# scientific reports

## OPEN



# Developing muscarinic receptor M1 classification models utilizing transfer learning and generative AI techniques

Souvik Dey<sup>1,2</sup>, Anders Wallqvist<sup>1</sup><sup>1</sup> & Mohamed Diwan M. AbdulHameed<sup>1,2</sup>

Muscarinic receptor subtype 1 (M1) is a G protein-coupled receptor (GPCR) and a key pharmacological target for peripheral neuropathy, chronic obstructive pulmonary disease, nerve agent exposures, and cognitive disorders. Screening and identifying compounds with potential to interact with M1 will aid in rational drug design for these disorders. In this work, we developed machine learning-based M1 classification models utilizing publicly available bioactivity data. As inactive compounds are rarely reported in the literature, we encountered the problem of imbalanced datasets. We investigated two strategies to overcome this bottleneck: 1) transfer learning and 2) using generative models to oversample the inactive class. Our analysis shows that these approaches reduced misclassification of the inactive class not only for M1 but also for other GPCR targets. Overall, we have developed classification models for M1 receptor that will enable rapid screening of large chemical databases and advance drug discovery.

Muscarinic receptors are a family of G protein-coupled receptors (GPCRs) that become activated in response to the neurotransmitter acetylcholine<sup>1-3</sup>. They play important roles in a variety of biological activities, including bladder muscle contraction, salivary gland secretion, and heart rate regulation, as well as control multiple cognitive processes<sup>4-7</sup>. However, overstimulation of muscarinic receptors, for instance by exposure to nerve agents, induces a cholinergic response that is characterized by narrowing of the airways, seizures, and even coma, making their inhibition relevant for military as well as civilian populations<sup>8,9</sup>. Recent studies have also identified muscarinic receptor subtype 1 (M1) (Fig. 1a) as a critical target for peripheral neuropathy, chronic obstructive pulmonary disease, and cognitive disorders. Pharmacological blockade of M1 receptors using antagonists has been proven to be effective in nerve regeneration, thereby reversing diabetes-, chemotherapy-, and HIV-induced neuropathies<sup>10,11</sup>.

The widespread distribution and impact of muscarinic receptors have prompted a multitude of efforts to discover novel therapeutics for M1. For example, atropine was found to be a nonspecific antagonist of muscarinic receptors and is now included in the current treatment for nerve-agent exposures<sup>12–17</sup>. Furthermore, using high-throughput screening (HTS), Merck identified a scaffold, benzyl quinolone carboxylic acid (BQCA), that can act as a positive allosteric modulator of M1<sup>18</sup>. Similarly, using HTS, Weaver et al. screened 63,656 compounds and identified 714 actives, three scaffolds of which were M1 selective. They further optimized these scaffolds and found that VU0255035 is an effective M1-selective antagonist<sup>19</sup>. Based on compounds that are similar to pirenzepine, using a scaffold-hopping approach, Millard et al. identified multiple muscarinic receptor antagonists with potential central nervous system activity<sup>20</sup>.

More recently, advances in artificial intelligence (AI) have enabled significant breakthroughs in drug discovery and development. AI-based models complement experimental studies by enabling screening of ultralarge libraries in the early stages of the drug discovery process. They also enable the identification of patterns in complex biological data that traditional Quantitative Structure Activity Relationship (QSAR) methods might overlook<sup>21–25</sup>. This increasing reliance on computational approaches is driven by the increased availability of bioactivities in public databases, such as ChEMBL, BindingDB, and PubChem, coupled with the expansion of the chemical space by generative models, such as recurrent neural networks, autoencoders, and transformers<sup>26–35</sup>.

<sup>1</sup>Department of Defense Biotechnology High Performance Computing Software Applications Institute, Defense Health Agency Research and Development, Medical Research and Development Command, 504 Scott Street, Fort Detrick, MD 21702-5012, USA. <sup>2</sup>The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD, USA. <sup>Semail:</sup> sven.a.wallqvist.civ@health.mil; mabdulhameed@bhsai.org



Fig. 1. (a) Structure of muscarinic receptor subtype 1 (M1) (PDB ID: 5CXV). (b) Distribution of M1 bioactivity data collected from public sources, such as ChEMBL and BindingDB, considering a cutoff of 1  $\mu$ M for actives and 10  $\mu$ M for inactives.

The improvements in computing power and advanced models have aided the drug discovery process, with several AI-predicted candidates advancing to clinical trials<sup>36,37</sup>.

Muscarinic receptors have been the focus of several computational studies as well. Tanczos et al. docked atropine and other similar compounds to wild-type and mutated rat M1, and the relative activities of these compounds agreed with previous experimental studies<sup>38</sup>. Montejo-López et al. docked 30 antagonists, 18 agonists, and 11 partial agonists and developed a QSAR model connecting the molecular volume of these compounds to their biological activities<sup>39</sup>. Mikurova et al. used free-energy calculations to develop models for calculating K<sub>i</sub> values for 42 compounds against M1-M4<sup>40</sup>. However, all these previous computational approaches for M1 have either been structure-based or focused solely on one particular series of compounds for lead optimization, and there are no global machine learning models to effectively screen new compounds.

The first step in developing a machine learning model for M1 is to assemble the publicly available data from various bioactivity databases, such as ChEMBL and BindingDB. These aggregated public datasets are often imbalanced as inactives are rarely reported for dose-response studies<sup>41</sup>. Conversely, HTS studies, which test 100,000s of compounds against a target at a single dose and may only have ~100-1,000 active compounds, also create an imbalanced dataset, but in the opposite direction<sup>42–44</sup>. Models developed exclusively on these imbalanced datasets can be biased, with every query compound predicted to belong to the majority class, and such biased models are not practically useful for screening large chemical libraries<sup>45</sup>. Therefore, it is imperative to implement corrective measures to address these imbalances and enhance the reliability of ligand-based virtual screening.

