Journal of Experimental and Clinical Medicine
https://dergipark.org.tr/omujecm

Research Article

J Exp Clin Med
2022; 39(4): 1043-1050
doi: 10.52142/omujecm.39.4.23

Investigation, design and synthesis of new anticancer agents with anticancer effect potential on MCF-7 breast cancer cells by machine learning method

Suat Utku KESKİN[1], Volkan BÜLBÜL[2], Mervenur KALENDER[3], Sümeyye ÖZYAMAN[3] Arif MERMER[4,*]

[1]Robert College, İstanbul, Türkiye
[2]Department of Histology and Embryology, Faculty of Medicine, Nişantaşı University, İstanbul, Türkiye
[3]Department of Histology and Embryology, Faculty of Medicine, İstanbul Medipol University, İstanbul, Türkiye
[4]Experimental Medicine Application and Research Center, University of Health Sciences, İstanbul, Türkiye

Received: 28.07.2022 • **Accepted/Published Online:** 28.08.2022 • **Final Version:** 29.10.2022

Abstract

Cancer is one of the diseases with a high mortality rate, which occurs when cells multiply uncontrollably, acquire an invasive character and metastasize. Breast cancer is one of the cancer types with an increasing incidence worldwide. Chemotherapy is a method used in the treatment of cancer diseases, and the chemotherapeutic drugs used inhibit the growth and proliferation of cancer cells due to their cytotoxic properties. Today, machine learning techniques offer significant advantages by helping several steps of the drug discovery process, reducing the time spent in the laboratory, the use of consumables and chemical materials, and the maximum time predicted for the discovery of a drug with traditional methods. In our study, it was aimed to determine the 3 Schiff base derivatives with the most active cytotoxic effect on breast cancer cells from the large data set using machine learning. In our study, 7 Schiff base derivatives were determined from a large data set containing 98 compounds, and the 3 most active compounds with cytotoxic properties on breast cancer cells and their IC50 values were determined by machine learning method. In the future, it is thought that compound 1 can be used as an alternative to pharmacological applications to be used in preclinical studies as a therapeutic agent, supported by *in vitro* and *in vivo* applications, in order to be used in cancer treatments.

Keywords: machine learning, drug discovery, breast cancer, MCF-7

1. Introduction

Breast cancer is the second most frequently diagnosed cancer worldwide, with a frequency of up to 11.9%. It constitutes 25.2% of all newly diagnosed cancers and is the leading type of cancer encountered in women. It has been observed that only 5-10% of these cases are caused by genetic disorders, while the remaining 90-95% are due to environmental factors and lifestyle (1, 2). Hormonal factors, overexposure to estrogen, and subsequent differentiation of breast cells can affect a woman's risk of developing breast cancer (3). While obesity and weight gain in the postmenopausal period are associated with a higher risk of breast cancer, this association is not present for premenopausal women. The number of first-degree female relatives with a history of breast cancer significantly influences an individual's risk of breast cancer (4-6).

The use of drugs in cancer treatment has been extensively studied since the 1940s (7). Chemotherapy is an important part of treatment for many types of cancer. For this reason, the effort to develop new anti-cancer drugs constitutes one of the largest areas in the pharmaceutical industry (8). It is known that a wide variety of chemotherapeutic agents are used clinically in sending cancer cells to death by apoptosis. Some of the most commonly used of these are Cisplatin, Taxol and Doxorubucin (9). While an increase has been observed in the number of anti-cancer agents developed in the last 10 years, the number of agents that can successfully progress clinically from these products is less than 10%. The two most important reasons why anti-cancer agents cannot be accepted for treatment are their lack of clinical efficacy and high toxicity values. The product intended to be used as an anti-cancer agent must pass the phase-III stage during clinical tests. Drug development studies have various difficulties due to the high cost of approximately 1.8 billion dollars and the 9 to 12 years to be approved (10).

Traditional methods of cancer treatment today, including radiotherapy and chemotherapy, are expensive and often have harmful side effects on healthy cells. In addition, cancer cells have the ability to develop resistance to existing chemotherapeutic drugs (11). Therefore, there is a constant need for the development of new anticancer drugs to reduce the proliferation of cancer cells. While traditional *in vitro* prediction strategies developed in this context face time and

*Correspondence: arif.mermer@sbu.edu.tr

cost constraints, potential interaction candidates can be predicted more efficiently by computational or *in silico* methods, which have recently become widespread (12). For all these reasons, the interest in the use of machine learning (ML) methods in the pharmaceutical industry is increasing in order to reduce the time spent and increased costs.

Machine learning algorithms and software have been developed and used at all stages of drug discovery and development, including clinical trials. ML provides various facilities to identify new targets, provide stronger evidence for target-disease relationships, improve small molecule compound design and optimization, increase understanding of disease mechanisms, increase understanding of disease and non-disease phenotypes, and develop new biomarkers for prognosis (13). ML, a branch of artificial intelligence, is based on the idea of learning data from systems, identifying patterns and making decisions with minimal human intervention. There are several different ML methods implemented within the framework of artificial intelligence. For example, drug candidates can be optimized in drug discovery through a combination of models that predict favorable physicochemical properties (eg, solubility and permeability), pharmacokinetic properties, safety, and likely efficacy (14). ML techniques are very powerful tools in helping several steps of the drug discovery process, reducing the time spent in the laboratory, the use of consumables and chemical materials, and the maximum time anticipated for the discovery of a drug by conventional methods (15).

