



OPEN

Structure-based design and classifications of small molecules regulating the circadian rhythm period

Seref Gul¹, Fatih Rahim², Safak Isin³, Fatma Yilmaz⁴, Nuri Ozturk⁴, Metin Turkey^{2✉} & Ibrahim Halil Kavakli^{1,3✉}

Circadian rhythm is an important mechanism that controls behavior and biochemical events based on 24 h rhythmicity. Ample evidence indicates disturbance of this mechanism is associated with different diseases such as cancer, mood disorders, and familial delayed phase sleep disorder. Therefore, drug discovery studies have been initiated using high throughput screening. Recently the crystal structures of core clock proteins (CLOCK/BMAL1, Cryptochromes (CRY), Periods), responsible for generating circadian rhythm, have been solved. Availability of structures makes amenable core clock proteins to design molecules regulating their activity by using in silico approaches. In addition to that, the implementation of classification features of molecules based on their toxicity and activity will improve the accuracy of the drug discovery process. Here, we identified 171 molecules that target functional domains of a core clock protein, CRY1, using structure-based drug design methods. We experimentally determined that 115 molecules were nontoxic, and 21 molecules significantly lengthened the period of circadian rhythm in U2OS cells. We then performed a machine learning study to classify these molecules for identifying features that make them toxic and lengthen the circadian period. Decision tree classifiers (DTC) identified 13 molecular descriptors, which predict the toxicity of molecules with a mean accuracy of 79.53% using tenfold cross-validation. Gradient boosting classifiers (XGBC) identified 10 molecular descriptors that predict and increase in the circadian period length with a mean accuracy of 86.56% with tenfold cross-validation. Our results suggested that these features can be used in QSAR studies to design novel nontoxic molecules that exhibit period lengthening activity.

The circadian clock is a biochemical oscillator that modulates several physiologic functions such as alertness, memory, heart rate, blood pressure, and immune responses through periodic transcriptional regulation^{1–5}. Additionally, genetic and epidemiologic studies have linked clock disruption with various adverse metabolic phenotypes⁶, sleep⁷ and mood disorders⁸.

At the molecular level, four core clock proteins are required to generate circadian rhythm, which are BMAL1, CLOCK, CRYPTOCHROMES (CRYs), and PERIODs (PERs). Among these BMAL1 and CLOCK form heterodimer and bind E-box on DNA (CACGTG) and in turn, initiate transcription of clock-controlled genes (CCGs) including *Pers* and *Crys*^{9–11}. Then, PERs and CRYs accumulate in the cytosol and form a trimeric complex with casein kinase I ϵ/δ (CKI) and then translocate into the nucleus. Trimeric complex interacts with BMAL1/CLOCK and inhibits transcription of CCGs¹². Period determination in the mammalian circadian clock involves the turnover rate of the CRY and PER via post-translational modifications. FBXL3 and FBXL21 mediate the degradation of CRY proteins^{13,14}. A recent next-generation RNA sequencing analysis indicated that 10% of all genes and 43% of all protein-coding genes are under the control of the circadian clock in at least one tissue¹⁵.

Several diseases are associated with disruption of circadian rhythm at genetic level^{16–20}. Studies show a broad role for the clock in normal physiology and its role in mediating pathophysiological conditions. The importance of a robust circadian clock for health is increasingly recognized, and therefore, the identification of molecules

¹Department of Chemical and Biological Engineering, Koc University, Rumelifeneri Yolu, Sariyer, Istanbul, Turkey. ²Department of Industrial Engineering, Koc University, Rumelifeneri Yolu, Sariyer, Istanbul, Turkey. ³Department of Molecular Biology and Genetics, Rumelifeneri Yolu, Sariyer, Istanbul, Turkey. ⁴Department of Molecular Biology and Genetics, Gebze Technical University, Gebze 41400, Kocaeli, Turkey. ✉email: mturkey@ku.edu.tr; hkavakli@ku.edu.tr

that modulate circadian clocks became a hot topic^{21–28}. High-throughput screening is currently instrumental for identifying the molecules that affect the circadian clock.

Structure-based drug discovery methods have advantageous in terms of saving time and reducing cost. The addition of classification methods to the drug discovery pipeline will eliminate inappropriate molecules such as toxic and inactive. The feasibility of such implementation between them is shown in quantitative structure–activity relationships (QSAR) models for many targets^{29–37}. Given the importance of circadian rhythm in human health, with the recent reports of resolved crystal structures of core clock proteins and their interacting partners (CRY-FBXL3 (pdbID: 4K0R)³⁸, BMAL1-CLOCK (pdbID:4F3L)³⁹, and CRY-PER (pdbID:4U8H)⁴⁰, now it is possible to perform in silico screening to find small molecules targeting core clock proteins. A recent study reports the discovery of a molecule that regulates CLOCK and BMAL1 interaction using the structure-based approach²².

In this study, we performed in silico screening using CRY1 crystal structure (ID: 4K0R) to find molecules that regulate circadian rhythm in U2OS cell line. We experimentally tested 171 molecules in terms of toxicity and activity. The 56 molecules were found to be toxic, and 115 molecules were nontoxic to the cell, and 22 molecules significantly lengthened the period of the circadian rhythm. To identify molecular features, using machine learning, we used 171 molecules and our result showed that 11 features among the available 1538 were the best to predict the toxicity of the molecules. Similarly, we determined 10 molecular descriptors that explain the period change in circadian rhythm. Our results suggest that these molecular descriptors can be used in QSAR studies for the identification of nontoxic and circadian period lengthener molecules using big libraries that can be used in various CRY1 related disorders.

Material and methods

Molecular dynamics simulation. Mouse-CRY1 (mCRY1) (PDB ID: 4K0R) which is 93% identical to human CRY1 protein was retrieved from the protein databank. The structure was solvated in a rectangular box with TIP3P water molecules with the size of $7.25 \times 10^5 \text{ \AA}^3$ and neutralized with counterions using the NAMD (v. 2.6)⁴¹ program packages. Then the system was minimized using the conjugate gradient method and kept the backbone atoms of the protein frozen. Then further minimization steps with relaxed backbone atoms were carried out. The system was heated up to physiological temperature with 10 K increments by running 10 ps simulation at each temperature. Constraints were applied during 1.4 ns equilibration simulation where the initial force constant on the C_α atoms of the protein was 2 kcal/mol/Å² and reduced by 0.5 kcal/mol/Å² for each 0.4 ns equilibration run. CHARMM-PARAM22 force field⁴² was used for the molecular dynamics (MD) simulations. After the equilibration of the system, MD simulation was run at 310⁰ K for 10 ns. The pressure was controlled by the Langevin piston method during the simulations. The timestep was set to 2 fs and the bonded interactions, the van der Waals interactions (12 Å cutoff), long-range electrostatic interactions with particle-mesh Ewald (PME) were included for calculating the total force acting on the system. The last frame of the simulation was used as the “receptor” for the docking simulations. RMSD values were obtained using the RMSD trajectory tool of VMD. Backbone atoms (C, CA, N, and O) of each residue were used for RMSD calculation by excluding the translational motions.

Molecular docking simulations. More than 8 million small molecules with non-identified functions were used as ligands for the docking. Molecules having the following criteria were filtered to eliminate non-relevant molecules: molecules having more than 7 H-bond donors, more than 12 H-bond acceptors, more than 600 Da molecular weight, logP > 7, more than 8 rotatable bonds, less than 3 aromatic rings⁴³, and less than total of 4 rings. Openbabel, Autodock4.2, Autodock Tools⁴⁴ and Autodock Vina⁴⁵ programs were utilized to prepare ligands (small molecules) for the docking. Finally, more than 1million compounds were docked to target pockets by using the Autodock Vina program. The target pocket for FAD and FBXL3 binding site was determined based on the CRY-FBXL3 crystal structure³⁸. The target pocket on CRY1 was constructed via Autodock Tools. The Center of the box was located on the side chain of Phe296 amino acid residue, and the grid box size was determined as $1.9 \times 10^4 \text{ \AA}^3$. Another target pocket was the secondary pocket of CRY1. The Center of the box was located on the side chain of Lys11 amino acid residue, and the grid box size was determined as $2.7 \times 10^4 \text{ \AA}^3$.

The binding energy of molecules to CRY1 was calculated by Autodock Vina which uses a novel scoring function combining the knowledge-based and empirical approaches.

$$\Delta G_{binding} = \Delta G_{vdw} + \Delta G_{elect} + \Delta G_{hbond} + \Delta G_{desolv} + \Delta G_{tors}$$

ΔG_{vdw} : 12–6 Lennard–Jones potential function; ΔG_{elect} : Coulombic with Solmajer–dielectric function; ΔG_{hbond} : 12–10 Potential with Goodford Directionality; ΔG_{desolv} : Stouten Pairwise Atomic Solvation Parameters; ΔG_{tors} : Number of rotatable bonds.

Autodock Tools4 or PyMol (<http://pymol.sourceforge.net/>) software were used to visualize the docking results and protein structure, respectively.

