Article

Discrete Fourier Transform based Multivariate Image Analysis: Application to Modeling Aromatase Inhibitory Activity.

Stephen J. Barigye, Matheus P. Freitas, Priscila Ausina, Patricia Zancan, Mauro Sola-Penna, and Juan Alberto Castillo-Garit

ACS Comb. Sci., **Just Accepted Manuscript** • DOI: 10.1021/acscombsci.7b00155 • Publication Date (Web): 03 Jan 2018

Discrete Fourier Transform based Multivariate Image Analysis: Application to Modeling Aromatase Inhibitory Activity.

Stephen J. Barigye*,[†]Matheus P. Freitas,[‡]Priscila Ausina,[£]Patricia Zancan,[§]Mauro Sola-Penna,[£]Juan A. Castillo-Garit.[‖]

[†]Department of Chemistry, McGill University, 801 Sherbrooke St. W., Montréal, QC, Canada.

[‡]Department of Chemistry, Federal University of Lavras,P.O.Box 3037, 37200-000, Lavras, MG, Brazil

[§]Laboratório de Oncobiologia Molecular (LabOMol), Departamento de Biotecnologia Farmacêutica, Faculdade de Farmácia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil

[£]Laboratório de Enzimologia e Controle do Metabolismo (LabECoM), Departamento de Biotecnologia Farmacêutica, Faculdade de Farmácia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil

[‖]Unidad de Toxicología Experimental, Universidad de Ciencias Médicas "Serafín Ruiz de Zárate Ruiz" Santa Clara, 50200, Villa Clara, Cuba.

ABSTRACT: We have recently generalized the formerly alignment-dependent **MIA-QSAR** (acronym for **M**ultivariate **I**mage **A**nalysis applied to **Q**uantitative **S**tructure **A**ctivity **R**elationships) method through the application of the Discrete Fourier Transform (DFT), allowing for its application to non-congruent and structurally diverse chemical compound datasets. Here, we report the first practical application of this method in the screening of molecular entities of therapeutic interest, with the human aromatase inhibitory activity as the case study. We develop an ensemble classification model based on 2D-DFT MIA-QSAR descriptors, with which we screen 34 chemical compounds with possible aromatase inhibitory activity from the NCI Diversity Set V (1593 compounds). These compounds are docked into the aromatase active site and 10 most promissory compounds selected for *in vitro* experimental validation. Of these compounds, **7419** (non-steroidal) and **89201** (steroidal) demonstrate satisfactory antiproliferative and aromatase inhibitory activities. The obtained results suggest that the 2D-DFT MIA-QSAR method may be useful in the ligand-based virtual screening of NMEs of therapeutic utility.

INTRODUCTION

Molecular modeling methods have over the years evolved, permeating all stages of the drug discovery process.[1-3] The extrapolation of concepts and/or methods particular to a wide range of disciplines such as biophysics, computer science, statistics, mathematics, structural biology and chemistry, to drug discovery paradigms, coupled with the availability of volumes of experimental data ranging from bioactivity parameters to crystallographic and NMR data, have fostered the advancement of existing computational tools to achieve more accurate models of chemical structures and their interactions with biological targets.[4-7]

We have recently developed a generalization scheme for the **MIA-QSAR** (acronym for **M**ultivariate **I**mage **A**nalysis-**Q**uantitative **S**tructure **A**ctivity **R**elationship) method based on the extrapolation of the 2D-Discrete Fourier Transform (2D-DFT) to chemical structural images.[8-10] The MIA-QSAR method is based on the use of pixels comprising 2D chemical structure images as descriptors of the topo-chemical properties of molecules. Notwithstanding the successful application of the MIA-QSAR approach in the modeling of numerous bioactivities for over a decade, this method has been until recently exclusively applicable to congeneric datasets, due to the requirement that chemical structural images are aligned with respect to the basic molecular scaffold.[10] This prerequisite as consequence rendered MIA-QSAR method inapplicable to virtual screening tasks for new molecular entities of therapeutic interest. We have demonstrated that the application of the 2D-DFT to chemical structural images eliminates the necessity of alignment of chemical images, allowing for the modeling of structurally diverse chemical datasets. Scheme 1 illustrates the workflow followed in the derivation of the 2D-DFT MIA-QSAR descriptors.

In the present report, we aim to demonstrate the possible utility of the 2D-DFT MIA-QSAR approach in drug discovery process, via integrated ligand- and structure-based virtual screening workflow for novel cytochrome P450 aromatase inhibitors, coupled with posterior *in vitro* experimental validation. Human aromatase cytochrome P450 catalyzes the final step of the conversion of androgens (*i.e.* androstenedione, 16α-hydroxytestosterone and testosterone) to estrogens (*i.e*. estrone, 17β, 16α-estriol and 17β-estradiol, respectively).[11] Clinical evidence has demonstrated that breast adenocarcinomas of postmenopausal women possess much higher levels of 17β-estradiol relative to plasma concentrations and thus the inhibition of the aromatase catalyzed biosynthesis of estrogens constitutes an attractive target for the prophylaxis and therapy of estrogen-dependent breast cancer.[12, 13]

RESULTS AND DISCUSSION

Classification Model Building. The built base classifier models for predicting the aromatase inhibitory activity, based on the 2D-DFT MIA-QSAR descriptors, on the whole demonstrated high accuracy, sensitivity and specificity, for the training and test sets, respectively, with each classifier exhibiting a particular strength. Table 1 shows the statistical parameters for the built classifier models. Data matrix employed in the model building is provided as supporting information, S1. For the training set, the Boosted Trees (BT) model yielded the highest accuracy (96.47%), sensitivity (93.75%) and specificity (98.62%) values. As for the test set, the Artificial Neural Networks (ANN) model produced the best accuracy (88.74%) and sensitivity (86.27%), followed by the k-Nearest Neighbor (k-NN) model. On the other hand, the highest specificity was obtained with Linear