Several previous cheminformatics studies have attempted to address this issue, both from a data and an algorithm perspective<sup>46</sup>. The data-driven methods focused on either oversampling the minority class by generating synthetic data or undersampling the majority class using methods such as clustering or cleaning up noisy data that hinder classification<sup>47-52</sup>. For example, Idakwo et al. applied several undersampling and oversampling methods on the well-studied Tox21 datasets and showed that oversampling the minority class using synthetic minority oversampling technique (SMOTE) followed by data cleaning using edited nearest neighbor (ENN) outperformed the other methods, although the performance deteriorated once the imbalance became more prominent<sup>52</sup>. Conversely, the algorithm-based methods, which focus on modifying an existing method by increasing the penalty for misclassifying the minority class, have been less frequent in the realm of virtual screening. For example, Li et al. applied the granular support vector machine repetitive undersampling method (GSVM-RU) to extract the most informative majority class samples to build a support vector machine model on a highly imbalanced luciferase HTS dataset<sup>53</sup>.

In this work, we compiled public bioactivity data for M1 to construct deep-learning-based classification models. The initial dataset had a disparate number of actives compared to inactives (Fig. 1b). The models performed well during 10-fold cross validation, but performance worsened significantly for both scaffold-split and HTS test sets, highlighting the commonly encountered problem of imbalanced datasets affecting model performance. In order to address this data imbalance problem when building effective M1 classification models, we employed two different methods: 1) developing a transfer-learning framework using additional GPCR bioactivities and 2) using two generative models, i.e., recurrent neural network (RNN) and transformer-based REINVENT4, to generate additional inactive compounds. These methods contributed to improving both the active and inactive classification results using diverse test sets, including one imbalanced in the opposite direction. We also show the applicability of our methods to other GPCR datasets. Overall, our work provides a well-validated computational tool for screening new compounds against M1 and highlights important strategies to tackle the issue of class imbalance while developing target-specific classification models. We envision that adopting such techniques will help advance future drug discovery efforts.

### Methods

### Data curation

We compiled bioactivity data ( $EC_{50}$ ,  $IC_{50}$ , and  $K_i$ ) for our target M1 from public databases, i.e., ChEMBL and BindingDB. We retrieved the compounds as Simplified Molecular Input Line Entry System (SMILES) strings, which we subsequently processed for validity, desalted, and standardized using the ChEMBL structure pipeline<sup>54</sup>. We excluded entries lacking absolute bioactivity values, except for compounds with values greater than the inactive cutoff, which we labelled as inactives. Our overall goal in this study is to predict the potential of a compound to interact with the M1 receptor, not to predict the implications of this binding interaction (for example, agonism or antagonism). Hence, we collected and combined all available M1 interaction data throughout this work.

We applied an activity cutoff of  $\leq 1 \ \mu$ M for actives and  $\geq 10 \ \mu$ M for inactives. We chose the cutoff to minimize the overlap between the two classes, based on the Tanimoto similarity (see Results and discussion section for a detailed explanation). After assigning classification labels, we removed duplicate SMILES within the same class. If a compound was found in both classes, both SMILES were excluded. Next, using RDKit55, we generated Morgan fingerprints<sup>56</sup> from the SMILES strings, with a length of 1,024 bits and a bond radius of 2. We divided the dataset into training and test sets using an 80:20 split, employing a scaffold-split strategy<sup>57</sup> in DeepChem<sup>58</sup>. This method clusters compounds with similar scaffolds into the same class, thus creating a more challenging test set for the model. In addition, to ensure that all duplicates were removed, we also checked for duplicate fingerprints among the training and test sets, and removed duplicate entries from the training set. The final cleaned dataset contained 1,844 actives and 275 inactives in the training set and 464 actives and 70 inactives in the scaffold-split test set (Fig. 1b). Additionally, we identified a public HTS assay for M1 in PubChem (AID: 588852) and applied the pre-processing steps of desalting, standardization, and duplicate removal to generate a HTS test set consisting of 4,516 actives and 345,301 inactives. We also collected a final test set of known M1 antagonists from DrugBank<sup>59</sup>. This set has only active compounds (Table 1). The details of the composition of our training and test sets are shown in Table 1. To check the reproducibility of our methods across other targets with imbalanced data, we repeated the same procedure on five randomly chosen imbalanced GPCR datasets. Four of the targets belonged to the same subfamily of class A/rhodopsin-like receptors as M1, beta-2 adrenergic receptor, adenosine receptor A2a, C-C chemokine receptor type 5, and gastrin/cholecystokinin type B receptor, while metabotropic glutamate receptor 5 belonged to class C. Adenosine receptor A2a even belonged to the same subgroup A18 as M1.

### Model building

We developed models using four different approaches: Bernoulli Naïve Bayes<sup>61</sup>, random forest<sup>62</sup>, XGBoost<sup>63</sup>, and deep neural network (DNN)<sup>64</sup>. We implemented the first three models using scikit-learn<sup>65</sup>. We constructed the DNN in Python 3.11, utilizing the Keras package with a TensorFlow 2.15 backend<sup>66</sup>. We adapted the DNN architecture from a previous study<sup>67</sup> and carried out hyperparameter optimization, the details of which are shown in Supplementary Table S1. We performed a 10-fold cross validation for each of the combinations and chose the best-performing architecture. The input layer contained 1,024 neurons, with each feature representing an entry from the 1,024-bit Morgan fingerprints, and the output layer consisted of a single neuron. Between these layers, the network included two fully connected hidden layers with 1,000 and 500 neurons. We employed the Adam optimizer<sup>68</sup> and used binary cross-entropy as the loss function, with a learning rate of 0.001, batch size of 64, and 2,000 epochs. We also applied an early stopping criterion with a patience of 50 such that the calculations stopped when the loss function did not improve after 50 epochs. To mitigate overfitting, we implemented dropout regularization with a rate of 0.25 in each layer. We used the rectified linear unit (ReLU) activation function for all the layers except the output layer, where we applied a sigmoid activation function, as this was a classification task requiring output probabilities between 0 and 1<sup>69</sup>.