MTT toxicity assay. Human osteosarcoma U2OS cell lines were used for the cytotoxicity assay. Cells were cultured and passaged at 37 °C under 5% CO₂ in 1X medium (filtered DMEM, 10% FBS, 100 µg/ml streptomycin, and 100 µg/ml penicillin and 2 mM L-Glutamine). Cells were seeded in triplicate to clear 96-well plates with 4000 cells/well then grown for 48 h. Cells were treated with molecules at desired concentrations (final DMSO concentration 0.5%) in DMEM and incubated for 48 h. Cell viability was measured by adding tetrazolium dye 3-[4,5-dimethylthiazol-2-yl]-2,5 diphenyl tetrazolium bromide (MTT) which is converted to insoluble purple color formazan because of the mitochondrial activity. Cells were incubated with MTT reagent for 4 h and then the medium was replaced with DMSO:EtOH (50:50) mixture. Purple salt was dissolved, and the absorbance of

Type of descriptors	Number of descriptors
E-state	568
Autocorrelation	346
Topological	266
Constitutional	120
Burden	96
Connectivity	56
Basak	42
Molecular property	15
Amino acid count	13
BCUT	6
Quantum chemical	6
Kappa	3
IP Molecular Learning	1

Table 1. Features of PaDEL descriptors used in this study.

wells was measured at 570 nm by the spectrophotometer. As a negative control, cells treated with 5% final DMSO concentration (known as toxic to cells). In each experiment 3-technical replicates were done.

Real time bioluminescence monitoring. 5×10^4 U2OS *Bmal1-dLuc* cells per well were seeded to an opaque 96-well plate and cultured overnight as described earlier⁴⁶. The next day cells were reset by adding dexamethasone (DXM) (0.1 μ M final) for 2 h. Then medium was changed to bioluminescence recording media which contains the following in 1L: DMEM powder (sigma D-2902, 10X 1L), 0.35 gr sodium bi-carbonate (tissue culture grade, sigma S5761), 3.5gr D(+) glucose powder (tissue culture grade, sigma G7021), 10 mL 1 M HEPES buffer (Gibco 15,140-122), 2.5 mL Pen/Strep (100 ug/ml), 50 mL 5% FBS and up to 1L sterile milliQ water. Luciferin is added freshly with 0.1 mM final concentration. Molecules were added to the bioluminescence recording media at the desired concentration (0.5% DMSO final concentration). Plates were sealed with optically clear film to prevent evaporation and gas exchange thereby maintaining homeostasis of the cells. Luminescence values were recorded at 32 °C for every 30 min with 15 s integration time via Synergy H1 luminometer for a week. The experiment was repeated three times with 3-technical replicates. To obtain the period values BioDare2 (biodare2.ed.ac.uk) was used⁴⁷. Significant analysis was performed by using the unpaired t-test with Welch's correction.

Establishment of CRY1-knockout U2OS cell line. *CRY1* knockout U2OS cell line was generated using the LentiCRISPRv2 system⁴⁸. In this study, we used the LentiCRISPRv2-CRY1-T1 construct which was described previously⁴⁹. This construct was generated using the following oligos: CRY1 Sense: 5' CACCGCCTT CAGG GCGGGGTTGTCG 3'; CRY1 Antisense: 5' AAACCGACAACCCCGCCCTGAAGGC 3'.

The lentivirus preparation, transduction of U2OS cells and selection of the knockout candidates with puromycin (at 0.5 mg/mL concentration) were performed as described previously⁴⁹. *CRY1* knockout candidates were screened with immunoblotting using anti-CRY1. To show the specificity of targeting *CRY1*, we also analyzed *CRY2* protein level and actin level, which was probed as the loading control. The antibodies used for this were as follow: anti-CRY1 (A302-614A, Bethyl Labs Inc. Montgomery, TX, USA), anti-CRY2 (A302-615A, Bethyl Labs), and anti-Actin (CST- 4967S, Cell Signaling Technology, Boston, MA, USA). HRP-labeled anti-rabbit antibody (Thermo Fisher Scientific, Waltham, MA, USA cat: 31460) were used at 1:5000 dilution. Chemiluminescence was developed using WesternBright Sirius HRP substrate (Advansta, San Jose, CA, USA, cat no: K-12043-D20) and images were captured using the ChemiDoc XRS + system (Bio-Rad).

Real time bioluminescence of CRY1-knockout cells. 40×10^4 *Cry1*^{-/-} U2OS cells were seeded to 35 mm clear plates. Then, cells were transduced with *Bmal1-dLuc* lentiviral particles as described in Doruk et al²². Next cells were reset with dexamethasone (0.1 μ M final) for 2 h and then media replaced with bioluminescence media described above with DMSO or molecules (final DMSO concentration 0.5%). Plates were sealed with vacuum grease and placed to luminometer LumiCycle (Actimetrics). Each plate was recorded continuously every 10 min for 70 s at 37 °C via photomultiplier tubes for a week. Raw luminescence data were analyzed using BioDare2 (biodare2.ed.ac.uk)⁵⁰. For each molecule, the experiment was performed three times with duplicates (at least 6 plates per molecule) The unpaired t-test with Welch's correction was used to evaluate the significance.

Classification. PaDEL descriptors of molecules were produced using ChemDes web server⁵¹. The 1538 descriptors were evaluated to describe the properties of molecules; details of molecular descriptors analyzed by PaDEL in ChemDes server were given in Table 1. The molecules both in the toxicity and period change datasets belong to two groups and we can categorize these datasets using binary classification, a machine learning approach to classify objects into two groups. The toxicity molecule set is composed of toxic and nontoxic molecules whereas in the period change dataset we have group of molecules that significantly change the period

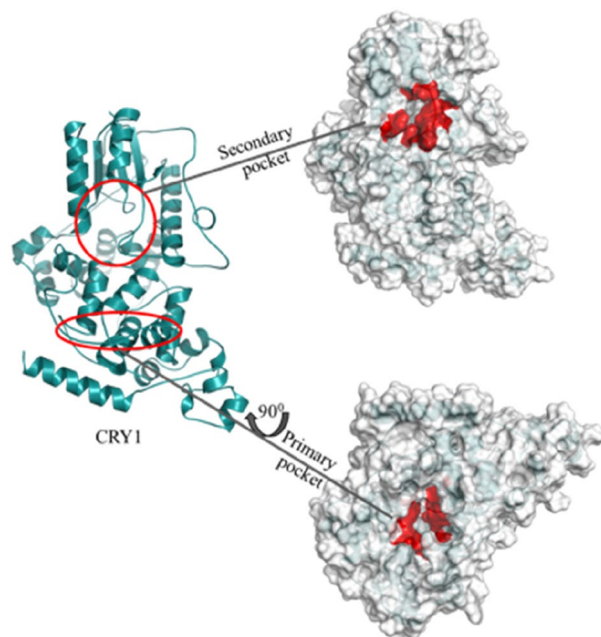


Figure 1. Crystal structure of Cryptochrome 1 (CRY1). There are two functionally important pockets, called primary and secondary pockets. Regions in primary and secondary pockets on CRY1 are shown in red color.

and another that does not affect it. The class membership of each molecule is explained in [Results and discussion](#) section.

The number of the molecular descriptors (and so is the size of the feature space) 1538, is high relative to the number of molecules in both datasets that may cause overfitting. As such, a classifier with a good fit on the training set may produce poor results on the test dataset. To prevent overfitting, it is necessary to select the best set of molecular descriptors and eliminate the redundant features. As an initial step, the features with a single value for all the molecules are discarded since they do not provide any information for classification. For feature selection, we used Recursive Feature Elimination (RFE)⁵², which is originally proposed for selection of gene subset from patterns of gene expression data.

RFE necessitates an external estimator to weigh the features with respect to their importance. Starting from initial feature set, the estimator is trained on the current set to get the importance of each feature and the features with the least importance are discarded. The process continues the reduced sets until a feature set with a predefined size is reached.

Decision Tree (DTC)⁵³, Random Forest (RFC)⁵⁴, Extra Trees (ETC)⁵⁵, and Gradient Boosting⁵⁶. Classifiers were used as classification methods, all of which can also work as an estimator for RFE. DTC assigns labels to samples on leaves of a decision tree by partitioning the feature space on each node and it is superior to other methods considering its interpretability. RFC is an ensemble classification method where multiple DTCs are trained on several subsets of the dataset and prediction is made based on the outcome of individual trees. Like RFC, ETC is based on training several DTCs. The main difference is that ETC uses the full learning set instead of its subsets. In addition, to find the best split at any node ETC uses randomly selected features. As in RFC, ETC does final prediction by majority voting of the individual trees. Gradient Boosting Classifier is a boosting algorithm that converts weak learners to stronger ones. Starting from a weak learner, decision trees, it adds new trees sequentially by minimizing a loss function using a gradient descent procedure. We used Extreme Gradient Boosting⁵⁷ (XGBC) which is an efficient, and flexible implementation of Gradient Boosting. We implemented RFE and the classification methods using the Scikit-learn package⁵⁸ and coded in Python.

