Enhancing Drug Discovery for Circadian Rhythm Regulation through Toxicity Classification of CRY1 Protein Molecules

Anagha Anil, Leila Michal, Xiao Lei (Brian) Zhang, Jonathan Zhu
Ray and Stephanie Lane Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

This study explores the relationship between circadian rhythms and human health, focusing on the classification of the toxicity of molecules within the Cryptochrome (CRY) functional domains, particularly CRY1. Utilizing a dataset of 171 molecules and 1538 features from the UC Irvine Machine Learning Repository, methods including Decision Trees, Random Forest, Logistic Regression, and Gradient Boosting were employed for feature elimination and model construction. Logistic Regression emerged as the most effective model across multiple performance metrics despite all models being tested under the same conditions of a 70 percent training-testing split and 5-fold cross-validation. The findings highlight the potential of these computational approaches to identify non-toxic molecules for therapeutic use, setting a foundation for future drug discovery aimed at modulating circadian rhythms. Future research will focus on refining these models through advanced heuristics and empirical validation to enhance their predictive power and reliability in clinical applications. Availability: https://github.com/jonazhu/02620_ml_group5

1 Introduction

Circadian rhythms are intrinsic 24-hour cycles that regulate a variety of biological processes including sleep, hormone production, and metabolism (Pittendrigh 1960; Vitaterna et al. 2001). These rhythms are primarily controlled by a set of core clock proteins, among which cryptochromes such as CRY1 (Figure 1) and CRY2 play a critical role (Horst et al. 1999; Griffin Jr et al. 1999). Disruptions in these rhythms are linked to a wide array of health issues, from mood disorders (Germain and Kupfer 2008; Walker et al. 2020) and various forms of cancer (Savvidis and Koutsilieris 2012; Gery and Koeffler 2010) to metabolic syndromes (Masri and Sassone-Corsi 2018; Rutter et al. 2002). Our study addresses the challenge of understanding how modifications in CRY proteins can potentially stabilize these rhythms, providing a pathway for therapeutic interventions.

1.1 Detailed motivation

The urgency and importance of the research are underscored by the increasing recognition of circadian misalignment as a factor in severe diseases (Walker et al. 2020; Gery and Koeffler 2010; Masri and Sassone-Corsi 2018). Modern lifestyles often disrupt these natural cycles through irregular sleep patterns and exposure to artificial lighting, exacerbating health risks (Pauley 2004; Tähkämö et al. 2019). With recent breakthroughs in molecular biology providing detailed insights into the structure and function of clock proteins like CRY1, there is now an unprecedented opportunity to pioneer treatments that can modulate these proteins (Gul et al. 2021). Thus, we seek to harness these

Preprint. Under review.

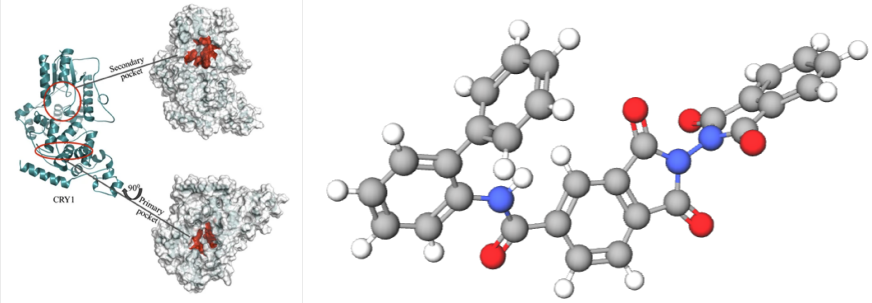

Figure 1: Figure from Gul et al. (2021) detailing the structure of CRY1, a vital protein in maintaining circadian rhythms, and its functional pockets (left), and potent CRY1 complex small molecule M17 (right).

advancements to identify non-toxic, effective molecules that can influence CRY function (Figure 1) and correct circadian disruptions at a molecular level.