### Methods to address class imbalance

We used two different methods to address the issue of class imbalance in the M1 dataset (transfer learning from GPCR data and using two generative models, RNN and REINVENT4, to generate additional inactive compounds), and we explain each method below. Since we wanted to assess the impact of our dataset-balancing approaches, further tuning of the DNN hyperparameters was not performed again. We repeated the same procedures for the five additional targets as well.

| Dataset (acronym)                                    | Origin   | No. of actives | No. of inactives |
|--|--|----------------|------------------|
| Original training set (baseline DNN)                 | ChEMBL <sup>26</sup> , BindingDB <sup>27</sup> | 1,844          | 275              |
| Training set with RNN-generated inactives (RNN)      | ChEMBL, BindingDB, RNN <sup>35</sup>           | 1,844          | 1,879            |
| Training set with REINVENT4-generated inactives (R4) | ChEMBL, BindingDB, REINVENT4 <sup>60</sup>     | 1,844          | 1,816            |
| Scaffold-split test set                              | ChEMBL, BindingDB                              | 464            | 70               |
| HTS test set   | PubChem <sup>31</sup>                          | 4,516          | 345,301          |
| DrugBank test set                                    | DrugBank <sup>59</sup>                         | 33             | 0                |

Table 1. Composition of the different datasets used in this study along with their source of origin.

------

### Transfer learning from GPCR data

We curated all the GPCR bioactivity data from the GPCR-Ligand Association (GLASS) database<sup>70</sup>, keeping the same criteria for classifying compounds, i.e., labeling compounds with  $EC_{50}$ ,  $IC_{50}$ , or  $K_i$  values  $\leq 1 \mu M$  as actives and  $\geq 10 \mu M$  as inactives. In order to prevent data leakage, we removed compounds that were present in the M1 bioactivity dataset, resulting in a dataset comprised of 118,865 actives and 16,203 inactives. Similar to our M1 dataset, the GLASS database is also imbalanced, but we expect model improvement due to the exploration of additional chemical space coverage provided by the inactives. We first developed a DNN for the GPCR data following the same protocol outlined above. We then transferred the parameters from the first hidden layers into our initial DNN to create the transfer learning framework (Fig. 2a), without modifying any other component of the model<sup>67</sup>.

### Using RNNs to oversample inactives

RNNs are among the various techniques used for generating new compounds, and they have been demonstrated to produce novel molecules with properties similar to those in a training set<sup>28,34,35</sup>. We employed a character-level RNN (CharRNN) specifically for this purpose, implementing it to oversample inactive compounds (Fig. 2b, top)<sup>35</sup>. This model learned the statistical structure of the SMILES syntax from a large number of SMILES and treated it as a language problem. We optimized the model parameters using maximum likelihood estimation and implemented the CharRNN using three-layered, long short-term memory (LSTM) RNN cells, each with a hidden dimension of 600. To mitigate overfitting, we inserted a dropout layer with a dropout probability of 0.2 between the intermediate layers. The output layer utilized a softmax activation function. We trained the model with a batch size of 64, using the Adam optimizer with a learning rate of 0.001 over 50 epochs. We evaluated the properties of the generated molecules according to the Molecular Sets (MOSES) benchmark<sup>32</sup>.

From this framework, we generated 200,000 compounds using a training set of 275 inactives and standardized them. We excluded SMILES strings that were shorter than the smallest compound in our training set, resulting in a final collection of 5,031 valid compounds. We clustered these RNN-generated compounds into 1,604 clusters and incorporated the cluster centroids into the training set, bringing the inactive set to 1,879 compounds.



**Fig. 2.** Overview of the two dataset-balancing methods: transfer learning and oversampling using generative models. (**a**) Transfer learning model architecture using G protein-coupled receptor (GPCR) bioactivity data from the GPCR-Ligand Association (GLASS) database<sup>70</sup>. (**b**) Overview of the two generative models. (Top) Recurrent neural network model architecture as adapted from Santana and Silva<sup>35</sup>. (Bottom) REINVENT4 model architecture as adapted from Loeffler et al.<sup>60</sup>.

### Using REINVENT4 to oversample inactives

The RNN method was able to generate diverse compounds, but one of the major issues we faced was their low validity, which could possibly be attributed to our small training set. Consequently, we applied a more recent method, REINVENT4<sup>60</sup>, to generate M1 inactive compounds (Fig. 2b, bottom). Specifically, we utilized the mol2mol generator function<sup>71</sup> in REINVENT4, which uses a transformer-based architecture. The model was pre-trained on 2.8 million pairs of molecules in the ChEMBL 28 database, originating from the same publication and containing the same Murcko scaffold, such that it could transform one of these molecules into another. The pre-trained model was available and fine-tuned on the M1 public inactives. Using a scaffold-based constraint allowed us to explore regions of the chemical space similar to the inactives, while also ensuring that the generated compounds were not too similar to the fine-tuning dataset. Our goal was to maximize exploration of new regions within the inactive space. Using a scaffold-based constraint also prevented us from adding molecules similar to our test set to avoid overfitting our model. We compared the compounds generated by the different REINVENT4 mol2mol foundation models,<sup>71</sup> such as scaffold, medium/high similarity, and matched molecular pairs, and found that the compounds generated by scaffold constraints had the least average maximum Tanimoto similarity to the inactives. The Tanimoto similarity was calculated in RDKit using 1,024-bit Morgan fingerprints with a radius of 2. A detailed description of the transformer network can be found in an earlier study<sup>72</sup>.

We followed the same strategy as previously described in the section Using RNNs to oversample inactives to add the REINVENT4-generated compounds to our inactive dataset. Using the same training set of 275 inactives, we generated 8,919 compounds with this method. Following removal of compounds with SMILES strings shorter than the smallest compound in our training set, we combined the remaining compounds into 1,541 cluster centroids and included them in the inactive dataset, bringing their number to 1,816 compounds.

### Comparison against traditional dataset-balancing approaches

We wanted to assess how our methods compare with a few traditional dataset-balancing approaches and implemented three methods paired to XGBoost: 1)  $ENN^{73}$  for undersampling actives, 2) SMOTE followed by ENN (SMOTE-ENN)<sup>74</sup> to oversample inactives and clean the data, and 3) KSMOTE<sup>75</sup> to oversample inactives. We adapted these models from the Imbalanced-learn Python library from scikit-learn.

