Ignition of Small Molecule Inhibitors in Friedreich's Ataxia with Explainable Artificial Intelligence

Kevser Kübra Kırboğa (✉ kubra.kirboga@yahoo.com)

Süleyman Demirel University

Ecir Uğur Küçüksille

Süleyman Demirel University

Utku Köse

Süleyman Demirel University

Research Article

Keywords: Explainable Artificial Intelligence, Friedreich Ataxia, Predictive accuracy, Quantitative structure-activity relationship, QSAR, Shapley values

Posted Date: March 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1408745/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Iron (Fe) chelating medicines and Histone deacetylase (HDAC) inhibitors are two therapy options for hereditary Friedreich's Ataxia that have been shown to improve clinical results (FA). Fe chelation molecules can be used to minimize the quantity of stored Fe, and HDAC inhibitors can be used to boost the expression of the Frataxin (FXN) gene in the process of enhancing FA. A complete quantitative structure-activity relationship (QSAR) search of inhibitors from the ChEMBL database is reported in this paper, which includes 437 compounds for Fe chelation and 1,354 compounds for HDAC inhibitors. For further investigation, the IC₅₀ was chosen as the unit of bioactivity, and following data refinement, a final dataset of 436 and 1,163 compounds for Fe chelation and HDAC inhibition, respectively, was produced. The Random Forest (RF) technique was used to generate models, and the models created using the PubChem fingerprint were the strongest of the 12 fingerprint kinds, hence that feature was chosen for interpretation. The results showed the importance of properties related to nitrogen-containing functional groups and aromatic rings. As a result, we explained the effect of the molecular fingerprints used on the models and therefore the effects on possible drugs that can be developed for FA with artificial intelligence (XAI), which can be explained through SHAP (Shapley Additive Explanations) values. Model scripts and fingerprinting methods are also available at <https://github.com/tissueandcells/XAI>.

Highlights

- Friedreich's Ataxia (FA/FRDA) is one of the most widespread hereditary diseases with a wide range of symptoms and no definitive proven treatment.
- Computational methods for the controlled and easy production of innovative chemical entities hold promise for drug discovery.
- In the process of defining target chemicals, the need for explicable deep learning models is increasing to increase the ability of human interpretation of mathematical models.

Introduction

The process that reveals a therapeutically useful compound for use in curing and treating diseases is termed drug discovery. It gets approximately 12-15 years from the first exploration of a drug to the marketing process. Identification, synthesis, characterization, optimization, and determination of therapeutic efficacy values of this compound is a very long and challenging process [1]. To overcome this complexity, the use of molecular fingerprints has been proposed to generalize previously defined patterns. In this process, the input data encoded with the SMILES (Simplified Molecular Input Line Entry Specification) information of the compounds can be trained. Therefore, the relationship between structure and biological, chemical, and physicochemical activities is revealed [2].

In recent years, the increasing interest in computer-assisted drug discovery (CADD) has resulted in drug discovery studies associated with artificial intelligence. A new compound design, de novo designs,

structure and ligand-based designs, estimation of the physicochemical and pharmacokinetic properties of the drug, and machine learning algorithms for drug reuse are applied for CADD [3-5].

In addition to deep learning methods used to predict properties based on input data and to determine nonlinear input-output relationships, machine learning approaches that emphasize molecular descriptors also increase the synergy of drug discovery and cheminformatics [6].

Explainable artificial intelligence (XAI) has great potential to increase the intelligibility of models emerging from input data and transform it into a more human-interpretable format. As AI becomes associated with drug discovery and its use in related fields, there is a growing demand for the use of XAI in the interpretation of fundamental models. Additionally, to mathematical models, XAI supports the design of new drugs, the extraction of pharmacological activities from molecular structures, and the design of new bioactive compounds with desired properties, with ways to make the basic decision-making process transparent, increase interpretability and avoid false predictions in the drug discovery process. XAI studies, which are at the beginning of the growth period, are progressing rapidly and it is thought that the use of XAI studies in drug discovery and pharmacological studies will increase in the coming years [7]. Friedreich's ataxia (FRDA/FA) is a progressive and neurodegenerative genetic disorder that usually occurs between the ages of 10-15 years. At the onset of FA, the first symptoms include unbalanced gait, frequent falls, and impaired movement coordination ability [8,9]. It is the most prevalent inherited disease that occurs with the expansion of the GAA triplet (triplet repeat) located in intron 1 of the Frataxin gene on chromosome 9q13 [10]. There is a linear correlation between the expanded repeats size and the phenotypic severity of FA. Frataxin (FXN) protein (218 aa., 18 kDa), which is the protein of the frataxin gene, is a mitochondrial protein involved in Fe metabolism. FXN deficiency in FA, which is defined as a nuclear-encoded mitochondrial disease, causes Fe accumulation in the mitochondria, disruption of mitochondrial enzymes, and creates a sensitivity to oxidative stress. Due to this sensitivity, free radical-mediated cell death occurs [11-13]

Furthermore, FA patients may have variable clinical diseases such as heart disease, diabetes mellitus, and glucose intolerance. The pathogenesis of FA has not yet been resolved. Therefore, there is no current treatment method for FA. Bulteau et al. found that mitochondrial Fe has a major key role in the development and advancement of the disease in their studies on *S. cerevisiae* yeast [14].

The clinical signs and symptoms associated with FA are the result of degenerative changes in structures known as dorsal root ganglia. Nerve fibers in the spinal cord degenerate and cause a lack of signals to the cerebellum, which is involved in the voluntary movement coordination of the brain [15]. Decreased iron-sulfur cluster and heme formation in FA cause Fe accumulation in mitochondria. When Boddaert et al. evaluated the reduction of Fe accumulation by using appropriate iron chelators, they showed that the use of chelators benefited the reduction in Fe accumulation [16]. It is also known that Fe chelator agents such as deferiprone, idebenone, and desferoxamine are effective in moderate Fe accumulation [17-19].

