Origin of aromatase inhibitory activity via proteochemometric modeling

Saw Simeon, Ola Spjuth, Maris Lapins, Sunanta Nabu, Virapong Prachayasittikul, Jarl ES Wikberg, Chanin Nantasenamat

Aromatase, which is a rate-limiting enzyme that catalyzes the conversion of androgen to estrogen, plays an essential role in the development of estrogen-dependent breast cancer. Side effects due to aromatase inhibitors (AIs) necessitate the pursuit of novel inhibitor candidates with high selectivity, lower toxicity and increased potency. Designing a novel therapeutic agent against aromatase could be achieved computationally by means of ligand-based and structure-based methods. For over a decade, we have utilized both approaches to design potential AIs for which quantitative structure-activity relationship and molecular docking were used to explore inhibitory mechanisms of AIs towards aromatase. However, such approaches do not consider the effects that aromatase variants have on different AIs. In this study, proteochemometrics modeling was applied to analyze the interaction space between AIs and aromatase variants as a function of their substructural and amino acid features. Good predictive performance was achieved, as rigorously verified by 10-fold cross-validation, external validation, leave-one-compound-out cross-validation, leave-one-protein-out cross-validation and Y-scrambling tests. The investigations presented herein provide important insights into the mechanisms of aromatase inhibitory activity that could aid in the design of novel potent AIs as breast cancer therapeutic agents.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1556v1 | CC-BY 4.0 Open Access | rec: 1 Dec 2015, publ: 1 Dec 2015

Origin of aromatase inhibitory activity via proteochemometric modeling

Saw Simeon[1], Ola Spjuth[3], Maris Lapins[3], Sunanta Nabu[1], Virapong Prachayasittikul[2], Jarl E. S. Wikberg[3], and Chanin Nantasenamat*[1]

[1]Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[2]Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[3]Department of Pharmaceutical Biosciences, Uppsala University, Uppsala SE751 24, Sweden

ABSTRACT

Aromatase, which is a rate-limiting enzyme that catalyzes the conversion of androgen to estrogen, plays an essential role in the development of estrogen-dependent breast cancer. Side effects due to aromatase inhibitors (AIs) necessitate the pursuit of novel inhibitor candidates with high selectivity, lower toxicity and increased potency. Designing a novel therapeutic agent against aromatase could be achieved computationally by means of ligand–based and structure–based methods. For over a decade, we have utilized both approaches to design potential AIs for which quantitative structure–activity relationship and molecular docking were used to explore inhibitory mechanisms of AIs towards aromatase. However, such approaches do not consider the effects that aromatase variants have on different AIs. In this study, proteochemometrics modeling was applied to analyze the interaction space between AIs and aromatase variants as a function of their substructural and amino acid features. Good predictive performance was achieved, as rigorously verified by 10-fold cross-validation, external validation, leave-one-compound-out cross-validation, leave-one-protein-out cross-validation and Y-scrambling tests. The investigations presented herein provide important insights into the mechanisms of aromatase inhibitory activity that could aid in the design of novel potent AIs as breast cancer therapeutic agents.

Keywords: aromatase, aromatase inhibitor, breast cancer, quantitative structure-activity relationship, QSAR, proteochemometrics, data mining

INTRODUCTION

Cancer exerts a great impact on the quality of life of patients and is the leading cause of death worldwide. Breast cancer is the most common cancer type and is the second most common cause of death in women worldwide (Fontham et al., 2009). Despite the continuous efforts being made towards improving diagnostic tests, the incidence rate of breast cancer has gradually increased (May, 2014). It is estimated that around two-thirds of breast cancers in women are dependent on the steroid hormone estrogen, which regulates tumor cell growth and drives the progression of the cancer (Lipton et al., 1992). Therefore, two major therapeutic approaches are involved in breast cancer treatment and prevention: the first involves the development of drugs that target the estrogen receptor, which are also known as selective estrogen receptor modulators (SERMs), whereas the second approach involves the development of drugs that target aromatase, i.e., the enzyme that converts androgens to estrogens, the latter of which are also known as aromatase inhibitors (AIs).

Aromatase, also known as cytochrome P450 19A1 (EC 1.14.14.1), is the expression product of the CYP19A1 gene. The enzyme comprises 503 amino acids spanning twelve α-helices and ten β-strands, inside which sits a heme co-factor that is coordinated by a cysteine residue at position 437 (Ghosh et al., 2009). Aromatase is a major producer of estrogen in post-menopausal women, and it catalyzes the

*Corresponding author. E-mail: chanin.nan@mahidol.ac.th

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1556v1 | CC-BY 4.0 Open Access | rec: 1 Dec 2015, publ: 1 Dec 2015

rate-limiting step of converting androgens to estrogens (Simpson et al., 1994). The aromatase conversion of androgens to estrogens involves three steps, whereby androgen's methyl group at carbon 19 is oxidized to form formic acid, which is followed by the aromatization of the A ring to the phenolic A ring of estrogen. (Eisen et al., 2008). As aromatase catalyzes the biosynthesis of estrogen from androgens, inhibition of aromatase activity has become the standard treatment for hormone-dependent breast cancers in women.

Previously, our group utilized the quantitative structure-activity relationship (QSAR) method in our efforts towards understanding the origin of aromatase inhibition (Nantasenamat et al., 2013a,b; Worachartcheewan et al., 2014a,b; Nantasenamat et al., 2014; Shoombuatong et al., 2015). We also used structure-based approaches to elucidate how selected compounds of interest interact with aromatase to give rise to their inhibitory activity (Suvannang et al., 2011; Worachartcheewan et al., 2014b; Pingaew et al., 2015). Although robust, both ligand–based and structure–based approaches have limitations: the former will only allow the study of how modifications to functional moieties of ligands influence the bioactivity, whereas the latter will only provide insights into how the spatial location of amino acid residues influences the bioactivity.

In this study, we developed a unified proteochemometric (PCM) model to investigate the interaction between a series of ligands and a series of aromatase variants. Such computational approaches present methodological differences with the systems-based approach (i.e., the PCM model) described herein. To this end, aromatase protein variants were represented using highly interpretable and position-specific z-scale descriptors, while AIs were represented using substructure fingerprint descriptors. Each interacting pair of AIs with aromatase variants was assigned a pIC_{50} value. Various machine learning methods were then employed to model the interaction between the ligands and the aromatase variants. Compared to the conventional ligand-based QSAR approach, the PCM technique represents a leap forward for structure-activity relationship investigations due to its ability to simultaneously consider descriptive information of several proteins and several ligands as well as its inherent interpretability in which the relative significance of descriptors in relation to the dependent variable (i.e., pIC_{50}) can be derived. Furthermore, such PCM strategy provided important insights into the molecular basis for the inhibition of a set of AIs against a set of aromatase variants and may aid in the combat against aromatase inhibitor resistance.

MATERIALS AND METHOD

Data Set

A data set of compounds, site-specific variations of residues, and bioactivity values for protein-compound pairs was obtained from previous studies by Kao et al. (1996) and Auvray et al. (2002). The general workflow for PCM modeling of this data set is summarized in Figure 1. The compounds included in this study are 4-OHA (**1**), MDL101, 103 (**2**), 7α-APTADD (**3**), aminoglutethimide (**4**), CGS 20267 (**5**), vorozole (**6**), ICI D1033 (**7**), MR20814 (**8**), MR20492 (**9**) and MR20494 (**10**), and their chemical structures are shown in Figure 2. These compounds interact with target proteins to induce pharmacological effects. However, the interaction occurs at the active site, where the compounds bind to only a small portion of residues in the target proteins. However, residues that are involved both near and far way from the active site can be considered in the PCM model. In this study, residues at positions K119, C124, K130, I133, F235, E302, P308, D309, T310, F320, I395, I474 and D476 were considered. These residues cover the AI binding site as well as residues near the aromatase active site. Aromatase inhibitory activities were originally defined using IC_{50} values, but to obtain a more distributed spread of the data points, they were subjected to negative logarithmic transformation, yielding pIC_{50} values. A summary table of the pIC_{50} values for each pair of aromatase variant and compound is provided in the Supplementary Data.

Compound descriptors

The chemical structures of the compounds were drawn using Marvin Sketch version 6.2.1 (ChemAxon Ltd., 2014) and subsequently pre-processed according to the QSAR data curation workflow described by Fourches et al. (2010). In the workflow, metal ions containing compounds were removed because reliable descriptors cannot be calculated when compounds contain metal ions. The second part involved removing the salts from the compounds, followed by the normalization of the chemotypes and standardization of tautomers using the built-in function of the software program PaDEL-Descriptor (Yap, 2011). The curated compounds were subsequently coded using substructure fingerprint counts (Laggner, 2009). Fingerprint descriptors are numerical values that are used to describe the structure of compounds, including the number of hydroxyl groups and the number of benzene rings. In particular, substructure fingerprints were chosen to describe the compounds because they are interpretable and can therefore pinpoint the substructures in compounds that are important for inhibiting aromatase.

Protein descriptors

Aromatase comprises a polypeptide chain of 503 amino-acid residues and a prosthetic heme group at its active site. An androgen-specific cleft, consisting of hydrophobic and polar residues, is situated at the aromatase binding site (Simpson et al., 1994). Of the 503 amino acids, 13 amino acid positions were found to be mutated in the investigated variants, as shown in Figure 3. Each of the amino acid positions was encoded using a set of three z-scale descriptors, thus giving 39 z-scale descriptors for each of the 22 aromatase proteins. z-scale descriptors characterize the 20 naturally occurring amino acids by encapsulating 29 physicochemical descriptors, comprising 9 experimentally determined values for retention times in thin-layer chromatography, 7 nuclear magnetic resonance shift values, 2 pK values of amino acids from amino groups and carboxylic acid groups, van der Waals volume, MW, isoelectric point, paper chromatography value, dG of the transfer of amino acids, hydration potential, salt chromatography value, and log P, log D and dG of accessible amino acids along three principal components. This high-dimensional set of values is reduced to a low-dimensional set of variables using principal component analysis, giving rise to a set of 3 z-scale descriptors, where z_1 essentially represents the hydrophobicity/hydrophilicity, z_2 represents the side-chain bulk volume, and z_3 represents the polarizability and charge of the amino acids (Hellberg et al., 1987).

Data partitioning

The K-means clustering algorithm was used to partition the data into two groups, the internal and external sets. The algorithm selects a set of cluster centers to start the K-means clustering directly in Euclidean space whereby samples closest to the center cluster are picked from each cluster. The *naes* function *prospectr* from the R package was used to split the data; 80% of the protein-ligand pairs were used as the internal set and the remaining 20% were used as the external set (Stevens and Ramirez-Lopez, 2013).

Feature Selection

Intercorrelation, also known as collinearity, is a condition in which pairs of descriptors are known to have substantial correlations. Because it adds more complexity to models than the information they provide and also could potentially give rise to bias, it therefore has a negative impact on PCM analysis. Thus, the *cor* function from the R package *stats* (R Core Team, 2014) was used to calculate the pairwise correlation between descriptors, and a descriptor in a pair with a Pearson's correlation coefficient greater than the threshold of 0.7 was filtered out using the *findCorrelation* function with the cutoff set at 0.7 from the R package *caret* so as to obtain a smaller subset of descriptors (Kuhn, 2008).