Although there are a wide variety of chemotherapeutics in current use, scientific studies continue rapidly to produce new agents. In these studies, it is aimed to develop new agents that are more effective and have fewer side effects than the agents currently in use. However, successful results cannot be obtained from the vast majority of new agents whose efficacy is estimated and *in vitro* cytotoxicity tests are performed after synthesis. In this direction, in our study, it was aimed to select new anticancer compounds that can be effective on breast cancer by using large data sets with machine learning method. In this way, it is planned to avoid cost, time loss and ethical concerns arising from unsuccessful experiments by conducting trials with compounds with cytotoxic potential in future *in vitro* and *in vivo* studies.

2. Materials and Methods

In this study, Health Sciences University, Experimental Medicine Application and Research Center Organic Synthesis laboratory facilities were used.

2.1. Data set

Compounds showing anti-cancer activity were obtained for the analyses, with the support of the literature. The collected data was uploaded to the OCHEM database, an easy-to-use and web-based platform designed to store experimental results and biological activity results of compounds that form the basis of *in silico* modeling. This dataset (98 compounds) prepared to develop the models consisted of carbazones, coumarins, quinolones, azoles, barbiturates and different chemical series. Concentrations of compounds with various IC50 values were applied to the data set in micromolar (Fig. S1). The dataset was used to develop both classification and regression models. To develop the regression models, the concentration of selected compounds against anti-cancer inhibition was determined in μM. After uploading the dataset to the website (www.ochem.eu), an independent test set was created by randomly selecting approximately 20%-25% of the compounds to obtain validation. Molecular formulas and anti-cancer activity results of all compounds generated in the training and validation test sets are available and accessible online.

2.2. Machine learning

It is the products that have the basic features that have the basic features that are basic in a data set consisting of those who have this collection prepared for a data set consisting of those in the literature. EU). ML method available in OCHEM is used to create in-silico models based on different sets (ASNNN, XGBOOST, WEKA-RF).

2.2.1. Validation of Models

Fivefold cross-validation and external reading set are used to validate models.

2.2.2. Statistical Parameters

Classification models (SN) and specificity (SP) will be calculated as follows;

SN=TP/(TP+FN) and SP=TN/(TN+FP)

In this case, it can be accurately calculated, verified as accurately as possible, accurately calculated, which is absolutely not true.

2.2.3. Molecular Identification

The following molecular identifiers from the OCHEM database are used.

-E-state indices: Electro-topological state indices are 2D descriptors that combine both electronic and topological properties of the compounds analyzed.

-ALogPS: The program calculates the 1-octanol/water distribution coefficient and solubility in water.

-ChemAxon descriptors: ChemAxon supports the calculation of six descriptive groups from 0D to 3D: fundamental analysis, charge, geometry, partitioning, protonation states.

-ADRIANA.Code: The software uses a set of methods for the creation of 3D-structures, calculation of physicochemical descriptors and molecular properties based on experimental models.

2.3. Classification methods

The data set (98 compounds) was randomly divided into two groups: the training (80) and the test set (18). The models were

developed after unidentified filtering of molecular descriptors. In addition, unsupervised forward selection (UFS) was used in models 1, 2 and 4 (Table 1) for better filtering of molecular descriptors. The RF algorithm used random subsets of descriptors as it was less subject to the problem of correlation between molecular descriptors. Therefore, the best WEKA-RF model was obtained without the use of UFS.

Initially, all descriptor sets available on the OCHEM website were scanned and the latest models with the highest predictive accuracy were calculated using four different descriptor sets. Improved models were summarized in Table 1 and Fig. S2. All models showed approximate results in terms of sensitivity, specificity and mean accuracy (BA), calculated by averaging all other methods with the training set consensus model.

Table 1. Statistical coefficients calculated for classification models obtained from the data set

		Sensitivity (%)		Specificity (%)		Average Accuracy (%)	
N	Methods	Practice[a]	Test[a]	Practice[a]	Test[a]	Practice	Test
1	ASNN	86.1	76.4	62.5	100.0	81.0 ± 6.0	90.0 ± 10.0
2	*k*NN	91.6	78.6	50.0	50.0	59.0 ± 7.0	70.0 ± 10.0
3	XGBOOST	86.5	76.4	53.8	100.0	75.0 ± 7.0	90.0 ± 10.0
4	WEKA-RF	89.7	81.1	66.7	100.0	70.0 ± 7.0	80.0 ± 10.0
5	Consensus	89.8	86.6	72.7	100.0	75.0 ± 3.0	89.0 ± 5.00

2.3.1. Regression models

The data set used in classification is also used in this model. Three methods for regression model analysis, ASNN, *k*NN and XGBOOST, have been developed to give the highest performance. 3 different molecular descriptors Adriana, ALOGPS and E-state were used on OCHEM to ensure better performance of the above methods used in the calculation of regression models. The results obtained (R^2, q^2) are summarized in Table 2.