### Model evaluation

We performed model evaluation using 10-fold cross validation of the training set by randomly splitting the training set into 10 groups and leaving one of them out in each iteration. Due to the imbalanced nature of the data, we used a stratified split and kept the ratio between the active and inactive compounds constant across each fold (Supplementary Figure S1). For each of the 10 iterations, we calculated sensitivity, specificity, area under the receiver operating characteristic curve (ROC AUC), Matthews correlation coefficient (MCC), and G-mean (geometric mean between sensitivity and specificity). The test sets consisted of scaffold-split, HTS, and DrugBank datasets. We evaluated these datasets using the same metrics during each of the 10 cross-validation iterations. These parameters are defined as follows:

Sensitivity = 
$$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$
, (1)

Specificity = 
$$\frac{\text{TN}}{\text{TN} + \text{FP}}$$
, (2)

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$
(3)

where TP represents true positive, TN denotes true negative, FP represents false positive, and FN denotes false negative. Given the need to compare performance across multiple models, and the possibility that one model may outperform the others by chance, we conducted a detailed statistical analysis using Friedman's test<sup>76</sup>. This non-parametric test, analogous to ANOVA, can detect differences across multiple datasets. To further identify which models contributed to significant performance differences, we employed the Conover-Friedman post-hoc test on three metrics, i.e., MCC, ROC AUC, and G-Mean<sup>77</sup>. We considered a model to be significantly better than another only when the Conover-Friedman test was satisfied for each of the three metrics.

### **Results and discussion**

In this study, we developed a classification model to predict the potential of compounds to interact with muscarinic receptor M1. Recognizing the inherent imbalance in the M1 dataset, we employed two strategies to address this issue: transferring model parameters from a DNN trained on GPCR data and oversampling inactives by incorporating two generative models, i.e., RNN and REINVENT4. We validated our findings using a 10-fold cross validation of the training set and scaffold-split, HTS, and DrugBank test sets.

### Evaluation of classification cutoffs and their impact on data curation

We carried out a detailed analysis of the M1 bioactivity distribution which revealed a bias towards lower activity values (Supplementary Figure S2). While it is a standard practice to use a fixed cutoff, such as 1 or 10  $\mu$ M, to distinguish between activity classes, we opted for a more refined approach to minimize class overlap. We designated an active cutoff at 1  $\mu$ M and varied the inactive cutoff from 2 to 10  $\mu$ M to exclude samples with ambiguous classifications. We noticed that as we increased the inactive cutoff, there was a reduction in the overlap between the active and inactive compounds, measured by a maximum Tanimoto similarity greater than

| Active cutoff (µM) | Inactive cutoff (µM) | No. of actives | No. of inactives | Max Tanimoto similarity>0.5 |
|--------------------|----------------------|----------------|------------------|-----------------------------|
| 1                  | 1                    | 2,219          | 1,591            | 1,112                       |
| 1                  | 2                    | 2,256          | 1,311            | 847                         |
| 1                  | 3                    | 2,279          | 1,181            | 737                         |
| 1                  | 4                    | 2,285          | 1,074            | 671                         |
| 1                  | 5                    | 2,288          | 1,007            | 611                         |
| 1                  | 6                    | 2,296          | 955              | 604                         |
| 1                  | 7                    | 2,299          | 915              | 581                         |
| 1                  | 8                    | 2,302          | 874              | 574                         |
| 1                  | 9                    | 2,306          | 852              | 548                         |
| 1                  | 10                   | 2,308          | 345              | 359                         |

**Table 2**. Variation of the number of compounds in the active dataset showing overlap (indicated by max Tanimoto similarity > 0.5) with increasing inactive cutoff.



**Fig. 3**. Distribution of six important physicochemical properties and their comparison among the active, inactive, recurrent neural network (RNN)-generated inactive, and REINVENT4 (R4)-generated inactive compounds. logP, logarithm of the octanol-water partition coefficient.

.....

0.5 (Table 2). This also led to a small increase in the number of active compounds, even though their cutoff remained fixed. We settled for an inactive cutoff of 10  $\mu$ M. Choosing a lower cutoff would have led to assigning inactive labels to many compounds that resembled the active class, affecting the built model. Moreover, previous studies have also shown that using a binary classification cutoff leads to better model performance as compared to a single cutoff<sup>78,79</sup>.

Moreover, we also evaluated whether our approach of utilizing all available data and data integration from multiple assay types led to addition of noise and decreased model performance<sup>80</sup>. We evaluated our approach by creating datasets that included only  $IC_{50}$  or  $K_i$  values and re-ran the models. The  $IC_{50}$  dataset consisted of 586 actives and 95 inactives, while the  $K_i$  dataset contained 837 actives and 149 inactives. We did not observe any significant improvement in classification performance for M1 (Supplementary Figure S3 and Supplementary Table S2) during 10-fold cross validation.

## Analysis of the molecular properties of active and inactive compounds interacting with the muscarinic receptor M1

We calculated six well-characterized physicochemical descriptors—molecular weight, logarithm of the octanolwater partition coefficient, number of rings, rotatable bonds, and hydrogen bond donors and acceptors—for both active and inactive compounds using RDKit (Fig. 3). We treated inactive compounds generated by the RNN and REINVENT4 as separate classes. A comparison of the distribution of these descriptors revealed that the inactives generally had lower molecular weight, reduced lipophilicity, fewer rings, and slightly fewer rotatable bonds as compared to actives. However, due to the shear inequality of the class sizes, we were unsure whether these differences were significant.

The RNN-generated inactives exhibited a narrower distribution, with more similarity to the public inactives, albeit with many outliers, and were consistently lower. We noted that the RNNs had difficulty in generating ringcontaining structures—only 22% of the generated compounds contained at least one ring—indicating potential challenges in capturing long-term dependencies. This limitation may originate from the small size of the training set, which included only 275 inactive compounds, compared to studies focused on larger datasets, such as the ZINC clean leads<sup>32</sup> or GDB-13<sup>81</sup>. A closer look at the properties of these generated compounds (Supplementary Table S3) supported this observation, with a validity rate of only 3%. Efforts to generate additional compounds did not result in proportional increases in validity. Nevertheless, despite the low validity, the generated compounds were unique, novel, and diverse.

Due to limitations in validity and ring formation observed with compounds generated using the RNN model, we explored an alternative generative model REINVENT4, which uses a transformer-based architecture. It is trained on pairs of molecules containing same scaffolds and equipped to explore similar regions of chemical space as the public M1 inactives. Compounds generated by REINVENT4 exhibited superior validity, achieving 100%, as well as enhanced uniqueness. However, due to the controlled nature of their generation, they had a higher similarity to their nearest neighbor. Additionally, they displayed a slightly broader distribution of physicochemical properties and were able to overcome the ring generation problem faced by RNN-generated compounds. Compared to publicly available inactive compounds, the REINVENT4-generated compounds exhibited similar physicochemical profiles, with the exception of a reduction in the number of rings.