Discriminant Analysis (LDA) model (95.00%), notwithstanding its much lower sensitivity (71.57%), followed by the Support Vector Machine (SVM) model (92.50%). In this sense, given the overall good performance of these base classifiers, a majority vote ensemble was built to balance out the weaknesses of individual base classifiers, and thus enhancing the robustness and predictive power of the classification models. The accuracy, sensitivity and specificity values for the built ensemble were 86.94%, 81.37% and 91.67%, respectively (Table 1). Moreover, the 2D-DFT MIA-QSAR models were validated using a set of decoys, yielding high specificity (93.79%) for the ensemble, as well as the base classifiers BT (98.81%), SVM(90.09%) and LDA (87.84%), while moderate specificity was obtained for ANN (59.41%) and k-NN (60.84%). The generally good performance of these models demonstrates the utility of the 2D-DFT MIA-QSAR descriptors in codifying vital chemical structural information and may thus be considered as suitable in the virtual screening of chemical compounds of therapeutic interest. Detailed data for the predictions of each classifier is provided as Supporting Information, S2 (modeling set) and S3 (decoy set).

Table 1 comes about here

Virtual Screening for Novel Aromatase Inhibitors. Using the built multi-classifier model, 34 compounds were screened from NCI Diversity Set V as potential aromatase inhibitors. The screened compounds were docked into the aromatase active site to assess their possible interaction modes and their binding affinities estimated according to the $PLANTS_{PLP}$ empirical scoring function.[14] Based on the obtained scores, 10 chemical compounds were selected for *in vitro* MCF-7 antiproliferative and aromatase inhibitory activity assays. The structures of the screened compounds, as well as the corresponding docking scores, are shown in Table 2 and demonstrate the structural diversity of the selected compounds.

Table 2 comes about here

Inhibition of MCF-7 Cells Proliferation and Aromatase Activity. First, we evaluated the viability of TST-stimulated MCF-7 cells treated for 24 hours with 50 μM for each of the10 chemical compounds. Tyrphostin A9 was used as a positive control. As shown in Figure 1a, 5 compounds (*i.e.* **7419**, **54709**, **89201**, **310354** and **661122**) were able to reduce cell viability by more than 50%. Subsequently, the inhibitor concentrations were lowered to 10 μM and the corresponding cell viability determined. At this concentration, only **7419** and **89201** negatively affected the viability of MCF-7 breast cancer cells (Figure 1b). In order to attest the cytochrome P450 aromatase inhibitory activity of the screened compounds, an evaluation of the enzymatic activity was performed using a cell-free extract of MCF-7 cells in presence of 10 μM for each of the compounds. As can be observed in Figure 2a, compounds **7419**, **89201** and Tyrphostin A9 exhibited aromatase inhibitory activity. Based on this result, **7419**, **89201** and Tyrphostin A9 were evaluated for their anti-proliferative activity against TST stimulated MCF-7 cells at 10 μM concentrations and all the three compounds strongly inhibited the proliferation of MCF-7 cells (Figure 2b). Finally, dose response curves were obtained for the antiproliferative activity of compounds **7419** [2-chloro-N-(2-(diethylamino)ethyl)-2-(perchlorocyclopenta-2,4-dien-1-ylidene)acetamide], **89201** [(8R,9S,13S,14S,17S)-17-hydroxy-13-methyl-7,8,9,11,12,13,14,15,16,17-decahydro-6H-cyclopenta[a]phenanthren-3-yl bis(2-chloroethyl)carbamate] and **Tyrphostin A9**, and the IC_{50} values were determined to be 4.9 ± 1.1 μM, 4.1 ± 0.9 μM and 0.2 ± 0.1μM, respectively (Figure 3). Note that compound **89201** is in fact a well-known alkylating agent (i.e. estramustine) whose phosphate ester is a clinically approved drug that has been employed for decades in prostate cancer therapy.[15, 16] On the other hand, although off-target activity evaluations are beyond the objectives of

the present study, it should be noted that compound **7419** is in fact a cyclic vinyl chloride derivative. This class of compounds is notorious for presenting several nonspecific interactions and therefore caution should be exercised in pursuing **7419** as a clinically viable candidate.

Figure 1 comes about here

Figure 2 comes about here

Figure 3 comes about here

CONCLUSION

The results obtained with the 2D-DFT based generalization scheme of the MIA-QSAR method in the screening of novel aromatase inhibitors, albeit the modest chemical space explored, demonstrate the utility of this approach in the modeling of the bioactivities of chemical datasets and in the virtual screening of chemical compounds with desired therapeutic profiles. Moreover, in the *in silico* screening experiment, we employed a combination of ligand-based and structure-based virtual screening approaches based on the notion that one method helps to minimize the weaknesses of the other. In a broader sense, the ligand-based VS techniques are based on the similarity principle in that similar compounds demonstrate similar bioactivity profiles. However, this concept may be flawed given that similar compounds may not yield similar interaction profiles with a given active site and thus structure-based VS serves as a validation of the ligand-based VS results. On the other hand, compounds with high binding affinity to an active site may not yield the desired activity (*i.e.* agonists instead of antagonists and *vice versa*). In this sense, a combined ligand and structure-based approach helps to limit the weaknesses of each

method. Future initiatives include the application of the inverse DFT to identify functional groups or substructures vital for modeled bioactivities. Moreover, other multivariate image processing techniques such as Discrete Cosine, Wavelets and Walsh-Hadamard transforms will be explored for their utility in the MIA-QSAR context.

