Catch the next designer drug: A comparison of methods for predicting in vitro pharmacological profiles using molecular fingerprints

Kedan He*[a], Ceana E. Rodriguez[b]

Abstract: Facing the growing trend of emerging new psychoactive substances (NPS) and their threat to public health, more effective methods of NPS prediction and identification are of critical importance. In this study, we aimed to compare methods for predicting the pharmacological profile of unknown compounds based on the chemical structures described by molecular fingerprints. We built predictive models based on the high-throughput screening (HTS) data sets using four different machine learning algorithms for a total of 10 targets and validated the performance of the models using in vitro bioassay data collected from the literature for an external NPS compound set. Clustering analysis revealed that the MACCS fingerprint may be more suitable for describing the similarity of pharmacological profiles of NPS, indicated by the highest adjusted Rand index (0.46) between the two clustering trees. The SVM classifiers validated by the external NPS set achieved ROC AUC above 0.80 and MCC above 0.45, therefore were used to generate the multi-target pharmacological profiles. The hit rate for retrieving pharmacologically similar compound pairs using MACCS calculated Tanimoto coefficient was below 1.85%, in contrast, the models were more successful in identifying similar compound pairs (MCC = 0.72), which were otherwise considered dissimilar by molecular fingerprints.

Keywords: New psychoactive substances, Pharmacological profile, Molecular fingerprints, Similarity search, Tanimoto similarity coefficient, Hierarchical clustering, Machine learning

[a] Eastern Connecticut State University, Physical Sciences, 83 Windham St, Willimantic, CT 06226, USA.
E-mail: hek@easternct.edu,
Phone: 860-465-5759

[b] Eastern Connecticut State University, Biology, 83 Windham St, Willimantic, CT 06226, USA.

Supporting Information for this article is available on the WWW under www.molinf.com
1 Introduction

"Designer drugs" or new psychoactive substances (NPS) are compounds that slightly modify the molecular structure of existing controlled substances to mimic their pharmacological effects and bypass legislation.\cite{1,2} Terms such as “research chemicals, bath salts, fertilizers, incense, and plant foods” are used to circumvent legislation designed to control the supply and distribution of these substances. According to the United Nations Office on Drugs and Crime (UNODC), 126 countries have reported a total of more than 1,047 NPS as of December 2020.\cite{1} Using a 24/7 web crawler to capture the real number of NPS shows over 4,000 unique substances of interest circulating in the online environment, a number roughly four times greater than that reported in known NPS databases.\cite{3} Many countries have used or amended existing legislation, or innovative legal instruments, as a way to address the prevalence of NPS. For example, the Controlled Substances Act, passed in 1986 in the United States, allows any chemical that is "substantially similar" to a Schedule I or II controlled substance to be treated as a Schedule I substance.\cite{2} In the UK, any substance that is not regulated by the Misuse of Drugs Act 1971 falls within the scope of the Psychoactive Substances Act 2016.\cite{4} However, the ban on any particular NPS or NPS category has led to a rapid substitution in the market. Given that these compounds will now reach users through more clandestine routes and that synthetic drug overdose mortality is increasing across all age groups, races, genders, and ethnicities, new tools and methods must be developed to more effectively address the current problem of NPS abuse.\cite{5}

The combination of chromatographic separation and mass spectrometric analysis, such as GC-MS and LC-MS, is the gold-standard technique for analyzing NPS.\cite{6} However, these highly accurate and sensitive techniques have disadvantages such as lack of portability, time-consuming, need for trained specialists, and expensive operation and maintenance. The use of
infrared,\cite{7} Raman vibrational spectroscopy,\cite{8} and surface-enhanced Raman scattering (SERS)\cite{9,10} for \textit{in situ} detection has attracted a lot of interest because they can provide a portable, fast, and cost-effective solution for providing actionable feedback to first responders. \textit{In situ} detection using handheld instruments often relies on built-in algorithms and the development of reference spectral libraries.\cite{11,12} To identify "unknown" compounds with confidence, there needs to be a high degree of structural similarity between the "unknown" and the library substance.\cite{13} Due to time constraints, availability and cost, it is difficult to obtain complete information on each known NPS. Coupled with the challenge that previously unseen NPSs continue to emerge rapidly, emerging NPS are often not identified by library-based matching algorithms. An alternative strategy is to select substances to create a diverse subset as a "representative" of the physicochemical properties of known NPS.\cite{14}

Representative NPS compounds can be selected from clusters formed based on structural similarity followed by the identification of medoids,\cite{14} based on the observation that structurally similar compounds may have similar biological activities and exhibit similar spectral behavior.\cite{15} However, such a subset does not adequately represent the diverse pharmacological space covered by the NPS. The current classification of NPS is based on their chemical scaffolding and/or pharmacological/clinical effects. A systematic classification of NPS based on pharmacological effects is very challenging because a drug often interacts with many different biological targets. For example, the main pharmacological mechanism of stimulants such as cathinones is to inhibit the reuptake of neurotransmitters and induce their release into the synaptic cleft.\cite{16-18} However, cathinones class also exhibit many other pharmacological responses, including hallucinogenic and antidepressant effects associated with the interaction of tyrosine and tryptophan hydroxylase.\cite{19,20} In contrast, synthetic cannabinoids interact less
ambiguously with CB1 receptors but containing very structurally diverse molecules. Similarly, clinical classification of NPS designed to effectively treat NPS intoxication faces the same problem, as some NPS may have overlapping effects in multiple clinical categories.

Despite these challenges, the use of structural similarities to identify compounds with similar biological activities has been the subject of virtual screening (VS) strategies.\textsuperscript{[21]} Molecular fingerprints have been used for quantitative similarity measurements to analyze the diversity of compound databases, in ligand-based virtual screenings to find near-neighbors of known active molecules, and for scaffold-hopping to identify molecules with different structures but similar biological activity for given targets.\textsuperscript{[22]} However, ligand-based similarity search methods perform poorly when the number of known ligands is insufficient, such as when there are far more unknown NPS compounds than known NPS compounds. In the meantime, molecular fingerprints have been used as descriptors for building predictive models with machine learning workflows for various endpoints such as target bioactivity, solubility, and toxicology.\textsuperscript{[23-26]} For compounds that interact with multiple targets, arrays of activity predictions generated from these models can be used as activity profile fingerprints, and then similarity searches are performed to identify compounds with similar multi-target affinity profiles, which may or may not be highly similar in structure.\textsuperscript{[27]}

For all the reasons mentioned above, a comprehensive assessment of the correspondence between structural and pharmacological profile similarities is essential for the effective detection of emerging NPS. The aim of this study was to investigate the relationship between structural similarities derived from two-dimensional molecular fingerprint methods and the corresponding similarities in pharmacological profiles derived from the multi-target bioassays of NPS compounds. In addition, we investigated the effectiveness of NPS-target interaction prediction
using a machine learning-based (ML-based) approach. For key targets within the central nervous system that are critical for the classification of stimulants and hallucinogens classes, we developed predictive models using the high-throughput screening (HTS) data from the PubChem bioassay database.[28] Using an external set of NPS compounds collected from literature with their \textit{in vitro} bioactivities measured with the same assay and procedure, we further validated the performance of the predictive models. To identify the appropriate multi-target array sets used to create pharmacological profile fingerprints (PFs), we used the external NPS set for hierarchical clustering analysis, and compared the agreement of clustering trees generated from structural similarities and pharmacological profile similarities. Finally, we evaluated the performance of retrieving pharmacologically similar compound pairs using both structural similarity-based and ML-based approaches.

