



OPEN

# Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery

Lun K. Tsou<sup>1,2</sup>, Shiu-Hwa Yeh<sup>1,2</sup>, Shau-Hua Ueng<sup>1,2</sup>, Chun-Ping Chang<sup>1,2</sup>, Jen-Shin Song<sup>1</sup>, Mine-Hsine Wu<sup>1</sup>, Hsiao-Fu Chang<sup>1</sup>, Sheng-Ren Chen<sup>1</sup>, Chuan Shih<sup>1</sup>, Chiung-Tong Chen<sup>1</sup> & Yi-Yu Ke<sup>1</sup>✉

Machine learning is a well-known approach for virtual screening. Recently, deep learning, a machine learning algorithm in artificial neural networks, has been applied to the advancement of precision medicine and drug discovery. In this study, we performed comparative studies between deep neural networks (DNN) and other ligand-based virtual screening (LBVS) methods to demonstrate that DNN and random forest (RF) were superior in hit prediction efficiency. By using DNN, several triple-negative breast cancer (TNBC) inhibitors were identified as potent hits from a screening of an in-house database of 165,000 compounds. In broadening the application of this method, we harnessed the predictive properties of trained model in the discovery of G protein-coupled receptor (GPCR) agonist, by which computational structure-based design of molecules could be greatly hindered by lack of structural information. Notably, a potent (~500 nM) mu-opioid receptor (MOR) agonist was identified as a hit from a small-size training set of 63 compounds. Our results show that DNN could be an efficient module in hit prediction and provide experimental evidence that machine learning could identify potent hits in silico from a limited training set.

## Abbreviations

DNN	Deep neural networks
LBVS	Ligand-based virtual screening
RF	Random forest
TNBC	Triple-negative breast cancer
GPCR	G-protein-coupled receptors
AI	Artificial intelligence
QSAR	Quantitative structure–activity relationship
SVMs	Support vector machine
ADME	Absorption, distribution, metabolism, and excretion
DT	Decision tree
K-NN	K-nearest neighbors
ANNs	Artificial neural networks
VS	Virtual screening
MOR	Mu-opioid receptor
SBVS	Structure-based virtual screening
PLS	Partial least squares
MLR	Multiple linear regression
ECFPs	Extended connectivity fingerprints
FCFPs	Functional-class fingerprints
FRU	Relative fluorescence units

<sup>1</sup>Institute of Biotechnology and Pharmaceutical Research, National Health Research Institutes, Zhunan 35053, Miaoli County, Taiwan, ROC. <sup>2</sup>These authors contributed equally: Lun K. Tsou, Shiu-Hwa Yeh, Chun-Ping Chang and Shau-Hua Ueng. ✉email: yiyuke@nhri.edu.tw

Implementation of “big data” with deep learning has created a paradigm shift in many scientific disciplines<sup>1–3</sup>. From the perspective of medicinal chemistry, predicting particular functions or properties, e.g., absorption, distribution, metabolism, and excretion (ADME), of a molecular entity might greatly increase the quality of hit compounds and quicken the drug-discovery process. The use of artificial intelligence (AI) in drug design to generate a prediction model, conduct virtual screening, and predict compounds’ activities has received much attention recently<sup>4–7</sup>. Traditionally, quantitative structure–activity relationship (QSAR) model was utilized by medicinal chemists and statisticians to associate bioactivities to particular functional group manipulations. In particular, a linear equation was generated to correlate the features and bioactivities for each compound, while different descriptors were employed to calculate the physical properties to merge with the 3D-structural information and generate 2D or 3D-QSAR models<sup>8–10</sup>. Nowadays the development of QSAR have apply to multi-target and multi-objective QSAR approaches to assist drug design<sup>11–13</sup>. These QSAR approaches are able to integrate multiple diverse chemical and biological data, being therefore capable of jointly making predictions ranging from in vitro and in vivo activities to ADMET properties<sup>14</sup>. Nonetheless, these QSAR models were hard to generate from random and diverse databases. In addition, to properly separate the training set and the test set was time consuming. To provide an alternative strategy, as reported by Zhavoronkov et al., they have successfully used the deep learning method in the designs of more potent compounds<sup>15</sup>. The incorporation of machine learning method for the progressive analysis of the active compounds and concurrent generation of the prediction model should address such limitations.

Lavecchia et al.<sup>16</sup> summarized applications of machine learning algorithms, such as support vector machine (SVM)<sup>17</sup> for ADME evaluation and decision tree (DT) in the classification of compounds<sup>18</sup>. Moreover, a Naïve Bayesian classifier is frequently used in chemoinformatics for predicting biological properties, while *k*-Nearest neighbors (*K*-NN) is a simple and rough method to predict and rank the molecule<sup>19,20</sup>. Others like the artificial neural networks (ANNs), is the popular technique for compound classification, QSAR studies, and primary virtual screening (VS) of compounds<sup>21</sup>. All these machine learning algorithms were programmed to pick out and reclassify important features of the molecules as instructed, the limitations of these algorithms stemmed from the intrinsic inability to “self-taught” and prioritize the features in relation to the activities. Improper combining of the compounds’ descriptors could increase the noise level in features learning that could result in swamping the classifier model and generate a misleading prediction<sup>22</sup>.

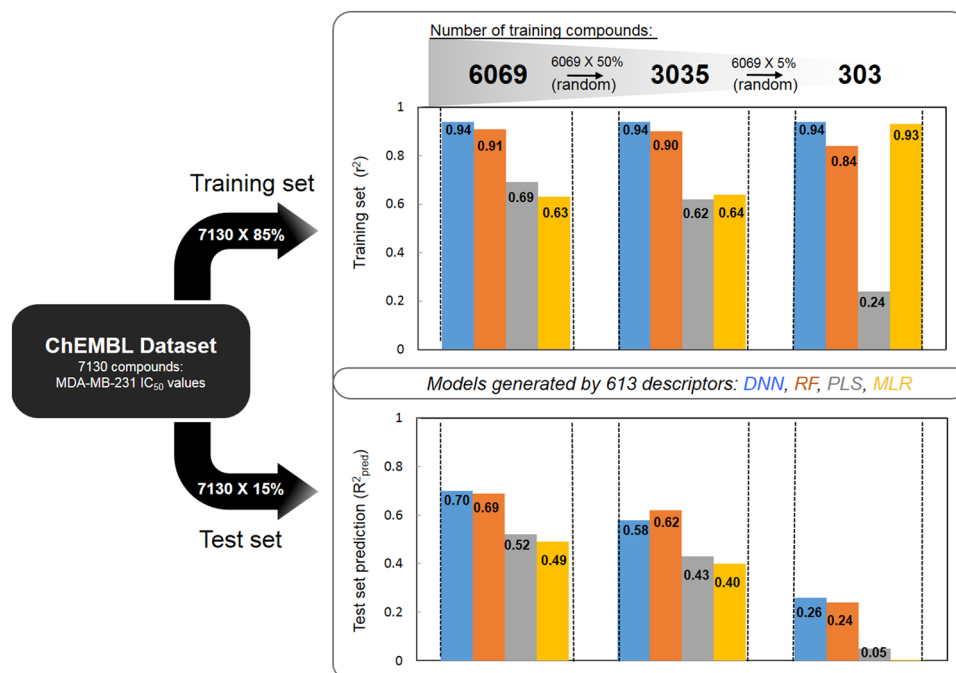
Herein, we employed deep learning algorithm to analyze the compound features, generate a first-hand model through 613 descriptors for training, and validated its findings through experimental confirmation. In addition, we compared its accuracy and efficiency with three other different virtual screening methods. After in silico screening of our in-house database of 165,000 compounds, by which different hit compounds were identified from<sup>15,23–25</sup>, 100 top-ranked newly identified TNBC inhibitors were subjected to the bioassay to cross-examine the model accuracy. Moreover, to extend the scope of this deep learning model in predicting meaningful hits, another case study for the search of novel G protein-coupled receptor (GPCR) agonist was carried out. By using a similar model, we only trained the model with a collection of 63 mu-opioid receptor (MOR) agonists to learn the importance of compound features for the given bioactivities. We then identified the nanomolar MOR agonist from the in-house compounds library. Our study suggested that deep learning method could generate potent hit compounds in different disease areas for the drug discovery process.