MATERIALS AND METHODS

Chemical Dataset and 2D-DFT MIA-QSAR. For this study, we extracted from the literature a dataset of 973 chemical compounds with known aromatase inhibitory activity profiles.[17] For each of these chemical structures, their corresponding structural images were retrieved and transformed into magnitude spectra whose coordinates correspond to the spatial frequencies and the pixel values the intensities (*i.e.* Fourier coefficient magnitudes) of these frequencies. Posteriorly, consistent with the MIA-QSAR approach, the pixels of the magnitude spectra were considered as the variables codifying topo-chemical information for each of these structures (scheme 1). For a detailed treatise on the theoretical aspects of the 2D-DFT MIA-QSAR method, see references.[8, 9] Given the characteristic symmetry presented by Fourier spectra, only half of the data points were employed. As anticipated, considering the pixels of magnitude spectra as descriptors generates high dimensional data matrices and thus dimensionality reduction procedures are necessary in order to work with manageable matrix dimensions. To this end, a combination of unsupervised and supervised feature selection procedures were performed using our freely available software, IMMAN (acronym for **I**nformation theory-based che**M**o**M**etrics **AN**alysis),[18] and a low-dimensional data was obtained with which the posterior experiments were performed.

Data Set Splitting and Classification Model Building. The chemical compounds were split into actives (pIC50 > 6.0) and inactives (pIC50 < 6.0), comprising of 390 and 483 chemical structures, respectively. Consistent with good practices of QSAR modeling, each group was divided into training and test sets, using hierarchical and *k*-means cluster analysis methods, with Ward's algorithm as the amalgamation rule and the squared Euclidean distance as the proximity function, respectively. Consequently, training and test set sizes of 651 and 222 chemical compounds, respectively, were obtained. We built a majority (hard) voting classifier comprising of 5 conceptually distinct base classifiers, *i.e.* Linear Discriminant Analysis (LDA), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Boosted Trees (BT) and k-Nearest Neighbor (k-NN). For the ANN approach, the MLP (multi-layer perceptron) model was employed with network architecture of 14-10-2, while the logistic and SOS functions were employed as the hidden activation and error functions, respectively. For the SVM model, the RBF kernel was employed with the tunable parameters γ and C equal to 0.55 and 13, respectively. The meta-parameters shrinkage (ν) and training set fraction (η) for the BT model were set to 0.13 and 0.30, respectively. For the LDA model, the tolerance parameter was fixed at 0.01 and the forward-step wise selection procedure was employed as the variable selection strategy. The dataset splitting and model building was performed using STATISTICA software.[19] Several reports in the literature have demonstrated that consensus models yield the most accurate predictions, in addition to eliminating the dependence on training and set statistics in the selection of models for virtual screening tasks.[20-22]

Virtual Screening and Molecular Docking Validation. The NCI Diversity Set V (1593 compounds) was screened for compounds with potential inhibitory activity against CYP19A1 using the built ensemble model. The choice of this dataset was attributed to its

inherent diversity and the availability of vialed samples from the DTP repository for posterior experimental validation. In the generation of this diversity dataset, priority was given to chemical compounds with pharmacologically desirable characters (*i.e.* devoid of weakly bonded heteroatoms, obvious leaving groups, polycyclic aromatic hydrocarbons, etc.) and thus constitutes an interesting source of possible novel lead compounds, notwithstanding the small size. The majority vote criterion was considered in the prediction of the aromatase inhibitory activity of the screened compounds [*i.e.* compounds were considered as active or inactive if the probability (active or inactive) ≥ 0.6].

On the other hand, while ligand-based virtual screening is in a broader sense, based on the principle of chemical structural similarity, in that compounds with similar chemical structural characteristics should exhibit similar bioactivity, favorable protein-ligand interactions play a central role in determining the possible activity of screened compounds. In this sense, the chemical compounds predicted as promissory aromatase inhibitors were posteriorly screened using a protein-ligand docking procedure and the ligands with the lowest free-energy binding were selected. For the docking procedure, the CLC Drug Discovery Workbench was used and a 2.75 Å crystal structure of human placental aromatase cytochrome P450 in complex with androstenedione (PDB code 3S79) was obtained from the Protein Data Bank. The carboxylate moiety of the Asp309 side chain of the aromatase structure was protonated consistent with previous reports that validated its role as a proton donor in the aromatization reaction.[23, 24] The most promissory chemical compounds based on the integrated ligand-based and structure-based virtual screening were posteriorly submitted to experimental validation to attest their antiproliferative and inhibitory activity. Moreover, retrospective validation of the built 2D-DFT MIA-QSAR models was performed using a decoy set comprising of 3348 chemical compounds

(supporting information, S3) generated based on the 50 most active aromatase inhibitors, using the Directory of Useful Decoys, Enhanced (DUD-E) utility.[25]

Chemical Compounds and MCF-7Cell Line. The screened chemical compounds were kindly provided by Development Therapeutics Program (DTP), National Cancer Institute, USA. The human breast cancer cell line MCF-7 was obtained from the Cell Bank of Hospital Universitário Clementino Fraga Filho, UFRJ, Brazil, and was maintained in DMEM (Dulbecco's modified Eagle's medium; Invitrogen, São Paulo, SP, Brazil) supplemented with 10% (v/v) FBS (fetal bovine serum; Invitrogen) and L-glutamine.[26] For all the experiments, cells growth was stimulated by 1 nM testosterone (TST). The TST - stimulated cell viability and cell proliferation experiments, the difference between the results in the presence and in the absence of TST were considered.