Table 2. Statistical coefficients of regression models

		Training Set		Test Set	
N	Methods	R^2	q^2	R^2	q^2
1	ASNN	0.09 ± 0.07	0.05 ± 0.06	0.6 ± 0.2	0.3 ± 0.1
2	*k*NN	0.02 ± 0.03	0.00 ± 0.00	0.3 ± 0.2	0.06 ± 0.04
3	XGBOOST	0.09 ± 0.01	0.04 ± 0.1	0.5 ± 0.3	0.4 ± 0.2
4	Consensus	0.08 ± 0.09	0.06 ± 0.09	0.5 ± 0.2	0.3 ± 0.1

2.3.2. Activity estimation of new compounds

A virtual database of 7 Schiff base derivatives with different substituents at the R1, R2 and R3 positions was created (Fig. 1). Compounds identified using the consensus classification model were screened for being active or inactive against the MCF-7 cancer cell line, and all of these compounds were found to be active (Table 3)

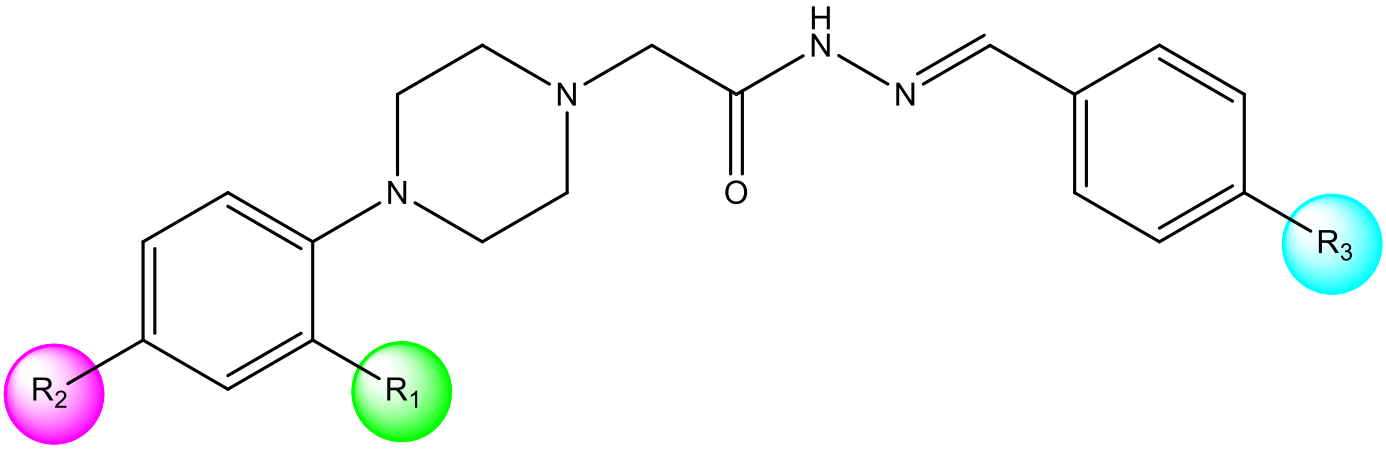
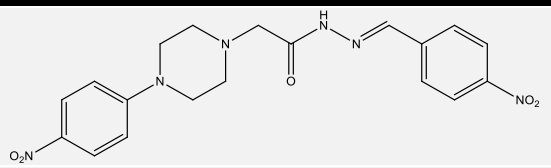
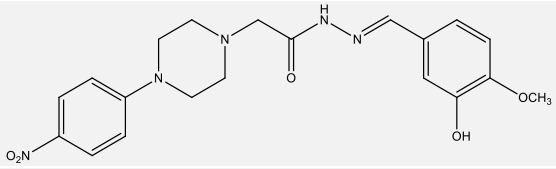
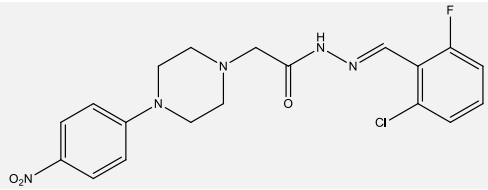
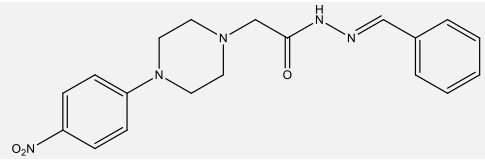
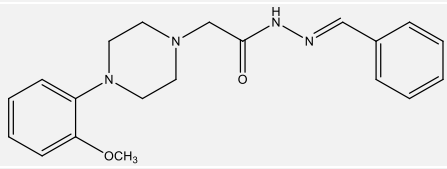
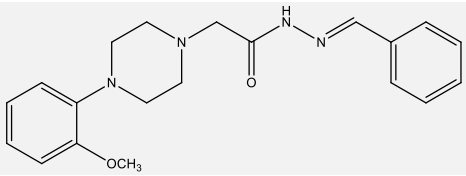
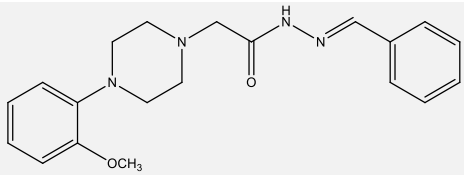


Fig. 1. 7 Virtual library of Schiff base derivative; R1, R2 and R3 were chosen based on the availability of starting materials for the synthesis

Table 3. Compounds tested against the MCF-7 cancer cell line and their estimated activity accuracies, the most active (marked in red), and their IC_{50} values (µM)

Compound	Active/Inactive	Predicted accuracy (%)	Predicted activity (IC_{50}, µM)
	active	88	29.50
	active	88	29.50
	active	96	38.00
	active	92	41.00
	active	88	39.00
	active	91	43.00
	active	87	28.00

2.4. Synthesis procedure

Compound 1 was selected among those with the highest activity results (compounds 1, 3 and 7), which were found to be active in our study, and its synthesis was carried out step by step as follows (Fig. 2).