Another treatment modality based on research and clinical results is gene silencing in FXN alleles with extended repeats. This heterochromatin-mediated method involves a histone deacetylase (HDAC)

inhibitor that reverses FXN gene silencing. These molecules provide therapeutic benefit by acetylating lysine residues on H4 and H4 [20,21]. Since Fe chelator molecules are effective in moderate Fe accumulation and HDACi's (HDAC inhibitors) are highly toxic or have low specificity, there is no current treatment method. Due to the need for the discovery of HDACi and Fe chelator molecules, the XAI method was used in this study to determine the central point required for the discovery of molecules and to target the central point for possible drug discovery for FA. Whether or not the molecules obtained from the databases carry molecular fingerprints from the bioactivity data were investigated and the importance of XAI and molecular fingerprints on these molecules was explained.

Research Gaps and Motivation

Due to the limited number of studies for FA and the inability to clearly define the function of the Frataxin (FXN) gene, there is no drug or gene therapy method yet.

- Frataxin protein, the product of the FXN gene, is thought to be involved in the Fe-S mechanism. Since the function studies of proteins are a difficult and complex process, there is no defined treatment method yet. Therefore, the selection of a faster and less costly computational procedure and the use of large datasets are considered a more appropriate ways to influence the symptomatic aspects of the disease.
- FA is a neuromuscular condition. Clarke's colon loses its lumbosacral and nerve cells, which are replaced with capsular cells. Loss of proprioception and sensory ataxia are symptoms of posterior colon degeneration. Similarly, the loss of sensory ganglia results in the lack of tendon reflexes. Kyphoscoliosis is a problem caused by a misalignment of the spinal muscles [22]. It may be a more accurate technique to analyze chemicals that alter the expression of the FXN gene to overcome this complex system.
- Although artificial intelligence developments are progressing rapidly, XAI in drug discovery is not yet at the desired stage. It can be developed by the synergetic efforts of research with different scientific backgrounds. In highly costly and time-sensitive situations such as drug discovery, deep learning, and machine learning applications, modeling choices and predictions need to be examined. With XAI, it makes sense to develop hybrid approaches that are easier to understand and computationally affordable without forgetting the limitations of drug discovery.

Related Works

Rapid advances in computational techniques and artificial intelligence have contributed significantly to medication discovery for a variety of genetic and metabolic illnesses in recent years. One of the biggest gifts computer technologies has given to drug discovery is the ability to anticipate the biological activity of molecules, opening up new avenues and possibilities for the development of new medications with stable features.

For studies in breast cancer, Schaduangrat et al. presented a comprehensive classification structure-activity relationship (CSAR) investigation of inhibitors from the ChEMBL database, which included a

starting set of 11,618 compounds for ER and 7,810 compounds for ER. They demonstrated the importance of nitrogen and aromatic bonded compounds [23].

Rodriguez-Perez et al. explored a variant for the precise calculation of Shapley values for decision tree methods and further extended the evaluation of the SHAP methodology by systematically comparing this variant with the model-free SHAP method in estimating the combined activity and potency value. In this way, the logic and convenience of the SHAP value method are shown [24].

In another similar study, they used the random forest algorithm to classify the antimicrobial activity and identify molecular identifiers that support the antimicrobial activity of the investigated peptides. As a result of the explanations of the key descriptors identified, it was revealed that polarity-resolution is required for membrane lytic antimicrobial activity [25].

Research Contributions

This study aims to predict bioactivity through machine learning for FA hereditary disease without drug and gene therapy and to detect molecular fingerprints affecting the model with XAI.

By looking at FA pathology and signal transduction pathways, molecules that can be used to reduce the amount of accumulated Fe and increase the expression level of the FXN gene were brought from the ChEMBL database. Fe chelation molecules have been introduced to reduce the amount of Fe, and HDAC inhibitors have been introduced to affect the gene expression level. While the bioactivity predictions of these molecules are made with machine learning, the effects of molecular fingerprints are also explained with XAI.

The following questions were asked for Fe chelation molecules and HDAC inhibitors that could be developed for FA:

- What is the distribution of bioactivity data of Fe chelation molecules and HDAC inhibitors that can be used for FA?
- What are the effects of the molecular fingerprints of Fe chelation molecules and HDAC inhibitors on the model?
- Is there a positive correlation between the SHAP values of each molecular fingerprint effective on the model and the data in the literature affecting the drug discovery and development process?

Materials And Detailed Methods

Data compilation and curation

Datasets were collected from the version 25 ChEMBL database for Fe chelation and HDAC [26]. A data improvement process was carried out with the IC50 selection for the bioactivity unit. Correspondingly, a data set of 436 and 1.163 compounds was obtained for Fe chelation and HDAC, respectively. Since this

study also aimed to establish a classification model for HDAC and Fe chelation to achieve the goal, we set threshold values <1 and $>10\ \mu\text{M}$ to distinguish active and inactive compounds. At the end of these processes, a final non-redundant and ameliorated dataset was obtained for Fe chelation and HDAC, consisting of 262 and 945 compounds, respectively.

