WASOT
www.wasot2012.org

Anireh V.I.E and Osegi E. N. Toxicology Digest Vol. 2 (1): 1-12 (2018)

Toxicology
Digest

An Online Cortical Machine Learning Artificial Intelligence Technique for Drug Discovery

Anireh, V.I.E PhD[1,*] and Osegi, E.N[2]

[1] **Department of Computer Science, Rivers State University, Port-Harcourt**
[2] **Department of Information and Communication Technology, National Open University of Nigeria.**

Abstract

Bioinformatics deals with the analysis and interpretation of biological data by using tools of information science. Drug discovery prediction which is a process of discovering new candidate medications from some molecular compounds has challenged professionals in the field of medical sciences. Tools that have assisted in drug discovery and have been reported by researchers includes the use of decision trees, induction programming logic, expert systems and supervised neural networks. In this research paper, we propose an approach to the drug discovery and prediction problem using a variant of an unsupervised online cortical machine learning artificial intelligence technique. The approach has an explicit tuning parameter called the relaxation factor used in determining possible new candidate sequence. Experiments on a popular DNA sequence dataset and a reversed-phase high-performance liquid chromatography (RP-HPLC) drug dataset were performed to determine whether the proposed technique can give effective predictions. The results showed that the approach compares favorably with the other methods reported in the literature but has a more promising performance when it is set to lower relaxation values.

Keywords: *Artificial Intelligence; bioinformatics; drug discovery prediction; DNA sequence; online cortical machine learning; relaxation factor; RP-HPLC*

Introduction

Drug discovery involves the determination of candidate drugs or chemical compounds that bind to a more stable and recognized expectation compound. Over the past three decades, drug discovery researches are being carried out by many scholars across the globe. Notable works include the use of inductive logic programming and machine learning to model Quantitative Structure-Activity Relationships (QSAR) of a class of drugs [1-3], the use of Knowledge Based Artificial Neural Networks (KBANNs) to model promoters in DNA sequences [4] and the use of ANNs to predict chromatographic retention of drugs based

[1,*] Corresponding Author

on certain predefined structural descriptors [5]. In order to accurately model QSARs, such approaches should as matter of requirement make effective decisions on the promoters or non-promoters of DNA sequences. Thus, it has become a key requirement for practicing pharmacologists and medical experts alike to use artificial intelligence tools to leverage their understanding of the composition and correlatedness of chemical compounds that aid drug discovery. This research study proposes a new artificial intelligence tool called HTM-MAT that will assist in this regard. HTM-MAT is a software framework inspired by the way the cortex functions, an idea derived from how the mammalian brain processes sensory information or data.

Recent works on drug discovery

Some recent studies have been conducted in relation to drug discovery. We categorize these researches under two headings namely drug discovery using conventional Artificial Intelligence (AI) and that using generative AI models.

Drug Discovery Prediction Using Conventional AI:

Recent researches using Artificial Intelligence (AI) include the works in [8] where genetic algorithms combined with ANNs were used to evolve a set of drug dataset in order to effectively predict a set of pharmacokinetic parameters and in [9] where ANNs were used to predict binding energies which are based on physicochemical molecular descriptions of certain selected drugs. However, as stressed in Kustrin and Beresford [10], the capacity of conventional ANNs is still very limited when compared to that of the human brain. Thus, such AI techniques may not scale well for more challenging tasks.

Drug Discovery Using Generative AI models:

More recently, generative models have been proposed as a useful tool to facilitate drug discovery. One of such model is the Recurrent Neural Networks (RNN). RNNs can be trained on molecular structures in such a way that they generate predictive models that decipher the data generating distribution (Segler et al, 2017). For instance, an RNN can generate new molecules that bind towards a target; however, RNNs require heavy hyper-parameter tuning to attain a useful solution space (Cui et al, 2016). The RNN an approach to drug design is highly desirable but is beyond the scope of the current research to consider this technique.