Results and discussion

Structure-based small molecule design. CRYs are core clock proteins that participate in generating circadian rhythm by acting as strong transcriptional repressors of BMAL1/CLOCK transactivation in mammals^{4,59,60}. Studies revealed that CRYs SNPs are associated with different types of diseases. For example, *CRY1* variants have been associated with depression and mood disorders^{8,61,62}, elevated blood pressure and hypertension⁶². Additionally, a *CRY1* variant is linked with familial delayed sleep phase disorder and attention deficit/hyperactivity disorder^{18,63}. We, therefore, selected mammalian *CRY1* as a target for in silico screening to find molecules that regulate the period of the circadian rhythm. The *CRY1* crystal structure (ID: 4K0R) is solved³⁸. Comparison of various CRYs from different organisms shows that CRYs have variable length of extended C-terminal domains that range from 30 to 300 amino acids^{3,64,65} (Fig. 1). N-terminal domain has high homology to photolyases and is called the PHR domain. The PHR domain consists of two important regions,

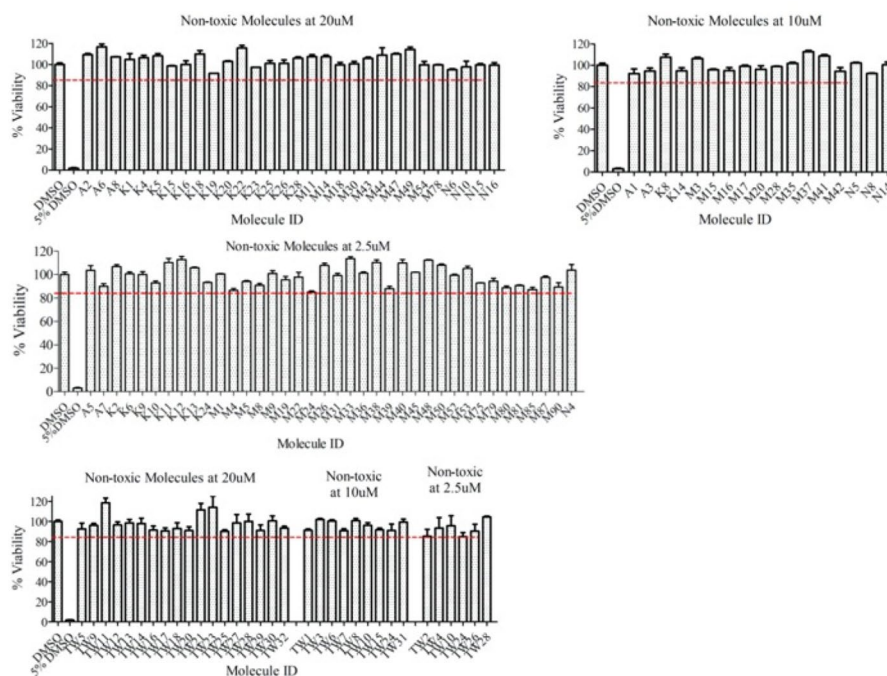


Figure 2. Non-toxic dosages of molecules determined by MTT toxicity assay. The cell viability was measured in cells treated with different concentrations of molecules. A dose of a molecule that allowed >90% of cell survival evaluated as non-toxic dose. (Data represent the mean \pm SEM $n = 3$). Cells were seeded 96-well plate and grown for 48 h. Then, cells were treated with molecules with indicated concentrations or solvent (DMSO) as control (final volume of DMSO is 0.5%). After 48 h of treatment, medium was replaced with DMEM: MTT reagent mix and incubated for 4 h. Finally, formazan salts dissolved in ethanol: DMSO mix and absorbance values of each well were measured at 570 nm using Synergy H1 (BioTek). The viability of DMSO-treated cells was normalized to 100% and the relative viability of cells treated with molecules was reported. A group of cells treated with 5% DMSO, known as toxic to cells, used as a positive control.

called the FAD-binding domain (primary pocket) and an α/β domain (secondary pocket) which are shown to be important for the interaction with the FBXL3 and the CLOCK PAS B domain, respectively⁶⁶. Therefore, the FAD-binding and secondary pockets were selected as targets which are shown to be important for regulating repressor activity of the CRYs^{67,68} (Fig. 1).

To bring CRY1 structure (PDB ID: 4K0R) near physiological conditions it was minimized and gradually heated to 310⁰ K. Then 10 ns MD simulation was run to obtain structure for the molecular docking simulations. To monitor the convergence of the simulation root mean square deviation (RMSD) of backbone atoms (C, N, C $_{\alpha}$) of amino acid residues were analyzed throughout the simulation (Fig. S1).

We initiated *in silico* screening using a commercially available small molecule library (which contains ~8 million molecules). Since docking pockets are large enough to accommodate relatively large molecules, we filtered the library to eliminate irrelevant molecules as described in the material-method section. Thus, nearly ~1 million molecules were docked to primary and secondary pockets of CRY1 by using AutodockVina. Then, molecules were ranked based on their Vina binding energies. Additionally, Pan Assay INterference compounds (PAINS) PAINS-Remover⁶⁹ was used to eliminate possible false-positive results. We tested 139 molecules designed for the primary packet of the CRY1 based on their availability. Similarly, 32 molecules designed for the secondary packet of the CRY1 were also tested for toxicity (Table S1).

Toxicity of molecules. The toxicity studies were conducted using the human osteosarcoma (U2OS) cell line, which was also employed in the circadian bioluminescence assay. We initially tested the toxicity of the 171 compounds using an MTT (3-[4,5-dimethylthiazol-2-yl]-2,5 diphenyl tetrazolium bromide)-based assay at 20 μ M and determined that 48 of them were non-toxic (Fig. 2). The remaining 123 molecules that show toxic effects at 20 μ M were further evaluated at 10 μ M. Results indicated that 26 molecules were not toxic at 10 μ M. Finally, the other 97 molecules were tested at 2.5 μ M and found that 41 molecules were non-toxic at this concentration (Fig. 2). The rest of 56 molecules with relative cell viability < 85% at 2.5 μ M were labeled as toxic and, therefore, eliminated from further characterization. As a control, cells treated with 5% DMSO known to be toxic. In summary, of 171 tested molecules, 56 were toxic to U2OS cell lines whereas the 115 molecules were evaluated as non-toxic molecules at different concentrations. Structures of all molecules were provided in the supplementary data (Fig. S2).

Classification of molecules based on toxicity. The toxicity data set is composed of 171 molecules with 1538 molecular descriptors. 334 features repeating the same value for all molecules were discarded. The remaining 1203 features were utilized to obtain the best feature set by Recursive Feature Elimination (RFE). Since RFE method⁵² and so the selected feature set is dependent on the estimator used, Decision Tree Classifier (DTC)⁵³, Random Forest Classifier (RFC)⁵⁴, Extra Trees Classifier (ETC)⁵⁵, and Gradient Boosting (XGBC)⁵⁶ were tested as the external estimators. To search for the promising regions in the space of the selected features, we generated feature sets with cardinality ranging from 2 to 20 in increments of 1 for each of these classifiers. To evaluate the potency of the selected features and compare the classifiers based on their prediction accuracy, 10-Fold cross-validation (CV) were run on all the generated sets and replicated 100 times.

The toxic and non-toxic groups of molecules are not evenly distributed. The 33% of all molecules are toxic and the rest, 67%, is non-toxic. To cope with the unbalanced groups, we used weights associated with each class which are inversely proportional to the class sizes. The weights corresponding to toxic and non-toxic molecules, 1.53 and 0.74 are calculated by $w_T = n/(2 * n_T)$, $w_{NT} = n/(2 * n_{NT})$, respectively, where n is the total number of molecules, n_T and n_{NT} are the number of toxic and non-toxic molecules in the dataset.

To optimize the performance in discriminating between toxic and non-toxic molecules, we tuned the hyperparameters of each classifier. We did a grid search within the space of all combinations of a selected set of values of parameters and optimized over 10-Fold CV. The parameters of DTC and their corresponding values evaluated in grid search are, max_depth (The maximum depth of the tree): [1, 2, 3, ..., 10, None], min_samples_splits (the minimum number of samples required to split an internal node): [2, 3, ..., 10], min_samples_leafs (The minimum number of samples required to be at a leaf node): [1, 2, 3, ..., 10], and max_features (the number of features to consider when looking for the best split): [1, 2, 3, ..., num_features]. RFE and ETC parameters and their alternative values are max_depth: [1, 2, 3, ..., 6, None], min_samples_splits: [2–5], min_samples_leafs: [1–5], max_features: [1, 2, 3, ..., sqrt(num_features)], and n_estimators (The number of trees in the forest): [100, 200]. Note that the first four parameters of RFC and ETC are common with DTC. However, we use a smaller space of values due to the computational complexity of RFC and ETC. The set of parameters of XGBC and their set of levels to optimize are learning_rate (step size shrinkage used in update to prevents overfitting): [0.01, 0.1], max_depth: [3, 5, 7, 10], min_child_weight (minimum sum of instance weight needed in a child): [1, 3, 5], subsample (subsample ratio of the training instances): [0.5, 0.7], colsample_bytree (subsample ratio of columns when constructing each tree): [0.5, 0.7], and n_estimators: [100, 200]. We used gbtrees as the booster of XGBC which employs tree-based models.

To search for the promising regions in the space of the selected features, we generated feature sets with cardinality ranging from 2 to 20 in increments of 1 for each of the classifiers. To evaluate the potency of the selected features and compare the classifiers based on their prediction accuracy, 10-Fold cross-validation (CV) were run on all the generated sets and replicated 100 times. The grid search for parameter tuning is made for each feature set independently and CV repetitions are implemented based on tuned hyperparameters.

The average accuracies of classifications with 100 replications on generated feature sets are given in Table 2. The rows represent the number of features selected and the mean accuracies for the generated feature sets. The classifiers used for 10-Fold CV are placed in columns. The maximum average accuracies for each of the classifiers are marked in bold numbers. Our analyses show that DTC attained the highest mean accuracy of 78.77% for 19 feature set and is by far superior to the other classifiers studied in terms of prediction power. RFC follows DTC with the highest mean accuracy 72.99% for a set having 9 features, while ETC and XGBC are inferior, resulting in 71.36% and 71.17% maximum accuracies with 20 and 17 features, respectively. The maximum and standard deviation of 100 CV accuracies for each feature set and classifier pair are presented in Tables S2 and S3, respectively. In line with the mean accuracy comparison, DTC attained the best with 84.80% maximum accuracy on a set with 19 features. RFC and XGBC reached the highest score of 77.78% while ETC stayed at 76.02% level. The standard deviation of DTC values is greater than 2 except for one feature set and is higher compared to the other classifiers (Table S3). However, this high variation in DTC results is compensated by higher mean accuracies. The lower variation in RFC, ETC and XGBC results, mostly less than 2, does not pose an advantage due to their lower mean accuracies.

