JOURNAL OF
EMERGING INVESTIGATORS

Article

Development of a novel machine learning platform to identify structural trends among NNRTI HIV-1 reverse transcriptase inhibitors

Bhavesh Ashok[1,8,9,*], Tanish Baranwal[2,8,*], Udbhav Avadhani[3,9,], Geethika Biddala[1,9,**], Tvisha Nepani[4,9,**], Vishak Srikanth[5,8], Luqman Zaceria[6,8], Natalia Williams[7,8], Robert Downing[8], Edward Njoo[9]**

[1]Amador Valley High School, Pleasanton, California
[2]Dublin High School, Dublin, California
[3]Leigh High School, San Jose, California
[4]Milpitas High School, Milpitas, California
[5]BASIS Independent Silicon Valley, San Jose, California
[6]James Logan High School, Union City, California
[7]Bentley School, Lafayette, California
[8]Dept. of Computer Science & Engineering, Aspiring Scholars Directed Research Program, Fremont, California
[9]Dept. of Chemistry, Biochemistry, & Physics, Aspiring Scholars Directed Research Program, Fremont, California
* equal first authors
**equal second authors

SUMMARY

With advancements in machine learning powered by on-demand computing and information processing at a large data scale, high throughput virtual screening has become a more attractive method for screening drug candidates, reducing both costs and the timeframe from hit-to-lead. The efficiency offingerprinting using cheminformatics-based approaches coupled with machine learning to model and identify structure-activity relationships (SAR) has immense potential to improve the time-to-market for the drug development process. This study compared the accuracy of molecular descriptors from two cheminformatics software libraries, Mordred and PaDEL, in characterizing the chemo-structural composition of 53 compounds from the non-nucleoside reverse transcriptase inhibitors (NNRTI) class. We built a logistic regression model to classify NNRTIs based on salient descriptors from each software. We hypothesized that the descriptor data generated by Mordred would be more accurate when characterizing the SAR between NNRTI compounds and the HIV-1 RT enzyme. We identified structural trends in potential inhibitors of the HIV-1 RT enzyme. The classification model built with the filtered set of descriptors from Mordred was superior to the model using PaDEL descriptors as it revealed significant cluster separation between the 53 NNRTI molecules and other drugs, while the filtered PaDEL descriptors model did not show a clear distinction between classes. This approach can accelerate the identification of hit compounds and improve the efficiency of the drug discovery pipeline.

INTRODUCTION

The introduction of computational methods to aid in drug discovery has greatly improved the efficiency in screening compounds for potential biological activity (1). Traditional routes of drug discovery entail intensive synthetic development requiring significant time, effort, and millions of dollars in research funding (2). With computational methods such as molecular docking, researchers have been able to identify compounds that have the potential to bind to and activate/inhibit specific biological processes before going through the lengthy process of synthesizing and performing *in vitro* screening on a large library of compounds (3). However, these methods have been limited by the high computational expenses and potential inaccuracies such as false negatives or false positives, necessitating a more advanced platform relying on data-driven techniques to identify structural trends in compounds that present a desired biological activity (3).

Human Immunodeficiency Virus 1, or HIV-1, is the most common type of human immunodeficiency virus. The reverse transcriptase (RT) enzyme aids retroviruses such as HIV in replication by synthesizing viral DNA from RNA templates, the latter of which are more susceptible to damage and alteration (4). As a result of its vital role in viral replication, RT is often the target of many antiretroviral drugs (5-7). Non-nucleoside reverse transcriptase inhibitors (NNRTIs) are a class of antiretroviral drugs that prevent the conversion of RNA to DNA by reversibly binding to and blocking the HIV RT enzyme (8). As NNRTIs interfere with the conversion from RNA to DNA, the genetic material of the retrovirus is unable to integrate with existing healthy genetic material in the cells, preventing replication and the formation of new viruses. NNRTIs serve as crucial components to antiviral drug combination methods for the HIV-1 infection because of their high specificity and low toxicity (9, 10).

Cheminformatics is the application of computational and informational techniques to solve key problems in chemistry, such as *in silico* mapping of possible molecules through storing, indexing, searching, retrieving, and applying information about physicochemical, structural, and biological properties, spectroscopic signatures, etc. of molecules (11). A molecular descriptor is a structural or physicochemical property of a molecule or part of a molecule that is generated computationally to describe the physical and chemical information contained within the molecules (11). Molecular descriptors derived from atomic or molecular properties provide various physicochemical, topological, and surface properties of compounds that play a vital role in modeling interactions and effects that are critical for *in silico* drug discovery. As an example, LogP is a commonly used molecular descriptor of the lipophilicity of molecules that measures the partitioning of the molecule between an aqueous phase and a lipophilic phase, which is an important characteristic *in vivo* as the molecule moves through biological membrane lipid layers (11). Cheminformatics approaches rely on many molecular descriptors that define various structural, molecular, and local quantities characterizing the reactivity, shape, binding properties, polarizability, and energy of a composite molecule as well as its molecular fragments and substructures (12). Molecular descriptors help in high throughput virtual screening of molecular libraries as they can find molecules with similar physical or chemical properties (12). Molecular fingerprints are a way of encoding the structure of a molecule mathematically to indicate the presence or absence of substructures in the molecule. One can determine the similarity between two molecules and find matches based on querying a substructure simply by comparing fingerprints (12). For example, substructure fingerprints are useful for querying small molecules such as drugs, while atom-pair fingerprints can be used for querying large molecules such as peptides. Molecular fingerprints can additionally aid in performing large-scale statistical and machine learning analyses on molecules in high throughput screening scenarios for drug candidates (12).