Descriptors of Molecular

Molecular fingerprint identifiers for the compounds in the obtained datasets were calculated using the PaDEL-Descriptor software [27]. SMILES indicators were used to calculate these molecular descriptors. Compound structures have been standardized using functions included in the PaDEL software. Molecular fingerprints play a key role in QSAR studies as they identify molecules and characterize chemical structure information quantitatively and qualitatively. As listed in Table 1, Malik et al. (2020) used 12 molecular fingerprints in 9 classes to accurately identify chemical structures [27,28]

Data splitting

After obtaining a set of data that ML models can handle, this dataset is divided into two subgroups. One of these subgroups is the training set used to train the model and has a higher data percentage. The second group is the smaller dataset and is dedicated to testing the model. Normally, there are many variables in the creation of mathematical descriptors. The training set, on the other hand, aims to search for the best variable subset with the correct and necessary information. In this way, unnecessary variables are reduced as much as possible. To provide an understandable reason at the biological level, a subset of features is added to the original set without changing the content of the variables [29]. The model is trained after determining the best subset of variables. Overtraining should be avoided in order to keep the model's validity when dealing with unknown data. Cross-validation (CV) techniques are commonly used in these situations. CV provides for the measurement of the model's generalization degree during the training phase, evaluation of the model's performance, and performance estimation with unknown data. The original dataset is separated into two subgroups during each execution of the experiment in the CV (training set and validation set). The 10-fold CV approach was used in our study. The purpose of a CV is to enable you to select the best set of parameters. The performance of each model is measured using these parameters, and the model with the best performance is chosen. The best model's final validation is carried out. It can be considered that a novel predictive drug model has been built if the validation results are statistically significant [30,31].

Statistical analysis

In the research, we used minimum (Min), maximum (Max), median, and mean parameters for statistical analysis to determine the orientation of active and inactive compounds on the basis of descriptors at the stage of classification of compounds. The results were visualized with the Seaborn v0.8.1 Python package.[32].

Explainability with Shapley additive explanations (SHAP)

In this work, we were finally interested in understanding the effect of molecular fingerprints on models. Among many explainability techniques [33,34], there are several features that are advantageous in our SHAP selection. The most important thing is that these values are independent of the model. Like RandomForestRegressor [35]. They are not tied to a particular model type. It will also be suitable for the RandomForestRegressor model we chose for HDAC inhibitors. It also offers SHAP values, accuracy, consistency, and incompleteness [34-36]. Finally, SHAP applications are very easy to use and can be conveniently documented with images. Shapley values were originally proposed as a game theory used to pay players fairly based on their contribution to their total earnings [37]. In estimating a model, it means assigning their quantitative significance based on contributions. Therefore, in our study, the SHAP value can be defined as the average marginal contribution of the feature value across all possible coalitions of possible features. That is, a SHAP value for a given molecular fingerprint value can be explained as the difference between the actual prediction and the average prediction for the entire set of molecular fingerprints. When working on SHAP values, we should say that the estimated value does not differ after removing a corresponding feature [38]

The SHAP method calculates the Shapley values of each feature and represents it as a linear model of all features. The SHAP value is calculated with the following formula.

$$\phi(f(X_i)) \equiv \phi_0 + \sum_{k=1}^K \phi_k(X_i), \forall i = 1, \dots, n$$

where k denotes a single property variable; K denotes the total number of explanatory variables available; n is the total number of units that should be. $\phi \in \mathbb{R}^K$; $\phi_k \in \mathbb{R}$. $\phi_k(X_i)$ are the Shapley values of local functions [34-37,39,40].

Also, SHAP values use a specific index not only on individual features, but also among all binary features. In this way, SHAP values can easily explain the modeling of interactions that would go unnoticed. This feature is important because it provides a clearer understanding of the variables in the model and the relationship between them [41].

Using the data of Fe chelation and HDAC inhibitors with the Python application, we calculated the SHAP values of the best performing ML-based model as a training and test set.

We used SHAP values to visualize the importance of molecular fingerprints for the models (RandomForestRegressor for Fe chelation, RandomForestRegressor for HDACi). Next, we generated SHAP value plots for both models and compared how these molecular fingerprints contribute to the output of these models and their significance in the models. We analyzed the important interactions between these molecular fingerprints and their targets [42].

In conclusion, we determined individual (local) and aggregated (global) plots to understand how variables affect model results in Fe chelation and HDAC inhibitors. Next, we determined the SHAP

findings, the effect of Fe chelation molecules and HDAC inhibitors on the drug discovery process.

Model Development

The current study uses Fe chelation and HDAC for FA in the prediction model and PubChem molecular fingerprints as predictors. The predictive capacity of the model is checked using three statistical measures: Coefficient of Determination (R^2), mean square error (MSE), and Mean-Absolute-Error (MAE). The mathematical formulas for these metrics are:

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = determination coefficient

RSS = sum of squares of residuals

TSS = total sum of squares

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

MAE = mean absolute error

y_i = prediction

x_i = true value

$$R^2 = 1 - \frac{RSS}{TSS}$$

= total number of data points

Mechanistic interpretations of SelectFromModel

SelectFromModel is a meta-converter that uses the attribute of the estimator to evaluate and sort features and converts them to a dataset according to their order. In this way, the estimator is adapted to the training data and the importance of each feature is calculated on the model. It ranks the meta-transformer properties according to their importance and selects the most suitable ones depending on the selected threshold value. In our study, the threshold value was chosen as 0.005 and 0.003 Fe chelation and HDACi, respectively [43].

To gain a better understanding of the mechanistic details ruling Fe chelation and HDACi compounds, a property significance analysis was performed on selected informative descriptors. Because of the built-in capability of feature importance estimation and excellent prediction performance of the RandomForestRegressor and RandomForestRegressor model for Fe chelation and HDACi, respectively, they were used for analysis in this action. We used the SelectFromModel method to rank the importance of PubChem property descriptors. The top 10 PubChem identifiers of the SHAP value deduced from the SHAP value derived from the RandomForestRegressor models can be found in Fig. 4 and their infrastructures will contribute to the overall functioning of the compounds. It is discussed in the section below. In the proposed methodology, the dataset was trained and tested with 8 different ML models (Table 2-3). To increase the prediction accuracy, the SelectFromModel feature selection technique was included in the methodology and important features were tested with machine learning models again [44].