To further evaluate the diversity of the compounds, we constructed a t-distributed stochastic neighbor embedding (t-SNE) plot in scikit-learn using 1,024-bit Morgan fingerprints with a radius of 2 (Fig. 4). The chemical space displayed a broad distribution, with the RNN and REINVENT4 inactives occupying the central region and the public actives and inactives distributed more peripherally. We observed considerable overlap between the public actives and inactives, whereas compounds generated by REINVENT4 showed more overlap with the public inactives as compared to RNN, with overlaps being seen in almost all unique scaffolds. However, both methods missed sampling a few inactives that were closer to the public actives. The overlap between the RNN- and REINVENT4-generated compounds and the public actives was minimal.

### Model evaluation by cross validation

We conducted our initial model evaluation using 10-fold cross validation of the training set. We tested four different models—Naïve Bayes, random forest, XGBoost, and DNN—using the cross-validation analysis (Supplementary Table S4). All models showed strong performance in predicting both active and inactive



**Fig. 4**. t-SNE plots of the chemical space occupied by the active (green dots), inactive (red dots), recurrent neural network (RNN)-generated inactive (grey dots), and REINVENT4 (R4)-generated inactive compounds (pink dots).

compounds accurately. Sensitivity values were close to 1, which is expected for models trained on an imbalanced dataset. Specificity was also high, with Naïve Bayes achieving an average of 0.91 and DNN reaching 0.80. MCC, which is considered a robust indicator of model performance on imbalanced datasets<sup>82</sup>, also reached a very high value, with values as high as 0.83 for DNN. Additionally, the ROC AUC and G-Mean values were notably high across all models. A close comparison of the four models revealed that Naïve Bayes and random forest slightly underperformed (Supplementary Figure S4) on this cross-validation dataset. XGBoost and DNN both produced very similar metrics, but we selected DNN as our baseline for further evaluation, as we built a transfer learning protocol using this method.

The performance of the models trained on the imbalanced dataset was already reasonable, and balancing the dataset led to further improvements, especially after incorporating the additional inactives from the generative models (Fig. 5 and Supplementary Table S5). Sensitivity remained high even with the addition of more inactive compounds, with a notable increase in specificity, reaching an average of 0.96 when incorporating either RNN- or REINVENT4-generated inactives. The MCC, ROC AUC, and G-Mean of the RNN- and REINVENT4-addition models were significantly better than the other two models, as indicated by the low *P*-values determined using Friedman's test (Fig. 5) and the Conover-Friedman test (Supplementary Figure S5).

### Model evaluation on test sets

We achieved strong performance metrics for our models during cross validation, particularly when the training set was enhanced with RNN- or REINVENT4-generated inactives. However, our hyperparameters were tuned during this process, making the baseline DNN optimized for this dataset. Moreover, we used random splits to create the cross-validation folds. These factors are very likely to cause model overfitting, and a more challenging task for the models would be to predict M1 binding for compounds not encountered during training. We constructed the first test set using scaffold-split of the public data. While sensitivity remained close to 1 for our baseline DNN model, specificity dropped significantly from 0.81 during cross validation to 0.37 for this test set (Supplementary Table S6). This decline was accompanied by decreases in MCC, ROC AUC, and G-Mean, suggesting that the model was misclassifying a substantial number of inactive compounds. Efforts to balance the training set led to marked improvements in predicting this test set (Fig. 6a). Both ROC AUC and G-Mean improved from 0.72 to 0.75 and from 0.60 to 0.67, respectively, after we implemented transfer learning from the GLASS database, and they increased further to 0.81 and 0.76, respectively, following the incorporation of additional inactive compounds from REINVENT4. Of these two generative methods, REINVENT4 slightly outperformed RNN, probably because it had more valid and unique compounds (Supplementary Table S2). In fact, REINVENT4 significantly outperformed all other methods in all the three metrics (Supplementary Figure S6)

We then wanted to test the generalizability of our models in a new chemical space, so we opted to apply our models to predict the HTS test set. The baseline DNN model's performance worsened even more significantly when predicting this dataset. HTS datasets typically are imbalanced in the opposite direction as compared to public datasets, and this was no exception, making them very challenging to begin with. Our baseline DNN model was not able to correctly identify the large number of inactive compounds, with specificity falling to a mere 0.08 (Supplementary Table S7). Sensitivity remained high, with average values reaching 0.94. Balancing the datasets improved specificity to 0.56 through REINVENT4-addition. Transfer learning and RNN-addition also improved specificity, and consequently G-Mean (Fig. 6b). However, these improvements were achieved at a cost of reduced sensitivity, decreasing to 0.56 upon REINVENT4-addition. MCC for this dataset was always very low, as even with a specificity of 0.56, the REINVENT4-addition approach misclassified 157,481 inactive compounds. Hence, when we compared two M1 test sets differing by orders of magnitude in size, we could not rely on MCC. A recent study also highlighted this same issue, showing that MCC is dependent on the ratio between the two classes and may be underestimated during extreme class imbalance<sup>83</sup>. ROC AUC was around 0.50 for most



**Fig. 5**. Boxplots showing comparisons of the baseline DNN against the two dataset-balancing methods (TL and RNN/R4) during 10-fold cross validation for MCC, ROC AUC, and G-Mean. The *P*-value determined using Friedman's test is shown above each plot. MCC: Matthews correlation coefficient; ROC AUC: area under the receiver operating characteristic curve; G-Mean: geometric mean of sensitivity and specificity; DNN: deep neural network; TL: transfer learning; RNN: recurrent neural network; R4: REINVENT4.

.....



**Fig. 6.** Boxplots showing comparisons of the baseline DNN against the two dataset-balancing methods (TL and RNN/R4) during (**a**) scaffold-split and (**b**) HTS test sets predictions for MCC, ROC AUC, and G-Mean. The *P*-value determined using Friedman's test is shown above each panel. MCC: Matthews correlation coefficient; ROC AUC: area under the receiver operating characteristic curve; G-Mean: geometric mean of sensitivity and specificity; DNN: deep neural network; TL: transfer learning; RNN: recurrent neural network; R4: REINVENT4.

models, reaching its highest value of 0.57 upon REINVENT4-addition, indicating the near-random nature of the predictions. The trend of the two dataset-balancing methods was the same as seen during the earlier prediction of the scaffold-split test set. REINVENT4 had the best performance, followed closely by RNN, and transfer learning showed the least improvement compared to the baseline DNN model and slightly reduced ROC AUC (Supplementary Figure S7).