Principal Component Analysis

Principal component analysis (PCA) is a widely used method for finding the linear combination of a set of observations with the most possible variance, and it can reveal important characteristics of the data structures, which are otherwise difficult to distinguish. PCA results in mutually orthogonal axes, called principal components (PCs), which are linearly uncorrelated. Two important features of PCA are the loadings and scores. The loadings reveal correlations between all variables simultaneously, whereas the scores reveal similarities and differences between samples. The fundamental assumption is that PCs with a high explained variance possess systematic variance, whereas PCs with a low explained variance represent noise. Thus, it is important to decide on the number of PCs that sufficiently represent the information present in the data. Including higher-order PCs may just over-fit a model and result in a poor generalization of the data structures. To obtain the optimal number of PCs, Horn's parallel analysis was applied to the biological space of aromatase variants (Zwick and Velicer, 1986). To allow comparisons, the same number of PCs as that obtained from Horn's parallel analysis of aromatase variants was used also for the chemical space of AIs. Four PCs were deemed as sufficient for providing meaningful information on the chemical space of both AIs and aromatase variants. PCA was performed using the R statistical programming language. Descriptors with a variance close to zero were removed using the *nearZeroVar* function of the R package *caret* (Kuhn, 2008). The *prcomp* and *kmeans* functions from the R package *stats* were used to perform PCA and K-Means clustering, respectively (R Core Team, 2014). Prior to PCA analysis, all the data were centered and scaled to have a unit variance using the *center* and *scale*

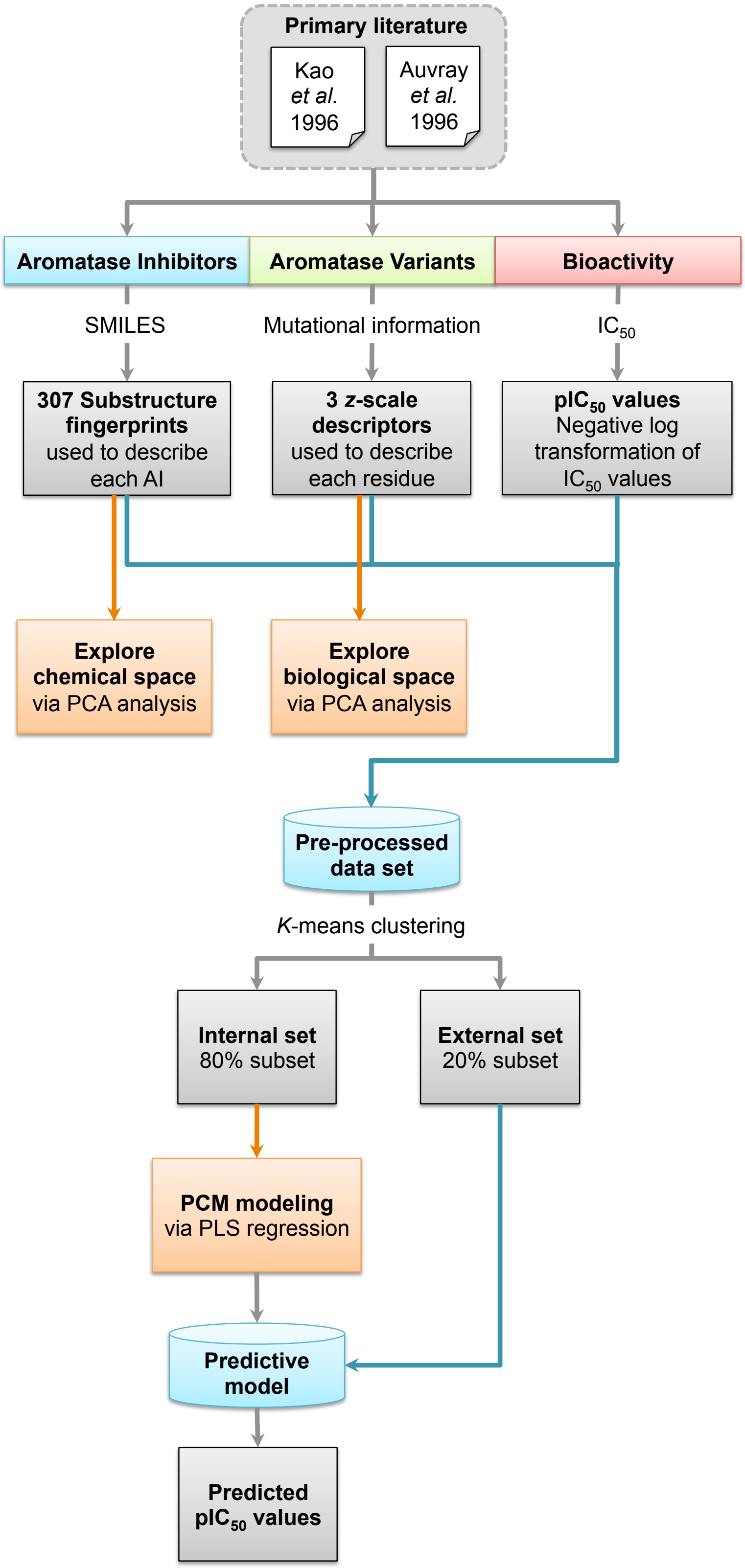
Primary literature
Kao et al. 1996
Auvray et al. 1996
Aromatase Inhibitors
Aromatase Variants
Bioactivity
SMILES
Mutational information
IC_{50}
307 Substructure fingerprints used to describe each AI
3 z-scale descriptors used to describe each residue
pIC_{50} values Negative log transformation of IC_{50} values
Explore chemical space via PCA analysis
Explore biological space via PCA analysis
Pre-processed data set
K-means clustering
Internal set 80% subset
External set 20% subset
PCM modeling via PLS regression
Predictive model
Predicted pIC_{50} values

Figure 1. Workflow for PCM modeling of aromatase inhibitory activity.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1556v1 | CC-BY 4.0 Open Access | rec: 1 Dec 2015, publ: 1 Dec 2015

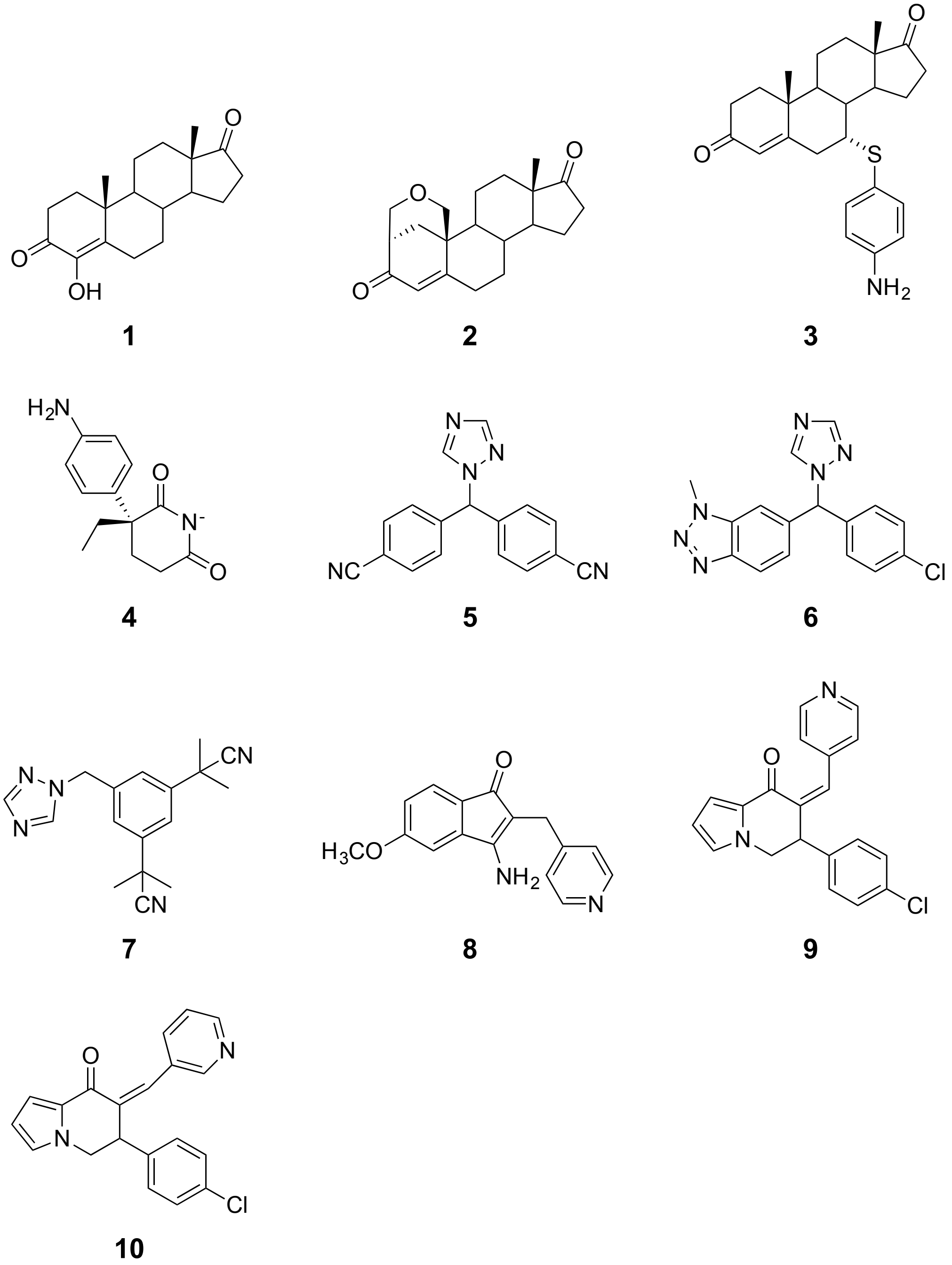

Figure 2. Chemical structures of aromatase inhibitors.

functions. The *paran* function with the argument for the *iterations* set as 5000 from the R package *paran* was utilized to perform Horn's parallel analysis to determine the optimal number of PCs (Dinno, 2012). Plots were created using the R package *ggplot2* with a 95% confidence ellipse drawn around the clusters (Wickham, 2009).

Compound-receptor cross-terms

The goal of PCM analysis is to relate the compound and target spaces with the interaction activity by creating a mathematical representation of the interaction space. Thus, unlike QSAR in which the

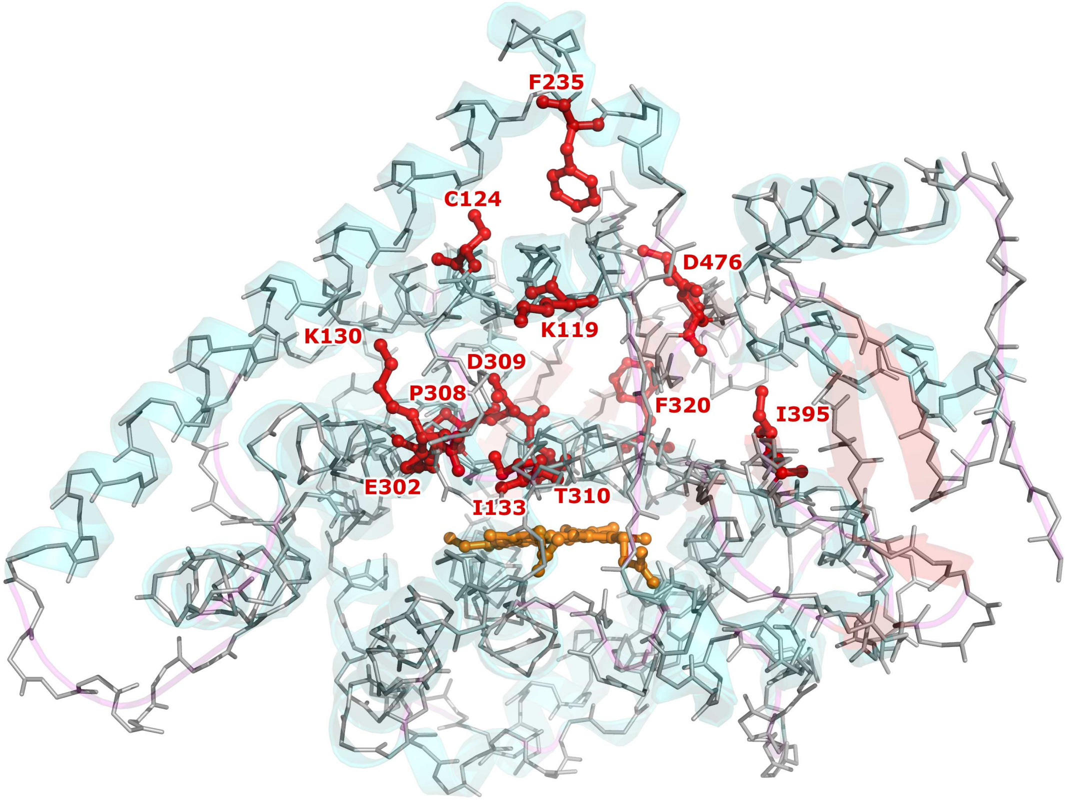

Figure 3. Three-dimensional structure of aromatase showing the investigated sites of mutations.

compounds' chemical spaces are independently related to biological activities, PCM links the unified compounds and protein space to represent their ability to form non-covalent interactions. In addition to compound descriptors and protein descriptors, PCM also makes use of cross-terms as a representation of interactions between compounds and proteins. In this study, cross-terms were calculated as the mathematical product of the compounds descriptors with those of the protein descriptors. Cross-terms were computed using the *getCPI* function from the R package *Rcpi* Cao et al. (2014). Moreover, the total number of cross-terms computed for self interaction (i.e., compound×compound and protein×protein) was obtained as follows:

$$\frac{N(N-1)}{2} \tag{1}$$

where N is the total number of descriptors of compounds or proteins.