Processes such as molecular docking predict stable, three-dimensional (3D) protein-ligand complexes with a high degree of accuracy, allowing researchers to design drugs *in silico* and perform a preliminary examination of the viability of the compound before synthetic development (13). In tandem with a large library of compounds, molecular docking allows for the discovery of existing compounds that can be docked to new targets, known as high throughput virtual screening (HTVS). This allows for the development of a computational structure-activity relationship (SAR), correlating chemical structures with binding affinity, which is useful for identifying "hit" compounds for hit-to-lead drug development (13). However, all this comes at a high computational expense, which greatly limits the use of HTVS when screening large chemical spaces for drug discovery. Fingerprinting and cheminformatics have been shown to be a low-expense alternative for the identification of computational SARs between a library of ligands and a target (14).

PaDEL-Descriptor (PaDEL) is a free software that generates 1,875 descriptors and 12 fingerprints for molecules, conveniently outputted into a CSV format for data parsing (15). Similarly, Mordred is a molecular descriptor calculator which computes around 1,800 descriptors for 42 attributes of a molecule, and its direct integration into Python makes it a valuable tool for data mining and processing (16). In this study, we compared the accuracy of molecular descriptors from Mordred and PaDEL in their ability to characterize the chemo-structural composition of 53 compounds from the NNRTI drug class and a database of FDA-approved drugs targeting the HIV-1 RT enzyme. We then built a machine learning model based on logistic regression to classify which molecules are NNRTIs based on salient descriptors from each software and compared the relative performance of the models built using each set of descriptors using a variety of metrics such as accuracy, precision, recall, and F1 scores. Mordred has noted improvements over PaDEL, in which many descriptors have invalid outputs. We therefore hypothesized that the descriptor data generated by Mordred would be more accurate and provide higher fidelity results in our SAR study of NNRTIs and the HIV-1 RT, given that invalid values would not be present or have to be accounted for in our study. We found that Mordred was the superior descriptor set in this task. The approach outlined in this work can be broadly applied as a process template to help identify hit compounds and improve the throughput and efficiency of the drug discovery pipeline.

Machine learning and statistical learning techniques have improved the success of HTVS by providing a data-driven algorithmic approach that does not require hand-crafted rules and manual thresholds by leveraging the power from distributed computing. Molecule classification is an essential segment of HTVS to estimate the identity of a molecule without having to run extensive simulation algorithms. A common and effective algorithm is Logistic regression (17). Logistic regression takes multi-dimensional data and uses the features to estimate the probability of the data point falling into a particular class. While a simple accuracy metric can be used to evaluate the performance of a logistic regression algorithm, it overlooks the amount of data available for each class. Because of this, confusion matrices and F1 scores are used to analyze the performance of classification models. A confusion matrix provides a visualization of the information that the F1 score encompasses, which is the precision and recall of the model.

Dimensionality reduction is often needed when working with several dimensions, in both helping the human visualize the data being worked with and in simplifying the data by removing redundancies. A popular linear dimensionality reduction technique is the Principal Component Analysis (PCA) (21). PCA is a fast, unsupervised technique for analyzing the variance structure of a high dimensional dataset. However, it

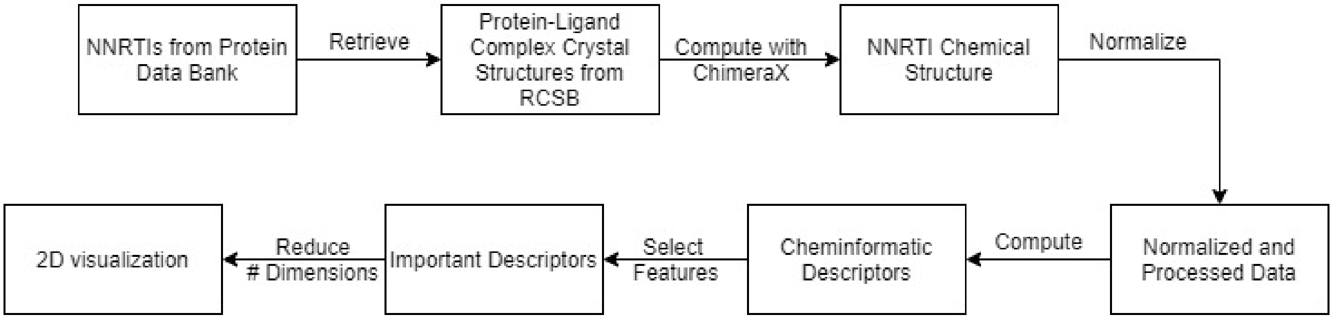

Figure 1. General algorithm used to train the machine learning algorithm to identify NNRTIs and the defining features of NNRTIs. NNRTIs showing bioactivity in HIV-1 RT are extracted from crystal structures in PDB, cheminformatic descriptors are computed, and salient features of NNRTIs are calculated through the machine learning algorithm.

only focuses on keeping the low dimensional representations of dissimilar points far apart. Therefore, we used t-Distributed Stochastic Neighbor Embedding (t-SNE). t-SNE measures and matches the distances between any two points in the high dimensional space with the corresponding points in the lower-dimensional subspace (22). t-SNE is primarily used to visualize high dimensional data by mapping the points into two- or three-dimensional subspaces that can be plotted to analyze the distribution of the data visually since it preserves both the local and global structure of the data.