## Results and discussion

**Model generation and comparative studies in efficiency.** An advancement in the virtual screening method was made to reduce the burden of the drug discovery/development processes in a cost-effective manner<sup>26</sup>. The virtual screening can be devised by using either structure-based virtual screening (SBVS) like docking screening methods<sup>27</sup> or LBVS like QSAR model screening<sup>28</sup>. To investigate the application and efficiency of the DNN approach in medicinal chemistry, we compared other contemporary QSAR method, such as RF approach<sup>29</sup>, with traditional QSAR methods, such as PLS and MLR. RF has been demonstrated to have high prediction accuracy and robustness with adjustable parameters. It has become a “gold standard” machine learning method. Meanwhile, partial least squares (PLS) and multiple linear regression (MLR) are methods used for large data manipulation and allow facile generation of the model unlike other 3D-QSAR methods. In the current study, the same data set and descriptors were systematically incorporated to generate the models.

The traditional QSAR model helps to identify the relationship between activities and the descriptors’ variables. In addition to the QSAR methods, RF and DNN from the machine learning approach were used to generate the prediction model. RF is an ensemble learning method to perform classification in a similar manner to that of the decision tree (DT). Yet, the major difference stems from the use of Bagging method (or Bootstrap Aggregating) to generate many individual trees<sup>30</sup>. Each tree could self-process samples from the training set data and provide a fixed number of random sampling data from the training set to generate a DT for voting. The final model was based on the highest score from individually developed trees in the forest. On the other hand, DNN are mathematical methods developed to mimic the neurons (nodes) of the human brain to recognize objects and analyze progressively, improving the efficiency of previously reported neural network algorithms<sup>1,31</sup>. Each neuron is treated as a particular feature to classify the complex factors. The system, in turn, learns from the training set and assigns different weights for each neuron as this model eventually facilitates a prediction following the different clusters. Taken together, DNN increase the hidden layer numbers by allowing each layer of the nodes to access different features based on the previous layer’s output. Consequently, as more executed nodes are added in each layer, more features are recognized, enhancing the overall decision process.

To compare the different methods of virtual screening, a database of 7130 molecules with previously reported MDA-MB-231 inhibitory activities were collected from the ChEMBL web service. As the model prediction accuracy is highly depended on the quality of the database. In this study, these compounds were then randomly



**Figure 1.** Comparative studies of classification methods. Models generated through 613 descriptors were trained and tested using the ChEMBL dataset of 7130 compounds that exhibited MDA-MB-231 IC<sub>50</sub> values. The training and test sets' prediction efficiencies between different models, DNN, RF, PLS, and MLR were compared with decreasing number of training compounds.

separated into 6069 compounds (the training set) and 1061 compounds (the test set) to evaluate which model can more efficiently analyze the database and generate more useful models (Fig. 1). We implemented the extended connectivity fingerprints (ECFPs), which are circular topological depictions of the molecules, as the major molecular descriptors. Specifically, ECFPs are generated in a molecule-directed manner by systematically recording the neighborhood of each non-hydrogen atom into multiple circular layers up to a given diameter of that molecule<sup>32</sup>. These atom-centered sub-structural features are then mapped into integer codes and the resulting identifiers shape the extended connectivity fingerprint. These identifiers capture the local information of the corresponding atom in such a way that various atom properties (e.g., atomic number, connection counts) are packed into a single integer value. The default identifier configuration of ECFP captures highly specific atomic information, enabling the representation of a large set of precisely defined structural features.

In some applications, however, different kinds of abstraction may be desirable. For example, a chlorine or a bromine substituent on a ring may be functionally equivalent but would be redundantly distinguished by ECFP. Alternatively, functional-class fingerprints (FCFPs)<sup>32</sup> detail circular fingerprints via the pharmacophore identification of atoms, which reports topological pharmacophore fingerprints. To perform the classifications comparisons, the software devised a total of 613 descriptors from AlogP\_count<sup>33</sup>, ECFP, and FCFP to generate the model (Fig. 1, and supplementary data Table S1).

Three distinct different numbers of training set (6069, 3035, and 303 compounds) were used to generate the models and their efficiencies were evaluated by the fixed test set (1061 compounds). R-square value ( $r^2$  value) was used to quantify the differential efficiencies between the training set and test set prediction in machine learning methods (DNN and RF) and the QSAR methods (PLS and MLR) (Fig. 1). With training set compounds fixed at 6069, the machine learning methods (DNN or RF) exhibited higher predicted  $r^2$  value near 90% than the traditional QSAR method (PLS or MLR) at 65%. In general, a good model was considered as having larger  $r^2$  and  $R^2_{pred}$  ( $r^2 > 80$ ,  $R^2_{pred} > 60$  is an assessable model)<sup>34–36</sup>. With the decrease of training set numbers, the machine learning methods sustained the overall higher  $r^2$  value. As the training set number decreases, the deviation only retained with DNN and RF at 0.84 to 0.94, while PLS and MLR dropped to 0.24 from 0.69. In particular, with significantly lower training set numbers, interestingly, the MLR method maintained a respectful  $r^2$  value near 0.93, but when running against the test set,  $R^2_{pred}$  was calculated to be zero. This implies that MLR could be an over-fitting model with a high false-positive rate, especially when the numbers of learning compounds are very limited. These results showed that the PLS and MLR methods could not efficiently distinguish the descriptors and were problematic in generating meaningful fitting equations. On the other hand, the DNN method with lower number of training sets, the data still held a higher  $r^2$  value of 0.94 than that of 0.84 by RF method (Fig. 1). Although the RF method could classify the features and select intrinsic feature for the analysis, DNN method was better in providing insights in weighting of important features. As a result, the DNN method held a higher  $r^2$  value with lower numbers of training data sets. Of the machine learning methods, the  $R^2_{pred}$  significantly improved with the increase in training set numbers, which is vastly different than the QSAR models

(Fig. 1). With routine sampling of large amount of molecular features against a target from the public domain might be limiting, the large spread or deviation of PLS and MLR processes could greatly hinder the potential of identifying potent hits. Taken together, DNN and RF exhibited better accuracy and efficiency in the prediction of hit compounds. As shown in Fig. 1, the  $R^2_{\text{pred}}$  of DNN (0.26) and RF (0.24) were much lower, which implies that the database quality might not be sufficient for learning. We envision that more datasets might be needed or the quality of the datasets in terms of structural information and their activities should be more correlated for better learning by the algorithm.