Multivariate analysis

Descriptors of the chemical compounds and investigated amino acids residues were modeled for the pIC_{50} activities using partial least squares (PLS) modeling. PLS is an extension of PCA that correlates the **X** matrix of predictors with the **Y** dependent variables by simultaneously projecting **X** onto the latent variables and finding linear relationships between them. PLS is a robust regression method that can handle a large amount of predictors without severely affecting the predictive power of its models. Briefly, PLS finds linear combinations of the predictors, called components or latent variables. The latent variables are chosen to maximally summarize the covariance with the response, thus yielding components that maximally summarize the variation of the data set in terms of the descriptors while simultaneously having these components correlated with the response. Therefore, PLS finds a compromise between predictor space dimension reduction and the predictability of the relationship with the response (i.e., pIC_{50}). Because PLS identifies the optimal predictor sample dimension reduction to perform regression with the response, it is important to select the optimal principle component. Each extracted component increases the explained variation of the predictors, where the first component normally identifies the real correlation between the predictors and response. The PLS model was fine-tuned with the train function from the caret package, and this operator was used to extract the optimal number of PCs for building the predictive model. Finally, the *plsr* function from the R package *pls* was used to build PLS models with different combinations of predictors (Mevik and Wehrens, 2007).

When the number of descriptors is large compared to the number of samples, linear regression tends to exhibit very high variance. Thus, a small number of changes in a few samples will produce substantial changes in the coefficient. Ridge regression is effective at reducing the predictive model variance by minimizing the residual sum of squares. This is done by dividing the values of all the descriptors by their variance. Ridge regression was performed using *linearRidge* from the R package *ridge*. The parameter for the model was fine-tuned with the *train* function from the R package *caret*. To avoid random seeds, the model was trained 100 times, and the values of the statistical assessment parameters (i.e., R^2, Q^2 and RMSE) were reported as the mean and standard deviation.

Random forest (RF) is an ensemble classifier that comprises multiple decision tress. Decision trees are powerful and transparent classifiers, which use a tree structure to model the relationship between the descriptors and the classes. The route towards an activity class of HDPs begins at the root node, where it is then passed through decision nodes that require choices to be made based on the features (i.e., compound, protein and cross–terms). These outcomes split the data across branches that indicate the potential class of a decision. The final decision can be made when the tree terminated by leaf nodes provides a particular expected class as the result of a series of decisions. This provides tremendous insights into how the model works for a particular task of prediction, which makes it especially appropriate for classification. In RF, the classification is obtained by averaging the results of all tress by a majority vote based on each tree. Optimal tuning parameters (i.e., mtry) for RF were obtained by training the model with different ranges of mtry accompanied with 5–fold cross validation. The *train* function from *caret* was used with the argument *trControl* set as 5–fold cross validation with 100 iterations. The *randomForest* function from the R package *randomForest* was used to build the predictive models with 500 decision tress (Liaw and Wiener, 2002). To avoid the possibility of chance correlation that may arise from random seed of a single data partition, the models were built from 100 independent data partitions as described above using *K*-means clustering.

Validation of model performance

The internal validation set (i.e., the 80% data subset) was subjected to 10-fold cross-validation (10-fold CV). This was performed by splitting the internal validation set further into 10 folds. Afterwards, 1 fold of the data was left out as the testing set, while the remaining were used as the training set for building the predictive model. This was repeated iteratively until all folds were left out once. The *defaultSummary* function from the R package *caret* was used to obtain statistical assessment parameters for validating the PCM models Kuhn (2008). The external set was used to validate the predictability of the constructed PCM models, and the goodness-of-fit (R^2), predictive ability (Q^2) and root mean squared error (RMSE) were determined.

In addition, leave-one-protein-out (LOPO) validation and leave-one-compound-out (LOCO) cross-validation were also used to externally validate the PCM models for their extrapolation abilities in terms of new proteins or compounds. In the LOPO scheme, data annotated for single protein are left out as the test set while the remaining data are used to build the predictive model. Similarly, in the LOCO scheme, one compound is iteratively left out as the test set and evaluated against the trained model. Both processes were repeated iteratively until each aromatase variant and compounds had a chance to be left out as the test set.

To assess the statistical significance of R^2 and Q^2, the $\mathbf{Y}$-scrambling test, a well-established statistical method also known as permutation testing, was used to ensure the robustness of the PCM models to rule out the possibility of chance correlations or redundant data sets. In the test, the true $\mathbf{Y}$-dependent variable is randomly shuffled, and the statistical assessment parameters are recalculated. The *permute* function from the R package *gtools* was used to scramble the $\mathbf{Y}$-dependent variables (i.e., pIC_{50}) Warnes et al. (2015).

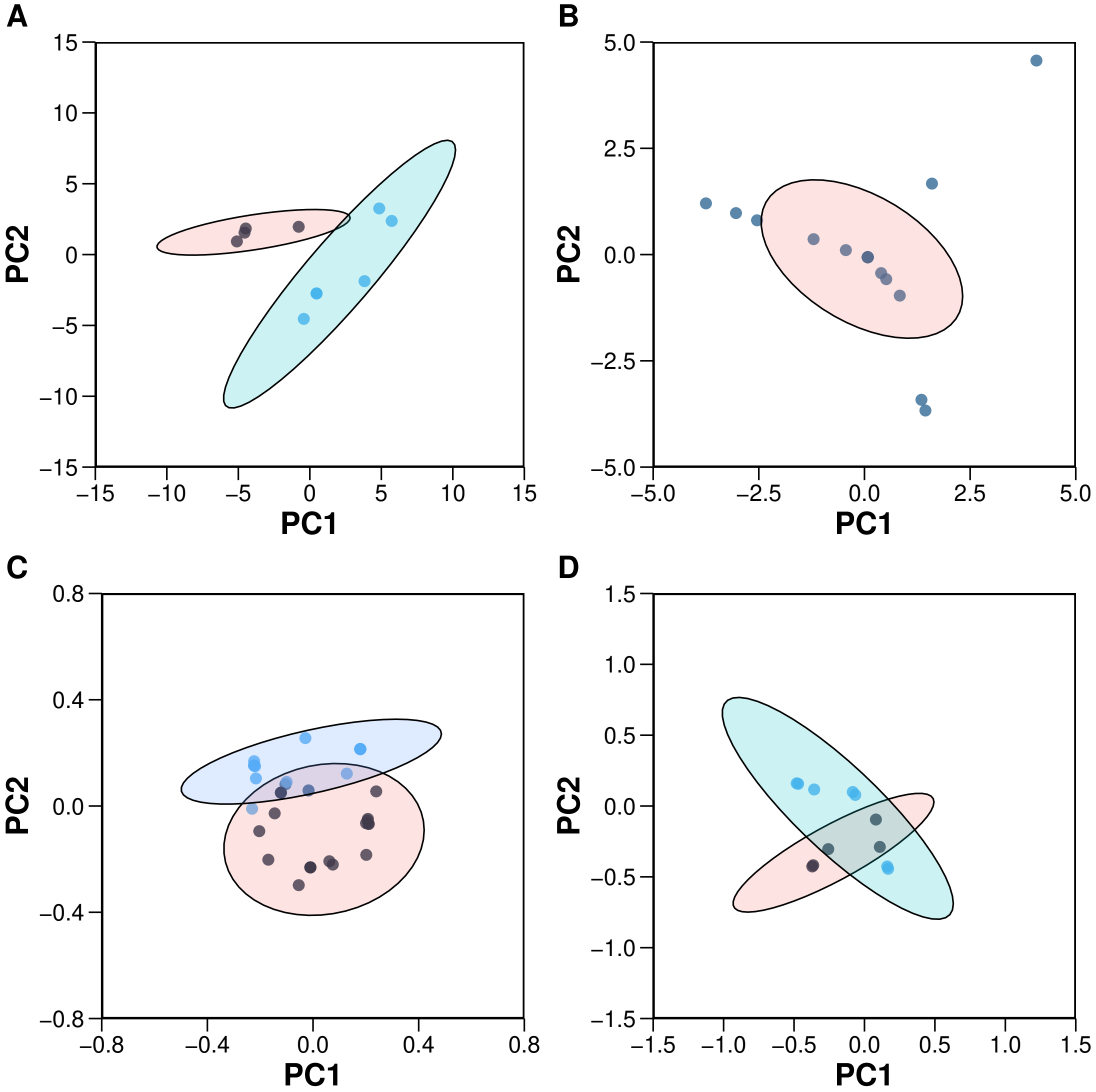

Figure 4. Plots of the PCA scores (A) and loadings (C) of 10 compounds. Plots of the PCA scores (B) and loadings (D) of 22 aromatase variants. In sub-plot (A), each dot represents an aromatase inhibitor derived from the first two PCs, while in sub-plot (C), each dot represents substructure fingerprint count descriptors. In sub-plot (B), each dot represent aromatase variants, and in sub-plot (D), each dot represents *z*-scale descriptors.

RESULTS AND DISCUSSION

Biological and chemical space of aromatase variants and compounds

PCA was utilized to analyze the *z*-scale descriptors of the aromatase variants for a better understanding of the biological space. Horn's parallel analysis deemed four PCs sufficient to yield information for satisfactorily explaining the biological space.The overall percentage of the total explained variance of the first four PCs was 75.02%, which is indicative of the good coverage of the data modeled by these PCs.

PC1 accounted for 22.07% of the data variance, in which the positive ends were dominated by p133z2 (side-chain bulk volume of the amino acid at position 133 of the aromatase variants), p133z3 (polarizability and charge of the amino acid at position 133 of the aromatase variants), and p133z1 (hydrophobicity/hydrophilicity of the amino acid at position133 of the aromatase variants), whereas p474z3 (polarizability of the amino acid at position 474 of the aromatase variants), p474z2 (side-chain bulk volume of the amino acid at position 474 of the aromatase variants), p476z3 (polarizability and charge of the amino acid at position 476 of the aromatase variants), p476z1 (hydrophobicity/hydrophilicity of the amino acid at position 476 of the aromatase variants) and p474z1 (hydrophobicity/hydrophilicity of the amino acid at position 474 of the aromatase variants) had high loadings for the negative ends. It can be observed that the physicochemical properties of position 133 have a strong influence, as they provide high loadings on one side, whereas the physiochemical properties of position 474 account for high loadings on the other side. The descriptors p119z3 (polarizability and charge of the amino acid at position 119) and p119z2 (side-chain bulk volume of the amino acid at position 119) did not provide much variance for PC1.

PC2 explained 21.21% of the variance for the protein descriptors. The descriptors with the highest loadings were p474z3 (polarizability and charge of the amino acid at position 474 of the aromatase variants), p474z2 (side-chain bulk volume of the amino acid at position 474 of the aromatase variants) and p474z1 (hydrophobicity/hydrophilicity of the amino acid at position 474 of the aromatase variants) for the positive ends, while the negative ends were dominated by p133z2 (side-chain bulk volume of the amino acid at position 133 of the aromatase variants), p133z3 (polarizability and charge of the amino acid at position 133 of the aromatase variants), p476z3 (polarizability and charge of the amino acid at position 476 of the aromatase variants) and p476z1 (hydrophobicity/hydrophilicity of the amino acid at position 476 of the aromatase variants).