Materials and Methods

In order to investigate the performance of the proposed technique, we have applied

HTM-MAT developed in [6]. The current version of HTM-MAT is based on the Cortical Learning Algorithms (CLA) proposed earlier in [7]. These algorithms are capable of online (continual) learning of streaming or sequential data with a temporal structure. It can also process spatial information as long as there is a way by which the examples (data inputs) can be read in sequentially. The default key parameters of HTM-MAT used in the present work are listed in Table 1. The HTM-MAT tool can be obtained from www.matlabcentral.com/ and includes all the methods and functions for conducting cortical based memory predictions on the aforementioned drug datasets.

The HTM algorithm used basically uses an intrinsic scoring metric called the 'overlap'. The overlap can be viewed as a metric that determines the magnitude of a matching pair i.e. the correlated-ness between an incoming sequence or sensory signal and a generative sequence. Using the overlap allows the formation of sparse distributed representations which is very important and efficient technique for information processing in the HTM. As an addition to this tool, we introduced a novel parameter or kernel operator called the relaxation factor for analyzing the DNA-sequence drug dataset that have been studied by previous researchers in [4]. The relaxation factor allows the HTM-MAT predictor to make inferences/decisions by computing a likelihood parameter or estimate using a sequence of succeeding matching prediction-test examples that meets a pre-specified relaxation threshold. The likelihood of a matching pair (obtained by extracting the target of a matched test example from the HTM-MAT memory) is computed as:

$$\rho_{likelihood} = \frac{\left(\sum_{i=1}^{N} \left(S_{pred_{(t\arg et)}} \equiv S_{test\_examples_{(t\arg et)}} \right) \right)}{N} \qquad (1)$$

The likelihood estimate can then be used for scoring the resulting performance of HTM-MAT meeting a relaxation factor requirement as:

$$S_{likelihood\_score} = \begin{cases} 1, & \rho_{likelihood} \geq \hbar_{thresh}, \{0.75 \leq \hbar_{thresh} \leq 1.0\} \\ 0, & otherwise \end{cases} \tag{2}$$

Note that S_{pred} and $S_{test\_examples}$ are assumed to be mixed-integer representations of the input-chain.

Using (1) and (2) it is easy to compute the error-rate as:

$$error_{rate} = \frac{\sum_{i=1}^{N} \left(S_{likelihood\_score} \equiv 0 \right)}{N} \tag{3}$$

Table1. Key parameters of the HTM-MAT used for the Simulations

HTM-MAT parameter/symbol	*Default Values*
No. of Monte Carlo Iterations (iters)	*10*
Minimum overlap (min_overlap)	*2*
Desired Local Activity (desired_localActivity)	*2*
Sequence Size (seq_size)	*200*
Percent Adjust (per_adjust)	*90*
Relaxation threshold (relax_thresh)	*0.90*

Figure 1 shows a flowchart of our proposed methodology. Further details of how the relaxation process is utilized are given in the experiments section. The process of drug prediction starts with the accumulation of data or information examples from a drug dataset in a sequential manner then these examples are predicted using the spatial-temporal pooler part of HTM-MAT sequence analyzer which generates a sparse distributed memory of predictions through time and space. A portion of the dataset examples is used for testing or validating the HTM-MAT sequence analyzer predictions with respect to a target class from which the percentage of correctly classified targets may be computed. This process continues until the number of examples or training iterations are completed.

Experiments and Results

Experiments were performed on a popular DNA sequence dataset and a reversed-phase high-performance liquid chromatography (RP-HPLC) drug dataset; the DNA dataset can be found in [4] while the RP-HPLC dataset is obtainable from [5].

Dataset description:

The DNA-sequence dataset contains 106 examples of which 53 samples are promoters and the other 53 samples non-promoters. The dataset target classes are (+) for promoters and (-) for non-promoters; these targets have been re-labeled to (2) and (1) for promoters and non-promoters respectively. All samples have been used selectively and independently for training and testing the examples in order to attain acceptable error rates.