Results And Discussion

The aim of this study is to evaluate the activity and bioactivity of Fe chelator molecules and HDA inhibitors for FA genetic disease. This study is also to determine which molecular fingerprints are dominant on the model with XAI in the development of molecules. As shown in Fig. 1 data were brought from the ChEMBL database after necessary targeting related to the deficiency of Frataxin protein due to Fe accumulation causing FA and insufficient expression of the FXN gene. After the bioactivity predictions of these molecules were made, a machine learning method was applied according to whether they contain molecular fingerprints or not. Finally, molecular fingerprints that contributed the most to the model were determined by visualizing with XAI.

Chemical Space Analysis

Chemical space analysis is used to find typical characteristics between active and inactive chemicals in the classification of substances. Ratio of molecular weight (MW) to Ghose–Crippen–Viswanadhan octanol-water partition coefficient (ALogP) for general analysis of compounds, Lipinski's five rules (Ro5) (weight (<500), octanol-water partition coefficient) for comparison of drug properties of compounds (ALogP < 5), number of hydrogen bond acceptors (<10), and hydrogen bond donors (<5)) were used as descriptors [45]. Chemical space analysis of MW-LogP is visualized in Fig 2. As can be observed from the

graphs generated for Fe chelation and HDAC, most of the compounds clustered in the 250-550 Da MW and 2-6 ALogP range.

Besides, it was observed that Fe chelation compounds have higher nHBD_{on} and nHB_{Acc} values in active compounds, while nHBD_{on} and nHB_{Acc} values do not present a significant difference for active and inactive compounds in HDAC_i.

QSAR modeling

We followed the Organization for Economic Co-operation and Development (OECD) guidelines to develop an interpretable QSAR model [46]. These guidelines consist of the following main points: (i) datasets have a defined endpoint, (ii) the algorithm for learning is clear, (iii) the area to which the QSAR model will be applied is well defined, (iv) measures of predictability, mechanistic interpretations, and robustness are available [28]. For a robust QSAR model, molecular fingerprint descriptors were applied in this study using PaDEL-Descriptor software. [27]. The interpretable properties from the 12 fingerprints (ie PubChem, Substructure, and Klekota–Roth) are listed on our Github page (<https://github.com/tissueandcells/XAI>).

Model Selection

The R², MAE and MSE values of the models created with the training data subjected to CV and the test data used to measure the performance of the models were measured. As shown in Table 2, the RandomForestRegressor model gave the highest R² values (0.701 and 0.572 for training and testing, respectively) for Fe chelation. It also has low MSE and MAE values. As shown in Table 3, for HDAC inhibition, the XGBRegressor model gave the highest R² values (0.983 and 0.619 for training and testing, respectively) and lowest MAE and MSE values.

Table 2 Statistical values of 8 models used according to training and test data for Fe chelation.

Model Name	Train R2 Value	Test R2 Value	Train MAE Value	Test MAE Value	Train MSE Value	Test MSE Value
RandomForestRegressor	0.701	0.572	0.375	0.511	0.254	0.534
XGBRegressor	0.746	0.428	0.299	0.550	0.216	0.697
DecisionTreeRegressor	0.749	0.392	0.280	0.542	0.214	0.740
MLPRegressor	0.434	0.549	0.571	0.563	0.482	0.549
BaggingRegressor	0.677	0.534	0.376	0.522	0.274	0.568
LinearRegression	0.422	-3.195	0.576	297	0.492	3.900
Support Vector Machine	0.514	0.523	0.494	0.572	0.414	0.581
Ridge Regression	0.421	0.576	0.578	0.539	0.494	0.517

Table 3 Statistical values of 8 models used according to training and test data for HDAC.

Model Name	Train R2 Value	Test R2 Value	Train MAE Value	Test MAE Value	Train MSE Value	Test MSE Value
RandomForestRegressor	0.892	0.460	0.274	0.575	0.163	0.587
XGBRegressor	0.912	0.386	0.219	0.132	0.588	0.668
DecisionTreeRegressor	0.915	-0.054	0.197	0.854	0.128	1.149
MLPRegressor	0.758	-0.06	0.442	0.906	0.367	1.158
BaggingRegressor	0.883	0.429	0.279	0.598	0.176	0.622
LinearRegression	0.523	0.241	0.638	0.670	0.725	0.826
Support Vector Machine	0.741	0.252	0.398	0.716	0.392	0.814
Ridge Regression	0.577	0.207	0.608	0.706	0.642	0.863

We used the SelectFromModel method to grade the importance of PubChem property descriptors. The importance levels of the selected molecular fingerprints are listed in the bar graphs in Fig. 4.

Explainability with Shapley additive explanations (SHAP)

With the SHAP method, the effect of the selected features on the functioning of the model was determined. The contribution of each attribute to the model is visualized with SHAP values. As shown in Fig. 5 and 6, bar and force graphs were created locally, that is, according to the degree of importance on each individual sample. Beeswarm, heatmap, and decision graphics are general graphics. Molecular fingerprints on all compounds are ranked according to their importance (Fig. 7,8,9).

A summary of the top 10 features in the iron chelation and HDAC model, along with their corresponding SMARTS patterns and explanations, is detailed on GitHub(<https://github.com/tissueandcells/XAI>).

Aromatic fingerprints

According to the values obtained from the SHAP values and images for Fe chelation, it was seen that the maximum number of PubChem fingerprints was associated with aromaticity. In other words, PubChemFP152, PubChemFP374, PubChemFP485, PubChemFP821, PubChemFP188, PubChemFP672, PubChemFP187 fingerprints were associated with aromaticity and ring structures as seen in the explanations. Deferiprone (Ferriprox) [47-49] and Clioquinol [50-52] drugs also support this situation. They are Fe chelation drugs that center the annular structures. Molecular fingerprints such as PubChemFP613, PubchemFP656, PubChemFP672, PubChemFP485 and PubChemFP821, which explain the importance of single and double bonds made by cyclic structures, also reflect the content of many existing drugs. For example, alpha-tocopherol quinone, which is synthetic molecule, and comparable to Coenzyme Q 10 but has a higher redox potential, is believed to reduce mitochondrial oxidative stress. Considering the structure of this compound, single and double bonds in aromatic rings draw attention. However, no studies are in progress despite significant improvement over 4 points in a dose-dependent manner, especially in the placebo group [53]. Thanks to the information given by XAI, we understand that we need to focus on pro-drug studies, which are in a positive process. Fe overload is a process that predisposes to oxidative stress and tissue damage. Since Fe overload paves the way for oxidative stress, information can also be obtained about drug molecules targeting oxidative stress in an indirect way.