Cell Viability. The MCF-7 cells viability was assessed through the MTT assay as described previously.[27] Cells were seeded in 96-well plates (2×10^3 cells/well) and incubated at 37 °C and 5% CO_2 for 48 hours. After this initial incubation, the medium was replaced by a fresh medium containing the 1 nMTST and the desired concentration of the drugs. This procedure was repeated 5-times at each 2 days of treatment. Then, the medium was removed, fresh medium was added, and the cells were returned to the incubator in the presence of different drugs used. In the end of the treatment, the medium was removed and the cells were incubated with 5 mg/mL MTT reagent (3,4,5-dimethiazol-2,5-diphenyltetrazolium bromide, Sigma–Aldrich Co., St. Louis, MO, USA) for 3 h. Thereafter, the formazan crystals formed were dissolved in DMSO, and the absorbance at 560 nm was evaluated using a VICTOR3 multilabel microplate reader (PerkinElmer, Waltham, MA, USA) with subtraction of the background absorbance at 670 nm.[27]

Cell Proliferation Assay. The MCF-7 cells proliferation was assessed through the CyQuant kit (Thermo Fisher Scientific, Waltham, MA, USA) that evaluates the DNA content of the cells as a measurement of cell number. For this, cells were seeded in 96-well plates (2 x 10^3 cells/well) and incubated at 37 °C and 5% CO_2 for 48 hours. After this initial incubation, the medium was replaced by a fresh medium containing the 1 nMTST and the desired concentration of the drugs. This procedure was repeated 5-times at each 2 days of treatment. Then, the medium was removed, a fresh medium added and the cells were returned to the incubator in the presence of different drugs used. At the end of the treatment, the medium was removed and 200 μL of the freshly prepared CyQuant reagent was added to each well and after 5 min the fluorescence emission at 520 nm was read in a VICTOR3 multilabel microplate reader (PerkinElmer, Waltham, MA, USA).

Aromatase Inhibitory Activity. Aromatase activity was assessed in a MCF-7 free cell homogenate. For this, MCF-7 cells (10^5 cells) were seeded in 75 cm^2 culture flasks and grown in DMEN supplemented with 10% FBS at 37 °C and 5% CO_2. When cells reached approximately 80% of confluence, cells were hashed and homogenized in Cell Lysis Buffer (Cell Signaling Technology, #9803, Danvers, MA, USA) in a proportion of 5 x 10^6 cells in 10 mL of buffer. Aromatase activity was assayed in a reaction mixture containing 50 mMTris-HCl (pH 7.4), 5 mM $MgCl_2$, 120 mMKCl, 0.1 mM TST and 0.2 mM NADPH. The reaction was initiated by the addition of 100 μL of the MCF-7 cell free homogenate. Appropriate controls were performed to consider the effects of TST vehicle (ethanol) and the drugs vehicle (DMSO). Oxidation of NADPH to $NADP^+$ was followed measuring the fluorescence emission at 340 nm in a VICTOR3 multilabel microplate reader (PerkinElmer, Waltham, MA, USA). The aromatase activity was calculated discounting the slope of the

fluorescence decay in the presence of TST from its counterpart in the absence of TST. All experiments were run in triplicate in a series of three independent experiments (n = 3).

ABBREVIATIONS

TST, Testosterone; DFT, Discrete Fourier Transform; MIA-QSAR, Multivariate Image Analysis applied to Quantitative Structure–Activity Relationships

ACKNOWLEDGMENT

The present work was supported by grants from Fundação Carlos Chagas Filho de Apoio à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). The authors would also like to thank the Drug Synthesis and Chemistry Branch (DSCB), Development Therapeutics Program (DTP), National Cancer Institute, USA for providing the chemical compounds for the *in vitro* assays.

Supporting Information. Data matrix employed in the model building, activity predictions for each compound in employed dataset. This material is available free of charge via the Internet at http://pubs.acs.org

AUTHOR INFORMATION

Corresponding Author

*Phone: 1-514-660-9351. E-mail: stephen.barigye@mcgill.ca

REFERENCES

1. Jorgensen, W. L., The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813-1818.

2. Kar, S.; Roy, K., How far can virtual screening take us in drug discovery? *Expert Opin Drug Discov.* **2013**, 245-261.

3. Passeri, G. I.; Trisciuzzi, D.; Alberga, D.; Siragusa, L.; Leonetti, F.; Mangiatordi, G. F.; Nicolotti, O., Strategies of Virtual Screening in Medicinal Chemistry. *Int. J. Quant. Struct. Prop. Relat.* **2018**, *3*, 134-160.

4. Todeschini, R.; Consonni, V., *Molecular Descriptors for Chemoinformatics*. Wiley-VCH: Weinheim, 2009; Vol. 1, p 1265.

5. Barigye, S. J.; Marrero-Ponce, Y.; Pérez-Giménez, F.; Bonchev, D., Trends in information theory-based chemical structure codification. *Mol. Divers.* **2014**, *18*, 673-686.

6. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Computational methods in drug discovery. *Pharmacol. Rev.* **2014**, *66*, 334-395.

7. Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R., QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **2014**, *57*, 4977-5010.

8. Barigye, S. J.; Freitas, M. P., 2D-Discrete Fourier Transform: Generalization of the MIA-QSAR strategy in molecular modeling. *Chemom. Intell. Lab. Syst.* **2015**, *143*, 79-84.