RESULTS

Using Python, we performed a logistic regression comparing 53 NNRTI molecules and a dataset of FDA-approved drugs to obtain 15 chemical properties that were found to be essential in differentiating whether a molecule was an NNRTI or not (**Figure 1**). Both Mordred and PaDEL were used as molecular descriptor software, and a variety of descriptors were shown to be influential with slight differences between these outlier descriptors dependent on the software used. Twenty-seven attributes from Mordred were important in determining whether a molecule was an NNRTI. Of those, the 15 most salient were MoeType, Aromatic, BondCount, CPSA, ExtendedTopochemicalAtom, CarbonTypes, EState, LogS, TopologicalIndex, Lipinski, RingCount, BaryszMatrix, BCUT, AcidBase, and Autocorrelation. Out of all PaDEL's attributes, 26 were shown to be important in determining an NNRTI. Of those 26, the 15 most salient were: Aromatic atoms count, Aromatic bonds count, Information content, Charged, partial surface area, Path counts, BaryszMatrix, Molecular linear free energy relation, Ring Count, Detour matrix, Petitjean number, Topological polar surface area, Autocorrelation, RDF, Burden modified eigenvalues, and Atom type electrotopological state. However, many of these properties are computational parameters, such as BCUT, that are not easily interpretable in a chemical sense and are instead quantities developed through specific calculations and matrices created by the software. Discrepancies in notable attributes were due to differences in the descriptors that each software calculates and potentially inaccurate data provided by these molecular descriptor tools. Thus, we sought to determine the most accurate tool.

The logistic regression trained on PaDEL descriptors had a 96.7 % classification accuracy and an F1 score of 0.9 while the regression trained on the PaDEL descriptors had a prediction accuracy of 100% and an F1 score of 1 indicating that logistic regression model trained on the Mordred descriptors was better than the PaDEL descriptors **(Figure 2)**. Since both models were trained with the optimal hyperparameters for each descriptor set, the nature of the logistic regression algorithm shows that there does not exist a linear relationship that correctly classifies all the molecules for the PaDEL descriptor set. The increase in prediction accuracy and the decrease in logistic likelihood loss were both statistically significant when using the Mordred descriptors as opposed to PaDEL descriptors with a significance level of 0.01. We generated t-SNE visualizations of between 150 and 200 PaDEL and Mordred descriptors (**Figure 3**), which were determined to be significant in classifying a molecule as an NNRTI. The 53 NNRTI molecules were visually distinguishable as a separate cluster from the other drugs when the filtered Mordred descriptors were reduced and plotted (**Figure 3B**), while the plot with the filtered PaDEL descriptors did not show any significant distinction between NNRTIs and other drugs (**Figure 3A**).

DISCUSSION

Our experiment was a comparison of the accuracy of PaDEL and Mordred's molecular descriptors in identifying structural trends among NNRTIs. We performed our experiment by first searching for and processing known NNRTIs into the Protein DataBank (PDB) format. The dummy dataset was gathered from the CureFFI dataset of FDA-approved drugs. We used the feature weights computed by a logistic regression model trained to predict NNRTIs from the dummy dataset to extract important descriptors. Then, in order to assess the success of extracting the correct weights, we used t-SNE to visualize the separation between the NNRTIs and other molecules. We found that the PaDEL descriptor set is inferior to the Mordred set in classifying NNRTIs from other drugs.

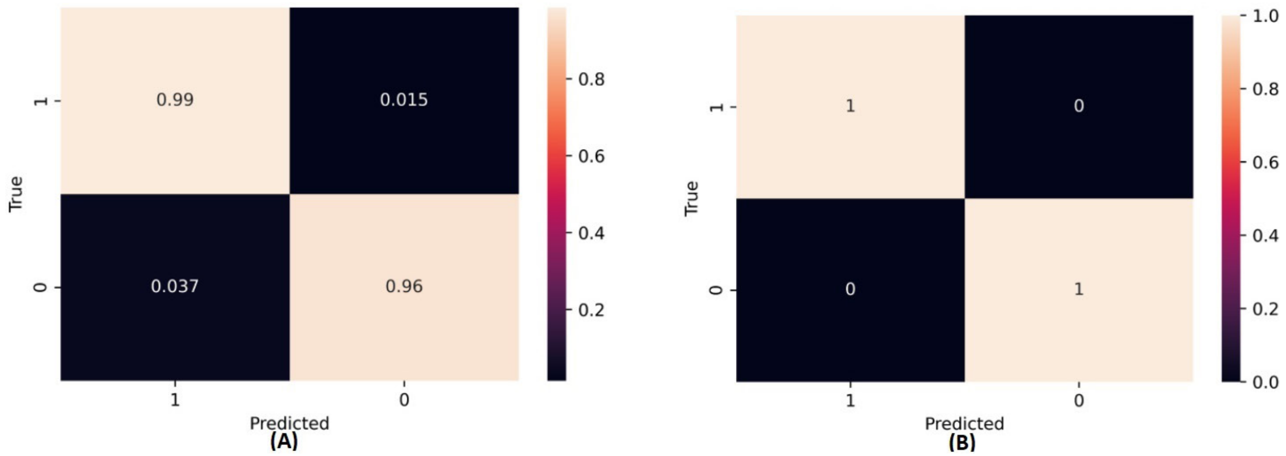

Figure 2. Confusion matrices of the logistic regression performed. A) PaDEL and **B)** Mordred. datasets. The y-axis shows whether the molecule was an NNRTI (1) or not (0), and the x-axis shows the regression's predicted classification of the molecule. The logistic regression performed better on the Mordred dataset, in which is classifies all the molecules correctly, while the model trained on the PaDEL dataset performs worse.