Seminal work by Grisoni and coworkers<sup>37,38</sup>, have indicated the  $R^2_{\text{pred}}$  or  $Q^2$  metrics (Eq. 1) should be optimized to  $Q^2_{F3}$  (Eq. 2) as it was more sensitive for comparing predicted abilities between different models with the same training set. The original  $R^2_{\text{pred}}$  metrics was shown below

$$Q^2 = 1 - \frac{\sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{\text{test}}} (y_i - \bar{y}_{\text{TR}})^2} \quad (1)$$

where  $y_i$  is the experimental result for  $i$ -th compounds not existing in the training set,  $\hat{y}_i$  is the predicted result of the  $i$ -th compound,  $\bar{y}_{\text{TR}}$  is the average value of the training set experimental results, and  $n_{\text{test}}$  is the test set numbers. Reported by Todeschini et al., the  $Q^2_{F3}$  should be calculated as

$$Q^2_{F3} = 1 - \frac{\sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2 / n_{\text{test}}}{\sum_{j=1}^{n_{\text{TR}}} (y_j - \bar{y}_{\text{TR}})^2 / n_{\text{TR}}} \quad (2)$$

By which,  $y_j$  is the experimental result for training set,  $\bar{y}_{\text{TR}}$  is the average value of the training set experimental result, and  $n_{\text{TR}}$  is the training set numbers. By applying this metric to our studies, the DNN and RF exhibited highest  $Q^2_{F3}$  value of 0.679 and 0.670, respectively (Supplementary data Table S2). In addition, Consonni et al. showed the calculation of Root-Mean-Square Error in prediction (RMSEP) and Root-Mean-Square Error in calculation (RMSEC) could quantify predictive abilities of QSAR model. The higher value of RMSEP led to higher chances of error. Our calculation results also showed that DNN method had the lowest value for RMSEC and RMSEP in comparison to those of other models (Supplementary data Table S2).

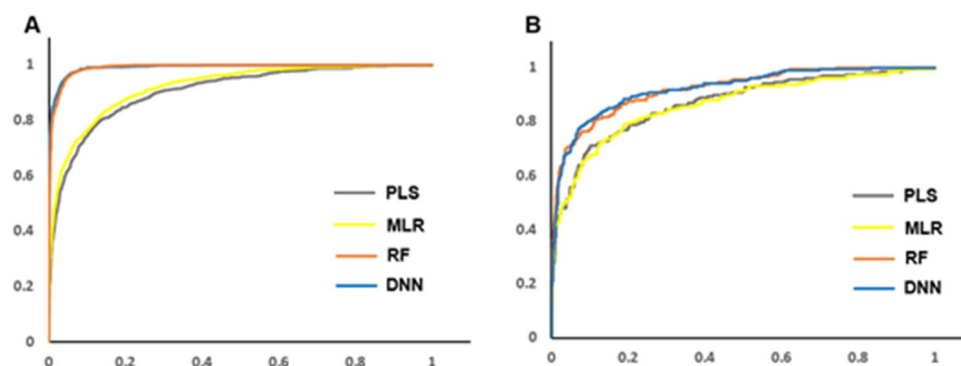
To further investigate the advantageous prediction ability of machine learning methods (DNN and RF) over the traditional QSAR methods (PLS and MLR), we analyzed the receiver operating characteristic (ROC) curve with the fix training set (6069 compounds) and fix test set (1061 compounds)<sup>39,40</sup>. ROC curve evaluates the performance of a binary classifier system and provides means in selecting optimal models. ROC curve was constructed by plotting a graph of sensitivity (Se, true positive rate) vs. 1-specificity (1-Sp, false positive rate). The measure of Se and Sp are defined as

$$\text{Se} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

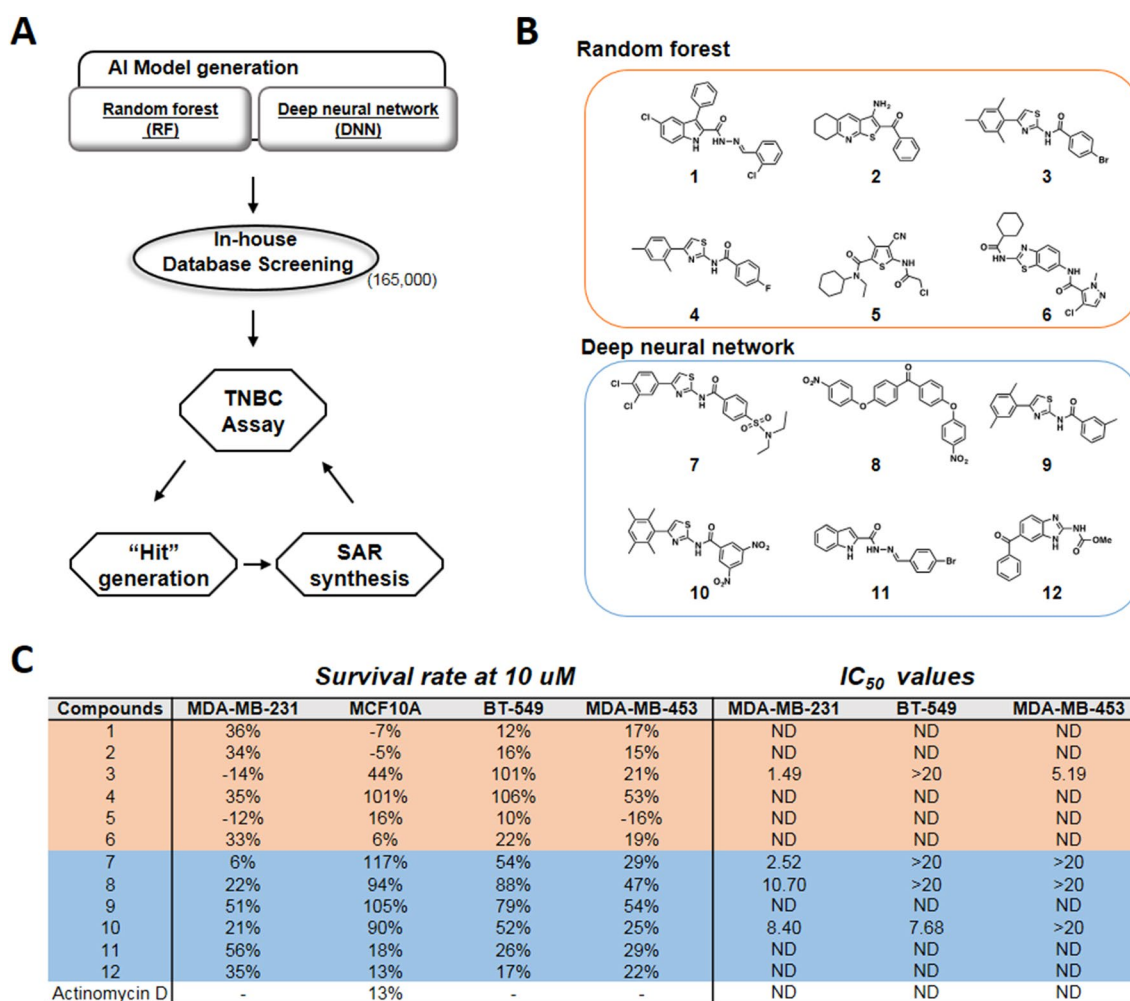
$$\text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4)$$

where TP is the number of correctly identified active ligands (true positives), TN is the number of correctly identified inactive ligands (true negatives), FP the number of incorrectly identified active ligands (false positives), and FN the number of incorrectly identified inactive ligands (false negatives). The area under the ROC curve (AUC) measures the performance of each virtual screening approaches. The ideal screening method results in an AUC value of 1, while a random screening method would lead to an AUC value of 0.5. As shown in Fig. 2A, the AUC calculated by the training set of the RF and DNN methods were 0.991 and 0.992, respectively. Interestingly, these values were higher than those of PLS and MLR methods with 0.907 and 0.922. To investigate the prediction ability of the test set, the respective AUC values of RF and DNN methods were 0.922 and 0.924. Also, they were expected to be superior than those of PLS and MLR methods with 0.870 and 0.865. These ROC curve analyses further potentiated the RF and DNN screening method might be more suitable than traditional QSAR methods (PLS and MLR).