2 Materials and Methods

2.1 Data sets and molecular fingerprints

All compounds were downloaded from the PubChem include the Simplified Molecular Input Line System String (SMILES) and the CID (PubChem Compound ID) for each molecule for unique identification.[28] Active or inactive labels for each compound were assigned using the reported activity outcomes. In addition, we also constructed a combined dataset (PubChem\textunderscore M) containing 163 compounds and their bioassay outcome array as PFs. The description of the data preprocessing is provided in Appendix A of the supporting documentation. For external validation of predictive models and hierarchical clustering analysis, an external NPS set was created containing a total of 117 NPS compounds collected from literature and their \textit{in vitro} pharmacological profiles were measured using the same assay methods and procedures.[29-39] Substances are pharmacologically classified as stimulants or hallucinogens based on whether
they primarily inhibit monoamine transporters reuptake or mostly bind to 5-hydroxytryptamine (5-HT) receptors, respectively.\cite{40} Twenty-eight cathinones, 5 aminoindanes, 8 benzofurans, 15 phenethylamines, 3 piperazines, 9 piperidines, 1 indole, and 3 other NPSs were categorized as stimulants. One benzodifuran, 1 ergoline, 7 tryptamines, and 36 phenethylamines were categorized as hallucinogens. The phenethylamines class was divided into 15 stimulants (amphetamine-type substances) and 36 hallucinogens (ring-substituted phenethylamines including 2C drugs and their methoxybenzyl [NBOMes] analogues).

The 2D molecular fingerprints used are listed in Table 1. Substructural key-based fingerprints, MACCS \textit{(maccsfp)} and PubChem \textit{(pubchemfp)}, the absence (0) and presence (1) of a specific chemical group is coded in each bit position.\cite{41} Morgan \textit{(morganfp)} is the RDKit implementation of the ECFP4 circular topological fingerprint with radius 2 as bit vector.\cite{42} The RDKit-specific fingerprint \textit{(rdkfp)} is inspired by public descriptions of the Daylight fingerprint, typically capturing all subgraphs in a molecule within a given number of connecting bonds, and is particularly useful for clustering compounds together, similar to how medicinal chemists would divide compounds into structurally related groups. PubChem fingerprints were downloaded from PubChem database, and all other fingerprints where calculated using RDKit software.\cite{43}

<table>
<thead>
<tr>
<th>Name</th>
<th>Acronym</th>
<th>Type</th>
<th>Bit length</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubChem</td>
<td>\textit{pubchemfp}</td>
<td>structural</td>
<td>881</td>
</tr>
<tr>
<td>MACCS</td>
<td>\textit{maccsfp}</td>
<td>structural</td>
<td>116</td>
</tr>
<tr>
<td>ECFP4</td>
<td>\textit{morganfp}</td>
<td>Hashed topological</td>
<td>1024</td>
</tr>
<tr>
<td>RDKit</td>
<td>\textit{rdkfp}</td>
<td>Hashed topological</td>
<td>2048</td>
</tr>
<tr>
<td>PFs\textsuperscript{[a,b]}</td>
<td>\textit{pf0 - 3}</td>
<td>Pharmacological</td>
<td>14, 10, 9, 7</td>
</tr>
</tbody>
</table>

\textsuperscript{[a]} Pharmacological profile fingerprints (PFs) are derived from HTS results downloaded from the PubChem bioassay database or literature data.  
\textsuperscript{[b]} Each bit in the fingerprint corresponds to an assay. The length of the bits corresponds to the total number of targets included. Please refer to Appendix A of the supporting documentation for more details.
In order to quantify the structural and pharmacological profile similarities, the binary Tanimoto coefficient \( T_c \) was calculated, which is defined as follows:

\[
T_c = \frac{c}{a + b - c}
\]  

(1)

where “c” is the number of bits common to the two fingerprints and “a” and “b” denote the number of bits set in each of the two fingerprints. The values of \( T_c \) range from 0 to 1, where 1 means identical. The pharmacological profile similarity is referred to as the PF similarity in following text. The maximum common substructure (MCS) similarity was also calculated used the \textit{rdFMCS} module implemented in RDKit for the MCS SMARTS string calculations by matching atomic elements and bond types. The MCS Tanimoto similarity coefficient \( T_{MCS} \) is calculated only on the matched heavy atoms as described:[44]

\[
T_{MCS} = \frac{N_c}{N_A + N_B - N_c}
\]  

(2)

where \( N_c \) is the number of matched heavy atoms in MCS of molecule A and B, \( N_A \) and \( N_B \) are the number of heavy atoms in molecule A and B, respectively.

2.2 Binary classification model training and validation

The PubChem data set for each target was divided into an 80% training and a 20% test set. Four different machine learning (ML) methods were selected: Random Forest (RF),[45, 46] Support vector machine (SVM)[47] with a linear kernel function, Multilayer perceptron neural networks (MLP), and similarity-weighted \textit{k}-Nearest neighbors (\textit{k}-NN).[48] The models were evaluated using 10-fold cross-validation to avoid overfitting, hyperparameters were optimized by randomized grid search. The quality of the model was evaluated by the quantities defined in Table 2. Accuracy and F1 fail to consider the ratio between positive and negative elements. The Matthews Correlation Coefficient (MCC) incorporates the imbalance of the data set and its
invariance to the exchange of classes and is therefore considered a balanced measure of the biased data set.[49] Independent of their ratio in the dataset, the classifier must make correct predictions for both negative and positive cases to obtain a high MCC. It ranges in the interval of $[-1, +1]$ and reaches the extreme values of $-1$ and $+1$ in the case of complete misclassification and perfect classification, respectively, while $MCC = 0$ is the expected value of the coin tossing classifier. In addition, the receiver characteristic operator (ROC) curves and their AUC (area under the curve score) were also calculated. A detailed ML workflow and the optimal hyperparameters chosen for each classifier can be found in Appendix B and GitHub repository.

### Table 2. Confusion table and performance metrics for binary classification

<table>
<thead>
<tr>
<th>Confusion table[a]</th>
<th>Metrics</th>
<th>Formulas</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>$\frac{tp + tn}{tp + tn + fp + fn}$</td>
</tr>
<tr>
<td>Actual label</td>
<td>Recall</td>
<td>$\frac{tp}{tp + fn}$</td>
</tr>
<tr>
<td>Predicted label</td>
<td>Precision</td>
<td>$\frac{tp + fp}{tp \cdot Pre \cdot Rec}$</td>
</tr>
<tr>
<td>Positive</td>
<td>F1 score</td>
<td>$\sqrt{\frac{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}{tn \cdot fn \cdot (tn - fp \cdot fn)}}$</td>
</tr>
<tr>
<td>Negative</td>
<td>MCC[b]</td>
<td>$\frac{tp \cdot tn - fp \cdot fn}{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}$</td>
</tr>
</tbody>
</table>

[a] $tp$: true positive, $tn$: true negative, $fp$: false positive (Type I error), $fn$: false negative (Type II error).