A significant issue found with the PaDEL descriptor set is that not all descriptors are available for all molecules. Performing data analytics on an unavailable descriptor set either directly introduces error or reduces the amount of data available for inference, which indirectly decreases the confidence in the inferences. Overall, Mordred was found to be the superior descriptor set to perform data analytics due to its superiority in representing a linear relationship in classifying molecules and its data quality.

The methods used for filtering the descriptors were inspired by the fundamental properties of statistical learning. From a data analytics perspective, the task at hand is a data compression problem, where given thousands of features about a certain data sample, the most important features need to be extracted. In this case, the measure of importance is the feature's utility in classifying whether a molecule is an NNRTI or not. Because of this, a directed, supervised approach to dimensionality reduction can be used in which a classification algorithm is used to classify a set of predefined NNRTIs against a dummy dataset. In this case a dataset of common drugs is used and the features that have higher weights associated with them in classification are significant. The choice of logistic regression as the classification algorithm was due to the importance of interpretability of the model, and specifically being able to infer which features were important. The initial assumption of a linear relationship

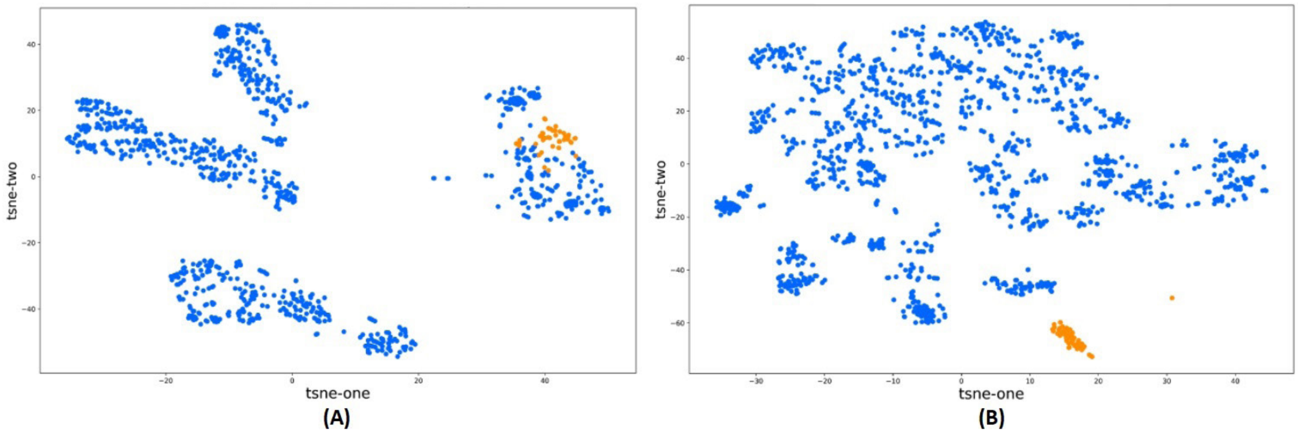

Figure 2. t-SNE visualizations of the filtered descriptors. A) PaDEL and **B)** Mordred. Orange points represent NNRTI molecules and blue points represent molecules that are not NNRTIs. The clustering displays the high dimensional structure of the data in 2D and allows the analysis of the distribution of each dataset. The NNRTI molecules are closely clustered together and separate from the other common drugs in the t-SNE visualization of the Mordred dataset, while the t-SNE visualization of the PaDEL dataset does not show this separation. This shows how the Mordred dataset has more information to differentiate between NNRTI molecules and other common drugs. The non-NNRTI clustering for the Mordred dataset shows that the Mordred dataset has descriptors that can split the common drugs into several groups (clusters) while the non-NNRTI clustering for the PaDEL dataset has fewer, larger groups of molecules. The Mordred dataset appears to match the variety in the common drugs better than the PaDEL dataset.

between the log-odds and the features allows the feature weights of logistic regression to be easily interpreted, as larger weights on a feature directly correspond to a higher impact of that feature in determining the classification outcome. A simple interquartile range-based outlier detection algorithm run on the weights can be used to get the upper outliers of features that were weighted highly. The simplicity of logistic regression paired with its easy interpretability made it the best classification algorithm to use for dimensionality reduction to get the most important features, or descriptors, in classifying NNRTIs. Once the important features are determined by the logistic regression-based dimensionality reduction algorithm, these features can be used in a clustering algorithm, such as a DBSCAN, on a large dataset of synthetically accessible molecules to find molecules with features similar to NNRTIs (23).

Because of the complexity associated with chemical molecules, a more powerful dimensionality reduction technique can be used in the future to more accurately model the distribution of chemical molecules in general. We can use the same logistic regression algorithm used in this paper and then run DBSCAN to find similar molecules. An effective and powerful method of dimensionality reduction is a neural network-based autoencoder, which would be trained on a massive dataset of chemical molecules to learn their distribution and the distribution of the descriptor set (24). An autoencoder would be able to directly consider the chemical structure and composition of the molecule and learn the salient features from the descriptor set, which would allow it to learn a more accurate representation of the data. Once a more accurate representation of the chemical molecule is obtained, DBSCAN can be used to find molecules with similar features as NNRTIs. Gaussian mixture models (GMMs) can augment autoencoders by considering labeled data instead of learning from unlabeled data (25). GMMs can analyze the 53 labeled NNRTI molecules to output higher fidelity inferences on the likelihood that a molecule is an NNRTI.