The similarity of the molecular background represented in PubChemFP116 with the structure of VP-20629 indole-3-propionic acid (also known as SHP622, OXIGON or OX1) is noteworthy. This antioxidant, which has neuroprotective properties, was developed for Alzheimer's disease with its capacity to prevent the production of beta-amyloid fibrils [54]. However, studies were conducted by including 46 FA participants. While tolerance was observed at all doses, great benefits were not observed.

Nitrogen-containing fingerprints

In the SHAP values for HDAC, nitrogen atoms appear in some molecular fingerprints. PubchemFP656, PubChemFP338, PubChemFP397, PubchemFP517, PubchemFP791, and PubchemFP645 belong to the nitrogen-containing class. These nitrogen-containing molecular fingerprints are in the group of amines and amides and constitute a high property number for HDAC. It is known that N atom is in aromatic rings

in hydroxamic acid and benzamide structures, which are dominant inhibitors. This also backs up the SHAP findings. Similarly, the presence of the CH group in aromatic ring systems with N atoms is the most common bioisosteric conversion utilized to simulate natural ligand binding while creating antagonistic consequences. [55].

The fact that 10 out of 10 feature selections are nitrogen-themed features also highlighted the importance of this. On the side, PubChemFP152, one of the leading molecular fingerprints for Fe chelation, represents 5-membered ring compounds containing at least one nitrogen and two heteroatoms. These groups are called azoles. Thiazoles and isothiazoles contain a nitrogen and a sulfur atom in their ring structure. It has been supported by various studies that thiazoles can be used in antiepileptic drugs [56]. Since the muscle symptoms in the mechanism of epilepsy are similar to FA, it seems that azole groups may actually be promising targets for drug discovery. The sulfur content of PubChemFP489, PubChemFP349 features in the prominent fingerprints for HDAC is also noteworthy. This result may be an effective molecular fingerprint in reducing the symptoms of FA genetic disease. Thiamine (Vitamin B1) will also be a shining example for molecular fingerprints where nitrogen-based fingerprints are intense. N atoms and cyclic structures in the structure of thiamine are dominant and their deficiency causes serious neurological changes in the nervous systems, both central and peripheral. The molecular and clinical changes of thiamine deficiency, such as impaired oxidative stress metabolism, increased oxidative stress, and selective neuronal loss, are like FA genetic disease. When Thiamine, which has N and ring structures, was evaluated in 34 FA patients, improvement was observed in tendon reflexes and thickening of ventricular septum. These results also showed that it can be a therapeutic [57].

Conclusions

FA, the most widespread inherited ataxia, accounting for around half of all ataxia cases and for seventy percent of individuals under the age of 25 [22]. However, current inhibitors and gene therapy, on the other hand, are insufficient since they are unable to effectively alter stored iron and create a change in gene expression. Therefore, this study addressed these issues qualitatively and quantitatively by constructing a QSAR model for Fe chelation and HDAC that could distinguish active compounds from inactive compounds. The activity prediction of these molecules was appraised through machine learning algorithm and various fingerprint identifiers classes. The results showed that combining the RF technique with PubChem fingerprints produced the most interpretable identifiers as well as the highest performing model. Aromaticity and amine groups are significant for active compounds, according to the property analysis of the important infrastructure contributions from the SHAP values. Much research has not focused on FA disease, thus therapeutic compounds are worth investigating further. As a result, the findings of this study can be used as a general guideline for developing potentially active and selective Fe chelation compounds and HDAC inhibitors based on data. Documentation of the work is available on Github (<https://github.com/tissueandcells/XAI>).

The primary purpose of this paper is to apply AI-based transparency and explainability models to expand the number of therapeutic compounds available for usage in FA hereditary disorder. This will support

potential drug development and discovery and help build confidence. The following are the study's main findings:

- Deep learning could be a useful tool for predicting pharmaceutical manufacturing and development, but its broad applicability will need to be verified in a variety of inherited disorders. The use of explainable algorithms like SHAP is critical for increasing trust and transparency in the usage of machine learning models so that data-driven models may be used in medication and chemical development. This study discovered the importance of Fe chelation and HDAC inhibitors as predictors in the prediction model of molecular fingerprints contained using the SHAP annotator.
- Future studies should focus on analyzing SHAP plots for long-term prediction as well as examining the SHAP properties of therapeutic studies and other contributing molecules that may influence FA.
- XAI, which has various benefits in terms of cost and time to the drug discovery process, also has negative aspects such as low model accuracy, conflicting molecular fingerprints or choosing the right threshold. Expanding computational studies, increasing their applicability, in vitro and in vivo experiments, and XAI and drug discovery studies will support each other and are thought to be promising for future drug studies.

Declarations

Author Contributions

KKK; researching, methodology, writing. EUK; software, editing, methodology, data analysis; UK; revising, methodology, data analysis, software.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data Availability

The data that support the findings of this study are available within the article and its supplementary material.

Acknowledgement

This research is the product of a thesis work.