PC3 accounted for 20.04% of the data variation. It can be observed that PC1 and PC2 have the same explained variance as PC3, accounting for a total explained variance of 63.31%. For PC3, the descriptor providing the highest loadings for the positive end was p119z3 (polarizability and charge of the amino acid at position 119 of the aromatase variants), whereas p199z1 (hydrophobicity/hydrophilicity of the amino acid at position 119 of the aromatase variants), p119z2 (side-chain bulk volume of the amino acid at position 119 of the aromatase variants) and p113z2 (side-chain bulk volume of the amino acid at position 113 of the aromatase variants) and p113z3 (polarizability and charge of the amino acid at position 113 of the aromatase variants) had a large influence on the negative ends.

PC4 accounted for 11.70% of the explained variance. For PC4, the descriptors with high loadings for the positive side were p474z3 (polarizability and charge of the amino acid at position 474 of the aromatase variants) and p474z2 (side-chain bulk volume of the amino acid at position 474 of the aromatase variants), whereas p119z1 (hydrophobicity/hydrophilicity of the amino acid at position 119 of the aromatase variants) and p119z2 (side-chain bulk volume of the amino acid at position 119) had the highest loadings for the negative side.

For a comparison, 4 PCs were selected from the PCA analysis of the substructure fingerprint descriptors of the chemical compounds in order to provide a general account of the chemical space. The cumulative proportion of the explained variance of the first 4 PCs was 81.22%, which can seem to provide enough information for insights on the data, as the data appear geometrical in the feature space. PC1 accounted for 38.89% of the data variance. It can be noted that the first PC was the most informative, as it explained the highest data variation among the PCs. It can be observed that the highest descriptor effects of PC1 were SubFPC49 (ketone), SubFPC300 (1,3-tautomerizable), SubFPC301 (1,5-tautomerizable), SubFPC4 (quaternary carbon), SubFP2 (secondary carbon) and SubFPC3 (tertiary carbon) on one end, while the other end was dominated by SubFPC295 (C ONS bond), SubFPC184 (heteroaromatic), SubFPC181 (hetero N nonbasic), SubFPC275 (heterocyclic) and SubFPC302 (rotatable bond). SubFPC12 (alcohol), SubFPC76 (enamine), SubFPC135 (vinylogous carbonyl or carboxyl derivative) and SubFPC13 (primary alcohol) had low loadings on PC1, suggesting that they only provide low data variation in terms of AI. It can be seen that in substructures, chemical conjugation, a phenomenon in which *p*-orbitals are connected, thereby allowing electrons to flow within the conjugated system, provided the highest afforded loadings in PC1.

PC2 accounted for 18.45% of the data variance, and descriptors providing the high loading on the positive ends were SubFPC1 (primary carbon), SubFPC35 (ammonium), SubFPC134 (isonitrile), SubFPC296 (charged), SubFPC297 (anion), SubFPC298 (cation) and SubFPC299 (salt), whereas SubFPC287 (conjugated double bond), SubFPC13 (primary alcohol), SubFPC12 (alcohol), SubFPC76 (enamine) and SubFPC135 (vinylogous carbonyl or carboxyl derivative) dominated the negative ends. Interestingly, the substructures associated with charge showed the most variance in describing the data variation at PC2. In contrast, SubFPC49 (ketone), SubFPC5 (alkene) and SubFPC275 (heterocyclic) provided little information.

PC3 accounted for 12.63% of the data variance for AI. PC3 thus represented just a small proportion of the data variance compared with the lower-order PCs. However, the spread of the data for PC3 was sufficiently large for it to be viewed as informative. The loadings of PC3 mainly comprised SubFPC13 (primary alcohol), SubFP12 (alcohol), SubFPC76 (enamine) and SubFPC135 (vinylogous carbonyl or carboxyl derivative) on the positive ends, whereas SubFPC307 (chiral center specified), SubFPC5 (alkene), SubFPC171 (arylchloride) and SubFPC180 (hetero N basic no H) dominated the negative ends.

PC4 had an explained variance of 11.25%. The descriptors that capture high loadings at the positive end were SubFPC20 (alkylarylthioether), SubFPC38 (alkylarylthioether), SubFPC96 (carbodithioic ester), SubFPC137 (vinylogous ester) and SubFPC303 (Michael acceptor). In contrast, the negative ends were dominated by SubFPC88 (carboxylic acid derivative), SubFPC105 (imide acidic), SubFPC171 (arylchloride), SubFPC275 (heterocyclic) and SubFPC72 (enol).

A closer look at the data structures for both chemical descriptors and protein descriptors revealed that the chemical descriptors provided better systemic data types when compared to the protein descriptors. It can be observed that of the overall explained variance of the first two PCs, 57.34% and 43.28% were accounted for by compound and protein descriptors, respectively. Thus, in comparison, it can be concluded that the compound descriptors represent data structures with more useful information, whereas the protein descriptors contain noise in the data. Noise in the data structure may just add to the complexity of the model, causing overfitting and thereby producing unstable models. Nevertheless, the first four PCs afforded overall variance in the data of 81.22%, and 75.02% for compounds and proteins, respectively.

PCM modeling of aromatase inhibitory activity

PCM allows the study of ligand-protein interactions by simultaneously investigating the interaction of several compounds against several proteins (i.e., in this case several aromatase variants). Our earlier QSAR models of the inhibitory properties of AI used only information from chemical compounds while the potential effects of protein binding sites and residues on the inhibitory properties of AI were not considered. This study addresses this issue by applying PCM modeling to integrate information on the interaction space of both proteins and ligands into one unified model.

The approach seems rational in view of an earlier PCM investigation by Prusis et al. (2006), where the amino acid position located very far from the binding site of a peptide hormone receptor could be effectively studied via PCM. One of the biggest problems with PCM modeling is that the data matrix tends to be very large, which leads to a high computational cost and may be prone to overfitting. To remove irrelevant descriptors that contribute more noise to the model than the information they provide, therefore feature selection was performed by removing descriptors that have pairwise Pearson's correlations higher than the cutoff threshold of 0.7. Such threshold was chosen because Pearson's correlation coefficients that are larger in value are indicative of high collinearity between descriptors (Booth et al., 1994).

The results from PCM modeling are shown in Table 1. It can be observed that the sizes of descriptor blocks, C, P, $C \times P$, $C \times C$ and $P \times P$ are 13, 18, 234, 78 and 153, respectively. As seen in Table 1, the predictive performances of the PCM models were $R^2 = 0.92 \pm 0.01/Q^2_{CV} 0.87 \pm 0.09$, $R^2 = 0.82 \pm 0.01/Q^2_{CV} = 0.62 \pm 0.22$ and $R^2 = 0.84 \pm 0.01/Q^2_{CV} = 0.74 \pm 0.19$ for models 6, 10 and 13, respectively. A closer inspection revealed that the linear models using PLS models 1, 2 and 6 showed R^2 values ranging from 0.20 ± 0.02 to 0.92 ± 0.01, Q^2_{CV} values ranging from 0.16 ± 0.20 to 0.87 ± 0.09 and Q^2_{Ext} values ranging from 0.21 ± 0.11 to 0.93 ± 0.01. Despite the low accuracy provided by the 10–fold CV set, the results were compared using the standard criteria described by Tropsha (2010), where $R^2 > 0.6$ and $Q^2 > 0.5$ are indicative of good, validated predictive models. The plot of predicted versus experimental pIC$_{50}$ for the 13 models is shown in Figure 5. As seen in Table 1, the differences between R^2–Q^2_{Ext} range from (–0.08) to (–0.32), whereas R^2–Q^2_{CV} ranges from (0.04–0.25). Generally speaking, the performance of the 10–fold CV and external sets should be lower than those of the training sets, as some samples were left out when training the models. However, models 1, 2, 4 and 5 showed differences of –0.05, –0.01, –0.06 and –0.08, respectively. Typically, the training set should not only be representative of the test set, but it should also be completely independent. This was ensured by applying the K-means clustering algorithm in which the algorithm selects training samples from the initial data set to construct a complete sample of independent variables. However, when the training samples are selected in such a way that they are representative of the test samples, the prediction error for the test set may be lower than expected. This may explain why the differences between R^2 and Q^2_{Ext} for some models are negative in value.

The PCM models after feature selection were then compared with other machine learning algorithms (i.e., ridge regression and random forest). The results of the ridge regression were comparable to those of the PLS model where the predictive performances of the PCM models were as follows: $R^2 = 0.93 \pm 0.01$ /

Table 1. Summary of the predictive performance of PCM models of pIC_{50} of aromatase after feature selection using PLS.

Model	Number of descriptors						Training set		10–fold CV		External set		$R^2–Q^2_{CV}$	$R^2–Q^2_{Ext}$
	C	P	C×P	C×C	P×P	Total	R^2	$RMSE_{Tr}$	Q^2	$RMSE_{CV}$	Q^2	$RMSE_{Ext}$		
1	13	0	0	0	0	13	0.88±0.01	0.43±0.01	0.86±0.11	0.46±0.11	0.93±0.01	0.42±0.03	0.04	–0.05
2	0	18	0	0	0	18	0.20±0.02	1.14±0.02	0.16±0.20	1.26±0.21	0.21±0.11	1.10±0.12	0.04	–0.01
3	0	0	234	0	0	234	0.86±0.02	0.48±0.03	0.61±0.22	0.79±0.24	0.54±0.12	0.88±0.15	0.25	0.32
4	0	0	0	78	0	78	0.87±0.05	0.43±0.01	0.86±0.11	0.46±0.11	0.93±0.01	0.42±0.03	0.01	–0.06
5	0	0	0	0	153	153	0.22±0.02	1.13±0.03	0.18±0.18	1.26±0.27	0.30±0.13	1.04±0.13	0.04	–0.08
6	**13**	**18**	**0**	**0**	**0**	**31**	**0.92±0.01**	**0.36±0.01**	**0.87±0.09**	**0.46±0.12**	**0.89±0.04**	**0.43±0.06**	**0.05**	**0.03**
7	13	18	234	0	0	165	0.87±0.01	0.46±0.02	0.69±0.20	0.73±0.25	0.63±0.16	0.77±0.18	0.18	0.24
8	13	18	0	78	0	109	0.90±0.01	0.40±0.01	0.81±0.13	0.55±0.14	0.88±0.06	0.44±0.08	0.09	0.02
9	13	18	0	0	153	184	0.87±0.01	0.44±0.01	0.72±0.16	0.70±0.21	0.74±0.08	0.70±0.11	0.15	0.13
10	**13**	**18**	**234**	**78**	**0**	**343**	**0.82±0.01**	**0.54±0.02**	**0.62±0.22**	**0.81±0.26**	**0.58±0.13**	**0.80±0.12**	**0.21**	**0.24**
11	13	18	234	0	153	418	0.90±0.01	0.41±0.02	0.72±0.20	0.69±0.23	0.63±0.12	0.77±0.14	0.18	0.27
12	13	18	0	78	153	262	0.83±0.01	0.52±0.01	0.72±0.19	0.67±0.21	0.79±0.09	0.60±0.09	0.11	0.04
13	**13**	**18**	**234**	**78**	**153**	**496**	**0.84±0.01**	**0.51±0.01**	**0.74±0.19**	**0.64±0.21**	**0.80±0.07**	**0.60±0.09**	**0.10**	**0.04**

Table 2. Summary of the predictive performance of PCM models of pIC_{50} of aromatase after feature selection using ridge regression.