Significant attributes by average descriptor weight from Mordred provided insights into the pharmacophoric relevance of the structure of NNRTIs. By averaging the weights of the descriptors comprising each of the attributes, the Mordred data indicated that, from most important to least important, the MoeType, Aromatic, BondCount, CPSA, ExtendedTopochemicalAtom, CarbonTypes, EState, LogS, TopologicalIndex, Lipinski, RingCount, BaryszMatrix, BCUT, AcidBase, and Autocorrelation attributes of the NNRTI compounds were important in determining the chemical composition of NNRTIs. From these attributes, we extrapolated information about structural trends in NNRTI compounds that may enable their ability to inhibit the HIV-1 RT enzyme. It is important to note that the Autocorrelation was not considered a part of the structure of the NNRTIs, so the descriptors were discounted from our analysis.

The n5ring and n6ring attributes in Mordred compute the number of five- and six-membered rings, respectively. Essentially, the NNRTI compounds seemed to have at least one or the other, with the average of NNRTIs with five membered rings being about 0.5 rings and the average for six membered rings being 2.5 rings. From the perspective of ligand-receptor interactions, the presence of these potentially aromatic cyclic structures, which are known to be the most stable ring structures that can form from organic compounds due to their relatively low ring strain of 6.1 kcal/mol (five-membered rings) and 0.1 kcal/mol (six-membered rings), can contribute to the formation of stable intermolecular interactions (18). These aromatic ring moieties can induce π-π stacking interactions with aromatic residues in the HIV-1 RT enzyme, which potentially allow for the blocking of DNA polymerization (19). Given that NNRTI compounds bind in a hydrophobic area of RT, we hypothesized that the nonpolar amino acids present in the RT enzyme, including aromatic residues, may interact through hydrophobic interactions with NNRTI compounds to induce conformational changes in the protein that inhibit protein activity (18).

The electrotopological state of the atom, or EState, was shown to be similar across many of the analyzed NNRTIs. Several Mordred descriptors are considered in the determination of the EState of molecules. Relatively similar EState values indicate similarities in both the electronic structure and the physical structure of the molecule. Our algorithm yielded several descriptors with relatively low deviations within the specific type of descriptor itself within EState, with the majority of relevant EState descriptors having 0 % variance from their mean. The other descriptors further revealed the similarities in EState between the NNRTIs, which is a multifaceted way to define and represent molecular structures and has been used in various QSAR studies across both biochemistry and biological activity of small molecules (20).

Using cheminformatics augmented with a machine learning approach, our methodology can identify the chemical properties which are vital to NNRTIs. This insight can be leveraged broadly with other classes of molecules. For example, researchers applied molecular fingerprinting data in a machine learning algorithm to identify novel Janus Kinase 2 (JAK2) inhibitors (26). It also allows more developments into the future of machine learning. The methodology described here can be implemented to identify any novel NNRTIs in any given dataset, which has a similar structure or possesses similar activity. The approach can be extended to drug discovery campaigns beyond just HIV-1 to all types of interactions between a small molecule and protein target. This technique of structural analysis can be applied to organize compounds that are effective for use as potential drugs as well as recognize key components and structural trends in various compounds. The approach outlined in this paper can be used to make synthetic chemistry and the production of novel or hit compounds for any protein drug screening more efficient.

MATERIALS AND METHODS

To obtain a list of training compounds for our algorithm, we conducted a search for NNRTIs on the Protein Data Bank (27). The filters for our search were that the ligand must be bound to the original non-mutated HIV-1 RT protein and not form a complex with DNA. We then compiled a list of 53 NNRTI HIV-1 RT complex crystal structures from the RCSB and using the software UCSF ChimeraX (release 1.1), extracted the NNRTIs from the complex (28-46). This was accomplished through the deletion of water and other undesired atoms bound to the reverse transcriptase protein, followed by the deletion of the enzyme to yield the cleaned ligand or NNRTI. This was used as our initial group for developing the descriptors for NNRTIs. To obtain a list of all two-dimensional (2D) descriptors and some 3D descriptors of each ligand, we fed the clean NNRTI file into two different software, Mordred (version 1.20) and PaDEL (version 2.21) (15, 16). Analyzing the results from both software programs yielded a list of descriptors used to uniquely identify a compound as an NNRTI (detailed further below). In order to corroborate our results from this training dataset and demonstrate the correlation between the descriptors obtained from our filter method, we then compiled a test dataset of 1,691 molecules, consisting of our extracted NNRTIs and the CureFFI dataset of common FDA-approved drugs (47). The FDA-approved drugs were used as a dummy dataset for our NNRTIs, as these drugs were a good match for the total scope of potential drugs. Once again, Mordred and PaDEL were used to generate a .csv file of all the descriptors of the objects. Computational comparison of the descriptors of the FDA dataset against the descriptors determined to be relevant by the training dataset of 53 NNRTIs was then conducted.

The original data obtained from Mordred and PaDEL had missing values and had varying feature magnitudes. To combat this so that we could obtain more effective results, we normalized the features by dividing each feature by the largest occurrence of the feature so that the values would be strictly between 0 and 1. We performed a K-Nearest-Neighbors based imputation where the values from similar data points were substituted in for the missing values. This normalization strategy allowed a direct comparison of the weights of the logistic regression without having to take into account the initial magnitude of the calculated descriptors.