Finally, we applied the models to a manually collected set of M1 antagonists from DrugBank. The baseline DNN model performed exceptionally well on this single-class dataset, achieving an accuracy of 0.97 (Supplementary Table S8). This was expected considering the imbalanced nature of this model and an all-active test set, and we wanted to evaluate whether our dataset-balancing strategies led to a decrease in accuracy. However, even after we balanced the dataset, the three models continued to perform well, with the lowest accuracy being 0.92 for REINVENT4. Transfer learning was the best-performing method, correctly predicting all the M1 active compounds.

We further assessed the performance of our transfer learning and generative AI approaches for balancing datasets in comparison to three traditional techniques: ENN, SMOTE-ENN, and KSMOTE. We evaluated these methods on both the scaffold-split and HTS test sets. Among the three, SMOTE-ENN showed the best performance, achieving MCC and G-Mean values comparable to those obtained with REINVENT4-addition on the scaffold-split test set (Supplementary Table S6). However, all three methods performed poorly on the HTS test set, underscoring limitations in real-world applications (Supplementary Table S7). These differences likely arise from the inability of traditional techniques to effectively expand the chemical space, as they primarily generate new data points within existing regions and lack the capability to explore more diverse areas compared to generative models.

In a nutshell, our baseline model, built on an imbalanced dataset, showed excellent performance on the crossvalidation set but struggled to replicate this success on the external test sets. The modifications we introduced to balance the training set not only improved the strong performance of the baseline model on the cross-validation dataset but also significantly improved its performance across the two external test sets, with the REINVENT4addition model outperforming the other models in both cases.

### Model evaluation on other datasets

Finally, we evaluated our methods on other datasets to show that the improvement in model performance is not unique to M1 and that our approaches are applicable for any target. We used our methods on five imbalanced GPCR datasets for both 10-fold cross validation and scaffold-split test sets and observed significant improvement in model performance in four out of the five datasets (Supplementary Figures S8-S12). Among our methods, transfer learning was slightly inconsistent and led to a performance deterioration in a few cases, but both generative models, RNN and REINVENT4, offered a consistent enhancement. These results further demonstrate the generalizability of our methods.

### Conclusion

In this study, we developed machine learning models to screen new compounds against M1. Our initial model developed using public datasets was imbalanced, leading to a biased model, and we presented two major strategies to address this issue. The first strategy utilizes the widely adopted technique of transfer learning by building a model with GPCR data and transferring the model parameters. This method showed modest improvements for our M1 dataset, across both external test sets. Our second strategy leverages the use of generative models, a rapidly evolving field, to augment the inactive class. By generating a diverse set of inactive compounds using RNN and REINVENT4, we observed a notable improvement in all our test sets, with better performance than transfer learning. Of these two generative models, RNN struggled to capture long-term dependencies, potentially due to the small training set of 275 compounds, and had low validity. To counter this issue, we implemented REINVENT4, which worked well even on a small training set, generating slightly more diverse compounds covering most unique scaffolds, and we noticed an immediate improvement in performance. Collectively, the solutions we propose enhanced our M1 classification model across three diverse test sets, underscoring the importance of balancing training sets. These approaches are not limited to just M1 but could also be translated to other targets.

This study primarily focused on model development for the muscarinic M1 receptor. Future work will explore the utility of the models in evaluating potential interaction with repurposed drugs. The newly developed models could also be used for screening large chemical databases, such as the 65 billion synthesizable compounds from the Enamine REAL database in future drug discovery efforts. Current computational efforts in GPCR hit discovery are predominantly structure-based, and our models provide a complementary ligand-based approach to enhance virtual screening.

### Data availability

The data employed to conduct our analysis are available on GitHub, at the following URL https://github.com/B HSAI/imbalanced\_data\_M1.

Received: 19 February 2025; Accepted: 2 May 2025 Published online: 12 May 2025

### References

- 1. Hammer, R. & Giachetti, A. Muscarinic receptor subtypes: M1 and M2 biochemical and functional characterization. *Life Sci.* 31, 2991–2998 (1982).
- Maeda, S., Qu, Q., Robertson, M. J., Skiniotis, G. & Kobilka, B. K. Structures of the M1 and M2 muscarinic acetylcholine receptor/ G-protein complexes. Science 364, 552–557 (2019).
- Scarr, E. Muscarinic receptors: Their roles in disorders of the central nervous system and potential as therapeutic targets. CNS Neurosci. Ther. 18, 369–379 (2012).
- 4. Eglen, R. M. Muscarinic receptor subtypes in neuronal and non-neuronal cholinergic function. *Auton. Autacoid. Pharmacol.* 26, 219–233 (2006).
- 5. Vandecasteele, G. et al. Muscarinic and beta-adrenergic regulation of heart rate, force of contraction and calcium current is preserved in mice lacking endothelial nitric oxide synthase. *Nat. Med.* **5**, 331–334 (1999).
- 6. Beroukas, D. et al. Up-regulation of M3-muscarinic receptors in labial salivary gland acini in primary Sjogren's syndrome. *Lab. Invest.* **82**, 203–210 (2002).
- 7. Hegde, S. S. Muscarinic receptors in the bladder: From basic research to therapeutics. Br. J. Pharmacol. 147, S80-S87 (2006).
- Aroniadou-Anderjaska, V., Apland, J. P., Figueiredo, T. H., De Araujo Furtado, M. & Braga, M. F. Acetylcholinesterase inhibitors (nerve agents) as weapons of mass destruction: History, mechanisms of action, and medical countermeasures. *Neuropharmacology* 181, 108298 (2020).
- 9. Amend, N. et al. Diagnostics and treatment of nerve agent poisoning-current status and future developments. *Ann. N Y Acad. Sci.* **1479**, 13–28 (2020).
- 10. Jolivalt, C. G. et al. Using corneal confocal microscopy to identify therapeutic agents for diabetic neuropathy. J. Clin. Med. 11, 2307 (2022).
- 11. Jolivalt, C. G. et al. Topical delivery of muscarinic receptor antagonists prevents and reverses peripheral neuropathy in female diabetic mice. J. Pharmacol. Exp. Ther. 374, 44–51 (2020).
- 12. Cannard, K. The acute treatment of nerve agent exposure. J. Neurol. Sci. 249, 86-94 (2006).
- 13. Wigenstam, E., Artursson, E., Bucht, A. & Thors, L. Supplemental treatment to Atropine improves the efficacy to reverse nerve agent induced bronchoconstriction. *Chem. Biol. Interact.* **364**, 110061 (2022).
- Cornelissen, A. S., Klaassen, S. D., van Groningen, T., Bohnert, S. & Joosen, M. J. A. Comparative physiology and efficacy of Atropine and scopolamine in sarin nerve agent poisoning. *Toxicol. Appl. Pharmacol.* 396, 114994 (2020).
- 15. Pauling, P. J. & Petcher, T. J. Interaction of Atropine with the muscarinic receptor. *Nature* 228, 673–674 (1970).
- Hagan, J. J., Jansen, J. H. & Broekkamp, C. L. Blockade of spatial learning by the M1 muscarinic antagonist pirenzepine. Psychopharmacology 93, 470–476 (1987).
- Lammers, J. W., Minette, P., McCusker, M. & Barnes, P. J. The role of pirenzepine-sensitive (M1) muscarinic receptors in vagally mediated bronchoconstriction in humans. Am. Rev. Respir Dis. 139, 446–449 (1989).
- Kuduk, S. D. & Beshore, D. C. SAR studies on carboxylic acid series M(1) selective positive allosteric modulators (PAMs). Curr. Top. Med. Chem. 14, 1738–1754 (2014).