The RP-HPLC dataset contains 52 samples of drug compounds of which one of the tasks is to predict the logarithm of HPLC retention factor of acid glycoprotein (AGP) column ($\log k_{AGP}$); in addition, 36 structural parameters including a target $\log k_{AGP}$, are defined quantitatively and serve as input to the HTM-MAT system. In accordance with the technique described in [5], 26 of these samples were randomly selected for training the HTM-MAT algorithm while 10 were used for testing.

For the DNA dataset, the error-rate is determined using the cross-validation procedure described in [4] while the RP-HPLC dataset is evaluated using the metric called the Mean Absolute Error (MAE). Figure 2 shows a scheme for processing the DNA sequence dataset and Figure 3 shows a possible transformation.

The results of HTM-MAT on the both datasets compared to that using the other techniques reported in [4, 5] are given in Table 2; the default settings of the HTM-MAT are used for the training/testing phase.

Further tests on the DNA sequence dataset have been performed on the basis of relaxation factor. These results showing the error rates at different relaxation factor and the likelihood plots are also given in Table 3 and Figure 4 respectively.

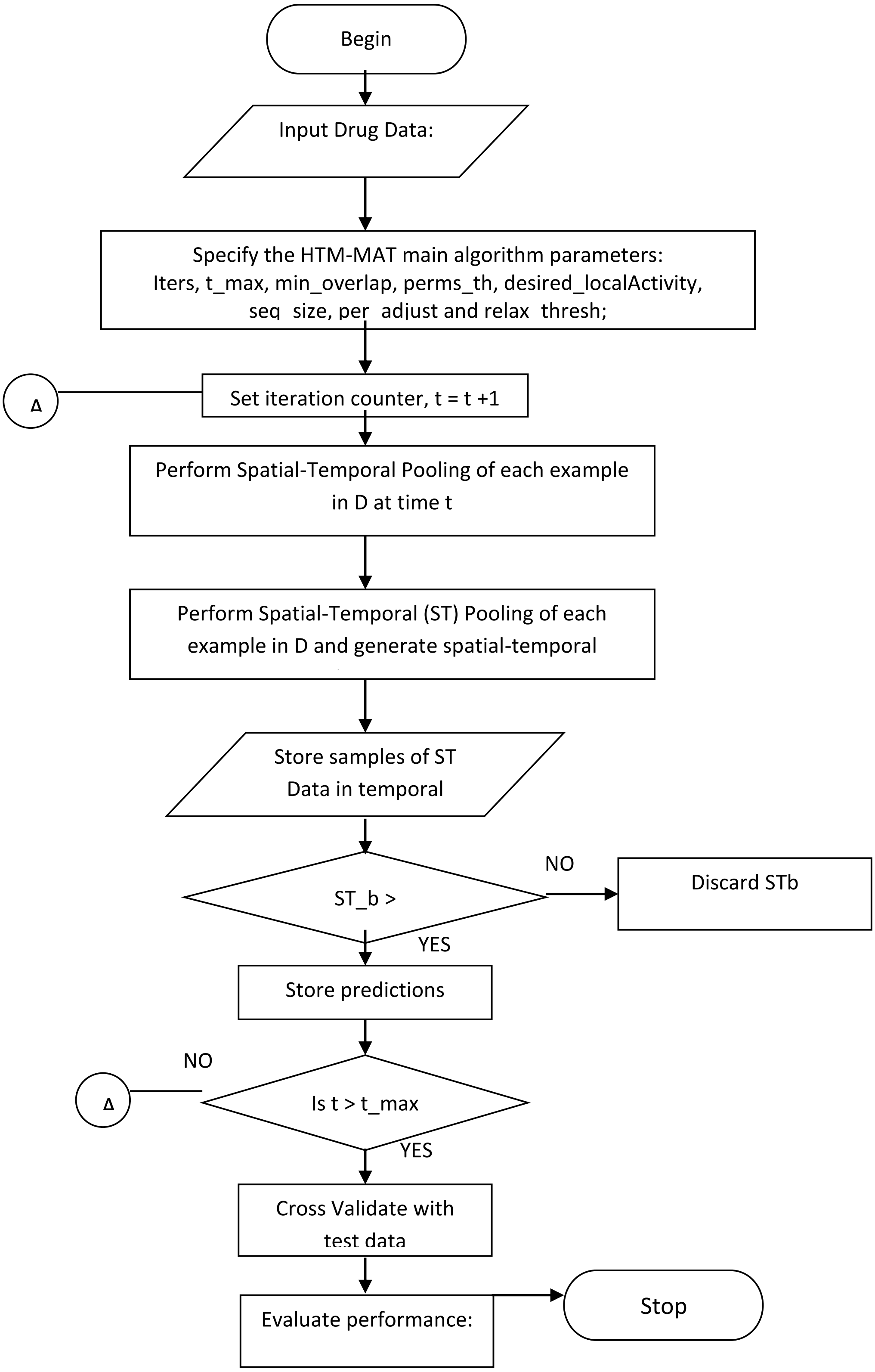
Begin
Input Drug Data:
Specify the HTM-MAT main algorithm parameters:
Iters, t_max, min_overlap, perms_th, desired_localActivity,
seq size, per adjust and relax thresh;
Δ
Set iteration counter, t = t +1
Perform Spatial-Temporal Pooling of each example in D at time t
Perform Spatial-Temporal (ST) Pooling of each example in D and generate spatial-temporal
Store samples of ST Data in temporal
ST_b >
NO
Discard STb
YES
Store predictions
NO
Δ
Is t > t_max
YES
Cross Validate with test data
Evaluate performance:
Stop