[b] $MCC$: Matthews correlation coefficient.

### 2.3 Hierarchical clustering analysis and comparison

The $n \times n$ Tanimoto coefficient matrices were used as input features and submitted to a ward linkage clustering with Euclidean distance as the similarity metric. The clustering performance using other linkage methods were assessed and more information could be found in Appendix C. The confidence of clustering results was quantitatively assessed by cophenetic correlation coefficient[50, 51] and the optimal number of clusters $K$ was determined by silhouette analysis,[52] which is a measure of how well cluster members belong to their respective clusters,
given the number of desired clusters $K$. Silhouette score ranges from $-1$ to $+1$, where $+1$ means clusters are well apart from each other and clearly distinguished, $-1$ indicates member is assigned to the wrong cluster. The structural similarity-based clustering trees were compared with the PF similarity-based clustering trees using the adjusted Rand-Index (ARI).\cite{53} All hierarchical clusterings were generated using the `fcluster` and `dendrogram` in `scipy.cluster.hierarchy` package, cophenetic correlation coefficients were calculated using `cophenet` in `scipy.cluster.hierarchy`. Heatmaps were generated using `seaborn` package, and both silhouette scores and ARI were computed using `sklearn.metrics` package.

2.4 Comparison of pharmacologically similar compound pairs prediction

In the final analysis, we performed similarity searches using both structural similarity-based and ML-based methods to identify compound pairs with similar pharmacological profiles. We defined the compound pairs that exhibited a PF similarity above a cutoff value of 0.7 as the positive set (1) and all the remaining pairs as the negative set (0). For predictions using the similarity-based approach, we calculated the maximal MCC score for each molecular fingerprint and at which structural similarity threshold value the maximal MCC occurred. When using the PFs derived using the ML classifiers in the prediction, the same PF similarity cutoff was used to distinguish between PF similar and unsimilar pairs, and the MCC score was calculated to assess the quality of the classification.
3 Results and Discussion

In this section, we provide (1) a summary of the training/test sets used to build the predictive models and the external NPS set curated from literature for external model validation, (2) a comparison of the test sets and the external NPS set prediction results to justify the selection of SVM-rdkfp classifiers, (3) the selection of target arrays used to create PFs, (4) the hierarchical clustering analysis of external NPS set to assess the performance of PFs in characterizing the pharmacological profile of NPS, and (5) the comparison of structural similarity-based and ML-based methods in finding pharmacologically similar compound pairs using similarity search.

3.1 Data curation and preprocessing

According to the recommendations of the Advisory Council on the Misuse of Drugs (ACMD), the in vitro testing should be used to demonstrate whether a substance is psychoactive.[54] The 117 compounds in the external NPS set can be classified as stimulants or hallucinogens[40] based on their in vitro pharmacological and psychoactive effect profiles.[29-39] The biological activities reported include: monoamine uptake transporters (NET, DAT, and SERT) inhibitions and binding affinities, 5-hydroxytryptamine receptors (5-HT1A, 5-HT2A, and 5-HT2C) binding affinities, adrenergic receptors (α1A and α2A) binding affinities, dopamine D2 receptor binding affinities, and trace amine-associated receptors (TAARs) binding affinities. The external NPS set compounds and their reported activities data can be found in the GitHub repository.

The training and test data sets for building classifiers were collected for all targets from the PubChem BioAssay database.[28] Targets (5-HT1A, α1A) with high active/inactive (1/0) ratios (> 10) or no inactives reported (TAARs) were excluded, resulted in a total of 10 data sets.
Redundant entries and compounds that were not confirmed in confirmatory assays or had conflicting activity outcomes were further excluded. A description of the data source and pre-processing are provided in Appendix A. Table 3 gives a summary of the preprocessed training/test sets and the external NPS set. The training/test sets active/inactive (1/0) ratio varied among target assays from 0.80/0.90 to 3.12/3.11, respectively. For data with this degree of imbalance, rebalancing techniques do not significantly improve model prediction performance and therefore were not used in this study. The distribution of bioactivities of different targets in the external NPS set correlates to some extent with their importance in describing the pharmacological profile space of NPS, therefore we did not perform any artificial resampling of this data set. In general, the number of active compounds that the model learns from the training set is much larger than the number of actives it needs to predict from the external NPS set.

**Table 3.** Class distribution and active/inactive (1/0) ratio of the preprocessed training/test datasets and the external NPS set.

<table>
<thead>
<tr>
<th>In vitro assay target</th>
<th>Total number of compounds</th>
<th>Active</th>
<th>Inactive</th>
<th>1/0 ratio (training)</th>
<th>1/0 ratio (test)</th>
<th>Active</th>
<th>Inactive</th>
<th>1/0 ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>NET_I</td>
<td>2357</td>
<td>1478</td>
<td>879</td>
<td>1.65</td>
<td>1.83</td>
<td>76</td>
<td>41</td>
<td>1.85</td>
</tr>
<tr>
<td>DAT_I</td>
<td>2334</td>
<td>1209</td>
<td>1125</td>
<td>1.07</td>
<td>1.09</td>
<td>44</td>
<td>73</td>
<td>0.60</td>
</tr>
<tr>
<td>SERT_I</td>
<td>3017</td>
<td>1986</td>
<td>1031</td>
<td>1.89</td>
<td>2.07</td>
<td>50</td>
<td>64</td>
<td>0.75</td>
</tr>
<tr>
<td>NET</td>
<td>2111</td>
<td>1229</td>
<td>882</td>
<td>1.41</td>
<td>1.32</td>
<td>12</td>
<td>105</td>
<td>0.11</td>
</tr>
<tr>
<td>DAT</td>
<td>2206</td>
<td>1183</td>
<td>1023</td>
<td>1.17</td>
<td>1.11</td>
<td>25</td>
<td>92</td>
<td>0.27</td>
</tr>
<tr>
<td>SERT</td>
<td>2829</td>
<td>1860</td>
<td>969</td>
<td>1.90</td>
<td>1.99</td>
<td>9</td>
<td>108</td>
<td>0.08</td>
</tr>
<tr>
<td>5-HT2A</td>
<td>4214</td>
<td>3148</td>
<td>1066</td>
<td>2.91</td>
<td>3.12</td>
<td>52</td>
<td>65</td>
<td>0.80</td>
</tr>
<tr>
<td>5-HT2C</td>
<td>2929</td>
<td>1894</td>
<td>1035</td>
<td>1.82</td>
<td>1.87</td>
<td>48</td>
<td>69</td>
<td>0.70</td>
</tr>
<tr>
<td>α2A</td>
<td>1519</td>
<td>702</td>
<td>817</td>
<td>0.85</td>
<td>0.90</td>
<td>45</td>
<td>72</td>
<td>0.62</td>
</tr>
<tr>
<td>D2</td>
<td>6814</td>
<td>5159</td>
<td>1655</td>
<td>3.12</td>
<td>3.11</td>
<td>6</td>
<td>111</td>
<td>0.05</td>
</tr>
</tbody>
</table>