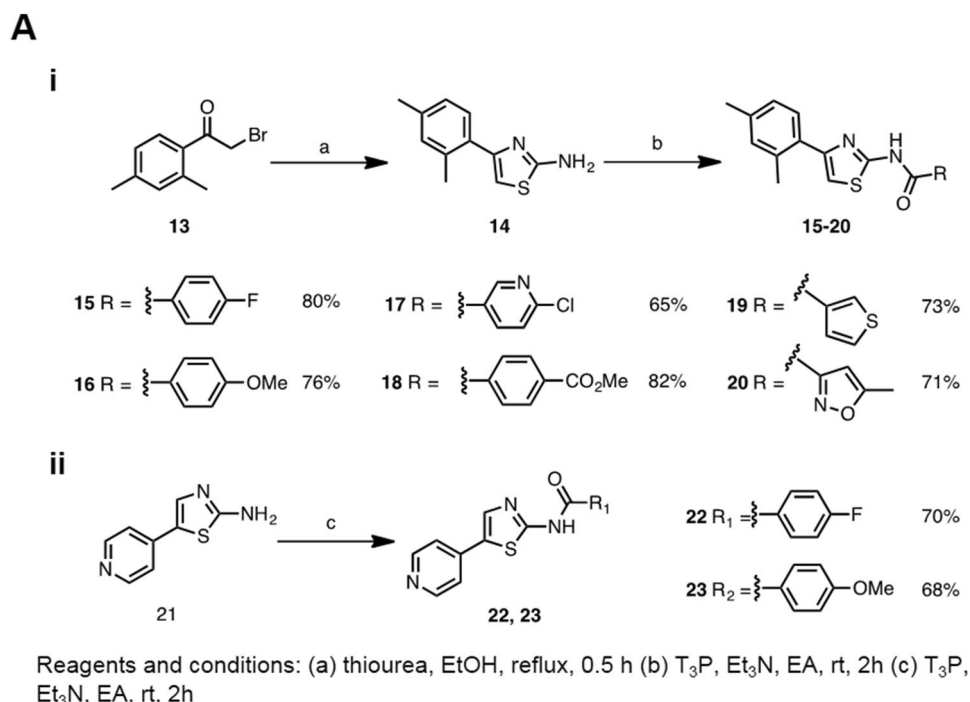
**Virtual screening and identification of TNBC inhibitors by DNN and RF models with experimental validation.** Based on the above information, the DNN and RF models were chosen as the preferred means to perform virtual screening. The identified compounds were then assayed for their corresponding bioactivities. Herein, we demonstrated two different cases for evaluating these models' accuracy. First, we successfully identified active hits for TNBC inhibition. The DNN and RF models were used to screen the in-house database (165,000 compounds), and the selected hits were assayed against the anti-TNBC cellular assay (Fig. 3A). The top predicted 100 compounds were selected and tested at 10  $\mu\text{M}$  concentration for MDA-MB-231 cell line inhibition (Supplementary data Table S3.1 and Table S3.2). Since the compound collection was acquired based on MDA-MB-231 inhibitory activities, other TNBC cell lines were also assayed to obtain selective TNBC inhibitors. Out of the multiple hits identified through both methods, six compounds from each classification (compounds 1–12) were assayed and showed low cytotoxicity to MCF10A, a nonmalignant mammary epithelial cell line (Fig. 3B,C). We then assayed these hits against other TNBC cell lines, BT-549 and MDA-MB-453. Compounds 3, 7, 8, 10, which exhibited broader TNBC inhibitions, were then subjected to  $\text{IC}_{50}$  determination (Supplementary data Figure S1). Notably, between RF and DNN, we obtained a thiazole core with selective anti-TNBC profiles over



**Figure 2.** Comparative studies of ROC calculation for DNN, RF, PLS, and MLR methods. Comparisons of ROC curves for (A) Using the fix number of training set (6069 compounds) to generate the model for the analysis of the training set itself. The AUC value of each ROC curve for PLS, MLR, RD, and DNN are 0.907, 0.922, 0.991, 0.922, respectively. (B) Using the fix number of training set (6069 compounds) to generate model for analysis of the test set. The AUC value of each ROC curve for PLS, MLR, RD, and DNN are 0.924, 0.922, 0.870, 0.865, respectively.



**Figure 3.** Model generation and discovery of TNBC inhibitors from in-house library. (A) Flow scheme of discovery of potent TNBC inhibitors. (B) Chemical structures of in-house identified TNBC inhibitors: 6 hits from random forest classifications and 6 hits from deep neural network. (C) Cellular survival rate at 10  $\mu$ M of the 12 hits against nonmalignant mammary epithelial cell line (MCF10A) and three TNBC cell lines: MDA-MB-231, BT-549, and MDA-MB-453. Actinomycin D, was used as control. Values are expressed as the mean of at least two independent determinations and are within  $\pm 15\%$ .



**B**

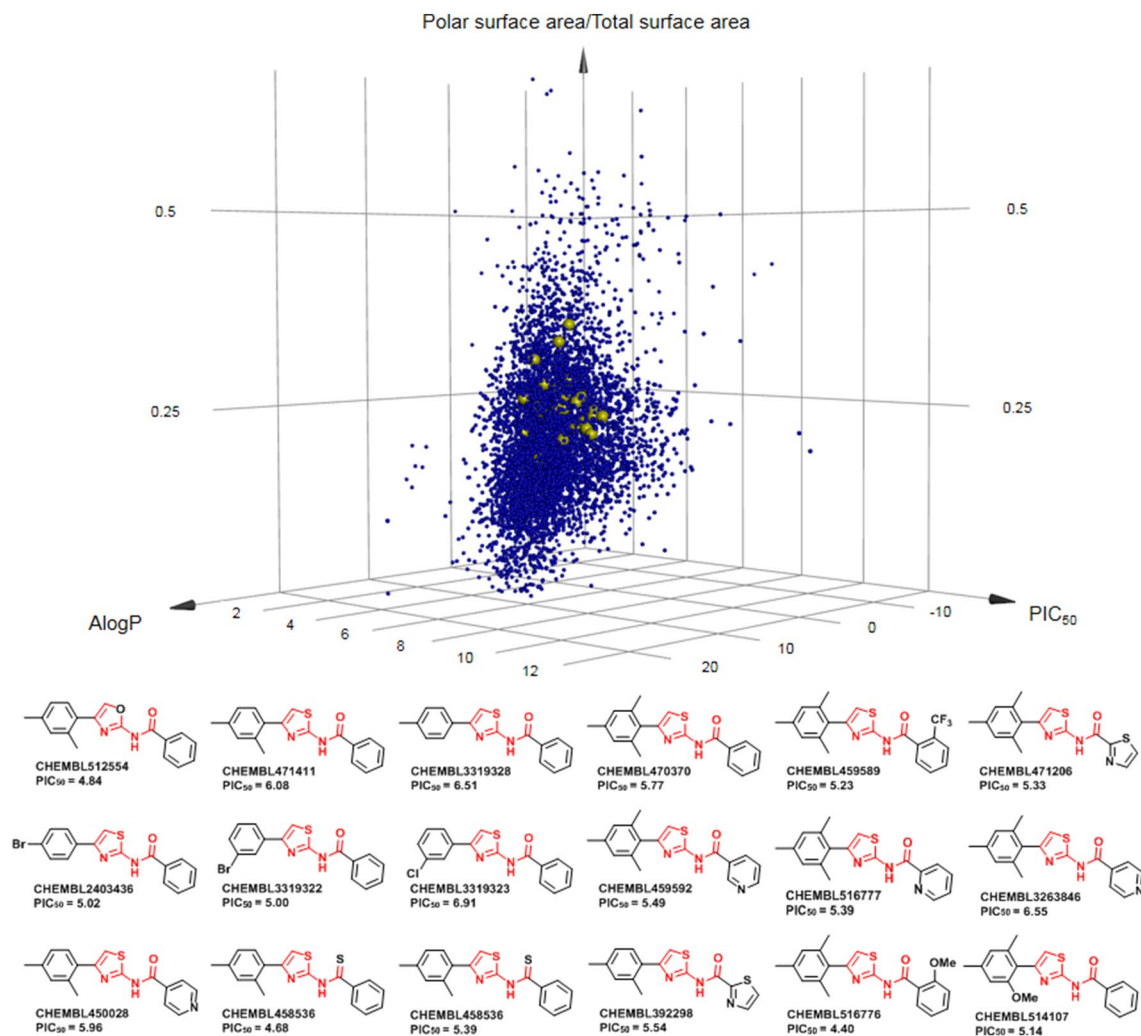
Compounds	MCF-10A	BT-549	MDA-MB-231	MDA-MB-453
	Inhibition rate	IC <sub>50</sub> ( $\mu$ M)	IC <sub>50</sub> ( $\mu$ M)	IC <sub>50</sub> ( $\mu$ M)
15	-8%	>10.00	1.89	9.28
16	8%	>10.00	1.4	6.3
17	2%	>10.00	3.41	>10.00
18	4%	>10.00	0.62	>10.00
19	14%	>10.00	2.68	>10.00
20	14%	>10.00	>10.00	>10.00
22	-5%	>10.00	>10.00	>10.00
23	35%	2.6	3.98	7.93