- Weaver, C. D. et al. Discovery and development of a potent and highly selective small molecule muscarinic acetylcholine receptor subtype I (mAChR 1 or M1) antagonist *in vitro* and *in vivo* probe. Curr. Top. Med. Chem. 9, 1217–1226 (2009).
- Millard, M. et al. Design, synthesis and preclinical evaluation of muscarine receptor antagonists via a scaffold-hopping approach. Eur. J. Med. Chem. 262, 115891 (2023).
- 21. Baskin, I. I., Winkler, D. & Tetko, I. V. A renaissance of neural networks in drug discovery. *Expert Opin. Drug Discov.* 11, 785–795 (2016).
- 22. Gawehn, E., Hiss, J. A. & Schneider, G. Deep learning in drug discovery. Mol. Inf. 35, 3-14 (2016).
- 23. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. Nature 616, 673-685 (2023).
- 24. Vazquez, J., Lopez, M., Gibert, E., Herrero, E. & Luque, F. J. Merging ligand-based and structure-based methods in drug discovery: An overview of combined virtual screening approaches. *Molecules* 25, 4723 (2020).
- Wilson, G. L. & Lill, M. A. Integrating structure-based and ligand-based approaches for computational drug design. *Future Med. Chem.* 3, 735–750 (2011).
- 26. Gaulton, A. et al. ChEMBL: A large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40, D1100–D1107 (2012).
- 27. Gilson, M. K. et al. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 44, D1045–D1053 (2016).
- Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent. Sci. 4, 120–131 (2018).
- 29. Bjerrum, E. J. & Sattarov, B. Improving chemical autoencoder latent space and molecular *de novo* generation diversity with heteroencoders. *Biomolecules* 8, 131 (2018).
- Brown, N., Fiscato, M., Segler, M. H. S. & Vaucher, A. C. GuacaMol: Benchmarking models for *de novo* molecular design. J. Chem. Inf. Model. 59, 1096–1108 (2019).
- 31. Kim, S. et al. PubChem 2019 update: Improved access to chemical data. Nucleic Acids Res. 47, D1102–D1109 (2019).
- 32. Polykovskiy, D. et al. Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Front. Pharmacol.* 11, 565644 (2020).
- Bagal, V., Aggarwal, R., Vinod, P. K. & Priyakumar, U. D. MolGPT: Molecular generation using a transformer-decoder model. J. Chem. Inf. Model. 62, 2064–2076 (2022).
- 34. Bort, W. et al. Discovery of novel chemical reactions by deep generative recurrent neural network. Sci. Rep. 11, 3178 (2021).
- 35. Santana, M. V. S. & Silva-Jr, F. P. *De novo* design and bioactivity prediction of SARS-CoV-2 main protease inhibitors using recurrent neural network-based transfer learning. *BMC Chem.* 15, 8 (2021).
- 36. Burki, T. A new paradigm for drug development. Lancet Digit. Health 2, e226-e227 (2020).
- Ren, F. et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nat. Biotechnol.* 43, 63–75 (2025).
  Tanczos, A. C., Palmer, R. A., Potter, B. S., Saldanha, J. W. & Howlin, B. J. Antagonist binding in the rat muscarinic receptor: A
- study by Docking and X-ray crystallography. *Comput. Biol. Chem.* 28, 375–385 (2004). 39. Montejo-López, W., Sampieri-Cabrera, R., Nicolas-Vazquez, M. I., Aceves-Hernandez, J. M. & Razo-Hernandez, R. S. Analysing the
- effect caused by increasing the molecular volume in M1-AChR receptor agonists and antagonists: A structural and computational study. *RSC Adv.* **14**, 8615–8640 (2024).
- Mikurova, A. V., Skvortsov, V. S. & Grigoryev, V. V. Generalized predictive model of estimation of inhibition of muscarinic receptors M1–M5. *Biomed. Chem. Res. Methods* 3, e00129 (2020).
- Cáceres, E. L., Mew, N. C. & Keiser, M. J. Adding stochastic negative examples into machine learning improves molecular bioactivity prediction. J. Chem. Inf. Model. 60, 5957–5970 (2020).
- 42. Bajorath, J. Integration of virtual and high-throughput screening. Nat. Rev. Drug Discov. 1, 882-894 (2002).
- Szymanski, P., Markowicz, M. & Mikiciuk-Olasik, E. Adaptation of high-throughput screening in drug discovery-toxicological screening tests. Int. J. Mol. Sci. 13, 427–452 (2012).
- 44. White, R. E. High-throughput screening in drug metabolism and pharmacokinetic support of drug discovery. Annu. Rev. Pharmacol. Toxicol. 40, 133-157 (2000).
- Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J. Med. Chem. 53, 2719–2740 (2010).
- 46. Korkmaz, S. Deep learning-based imbalanced data classification for drug discovery. J. Chem. Inf. Model. 60, 4180-4190 (2020).
- Zakharov, A. V., Peach, M. L., Sitzmann, M. & Nicklaus, M. C. QSAR modeling of imbalanced high-throughput screening data in pubchem. J. Chem. Inf. Model. 54, 705–712 (2014).
- Antelo-Collado, A., Carrasco-Velar, R., Garcia-Pedrajas, N. & Cerruela-Garcia, G. Effective feature selection method for classimbalance datasets applied to chemical toxicity prediction. J. Chem. Inf. Model. 61, 76–94 (2021).
- 49. Hussin, S. K. et al. Handling imbalance classification virtual screening big data using machine learning algorithms. *Complexity* 6675279 (2021).
- Casanova-Alvarez, O., Morales-Helguera, A., Cabrera-Perez, M. A., Molina-Ruiz, R. & Molina, C. A novel automated framework for QSAR modeling of highly imbalanced *Leishmania* high-throughput screening data. *J. Chem. Inf. Model.* 61, 3213–3231 (2021).
- Hemmerich, J., Asilar, E. & Ecker, G. F. COVER: Conformational oversampling as data augmentation for molecules. J. Cheminform. 12, 18 (2020).
- 52. Idakwo, G. et al. Structure-activity relationship-based chemical classification of highly imbalanced Tox21 datasets. *J. Cheminform.* **12**, 1–19 (2020).
- Li, Q., Wang, Y. & Bryant, S. H. A novel method for mining highly imbalanced high-throughput screening data in pubchem. Bioinformatics 25, 3310–3316 (2009).
- 54. Bento, A. P. et al. An open source chemical structure curation pipeline using RDKit. J. Cheminform. 12, 51 (2020).
- 55. RDKit. open-source cheminformatics (2025). https://www.rdkit.org. Accessed April 11, 2025.
- 56. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. J. Chem. Inf. Model. 50, 742-754 (2010).
- 57. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. J. Med. Chem. 39, 2887-2893 (1996).
- 58. Ramsundar, B. et al. Deep Learning for the Life Sciences (O'Reilly Media, 2019).
- 59. Knox, C. et al. DrugBank 6.0: The drugbank knowledgebase for 2024. Nucleic Acids Res. 52, D1265–D1275 (2024).
- 60. Loeffler, H. H. et al. Reinvent 4: Modern AI-driven generative molecule design. J. Cheminform. 16, 20 (2024).
- 61. Schütze, H., Manning, C. D. & Raghavan, P. Introduction To Information Retrieval Vol. 39 (Cambridge University Press, 2008).
  - 62. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
  - 63. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (2016).
  - 64. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* 23, 1241–1250 (2018).
  - 65. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
- 66. Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467 (2016).
- 67. Liu, R., Laxminarayan, S., Reifman, J. & Wallqvist, A. Enabling data-limited chemical bioactivity predictions through deep neural network transfer learning. *J. Comput. Aided Mol. Des.* **36**, 867–878 (2022).
- 68. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv:1412.6980 (2014).
- 69. Agarap, A. Deep learning using rectified linear units (ReLU). arXiv:1803.08375 (2018).