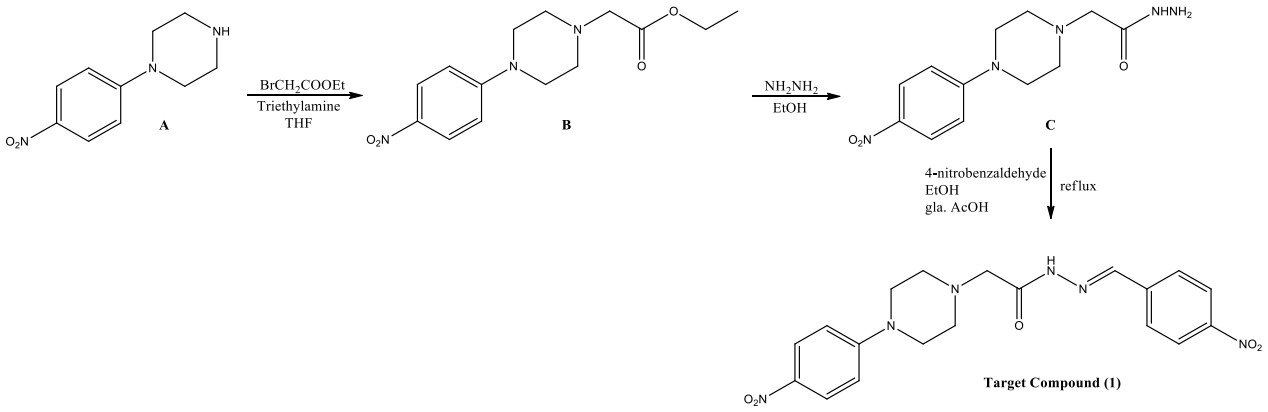


Fig. 2. Synthetic pathway for the synthesis of target compound (1)

Synthesis of Compound B (Ethyl 2-(4-(4-nitrophenyl)piperazin-1-yl)acetate) [16]

Ethyl bromoacetate (10 mmol) was added dropwise to the solution of 1-(4-nitrophenyl) piperazine (10 mmol) in tetrahydrofuran in a round-bottom flask, and the resulting mixture was stirred at room temperature in the presence of triethylamine (10 mmol) for 24 hours. After the salt formed was removed by filtration, the solvent was evaporated under reduced pressure and the solid obtained was crystallized from ethanol and purified.

Melting point: 115-116 ºC, Reaction yield: 99%. Yellow solid. FT-IR (υ_{max}, cm^{-1}): 3080 (ar-CH), 1742 (C=O). ^{1}H NMR (DMSO-d_6, δppm): 1.24 (t, 3H, *J*=8.0 Hz, CH_3), 2.69 (s, 4H, $2CH_2$), 3.34 (s, 2H, CH_2), 3.50 (s, 4H, $2CH_2$), 4.14 (q, 2H, *J*=8.0 Hz, CH_2), 7.06 (d, 2H, *J*=8.0 Hz, arH), 8.09 (d, 2H, *J*=8.0 Hz, arH). ^{13}C NMR (DMSO-d_6, δppm): 14.70, 46.90, 52.00, 58.64, 60.48, 113.20, 113.24, 126.28, 126.30, 137.44, 155.29, 170.41. LC-MS m/z: 316.12 ($[M+Na]^+$).

Synthesis of Compound C (2-(4-(4-nitrophenyl)piperazin-1-yl)acetohydrazide) [16]

Hydrazine hydrate (30 mmol) was added to the solution of compound no. 2 (10 mmol) in ethanol in a round-bottom flask, and the mixture was boiled under reflux for 5 hours. After completion of the reaction, the solvent was evaporated under reduced pressure to give a solid. The resulting solid was crystallized from ethanol and purified.

Melting point: 163-164 ºC, Reaction yield: 97%. Yellow solid. FT-IR (υ_{max}, cm^{-1}): 3328 and 3248 (NH_2 + NH), 3006 (ar-CH), 1631 (C=O). ^{1}H NMR (DMSO-d_6, δppm): 2.58-2.63 (m, 4H, $2CH_2$), 3.06 (s, 2H, CH_2), 3.55 (s, 4H, $2CH_2$), 4.32 (s, 2H, NH_2), 7.09 (s, 2H, arH), 8.11 (s, 2H, arH), 9.06 (s, 1H, NH). ^{13}C NMR (DMSO-d_6, δppm): 46.88, 52.89, 60.16, 113.22, 126.36, 137.39, 155.31, 168.65. LC-MS m/z: 302.11 ($[M+Na]^+$).