A logistic regression model was built with the data, with the 55 NNRTIs being the true positive labels and the CureFFI dataset as the true negative labels. This was implemented in Tensorflow (version 2.40), an end-to-end open-source platform for machine learning (48). The log-odds of an observation was expressed as a linear function of the *K* input variables x, as can be seen in **Equation 1:**

$$\log\left(\frac{P(X)}{1-P(X)}\right) = \sum_{j=0}^{K} w_j x_j$$

Solving for *P(X)*, which is the probability that a data sample is a certain class, yields:

$$P(X) = \frac{\exp(z)}{1+\exp(z)}$$

Where *z* is the right-hand side of **Equation 1**. The solution to logistic regression is the representation of *P(X)* that maximizes the likelihood of the data, which is the product of the predicted probabilities of *N* data samples. In practice, evaluating that many products cause errors due to lack of numerical precision, so the products are converted into sums by maximizing the natural log of the likelihood of the data. This can be done because the logarithm function increases monotonously. L2 regularization was also added, which adds the two norms of the weight vector multiplied by a coefficient, in this case, 1, which yields a final loss function as follows:

$$L(X|P) = \log P(x_i) + \sum_{i=1, y_i=0}^{N} \log\big(1 - P(x_i)\big) + ||w||_2$$

The weight vector *w* was initialized according to a random normal distribution to have a unique non-zero gradient for each feature weight. **Equation 3** was minimized by the Limited Memory Broyden–Fletcher–Goldfarb–Shanno algorithm, which is a quasi-Newtonian optimization algorithm based on the Hessian matrix (49, 50).

Once the logistic regression was computed, a confusion matrix was created, and the models' F1 scores were calculated to analyze the performance of the model. Confusion matrices plot the predicted classes against the labels, and the F1 score is the harmonic mean of the precision, or the proportion of the data points our model says were in a class that was actually in that class. We also determined our model's recall, or its ability to find all the data points that were a positive class in a dataset. Then, the distribution of the squares of the weights was examined, and the upper outlier weights were determined using a threshold of the third quartile of the data plus 1.5 times the interquartile range. Any weights above this threshold were marked as important descriptors in determining whether a molecule is an NNRTI or not.

The t-SNE algorithm was run on only the descriptors deemed important by the logistic regression-IQR algorithm to visualize the success in emphasizing the unique descriptors in NNRTIs. The high dimensional Euclidean distances between data points were converted into conditional probabilities. These probabilities represent the similarity of datapoint x_j to datapoint x_i using the conditional probability $p_{j|i}$ that x_i would pick x_j as its neighbor. Neighbors were picked according to their probability density under a t-distributed probability density function with one degree of freedom centered at x_i.

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

A similar function $q_{j|i}$ was calculated for the low dimensional counterparts of x_i and x_j, y_i and y_j.

$$q_{j|i} = \frac{\exp\left(-||y_i - y_j||^2/2\sigma_i{}^2\right)^{-1}}{\sum_{k\neq i}\exp\left(-||y_i - y_k||^2/2\sigma_i{}^2\right)^{-1}}$$

For the low dimensional counterparts of x_i and x_j to correctly model the similarity between the high dimensional data, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ must be equal. A measure of the effectiveness of $q_{j|i}$ modeling $p_{j|i}$ is the Kullback-Leibler divergence, an information-based measure of disparity among probability distributions (51). The cost function thus consisted of the sum of the Kullback-Leibler divergences over all data points.

This cost function was minimized using Gradient Descent (52).

$$C = \sum_i KL(P||Q) = \sum_i \sum_i p_{j|i} \log\left(\frac{p_{j|i}}{1_{j|i}}\right)$$

ACKNOWLEDGEMENTS

We would like to acknowledge and thank the developers and authors of all open-source packages used in our research. Molecular graphics and analyses were performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from the National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

The authors declare no competing conflicts of interests in the work presented. All code will be provided upon request. We gratefully acknowledge the Aspiring Scholars Directed Research Program for funding for our research.