**Figure 4.** Structure–activity relationship studies of thiazole-based TNBC inhibitors. (A) Synthetic routes of thiazole (i) and its regioisomers (ii). (B) Cellular cytotoxicity of the inhibitors against nonmalignant mammary epithelial cell line (MCF10A) at 10  $\mu$ M and IC<sub>50</sub> values against three TNBC cell lines: MDA-MB-231, BT-549, and MDA-MB-453.

the normal mammary cells (Fig. 3B,C). Synthesis of the thiazole-based inhibitors were carried out and several potent TNBC inhibitors were identified (Fig. 4). Compound 18, which showed good selectivity over nonmalignant mammary epithelial cell, had an IC<sub>50</sub> of 0.62  $\mu$ M against MDA-MB-231. Interestingly, regioisomeric controls in compounds 22 and 23 were synthesized. Compound 22 did not show activities toward the TNBC and 23, although it possessed moderate micro molar activities and also exhibited cytotoxicity to MCF-10A. This study serves as a good example of hit generation from an unknown target with good cellular selectivity and functional manipulatable core.

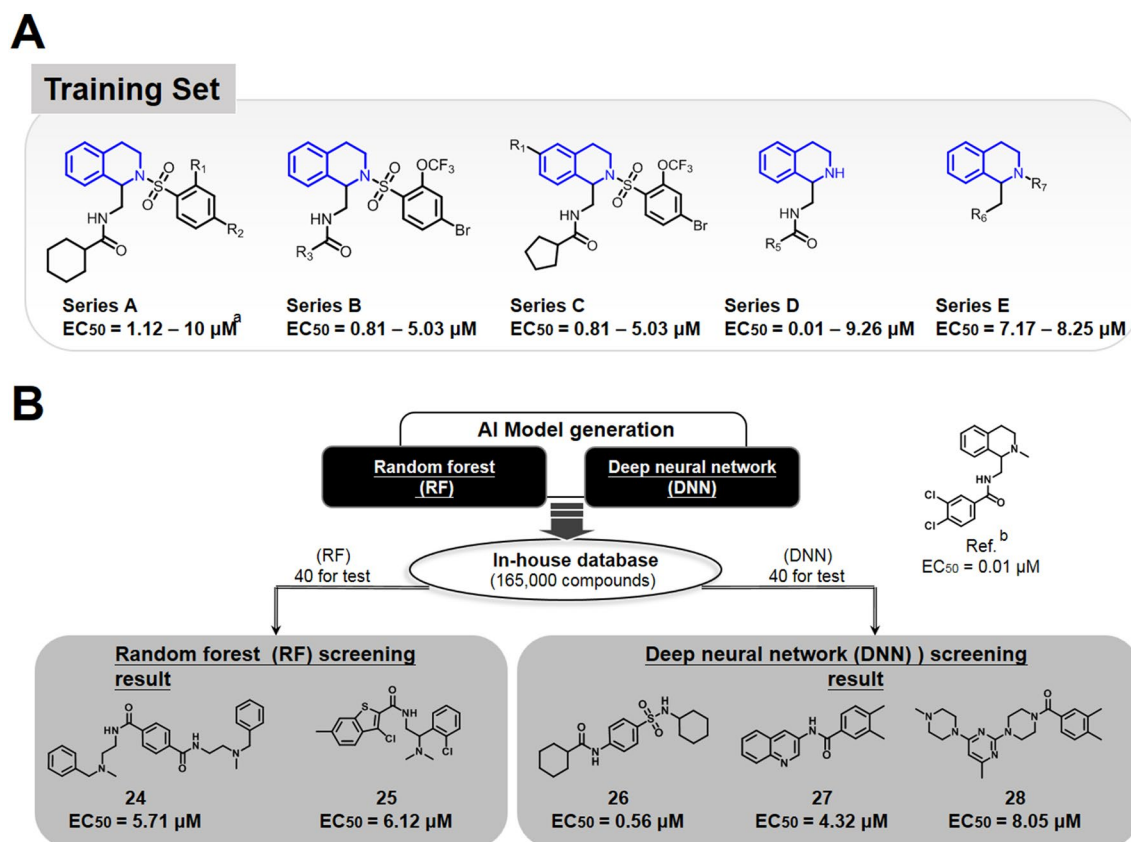
**Analysis between model-identified compounds and database compounds.** To address the availability of thiazole core in the original set of 7130 compounds, we devised a principle component analysis of the database with PIC50, AlogP, and polar surface versus total surface area (Fig. 5). These properties were chosen to fulfill the characteristics of a hit compound in a common drug-discovery campaign. The 7130 compounds were then mapped and showed that compounds consisting of the thiazole core are clustered in the quadrant with activities ranging from PIC50 4.8–6.5 (10  $\mu$ M to 0.3  $\mu$ M). Moreover, the AlogP and polar surface versus total surface area values were in the satisfactory range for a hit compound (Fig. 5). Gratifyingly, this finding correlates well to the experimental results from our SAR studies of the TNBC inhibitors (Fig. 4). Our findings suggest that both RF and DNN can be adapted to generate meaningful models and identify functional hits for the later optimization process.