References

1. Deore A, Dhumane J, Wagh R, Sonawane R (2019) Asian J Pharm Res Dev 7:62
2. Staszak M, Staszak K, Wieszczycka K, Bajek A, Roszkowski K, Tylkowski B (2021) Wiley Interdisciplinary Reviews:Computational Molecular Science

3. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, Zhao S (2019) *Nat Rev Drug Discov* 18(6):463
4. Lavecchia A (2015) *Drug Discov Today* 20(3):318
5. Lo YC, Rensi SE, Torng W, Altman RB (2018) *Drug Discov Today* 23(8):1538
6. Xue L, Bajorath J (2000) *Comb Chem High Throughput Screen* 3(5):363
7. Jiménez-Luna J, Grisoni F, Schneider G (2020) *Nat Mach Intell* 2(10):573
8. Cook A, Giunti P (2017) *Br Med Bull* 124(1):19
9. Lew S-Y, Yow Y-Y, Lim L-W, Wong K-H (2020) *Food Sci Technol* 40(suppl 1):264
10. Alper G, Narayanan V (2003) *Pediatr Neurol* 28(5):335
11. Wilson RB (2006) *Semin Pediatr Neurol* 13(3):166
12. Seznec H, Simon D, Bouton C, Reutenauer L, Hertzog A, Golik P, Procaccio V, Patel M, Drapier JC, Koenig M, Puccio H (2005) *Hum Mol Genet* 14(4):463
13. Gottesfeld JM (2019) *Neurotherapeutics* 16(4):1032
14. Bulteau AL, Dancis A, Gareil M, Montagne JJ, Camadro JM, Lesuisse E (2007) *Free Radic Biol Med* 42(10):1561
15. Pandolfo M (1999) *Semin Neurol* 19(3):311
16. Boddaert N, Le Quan Sang KH, Rötig A, Leroy-Willig A, Gallet S, Brunelle F, Sidi D, Thalabard J-C, Munnich A, Cabantchik ZI (2007) *Blood* 110(1):401
17. Soriano S, Llorens JV, Blanco-Sobero L, Gutiérrez L, Calap-Quintana P, Morales MP, Moltó MD, Martínez-Sebastián MJ (2013) *Gene* 521(2):274
18. Goncalves S, Paupe V, Dassa EP, Rustin P (2008) *BMC Neurol* 8(1):20
19. Pandolfo M, Hausmann L (2013) *J Neurochem* 126:142
20. Soragni E, Xu C, Cooper A, Plasterer HL, Rusche JR, Gottesfeld JM (2011) *Methods in molecular biology* 793:495
21. Herman D, Jenssen K, Burnett R, Soragni E, Perlman SL, Gottesfeld JM (2006) *Nat Chem Biol* 2(10):551
22. Aranca TV, Jones TM, Shaw JD, Staffetti JS, Ashizawa T, Kuo S-H, Fogel BL, Wilmot GR, Perlman SL, Onyike CU, Ying SH, Zesiewicz TA (2016) *Neurodegenerative Disease Management* 6(1):49
23. Schaduangrat N, Malik AA, Nantasenamat C (2021) *PeerJ* 9:e11716
24. Rodríguez-Pérez R, Bajorath J (2020) *J Comput Aided Mol Des* 34(10):1013
25. Li H, Nantasenamat C (2019) *PeerJ* 7:e8265
26. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR (2017) *Nucleic Acids Res* 45(D1):D945
27. Yap CW (2011) *J Comput Chem* 32(7):1466

28. Malik AA, Phanus-Umporn C, Schaduangrat N, Shoombuatong W, Isarankura-Na-Ayudhya C, Nantasenamat C (2020) *J Comput Chem* 41(20):1820
29. Saeys Y, Inza I, Larranaga P (2007) *Bioinformatics* 23(19):2507
30. Carracedo P, Blanco J, Rodriguez-Fernandez N, Cedrón-Santaeufemia F, Nóvoa F, Carballal A, Maojo V, Pazos A, Fernandez-Lozano C (2021) *Comput Struct Biotechnol J* 19:4538
31. Gramatica P, Sangion A (2016) *J Chem Inf Model* 56(6):1127
32. Waskom M, Botvinnik C, Qalieh (2017) Adel. Zenodo
33. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2019) *ACM-CSUR* 51(5):1
34. Lundberg S, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions
35. Hu L, Liu B, Ji J, Li Y (2020) *Journal of the American Heart Association* 9(22)
36. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK-W, Newman S-F, Kim J, Lee S-I (2018) *Nat Biomedical Eng* 2(10):749
37. Shapley L (1953) Princeton University Press, Princeton:307
38. Molnar C (2020) Self published
39. Lundberg SM, Erion GG, Lee (2019) S-I
40. Shapley LS (2016) 17. A Value for n-Person Games. In: Harold William K, Albert William T (eds) *Contributions to the Theory of Games (AM-28)*. Princeton University Press, Volume II, p 307
41. Li R, Shinde A, Liu A, Glaser S, Lyou Y, Yuh B, Wong J, Amini A (2020) *JCO Clinical Cancer Informatics*(4):637
42. Moncada-Torres A, Van Maaren MC, Hendriks MP, Siesling S, Geleijnse G (2021) *Scientific Reports* 11(1)
43. Stefanidou-Voziki P, Cardoner-Valbuena D, Villafafila-Robles R, Dominguez-Garcia JL (2021) Feature Selection and Optimization of a ML Fault Location Algorithm for Low Voltage Grids. 2021 IEEE International Conference on Environment and Electrical Engineering and 2021 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe): IEEE,
44. Shobana G, Priya DN (2021) Cancer Drug Classification using Artificial Neural Network with Feature Selection. Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV): IEEE, 2021
45. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) *Adv Drug Deliv Rev* 46(1):3
46. OECD O (2007) Environment health and safety publications series on testing and assessment No. 69, Guidance document on the validation of (quantitative) structure-activity relationships [(Q) SAR] models. OECD, Paris, France
47. Beutler E (2007) *Blood Cells. Molecules and Diseases* 39(2):140
48. Molina-Holgado F, Gaeta A, Francis PT, Williams RJ, Hider RC (2008) *J Neurochem* 105(6):2466
49. Reeder BJ, Wilson MT (2005) *Chem Res Toxicol* 18(6):1004