The feature set with cardinality 14 results in a mean accuracy of 78.49% by DTC and it is very close to the highest score of 78.77% for 19 features. We concluded that the additional 5 features do not provide significant improvement in the prediction power of DTC and we continued our study with 14 features. Tuning Hyperparameters of DTC by grid search for 14 features resulted in the optimized values: max_depth = None, max_features = 13, min_samples_leaf = 1, min_samples_split = 5.

The RFE method iteratively prunes the least important features to get the set with preferred cardinality. However, the generated set is not guaranteed to be optimal. To determine the most essential features, we iteratively pruned the features in a similar approach with RFE in the selected 14 features. At each iteration, we did 100 CV repetitions on the reduced sets obtained by dropping every feature one at a time. The feature that provides the highest mean accuracy among the reduced sets was pruned. Our analysis for the reduced sets together with the pruned molecular descriptor showed that removing the descriptor ATSC8v to get 13 features increased the mean accuracy from 78.49 to 79.63% (Table 3). Further reduction in the size of the feature set decreased the mean accuracies. This is probably due to excluding the informative descriptors. We concluded that 13 features are the best descriptive set among 1203 descriptors with DTC to classify the toxicity data. Note that since there is no max_depth limit for 14 features DTC parameters, at each pruning step we additionally tuned the max_depth parameter to get the best max_depth level of 10.

The selected 13 molecular descriptors are: “MDEC-23, MATS2v, ATSC8s, VE3_Dt, CrippenMR, SpMax7_Bhe, SpMin1_Bhs, C1SP2, GATS8e, GATS8s, SpMax5_Bhv, VE3_Dzi, VPC-4.” (Table 4). Finally, 10,000 CV repetitions were run to get the maximum and mean accuracies and the standard deviation of accuracies i.e.

Features	10 Fold CV—accuracy (%)			
	DTC	RFC	ETC	XGBC
2	60.22	68.63	70.32	65.06
3	59.20	68.68	70.40	67.64
4	60.39	71.97	68.82	68.10
5	71.93	69.47	69.88	68.88
6	69.87	71.18	67.99	68.67
7	71.73	72.70	68.12	68.20
8	72.81	72.78	69.16	67.37
9	75.87	72.99	68.16	68.32
10	75.65	72.68	70.21	67.92
11	76.75	71.71	68.60	70.18
12	75.33	71.43	68.83	70.16
13	76.46	72.36	68.57	68.87
14	78.49	72.54	68.23	68.63
15	77.22	70.76	69.67	69.08
16	77.21	72.41	68.73	71.16
17	75.83	72.81	68.77	71.17
18	78.75	72.37	69.49	71.05
19	78.77	70.03	70.32	70.11
20	78.02	70.73	71.36	70.73

Table 2. Toxicity dataset, mean accuracy of 10-Fold CV with 100 replications for feature sets with cardinality ranging from 2 to 20. DTC, RFC, ETC, and XGBC trained and tested on feature sets with cardinality between 2 and 20.

Features	Removed	Max	Mean	Std. Dev
14	-	83.04	78.49	2.36
13	ATSC8v	84.21	79.63	2.01
12	VE3_Dt	84.21	79.41	2.50
11	SpMin1_Bhs	82.46	78.90	1.77
10	SpMax5_Bhv	83.63	78.95	1.86
9	GATS8e	83.04	77.82	2.17

Table 3. Toxicity Dataset, maximum, mean, and standard deviation of 10-Fold CV accuracies with 100 repetitions. DTC applied to reduced feature sets obtained by removal of a single feature at a time.

Descriptor name	Type	Description
MDEC-23	MDEDescriptor	Molecular distance edge between all secondary and tertiary carbons
MATS2v	Moran Autocorrelation Descriptor	Moran autocorrelation—lag 2/weighted by van der Waals volumes
ATSC8s	Centered Broto-Moreau Autocorrelation Descriptor	Centered Broto-Moreau autocorrelation—lag 8/weighted by I-state
VE3_Dt	Detour Matrix Descriptor	Logarithmic coefficient sum of the last eigenvector from detour matrix
CrippenMR	Crippen Descriptor	Crippen's molar refractivity
SpMax7_Bhe	Burden Modified Eigenvalues Descriptor	Largest absolute eigenvalue of Burden modified matrix—n 7/weighted by relative Sanderson electronegativities
SpMin1_Bhs	Burden Modified Eigenvalues Descriptor	Smallest absolute eigenvalue of Burden modified matrix—n 1/weighted by relative I-state
CISP2	Carbon Types Descriptor	Doubly bound carbon bound to one other carbon
GATS8e	Geary Autocorrelation Descriptor	Geary autocorrelation—lag 8/weighted by Sanderson electronegativities
SpMax5_Bhv	Burden Modified Eigenvalues Descriptor	Largest absolute eigenvalue of Burden modified matrix—n 5/weighted by relative van der Waals volumes
VE3_Dzi	Barysz Matrix Descriptor	Logarithmic coefficient sum of the last eigenvector from Barysz matrix/weighted by first ionization potential
VPC-4	ChiPath Cluster Descriptor	Valence path cluster, order 4

Table 4. Name, type, and description of selected 13 features determining the toxicity of a molecule.

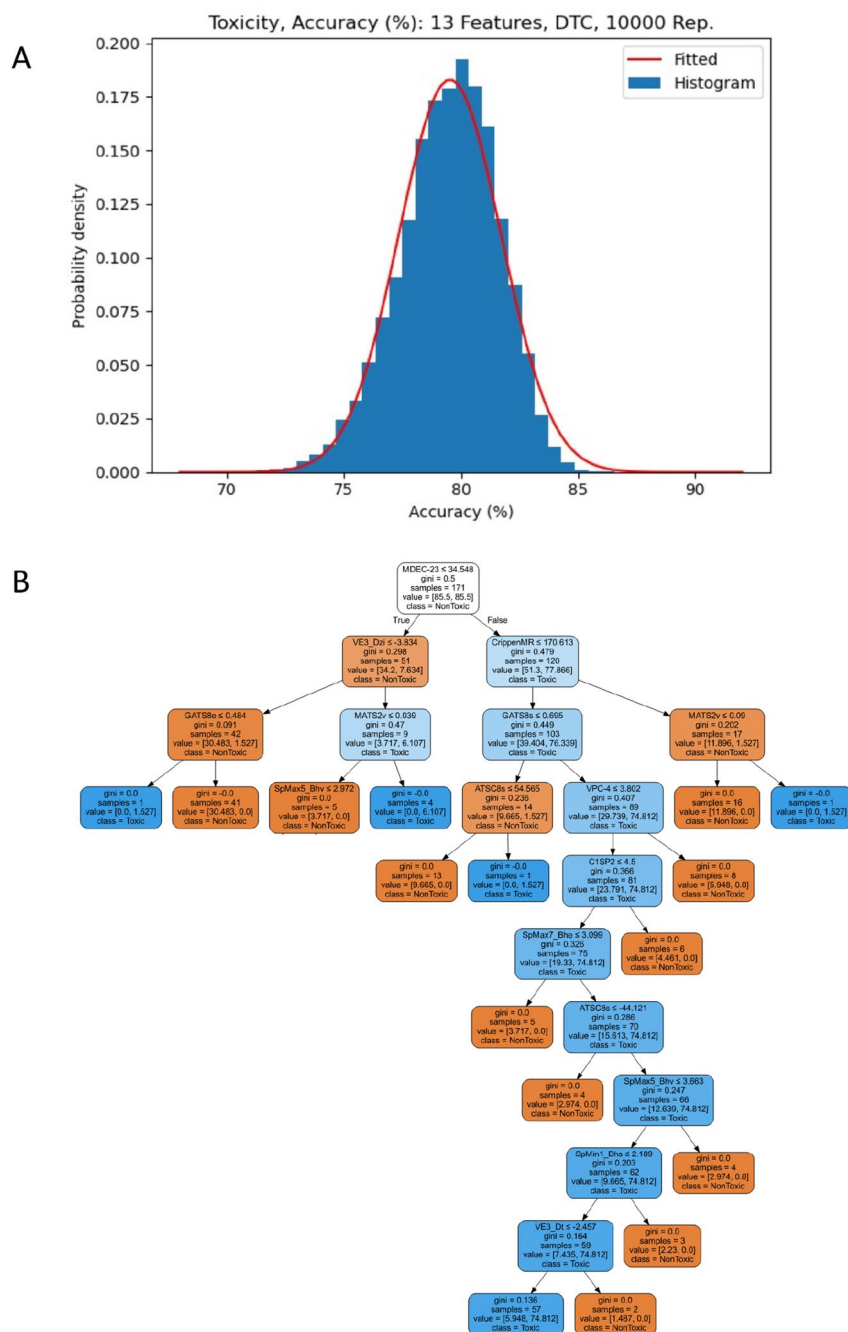


Figure 3. (A) Toxicity Dataset, histogram and fitted normal probability density function of the accuracies for 10,000 replications of DTC (Max. Depth, 10) applied to final 13 molecular descriptors. (B) Toxicity Dataset, DTC plot⁷⁰, which is generated by graphviz.org, with maximum depth 10 trained on final 13 features.