Synthesis of Compound 1 ((E)-N'-(4-nitrobenzylidene)-2-(4-(4-nitrophenyl)piperazin-1-yl) acetohydrazide)

4-nitrobenzaldehyde (10 mmol) and 3-4 drops of acetic acid were added to the solution of compound 3 (10 mmol) in ethanol in a round-bottom flask, and the mixture was refluxed for 6 hours. After completion of the reaction, the solid formed was filtered and dried. Purification was carried out by crystallization from ethyl acetate.

Melting point: 236-237 ºC, Reaction yield: 91%. Yellow solid. FT-IR (υ_{max}, cm^{-1}): 3380 (NH), 1684 (C=O), 1584 (C=N). ^{1}H NMR (DMSO-d_6, δppm): 2.59 (t, 2H, *J*=4.0 Hz, CH_2), 2.74 (t, 2H, *J*=4.0 Hz, CH_2), 3.18 (s, 2H, CH_2), 3.46 (t, 2H, *J*=4.0 Hz, CH_2), 3.60 (t, 2H, *J*=4.0 Hz, CH_2), 6.85 (d, 2H, *J*=4.0 Hz, arH), 7.80 (d, 2H, *J*=8.0 Hz, arH), 7.87 (d, 2H, *J*=4.0 Hz, arH), 8.22 (d, 2H, *J*=4.0 Hz, arH), 8.50 (s, 1H, CH), 10.74 (s, 1H, NH). ^{13}C NMR (DMSO-d_6, δppm): 50.36, 51.54, 57.62, 112.41, 124.48, 126.40, 127.95, 138.88, 139.94, 148.02, 148.64, 158.23, 165.42. LC-MS m/z: 413.23 ($[M+1]^+$).

3. Results

3.1. Calculation of IC_{50} Value with Machine Learning Method

Cross-validated average accuracy results for all models (5-fold cross validation) for all models using the ML method range from 59% to 81%. With the same models, the mean accuracy was also calculated for the test sets, and significant values were found in the range of 70%-89%. Consensus model, which is calculated by averaging all models, has been chosen for the highest results efficiency.

Statistical calculations of the regression models were made by applying ASNN, *k*NN, XGBOOST and consensus models as the average of the three methods. R2 and q2 values in the training and test sets are shown in Table 2. Anti-cancer activity predictions of new compounds were made with the help of classification and regression models created by ML method. All 7 determined compounds were found to be active with the models created, and 3 compounds were determined to be synthesized with an estimated accuracy result (≥87) and estimated activity value (≤30 μM) by making an evaluation from these results (Table 3).

4. Discussion

Using the OCHEM web-based platform, in silico models based on different ML techniques and various molecular descriptors were created. The models created, cross validation methods and estimation of external test sets were made to ensure the validation of the model. With these methods, the activities of the new Schiff base derivatives designed based on the literature data against the MCF-7 cancer cell line were estimated and it was determined that the models created showed high stability, robustness and predictive power. All compounds engineered as a result of the ML models were found to be active against the MCF-7 cancer cell line. This result confirms that ML approaches facilitate a rational search for active molecules within budget and time constraints, which is especially important in academic settings. In addition, the developed models are available to researchers working in this field and can be used to predict the anti-cancer activity of new compounds. Compound 1 was synthesized by ML method among the 3 most active compounds on MCF-7 cells. Traditional treatment methods are now being replaced by personalized cancer treatment. The structure of the cells taken from the tumour tissue of the patient diagnosed with cancer can be examined under laboratory conditions, and the treatment of the patient can be directed. With this method, it is aimed to find the right treatment method without time, energy and financial loss. Personalized cancer treatment is a more effective method than traditional therapies (17). In parallel, in our study, the anticancer effect of compound 1, whose cytotoxic effect was determined by machine learning method, was also demonstrated in vitro. Therefore, it has been demonstrated that personalized treatments can also be performed using machine learning method. One of the main problems with the

development of cancer therapy is the low reproducibility of results observed in animal models and patients. It has been reported that the correlation of data obtained from animal models with human tissue is less than 10% (18). More physiological human models are needed to reduce this attrition rate, improve preclinical screening, and reduce animal use. Monolayer and three-dimensional cell culture methods, such as cancer spheroids, are emerging as an important tool for high-throughput screening (19, 20). There is a large gap between in vitro two-dimensional cell culture and in vivo. 3D cell culture models provide more realistic spatial, biochemical and cellular parameters compared to 2D models and can bridge this gap. It has been shown that 3D cell culture models are pioneers in elucidating molecular and cellular mechanisms, facilitating the development and screening of new drugs because they reflect intercellular interactions more realistically (21-23). It is planned to switch to in vivo studies after the compounds synthesized by showing the potential for cytotoxic effects on MCF-7 will be used in 2D and 3D in vitro studies in the future.

Conflict of interest

The authors have no conflict of interest.

Funding

No funding was used for the study.

Acknowledgments

None to declare.

Authors' contributions

Concept: S.K., A.M., M.B., Design: S.K., A.M., M.B Data Collection or Processing: S.K., A.M., M.B., M.K., Analysis or Interpretation: S.K., A.M., M.B., M.K., S.O., Literature Search: S.K., A.M., M.B., M.K., S.O., Writing: S.K., A.M., M.B.