Received: May 29, 2021
Accepted: June 10, 2022
Published: June 24, 2022

REFERENCES

1. Karkoutly, Omar, *et al*. “Molecular Modelling a Key Method for Potential Therapeutic Drug Discovery.” Biomedical Journal of Scientific & Technical Research, vol. 37, no. 3, 2021, https://doi.org/10.26717/bjstr.2021.37.006000.
2. Morgan, Steve, *et al*. “The Cost of Drug Development: A Systematic Review.” Health Policy, vol. 100, no. 1, 2011, pp. 4–17., https://doi.org/10.1016/j.healthpol.2010.12.002.
3. Salmaso, Veronica, *et al*. “Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview.” Frontiers in Pharmacology, vol. 9, 2018, https://doi.org/10.3389/fphar.2018.00923.
4. Goodsell, D.S. “Reverse Transcriptase.” *RCSB Protein Data Bank*, 2002, doi:10.2210/rcsb_pdb/mom_2002_9.
5. Das, K., *et al*. “HIV-1 Reverse Transcriptase Structures.” *Encyclopedia of Biological Chemistry*, 2013, pp. 548–553., doi:10.1016/b978-0-12-378630-2.00247-4.
6. Andréola, M.-L., *et al*. “DNA Polymerases: Reverse Transcriptase Integrase, and Retrovirus Replication.” *Encyclopedia of Biological Chemistry*, 2013, pp. 101–107., doi:10.1016/b978-0-12-378630-2.00258-9.
7. Bhagavan, N.V., *et al*. “DNA Replication, Repair, and Mutagenesis.” *Essentials of Medical Biochemistry*, 2015, pp. 401–417., doi:10.1016/b978-0-12-416687-5.00022-1.
8. Joly, V, *et al*. “Inhibiteurs non nucléosidiques de la transcriptase inverse” [Non-nucleoside reverse transcriptase inhibitors]. *Annales de medecine interne* vol. 151,4 (2000): 260-7.
9. Sood, Shivani, *et al*. “Drug design strategies using non nucleoside reverse transcriptase inhibitors (NNRTI): current challenges and future perspectives.”, 1, 1-12.
10. Zhuang, Chunlin, *et al*. “Development of Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs): Our Past Twenty Years.” *Acta Pharmaceutica Sinica B*, vol. 10, no. 6, 2020, pp. 961–978., doi:10.1016/j.apsb.2019.11.010.
11. Xu, Jun, *et al*. “Chemoinformatics and Drug Discovery.” Molecules, vol. 7, no. 8, 2002, pp. 566–600., https://doi.org/10.3390/70800566.
12. Xue, Ling, *et al*. “Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening.” Combinatorial Chemistry & High Throughput Screening, vol. 3, no. 5, 2000, pp. 363–372., https://doi.org/10.2174/1386207003331454.
13. Torres, Pedro H. M. *et al*. “Key Topics in Molecular Docking for Drug Design.” International Journal of Molecular Sciences 20.18 (2019): 4574. Crossref. Web.
14. Jun Xu, *et al*. “Chemoinformatics and Drug Discovery.” Molecules, vol. 7, no. 8, 2002, pp. 566–600., doi:10.3390/70800566.
15. Yap, Chun Wei. “Padel-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints.” Journal of Computational Chemistry, vol. 32, no. 7, 2010, pp. 1466–1474., https://doi.org/10.1002/jcc.21707.
16. Moriwaki, Hirotomo, *et al*. “Mordred: A Molecular Descriptor Calculator.” Journal of Cheminformatics, vol. 10, no. 1, 2018, https://doi.org/10.1186/s13321-018-0258-y.
17. Talevi, Alan *et al*. “Machine Learning in Drug Discovery and Development Part 1: A Primer.” CPT: pharmacometrics & systems pharmacology vol. 9,3 (2020): 129-142. doi:10.1002/psp4.12491
18. Pagni, Richard. “Modern Physical Organic Chemistry (Eric V. Anslyn and Dennis A. Dougherty).” Journal of Chemical Education, vol. 83, no. 3, 2006, p. 101., https://doi.org/10.1021/ed083p387.
19. Sluis-Cremer, Nicolas, *et al*. “Conformational Changes in HIV-1 Reverse Transcriptase Induced by Nonnucleoside Reverse Transcriptase Inhibitor Binding.” Current HIV

Research, vol. 2, no. 4, 2004, pp. 323–332., https://doi.org/10.2174/1570162043351093.

20. Hall, Lowell H., *et al*. "The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs." Journal of Chemical Information and Modeling, vol. 31, no. 1, 1991, pp. 76–82., https://doi.org/10.1021/ci00001a012.
21. Giuliani, Alessandro. "The Application of Principal Component Analysis to Drug Discovery and Biomedical Data." Drug Discovery Today, vol. 22, no. 7, 2017, pp. 1069–1076., doi:10.1016/j.drudis.2017.01.005.
22. Van der Maaten, Laurens, *et al*. "Visualizing Data Using t-SNE." Journal of Machine Learning Research, vol. 9, no. 86, 2008, pp. 2579–2605.
23. Ester, Martin, *et al*. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.
24. Hinton, G. E., *et al*. "Reducing the Dimensionality of Data with Neural Networks." Science, vol. 313, no. 5786, 2006, pp. 504–507., https://doi.org/10.1126/science.1127647
25. Reynolds, Douglas. "Gaussian Mixture Models." Encyclopedia of Biometrics, 2009, pp. 659–663., https://doi.org/10.1007/978-0-387-73003-5_196.
26. Yang, Minjian, *et al*. "Machine Learning Models Based on Molecular Fingerprints and an Extreme Gradient Boosting Method Lead to the Discovery of JAK2 Inhibitors." Journal of Chemical Information and Modeling, vol. 59, no. 12, 2019, pp. 5002–5012., https://doi.org/10.1021/acs.jcim.9b00798.
27. Berman, H. M. "The Protein Data Bank." Nucleic Acids Research, vol. 28, no. 1, 2000, pp. 235–242., https://doi.org/10.1093/nar/28.1.235.
28. Chan, Joseph H., *et al*. "2-Amino-6-Arylsulfonylbenzonitriles as Non-Nucleoside Reverse Transcriptase Inhibitors of HIV-1." *Journal of Medicinal Chemistry*, vol. 44, no. 12, 2001, pp. 1866–1882., doi:10.1021/jm0004906.
29. Das, Kalyan, *et al*. "High-Resolution Structures of HIV-1 Reverse Transcriptase/TMC278 Complexes: Strategic Flexibility Explains Potency against Resistance Mutations." Proceedings of the National Academy of Sciences, vol. 105, no. 5, 2008, pp. 1466–1471., https://doi.org/10.1073/pnas.0711209105.
30. Freeman, George A., *et al*. "Design of Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase with Improved Drug Resistance Properties. 2." *Journal of Medicinal Chemistry*, vol. 47, no. 24, 2004, pp. 5923–5936., doi:10.1021/jm040072r.
31. Himmel, Daniel M., *et al*. "Crystal Structures for HIV-1 Reverse Transcriptase in Complexes with Three Pyridinone Derivatives: A New Class of Non-Nucleoside Inhibitors Effective against a Broad Range of Drug-Resistant Strains." *Journal of Medicinal Chemistry*, vol. 48, no. 24, 2005, pp. 7582–7591., doi:10.1021/jm0500323.
32. Hopkins, Andrew L., *et al*. "Design of Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase with Improved Drug Resistance Properties. 1." *Journal of Medicinal Chemistry*, vol. 47, no. 24, 2004, pp. 5912–5922., doi:10.1021/jm040071z.
33. Hopkins, Andrew L., *et al*. "Design of Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase with Improved Drug Resistance Properties. 1." *Journal of Medicinal Chemistry*, vol. 47, no. 24, 2004, pp. 5912–5922., doi:10.1021/jm040071z.
34. Ren, Jingshan, *et al*. "Binding of the Second Generation Non-Nucleoside Inhibitor S-1153 to HIV-1 Reverse Transcriptase Involves Extensive Main Chain Hydrogen Bonding." *Journal of Biological Chemistry*, vol. 275, no. 19, 2000, pp. 14316–14320., doi:10.1074/jbc.275.19.14316.
35. Ren, Jingshan, *et al*. "Structural Basis for the Resilience of Efavirenz (DMP-266) to Drug Resistance Mutations in HIV-1 Reverse Transcriptase." *Structure*, vol. 8, no. 10, 2000, pp. 1089–1094., doi:10.1016/s0969-2126(00)00513-x.
36. Rodgers, D. W., *et al*. "The Structure of Unliganded Reverse Transcriptase from the Human Immunodeficiency Virus Type 1." *Proceedings of the National Academy of Sciences*, vol. 92, no. 4, 1995, pp. 1222–1226., doi:10.1073/pnas.92.4.1222.
37. Ren, Jingshan, *et al*. "Crystal Structures of HIV-1 Reverse Transcriptase in Complex with Carboxanilide Derivatives." Biochemistry, vol. 37, no. 41, 1998, pp. 14394–14403., https://doi.org/10.1021/bi981309m.
38. Hopkins, Andrew L., *et al*. "Complexes of HIV-1 Reverse Transcriptase with Inhibitors of the HEPT Series Reveal Conformational Changes Relevant to the Design of Potent Non-Nucleoside Inhibitors." *Journal of Medicinal Chemistry*, vol. 39, no. 8, 1996, pp. 1589–1600., doi:10.1021/jm960056x.
39. Hopkins, Andrew L., *et al*. "Design of Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase with Improved Drug Resistance Properties. 1." *Journal of Medicinal Chemistry*, vol. 47, no. 24, 2004, pp. 5912–5922., doi:10.1021/jm040071z.
40. Kuroda, Daniel G., *et al*. "Snapshot of the Equilibrium Dynamics of a Drug Bound to HIV-1 Reverse Transcriptase." *Nature Chemistry*, vol. 5, no. 3, 2013, pp. 174–181., doi:10.1038/nchem.1559.
41. Mislak, Andrea C., *et al*. "A Mechanistic and Structural Investigation of Modified Derivatives of the Diaryltriazine Class of NNRTIs Targeting HIV-1 Reverse Transcriptase." *Biochimica Et Biophysica Acta (BBA) - General Subjects*, vol. 1840, no. 7, 2014, pp. 2203–2211., doi:10.1016/j.bbagen.2014.04.001.
42. Mislak, Andrea C., *et al*. "A Mechanistic and Structural Investigation of Modified Derivatives of the Diaryltriazine Class of NNRTIs Targeting HIV-1 Reverse Transcriptase." *Biochimica Et Biophysica Acta (BBA) - General Subjects*, vol. 1840, no. 7, 2014, pp. 2203–2211., doi:10.1016/j.bbagen.2014.04.001.