86.55, 79.53, 2.18, respectively. The accuracies displayed an approximately normal distribution for the histogram of accuracies and probability density function of the fitted normal distribution (Fig. 3A). The plot of DTC with d_{max} 10 trained on 13 features showed that among 16 leaf nodes, 11 (orange color) conclude that a new molecule is nontoxic, and the rest 5 (blue color) results in the decision that it is toxic (Fig. 3B).

Circadian bioluminescence assay. U2OS cell is a commonly used cell line in the circadian rhythm field due to its robust rhythm^{22,25,46,71}. Any agents such as small interfering RNA (siRNA) and chemicals or gene knockout (KO) affecting the stability or activity of clock proteins change the parameters (period, amplitude, and phase) of the circadian rhythm^{25,72}. We analyzed the effect of non-toxic molecules on the period length of circadian rhythm in U2OS cells stably expressing destabilized firefly Luciferase (*dluc*) under the control of the *Bmal1* promoter (U2OS *Bmal1-dluc*).

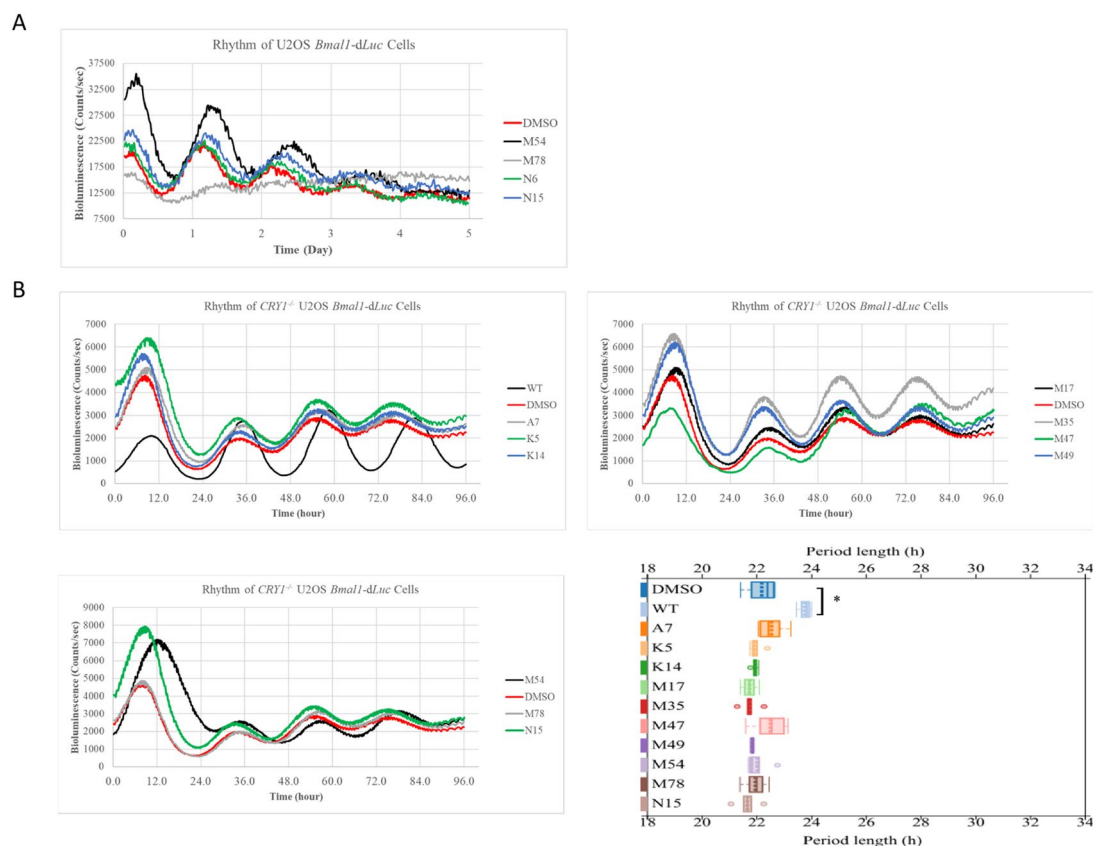


Figure 4. The effect of the molecules on circadian rhythm. **(A)** 5×10^4 U2OS *Bmal1-dLuc* cells were seeded to an opaque 96-well plate. Next day cells were synchronized by dexamethasone for 2 h. Then the medium was replaced with luminescence recording medium having molecules or DMSO. Bioluminescence readings were recorded for a week using Synergy H1 (BioTek). Period data was calculated using Biodare2 web-server (biodare2.ed.ac.uk) (all results were given in Fig. S3). To determine molecules that are changing the period of the rhythm significantly, the period length of molecule-treated cells was compared to that of DMSO control using unpaired t-test with Welch's correction ($****p < 0.0001$ $***p = 0.001$ $**p < 0.01$, $*p < 0.05$, $n = 3$). Each biological replicate was the average of the 3 technical replicates. **(B)** 1×10^5 *CRY1*^{-/-}U2OS transduced with *Bmal1-dLuc* reporter. WT represents U2OS *Bmal1-dLuc*. Bioluminescence readings were recorded for a week using LumiCycle (Actimetrics, USA). Period data was calculated using Biodare2 web-server (biodare2.ed.ac.uk). Unpaired t-test with Welch's correction was used for significant analysis ($*p < 0.05$, $n = 3$). Each biological replicate was the average of the two technical replicates.

Since the primary and secondary pockets of CRY1 are critical to interact with different proteins e.g. CLOCK and FBXL3, respectively, molecules designed for these two pockets might have differential impacts on the circadian rhythm. Thus, we focused only on the effect of 85 non-toxic molecules designed for the primary pocket of CRY1. U2OS *Bmal1-dLuc* cells treated with these molecules and their effect on circadian period length was analyzed by BioDare2 (biodare2.ed.ac.uk)⁵⁰. Analysis revealed that 21 molecules significantly lengthen the period of circadian rhythm (Fig. S3). One molecule, N8, shortened the period and was excluded from further classification studies. The representative figure for period lengthener molecules is shown in Fig. 4A. Circadian rhythm results of all period lengtheners were given in Fig. S3. All period values were provided in Table S4. To verify the CRY1 dependency of molecules, we generated U2OS *CRY1*^{-/-} *Bmal1-dLuc* cells by utilizing CRISPR/Cas9 technology (Fig. S4). Knocking out the *CRY1* in this cell line resulted in a shorter period (indicated with red line) compared to wild-type controls (indicated with black line) as in agreement with previously published data⁷³ (Fig. 4B). Notably, when U2OS *CRY1*^{-/-} *Bmal1-dLuc* cells were treated with potent molecules A7, K5, K14, M17, M35, M47, M49, M54, M78, and N15 no change was observed in the period length of the circadian rhythm (Fig. 4B). We, then, performed a classification study to determine the molecular characteristics leading to period change.

Next, we analyzed the stability of the interaction between molecules and CRY1 using the MD simulations. The seven of the most potent (A7, M17, M35, M47, M49, M54, and M78) molecules in complex with CRY1 were simulated. Parameters for molecules were generated using CHARMM-GUI server. The CRY1-molecule complexes, obtained from docking analysis, were simulated for 20 ns. The initial docking position of molecules and nearby amino acid residues on CRY1 were shown in (Fig. 5A). We identified the nature of interactions between CRY1 and molecules as followings. Ring structures in molecules generated pi-type (Pi-cation, Pi-alkyl, Pi-Pi stacked or T-shaped) interactions with at least one of either Arg-293, His-355 and Trp-399 of CRY1. In addition, all molecules interacted with Leu-255 and Ile-392 of the CRY1 through van der Waals forces. To