References

1. Kolak A, Kamińska M, Sygit K, Budny A, Surdyka D, Kukiełka-Budny B, et al. Primary and secondary prevention of breast cancer. Ann Agric Environ Med. 2017;24(4):549-53.
2. Anastasiadi Z, Lianos GD, Ignatiadou E, Harissis HV, Mitsis M. Breast cancer in young women: an overview. Updates Surg. 2017 Sep;69(3):313-317.
3. Enger SM, Ross RK, Paganini-Hill A, Bernstein L. Breastfeeding experience and breast cancer risk among postmenopausal women. Cancer Epidemiol Biomarkers Prev. 1998 May;7(5):365-9.
4. Colditz GA, Willett WC, Hunter DJ, Stampfer MJ, Manson JE, Hennekens CH, et al. Family history, age, and risk of breast cancer: prospective data from the Nurses' Health Study. Jama. 1993;270(3):338-43.
5. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet. 2001 Oct 27;358(9291):1389-99.
6. Pharoah PD, Day NE, Duffy S, Easton DF, Ponder BA. Family history and the risk of breast cancer: a systematic review and meta-analysis. Int J Cancer. 1997 May 29;71(5):800-9.
7. Lu Z-R, Steinmetz NF, Zhu H. New Directions for Drug Delivery in Cancer Therapy. ACS Publications; 2018.
8. Chabner BA, Roberts TG. Chemotherapy and the war on cancer. Nature Reviews Cancer. 2005;5(1):65-72.
9. Fuertes MA, Alonso C, Pérez JM. Biochemical modulation of cisplatin mechanisms of action: enhancement of antitumor activity and circumvention of drug resistance. Chemical reviews. 2003;103(3):645-62.
10. Breslin S, O'Driscoll L. Three-dimensional cell culture: the missing link in drug discovery. Drug discovery today. 2013;18(5-6):240-9.
11. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. 2018;68(6):394-424.
12. Chen R, Liu X, Jin S, Lin J, Liu J. Machine learning for drug-target interaction prediction. Molecules. 2018;23(9):2208.
13. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery. 2019;18(6):463-77.
14. Talevi A, Morales JF, Hather G, Podichetty JT, Kim S, Bloomingdale PC, et al. Machine Learning in Drug Discovery and Development Part 1: A Primer. CPT: pharmacometrics & systems pharmacology. 2020;9(3):129-42.
15. Lima AN, Philot EA, Trossini GHG, Scott LPB, Maltarollo VG, Honorio KM. Use of machine learning approaches for novel drug discovery. Expert opinion on drug discovery. 2016;11(3):225-39.
16. Mentese M, Demirbas N, Mermer A, Demirci S, Demirbas A, Ayaz FA. Novel Azole-Functionalited Flouroquinolone Hybrids: Design, Conventional and Microwave Irradiated Synthesis, Evaluation as Antibacterial and Antioxidant Agents. Letters in Drug Design and Discovery. 2018;15:46-64.
17. Jackson SE, Chester JD. Personalised cancer medicine. Int J Cancer. 2015 Jul 15;137(2):262-6.
18. Mak IW, Evaniew N, Ghert M. Lost in translation: animal models and clinical trials in cancer treatment. Am J Transl Res. 2014 Jan 15;6(2):114-8.
19. Anastasov N, Höfig I, Radulović V, Ströbel S, Salomon M, Lichtenberg J, et al. A 3D-microtissue-based phenotypic screening of radiation resistant tumor cells with synchronized chemotherapeutic treatment. BMC Cancer. 2015 Jun 10;15:466.
20. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006 Oct;6(10):813-23.
21. Kim MJ, Chi BH, Yoo JJ, Ju YM, Whang YM, Chang IH. Structure establishment of three-dimensional (3D) cell culture printing model for bladder cancer. PLoS One. 2019 Oct 22;14(10):e0223689.
22. Knight E, Przyborski S. Advances in 3D cell culture technologies enabling tissue-like structures to be created in vitro. J Anat. 2015 Dec;227(6):746-56.
23. Sokolova V, Rojas-Sánchez L, Białas N, Schulze N, Epple M. Calcium phosphate nanoparticle-mediated transfection in 2D and 3D mono- and co-culture cell models. Acta Biomater. 2019 Jan 15;84:391-401.