Fig.1. Flowchart of proposed technique for drug discovery prediction

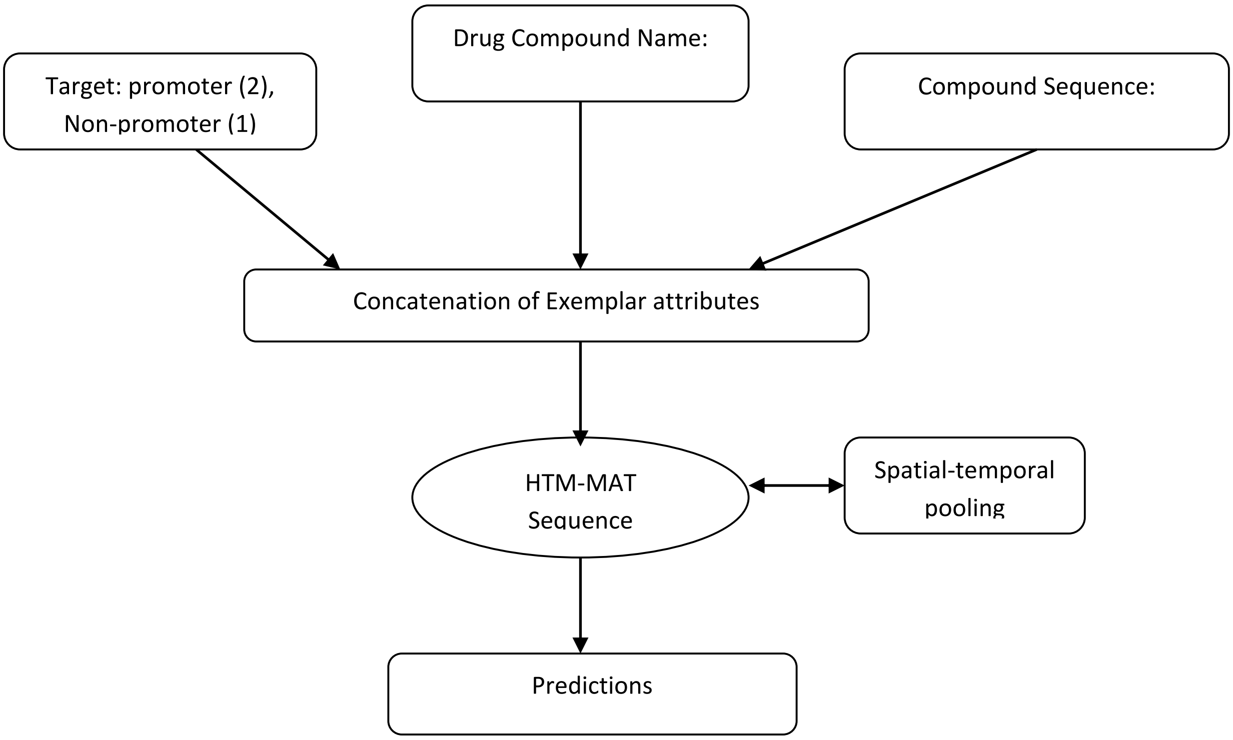

Fig. 2. A scheme for drug discovery prediction of the DNA-sequence dataset using HTM-MAT

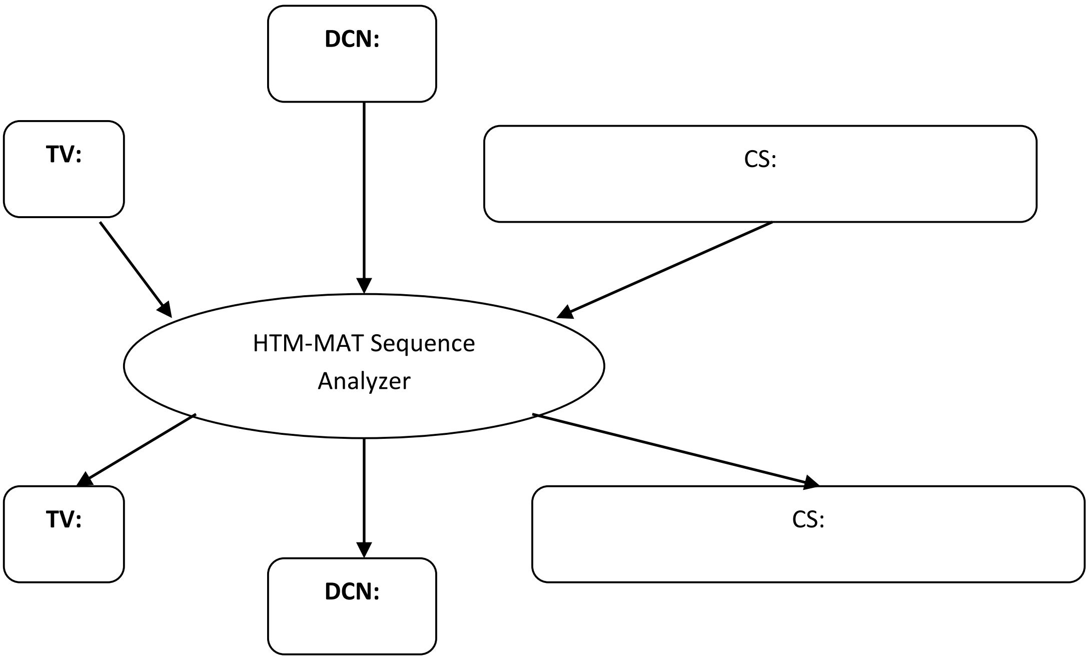

Fig.3. The diagram represents an instance of a transformation of the DNA-sequence drug pattern to a sparse distributed representation.

Here, TV is the target value, DCN the compound name, and CS, compound sequences. Permanence threshold was set to a value of 0.24. Note that, even when the permanence threshold was increased from the default value of 0.21 to 0.24, the HTM-MAT algorithm still attained a zero error rate.