9. Barigye, S. J.; Freitas, M. P., Is molecular alignment an indispensable requirement in the MIA□QSAR method? *J Comp Chem* **2015**, *36*, 1748-1755.

10. Barigye, S. J.; de Freitas, M. P., Ten years of the MIA-QSAR strategy: historical development and applications. *Int. J. Quant. Struct. Prop. Relat.* **2016**, *1*, 64-77.

11. Ghosh, D.; Griswold, J.; Erman, M.; Pangborn, W., Structural basis for androgen specificity and oestrogen synthesis in human aromatase. *Nature* **2009**, *457*, 219-223.

12. Yadav, M. R.; Barmade, M. A.; Tamboli, R. S.; Murumkar, P. R., Developing steroidal aromatase inhibitors-an effective armament to win the battle against breast cancer. *Eur. J. Med. Chem.* **2015**, *105*, 1-38.

13. Ghosh, D.; Lo, J.; Egbuta, C., Recent progress in the discovery of next generation inhibitors of aromatase from the structure–function perspective. *J. Med. Chem.* **2016**, *59*, 5131-5148.

14. Korb, O.; Stutzle, T.; Exner, T. E., Empirical scoring functions for advanced protein− ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84-96.

15. Di Lorenzo, G., Estramustine in prostate cancer: new look at an old drug. *Lancet Oncol.* **2007**, *8*, 959-961.

16. Rautio, J.; Kumpulainen, H.; Heimbach, T.; Oliyai, R.; Oh, D.; Järvinen, T.; Savolainen, J., Prodrugs: design and clinical applications. *Nat. Rev. Drug Discov.* **2008**, *7*, 255-270.

17. Worachartcheewan, A.; Mandi, P.; Prachayasittikul, V.; Toropova, A. P.; Toropov, A. A.; Nantasenamat, C., Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors. *Chemom. Intell. Lab. Syst.* **2014**, *138*, 120-126.

18. Urias, R. W. P.; Barigye, S. J.; Marrero-Ponce, Y.; García-Jacas, C. R.; Valdes-Martiní, J. R.; Perez-Gimenez, F., IMMAN: free software for information theory-based chemometric analysis. *Mol. Divers.* **2015**, *19*, 305-319.

19. Statsoft-Team *STATISTICA. Data Analysis Software System*, version 6.0; StatSoft: Tulsa, 2001.

20. Rokach, L., Ensemble-based classifiers. *Artif Intell Rev* **2010**, *33*, 1-39.

21. Marrero-Ponce, Y.; Siverio-Mota, D.; Gálvez-Llompart, M.; Recio, M. C.; Giner, R. M.; García-Domènech, R.; Torrens, F.; Arán, V. J.; Cordero-Maldonado, M. L.; Esguera,

C. V., Discovery of novel anti-inflammatory drug-like compounds by aligning in silico and in vivo screening: the nitroindazolinone chemotype. *Eur. J. Med. Chem.* **2011**, *46*, 5736-5753.

22. Ventura, C.; Latino, D. A.; Martins, F., Comparison of multiple linear regressions and neural networks based QSAR models for the design of new antitubercular compounds. *Eur. J. Med. Chem.* **2013**, *70*, 831-845.

23. Di Nardo, G.; Breitner, M.; Bandino, A.; Ghosh, D.; Jennings, G. K.; Hackett, J. C.; Gilardi, G., Evidence for an elevated aspartate pKa in the active site of human aromatase. *J. Biol. Chem* **2015**, *290*, 1186-1196.

24. Adhikari, N.; Amin, S. A.; Saha, A.; Jha, T., Combating breast cancer with non-steroidal aromatase inhibitors (NSAIs): Understanding the chemico-biological interactions through comparative SAR/QSAR study. *Eur. J. Med. Chem.* **2017**, 365-438.

25. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K., Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582-6594.

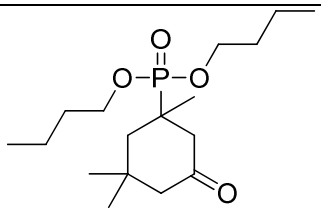
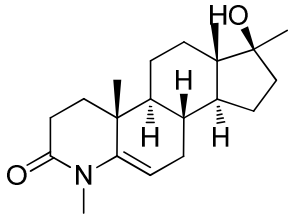
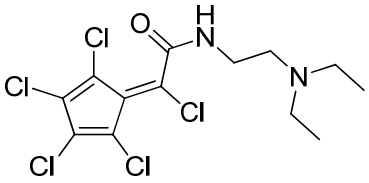
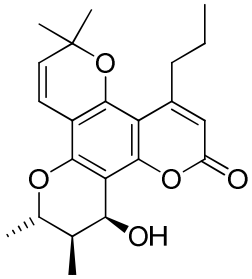
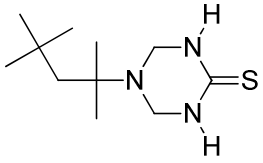
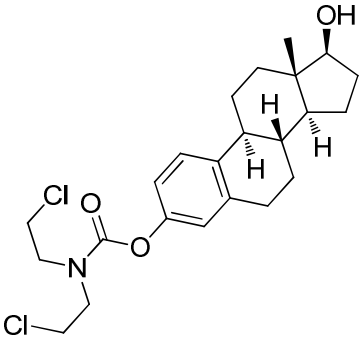
26. Zancan, P.; Sola-Penna, M.; Furtado, C. M.; Da Silva, D., Differential expression of phosphofructokinase-1 isoforms correlates with the glycolytic efficiency of breast cancer cells. *Mol. Genet. Metab.* **2010**, *100*, 372-378.