43. Ren, Jingshan, *et al.* "Crystal Structures of HIV-1 Reverse Transcriptase in Complex with Carboxanilide Derivatives†,‡." *Biochemistry*, vol. 37, no. 41, 1998, pp. 14394–14403., doi:10.1021/bi981309m.
44. Ren, Jingshan, *et al.* "Crystallographic Analysis of the Binding Modes of Thiazoloisoindolinone Non-Nucleoside Inhibitors to HIV-1 Reverse Transcriptase and Comparison with Modeling Studies." *Journal of Medicinal Chemistry*, vol. 42, no. 19, 1999, pp. 3845–3851., doi:10.1021/jm990275t.
45. Pettersen, Eric F., *et al.* "UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers." Protein Science, vol. 30, no. 1, 2020, pp. 70–82., https://doi.org/10.1002/pro.3943.
46. Goddard, Thomas D., *et al.* "UCSF Chimerax: Meeting Modern Challenges in Visualization and Analysis." Protein Science, vol. 27, no. 1, 2017, pp. 14–25., https://doi.org/10.1002/pro.3235
47. Minikel, Eric. "List of FDA-Approved Drugs and CNS Drugs with SMILES." List of FDA-Approved Drugs and CNS Drugs with Smiles, 2013, http://www.cureffi.org/2013/10/04/list-of-fda-approved-drugs-and-cns-drugs-with-smiles/.
48. Abadi, Martin, *et al.* "Tensorflow: A System for Large-Scale Machine Learning." USENIX, 2016, https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.
49. Byrd, R. H., *et al.* "A Stochastic Quasi-Newton Method for Large-Scale Optimization." SIAM Journal on Optimization, vol. 26, no. 2, 2016, pp. 1008–1031., https://doi.org/10.1137/140954362.
50. Andrew, Galen, *et al.* "Scalable Training ofl1-Regularized Log-Linear Models." Proceedings of the 24th International Conference on Machine Learning - ICML '07, 2007, https://doi.org/10.1145/1273496.1273501. Kullback, S., and R. A. Leibler. "On Information and Sufficiency." *The Annals of Mathematical Statistics*, vol. 22, no. 1, 1951, pp. 79–86., doi:10.1214/aoms/1177729694.
51. Robbins, Herbert, *et al.* "A Stochastic Approximation Method." *The Annals of Mathematical Statistics*, vol. 22, no. 3, 1951, pp. 400–407., doi:10.1214/aoms/1177729586.,