[a] The highest 1/0 ratio are in bold, and the lowest 1/0 ratio are in bold italic.
[b] Compounds reported in PubChem Bioassay data as “Unspecified”, “Inconclusive” were also labeled as inactive (0) in the final training data sets.
[c] For original data collected from Liechti et al. [29-39] an inhibition IC50 greater than 10 μM or binding affinity Ki greater than 1 μM is assigned as inactive (0) and lower than this threshold is assigned as active (1).
Several issues are worth mentioning. The PubChem Bioassay HTS data often lack the standardized activity values (such as IC50 or $K_i$, etc.) but instead the “activity outcomes”. However, the external NPS set contains the IC50 or $K_i$ values in micromolar units, thus an arbitrary cutoff value needs to be chosen for assigning the active or inactive labels. The use of the cutoff approach would result in the loss of some valuable information that is meaningful for describing the pharmacological/clinical profile of NPS. For example, DAT/SERT inhibition ratio has been identified as a useful marker for assessing the potential clinical psychotropic and acute toxic effects of NPS.$^{[29]}$ Based on the value of this ratio, compounds can be broadly categorized as MDMA/empathogens-like or methamphetamine/stimulant-like. Compounds with DAT/SERT inhibition ratios $>1$, such as cocaine, are also associated with high abuse potential.$^{[29, 30]}$ The second challenge is that except for 13 stimulants and 5 hallucinogens, most external NPS set compounds don't have HTS bioassay results reported in the PubChem database. The choice of this cutoff value affects the active/inactive ratio of the external NPS set. Due to the very limited number of compounds common to both data sources, it was difficult to cross-reference the consistency of activity values, therefore a uniform activity cutoff could not be found to assign activity labels. Finally, we simulated the use of the external NPS set as query molecules to identify emerging unknown NPS using similarity searches. This PubChem_M set was created by including training/test compounds that were labeled as active against at least one or more targets. In total, 163 compounds were included. The distribution of structural similarity described by molecular fingerprints and calculated $T_c$ are shown in Figure 1. The narrower and left-shifted distribution of the pairwise comparison between the external NPS set and the PubChem_M set compounds indicate increased chemical diversity, or in another word, lowered degree of structural similarity. It serves as a good hypothetical simulation of the emerging unknown NPS
as it might not necessarily be chemically similar to the known NPS. In this case, the structural similarity-based virtual screening would have poor performance in the identification of the likely new NPS.

![Histograms of all pairwise structural similarities.](image)

**Figure 1.** Histograms of all pairwise structural similarities.

### 3.2 Classification models for single target prediction

A total of 16 models were developed for each of the 10 targets using four ML methods with four molecular fingerprints (see “Materials and Methods” section for detail). The performance metrics and optimized hyperparameters of all classifiers can be found on GitHub repository. According to our analysis using cross-validation and average performance, the SVM and RF models have the highest MCC (0.76 and 0.75) and AUC (0.94 and 0.95), while the worst k-NN models have MCC and AUC of 0.73 and 0.92, respectively. The MCC and AUC values for all 16 classifiers for each target are given in Figure 1. Despite the differences among targets, the
same trend was observed according to both metrics: the SVM-\textit{morganfp} and RF-\textit{morganfp} have the best performances, with MCC over 0.74 and AUC over 0.94. Even though the mean MCC and AUC of \textit{k}-NN were the lowest, the resulted MCC was above 0.70 and AUC above 0.93 when \textit{morganfp} was used.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline
\textbf{Method} & \textbf{MCC} & \textbf{ROC AUC} \\
\hline
\textit{SVM-pubchemfp} & 0.72 & 0.73 & 0.72 & 0.76 & 0.74 & 0.82 & 0.80 & 0.74 & 0.72 & 0.93 & 0.93 & 0.94 & 0.95 & 0.94 & 0.97 & 0.95 & 0.95 & 0.94 & 0.92 \\
\hline
\textit{SVM-maccsfp} & 0.74 & 0.70 & 0.74 & 0.81 & 0.79 & 0.83 & 0.80 & 0.77 & 0.73 & 0.89 & 0.94 & 0.93 & 0.95 & 0.94 & 0.97 & 0.95 & 0.94 & 0.93 & 0.91 \\
\hline
\textit{SVM-morganfp} & 0.80 & 0.76 & 0.75 & 0.79 & 0.79 & 0.87 & 0.82 & 0.77 & 0.76 & 0.95 & 0.94 & 0.94 & 0.96 & 0.96 & 0.98 & 0.95 & 0.94 & 0.94 & 0.95 \\
\hline
\textit{SVM-rdkfp} & 0.74 & 0.72 & 0.70 & 0.82 & 0.76 & 0.81 & 0.79 & 0.77 & 0.71 & 0.93 & 0.92 & 0.93 & 0.95 & 0.95 & 0.97 & 0.95 & 0.95 & 0.94 & 0.94 \\
\hline
\textit{RF-pubchemfp} & 0.71 & 0.71 & 0.75 & 0.79 & 0.72 & 0.79 & 0.79 & 0.74 & 0.75 & 0.68 & 0.94 & 0.92 & 0.93 & 0.95 & 0.94 & 0.97 & 0.96 & 0.94 & 0.94 \\
\hline
\textit{RF-morganfp} & 0.76 & 0.68 & 0.69 & 0.76 & 0.74 & 0.84 & 0.80 & 0.75 & 0.73 & 0.68 & 0.94 & 0.93 & 0.93 & 0.95 & 0.95 & 0.97 & 0.96 & 0.96 & 0.93 & 0.92 \\
\hline
\textit{RF-rdkfp} & 0.76 & 0.77 & 0.74 & 0.80 & 0.76 & 0.84 & 0.80 & 0.77 & 0.76 & 0.76 & 0.95 & 0.95 & 0.94 & 0.96 & 0.96 & 0.98 & 0.96 & 0.96 & 0.94 & 0.94 \\
\hline
\textit{MLP-pubchemfp} & 0.69 & 0.66 & 0.72 & 0.81 & 0.76 & 0.78 & 0.75 & 0.74 & 0.73 & 0.73 & 0.93 & 0.92 & 0.93 & 0.95 & 0.95 & 0.97 & 0.96 & 0.96 & 0.94 & 0.94 \\
\hline
\textit{MLP-maccsfp} & 0.73 & 0.75 & 0.73 & 0.77 & 0.72 & 0.82 & 0.78 & 0.74 & 0.70 & 0.69 & 0.94 & 0.94 & 0.93 & 0.95 & 0.93 & 0.97 & 0.96 & 0.96 & 0.92 & 0.93 \\
\hline
\textit{MLP-morganfp} & 0.76 & 0.71 & 0.72 & 0.76 & 0.72 & 0.62 & 0.70 & 0.77 & 0.71 & 0.60 & 0.94 & 0.93 & 0.94 & 0.94 & 0.94 & 0.97 & 0.95 & 0.95 & 0.93 & 0.92 \\
\hline
\textit{MLP-rdkfp} & 0.67 & 0.69 & 0.68 & 0.78 & 0.74 & 0.72 & 0.67 & 0.70 & 0.64 & 0.64 & 0.91 & 0.89 & 0.90 & 0.91 & 0.89 & 0.94 & 0.92 & 0.91 & 0.91 & 0.91 \\
\hline
\textit{k-NN-pubchemfp} & 0.65 & 0.60 & 0.68 & 0.74 & 0.67 & 0.75 & 0.74 & 0.72 & 0.70 & 0.64 & 0.89 & 0.86 & 0.90 & 0.93 & 0.90 & 0.93 & 0.93 & 0.92 & 0.91 & 0.91 \\
\hline
\textit{k-NN-maccsfp} & 0.68 & 0.69 & 0.66 & 0.73 & 0.76 & 0.79 & 0.79 & 0.71 & 0.71 & 0.64 & 0.93 & 0.85 & 0.89 & 0.90 & 0.93 & 0.93 & 0.93 & 0.92 & 0.88 \\
\hline
\textit{k-NN-morganfp} & 0.75 & 0.75 & 0.74 & 0.80 & 0.77 & 0.83 & 0.79 & 0.76 & 0.73 & 0.70 & 0.94 & 0.93 & 0.93 & 0.95 & 0.95 & 0.97 & 0.96 & 0.94 & 0.93 & 0.93 \\
\hline
\textit{k-NN-rdkfp} & 0.72 & 0.70 & 0.71 & 0.78 & 0.75 & 0.80 & 0.78 & 0.75 & 0.72 & 0.68 & 0.92 & 0.92 & 0.91 & 0.93 & 0.94 & 0.96 & 0.93 & 0.93 & 0.92 & 0.90 \\
\hline
\end{tabular}
\caption{Heat map of the MCC and ROC AUC results models trained with 10-fold cross validation, evaluated using the test sets.}
\end{table}