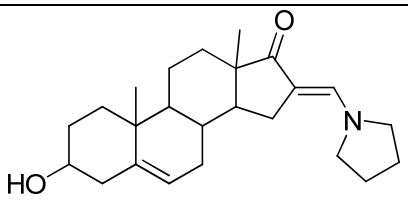
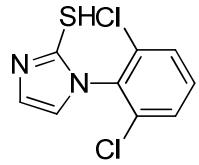
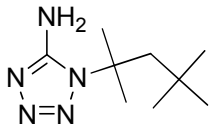
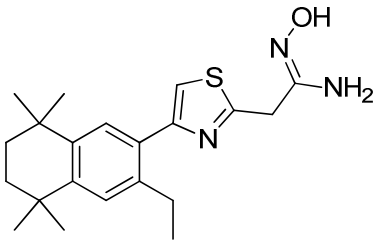
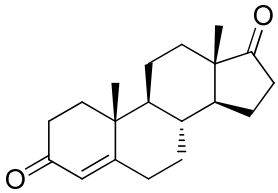
27. Spitz, G. A.; Furtado, C. M.; Sola-Penna, M.; Zancan, P., Acetylsalicylic acid and salicylic acid decrease tumor cell viability and glucose metabolism modulating 6-phosphofructo-1-kinase structure and activity. *Biochem. Pharmacol.* **2009**, *77*, 46-53.

Table 1. Statistical parameters of the base classifier models and majority-vote ensemble model.

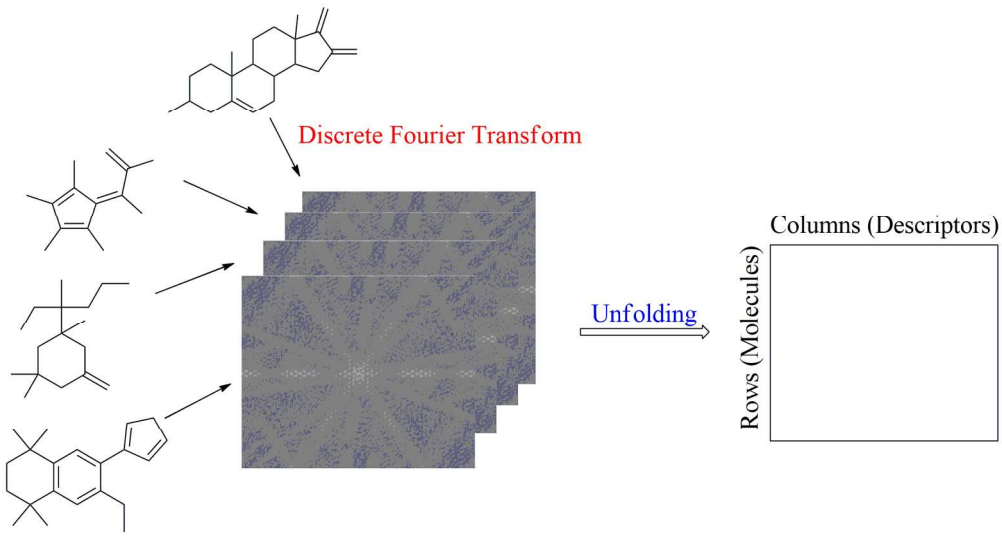
Classification Models		ANN	LDA	SVM	BT	k-NN	Ensemble
Accuracy	Training	87.71%	80.34%	87.71%	96.47%	-	-
	Test	88.74%	84.23%	85.14%	84.68%	85.59%	86.94%
Sensitivity	Training	87.15%	71.88%	80.56%	93.75%	-	-
	Test	86.27%	71.57%	76.47%	81.37%	82.35%	81.37%
Specificity	Training	88.15%	87.05%	93.39%	98.62%	-	-
	Test	90.83%	95.00%	92.50%	87.50%	88.33%	91.67%
	Decoy	59.41%	87.84%	90.09%	98.81%	60.84%	93.79%

Table 2. Chemical structures, docking scores and experimental IC_{50} values for screened compounds with possible aromatase inhibitory activity, based on the 2D-DFT MIA-QSAR ensemble classification model and the molecular docking technique.

ID	Chemical Structure	$PLANTS_{PLP}$ Score	IC_{50}, μM
14506		-73.43	> 50.0
36923		-68.31	> 50.0
7419		-64.24	**4.9**±1.1
661122		-62.36	> 10.0
319034		-61.69	> 50.0
89201		-61.43	**4.1**±0.9

54709		-57.03	> 10.0
321506		-54.69	> 50.0
11128		-53.39	> 50.0
310354		-51.46	> 10.0
Androstenedione		-84.54	-

Scheme1. Workflow followed in the derivation of the 2D-DFT MIA-QSAR descriptors.



Scheme 1

168x90mm (300 x 300 DPI)