**Identification and experimental validation of novel GPCR agonists by the DNN and RF models.** We envision that predicting new scaffold with the experimental validation should render the greatly



**Figure 5.** Chemoinformatics of thiazole-based inhibitors in the ChEMBL dataset. Analysis of compound properties is characterized by AlogP, PIC<sub>50</sub>, and percent polar surface area of the molecules to address solubility, potency, and cellular properties. The thiazole compounds were clustered in the center of the matrix.

expand the application of this deep learning approach. We then adapted this classification for GPCR agonist generation, where structure-based designs are limited without a known information of the core structure due to the membrane associated nature of many GPCRs (Fig. 6). To evaluate the scope of the model, the MOR agonist was also identified via virtual screening of the same in-house database with the DNN and RF models. In our previous studies on MOR agonist<sup>41</sup>, we synthesized 63 compounds and tested by FLIPR calcium assay (Supplementary data Table S4 and Figure S2). We used MOR as an example to demonstrate the predictive power of this approach. To train the learning system, we provided a small sample collection of 63 compounds<sup>41</sup>. The total 63 compounds, divided into series A–E clusters, were used as the training set to generate the DNN and RF models (Fig. 6A). We envision that incorporation of molecular diversity with large spread of bioactivities in series A–E should minimize deviation of the  $r^2$  with DNN and RF and improve the learning process. Model generation was performed with the same 613 descriptors, and then new cores in the 165,000 in-house pool were processed. The top 40 compounds predicted by RF and another top 40 by DNN (Supplementary data Table S5.1 and Table S5.2) were subjected to the FLIPR calcium assay (Fig. 6B). The CHO-K1 cell line, stably expressing MOR and Ga15 (GenScript), was used to evaluate the selected compounds. In the FLIPR calcium assay of CHO-K1/MOR/Ga15 cells, activation of MOR elicits an intracellular calcium release, leading to an increase in the relative fluorescence units (RFU). Five compounds, **24–28**, were identified as potential hits by these two different screening models. As shown in Fig. 6B, in addition to hit **26** identified from DNN method exhibited potent agonist activities (EC<sub>50</sub> = 560 nM), these models provided great molecular diversities over the training set of compounds. To the best of our knowledge, this is the first example correlating prediction and validation of a GPCR agonist discovery where structure-based design is limited. Notably, only a small training set of 63 compounds (Supplementary data Table S4) was employed, and a set of five structurally distinct hits was identified. This result provided strong support in that DNN and RF methods could still sustained high predicted  $r^2$  value in low numbers of training data set.



**Figure 6.** Prediction and validation of novel MOR agonist from RF and DNN classifications. (A) Molecular descriptions of training set compounds of MOR agonist, Series A–E, and their corresponding  $EC_{50}$  values. (B) Flow scheme of model generation and novel hits identified from RF and DNN prediction. By FLIPR calcium assay,  $EC_{50}$  values are the means of at least three independent experiments. Reference compound was published by Chen et al. and assigned as compound 46 in the publication.

	[ <sup>3</sup> H]diprenorphine binding, $K_i$ (nM)
Morphine	$6 \pm 1.0$
ME	$3.5 \pm 0.8$
24	$830 \pm 50.0$
25	$625 \pm 79.0$
26	$535 \pm 50.0$
27	$720 \pm 95.0$
28	$965 \pm 110.0$

**Table 1.** The binding affinity assay of compounds 24, 25, 26, 27, 28 and morphine on MOR. [Met<sup>5</sup>] Enkephalin (ME) is an opioid pentapeptide.

The Opioid receptor binding affinity assay was performed to further confirm these compounds direct bind to MOR. The MOR membranes was detected by measuring the competitive inhibition ratio of [<sup>3</sup>H]diprenorphine binding assessment.  $K_i = IC_{50}/(1 + L/K_d)$ , where L is the concentration of [<sup>3</sup>H]diprenorphine (1 nM) used, and the  $K_d$  value in MOR is 0.4 nM. All assays were carried out independently and at least in triplicate. The values indicate the mean  $\pm$  SEM. MOR =  $\mu$  opioid receptor; ME = [Met<sup>5</sup>]Enkephalin; N.D. = not determined; SEM = standard error of the mean. As shown in Fig. 6, the compounds 24–28 has no structural similarity to morphine or any other previously described opioid receptor agonist. In the receptor binding assay, membrane proteins from HEK-MOP were used for detecting the binding affinity of these compounds by comparing with the morphine (Table 1).



## Conclusion

Hit identification is an important step in the early stages of drug discovery. Virtual screening is extensively used to identify suitable hits, and such methods to improve the hit rate are much sought after. In this study, we report comparative studies between traditional QSAR methods and machine learning methods applied in VS. The results showed that machine learning methods could achieve a higher predicted  $r^2$  value with fewer compounds required in the training set. In our work, DNN and RF predicted the selective TNBC inhibitors from the our in-house database. In case of identifying novel MOR agonist, 5 hit compounds were readily found from only a 63-compound training set. The diversified chemical structures of the 5 hits identified by the DNN method showed good potency as a hit with an  $EC_{50} = 560$  nM. This is an interesting application of the deep learning classification as structure-based design of GPCR agonist are limited with limited information of the core structure due to the membrane associated nature of many GPCR. Taken together, this study demonstrated the efficiency of DNN and RF machine learning methods for VS and provided experimental evidences that this application can be adapted to identify hit compounds among different diseases.

## Experimental procedures

**Data set collection for TNBC and MOR.** For the TNBC inhibitor identification studies, 7130 compounds that contain MDA-MB-231 bioassay activity data were collected from the ChEMBL database (<https://www.ebi.ac.uk/chembl/>). The database was randomly separated into two parts. One part contained 85% of the compounds (6069 compounds), which were used as the training set; the other 15% of compounds (1061 compounds) were used as a test set in our studies. However, for the MOR agonist discovery studies, 63 compounds were collected from the publication of Chen et al.<sup>41</sup> as a training set database (Supplementary data Table S4).

**Descriptors and model generation.** All models were generated by BIOVIA pipeline pilot V18.1 platform with R statistic software V 3.4.1<sup>42,43</sup>. These models were generated by the same descriptors from the Discovery Studio/Calculates ligand properties program (BIOVIA, Inc., San Diego, CA), including ALogP\_count (101 descriptors), ECFP\_4 (256 descriptors), and FCFP\_4 (256 descriptors). The RF model use a recursive partitioning (decision tree) forest model by R package “randomForest”. The number of trees was set for 500. The fraction of descriptors to use for each tree in the forest was set to 0.3. A deep neural network model using R package “deepnet” performed the DNN model. Three hidden layers were used and each layer with 80 nodes. The learning rate of every epoch was 0.1 with the momentum for 0.9, the maximum number of iterations for network training was 5000. To prevent the model + over-fitting, the fraction of hidden layer to be dropped out for model training was set for 0.25. The traditional QSAR model, like multiple linear regression analysis (MLR), is a equation to describe the dependent variable Y with independent variables,  $X_1, X_2, \dots$ , etc. For example,  $Y(\text{pred})^i = b_0 + b_1 * X_1 + b_2 * X_2 + \dots + b_p * X_p$ , where the  $b_1, b_2, \dots, b_n$  are the regression coefficients,  $Y(\text{pred})^i$  can be project as ith bioactivities, and  $X_1, X_2, \dots, X_p$  can apply to different descriptors<sup>44</sup>. The PLS regression is using the orthogonal matrices (T) to determine the fundamental relations between dependent variable Y and independent variables X. For example,  $Y = X \times W \times Q + E$ ,  $T = X \times W$ , where Y is a response matrix for the dependent variables like bioactivities result, T is a extraction matrix for the independent variables like descriptors, Q is a matrix of the regression coefficients, W are the factor score matrix and the weight matrix, and E is an error term for the model<sup>45,46</sup>. The PLS and MLR models were also conduct by pilot V18.1 platform with the default protocol and evaluate by fivefold cross-validated method.

**Cell viability assay for TNBC inhibitors.** The cells were seeded in 384-well clear plates with a density of  $8 \times 10^2$  cells/well for MCF-10A and BT-549 cell lines,  $1 \times 10^3$  cells/well for MDA-MB-453, and  $2 \times 10^3$  cells/well for MDA-MB-231 overnight. Then cells were treated with the indicated concentrations of test compounds for 72 h. At the end of incubation, 5  $\mu$ L of PrestoBlue Cell Viability Reagent (Invitrogen, Carlsbad, CA, USA) was added to each well with 50  $\mu$ L medium. The plates were incubated for an additional 1.5 h at 37 °C in a humidified 5% CO<sub>2</sub> atmosphere; the relative fluorescence unit (RFU) in the reaction mixture will then be recorded (Ex560/Em590) by Victor<sup>2</sup>-Vplate reader (PerkinElmer, Waltham, MA, USA). The cell lines were chosen based on the mutation status of PTEN and/or TP53: MCF-10A, the nonmalignant mammary epithelial cell line; BT-549 with mutation of PTEN and TP53; MDA-MB-453 with mutation of PTEN; MDA-MB-231 with mutation of TP53<sup>47</sup>.