50. Cherny RA, Atwood CS, Xilinas ME, Gray DN, Jones WD, Mclean CA, Barnham KJ, Volitakis I, Fraser FW, Kim Y-S, Huang X, Goldstein LE, Moir RD, Lim JT, Beyreuther K, Zheng H, Tanzi RE, Masters CL, Bush AI (2001) *Neuron* 30(3):665
51. Hider Robert C, Ma Y, Molina-Holgado F, Gaeta A, Roy S (2008) *Biochem Soc Trans* 36(6):1304
52. Kaur D, Yantiri F, Rajagopalan S, Kumar J, Mo JQ, Boonplueang R, Viswanath V, Jacobs R, Yang L, Beal MF, Dimonte D, Volitaskis I, Ellerby L, Cherny RA, Bush AI, Andersen JK (2003) *Neuron* 37(6):899
53. Rodríguez LR, Lapeña T, Calap-Quintana P, Moltó MD, Gonzalez-Cabo P, Navarro Langa JA (2020) *Antioxidants* 9(8):664
54. Bendheim PE, Poeggeler B, Neria E, Ziv V, Pappolla MA, Chain DG (2002) *J Mol Neurosci* 19(1):213
55. Kumar R, Zakharov MN, Khan SH, Miki R, Jang H, Toraldo G, Singh R, Bhasin S, Jasuja R (2011) *J Amino Acids* 2011:812540
56. Işık M, Demir Y, Kırıcı M, Demir R, Şimşek F, Beydemir Ş (2015) *Arch Physiol Biochem* 121(3):97
57. Costantini A, Laureti T, Pala MI, Colangeli M, Cavalieri S, Pozzi E, Brusco A, Salvarani S, Serrati C, Fancellu R (2016) *J Neurol* 263(11):2170.

Tables

Table 1 List of 12 sets of fingerprint descriptors calculated from the PaDEL-Descriptor software.

Fingerprint	Number	Description
2D Atom Pairs	780	Presence/absence of atom pairs for various topological distances
2D Atom Pairs Count	780	Frequency count of atom pairs for various topological distances
E-state	79	Electrotopological state atom types
CDK	1,024	Fingerprint of length 1,024 and search depth of 8
CDK Extended	1,024	Extends the fingerprint with additional bits describing ring features
CDK Graph Only	1,024	A special version considering only the connectivity and not the bond order
Klekota-Roth	4,860	Presence/absence of SMARTS patterns for functional groups
Klekota-Roth Count	4,860	Frequency count of SMARTS patterns for functional groups
MACCS	166	Binary representation of chemical features defined by MACCS keys
PubChem	881	Binary representation of substructures as defined by PubChem
Substructure	307	Presence/absence of chemical substructures
Substructure Count	307	Frequency count of chemical substructures

Table 2 Statistical values of 8 models used according to training and test data for Fe chelation.

Model Name	Train R2 Value	Test R2 Value	Train MAE Value	Test MAE Value	Train MSE Value	Test MSE Value
RandomForestRegressor	0.701	0.572	0.375	0.511	0.254	0.534
XGBRegressor	0.746	0.428	0.299	0.550	0.216	0.697
DecisionTreeRegressor	0.749	0.392	0.280	0.542	0.214	0.740
MLPRegressor	0.434	0.549	0.571	0.563	0.482	0.549
BaggingRegressor	0.677	0.534	0.376	0.522	0.274	0.568
LinearRegression	0.422	-3.195	0.576	297	0.492	3.900
Support Vector Machine	0.514	0.523	0.494	0.572	0.414	0.581
Ridge Regression	0.421	0.576	0.578	0.539	0.494	0.517

Table 3 Statistical values of 8 models used according to training and test data for HDAC.

Model Name	Train R2 Value	Test R2 Value	Train MAE Value	Test MAE Value	Train MSE Value	Test MSE Value
RandomForestRegressor	0.892	0.460	0.274	0.575	0.163	0.587
XGBRegressor	0.912	0.386	0.219	0.132	0.588	0.668
DecisionTreeRegressor	0.915	-0.054	0.197	0.854	0.128	1.149
MLPRegressor	0.758	-0.06	0.442	0.906	0.367	1.158
BaggingRegressor	0.883	0.429	0.279	0.598	0.176	0.622
LinearRegression	0.523	0.241	0.638	0.670	0.725	0.826
Support Vector Machine	0.741	0.252	0.398	0.716	0.392	0.814
Ridge Regression	0.577	0.207	0.608	0.706	0.642	0.863

Table 4 The top features were obtained from the SelectFromModel of Random Forest Regressor model for Fe chelation.

Features	Bit Substructure	Section description
PubChemFP152	<chem>>= 2 saturated or aromatic nitrogen-containing ring size 5</chem>	<p>Rings in a canonic Extended Smallest Set of Smallest</p> <p>Rings (ESSSR) ring set - These bits test for the presence</p> <p>or count of the described chemical ring system. An ESSSR</p> <p>ring is any ring which does not share three consecutive</p> <p>atoms with any other ring in the chemical structure. For</p> <p>example, naphthalene has three ESSSR rings (two phenyl</p> <p>fragments and the 10-membered envelope), while biphenyl</p> <p>will yield a count of only two ESSSR rings.</p>
PubChemFP485	<chem>N:C:C:N</chem>	<p>These bits test for the</p> <p>presence of simple SMARTS patterns, regardless of count,</p> <p>but where bond orders are specific and bond aromaticity</p> <p>matches both single and double bonds.</p>
PubChemFP438	<chem>C(-C)(-N)(=N)</chem>	<p>Detailed atom neighborhoods - These bits test for the</p> <p>presence of detailed atom neighborhood patterns,</p> <p>regardless of count, but where bond orders are specific,</p> <p>bond aromaticity matches both single and double bonds,</p> <p>and where "-", "=", and "#" matches a single bond, double</p> <p>bond, and triple bond order, respectively.</p>