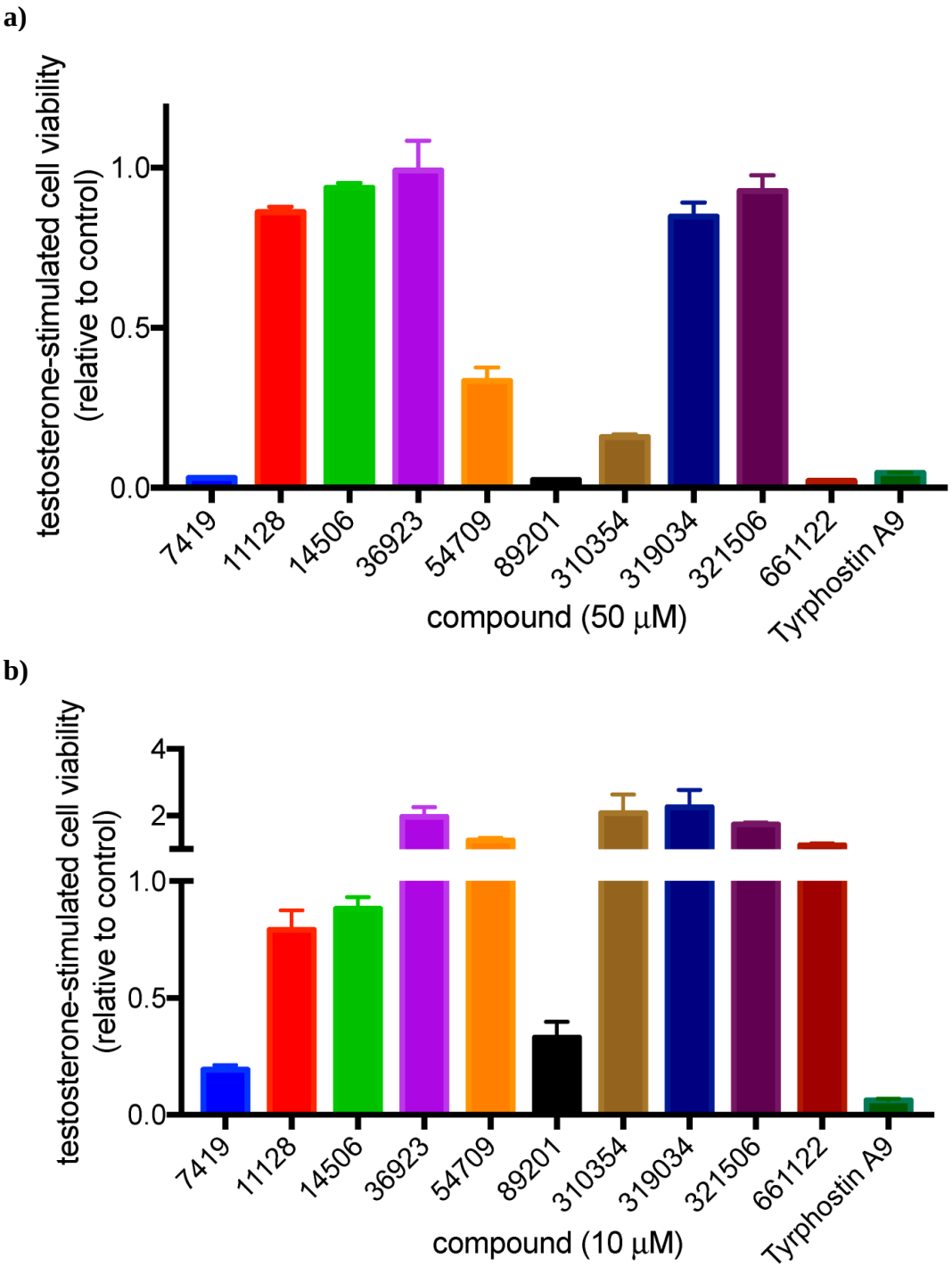


Figure 1.Measurement of TSTstimulated MCF-7 cell anti-proliferative activity of 10 compounds screened from NCI Diversity Set V at a) 50 μM concentrations b) 10 μM concentrations. Tyrphostin A9 is employed as a positive control.