1.2 The dataset

The dataset analyzed in this project was used by Gul et al. (2021) obtained via the University of California - Irvine (UCI) Machine Learning Repository (Gul and Rahim 2022) and comprises data on 171 distinct molecules, each characterized by a total of 1538 features. These features are all numerical PaDEL molecular descriptors (Yap 2011), primarily consisting of 1003 floating-point and 200 integer data types (with 335 duplicate features), along with one categorical or identifier feature, likely serving as a unique identifier for each molecule. The feature set encompasses a variety of molecular descriptors that detail both the chemical and physical properties of the molecules. Notably, the dataset includes Molecular Connectivity Indices (e.g., MATS3v, MATS3s, MATS3p) which describe the topology of the molecules and are crucial for understanding their physical and chemical behavior. Additionally, hydrogen bond features such as nHBint10, minHBint8, and minHBint2 indicate the number and types of potential hydrogen bonding interactions, essential for assessing binding affinity and specificity. Features related to Lipinski's rules, such as the number of hydrogen bond donors, help predict the drug-likeness of the molecules. Electronegativity and electron affinity descriptors (e.g., MATS3e, MATS3c, MATS3m) are also included, reflecting the electronic distribution and potential reactivity of the molecules. This comprehensive collection of descriptors is aimed at deeply characterizing each molecule's structural integrity and biological activity potential, pivotal for evaluating their therapeutic viability and toxicity. By analyzing these descriptors, our project aims to uncover predictive patterns that can determine molecular interactions and effectively classify these compounds based on their safety and functional efficacy, particularly in their role as potential therapeutic agents targeting the CRY1 protein.

2 Methods

2.1 Feature elimination

Due to the complexity and high dimensionality of the data, the first step involves dimensionality reduction. PCA alone does not adequately reduce the feature space to acceptable levels (Figure 2); while PCA can get to 80% variance in around 50 features, Gul et al. (2021) identified just 13 that were useful for predicting toxicity. As such, our process begins with feature pruning, where features that show no variation across samples or offer redundant information are eliminated. Following this, recursive feature elimination (RFE) using logistic regression is implemented to iteratively discard the least significant features. This selective reduction refines the feature set to the 20-30 most impactful for predicting molecular toxicity (Figure 3), thereby streamlining the dataset for more effective analysis and enhancing model performance by concentrating on the most relevant data.

Thus, when comparing PCA on the full dataset to similar PCAs conducted on the RFE-selected feature dataset and the paper-selected feature dataset, we see similar levels of variance reached in far fewer features (close to 8), as demonstrated in Figure 2.

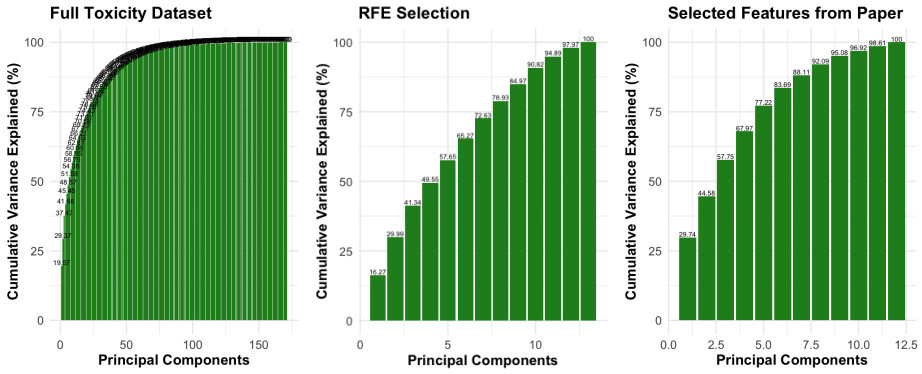

Figure 2: PCA Plots of cumulative variance explained by principal components performed on the full dataset (left), the recursive feature elimination features (middle) and the paper's selected features (right). 80% variance is reached in around 50, 8, and 7 PCs, respectively.

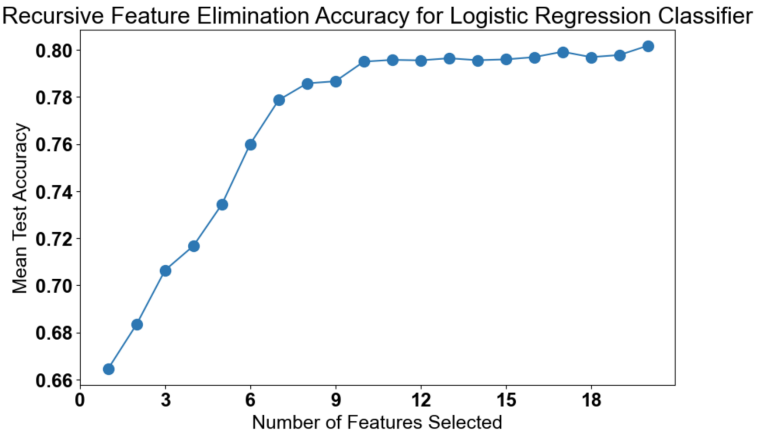

Figure 3: Plot of mean test accuracy for initial 20 features selected by recursive feature elimination. Mean test accuracy of 80% achieved in approximately 10 selected features.