3.3 Selecting SVM-\textit{rdkfp} as the classifier

To determine the best combination of ML methods and molecular fingerprints validated by the external NPS set, the SVM and RF models determined in the previous step were evaluated with test sets and repeated 10 times with different random splits. Finally, the results were calculated using the average of the 10 runs. Table 4 shows the average performance of the SVM and RF models summarized by the molecular fingerprints used. Based on the test set performance, there was no significant difference between the SVM and RF models, with slightly better results for \textit{morganfp} and \textit{rdkfp} according to MCC and AUC. However, since the ratio of active to inactive of the external NPS set differs from that of the training and test data sets, as expected, a more pronounced difference was observed among the external NPS set results. The
SVM-rdkfp classifier showed the highest MCC, AUC, F1, and Precision of 0.47, 0.83, 0.61, and 0.54, respectively. Therefore, we chose SVM-rdkfp to create and predict the multitarget PFs.

Table 4. Average performance metrics of SVM and RF models on test set and external NPS set

<table>
<thead>
<tr>
<th></th>
<th>Test Set</th>
<th></th>
<th></th>
<th></th>
<th>External NPS Set</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MCC</td>
<td>AUC</td>
<td>F1</td>
<td>Rec</td>
<td>MCC</td>
<td>AUC</td>
<td>F1</td>
<td>Rec</td>
</tr>
<tr>
<td>SVM-pubchemfp</td>
<td>0.76</td>
<td>0.94</td>
<td>0.91</td>
<td>0.90</td>
<td>0.90</td>
<td>0.93</td>
<td>0.36</td>
<td>0.80</td>
</tr>
<tr>
<td>SVM-maccsfp</td>
<td>0.76</td>
<td>0.94</td>
<td>0.91</td>
<td>0.90</td>
<td>0.92</td>
<td>0.93</td>
<td>0.34</td>
<td>0.76</td>
</tr>
<tr>
<td>SVM-morganfp</td>
<td>0.79</td>
<td>0.95</td>
<td>0.92</td>
<td>0.92</td>
<td>0.93</td>
<td>0.93</td>
<td>0.35</td>
<td>0.81</td>
</tr>
<tr>
<td>SVM-rdkfp</td>
<td>0.77</td>
<td>0.95</td>
<td>0.91</td>
<td>0.91</td>
<td>0.92</td>
<td>0.92</td>
<td>0.47</td>
<td>0.83</td>
</tr>
<tr>
<td>RF-pubchemfp</td>
<td>0.75</td>
<td>0.95</td>
<td>0.91</td>
<td>0.89</td>
<td>0.92</td>
<td>0.92</td>
<td>0.27</td>
<td>0.76</td>
</tr>
<tr>
<td>RF-maccsfp</td>
<td>0.75</td>
<td>0.95</td>
<td>0.91</td>
<td>0.90</td>
<td>0.92</td>
<td>0.92</td>
<td>0.27</td>
<td>0.75</td>
</tr>
<tr>
<td>RF-morganfp</td>
<td>0.78</td>
<td>0.96</td>
<td>0.92</td>
<td>0.91</td>
<td>0.93</td>
<td>0.93</td>
<td>0.35</td>
<td>0.80</td>
</tr>
<tr>
<td>RF-rdkfp</td>
<td>0.75</td>
<td>0.95</td>
<td>0.91</td>
<td>0.90</td>
<td>0.91</td>
<td>0.91</td>
<td>0.40</td>
<td>0.79</td>
</tr>
</tbody>
</table>

[a] All scores are averaged for all 10 targets for each ML method-fingerprint combination.

[b] SVM: support vector machine, RF: random forest, MCC: Matthews correlation coefficient, AUC: under the curve of the receiver characteristic operator curve, Pre: precision, Rec: recall (sensitivity).

3.4 Selecting target arrays for creating PFs

Figure 3 shows the SVM classifiers test set MCC and AUC scores, broken down by each target. The average MCC, AUC, and Precision for the SERT binding were 0.83, 0.97, and 0.94, respectively, while the 5-HT2A binding had the highest average F1 and Recall of 0.95 and 0.97, respectively. The predictions made by the RF models also yielded the best results for these two targets. Overall, all metrics were within a standard deviation of 0.01-0.04. The average MCC and AUC using SVM models were 0.77 and 0.95, respectively, and 0.76 and 0.95 using RF, where the average Precision and Recall were all higher than 0.90.