**FLIPR calcium assay.** Black with clear flat bottom 96-well assay plates (Corning) were coated with a 0.1 mg/mL Poly-L-Lysine solution a day prior to the assay. CHO-K1/MOR/Ga15 cells were suspended in the F12 medium and plated at a density of  $\sim 8 \times 10^4$  cells/well in 200  $\mu$ L medium. Cells were incubated in a humidified atmosphere of 10% CO<sub>2</sub> at 37 °C overnight to reach an 80–90% confluence cell monolayer before assay. On the day of assay, 150  $\mu$ L medium/well was removed from the plate. To each well, 50  $\mu$ L FLIPR calcium assay reagent dissolved in 1  $\times$  assay buffer (HBSS: KCl 5 mM, KH<sub>2</sub>PO<sub>4</sub> 0.3 mM, NaCl 138 mM, NaHCO<sub>3</sub> 4 mM, Na<sub>2</sub>HPO<sub>4</sub> 0.3 mM, D-glucose 5.6 mM, with an additional 20 mM HEPES and 13 mM CaCl<sub>2</sub>, pH 7.4), with 2.5 mM probenecid added; then the plate was incubated at 37 °C for 1 h. Compounds (30  $\mu$ M) and other reagents were dissolved in the assay buffer. Using a FlexStationIII (Molecular Devices Corp.), the increase of fluorescence after robotic injections of compounds or other reagents were monitored every 1.52 s interval with excitation wavelength at 485 nm and emission wavelength at 525 nm. The  $[Ca^{2+}]_i$  fluorescence was measured up to 90 s after agonist injection. The relative fluorescence intensity from 2 wells of cells were averaged and the relative amount of  $[Ca^{2+}]_i$  release was determined by integrating the area under the curve (AUC) with Prism software (GraphPad). The AUC of each compound was then subtracted from the response in the presence of MOR agonist naloxone (20 nM) to obtain the specific MOR responses<sup>48</sup>.

**Radioligand binding assay.** Human embryonic kidney 293 cells constitutively expressing MOR (HEK-MOR) (Dr. Ping-Yee Law; University of Minnesota Medical School) were harvested and homogenized in membrane preparation buffer (50 mM Tris-HCl at pH 7.4, containing 2 mM ethylenediaminetetraacetic acid [EDTA]) containing a fresh protease inhibitor cocktail (Roche, Basel, Switzerland) and then centrifuged at 30,000g for 30 min. The pellets were resuspended, aliquoted, and stored at  $-80^{\circ}\text{C}$ . For the [ $^3\text{H}$ ]diprenorphine saturation binding assays, membranes (containing 25  $\mu\text{g}$  of protein) were incubated with different concentrations (0.5–5 nM) of [ $^3\text{H}$ ]diprenorphine in binding buffer (50 mM Tris-HCl at pH 7.4, containing 2 mM EDTA) at  $25^{\circ}\text{C}$  for 1 h. For the competitive binding experiments, [ $^3\text{H}$ ]diprenorphine (1 nM) was incubated with membranes (containing 25  $\mu\text{g}$  of protein) in the absence or presence of various concentrations of compounds at  $25^{\circ}\text{C}$  for 1 h. The samples were then rapidly filtered onto glass-fiber filters (Millipore, Billerica, MA, USA) and washed three times with ice-cold phosphate-buffered saline. The radioactivity was quantified using a liquid scintillation counter<sup>49</sup>.

Received: 23 March 2020; Accepted: 18 September 2020

Published online: 08 October 2020

## References

- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444. <https://doi.org/10.1038/nature14539> (2015).
- Aliper, A. *et al.* Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* **13**, 2524–2530. <https://doi.org/10.1021/acs.molpharmaceut.6b00248> (2016).
- Jing, Y., Bian, Y., Hu, Z., Wang, L. & Xie, X. S. Deep learning for drug design: An artificial intelligence paradigm for drug discovery in the big data era. *AAPS J.* **20**, 58. <https://doi.org/10.1208/s12248-018-0210-0> (2018).
- Gawehn, E., Hiss, J. A. & Schneider, G. Deep learning in drug discovery. *Mol. Inform.* **35**, 3–14. <https://doi.org/10.1002/minf.201501008> (2016).
- Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* <https://doi.org/10.1126/sciadv.aap7885> (2018).
- Lavecchia, A. Deep learning in drug discovery: Opportunities, challenges and future prospects. *Drug Discov. Today* **24**, 2017–2032. <https://doi.org/10.1016/j.drudis.2019.07.006> (2019).
- Stahl, N., Falkman, G., Karlsson, A., Mathiason, G. & Boström, J. Deep Reinforcement learning for multiparameter optimization in de novo drug design. *J. Chem. Inf. Model.* **59**, 3166–3176. <https://doi.org/10.1021/acs.jcim.9b00325> (2019).
- Verma, J., Khedkar, V. M. & Coutinho, E. C. 3D-QSAR in drug design—A review. *Curr. Top. Med. Chem.* **10**, 95–115. <https://doi.org/10.2174/156802610790232260> (2010).
- Ke, Y. Y. *et al.* 3D-QSAR assisted drug design: Identification of a potent quinazoline based Aurora kinase inhibitor. *ChemMedChem* **8**(1), 136–148 (2013).
- James, N., Shanthi, V. & Ramanathan, K. Drug design for ALK-positive NSCLC: An integrated pharmacophore-based 3D QSAR and virtual screening strategy. *Appl. Biochem. Biotechnol.* **185**, 289–315. <https://doi.org/10.1007/s12010-017-2650-x> (2018).
- Ambure, P., Halder, A. K., Diaz, H. G. & Cordeiro, M. N. D. S. QSAR-Co: An open source software for developing robust multitasking or multitarget classification-based QSAR models. *J. Chem. Inf. Model.* **59**, 2538–2544 (2019).
- Cruz-Monteagudo, M., Borges, F. & Cordeiro, M. N. D. S. Desirability-based multiobjective optimization for global QSAR studies: Application to the design of novel NSAIDs with improved analgesic, antiinflammatory, and ulcerogenic profiles. *J. Comput. Chem.* **29**, 2445–2459 (2008).
- Cruz-Monteagudo, M. *et al.* Desirability-based methods of multiobjective optimization and ranking for global QSAR studies. Filtering safe and potent drug candidates from combinatorial libraries. *J. Comb. Chem.* **10**, 897–913 (2008).
- Nicolaou, C. A., Kannas, C. & Loizidou, E. Multi-objective optimization methods in de novo drug design. *Mini-Rev. Med. Chem.* **12**, 979–987 (2012).
- Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038. <https://doi.org/10.1038/s41587-019-0224-x> (2019).
- Lavecchia, A. Machine-learning approaches in drug discovery: Methods and applications. *Drug Discov Today* **20**, 318–331. <https://doi.org/10.1016/j.drudis.2014.10.012> (2015).
- Hou, T., Wang, J. & Li, Y. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *J. Chem. Inf. Model.* **47**, 2408–2415. <https://doi.org/10.1021/ci7002076> (2007).
- Klekota, J. & Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics* **24**, 2518–2525. <https://doi.org/10.1093/bioinformatics/btn479> (2008).
- Koutsoukas, A. *et al.* In silico target predictions: Defining a benchmarking data set and comparison of performance of the multiclass Naive Bayes and Parzen-Rosenblatt window. *J. Chem. Inf. Model.* **53**, 1957–1966. <https://doi.org/10.1021/ci300435j> (2013).
- Nigsch, F., Bender, A., Jenkins, J. L. & Mitchell, J. B. O. Ligand-target prediction using winnow and naive Bayesian algorithms and the implications of overall performance statistics. *J. Chem. Inf. Model.* **48**, 2313–2325. <https://doi.org/10.1021/ci800079x> (2008).
- Patel, J. L. & Goyal, R. K. Applications of artificial neural networks in medical science. *Curr. Clin. Pharmacol.* **2**, 217–226 (2007).
- Goodarzi, M., Dejaegher, B. & Vander Heyden, Y. Feature selection methods in QSAR studies. *J. AOAC Int.* **95**, 636–651 (2012).
- Wu, C. H. *et al.* Design and synthesis of tetrahydropyridothieno[2,3-d]pyrimidine scaffold based epidermal growth factor receptor (EGFR) kinase inhibitors: The role of side chain chirality and Michael acceptor group for maximal potency. *J. Med. Chem.* **53**, 7316–7326. <https://doi.org/10.1021/jm100607r> (2010).
- Yeh, J. Y. *et al.* Anti-influenza drug discovery: Structure–activity relationship and mechanistic insight into novel angelicin derivatives. *J. Med. Chem.* **53**, 1519–1533. <https://doi.org/10.1021/jm901570x> (2010).
- Ke, Y. Y. *et al.* Ligand efficiency based approach for efficient virtual screening of compound libraries. *Eur. J. Med. Chem.* **83**, 226–235. <https://doi.org/10.1016/j.ejmech.2014.06.029> (2014).
- Ripphausen, P., Nisius, B., Peltason, L. & Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **53**, 8461–8467. <https://doi.org/10.1021/jm101020z> (2010).
- Ripphausen, P., Stumpfe, D. & Bajorath, J. Analysis of structure-based virtual screening studies and characterization of identified active compounds. *Future Med. Chem.* **4**, 603–613. <https://doi.org/10.4155/fmc.12.18> (2012).
- Ripphausen, P., Nisius, B. & Bajorath, J. State-of-the-art in ligand-based virtual screening. *Drug Discov. Today* **16**, 372–376. <https://doi.org/10.1016/j.drudis.2011.02.011> (2011).
- Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. <https://doi.org/10.1023/A:1010933404324> (2001).
- Efron, B. 1977 Rietz Lecture—bootstrap methods—another look at the Jackknife. *Ann. Stat.* **7**, 1–26. <https://doi.org/10.1214/aos/1176344552> (1979).