Model	Number of descriptors						Training set		10–fold CV		External set		R^2–Q^2_{CV}	R^2–Q^2_{Ext}
	C	P	C×P	C×C	P×P	Total	R^2	$RMSE_{Tr}$	Q^2	$RMSE_{CV}$	Q^2	$RMSE_{Ext}$		
1	13	0	0	0	0	13	0.88±0.01	0.43±0.01	0.86±0.11	0.46±0.11	0.93±0.01	0.42±0.03	0.02	–0.05
2	0	18	0	0	0	18	0.34±0.03	1.04±0.03	0.20±0.23	1.26±0.23	0.17±0.10	1.15±0.12	0.14	0.17
3	0	0	234	0	0	234	0.96±0.01	0.25±0.03	0.53±0.26	1.17±0.59	0.63±0.15	0.95±0.26	0.43	0.33
4	0	0	0	78	0	78	0.87±0.01	0.43±0.01	0.86±0.11	0.46±0.11	0.93±0.01	0.42±0.03	0.01	–0.06
5	0	0	0	0	153	153	0.35±0.03	1.03±0.03	0.19±0.21	1.28±0.27	0.33±0.12	1.03±0.13	0.16	0.02
6	**13**	**18**	**0**	**0**	**0**	**31**	**0.93±0.01**	**0.33±0.02**	**0.86±0.10**	**0.47±0.14**	**0.87±0.05**	**0.47±0.08**	**0.07**	**0.06**
7	13	18	234	0	0	165	0.91±0.01	0.38±0.02	0.63±0.23	0.83±0.37	0.62±0.16	0.77±0.16	0.28	0.29
8	13	18	0	78	0	109	0.90±0.01	0.42±0.01	0.75±0.16	0.65±0.18	0.82±0.06	0.59±0.10	0.15	0.08
9	13	18	0	0	153	184	0.74±0.02	0.71±0.03	0.70±0.15	0.66±0.23	0.64±0.08	0.90±0.13	0.04	0.10
10	**13**	**18**	**234**	**78**	**0**	**343**	**0.93±0.01**	**0.34±0.01**	**0.67±0.24**	**0.74±0.30**	**0.63±0.12**	**0.75±0.11**	**0.26**	**0.30**
11	13	18	234	0	153	418	0.78±0.01	0.69±0.02	0.65±0.24	0.79±0.31	0.62±0.15	0.77±0.17	0.13	0.16
12	13	18	0	78	153	262	0.91±0.01	0.38±0.01	0.75±0.18	0.62±0.20	0.82±0.07	0.55±0.08	0.16	0.09
13	**13**	**18**	**234**	**78**	**153**	**496**	**0.84±0.01**	**0.53±0.01**	**0.78±0.18**	**0.59±0.19**	**0.83±0.06**	**0.56±0.07**	**0.06**	**0.01**

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1556v1 | CC-BY 4.0 Open Access | rec: 1 Dec 2015, publ: 1 Dec 2015

Table 3. Summary of the predictive performance of PCM models of pIC_{50} of aromatase after feature selection using random forest.

Model	Number of descriptors						Training set		10–fold CV		External set		R^2–Q^2_{CV}	R^2–Q^2_{Ext}
	C	P	C×P	C×C	P×P	Total	R^2	$RMSE_{Tr}$	Q^2	$RMSE_{CV}$	Q^2	$RMSE_{Ext}$		
1	13	0	0	0	0	13	0.87±0.00	0.43±0.01	0.86±0.12	0.46±0.11	0.93±0.01	0.43±0.03	0.01	–0.06
2	0	18	0	0	0	18	0.35±0.02	1.06±0.02	0.25±0.22	1.18±0.23	0.25±0.11	1.08±0.14	0.10	0.10
3	0	0	234	0	0	234	0.95±0.01	0.28±0.02	0.84±0.14	0.52±0.16	0.90±0.03	0.42±0.06	0.11	0.05
4	0	0	0	78	0	78	0.88±0.01	0.43±0.01	0.85±0.12	0.46±0.12	0.93±0.01	0.42±0.03	0.03	–0.05
5	0	0	0	0	153	153	0.32±0.03	1.06±0.03	0.18±0.19	1.25±0.23	0.33±0.12	1.01±0.14	0.14	–0.01
6	**13**	**18**	**0**	**0**	**0**	**31**	**0.93±0.01**	**0.35±0.01**	**0.85±0.11**	**0.48±0.14**	**0.90±0.04**	**0.40±0.07**	**0.08**	**0.03**
7	13	18	234	0	0	165	0.96±0.01	0.27±0.02	0.83±0.15	0.50±0.14	0.89±0.05	0.41±0.08	0.13	0.07
8	13	18	0	78	0	109	0.96±0.01	0.25±0.02	0.83±0.15	0.48±0.14	0.89±0.05	0.41±0.06	0.13	0.07
9	13	18	0	0	153	184	0.96±0.01	0.25±0.02	0.85±0.12	0.45±0.14	0.89±0.04	0.44±0.08	0.11	0.07
10	**13**	**18**	**234**	**78**	**0**	**343**	**0.96±0.01**	**0.27±0.02**	**0.84±0.15**	**0.46±0.15**	**0.90±0.04**	**0.39±0.06**	**0.12**	**0.06**
11	13	18	234	0	153	418	0.96±0.01	0.27±0.02	0.85±0.12	0.50±0.16	0.88±0.04	0.43±0.06	0.11	0.08
12	13	18	0	78	153	262	0.94±0.01	0.31±0.01	0.86±0.11	0.46±0.12	0.90±0.04	0.39±0.06	0.08	0.04
13	**13**	**18**	**234**	**78**	**153**	**496**	**0.94±0.01**	**0.31±0.01**	**0.86±0.11**	**0.48±0.14**	**0.90±0.04**	**0.40±0.05**	**0.08**	**0.04**

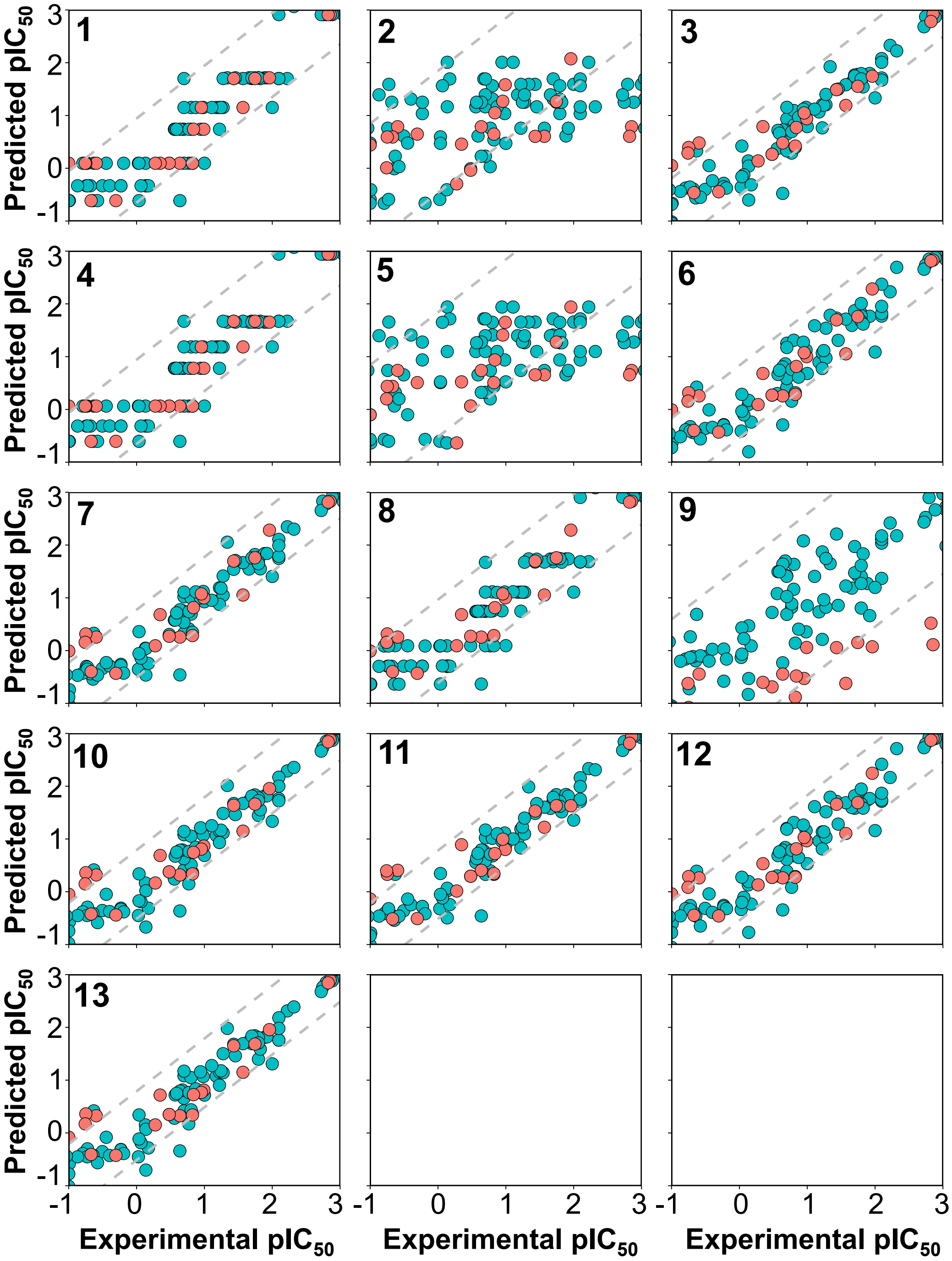

Figure 5. Plot of the experimental versus predicted pIC$_{50}$ values for 13 PCM models. Blue circles represent internal sets while the red circles correspond to external tests.

$Q^2_{CV} = 0.86 \pm 0.10$, $R^2 = 0.93 \pm 0.01$ / $Q^2_{CV} = 0.67 \pm 0.24$ and $R^2 = 0.84 \pm 0.01$ / $Q^2_{CV} = 0.78 \pm 0.18$ for models 6, 10 and 13, respectively. However, when the PLS models were compared with that of the random forest models, it is apparent that PCM models built using random forest are highly robust. In particular, models 10 and 13 yielded superior predictive results when compared with both the PLS and ridge models where values of $R^2 = 0.96 \pm 0.01 / Q^2_{CV} = 0.84 \pm 0.15$ and $R^2 = 0.94 \pm 0.01 / Q^2_{CV} = 0.86 \pm 0.11$, respectively, were observed. This may be attributed to the fact that random forest is an ensemble machine learning method employing multiple decision trees in which the bagging of trees improves the predictive

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1556v1 | CC-BY 4.0 Open Access | rec: 1 Dec 2015, publ: 1 Dec 2015

performance over that of a single model. As can be see in Table 3, the predictive performance of the 10–fold cross-validation as deduced from Q^2_{CV} ranges from 0.83±0.15 to 0.86±0.11, with exception of models 2 and 5, which were composed of protein descriptor blocks and their cross–terms.

External validation is an important process for assessing the predictive ability of PCM models. As can be seen in Table 1, results from the external validation using PLS showed $Q^2_{Ext} = 0.89 \pm 0.04, 0.58 \pm 0.13$ and 0.80 ± 0.07 for models 6, 10 and 13, respectively. However, for random forest the respective Q^2_{Ext} values for models 6, 10 and 13 were 0.90±0.04, 0.90±0.04 and 0.90±0.04, respectively. Thus, it is apparent that external validation for random forest yielded a superior performance and were thus subjected to further investigation. Subsequently, the PCM models built from random forest were then further validated using LOCO and LOPO cross–validations to evaluate their ability to extrapolate and predict the inhibitory activities for unknown compounds and aromatase variants, respectively. Table 4 summarizes the comparison of the performances of the training set and 10–fold CV set along with LOPO and LOCO sets. It can be seen that models 6, 10 and 13 performed well on both LOPO with $Q^2_{LOPO} = 0.88 \pm 0.07, Q^2_{LOPO} = 0.89 \pm 0.0.06$ and $Q^2_{LOPO} = 0.88 \pm 0.07$, respectively. In parallel, the predictive performances of LOCO were $Q^2_{LOCO} = 0.88 \pm 0.07, Q^2_{LOCO} = 0.89 \pm 0.06$ and $Q^2_{LOCO} = 0.89 \pm 0.06$, respectively. In contrast, the predictive performances of models 2 and 5 are rather poor as deduced from $Q^2_{LOPO} = 0.22 \pm 0.17 / Q^2_{LOPO} = 0.22 \pm 0.0.17$ and $Q^2_{LOCO} = 0.21 \pm 0.16 / Q^2_{LOCO} = 0.21 \pm 0.0.17$. This may be ascribed to the fact that models 2 and 3 do not contain the C descriptor block, thereby leading to poor predictability.