2.1.1 Logistic regression

For the logistic regression model, we implemented a version of logistic regression in `numpy` and `pandas`. Due to the presence of extreme outliers for some features, predictor variable values were normalized prior to model evaluation using the Standard Scaler technique. We performed an initial round of feature selection by simply running our logistic regression algorithm on the complete dataset and naively selecting the top 20-30 features based on coefficient weighting; however, this method of feature selection yielded poor performance with our classifier models. Subsequently, we employed recursive feature elimination to iteratively eliminate the least important features until we determined the 20-30 most important features for classification purposes.

2.1.2 Gradient boosting

Gul et al. (2021) also performed gradient boosting to reduce the feature space; we did the same with a gradient boosting package from `scikit-learn`. After running Gradient Boosting, the top ranked feature importances were initially used as the selected features for classification. Performance of the model was compared to the set of selected features obtained from RFE.

2.2 Classifier models

After feature elimination, we created a variety of classification models to predict the toxicity of the molecules in the dataset. All models were trained on the full dataset, the dataset made by RFE, the

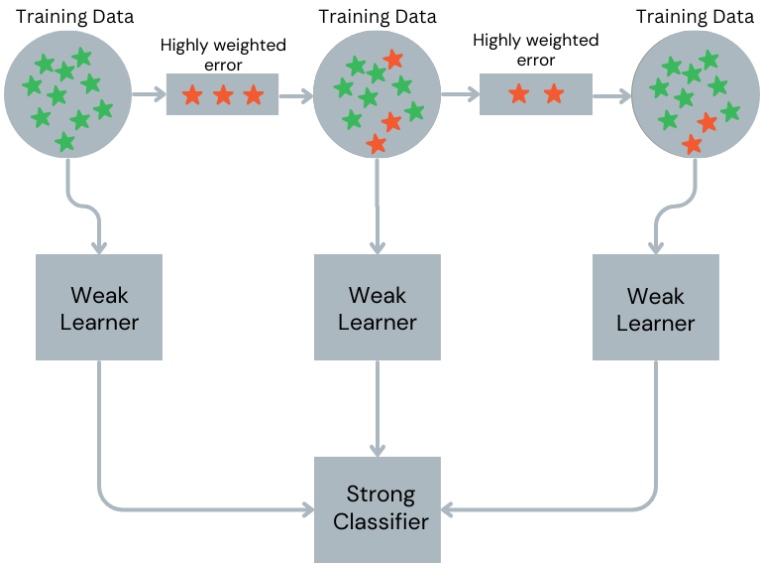

Figure 4: Diagrammatic explanation of the gradient boosting algorithm, a widely used high-performance algorithm that builds off of standard random forests and decision trees.

dataset made by Gradient Boosting feature eleimination, and the dataset selected by Gul et al. (2021). Additionally, the models we trained were the ones used by Gul et al. (2021), those being decision trees, random forests, and gradient boosted forests. We also applied logistic regression since we had the algorithm available and sought to test its efficacy.

Models were built with a 70/30 training-testing split. We used (when applicable) 5-fold cross-validation used for model selection; these folds were used to tune on a variety of parameter ranges. Once the optimal set of parameters was selected by cross-validation, we used the entire training dataset on these parameters and collected metrics based on the previously untouched testing set.

2.2.1 Decision tree classifiers

Decision Tree Classifiers (DTC) were our first model deployed to categorize each molecule as toxic or non-toxic. In general, DTCs operate through hierarchical decision-making, where each decision node within the tree uses the selected features to guide binary decisions, such as (in this case) thresholds of chemical properties or molecular weight. Normally, the next split in the tree is decided by the split that results in the highest information gain (equivalent to the largest reduction in entropy, or the most homogeneity of the resulting groups post-split).

While a variety of decision tree algorithms exist and have been refined to take a mixture of categorical and numerical data, we opted to implement a heuristic version of this algorithm directly using numpy and pandas. In short, our algorithm (working only on numerical features) decides splits for each feature by taking the minimum value for that feature for one class and the maximum value for that feature for the other class, and then taking the midpoint. Additionally, after deciding the split for the current node, the feature used to decide the split is removed. This method attempts to reduce complexity in the implementation of the algorithm at the cost of potentially finding an optimal tree.