31. Ma, J. S., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **55**, 263–274. <https://doi.org/10.1021/ci500747n> (2015).
32. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754. <https://doi.org/10.1021/ci100050t> (2010).
33. Ghose, A. K. & Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* **27**, 21–35. <https://doi.org/10.1021/ci00053a005> (1987).
34. Dearden, J. C., Cronin, M. T. D. & Kaiser, K. L. E. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **20**, 241–266. <https://doi.org/10.1080/10629360902949567> (2009).
35. Ke, Y. Y. & Lin, T. H. Modeling the ligand–receptor interaction for a series of inhibitors of the capsid protein of enterovirus 71 using several three-dimensional quantitative structure–activity relationship techniques. *J. Med. Chem.* **49**, 4517–4525. <https://doi.org/10.1021/jm0511886> (2006).
36. Cherkasov, A. *et al.* QSAR modeling: Where have you been? Where are you going to?. *J. Med. Chem.* **57**, 4977–5010. <https://doi.org/10.1021/jm4004285> (2014).
37. Todeschini, R., Ballabio, D. & Grisoni, F. Beware of unreliable Q(2)! A comparative study of regression metrics for predictivity assessment of QSAR models. *J. Chem. Inf. Model.* **56**, 1905–1913. <https://doi.org/10.1021/acs.jcim.6b00277> (2016).
38. Consonni, V., Todeschini, R., Ballabio, D. & Grisoni, F. On the misleading use of QF32 for QSAR model comparison. *Mol. Inform.* <https://doi.org/10.1002/Minf.201800029> (2019).
39. Truchon, J. F. & Bayly, C. I. Evaluating virtual screening methods: Good and bad metrics for the “early recognition” problem. *J. Chem. Inf. Model.* **47**, 488–508. <https://doi.org/10.1021/ci600426e> (2007).
40. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **16**, 412–424. <https://doi.org/10.1093/bioinformatics/16.5.412> (2000).
41. Chen, S. R. *et al.* Discovery, structure–activity relationship studies, and anti-nociceptive effects of *N*-(1,2,3,4-tetrahydro-1-isoquinolinylmethyl)benzamides as novel opioid receptor agonists. *Eur. J. Med. Chem.* **126**, 202–217. <https://doi.org/10.1016/j.ejmech.2016.09.003> (2017).
42. Gentleman, R., Hornik, K. & Leisch, F. R 1.5 and the Bioconductor 1.0 releases. *Comput. Stat. Data An.* **39**, 557–558 (2002).
43. Warr, W. A. Scientific workflow systems: Pipeline Pilot and KNIME. *J. Comput. Aid Mol. Des.* **26**, 801–804. <https://doi.org/10.1007/s10822-012-9577-7> (2012).
44. Wold, S. & Dunn, W. J. Multivariate quantitative structure activity relationships (QSAR)—conditions for their applicability. *J. Chem. Inf. Comput. Sci.* **23**, 6–13. <https://doi.org/10.1021/Ci00037a002> (1983).
45. Hellberg, S., Wold, S., Dunn, W. J., Gasteiger, J. & Hutchings, M. G. The anesthetic activity and toxicity of halogenated ethyl methyl ethers, a multivariate QSAR modeled by Pls. *Quant. Struct. Act. Rel.* **4**, 1–11. <https://doi.org/10.1002/qsar.19850040102> (1985).
46. Luco, J. M. & Ferretti, F. H. QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **37**, 392–401 (1997).
47. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Investig.* **121**, 2750–2767. <https://doi.org/10.1172/JCI45014> (2011).
48. Lin, S. Y. *et al.* The in vivo antinociceptive and mu-opioid receptor activating effects of the combination of *N*-phenyl-2',4'-dimethyl-4,5'-bi-1,3-thiazol-2-amines and naloxone. *Eur. J. Med. Chem.* **167**, 312–323. <https://doi.org/10.1016/j.ejmech.2019.01.063> (2019).
49. Chao, P. K. *et al.* 1-(2,4-dibromophenyl)-3,6,6-trimethyl-1,5,6,7-tetrahydro-4*H*-indazol-4-one a novel opioid receptor agonist with less accompanying gastrointestinal dysfunction than morphine. *Anesthesiology* **126**, 952–966. <https://doi.org/10.1097/Aln.0000000000001568> (2017).

## Acknowledgements

This work was supported by the grants of Intramural Research Program of the National Health Research Institutes (06A1-BPAP-01-033) and the Ministry of Economic Affairs, Taiwan R.O.C (106-EC-17-A-22-0624).

## Author contributions

L.T. wrote and organized the manuscript. S.H.Y. and H.F.C. carried out the FLIPR calcium assay to evaluate the MOR experiments. S.H.U. and S.R.C. help to synthesize the MOR agonist. J.S.S. and M.H.W. maintained the high-throughput screening core facility and aided in TNBC cellular assay. C.P.C. carried out the synthesis of the TNBC inhibitors. Y.Y.K. was the team leader, carried out the computational experiments, and analyzed the results. C.S. suggest some concepts and help to revise the manuscript. C.T.C. was the Chief Investigator and guided the development of the project. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-73681-1>.

**Correspondence** and requests for materials should be addressed to Y.-Y.K.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.