Y-scrambling was performed 50 times to assess the possibility of chance correlations for 13 PCM models. Scatter plots of R^2 versus Q^2 are shown in Figure 6 for the **Y**–permutated data set comprising various combinations of descriptors. It can be seen that the actual **X**-**Y** pairs from the PCM models (i.e., models 1, 3, 4, 6, 8, 10, 12 and 13) are distinctly separated from the scrambled **X**-**Y** pairs.

Interpretation of the PCM models

It is important to select the PCM model that best represents the inhibitory properties of AI. This was initially performed by selecting the top three PCM models in terms of performance. The reliability of the PCM models can be statistically assessed based on the differences between the goodness of fit and the predictive ability. From the top three models (highlighted using bold text in Table 1), the most reliable models were those for which R^2 was not greater by 0.2-0.3 units than Q^2. This is because a higher margin in the differences between R^2 and Q^2 is indicative of overfitted models either due to outliers or irrelevant descriptors. In addition, differences in R^2 and Q^2 can be used to explain the accumulated chance of correlations. Thus, PCM models with slightly similar R^2 and Q^2 values were considered.

Analysis of the feature importance can provide a better understanding on the underlying features that may strongly contribute to the inhibitory properties (i.e., pIC_{50}). The efficient and effective built-in feature importance estimators of the RF method was utilized to identify informative features. In general, two measures (i.e., the mean decrease in the Gini index and the mean decrease in prediction accuracy) are used for ranking important features. Because the mean decrease in the Gini index is reported to be robust when compared with the mean decrease in accuracy (Calle and Urrea, 2011), therefore the mean decrease in the Gini index was used to rank features. To avoid possible bias due to random seed of a single data partition, the mean and standard deviation values of the Gini index was calculated from the aforementioned 100 data partitions.

The top 10 descriptors are SubFPC16_SubFPC300 (43.79±12.46), SubFPC72_SubFPC300 (17.08±3.58), SubFPC28_SubFPC300 (14.66±2.40), SubFPC12_SubFPC88 (10.69±3.13), SubFPC1_SubFPC5 (8.91±1.87), SubFPC5_SubFPC287 (7.29±1.00), SubFPC1_SubFPC296 (6.14±2.66), SubFPC5_SubFPC88 (4.71±1.51), SubFPC288_SubFPC303 (4.53±2.28) and SubFPC35_SubFPC303 (3.58±1.36), which correspond to the following cross–terms: dialkylether × 1,3–tautomerizable, enol × 1,3–tautomerizable, primary aromatic amine × 1,3–tautomerizable, alcohol × carboxylic acid derivative, primary carbon × alkene, alkene × conjugated double bond, primary carbon × charged, alkene × carboxylic acid derivative, conjugated triple bond × Michael acceptor and ammonium × Michael acceptor, respectively.

It can be seen that the descriptors with cross–term features involving substructure fingerprints were among the top 10 descriptors thereby suggesting the importance of compound descriptors. As shown in Table 3, a predictive model built using compound descriptors and their associated cross-terms descriptors show superior performance when compared to that of the protein descriptors. The feature importance

Table 4. Summary of the predictive performance of PCM models of pIC_{50} of aromatase as assessed by 10-fold, LOPO and LOCO cross-validations.

Model	Number of Descriptors						Training Set		Cross-validation set		Leave-One-Compound-Out		Leave-One-Protein-Out	
	C	P	C×P	C×C	P×P	Total	R^2_{Tr}	RMSE_{Tr}	Q^2_{CV}	RMSE_{CV}	Q^2_{LOCO}	RMSE_{LOCO}	Q^2_{LOPO}	RMSE_{LOPO}
1	13	0	0	0	0	13	0.87±0.00	0.43±0.01	0.86±0.12	0.46±0.11	0.88±0.06	0.45±0.10	0.89±0.05	0.45±0.09
2	0	18	0	0	0	18	0.35±0.02	1.06±0.02	0.25±0.22	1.18±0.23	0.22±0.17	1.15±0.17	0.21±0.16	1.16±0.16
3	0	0	234	0	0	234	0.95±0.01	0.28±0.02	0.84±0.14	0.52±0.16	0.89±0.06	0.46±0.08	0.89±0.06	0.46±0.08
4	0	0	0	78	0	78	0.88±0.01	0.43±0.01	0.85±0.12	0.46±0.12	0.89±0.06	0.45±0.09	0.88±0.057	0.45±0.09
5	0	0	0	0	153	153	0.32±0.03	1.06±0.03	0.18±0.19	1.25±0.23	0.22±0.17	1.17±0.18	0.21±0.17	1.17±0.18
6	**13**	**18**	**0**	**0**	**0**	**31**	**0.93±0.01**	**0.35±0.01**	**0.85±0.11**	**0.48±0.14**	**0.88±0.07**	**0.45±0.10**	**0.88±0.07**	**0.44±0.11**
7	13	18	234	0	0	165	0.96±0.01	0.27±0.02	0.83±0.15	0.50±0.14	0.88±0.07	0.44±0.10	0.89±0.06	0.44±0.10
8	13	18	0	78	0	109	0.96±0.01	0.25±0.02	0.83±0.15	0.48±0.14	0.88±0.06	0.45±0.10	0.88±0.06	0.45±0.10
9	13	18	0	0	153	184	0.96±0.01	0.25±0.02	0.85±0.12	0.45±0.14	0.89±0.07	0.44±0.12	0.88±0.068	0.44±0.12
10	**13**	**18**	**234**	**78**	**0**	**343**	**0.96±0.01**	**0.27±0.02**	**0.84±0.15**	**0.46±0.15**	**0.89±0.06**	**0.46±0.08**	**0.89±0.06**	**0.46±0.08**
11	13	18	234	0	153	418	0.96±0.01	0.27±0.02	0.85±0.12	0.50±0.16	0.88±0.06	0.46±0.10	0.88±0.06	0.46±0.10
12	13	18	0	78	153	262	0.94±0.01	0.31±0.01	0.86±0.11	0.46±0.12	0.89±0.06	0.44±0.10	0.89±0.06	0.44±0.10
13	**13**	**18**	**234**	**78**	**153**	**496**	**0.94±0.01**	**0.31±0.01**	**0.86±0.11**	**0.48±0.14**	**0.89±0.06**	**0.44±0.10**	**0.88±0.07**	**0.44±0.11**

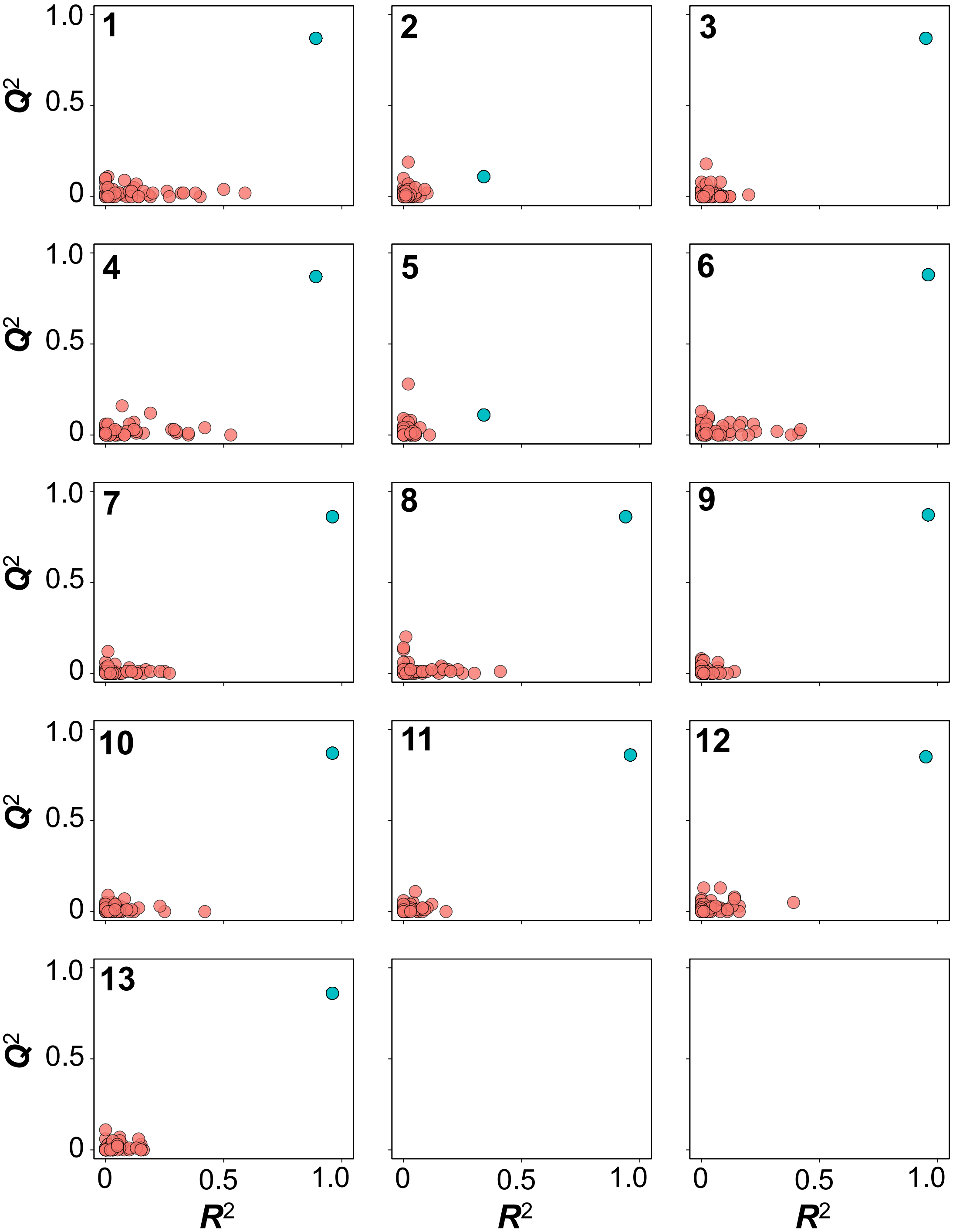

Figure 6. **Y**-scrambling plots of pIC_{50} as obtained from PCM models after feature selection.

as deduced from the Gini index is provided in Figure 7 where features having high values for the Gini index are considered to be important. It can be observed that the top 3 cross-terms consisted of 1,3–tautomerizable substructures. It has been known that the triazole moiety of compounds could interact strongly with the heme iron and thus is responsible for interacting at the active site of aromatase. Triazoles are able to undergo tautomerization, for which two constitutional isomers can be formed. In fact, compounds containing triazoles include vorozole, anastrozole and letrozole, which appear to be highly effective against aromatase. Letrozole, in particular, is marketed as an effective breast cancer