However, shown in Figure 4, the prediction performance of the external NPS set decreased compared to that of the test set and varied among targets. The variance in the predicted results is partly due to the difference in the active/inactive ratios compare to that of the training/test sets. A total of seven targets had an MCC above 0.43 and AUC higher than 0.80 when using the SVM-rdkfp classifier for the external NPS set. The 5-HT2A binding had the
highest average MCC, AUC, Precision, Recall, and F1 of 0.70, 0.94, 0.76, 0.94, and 0.84, followed by 5-HT$_{2C}$ and α$_{2A}$ binding, as well as NET and DAT inhibitions. In contrary, the D$_2$ binding had the lowest MCC and Precision of 0.03 and 0.06, SERT inhibition had the lowest AUC of 0.54, SERT binding affinities had MCC and Precision of 0.28 and 0.18 respectively. Of the 117 NPS compounds in the external set, only 9 and 6 were labeled as against SERT and D2 based on the reported $K_i$ values, consequently, the classifiers are expected to perform the worst for both targets, as shown from our results. Since the active is a minority class of the external NPS set, it is consistent with the fact that the prediction resulted in higher Recall than Precision. It could also be indicating that those two targets are not offering valuable information on distinguishing compounds with similar or unsimilar pharmacological profiles, which might contribute to this large decrease in the performances. The other interesting case is the SERT inhibition, which is not as biased as SERT and D2 binding within the external NPS set. The active/inactive ratio of SERT inhibition are 1.93 and 0.75 in the training/test and external NPS sets, respectively. The SERT inhibition had a low MCC of 0.09 and AUC of 0.54, and comparable Precision and Recall of 0.49 and 0.43. For following analysis, we created four versions of the PFs: pf0, pf1, pf2, and pf3, each is an array include the total 14 targets, 10 targets, 9 targets (without SERT Inhibition), and 7 targets, respectively (see Table 1). The shortest pf3 included bioassays of NET, DAT inhibitions, NET, DAT, 5-HT$_{2a}$, 5-HT$_{2C}$, and α$_{2A}$ binding affinities.
Figure 3. The performance of SVM models evaluated using test sets.
**Figure 4.** The performance of SVM models evaluated using the external NPS set. MCC, Matthews correlation coefficient; ROC, receiver operating characteristic curve; AUC, area under the curve. Molecular fingerprints: MACCS (maccsfp), PubChem (pubchemfp), Morgan (morganfp), and RDKit (rdkfp). The target name is listed in each column. NET_I, DAT_I, and SERT_I represent the inhibition bioassays of these monoamine transporters, respectively. The red dashed lines correspond to MCC = 0.43 and ROC AUC = 0.80, respectively.

### 3.5 Clustering with pharmacological profile fingerprints (PFs)

To compare the performance of clustering external NPS set for categorizing PF similar compounds, four versions of PFs were used to calculate the 117×117 PF similarity matrices. The pf3 with the shortest bit length was created by including targets with MCC above 0.45 and AUC above 0.80 from the SVM-rdkfp prediction of the external NPS set shown in Figure 4. The optimal number of clusters $K$ was determined by silhouette analysis (see “Materials and
Methods” section), which is a measure of the extent to which cluster members belong to their respective clusters given the desired number clusters $K$. Silhouette values were computed for cluster numbers $K = 2$ to $K = 50$ and the results are plotted in Figure 5. We also performed an analysis by varying the distance cutoff $t \in [0.1, 5.0]$ to reduce the amount of overlap between the different clusters and the resulted dendrograms are provided Appendix C. The resulted optimal clusters $K$ and corresponding maximal silhouette scores are in agreement between both analyses, indicating that the clusters are sufficiently distinct.

![Silhouette analysis to determine optimal number of hierarchical clusters with PFs](image)

**Figure 5.** Silhouette analysis for determining optimal clusters $K$ using PFs similarity matrices.

The average silhouette score increases as the bit length of PFs shorten from $pf0$ to $pf3$, while the number of clusters formed at which the maximal silhouette score was found decreases. This confirms that the four targets dropped from $pf0$ do not provide useful information to distinguish the unique pharmacological profile of NPS compounds. The optimal distance cutoff is located at 1.7 using $pf3$ with a corresponding maximal silhouette score of 0.94 (**Figure 6**), and this hierarchy yielded a total of $K = 17$ clusters. Looking at the comparison of cluster membership at $K = 2$ (stimulants vs. hallucinogens), the results were mainly influenced by the
inclusion/absence of SERT inhibition of PFs. Five tryptamines and 5 phenethylamines (NBOMes series), which inhibit SERT in addition to binding to the 5-HT$_{2A}$ and/or 5-HT$_{2C}$ receptors, were classified as stimulants using $pf0$. Conversely, when SERT was excluded from $pf3$, 7 benzofurans, 3 aminoindanes, 2 piperazines, and 3 other stimulants were classified as hallucinogens instead, because these compounds showed binding affinity to at least one or more targets of 5-HT$_{2A}$, 5-HT$_{2C}$, or $\alpha_{2A}$.

**Figure 6.** Average silhouette score at optimal clusters $K$ determined using distance cutoff.

### 3.6 PFs cluster membership analysis

Hierarchical clustering groups items based on a similarity metric, and the relationships between the various clusters can be visualized on a dendrogram, as well as their comparative sizes. Clustering with PFs aimed at grouping NPS compounds that are likely to have similar pharmacological and therefore psychoactive profiles. Of the 17 clusters identified using $pf3$, 2 were orphan clusters or “singletons” that contained only one molecule. Compared to the clustering using $pf1$ and $pf2$, which resulted in 29 and 25 clusters, respectively, and 7 singletons.
in both cases. We argue that pf3 provides a reasonable balance between the representation of diverse pharmacological characteristics of NPS, while still ensuring that only compounds with similar activity in the same assays are grouped together. The pf3 clustering heatmap in Figure 7 shows multiple supergroups: 57 stimulants of clusters 1 – 5 that mainly inhibit the monoamine uptake transporters NET, DAT, and SERT, with one hallucinogen (DiPT) placed under this supergroup. Interestingly, the cathinones class compounds were split among four clusters: together with amphetamine-type phenethylamines as monoamine uptake transporters inhibitors (cluster 1 and 2); with piperidines that show binding affinities to NET and DAT (cluster 3 and 4). Majority of the ring-substituted phenethylamines (2C and NBOMes series) were grouped in cluster 6 as hallucinogens due to their binding affinity to 5-HT$_{2A}$, 5-HT$_{2C}$, and $\alpha_{2A}$. The clusters 7 – 9 formed a supergroup of hallucinogens including 5 tryptamines and 3 2C-phenethylamines. Finally, there is a diverse supergroup consisting of members of clusters 10 – 17, showing overlapping pharmacological profiles of both stimulants and hallucinogens. Eight among the 9 benzofurans were classified to this supergroup.
3.7 Comparison of structural similarity-based and PF similarity-based clustering

The maximal ARI was calculated to compare the clustering generated based on structural similarity with those based on PF similarity, as described in the Materials and Methods section. In order to gain further insights into the behavior of the clustering performance, the MCS Tanimoto ($T_{MCS}$) similarity matrix was also included. All structural similarity-based clustering trees were compared with the optimal PFs clustering trees determined using the distance cutoffs identified in the previous section as external references. For ease of discussion, we will refer to each clustering tree by the original molecular fingerprints and PFs used. Figure 8 shows the maximal ARI for each fingerprint method as well as the distance cutoff value at which the maximal ARI occurs and the corresponding number of clusters $K$.