Table 2. Comparative results of simulation experiments

Dataset	*HTM-MAT*	*ANN* **[5]**	*KBANN* **[1]**
drug_data1	*0.2449* [a], *0.0265* [b]	*0.144* [c]	-
drug_data2	*0.0283* [d]	-	*0.038* [e]

[a] *Mean Absolute Error (MAE)is reported for 10 MC iterations after 10 consecutive simulation runs using data from* [5].

[b] *Mean Absolute Error (MAE)is reported for 5 MC iterations after 10 consecutive simulation runs using data from* [5].

[c] *Reported MAE using a conventional multilayer perceptron ANN from* [5].

[d] *Reported error rate for HTM-MAT using the procedure and data reported in* [1]

[e] *Reported error rate from* [1] *using the KBANN*

Table 3. Simulation results of error accuracy using different relaxation factors after 5 consecutive trials

Relaxation factors	*Average Error-rates*
1.000	*0.3283*
0.950	*0.1075*
0.900	*0.0057*
0.850	*0.0019*
0.800	*0.000*
0.750	*0.000*

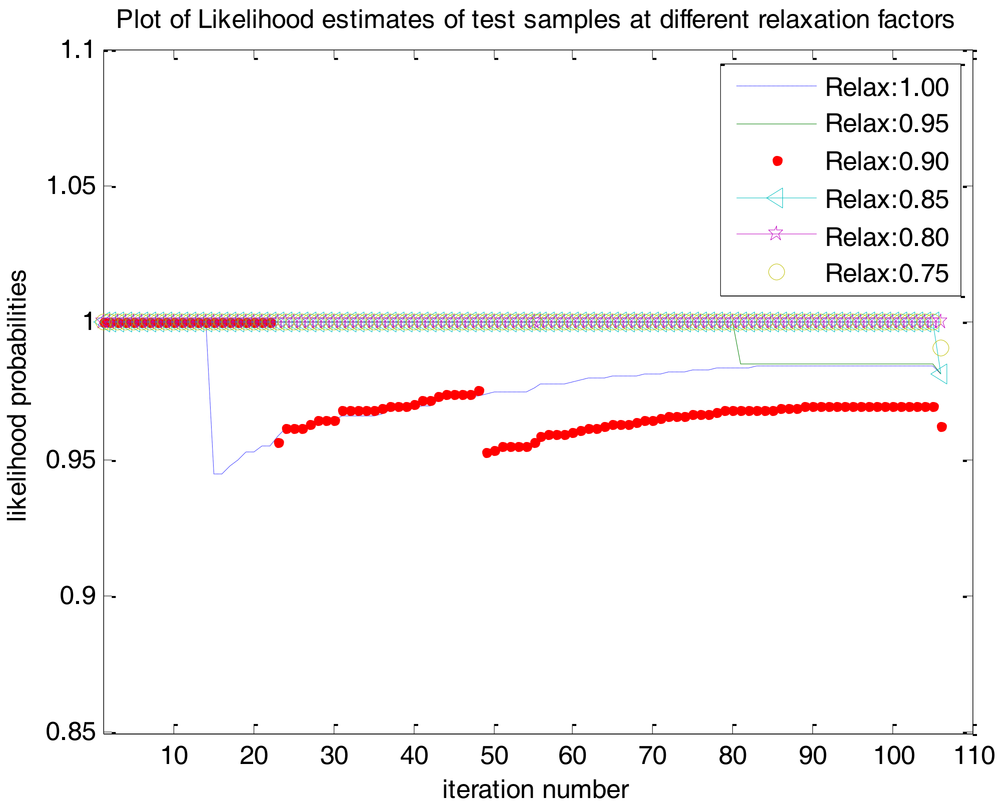

Fig.4. Possibility plots of the HTM-MAT using drug_data2 for the 5th trials

Discussions

In our experiments, we found out that the higher the iteration (trial) number the more likely the learning network generalizes better. However, the absolute errors may likely grow. Table2 shows that fine tuning HTM-MAT predictions and hence scoring based on the relaxation factors gives a mean error rate of around 0.0739. One interpretation for this phenomenon is the notion of a matching confusion matrix. Lower iteration numbers give more precise absolute errors though with lower generalization capability. In addition, reducing the relaxation threshold factor leads to better error-rate responses for the DNA-sequence dataset but this comes with the drawback of leading to false matches or predictions. Thus, it is recommended that a trade off be made between a more realistic and better generalization with high number of iterations (trials)/high relaxation threshold factors and lower error rates with small number iteration steps and a lower relaxation factor.