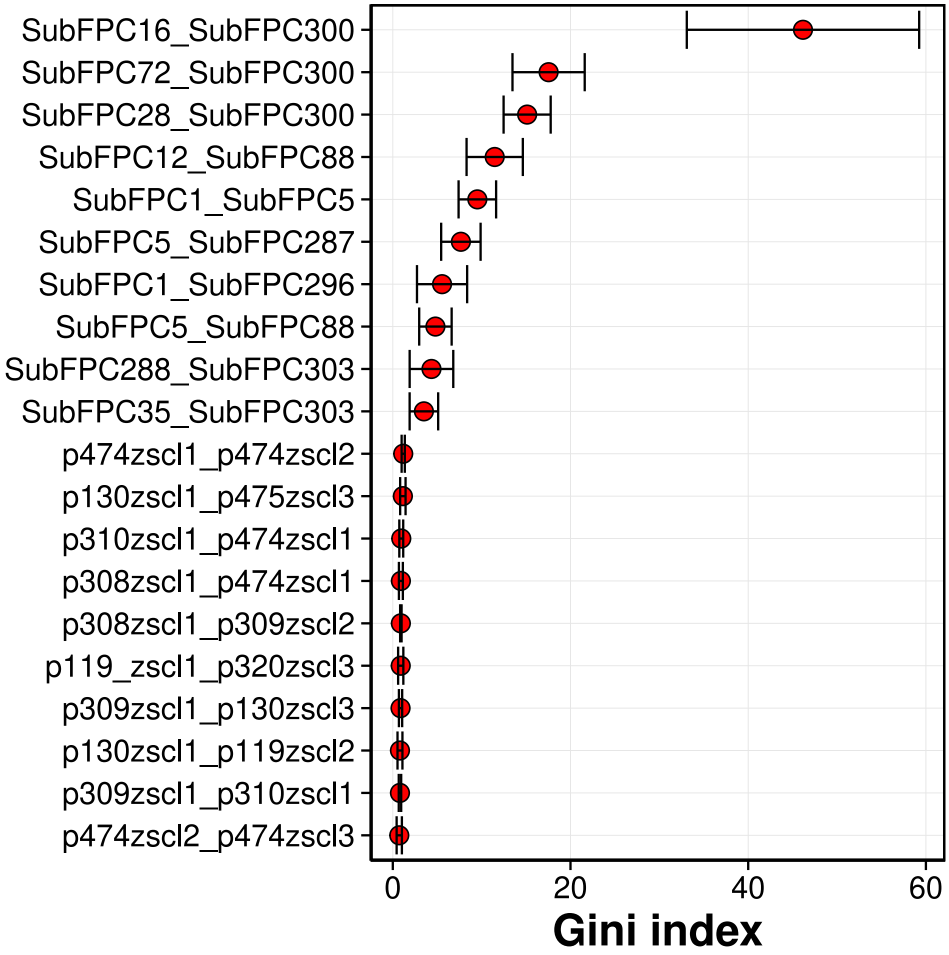

Figure 7. Plot of feature importance for RF model 13. High Gini index values are indicative of important descriptors.

drug. In the feature importance analysis, the top self cross–terms was dialkylether×1,3–tautomerizable (43.79±12.46), suggesting that this feature contributed strongly to the pIC_{50}. In general, aromatase inhibitors can be classified into two major types according to their chemical structures, steroids and non–steroids inhibitors. The steroid inhibitors are also known as mechanism–based inhibitors, as they bind covalently to aromatase, thus destroying the enzymes by forming irreversible interactions. On the other hand, non-steroidal inhibitors have reversible inhibitory interactions with the heme co–factor of the aromatase, thereby preserving the enzyme while also limiting its actions. The first generation of non–steroid inhibitors was aminoglutethimide, shown in Figure 2. Although aminoglutethimide is able to inhibit the action of aromatase, it exhibits poor specificity as it can also inhibit other cytochrome P450 enzymes, which are involved in the biosynthesis of cortisol aldosterone, leading to severe side effects. Because of these side effects, aminoglutethimide was withdrawn from clinical use. The second-generation aromatase inhibitors consist of fadrozole and formestane, which are non-steroidal imidazole derivatives and steroidal analogs. Although fadrozole was more selective and potent than aminoglutethimide, it still has undesirable effects, including inhibitory action against the production of aldosterone, corticosterone and progesterone. Formestane was the first aromatase to be used clinically, but the effects of covalently binding to aromatase led to its name of suicide inhibitor. The third-generation non-steroidal aromatase inhibitors include vorozole, anastrozole and letrozole,and the latter two are marketed under the trade names of Arimidex and Femara, respectively. The current standard–of–care compounds for preventing relapse of breast tumors are anastrozole, letrozole and exemestane (Ma et al., 2015). However, in the early and advanced stages of breast cancer, 20% of patients suffer relapse of the disease (Group et al., 2011), and the disease eventually progress despite AI therapy, leading to the disease becoming incurable, lethal and systemic. The mechanisms of aromatase resistance are heterogeneous, and the hallmarks range from changes in the tumor microenvironment, deregulation of the ER pathway, decrease in apoptosis and senescence, abnormality in the cell cycle machinery, increase in cancer stem cells, overexpression of EGFR in the growth factor receptor pathway and mutations in PIK3CA, PTEN and AKT1 through secondary messengers (Ma et al., 2015). Nevertheless, it can be observed that triazole, which can undergo tautomerization, is one of the building blocks of highly selective and potent aromatase inhibitors. Feature importance analysis also revealed that the 1,3–tautomerizable substructure fingerprint has a high weight in terms of the inhibitory properties of aromatase (i.e., pIC_{50}), as the three top features were composed of 1,3–tautomerizable. The fourth-ranked substructure included the self cross–terms of alcohol×carboxylic acid derivatives. Interestingly, the carboxylic acid derivatives were used as a substructure when combating endocrine therapy resistance. Antoon et al. (2011) selected a sphingosine kinase–2 of MAPK pathway for the treatment of endocrine therapy–resistance breast cancer and stressed that the novel selective Sphk2 inhibitor, ABC294640 (3-(4-chlorophenyl)-adamantane-1-carboxylic acid), is a potential therapeutic agent. Cadoo et al. (2014) claimed that cell cycle regulatory processes play an important role in the development of resistance in breast cancer and showed that a carboxylic acid derivative named Palbociclib is a promising therapy compound for dealing with endocrine therapy resistance. It can be observed that the top 10 features consisted of only compound descriptors, suggesting that compounds were dominant factors in terms of the inhibitory properties of aromatase. However, protein descriptors were found to have low weights for predicting activity. Recently, Ma et al. (2015) reviewed the mechanisms of aromatase inhibitor resistance, and it seems that aromatase inhibitor resistance does not just merely involve the mutation of the aromatase enzyme but also includes heterogeneous mechanisms that involve alteration of the carboxy–terminal ligand–binding domain region of estrogen receptor 1 (ER), cross-talk between growth factor receptors (GFR) and ER, mutation in the α– catalytic subunit of PI3K in ER, upregulation of cyclin dependent kinase 4 (CDK4) and modification of epigenetic regulators.

Interestingly, it can be observed that the top descriptors with large positive values are electron–rich structures, which makes the associated compounds have a more hydrophobic portion that may interact with the hydrophobic core of the protein backbone through hydrophobic effects. It has been known that the active site of proteins are highly hydrophobic in nature. Thus, hydrophobicity is important for the compound–protein interaction of aromatase with its inhibitors. Interestingly, Bansal et al. (2012) synthesized several steroid aromatase inhibitors, including 3-keto-4-ene steroid variants, and reported that compounds with heteroaromatic pyridine ring were the most potent ones. Similarly, Khodarahmi et al. (2015) utilized quantum mechanical/molecular mechanical (QM/MM)-based docking to identify the strength of compounds in acting as a potential inhibitors of aromatase and stressed that the necessary hydrophobic interactions between aromatase and its inhibitors are facilitated via heteroaromatic rings. This feature reflects the binding mechanism by which ligands with the heterocyclic aromatic ring with an azole moiety is coordinated to the heme iron of the aromatase active site while also forming a $\pi - \pi$ interaction with F221, W224, and I133 and hydrophobic interaction with W224, V369 and T310.

PLS Model 13 showed promising predictive performance with Q^2 values of 0.74±0.19 and 0.80±0.07 for the cross-validation and external sets, respectively, and were therefore selected for further investigation. Figure 8 shows the feature importance of the PLS model as deduced from their coefficients, which can be used to explain the relative contribution to pIC_{50} values. It should be noted that a positive coefficient of substructure descriptor corresponds to an increase in the pIC_{50} value while negative PLS coefficient values contribute negatively to pIC_{50} values. Such knowledge could be useful for designing compounds to modulate the aromatase enzyme.

Positive values of the PLS coefficient were seen for SubFP12_SubFPC88 (93.22±65.80), SubFPC5_SubFPC88 (88.42±62.14), SubFPC1_SubFPC5 (61.37±44.87), p130zscl1_p119zscl2 (56.75±33.96), p119zscl1_p119zscl2 (41.96±39.47), SubFPC16_SubFPC300 (28.83±17.24), SubFPC5_SubFPC287 (25.21±16.02), SubFPC72_SubFPC300 (24.73±17.08), p130zscl1_p124zscl3 (17.35±12.30) and SubFPC1_SubFPC296 (15.69±11.33). The top 3 features were those related to cross–terms of compounds: (i) alcohol×carboxylic acid derivative, (ii) alkene×carboxylic acid derivative

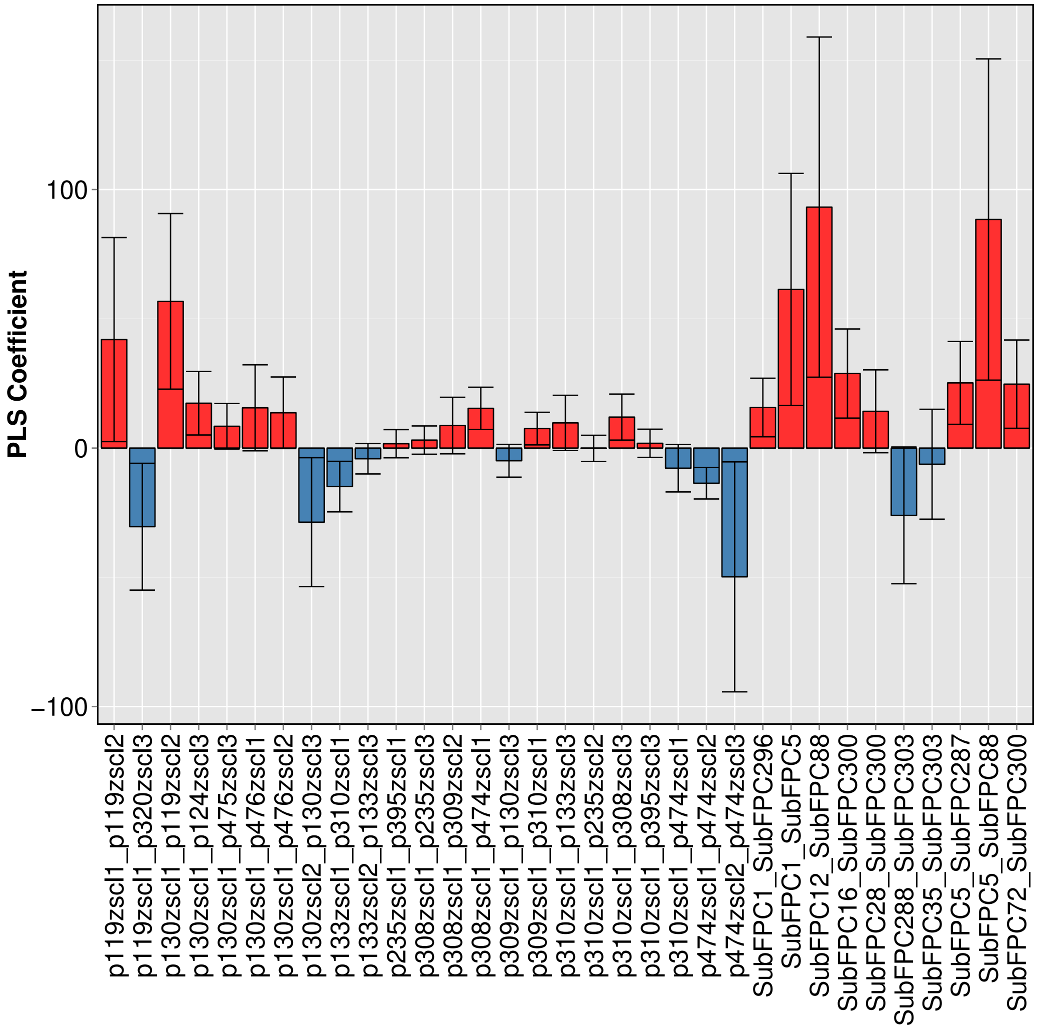

Figure 8. Plot of feature importance for PLS model 13 obtained using the regression coefficients. Positive PLS coefficients are shown in red and the negative PLS coefficients are shown in blue.