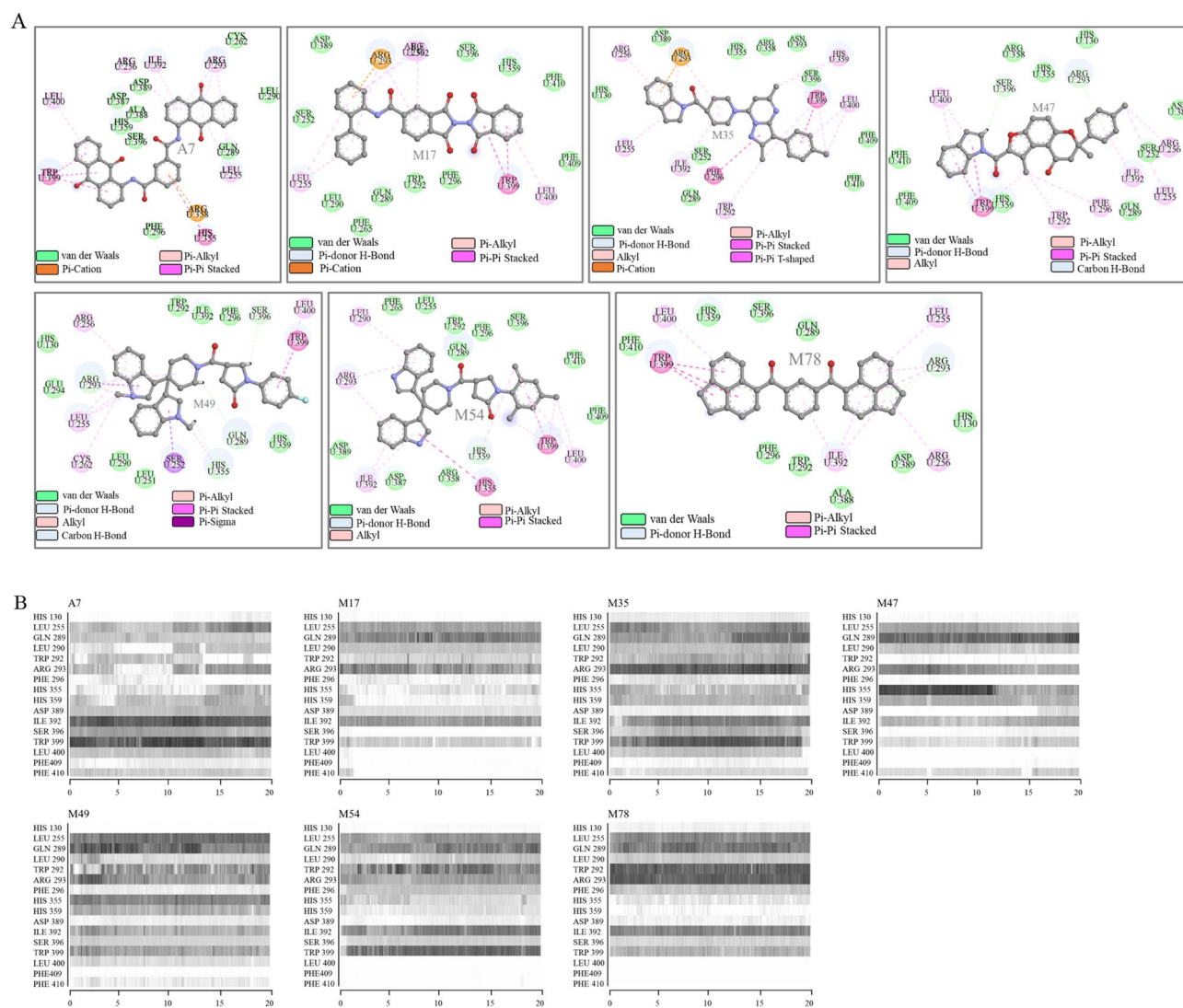


Figure 5. Binding mode of molecules on CRY1 and analysis of molecular dynamic simulations of CRY1-molecule complexes. (A) Docked conformation of molecules was analyzed and 2D interaction map was generated using Discovery Studio Visualizer. Type of interactions was given in color-coded format. (B) Interaction maps of molecules with CRY1 through 20 ns MD simulation were generated using VMD timeline applet. It has been evaluated as interaction if molecule and amino acid residues get 5 Å or closer and shown with a black bar.

to evaluate the persistence of these interactions we generated a contact map of molecules with nearby amino acids through 20 ns MD simulations. RMSD values of backbone atoms of proteins showed that simulations reached the equilibrium (Fig. S5). Contact maps showed that these interactions were maintained during the simulations (Fig. 5B). Visual inspection showed that molecules did not cause any conformational changes and their interactions were maintained throughout the simulation. A list of highly interacting amino acids with each molecule was given in Table 5. In addition, amino acid residues, which formed hydrogen bonds, were determined from the initial docking position of molecules (Table 5) and the persistence of these interactions was confirmed from interaction maps (Fig. 5B).

Classification of period lengthening molecules. Factors determining the period length in the circadian rhythm are quite complex. For example, the deletion of analogous *Cry1* and *Cry2* genes in mice causes short and long period phenotypes, respectively⁷⁴. CRY binding small molecules discovered by high-throughput screening were reported to stabilize the CRY1. Interestingly, these molecules caused differential circadian phenotype in treated cells e.g. shorter or longer period length^{25,27}. Thus, we focused only on the period lengthening molecules. Previously reported 6 molecules binding to CRY1 and lengthened the period of circadian rhythm (KL001, GO058, GO061, GO152, GO214, GO216)^{25,27} were included in the classification analysis, GO203 which does not change the rhythm included as no-changer. We started the classification with 90 molecules of which 27 are period lengthening, 63 are no-changers. 1538 molecular descriptors were generated for all molecules.

Molecules	Interacting amino acids	Hydrogen-bond forming amino acids
A7	Leu-255, Gln-289, Asp-389, Ile-392, Ser-396, Trp-399	Arg-293, His-359
M17	Leu-255, Gln-289, Leu290, Trp292, Arg-293, Ile-392	Arg-293
M35	Leu-255, Gln-289, Trp-292, Arg-293, Ile-392, Trp-399	Arg-293, His-359
M47	Leu-255, Gln-289, Leu-290, Arg-293, His-355, His-359	Arg-293, His-359
M49	Leu-255, Gln-289, Arg-293, His-355, Ile-392, Trp-399	Arg-293, Gln-289, His-355
M54	Leu-255, Gln-289, Trp-290, Arg-293, Ile-392, Trp-399	His-359, Ser-395
M78	Leu-255, Gln-289, Trp-292, Arg-293, Ile-392, Trp-399	Ser-396

Table 5. Highly interacting and hydrogen bond generating amino acid residues of CRY1 with molecules.

Features	10 Fold CV—accuracy (%)			
	DTC	RFC	ETC	XGBC
2	74.63	75.78	69.19	63.17
3	70.00	74.48	76.78	70.42
4	70.30	82.17	78.24	71.74
5	79.07	78.22	78.29	75.31
6	80.92	79.57	79.68	73.66
7	79.87	80.04	79.82	74.92
8	82.18	79.14	81.77	75.37
9	82.10	82.72	81.91	78.88
10	80.82	81.30	78.11	80.50
11	81.93	80.83	80.24	80.43
12	81.23	81.13	79.89	81.63
13	82.13	80.93	77.34	83.28
14	78.52	81.02	77.67	82.69
15	80.66	81.16	75.08	83.69
16	78.70	83.62	77.49	82.59
17	75.21	82.20	79.61	80.94
18	78.32	82.74	81.60	81.76
19	76.88	79.60	78.34	82.97
20	77.41	79.62	81.16	82.53

Table 6. Period lengthening dataset, mean accuracy of 10-Fold CV with 100 repetitions. DTC, RFC, ETC, and XGBC trained and tested on feature sets with cardinality between 2 and 20.

However, 360 of them have the same value for all molecules and were discarded. The remaining 1177 features were used to train the dataset.

We followed a similar approach with the toxicity dataset for the classification of the set of period-lengthening molecules. The period dataset is also unbalanced as in toxicity since 30% of the molecules are period changers and the rest 70% are no-changer molecules. To deal with the possible bias, we set the weights of the period-lengthening and no-changer molecules as 1.67 and 0.71, respectively. We generated feature sets with cardinalities between 2 and 20 from the period dataset by RFE using DTC, RFC, ETC and XGBC as external estimators. Next, we tuned the parameters of the given classifiers on each of the feature sets and did 100 CV with the optimized parameters.

Mean accuracy levels for each feature set and classifier pair are presented in Table 6. All of the classifiers achieved mean accuracies greater than 80% for multiple numbers of feature sets and the highest mean accuracies for each classifier are marked in bold. RFC and XGBC are the best of all with highest mean accuracies, 83.62% for 16 features and 83.69% for 15 features. Since XGBC provided slightly higher mean accuracy than RFC with 1 less feature, we selected XGBC as the most promising classifier for the period dataset.

The maximum and standard deviations of 100 CV accuracies for each feature set and classifier pair are given in Tables S5 and S6 respectively. Among all, XGBC provided the maximum accuracy of 90% again with the 15 features. XGBC Parameters tuned for 15 features are such that, `colsample_bytree = 0.5`, `learning_rate = 0.1`, `max_depth = 3`, `min_child_weight = 1`, `n_estimators = 100`, and `subsample = 0.7`.

As performed in the toxicity dataset, we iteratively pruned the features in the selected 15 features to eliminate the redundant ones, this time using XGBC with the tuned hyperparameters. In Table 7 maximum, mean and standard deviation of accuracies of 100 CV applied on reduced feature sets are presented. Reducing the set from

Features	Removed	Max	Mean	Std. Dev
15	–	90.00	83.69	2.41
14	ATS3m	90.00	85.11	2.11
13	AATSC4m	90.00	86.11	2.07
12	MATS5m	91.11	86.17	2.52
11	minHBint2	91.11	86.48	2.27
10	AATS4p	92.22	86.94	2.18
9	minsCH3	91.11	86.82	2.21
8	ATSC4p	90.00	85.47	2.16
7	MLFER_S	88.89	84.74	2.49

Table 7. Period lengthening dataset, maximum, mean, and standard deviation of 10-Fold CV accuracies with 100 repetitions. XGBC applied to reduced feature sets obtained by removal of a single feature at a time.

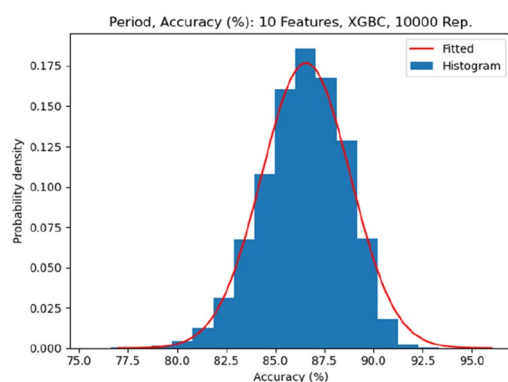


Figure 6. Period Change Dataset, histogram, and fitted normal probability density function of the accuracies for 10,000 replications of XGBC applied to final 10 molecular descriptors.

15 to 10 features increased the mean accuracy from 83.69 to 86.94%. Dropping further features reduced the mean accuracy since all the features in 10 features are informative.

We concluded that XGBC with the tuned parameters coupled with the reduced set with 10 features is the best classifier for the period changer dataset. We did a final 10,000 repetition of 10-Fold CV with XGBC on 10 features to get the maximum and mean accuracies of 93.33%, and 86.56% respectively. The histogram and fitted normal probability density function of the accuracies for 10,000 replications are presented in Fig. 6. The selected final 10 features includes: “ATSC8c, MATS1e, minsCH3, MATS4e, MATS4s, ATSC7i, SpMin4_Bhp, MLFER_S, ATSC4p, SpMax2_Bhm” (Table 8).