Supplementary materials

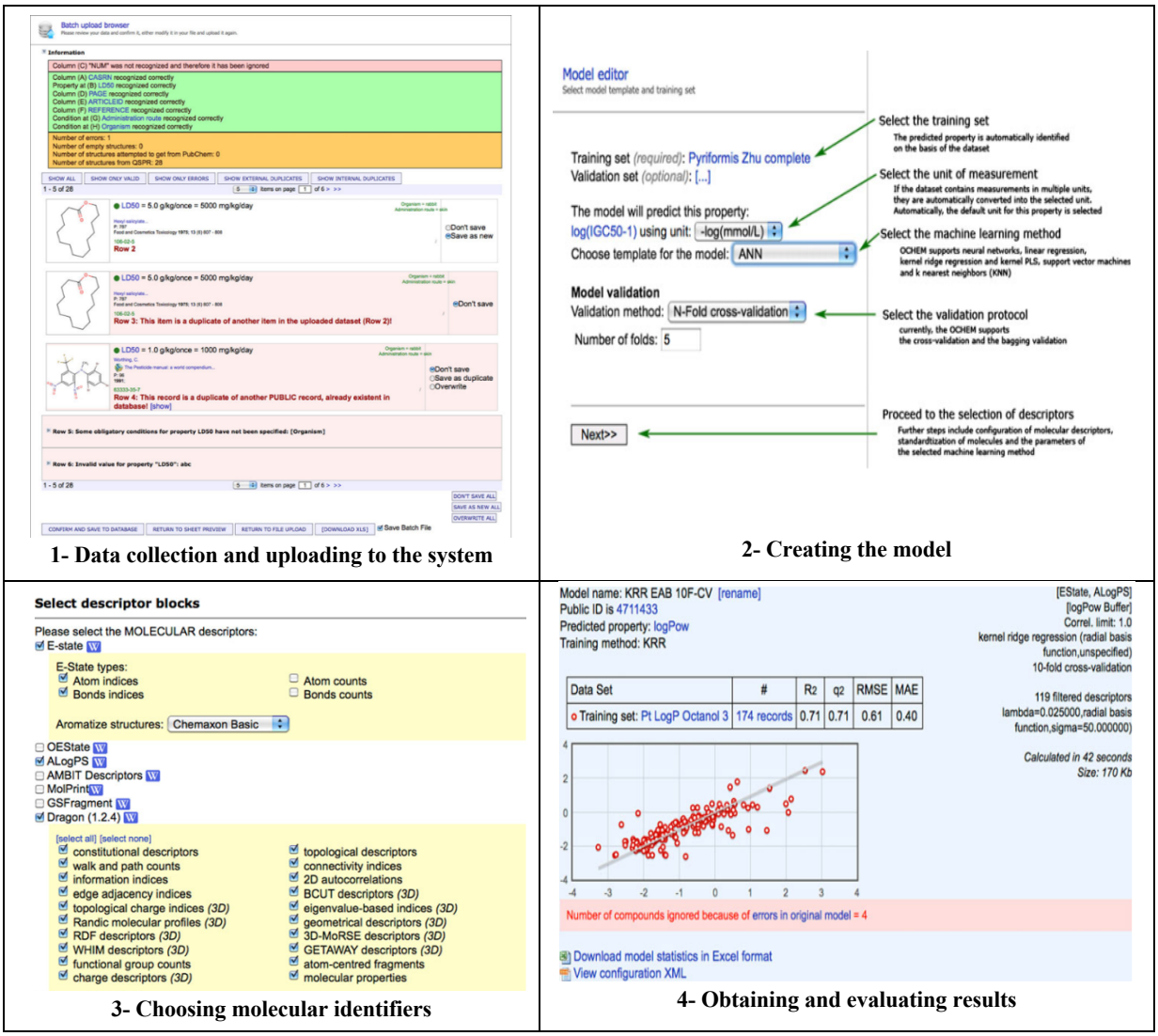
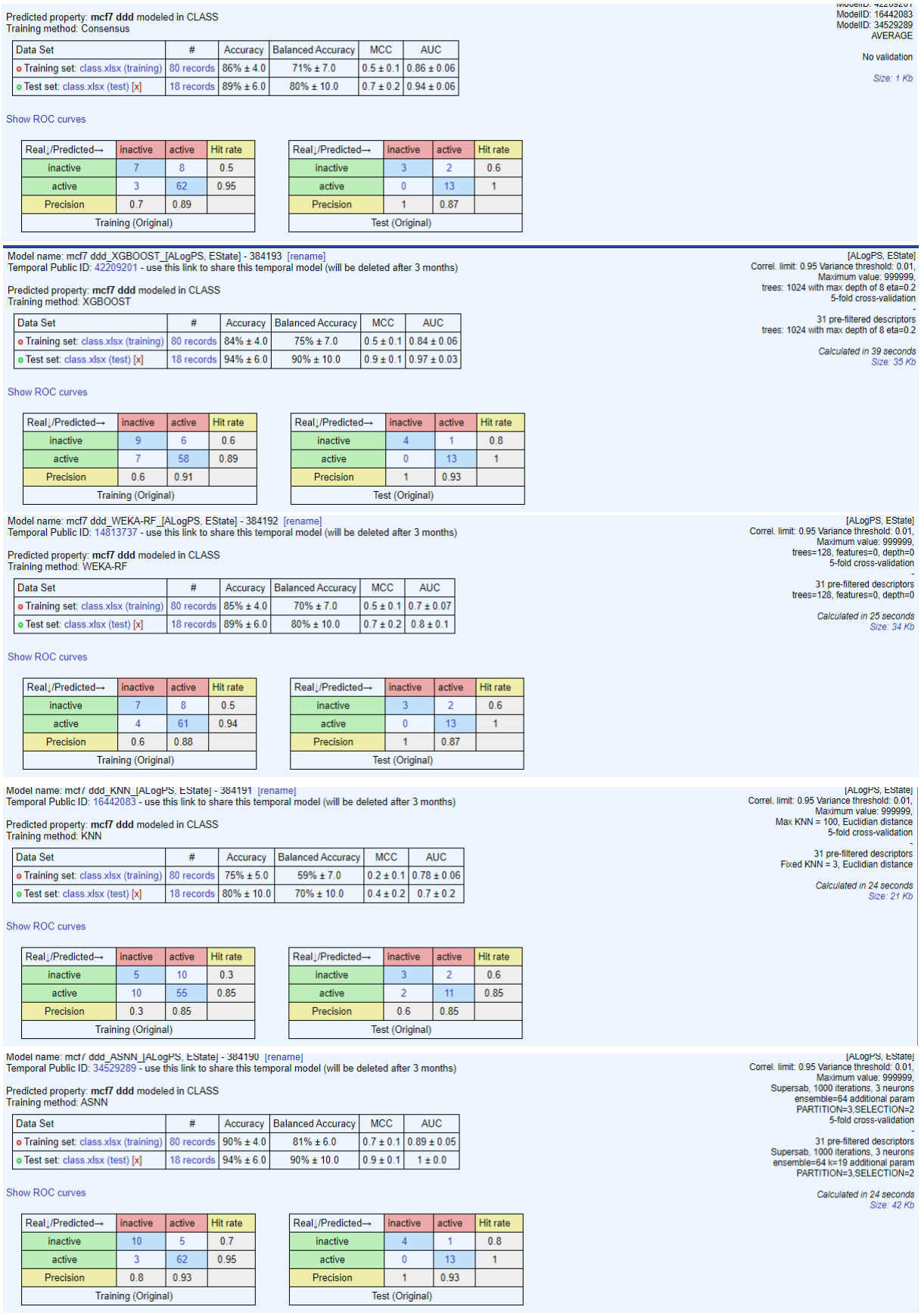