To avoid overfitting, the decision tree was limited to a maximum depth of ten levels and cross-validation is employed across various data subsets. Additionally, parameter tuning was performed on three parameters for the tree: the maximum fraction of outcome required to make a further split in the tree, the minimum number of entries to signal the creation of a final leaf, and the depth of the tree itself. This methodology not only ensures that the model remains generalizable and accurate in its predictions but also maintains simplicity in its interpretative capacity; it also helps to optimize accuracy and other metrics despite the heuristic method.

Table 1: Key metrics for the models tested, with the best metric for the given algorithm bolded. Clearly, logistic regression trained on the recursive feature elimination dataset performed the best across all metrics.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree - Full Dataset	**0.6923**	0.4118	**0.5385**	0.4667
Decision Tree - RFE Features	0.6538	0.2941	0.4545	0.3571
Decision Tree - Gradient Features	0.6731	0.2353	0.5	0.3199
Decision Tree - Paper Features	0.6538	**0.5294**	0.4737	**0.5**
Random Forest - Full Dataset	0.5962	**0.2353**	0.3333	0.2759
Random Forest - RFE Features	**0.6731**	0.1765	**0.5**	0.2609
Random Forest - Gradient Features	0.6346	0.1176	0.3333	0.1739
Random Forest - Paper Features	0.6346	**0.2353**	0.4	**0.2963**
Logistic Regression - Full Dataset	0.4286	0.2352	0.3636	0.2857
Logistic Regression - RFE Features	**0.6923**	**0.5**	**0.375**	**0.4286**
Logistic Regression - Gradient Features	0.5578	0.1818	0.125	0.1481
Logistic Regression - Paper Features	0.6731	0.3333	0.0625	0.1053
Gradient Boosting - Full Dataset	0.5428	**0.6923**	0.75	**0.6428**
Gradient Boosting - RFE Features	**0.6857**	0.5	0.1818	0.2666
Gradient Boosting - Paper Features	0.5714	0.66	**0.79**	0.1176

2.2.2 Random forest classification

Following this, a Random Forest classifier was implemented and deployed to increase the robustness and potential accuracy of the model. This involved a simple extension of the previous DTC, taking a random subset of the feature space each time and creating a tree based on that for some number of trees. As with DTCs, cross-validation and parameter tuning were performed; in addition to all of the parameters required by DTCs, the random forest tuning process looked at possibilities for the number of trees itself, as well as the proportion of features to take when building each tree.

2.2.3 Gradient boosted trees

To further enhance the accuracy and stability of the predictions, Gradient Boosted Trees (GBT) are integrated (Figure 4). This technique improves upon the decision tree approach by sequentially adding new trees focused on correcting misclassifications from previous iterations. Each new tree incrementally refines the accuracy of the overall model, addressing any biases or errors that might have arisen. Rigorous cross-validation was applied, similar to the DTC approach, to test the GBT model under varied data splits. This strategy ensures the model's consistent performance and robustness and aims to provide a powerful tool against overfitting and enhance predictive reliability in classifying molecular toxicity.

3 Results

For all models, we collected standard metrics to evaluate binary classification models, those being accuracy, precision, recall, and F1-score (Table 1). While we had initially expected Gradient Boosting models to have the strongest performance, logistic regression trained on the RFE-selected features had the strongest overall performance across all metrics, with close to 70% accuracy.

4 Discussion

Overall, a logistic regression model trained on RFE-selected features performed the best. While we reached acceptable results on all metrics, we were unable to replicate the accuracy reached by Gul et al. (2021), which is close to 85% accuracy; we did not come close. Possible explanations for this include the use of heuristic methods (especially for decision trees and random forests), improper tuning parameters, different cross-validation methods, etc. These heuristic methods present a notable

avenue for further directions of this work, and preliminary usage of stronger methods in R regularly yielded better results across all models.

Additionally, when looking at model metrics, we did not take into account the possibility of false negatives being worse than false positives (which would very much be the case in the context of toxicity, as a false negative could potentially cause harm to patients during drug development). As such, recall may be a better metric to examine than sheer accuracy, in which case Gradient Boosting recall far exceeds others. However, logistic regression performed better on more metrics and is a simpler model; with a relatively poor recall, this forms a trade-off to consider for model deployment and usage.