Conclusions

Drug discovery is a very expensive and time-consuming process posing several daunting challenges. Compared to the classical high-throughput approach to computer-assisted drug discovery, employing virtual screening (VS) is a promising approach to reduce the cost of the initial drug discovery. VS allows identifying hit compounds from large databases of drug-like molecules much faster and cheaper than traditional approaches. VS utilizes comprehensive evaluation of ADMET parameters by pharmacophore modeling⁷⁵ and quantitative structure–activity relationship (QSAR) analysis⁷⁶. In addition to these, toxicity prediction is becoming a more significant part of current computer-assisted drug development, especially when libraries contain tens of millions of untested compounds. As a result, quick and inexpensive computational algorithms are frequently used to eliminate potentially toxic compounds and reduce the number of experimental tests required. Here we identified small molecules that bind functionally important regions of a core clock protein CRY1. First, of tested 171 molecules, 115 molecules are nontoxic while 56 molecules are toxic. Then we performed machine learning methods to classify toxic and nontoxic molecules. DTC identified 13 features that can predict the toxicity with accuracy of about 80%. Second, 21 molecules were identified as period lengthener among 85 molecules. Furthermore, machine learning approach using XGBC determined 10 molecular descriptors that can predict period lengthener molecules about 87% accuracy. These descriptors can be implemented in future VS studies on CRY1 to predict the toxicity and period lengthener effect of molecules from libraries containing several hundred million compounds.

Descriptor name	Type	Description
ATSC8c	Centered Broto-Moreau Autocorrelation Descriptor	Centered Broto-Moreau autocorrelation—lag 8/weighted by charges
MATS1e	Moran Autocorrelation Descriptor	Moran autocorrelation—lag 1/weighted by Sanderson electronegativities
minsCH3	Electrotopological State Atom Type Descriptor	Minimum atom-type E-State: -CH3
MATS4e	Moran Autocorrelation Descriptor	Moran autocorrelation—lag 4/weighted by Sanderson electronegativities
MATS4s	Moran Autocorrelation Descriptor	Moran autocorrelation—lag 4/weighted by I-state
ATSC7i	Centered Broto-Moreau Autocorrelation Descriptor	Centered Broto-Moreau autocorrelation—lag 7/weighted by first ionization potential
SpMin4_Bhp	Burden Modified Eigenvalues Descriptor	Smallest absolute eigenvalue of Burden modified matrix—n 4/ weighted by relative polarizabilities
MLFER_S	MLFER Descriptor	Combined dipolarity/polarizability
ATSC4p	Centered Broto-Moreau Autocorrelation Descriptor	Centered Broto-Moreau autocorrelation—lag 4/weighted by polarizabilities
SpMax2_Bhm	Burden Modified Eigenvalues Descriptor	Largest absolute eigenvalue of Burden modified matrix—n 2/ weighted by relative mass

Table 8. Name, type, and description of selected 10 features determining the period changing information of a molecule.

Received: 3 November 2020; Accepted: 27 August 2021

Published online: 16 September 2021

References

- Franken, P. & Dijk, D. J. Circadian clock genes and sleep homeostasis. *Eur. J. Neurosci.* **29**, 1820–1829. <https://doi.org/10.1111/j.1460-9568.2009.06723.x> (2009).
- Gachon, F., Nagoshi, E., Brown, S. A., Ripperger, J. & Schibler, U. The mammalian circadian timing system: From gene expression to physiology. *Chromosoma* **113**, 103–112. <https://doi.org/10.1007/s00412-004-0296-2> (2004).
- Kavakli, I. H. *et al.* The photolyase/cryptochrome family of proteins as DNA repair enzymes and transcriptional repressors. *Photochem. Photobiol.* **93**, 93–103. <https://doi.org/10.1111/php.12669> (2017).
- Kavakli, I. H. & Sancar, A. Circadian photoreception in humans and mice. *Mol. Interv.* **2**, 484–492. <https://doi.org/10.1124/mi.2.8.484> (2002).
- Kyriacou, C. P. & Hastings, M. H. Circadian clocks: genes, sleep, and cognition. *Trends Cogn. Sci.* **14**, 259–267. <https://doi.org/10.1016/j.tics.2010.03.007> (2010).
- Sahar, S. & Sassone-Corsi, P. Metabolism and cancer: The circadian clock connection. *Nat. Rev. Cancer* **9**, 886–896. <https://doi.org/10.1038/nrc2747> (2009).
- Steele, T. A., St Louis, E. K., Videnovic, A. & Auger, R. R. Circadian rhythm sleep-wake disorders: A contemporary review of neurobiology, treatment, and dysregulation in neurodegenerative disease. *Neurotherapeutics* <https://doi.org/10.1007/s13311-021-01031-8> (2021).
- Soria, V. *et al.* Differential association of circadian genes with mood disorders: CRY1 and NPAS2 are associated with unipolar major depression and CLOCK and VIP with bipolar disorder. *Neuropsychopharmacology* **35**, 1279–1289. <https://doi.org/10.1038/npp.2009.230> (2010).
- Gekakis, N. *et al.* Role of the CLOCK protein in the mammalian circadian mechanism. *Science* **280**, 1564–1569. <https://doi.org/10.1126/science.280.5369.1564> (1998).
- Hogenesch, J. B., Gu, Y. Z., Jain, S. & Bradfield, C. A. The basic-helix-loop-helix-PAS orphan MOP3 forms transcriptionally active complexes with circadian and hypoxia factors. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 5474–5479 (1998).
- King, D. P. *et al.* Positional cloning of the mouse circadian clock gene. *Cell* **89**, 641–653 (1997).
- Vielhaber, E. L., Duricka, D., Ullman, K. S. & Virshup, D. M. Nuclear export of mammalian PERIOD proteins. *J. Biol. Chem.* **276**, 45921–45927. <https://doi.org/10.1074/jbc.M107726200> (2001).
- Busino, L. *et al.* SCFFbxl3 controls the oscillation of the circadian clock by directing the degradation of cryptochrome proteins. *Science* **316**, 900–904. <https://doi.org/10.1126/science.1141194> (2007).
- Xing, W. *et al.* SCF(FBXL3) ubiquitin ligase targets cryptochromes at their cofactor pocket. *Nature* **496**, 64–68. <https://doi.org/10.1038/nature11964> (2013).
- Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E. & Hogenesch, J. B. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 16219–16224. <https://doi.org/10.1073/pnas.1408886111> (2014).
- Dodson, E. R. & Zee, P. C. Therapeutics for circadian rhythm sleep disorders. *Sleep Med. Clin.* **5**, 701–715. <https://doi.org/10.1016/j.jsmc.2010.08.001> (2010).
- Hou, Y., Liu, L., Chen, X., Li, Q. & Li, J. Association between circadian disruption and diseases: A narrative review. *Life Sci.* **262**, 118512. <https://doi.org/10.1016/j.lfs.2020.118512> (2020).
- Onat, O. E. *et al.* Human CRY1 variants associate with attention deficit/hyperactivity disorder. *J. Clin. Invest.* **130**, 3885–3900. <https://doi.org/10.1172/JCI135500> (2020).
- Pellegrino, R. *et al.* A novel BHLHE41 variant is associated with short sleep and resistance to sleep deprivation in humans. *Sleep* **37**, 1327–1336. <https://doi.org/10.5665/sleep.3924> (2014).
- Wallach, T. & Kramer, A. Chemical chronobiology: Toward drugs manipulating time. *FEBS Lett.* **589**, 1530–1538. <https://doi.org/10.1016/j.febslet.2015.04.059> (2015).
- Chen, Z. *et al.* Identification of diverse modulators of central and peripheral circadian clocks by high-throughput chemical screening. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 101–106. <https://doi.org/10.1073/pnas.1118034108> (2012).