Fig. S1. Application of machine learning method



Data Set	#	Accuracy	Balanced Accuracy	MCC	AUC
Training set: class.xlsx (training)	80 records	86% ± 4.0	71% ± 7.0	0.5 ± 0.1	0.86 ± 0.06
Test set: class.xlsx (test) [x]	18 records	89% ± 6.0	80% ± 10.0	0.7 ± 0.2	0.94 ± 0.06

Real↓/Predicted→	inactive	active	Hit rate
inactive	7	8	0.5
active	3	62	0.95
Precision	0.7	0.89	

Training (Original)

Real↓/Predicted→	inactive	active	Hit rate
inactive	3	2	0.6
active	0	13	1
Precision	1	0.87	

Test (Original)

Data Set	#	Accuracy	Balanced Accuracy	MCC	AUC
Training set: class.xlsx (training)	80 records	84% ± 4.0	75% ± 7.0	0.5 ± 0.1	0.84 ± 0.06
Test set: class.xlsx (test) [x]	18 records	94% ± 6.0	90% ± 10.0	0.9 ± 0.1	0.97 ± 0.03

Real↓/Predicted→	inactive	active	Hit rate
inactive	9	6	0.6
active	7	58	0.89
Precision	0.6	0.91	

Training (Original)

Real↓/Predicted→	inactive	active	Hit rate
inactive	4	1	0.8
active	0	13	1
Precision	1	0.93	

Test (Original)

Data Set	#	Accuracy	Balanced Accuracy	MCC	AUC
Training set: class.xlsx (training)	80 records	85% ± 4.0	70% ± 7.0	0.5 ± 0.1	0.7 ± 0.07
Test set: class.xlsx (test) [x]	18 records	89% ± 6.0	80% ± 10.0	0.7 ± 0.2	0.8 ± 0.1

Real↓/Predicted→	inactive	active	Hit rate
inactive	7	8	0.5
active	4	61	0.94
Precision	0.6	0.88	

Training (Original)

Real↓/Predicted→	inactive	active	Hit rate
inactive	3	2	0.6
active	0	13	1
Precision	1	0.87	

Test (Original)

Data Set	#	Accuracy	Balanced Accuracy	MCC	AUC
Training set: class.xlsx (training)	80 records	75% ± 5.0	59% ± 7.0	0.2 ± 0.1	0.78 ± 0.06
Test set: class.xlsx (test) [x]	18 records	80% ± 10.0	70% ± 10.0	0.4 ± 0.2	0.7 ± 0.2

Real↓/Predicted→	inactive	active	Hit rate
inactive	5	10	0.3
active	10	55	0.85
Precision	0.3	0.85	

Training (Original)

Real↓/Predicted→	inactive	active	Hit rate
inactive	3	2	0.6
active	2	11	0.85
Precision	0.6	0.85	

Test (Original)

Data Set	#	Accuracy	Balanced Accuracy	MCC	AUC
Training set: class.xlsx (training)	80 records	90% ± 4.0	81% ± 6.0	0.7 ± 0.1	0.89 ± 0.05
Test set: class.xlsx (test) [x]	18 records	94% ± 6.0	90% ± 10.0	0.9 ± 0.1	1 ± 0.0

Real↓/Predicted→	inactive	active	Hit rate
inactive	10	5	0.7
active	3	62	0.95
Precision	0.8	0.93	

Training (Original)

Real↓/Predicted→	inactive	active	Hit rate
inactive	4	1	0.8
active	0	13	1
Precision	1	0.93	

Test (Original)

Fig. S2. Classification method results