We also take into account the possibility of some form of overfitting due to the usage of the entire dataset to perform RFE. Since the samples in the test sets, on which the metrics in Table 1 were calculated, were also used in the RFE process, these samples also inform which of the features are most useful and therefore would cause overfitting. Moreover, it was observed that the features selected through RFE were vastly different from those indicated as significant in Gul et al. (2021). This discrepancy suggests that our feature selection process may be capturing different aspects of the data or that the paper's features, while effective in their study, may not generalize to other datasets or conditions. However, RFE is difficult to perform on a dataset with as few samples as ours, especially when separating a dataset for testing. As such, another further direction of research would be to possibly gather more data points on more molecules so as to produce better trained, more reliable models.

Interestingly, attempting to eliminate features using RFE with a logistic regression classifier, logistic regression itself, and Gradient Boosting each yielded completely disjoint sets of selected features. When testing each of these sets of selected features, the set of features selected by RFE with logistic regression yielded the best performance with our classifier models. One possible explanation for this disparity across feature importances are the differences in model complexity, with a higher possibility of overfitting with RFE as mentioned previously. To address this issue, we could employ an ensemble feature selection model that incorporates the findings of both RFE and Gradient Boosting into a consensus set of important features, then assess the performance of our classifier models across various combinations of feature sets to improve the generalization performance of our feature selection process.

Final further directions to take for this research include the connection of this model to the actual drug development process and its use alongside clinicians and drug discoverers for circadian rhythm therapeutics. Additionally, CRY1 is not the only protein involved in regulating circadian rhythms; in fact, the related CRY2 is also essential in maintaining circadian rhythms. Expanding our model to include CRY2 functional domain molecules could also provide us with more data while also making a more generalized model.

Acknowledgments and Disclosure of Funding

We thank Dr. Min Xu, Dr. Martin Zhang, Alistair Turcan, and Ruiqi Liu for their guidance during this project. We also thank the Carnegie Mellon University Ray and Stephanie Lane Computational Biology Department for their support.

References

[1] Gul, S., Rahim, F., Isin, S., Yilmaz, F., Ozturk, N., Turkay, M., & Kavakli, I. H. (2021). Structure-based design and classifications of small molecules regulating the circadian rhythm period. *Scientific reports, 11*(1), 18510.

[2] Gül, Ş. & Rahim, F. (2022). Period Changer. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5B31D.

[3] Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry, 32*(7), 1466-1474.

[4] Horst, G. T. V. D., Muijtjens, M., Kobayashi, K., Takano, R., Kanno, S. I., Takao, M., ... & Yasui, A. (1999). Mammalian Cry1 and Cry2 are essential for maintenance of circadian rhythms. *Nature, 398*(6728), 627-630.

[5] Griffin Jr, E. A., Staknis, D., & Weitz, C. J. (1999). Light-independent role of CRY1 and CRY2 in the mammalian circadian clock. *Science, 286*(5440), 768-771.

[6] Vitaterna, M. H., Takahashi, J. S., & Turek, F. W. (2001). Overview of circadian rhythms. *Alcohol research & health, 25*(2), 85.

[7] Pittendrigh, C. S. (1960, January). Circadian rhythms and the circadian organization of living systems. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 25, pp. 159-184). Cold Spring Harbor Laboratory Press.

[8] Germain, A., & Kupfer, D. J. (2008). Circadian rhythm disturbances in depression. *Human Psychopharmacology: Clinical and Experimental, 23*(7), 571-585.

[9] Walker, W. H., Walton, J. C., DeVries, A. C., & Nelson, R. J. (2020). Circadian rhythm disruption and mental health. *Translational psychiatry, 10*(1), 1-13.

[10] Savvidis, C., & Koutsilieris, M. (2012). Circadian rhythm disruption in cancer biology. *Molecular medicine, 18*, 1249-1260.

[11] Gery, S., & Koeffler, H. P. (2010). Circadian rhythms and cancer. *Cell cycle, 9*(6), 1097-1103.

[12] Masri, S., & Sassone-Corsi, P. (2018). The emerging link between cancer, metabolism, and circadian rhythms. *Nature medicine, 24*(12), 1795-1803.

[13] Rutter, J., Reick, M., & McKnight, S. L. (2002). Metabolism and the control of circadian rhythms. *Annual review of biochemistry, 71*(1), 307-331.

[14] Pauley, S. M. (2004). Lighting for the human circadian clock: recent research indicates that lighting has become a public health issue. *Medical hypotheses, 63*(4), 588-596.

[15] Tähkämö, L., Partonen, T., & Pesonen, A. K. (2019). Systematic review of light exposure impact on human circadian rhythm. *Chronobiology international, 36*(2), 151-170.