- Chan, W. K. et al. GLASS: A comprehensive database for experimentally validated GPCR-ligand associations. *Bioinformatics* 31, 3035–3042 (2015).
- 71. He, J. et al. Transformer-based molecular optimization beyond matched molecular pairs. J. Cheminform. 14, 18 (2022).
- 72. He, J. et al. Molecular optimization by capturing chemist's intuition using deep neural networks. J. Cheminform. 13, 1–17 (2021).
- Wilson, D. L. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Syst. Man. Cybernet. SMC 2, 408–421 (1972).
- Batista, G. E., Prati, R. C. & Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newsl. 6, 20–29 (2004).
- 75. Douzas, G., Bacao, F. & Last, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inform. Sci.* 465, 1–20 (2018).
- 76. Pereira, D. G., Afonso, A. & Medeiros, F. M. Overview of Friedman's test and post-hoc analysis. Comm. Stat. Simul. Comput. 44, 2636-2653 (2015).
- 77. Conover, W. J. & Iman, R. L. Multiple-Comparisons Procedures. Informal Report (Los Alamos National Lab, 1979).
- Siramshetty, V. B., Chen, Q., Devarakonda, P. & Preissner, R. The Catch-22 of predicting hERG Blockade using publicly accessible bioactivity data. J. Chem. Inf. Model. 58, 1224–1233 (2018).
- 79. Chen, Y., Yu, X., Li, W., Tang, Y. & Liu, G. In silico prediction of hERG blockers using machine learning and deep learning approaches. J. Appl. Toxicol. 43, 1462–1475 (2023).
- Landrum, G. A. & Riniker, S. Combining IC<sub>50</sub> or K<sub>i</sub> values from different sources is a source of significant noise. J. Chem. Inf. Model. 64, 1560–1567 (2024).
- Zhang, J., Mercado, R., Engkvist, O. & Chen, H. Comparative study of deep generative models on chemical space coverage. J. Chem. Inf. Model. 61, 2572–2581 (2021).
- 82. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **21**, 1–13 (2020).
- 83. Guesné, S. J., Hanser, T., Werner, S., Boobier, S. & Scott, S. Mind your prevalence! J. Cheminform. 16, 43 (2024).

### Disclaimer

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Defense Health Agency, the U.S. Department of Defense, or The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. Distribution Statement A. Approved for public release: Distribution is unlimited.

### Author contributions

S.D., A.W., and M.D.M.A. designed the study. S.D. performed the computations and data analysis. S.D. wrote the first draft of the manuscript. All authors commented on previous versions of the manuscript and approved the final manuscript.

### Funding

This research was funded by the U.S. Army Medical Research and Development Command under Contract Numbers W81XWH20C0031 and HT942524F0189 and by Defense Threat Reduction Agency Grant CB-Call14-CBS-05-2-0007.

### Declarations

### **Competing interests**

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/1 0.1038/s41598-025-00972-w.

Correspondence and requests for materials should be addressed to A.W. or M.D.M.A.

Reprints and permissions information is available at 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/.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2025