22. Doruk, Y. U. *et al.* A CLOCK-binding small molecule disrupts the interaction between CLOCK and BMAL1 and enhances circadian rhythm amplitude. *J. Biol. Chem.* **295**, 3518–3531. <https://doi.org/10.1074/jbc.RA119.011332> (2020).
23. He, B. *et al.* The small molecule nobiletin targets the molecular oscillator to enhance circadian rhythms and protect against metabolic syndrome. *Cell Metab.* **23**, 610–621. <https://doi.org/10.1016/j.cmet.2016.03.007> (2016).
24. Hirota, T. *et al.* High-throughput chemical screen identifies a novel potent modulator of cellular circadian rhythms and Reveals CKI alpha as a clock regulatory kinase. *Plos Biol.* **8**, e1000559. <https://doi.org/10.1371/journal.pbio.1000559> (2010).
25. Hirota, T. *et al.* Identification of small molecule activators of cryptochrome. *Science* **337**, 1094–1097. <https://doi.org/10.1126/science.1223710> (2012).
26. Miller, S. *et al.* Isoform-selective regulation of mammalian cryptochromes. *Nat. Chem. Biol.* **16**, 676–685. <https://doi.org/10.1038/s41589-020-0505-1> (2020).
27. Oshima, T. *et al.* C-H activation generates period-shortening molecules that target cryptochrome in the mammalian circadian clock. *Angew. Chem. Int. Ed. Engl.* **54**, 7193–7197. <https://doi.org/10.1002/anie.201502942> (2015).
28. Stein, R. M. *et al.* Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* **579**, 609–614. <https://doi.org/10.1038/s41586-020-2027-0> (2020).
29. Lo, Y. C., Rensi, S. E., Tornig, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **23**, 1538–1546. <https://doi.org/10.1016/j.drudis.2018.05.010> (2018).
30. Polishchuk, P. Interpretation of quantitative structure-activity relationship models: Past, present, and future. *J. Chem. Inf. Model.* **57**, 2618–2639. <https://doi.org/10.1021/acs.jcim.7b00274> (2017).
31. Sakakibara, Y. *et al.* COPICAT: A software system for predicting interactions between proteins and chemical compounds. *Bioinformatics* **28**, 745–746. <https://doi.org/10.1093/bioinformatics/bts031> (2012).
32. Kauffman, G. W. & Jurs, P. C. QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *J. Chem. Inf. Comp. Sci.* **41**, 1553–1560. <https://doi.org/10.1021/ci010073h> (2001).
33. Kotzabasaki, M. I., Sotiropoulos, I. & Sarimveis, H. QSAR modeling of the toxicity classification of superparamagnetic iron oxide nanoparticles (SPIONs) in stem-cell monitoring applications: an integrated study from data curation to model development. *Rsc Adv.* **10**, 5385–5391. <https://doi.org/10.1039/c9ra09475j> (2020).
34. Nagai, J., Imamura, M., Sakagami, H. & Uesawa, Y. QSAR prediction model to search for compounds with selective cytotoxicity against oral cell cancer. *Medicines (Basel)* <https://doi.org/10.3390/medicines6020045> (2019).
35. Nastase, A. F. & Boyd, D. B. Simple structure-based approach for predicting the activity of inhibitors of beta-secretase (BACE1) associated with Alzheimer's disease. *J. Chem. Inf. Model.* **52**, 3302–3307. <https://doi.org/10.1021/ci300331d> (2012).
36. Ponzoni, I. *et al.* QSAR classification models for predicting the activity of inhibitors of beta-secretase (BACE1) associated with Alzheimer's disease. *Sci. Rep.* **9**, 9102. <https://doi.org/10.1038/s41598-019-45522-3> (2019).
37. Santana, L. *et al.* A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. *J. Med. Chem.* **49**, 1149–1156. <https://doi.org/10.1021/jm0509849> (2006).
38. Xing, W. M. *et al.* SCFFBXL3 ubiquitin ligase targets cryptochromes at their cofactor pocket. *Nature* **496**, 64. <https://doi.org/10.1038/nature11964> (2013).
39. Huang, N. *et al.* Crystal structure of the heterodimeric CLOCK: BMAL1 transcriptional activator complex. *Science* **337**, 189–194. <https://doi.org/10.1126/science.1222804> (2012).
40. Nangle, S. N. *et al.* Molecular assembly of the period-cryptochrome circadian transcriptional repressor complex. *Elife* **3**, e03674. <https://doi.org/10.7554/eLife.03674> (2014).
41. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802. <https://doi.org/10.1002/jcc.20289> (2005).
42. MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616. <https://doi.org/10.1021/jp973084f> (1998).
43. Dagliyan, O., Uney-Yuksektepe, F., Kavakli, I. H. & Turkyay, M. Optimization based tumor classification from microarray gene expression data. *PLoS ONE* **6**, e14579. <https://doi.org/10.1371/journal.pone.0014579> (2011).
44. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791. <https://doi.org/10.1002/jcc.21256> (2009).
45. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461. <https://doi.org/10.1002/jcc.21334> (2010).
46. Cal-Kayitmazbatir, S. *et al.* CRY1-CBS binding regulates circadian clock function and metabolism. *FEBS J.* **288**, 614–639. <https://doi.org/10.1111/febs.15360> (2021).
47. Zielinski, T., Moore, A. M., Troup, E., Halliday, K. J. & Millar, A. J. Strengths and limitations of period estimation methods for circadian data. *PLoS ONE* **9**, e96462. <https://doi.org/10.1371/journal.pone.0096462> (2014).
48. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784. <https://doi.org/10.1038/nmeth.3047> (2014).
49. Korkmaz, T. *et al.* Opposite carcinogenic effects of circadian clock gene BMAL1. *Sci. Rep.* **8**, 16023. <https://doi.org/10.1038/s41598-018-34433-4> (2018).
50. Zielinski, E. L., Deibel, S. H. & McDonald, R. J. The trouble with circadian clock dysfunction: Multiple deleterious effects on the brain and body. *Neurosci. Biobehav. Rev.* **40**, 80–101. <https://doi.org/10.1016/j.neubiorev.2014.01.007> (2014).
51. Dong, J. *et al.* ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminform.* **7**, 60. <https://doi.org/10.1186/s13321-015-0109-z> (2015).
52. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422. <https://doi.org/10.1023/A:1012487302797> (2002).
53. Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. *Classification and Regression Trees* (CRC Press, 1984).
54. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. <https://doi.org/10.1023/A:1010933404324> (2001).
55. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42. <https://doi.org/10.1007/s10994-006-6226-1> (2006).
56. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232. <https://doi.org/10.1214/aos/1013203451> (2001).
57. Chen, T. Q. & Guestrin, C. XGBoost: A scalable tree boosting system. *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785> (2016).
58. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
59. Sato, T. K. *et al.* Feedback repression is required for mammalian circadian clock function. *Nat. Genet.* **38**, 312–319. <https://doi.org/10.1038/ng1745> (2006).
60. Ueda, H. R. *et al.* System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nat. Genet.* **37**, 187–192. <https://doi.org/10.1038/ng1504> (2005).
61. Hua, P. *et al.* Cry1 and Tef gene polymorphisms are associated with major depressive disorder in the Chinese population. *J. Affect. Disord.* **157**, 100–103. <https://doi.org/10.1016/j.jad.2013.11.019> (2014).
62. Kovanen, L., Donner, K., Kaunisto, M. & Partonen, T. CRY1, CRY2 and PRKCDBP genetic variants in metabolic syndrome. *Hypertens. Res.* **38**, 186–192. <https://doi.org/10.1038/hr.2014.157> (2015).

63. Patke, A. *et al.* Mutation of the human circadian clock gene CRY1 in familial delayed sleep phase disorder. *Cell* **169**, 203–215. <https://doi.org/10.1016/j.cell.2017.03.027> (2017).
64. Ozber, N. *et al.* Identification of two amino acids in the C-terminal domain of mouse CRY2 essential for PER2 interaction. *BMC Mol. Biol.* **11**, 69. <https://doi.org/10.1186/1471-2199-11-69> (2010).
65. Partch, C. L., Clarkson, M. W., Ozgur, S., Lee, A. L. & Sancar, A. Role of structural plasticity in signal transduction by the cryptochrome blue-light photoreceptor. *Biochemistry* **44**, 3795–3805. <https://doi.org/10.1021/bi047545g> (2005).
66. Michael, A. K. *et al.* Formation of a repressive complex in the mammalian circadian clock is mediated by the secondary pocket of CRY1. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1560–1565. <https://doi.org/10.1073/pnas.1615310114> (2017).
67. Gul, S. *et al.* The Arg-293 of cryptochrome 1 is responsible for the allosteric regulation of CLOCK-CRY1 binding in circadian rhythm. *J. Biol. Chem.* **295**, 17187–17199. <https://doi.org/10.1074/jbc.RA120.014333> (2020).
68. Rosensweig, C. *et al.* An evolutionary hotspot defines functional differences between CRYPTOCHROMES. *Nat. Commun.* **9**, 1138. <https://doi.org/10.1038/s41467-018-03503-6> (2018).
69. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**, 2719–2740. <https://doi.org/10.1021/jm901137j> (2010).
70. Ellson, J., Gansner, E., Koutsofios, L., North, S. C. & Woodhull, G. Graphviz-open source graph drawing tools. *Lect. Notes Comput. Sci.* **2265**, 483–484 (2002).
71. Anafí, R. C. *et al.* Machine learning helps identify CHRONO as a circadian clock component. *Plos. Biol.* **12**, e1001840. <https://doi.org/10.1371/journal.pbio.1001840> (2014).
72. Zhang, E. E. *et al.* A genome-wide RNAi screen for modifiers of the circadian clock in human cells. *Cell* **139**, 199–210. <https://doi.org/10.1016/j.cell.2009.08.031> (2009).
73. Bording, T., Abdo, A. N., Maier, B., Gabriel, C. & Kramer, A. Generation of human CRY1 and CRY2 knockout cells using duplex CRISPR/Cas9 technology. *Front. Physiol.* **10**, 577. <https://doi.org/10.3389/fphys.2019.00577> (2019).
74. Vitaterna, M. H. *et al.* Differential regulation of mammalian period genes and circadian rhythmicity by cryptochromes 1 and 2. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 12114–12119. <https://doi.org/10.1073/pnas.96.21.12114> (1999).
75. Yang, S. Y. Pharmacophore modeling and applications in drug discovery: Challenges and recent advances. *Drug Discov. Today* **15**, 444–450. <https://doi.org/10.1016/j.drudis.2010.03.013> (2010).
76. Perkins, R., Fang, H., Tong, W. & Welsh, W. J. Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem.* **22**, 1666–1679. <https://doi.org/10.1897/01-171> (2003).

Acknowledgements

This work is supported by TUBITAK-SBAG 215S021. The data used in this study was derived from PhD thesis of Şeref Gül.

Author contributions

S.G. and FR performed work and wrote draft of the manuscript. SI performed circadian rhythm works. FY and NO generated CRY1 knockout cell line. IHK and MT wrote paper and analyzed the data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-97962-5>.

Correspondence and requests for materials should be addressed to M.T. or I.H.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021