The average ARI increases from 0.27 to 0.38 when the external clustering reference changes from $pf0$ to $pf1$. This may be because the pharmacological profile described by $pf1$ is more suitable for grouping similar compounds with a common core chemical structure. This is consistent with the result that the maximal average silhouette score increased from 0.69 to 0.84 for clustering from $pf0$ to $pf1$ (Figure 6). However, although clustering using $pf3$ proved to have the best clustering performance in classifying NPS compounds with similar pharmacological profiles, molecular fingerprint clusterings did not further improve the agreement with the PF similarity clusterings, except for macsfp, which had an ARI of 0.46 when compared with the $pf3$ clustering. The overall low level of clustering agreement (ARI below 0.50) suggested that, in general, the structural description of molecules is not particularly well suited to be used to group NPS compounds to reflect the diversity of their intrinsic pharmacological characteristics. This discrepancy could be partially accounted for by the “activity cliff”, which is defined as a pair of
structurally similar compounds that differ significantly in potency. Undoubtedly, the assessment of compound similarity depends to a large extent on the chosen molecular representation. A $T_C$ value of 0.85 using MACCS has often been applied as a similarity criterion for activity cliff formation.\[^{[55]}\] In comparison to topological fingerprints such as ECFP4 (equivalence of \textit{morganfp} in this study), this value is approximately equivalent to a $T_C$ value of 0.56.\[^{[56]}\] As we noted in the previous analysis, structurally different compounds could also show similar biological activity against an array of targets. This observation suggests the necessity to consider both pharmacological and chemical structural landscapes when building a representative NPS database.

![Figure 8. The adjusted Rand-Index of clustering trees comparison.](image)

3.8 Comparison of structural similarity-based and ML-based prediction of PFs on retrieval of PF similar compound pairs

The previously unseen NPS continue to emerge at an alarming rate posing additional challenges to their accurate and rapid detection. The difficulty in obtaining commercially available NPS reference standards has greatly hindered the establishment of extended instrumental libraries. We demonstrated using stimulants and hallucinogens compounds that clustering by chemical structural similarity to build a representative NPS library, while valuable,
does not fully capture the entire complex pharmacological profile space. However, with the increasing availability of HTS data, predictive models can be constructed for each target individually and then subsequently used to predict the multi-target pharmacological profile of sample compounds. To assess how effective the PFs generated using predictive models compared to structural similarity approaches in retrieving PF similar compound pairs, we performed similarity searches using the SVM-rdkfp classifiers generated pf3 (pf3_pred) and molecular fingerprints, respectively. The external NPS set compounds were used as query molecules to search among the PubChem_M set. The pf3 for the PubChem_M set compounds were created using the HTS data.

Compound pairs with a pf3 similarity greater than or equal to 0.7 cutoff were defined as the positive set (1) and all remaining pairs are the negative set (0). PF similar compound pairs were retrieved by comparing the molecular fingerprint $T_C$ to a threshold value. The maximal MCC score was calculated by varying the molecular fingerprint $T_C$ threshold from 0.1 – 0.9 in steps of 0.01. Different pf3 similarity cutoff value affects the true positive/negative pair ratios. The closer this cutoff value is to 0, the increasing number of compound pairs are defined as positive sets, leading to higher F1 and Precision. However, the MCC scores are expected to be relatively independent on this cutoff. We analyzed the dependence of the prediction results on the choice of pf3 similarity cutoff values using the external NPS set pairwise comparison, where the results are shown in Appendix D. The maximal MCC plateaued after the pf3 similarity cutoff became equal to or greater than 0.7. Therefore, this pf3 similarity cutoff value and the corresponding molecular fingerprint $T_C$ threshold values are chosen for subsequent analysis. The molecular fingerprint $T_C$ threshold identified for pubchemfp, maccsfp, morganfp, rdkfp were 0.72, 0.60, 0.24, 0.42, respectively. The MCSS similarity threshold used was 0.68.
Table 5 summarized the similarity search results using the \textit{pf3} similarity cutoff greater than or equal to 0.7. When searching among the external NPS set, \textit{maccsfp} performed the best with the highest MCC of 0.52, followed by the two hashed fingerprints (\textit{morganfp} and \textit{rdkfp}), where the substructural fingerprint \textit{pubchemfp} had the lowest maximal MCC score of 0.40. This trend is consistent with the ARI derived from the comparison of clustering trees, with the \textit{maccsfp} clustering showing the highest degree of agreement with the \textit{pf3} clustering tree.

Interestingly, although MCSS was useful in forming NPS clusters based on structural similarity, it was the worst predictor of PF similarity. Finally, the prediction using \textit{pf3\_pred} similarity has MCC of 0.53, which is higher than the prediction using \textit{maccsfp}.

\textbf{Table 5.} Number of PF similar pairs retrieved using structural similarity and SVM-\textit{rdkfp} classifier.

\begin{tabular}{lcccccc}
\hline
\textbf{External NPS set vs. External NPS set: 117 × 117 = 6786 pairs} & \textit{pf3} & \textit{pf3\_pred} & \textit{pubchemfp} & \textit{maccsfp} & \textit{morganfp} & \textit{rdkfp} & \textit{MCSS} \\
\hline
Positive & 1244 & 1637 & 1187 & 904 & 1447 & 1247 & 946 \\
Negative & 5542 & 5149 & 5389 & 6212 & 6718 & 6427 & 5941 \\
MCC & --- & 0.53 & 0.40 & 0.52 & 0.47 & 0.47 & 0.27 \\
\hline
\textbf{External NPS set vs. PubChem\_M: 117 × 163 = 19071 pairs} & \textit{pf3} & \textit{pf3\_pred} & \textit{pubchemfp} & \textit{maccsfp} & \textit{morganfp} & \textit{rdkfp} & \textit{MCSS} \\
\hline
Positive & 1352 & 1531 & 216 & 113 & 282 & 129 & 35 \\
Negative & 17719 & 17540 & 18751 & 19056 & 19070 & 19064 & 19047 \\
\%Similar & 7.09 & 8.03 & 1.13 & 0.59 & 1.48 & 0.68 & 0.18 \\
MCC & --- & 0.72 & 0.02 & 0.05 & 0.02 & 0.03 & 0.01 \\
\hline
\end{tabular}

\textsuperscript{[a]} \textit{pf3} similarity cutoff $\geq 0.7$ was used to define the positive set.

\textsuperscript{[b]} The molecular fingerprint $T_c$ threshold used for \textit{pubchemfp}, \textit{maccsfp}, \textit{morganfp}, \textit{rdkfp} were 0.72, 0.60, 0.24, 0.42, respectively. The MCSS similarity threshold used was 0.68.

