut, R.W. Hartmann, Preparation and pharmacological profile of 7-(alpha-azolylbenzyl)-1H-indoles and indolines as new aromatase inhibitors, Bioorg. Med. Chem. Lett. 13 (2003) 1553–1555.
[35] M.P. Leze, M. Le Borgne, P. Pinson, A. Palusczak, M. Duflos, G. Le Baut, R.W. Hartmann, Synthesis and biological evaluation of 5-[(aryl)(1H-imidazol-1-yl)methyl]-1H-indoles: potent and selective aromatase inhibitors, Bioorg. Med. Chem. Lett. 16 (2006) 1134–1137.
[36] M.P. Leze, A. Palusczak, R.W. Hartmann, M. Le Borgne, Synthesis of 6- or 4-functionalized indoles via a reductive cyclization approach and evaluation as aromatase inhibitors, Bioorg. Med. Chem. Lett. 18 (2008) 4713–4715.
[37] A.M. Farag, K.A. Ali, T.M. El-Debss, A.S. Mayhoub, A.G. Amr, N.A. Abdel-Hafez, M.M. Abdulla, Design, synthesis and structure–activity relationship study of novel pyrazole-based heterocycles as potential antitumor agents, Eur. J. Med. Chem. 45 (2010) 5887–5898.
[38] A.M. Farag, A.S. Mayhoub, T.M. Eldebss, A.G. Amr, K.A. Ali, N.A. Abdel-Hafez, M.M. Abdulla, Synthesis and structure–activity relationship studies of pyrazole-based heterocycles as antitumor agents, Arch. Pharm. Chem. Life Sci. 343 (2010) 384–396.
[39] L.W. Woo, O.B. Sutcliffe, C. Bubert, A. Grasso, S.K. Chander, A. Purohit, M.J. Reed, B.V. Potter, First dual aromatase–steroid sulfatase inhibitors, J. Med. Chem. 46 (2003) 3193–3196.
[40] L.W. Woo, C. Bubert, O.B. Sutcliffe, A. Smith, S.K. Chander, M.F. Mahon, A. Purohit, M.J. Reed, B.V. Potter, Dual aromatase–steroid sulfatase inhibitors, J. Med. Chem. 50 (2007) 3540–3560.

[41] P.M. Wood, L.W. Woo, M.P. Thomas, M.F. Mahon, A. Purohit, B.V. Potter, Aromatase and dual aromatase–steroid sulfatase inhibitors from the letrozole and vorozole templates, ChemMedChem 6 (2011) 1423–1438.
[42] P.M. Wood, L.W. Woo, J.R. Labrosse, M.N. Trusselle, S. Abbate, G. Longhi, E. Castiglioni, F. Lebon, A. Purohit, M.J. Reed, B.V. Potter, Chiral aromatase and dual aromatase–steroid sulfatase inhibitors from the letrozole template: synthesis, absolute configuration, and in vitro activity, J. Med. Chem. 51 (2008) 4226–4238.
[43] T. Jackson, L.W. Woo, M.N. Trusselle, S.K. Chander, A. Purohit, M.J. Reed, B.V. Potter, Dual aromatase–sulfatase inhibitors based on the anastrozole template: synthesis, in vitro SAR, molecular modelling and in vivo activity, Org. Biomol. Chem. 5 (2007) 2940–2952.
[44] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH Verlag, Weinheim, Germany, 2009.
[45] M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum–chemical descriptors in QSAR/QSPR studies, Chem. Rev. 96 (1996) 1027–1044.
[46] C. Nantasenamat, H. Li, P. Mandi, A. Worachartcheewan, T. Monnor, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Exploring the chemical space of aromatase inhibitors, Mol. Diversity (2013), http://dx.doi.org/10.1007/s11030-013-9462-x.
[47] OpenEye Scientific Software, VIDA, Version 4.2.1, Santa Fe, NM,
[48] OpenEye Scientific Software, Babel, Version 3.3, Santa Fe, NM,
[49] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery, J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J.M. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox, Gaussian 09, Revision A.1, Wallingford, Connecticut, 2009.