Features	Bit Substructure	Section description
PubChemFP374	<chem>C(~H)(~H)(~H)</chem>	These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant.
PubChemFP821	<chem>CC1C(N)CCCC1</chem>	These bits test for the presence of complex SMARTS patterns, regardless of count, but where bond orders and bond aromaticity are specific.
PubChemFP359	≥ 2 saturated or aromatic heteroatom-containing ring size 6	<p>Rings in a canonic Extended Smallest Set of Smallest</p> <p>Rings (ESSSR) ring set - These bits test for the presence or count of the described chemical ring system. An ESSSR ring is any ring which does not share three consecutive atoms with any other ring in the chemical structure. For example, naphthalene has three ESSSR rings (two phenyl fragments and the 10-membered envelope), while biphenyl will yield a count of only two ESSSR rings.</p>
PubChemFP672	<chem>O=C-C=C-[#1]</chem>	<p>Simple SMARTS patterns - These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.</p>
PubChemFP19	≥ 2 O	Greater than or equal to two oxygen atoms

Features	Bit Substructure	Section description
PubChemFP187	≥ 2 saturated or aromatic nitrogen-containing ring size 6	<p>Rings in a canonic Extended Smallest Set of Smallest</p> <p>Rings (ESSSR) ring set - These bits test for the presence</p> <p>or count of the described chemical ring system. An ESSSR</p> <p>ring is any ring which does not share three consecutive</p> <p>atoms with any other ring in the chemical structure. For</p> <p>example, naphthalene has three ESSSR rings (two phenyl</p> <p>fragments and the 10-membered envelope), while biphenyl</p> <p>will yield a count of only two ESSSR rings.</p>
PubChemFP643	[#1]-C-C-N-[#1]	<p>Simple SMARTS patterns - These bits test for the</p> <p>presence of simple SMARTS patterns, regardless of count,</p> <p>but where bond orders are specific and bond aromaticity</p> <p>matches both single and double bonds.</p>

Summary of top 10 features from the HDAC model along with their corresponding SMARTS patterns and description.

Table 5 The top features were obtained from SelectFromModel of RandomForestRegressor model for HDAC.

Features	Bit Substructure	Section description
PubChemFP338	<chem>C(~C)(~C)(~H)(~N)</chem>	Simple atom nearest neighbors - These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant.
PubChemFP656	<chem>C-C-N-C-C</chem>	Simple SMARTS patterns - These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.
PubChemFP397	<chem>N(~C)(:C)(:C)</chem>	Simple atom nearest neighbors - These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant.
PubChemFP489	<chem>N-C-S-C</chem>	Simple SMARTS patterns - These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.
PubChemFP349	<chem>C(~C)(~H)(~S)</chem>	Simple atom nearest neighbors - These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant.
PubChemFP116	>= 1 saturated or aromatic carbon-only ring size 3	Rings in a canonic Extended Smallest Set of Smallest Rings (ESSSR) ring set - These bits test for the presence or count of the described chemical ring system. An ESSSR ring is any ring which does not share three consecutive atoms with any other ring in the chemical structure. For example, naphthalene has three ESSSR rings (two phenyl fragments and the 10-membered envelope), while biphenyl will yield a count of only two ESSSR rings.

Features	Bit Substructure	Section description
PubChemFP335	<chem>C(~C)(~C)(~C)(~H)</chem>	Simple atom nearest neighbors - These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant.
PubChemFP517	<chem>N-N-C-N</chem>	Simple SMARTS patterns - These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.
PubChemFP791	<chem>NC1CCC(N)CC1</chem>	Complex SMARTS patterns - These bits test for the presence of complex SMARTS patterns, regardless of count, but where bond orders and bond aromaticity are specific.
PubChemFP645	<chem>O=C-N-C-C</chem>	Simple SMARTS patterns - These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.

Figures

Figure 1

This study's methodological path is depicted as a diagram.

Figure 2

For compounds in the iron chelation and HDACi datasets, plot MW vs ALogP. The graph depicts the chemical space of inhibitors against iron (left) and HDAC (right) (right). Orange and blue represent active and inactive molecules, respectively. However, ALogP did not differ significantly in Fe chelation and HDACi molecules ($p=0.135378$, $p=0.119859$, respectively). While nHBDon and nHBAcc are higher in active groups, they are lower in inactive groups in Fe chelation. For HDACi, nHBDon was equal in both active and inactive groups, while nHBAcc was slightly higher in inactive groups. In Fig. 3, according to the Ro5 descriptors, illustrates the distribution of active and inactive substances.

Figure 3

Lipinski's rule-of-five descriptions are plotted in a box. For the iron chelation (A) and HDACi (B) datasets, the four rule-of-five descriptors are shown. Blue and orange represent active and inactive chemicals, respectively. Compounds with a molecular weight not more than 500 Da, a logP of less than 5, and nHBDon and nHBAcc of less than 10 are the most common. Furthermore, employing the Mann-Whitney U test, statistical analysis revealed no significant difference between active and inactive substances. Inactive compounds' ALogP and MW values were not significantly different from active ones.

Figure 4

Feature importance plot from Fe chelation(left) and HDAC (right) models.

Figure 5

Bar graphs created according to SHAP Values for Fe chelation and HDAC inhibition, respectively (from left to right)

Figure 6

Force graph created according to SHAP Values for Fe chelation and HDAC inhibition, respectively (from top to bottom).

Figure 7

Beeswarm graphs created according to SHAP Values for Fe chelation and HDAC inhibition, respectively (from left to right).

Figure 8

Heatmap created according to SHAP Values for Fe chelation and HDAC inhibition, respectively (from left to right)

Figure 9

Decision plot created according to SHAP Values for Fe chelation and HDAC inhibition, respectively (from left to right).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.pdf](#)