and (iii) primary carbon×alkene. This indicates that the compounds have a substantial influence on the increase in pIC$_{50}$ values. It is worthy to note that NMR studies suggests that compounds with similar substructures bind selectively to the target protein (McGovern et al., 2002). The analysis revealed that conjugated triple bond substructures have a huge impact on the increase in pIC$_{50}$ values. In a conjugated system, an electron can delocalize around the ring through p orbitals. It can be observed that compounds with conjugated bonds as a substructure are able to modulate the inhibition of aromatase and its variants. Albrecht et al. (2011) stressed that compounds containing conjugated systems (e.g., N-fused heteroaromatic compounds) are considered to be privileged compounds in drug discovery with notable examples such as Zolpidem (i.e., hypnotic properties) and Alpidem (i.e., anxiolytic properties), which are commercially available drugs that contain heteroaromatics as their substructures. This may therefore indicate that chemical conjugations are indeed a privileged substructure that are important for the inhibitory property against aromatase. Indeed, nitrogen-containing ring structures are found in both anastrozole and letrozole, which are drugs used as standard treatment for preventing the relapse of breast cancer, under the trademark names Arimidex and Femara, respectively. Furthermore, it can be seen that the highest PLS coefficient is that of p474zscl2_p474zscl3, which has a negative coefficient value, which suggested that amino acid at position 474 contribute to decreased pIC$_{50}$ values (Zhou et al., 1994). Thus, results from the feature analysis of PLS coefficients are consistent with the aforementioned findings from medicinal chemistry and computational studies.

The following substructures with negative PLS coefficients contribute to a negative pIC_{50}: p474zscl2_p474zscl3 (–49.83±44.49), p119zscl1_p320zscl3 (–30.43±24.53), p130zscl2_p130zscl3 (–28.68±24.94), SubFPC288_SubFPC303 (–26.04±26.45), p133zscl1_p310zscl1 (–14.91±9.78), p474zscl1_p474zscl2 (–13.61±6.09), p310zscl1_p474zscl1 (-7.79±9.19), SubFPC35_SubFPC303 (-6.26±21.25), p309zscl1_p130zscl3 (-4.92±6.34) and p133zscl2_p133zscl3 (–4.15±5.88). It can be observed that most of the descriptors with negative values are self cross–terms of proteins, which suggests the importance of intramolecular interaction within the protein in contributing to decreased pIC_{50} values, which makes the compound less potent. Nevertheless, it should be noted that the mechanisms contributing to aromatase inhibitor resistance may be of heterogeneous nature.

CONCLUSIONS

Computational approaches for predicting the activities of AIs can facilitate drug discovery efforts by saving cost and time. The continual increase in breast cancer prevalence has led to the necessity for discovery of novel compounds with strong inhibitory properties towards aromatase. To consider possible effects of aromatase on different AIs, we present a PCM study on aromatase inhibitory activity of AI along with amino acid residues that are at the binding sites and/or near the binding sides. By utilizing an efficient feature importance estimator, we find that the tautomerizable substructures containing nitrogen and carboxylic derivatives are highly important based on the pIC_{50} value. These findings may aid in the design of novel compounds that not only are capable of inhibiting aromatase but can also address the issue of aromatase inhibitor resistance.

ACKNOWLEDGMENTS

JESW and CN are supported by a joint grant from the Swedish Research Links program (no. C0610701) from the Swedish Research Council. This work is also partially supported by the Office of Higher Education Commission and Mahidol University under the National Research University Initiative.

REFERENCES

Albrecht, Ł., Albrecht, A., Ransborg, L. K., and Jørgensen, K. A. (2011). Asymmetric organocatalytic [3 + 2]-annulation strategy for the synthesis of N-fused heteroaromatic compounds. *Chem Sci*, 2(7):1273–1277.

Antoon, J. W., White, M. D., Slaughter, E. M., Driver, J. L., Khalili, H. S., Elliott, S., Smith, C. D., Burow, M. E., and Beckman, B. S. (2011). Targeting nfκb mediated breast cancer chemoresistance through selective inhibition of sphingosine kinase-2. *Cancer Biol Ther*, 11(7):678–689.

Auvray, P., Nativelle, C., Bureau, R., Dallemagne, P., Séralini, G.-E., and Sourdaine, P. (2002). Study of substrate specificity of human aromatase by site directed mutagenesis. *Eur J Biochem*, 269(5):1393–1405.

Bansal, R., Thota, S., Karkra, N., Minu, M., Zimmer, C., and Hartmann, R. W. (2012). Synthesis and aromatase inhibitory activity of some new 16e-arylidenosteroids. *Bioorg Chem*, 45:36–40.

Booth, G. D., Niccolucci, M. J., and Schuster, E. G. (1994). Identifying proxy sets in multiple linear regression: an aid to better coefficient interpretation. *US Dept of Agriculture Forest Service*.

Cadoo, K. A., Gucalp, A., and Traina, T. A. (2014). Palbociclib: an evidence-based review of its potential in the treatment of breast cancer. *Breast Cancer (Dove Med Press)*, 6:123.

Calle, M. L. and Urrea, V. (2011). Letter to the Editor: Stability of Random Forest importance measures. *Brief Bioinformatics*, 12(1):86–89.

Cao, D.-S., Xiao, N., Xu, Q.-S., and Chen, A. F. (2014). Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds, and their interactions. *Bioinformatics*, 31(2):279–281.

ChemAxon Ltd. (2014). *MarvinSketch, version 6.2.1, Budapest, Hungary*.

Dinno, A. (2012). *paran: Horn's Test of Principal Components/Factors*. R package version 1.5.1.

Eisen, A., Trudeau, M., Shelley, W., Messersmith, H., and Pritchard, K. I. (2008). Aromatase inhibitors in adjuvant therapy for hormone receptor positive breast cancer: a systematic review. *Cancer Treat Rev*, 34(2):157–174.

Fontham, E. T., Thun, M. J., Ward, E., Balch, A. J., Delancey, J. O. L., and Samet, J. M. (2009). American Cancer Society perspectives on environmental factors and cancer. *CA Cancer J Clin*, 59(6):343–351.

Fourches, D., Muratov, E., and Tropsha, A. (2010). Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model*, 50(7):1189–1204.

Ghosh, D., Griswold, J., Erman, M., and Pangborn, W. (2009). Structural basis for androgen specificity and oestrogen synthesis in human aromatase. *Nature*, 457(7226):219–223.

Group, E. B. C. T. C. et al. (2011). Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *Lancet*, 378(9793):771–784.

Hellberg, S., Sjoestroem, M., Skagerberg, B., and Wold, S. (1987). Peptide quantitative structure-activity relationships, a multivariate approach. *J Med Chem*, 30(7):1126–1135.

Kao, Y.-C., Cam, L. L., Laughton, C. A., Zhou, D., and Chen, S. (1996). Binding characteristics of seven inhibitors of human aromatase: a site-directed mutagenesis study. *Cancer Res*, 56(15):3451–3460.

Khodarahmi, G., Asadi, P., Farrokhpour, H., Hassanzadeh, F., and Dinari, M. (2015). Design of novel potential aromatase inhibitors via hybrid pharmacophore approach: docking improvement using the qm/mm method. *RSC Advances*, 5(71):58055–58064.

Kuhn, M. (2008). Building predictive models in R using the caret package. *J Stat Softw*, 28(5):1–26.

Laggner, C. (2009). SMARTS Patterns for Functional Group Classification.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Lipton, A., Santen, R. J., Santner, S. J., Harvey, H. A., Sanders, S. I., and Matthews, Y. L. (1992). Prognostic value of breast cancer aromatase. *Cancer*, 70(7):1951–1955.

Ma, C. X., Reinert, T., Chmielewska, I., and Ellis, M. J. (2015). Mechanisms of aromatase inhibitor resistance. *Nat Rev Cancer*, 15(5):261–275.

May, F. E. (2014). Novel drugs that target the estrogen-related receptor alpha: their therapeutic potential in breast cancer. *Cancer Manag Res*, 6:225–252.

McGovern, S. L., Caselli, E., Grigorieff, N., and Shoichet, B. K. (2002). A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J Med Chem*, 45(8):1712–1722.

Mevik, B.-H. and Wehrens, R. (2007). The pls package: principal component and partial least squares regression in R. *J Stat Softw*, 18(2):1–24.

Nantasenamat, C., Li, H., Mandi, P., Worachartcheewan, A., Monnor, T., Isarankura-Na-Ayudhya, C., and Prachayasittikul, V. (2013a). Exploring the chemical space of aromatase inhibitors. *Mol Div*, 17(4):661–677.

Nantasenamat, C., Worachartcheewan, A., Mandi, P., Monnor, T., Isarankura-Na-Ayudhya, C., and Prachayasittikul, V. (2014). QSAR modeling of aromatase inhibition by flavonoids using machine learning approaches. *Chem Pap*, 68(5):697–713.

Nantasenamat, C., Worachartcheewan, A., Prachayasittikul, S., Isarankura-Na-Ayudhya, C., and Prachayasittikul, V. (2013b). QSAR modeling of aromatase inhibitory activity of 1-substituted 1, 2, 3-triazole analogs of letrozole. *Eur J Med Chem*, 69:99–114.

Pingaew, R., Prachayasittikul, V., Mandi, P., Nantasenamat, C., Prachayasittikul, S., Ruchirawat, S., and Prachayasittikul, V. (2015). Synthesis and molecular docking of 1,2,3-triazole-based sulfonamides as aromatase inhibitors. *Bioorg Med Chem*, 23:3472–3480.

Prusis, P., Uhlén, S., Petrovska, R., Lapinsh, M., and Wikberg, J. E. (2006). Prediction of indirect interactions in proteins. *BMC Bioinformatics*, 7(1):167.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shoombuatong, W., Prachayasittikul, V., Prachayasittikul, V., and Nantasenamat, C. (2015). Prediction of aromatase inhibitory activity using the efficient linear method (ELM). *EXCLI J*, 13:452–464.

Simpson, E. R., Mahendroo, M. S., Means, G. D., Kilgore, M. W., Hinshelwood, M. M., Graham-Lorence, S., Amarneh, B., Ito, Y., Fisher, C. R., Michael, M. D., et al. (1994). Aromatase Cytochrome P450, The Enzyme Responsible for Estrogen Biosynthesis. *Endocr Rev*, 15(3):342–355.

Stevens, A. and Ramirez-Lopez, L. (2013). *An introduction to the prospectr package*. R package version 0.1.3.

Suvannang, N., Nantasenamat, C., Isarankura-Na-Ayudhya, C., and Prachayasittikul, V. (2011). Molecular docking of aromatase inhibitors. *Molecules*, 16(5):3597–3617.

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Mol Inf*, 29(6-7):476–488.

Warnes, G. R., Bolker, B., and Lumley, T. (2015). R package version 3.4.2.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*.

Worachartcheewan, A., Mandi, P., Prachayasittikul, V., Toropova, A. P., Toropov, A. A., and Nantasenamat, C. (2014a). Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors. *Chemometr Intell Lab Syst*, 138:120–126.

Worachartcheewan, A., Suvannang, N., Prachayasittikul, S., Prachayasittikul, V., and Nantasenamat, C. (2014b). Probing the origins of aromatase inhibitory activity of disubstituted coumarins via QSAR and molecular docking. *EXCLI J*, 13:1259–1274.

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*, 32(7):1466–1474.

Zhou, D., Cam, L. L., Laughton, C. A., Korzekwa, K. R., and Chen, S. (1994). Mutagenesis study at a postulated hydrophobic region near the active site of aromatase cytochrome p450. *J Biol Chem*, 269(30):19501–19508.

Zwick, W. R. and Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychol Bull*, 99(3):432.