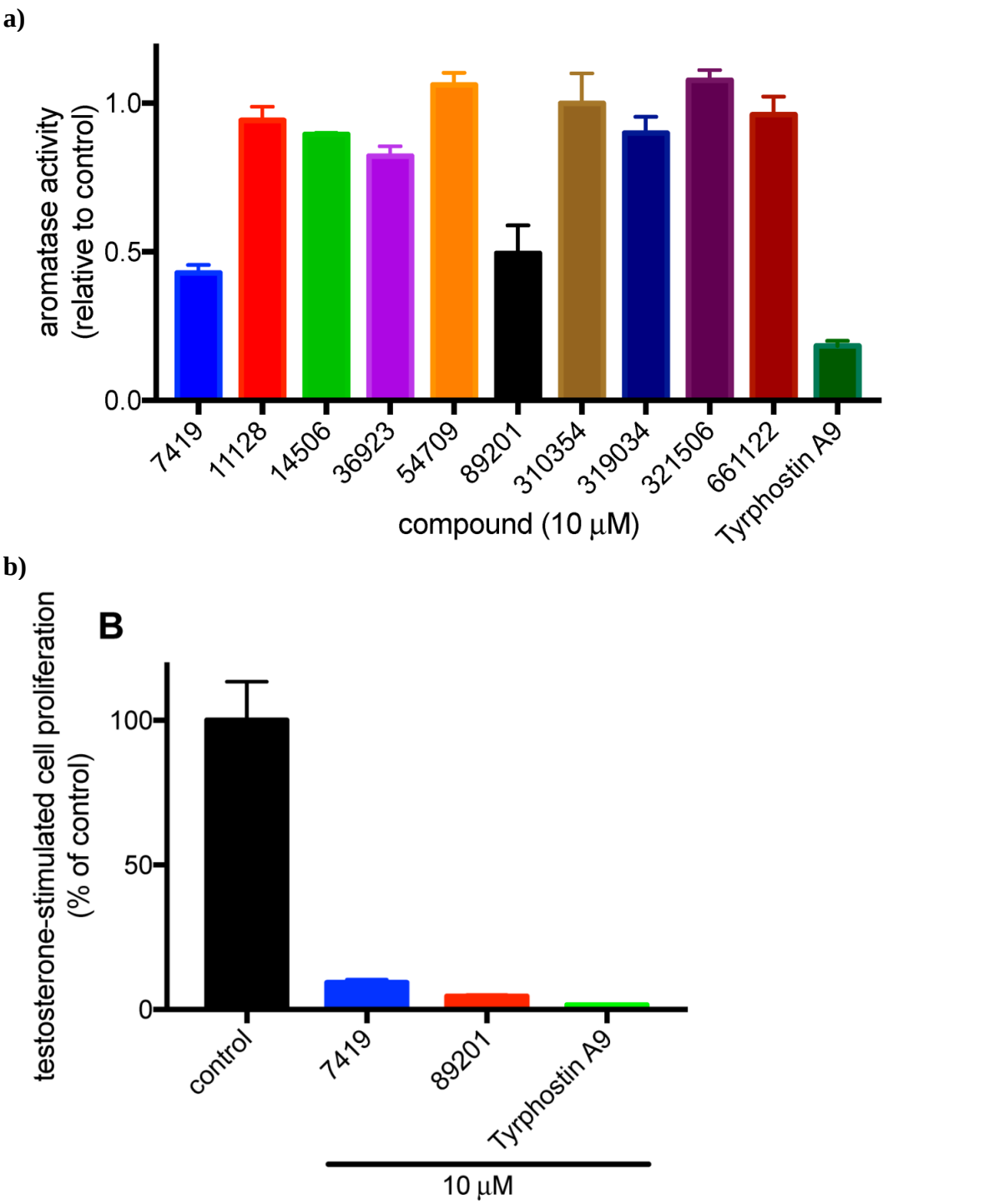


Figure 2.a) Aromatase inhibitory activity of 10 compounds screened from NCI Diversity Set V at 10 μM concentrations.b)Anti-proliferative activity of compounds 7419, 89201 and Tyrphostin A9 at 10 μM concentrations.

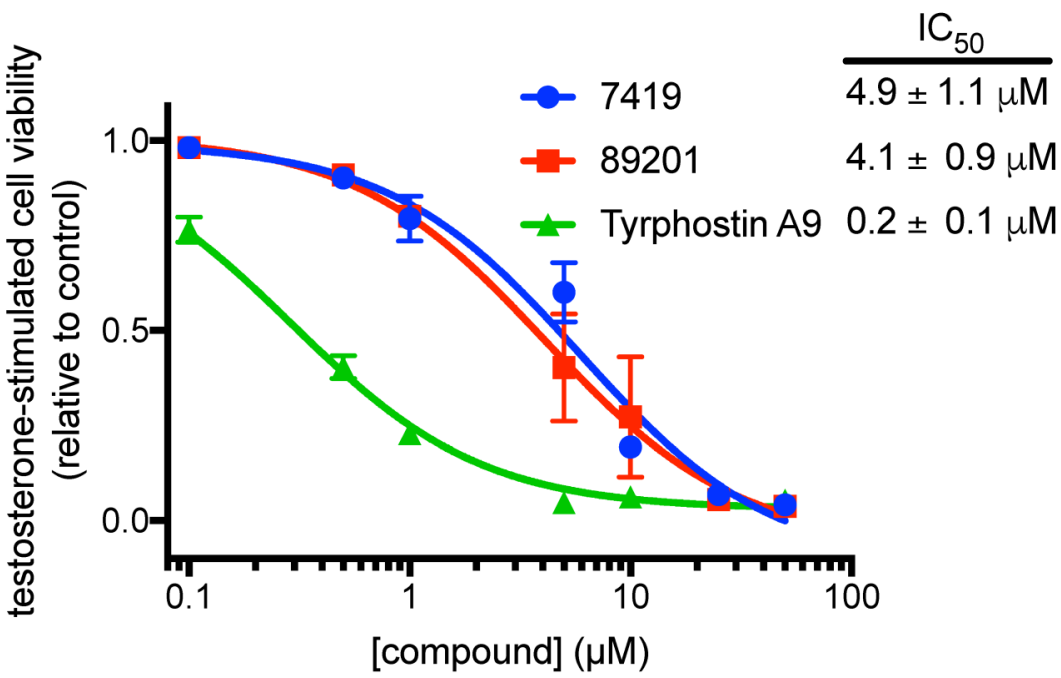
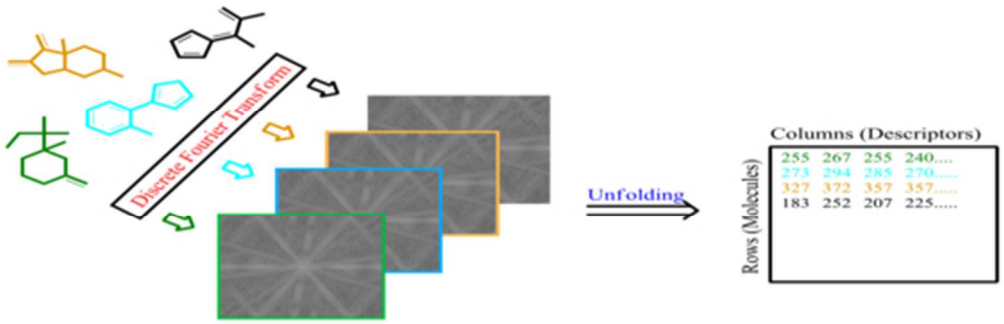


Figure 3. Dose response curve of TSTstimulated MCF-7 cell antiproliferative activity for 7419, 89201 and Tyrphostin A9. All experiments were run in triplicate in a series of three independent experiments.



TOC Graphic

148x47mm (96 x 96 DPI)