Conclusions and Recommendations for Future Work

A novel Artificial Intelligence tool (HTM-MAT) has been applied to the problem of drug prediction analysis. The tool has been applied to two datasets that facilitate drug discovery.

In terms of the reported error rates (ER) and mean absolute errors (MAE), it has shown very promising results as a candidate tool for discovering new drug sites or candidate compounds. In future, the HTM-MAT learning algorithms need to be tried and tested with more (recent) drug datasets and diverse learning tasks. Future work should also investigate the potentials of HTM-MAT on real-time predictions of drug sites. In addition, exploiting the generative feature in HTM-MAT for drug design is planned.

References:

King, R. D., Muggleton, S., Lewis, R. A., & Sternberg, M. J. (1992). Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proceedings of the national academy of sciences*, *89*(23), 11322-11326.

Hirst, J. D., King, R. D., & Sternberg, M. J. (1994). Quantitative structure-activity relationships by neural networks and inductive logic programming. I. The inhibition of dihydrofolate reductase by pyrimidines. *Journal of Computer-Aided Molecular Design*, *8*(4), 405-420.

King, R. D., Hirst, J. D., & Sternberg, M. J. (1995). Comparison of artificial intelligence methods for modeling pharmaceutical QSARS. *Applied Artificial Intelligence an International Journal*, *9*(2), 213-233.

Towell, G. G., Shavlik, J. W., & Noordewier, M. O. (1990, July). Refinement of approximate domain theories by knowledge-based neural networks. In *Proceedings of the eighth National conference on Artificial intelligence* (Vol. 861866).

Buciński, A., Wnuk, M., Goryński, K., Giza, A., Kochańczyk, J., Nowaczyk, A., & Nasal, A. (2009). Artificial neural networks analysis used to evaluate the molecular interactions between selected drugs and human α 1-acid glycoprotein. *Journal of pharmaceutical and biomedical analysis*, *50*(4), 591-596.

Anireh, V., & Osegi, E. (2017). HTM-MAT: An online prediction software toolbox based on cortical machine learning algorithm. https://hal.archives-ouvertes.fr/hal-01550944

Hawkins, J., Ahmad, S., & Dubinsky, D. (2010). Hierarchical temporal memory including HTM cortical learning algorithms. Techical report, Numenta. *Inc, Palto Alto*.

Zandkarimi, M., Shafiei, M., Hadizadeh, F., Darbandi, M. A., & Tabrizian, K. (2013). Prediction of pharmacokinetic parameters using a genetic algorithm combined with an artificial neural network for a series of alkaloid drugs. *Scientia pharmaceutica*, *82*(1), 53-70.

Tayarani, A., Baratian, A., Sistani, M. B. N., Saberi, M. R., & Tehranizadeh, Z.

(2013). Artificial neural networks analysis used to evaluate the molecular interactions between selected drugs and human cyclooxygenase2 receptor. *Iranian journal of basic medical sciences*, *16*(11), 1196.

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, *22*(5), 717-727.

Segler, M. H., Kogej, T., Tyrchan, C., & Waller, M. P. (2017). Generating focussed molecule libraries for drug discovery with recurrent neural networks. *arXiv preprint arXiv:1701.01329*.

Cui, Y., Ahmad, S., & Hawkins, J. (2016). Continuous online sequence learning with an unsupervised neural network model. *Neural computation*.