The PubChem\_M set contains compounds that are more structurally diverse when compared to the external NPS set compounds (see Figure 1). Only 7.09\% of the compound pairs between both sets are considered to be pharmacologically similar using \textit{pf3} similarity cutoff $\geq 0.7$. In comparison, the structural similarity-based approach consistently retrieved a much smaller fraction from 8.03\% to as low as 0.18\% of compound pairs as PF similar. The MCC
using SVM-rdkfp derived pf3_pred is 0.72, and comparatively, all molecular fingerprint $T_C$-based predictions had MCC scores between 0.01 – 0.05, indicating that relying on a single structural similarity $T_C$ value does not predict better than random guessing. Chart 1 shows examples of four compound pairs that were predicted to have identical pf3 but were not retrieved by any molecular fingerprint methods. Both ethcathinone and sibutramine inhibit and bind to 5-HT and monoamine reuptake transporters (SERT and NET) and therefore are listed as DEA Schedule I and IV controlled substances, respectively. 2C-B (Nexus) and phentolamine pair share the lowest morganfp $T_C$ of 0.08 among all compound pairs in Figure 9. 2C-B is one of the most widespread NPS, inducing euphoria and mild hallucinogenic states, and is classified as hallucinogen due to its binding affinity for 5-HT and $\alpha$-adrenergic receptors. Phentolamine has a markedly different structure, is a synthetic imidazoline that acts as a competitive $\alpha$-adrenergic antagonist, and also may block 5-HT receptors. Nortriptyline, a tricyclic compound, is similar to diclofenacine and shows broad biological activity against several targets, such as inhibition of monoamine reuptake transporters and binding to $\alpha$-adrenergic, histaminergic, and cholinergic receptors. Finally, olanzapine binds with high affinity to 5-HT, dopaminergic, and $\alpha$-adrenergic receptors and therefore is clustered with the classical hallucinogen LSD in this study.

Figure 10 a and b shows the distribution of maccsfp and morganfp similarities for all compound pairs, where the pairs with pf3 similarity cutoff values equal to or higher than 0.7 are also shown in the bar chart for comparison. Using a maccsfp $T_C \geq 0.60$, only 0.59% of compound pairs are predicted to be PF similar, which corresponds to a hit rate of 1.85%. Using a morganfp $T_C \geq 0.24$, only 282 compound pairs were regarded as PF similar, however, 94.7% of the pf3 similar pairs were calculated to have a morganfp $T_C \leq 0.20$. In contrast, 79% of the pf3 similar pairs are reported to have pf3_pred similarity above 0.7 (see Figure 10.c). This suggests
that although similar pharmacological profile does not necessarily correspond to structural similarity, especially when evaluated with a single score such as Tanimoto coefficient, the machine learning model trained with structural molecular fingerprints as input features can greatly improve the performance of identifying compound pairs that are likely to share similar pharmacological profile.
Figure 9. Compound pairs that were not retrieved using any molecular fingerprint methods.
Figure 10. Distribution of the maccsf, morganfp, and pf3_pred similarity values between pf3 similar and pf3 unsimilar compound pairs.

4 Conclusion

Given the rapid growth in the number of newly synthesized NPS, it is impractical to study all of them in detail. As the chemical space of emerging NPS becomes increasingly diverse compare to that of known NPS, attempts to identify potential new NPS using virtual screening based on the similarities calculated using molecular fingerprinting become ineffective. As demonstrated by our analysis using the external NPS set compounds of stimulants and
hallucinogens, there is an overall low degree of agreement between the clustering generated using structural similarity and pharmacological profile similarity. It would be interesting to compare the performance of representative databases constructed using structure-based clustering alone or in combination with the pharmacological space of the NPS in identifying unknown samples. Comparative analysis of the retrieval of similar compound pairs suggests that predictive models trained with the HTS data sets can be used to generate multi-target pharmacological profile fingerprints of samples that can subsequently be used for similarity search. A more economical approach to mitigate the public health threat of NPS would be to quickly screening compounds for biological activity against NPS targets and build predictive models using molecular fingerprints as input features.

**Acknowledgements**
Resources were provided in part by the MERCURY consortium ([http://mercuryconsortium.org/](http://mercuryconsortium.org/)) under NSF grants CHE-1229354, CHE-1662030, and CHE-2018427.

The authors declare no competing financial interest.

**Data and Software Availability**
Supporting information is provided, and the code used is available upon request to the corresponding author.

**Author Contributions**
KH designed the study, analyzed and organized the data, CR assisted in the collection of data and manuscript preparation. All authors have given approval to the final version of the manuscript.

**Abbreviations**
NPS: new psychoactive substances; SERS: surface-enhanced Raman scattering; VS: virtual screenings; ML-based: machine learning-based; PFs: pharmacological profile fingerprints; PF similarity: pharmacological profile similarity; MCS: maximum common substructure; MCC: Matthews correlation coefficient; 5-HT: 5-hydroxytryptamine; SMILES: Simplified Molecular

References


Figure 1. Histograms of all pairwise structural similarities.
[b] The modes of histograms are labeled for each distribution.

Figure 2. Heat map of the MCC and ROC AUC results models trained with 10-fold cross validation, evaluated using the test sets.
[a] MCC, Matthews correlation coefficient; ROC, receiver operating characteristic curve; AUC, area under the curve.
[b] The target name is listed in each column. NET_I, DAT_I, and SERT_I represent the inhibition bioassays of these monoamine transporters, respectively.
[c] The methods used are: SVM, support vector machine; RF, random forest; MLP, multilayer perceptron; k-NN, weighted k-nearest neighbor. Each method was combined with four molecular fingerprints: MACCS (maccsfp), PubChem (pubchemfp), Morgan (morganfp), and RDKit (rdkfp).

Figure 3. The performance of SVM models evaluated using test sets.
[a] MCC, Matthews correlation coefficient; ROC, receiver operating characteristic curve; AUC, area under the curve.
[b] Molecular fingerprints: MACCS (maccsfp), PubChem (pubchemfp), Morgan (morganfp), and RDKit (rdkfp).
[c] The target name is listed in each column. NET_I, DAT_I, and SERT_I represent the inhibition bioassays of these monoamine transporters, respectively.

Figure 5. Silhouette analysis for determining optimal clusters $K$ using PFs similarity matrices.
[a] Silhouette values as a function of number of clusters $K$ in ward linkage clustering.
[b] The optimal clusters $K$ is determined corresponding to the maximum silhouette value.

Figure 6. Average silhouette score at optimal clusters $K$ determined using distance cutoff.
[a] The total number of clusters corresponding to the maximal average silhouette score are shown for each PFs.

Figure 7. Heatmap of the $pβ$ similarity clustering of external NPS compound.

Figure 8. The adjusted Rand-Index of clustering trees comparison.
[a] The bars show the maximal ARI values derived for each structural similarity-based clustering compared with the PFs-based clustering.

Figure 9. Compound pairs that were not retrieved using any molecular fingerprint methods.
[a] The highest and lowest molecular fingerprint similarities values are shown for each compound pairs.

**Figure 10.** Distribution of the `maccsfp`, `morganfp`, and `pf3_pred` similarity values between `pf3` similar and `pf3` unsimilar compound pairs.