[50] A.D. Becke, A new mixing of Hartree–Fock and local density-functional theories, J. Chem. Phys. 98 (1993) 1372–1377.
[51] C. Lee, W. Yang, R.G. Parr, Development of the Colle–Salvetti correlation-energy formula into a functional of the electron density, Phys. Rev. B 37 (1988) 785–789.
[52] T. Piacham, C. Isarankura-Na-Ayudhya, C. Nantasenamat, S. Yainoy, L. Ye, L. Bülow, V. Prachayasittikul, Metalloantibiotic Mn(II)–bacitracin complex mimicking manganese superoxide dismutase, Biochem. Biophys. Res. Commun. 341 (2006) 925–930.
[53] V. Prachayasittikul, C. Isarankura-Na-Ayudhya, T. Tantimongcolwat, C. Nantasenamat, H.J. Galla, EDTA-induced membrane fluidization and destabilization: biophysical studies on artificial lipid membranes, Acta Biochim. Biophys. Sin. 39 (2007) 901–913.
[54] T. Suksrichavalit, S. Prachayasittikul, T. Piacham, C. Isarankura-Na-Ayudhya, C. Nantasenamat, V. Prachayasittikul, Copper complexes of nicotinic–aromatic carboxylic acids as superoxide dismutase mimetics, Molecules 13 (2008) 3040–3056.
[55] T. Suksrichavalit, S. Prachayasittikul, C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Copper complexes of pyridine derivatives with superoxide scavenging and antimicrobial activities, Eur. J. Med. Chem. 44 (2009) 3259–3265.
[56] S. Prachayasittikul, O. Wongsawatkul, A. Worachartcheewan, C. Nantasenamat, S. Ruchirawat, V. Prachayasittikul, Elucidating the structure–activity relationships of the vasorelaxation and antioxidation properties of thionicotinic acid derivatives, Molecules 15 (2010) 198–214.
[57] C. Isarankura-Na-Ayudhya, C. Nantasenamat, P. Buraparuangsang, T. Piacham, L. Ye, L. Bülow, V. Prachayasittikul, Computational insights on sulfonamide imprinted polymers, Molecules 13 (2008) 3077–3091.
[58] T. Piacham, C. Nantasenamat, T. Suksrichavalit, C. Puttipanyalears, T. Pissawong, S. Maneewas, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Synthesis and theoretical study of molecularly imprinted nanospheres for recognition of tocopherols, Molecules 14 (2009) 2985–3002.
[59] C. Nantasenamat, H. Li, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Exploring the physicochemical properties of templates from molecular imprinting literature using interactive text mining approach, Chemometr. Intell. Lab. Syst. 116 (2012) 128–136.
[60] DRAGON for Windows (Software for Molecular Descriptor Calculations). Version 5.5, Talete srl, Milano, Italy, 2007.
[61] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, third ed., Morgan Kaufmann Publishers, Burlington, MA, 2011.
[62] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, Data mining in bioinformatics using Weka, Bioinformatics (Oxf., England) 20 (2004) 2479–2481.
[63] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[64] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 2000.
[65] J.C. Platt, Sequential Minimal Optimization: a Fast Algorithm for Training Support Vector Machines. Technical Report, Microsoft Research, 1998. MSR-TR-98-14.
[66] V.D.A. Sánchez, Advanced support vector machines and kernel methods, Neurocomputing 55 (2003) 5–20.
[67] A. Mammone, M. Turchi, N. Cristianini, Support vector machines, Wiley Interdisp. Rev. Comput. Stat. 1 (2009) 283–289.
[68] R.G. Brereton, G.R. Lloyd, Support vector machines for classification and regression, Analyst 135 (2010) 230–267.
[69] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, 2000.
[70] R. Herbrich, Learning Kernel Classifiers: Theory and Algorithms, The MIT Press, Cambridge, 2002.