**BRAIN. Broad Research in Artificial Intelligence and Neuroscience**
ISSN: 2068-0473 | e-ISSN: 2067-3957
Covered in: Web of Science (WOS); PubMed.gov; IndexCopernicus; The Linguist List; Google Academic; Ulrichs; getCITED; Genamics JournalSeek; J-Gate; SHERPA/RoMEO; Dayang Journal System; Public Knowledge Project; BIUM; NewJour; ArticleReach Direct; Link+; CSB; CiteSeerX; Socolar; KVK; WorldCat; CrossRef; Ideas RePeC; Econpapers; Socionet.

2023, Volume 14, Issue 3, pages: 287-313 | https://doi.org/10.18662/brain/14.3/475
Submitted: September 3rd, 2023 | *Accepted for publication:* September 6th, 2023

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

**Kevser Kübra KIRBOĞA** [1]
**Ecir Uğur KÜÇÜKSILLE** [2]
**Utku KOSE** [3]

[1] PhD student at the Department of Computational Science and Engineering at Istanbul Technical University, Bilecik Seyh Edebali University, Department of Bioengineering, Bilecik, Turkey, ORCID ID:0000-0002-2917-8860, kubra.kirboga@bilecik.edu.tr
[2] Suleyman Demirel University, Engineering Faculty, Computer Engineering Department, 32260,Isparta, Turkey, ORCID ID:0000-0002-3293-9878, ecirkucuksille@sdu.edu.tr
[3] Suleyman Demirel University, Turkey, University of North Dakota, USA, ORCID ID: https://orcid.org/0000-0002-9652-6415 , utkukose@gmail.com, utkukose@sdu.edu.tr, utku.kose@und.edu

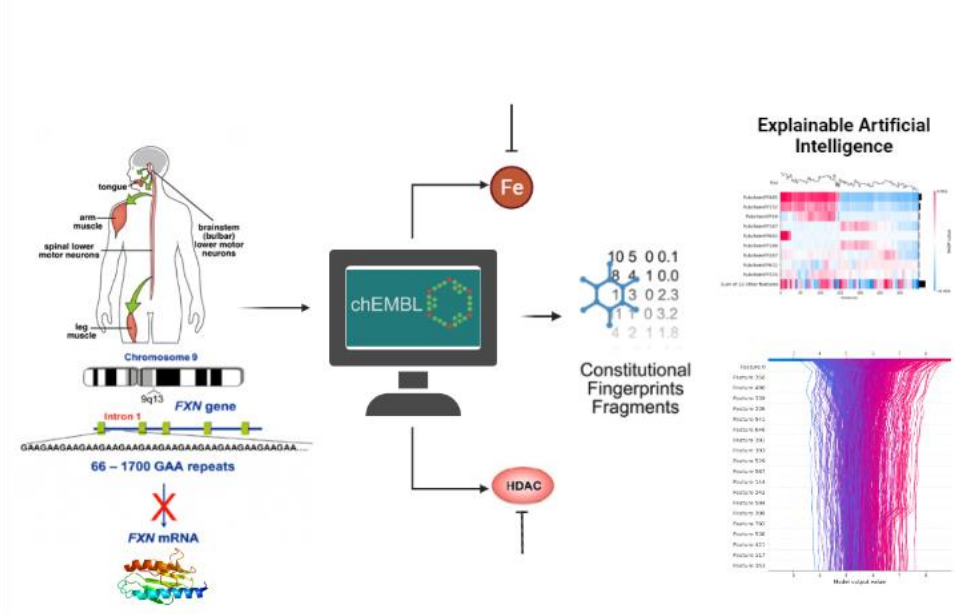
**Abstract:** *Iron (Fe) chelating medicines and Histone deacetylase (HDAC) inhibitors are two therapy options for hereditary Friedreich's Ataxia that have been shown to improve clinical results (FA). Fe chelation molecules can minimize the quantity of stored Fe, and HDAC inhibitors can boost the expression of the Frataxin (FXN) gene in 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 (train $R^2$ score, 0.701 and 0.892; test $R^2$ score 0.572 and 0.460, for Fe and HDAC, respectively). 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 (SHAP value of PubchemFP656 is -0.29) and aromatic rings (SHAP value of PubchemFP12 is -0.16). As a result, we explained the effect of the molecular fingerprints on the models and the impact 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.*

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

**How to cite:** Kırboğa, K. K., Küçüksille, E. U., Köse, U. (2023). Ignition of Small Molecule Inhibitors in Friedreich's Ataxia with Explainable Artificial Intelligence. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 14*(3), 287-313. https://doi.org/10.18662/brain/14.3/475

©2023 Published by LUMEN Publishing. This is an open access article under the CC BY-NC-ND license

## GRAPHICAL ABSTRACT




## *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 defining target chemicals, the need for explicable deep learning models is increasing to increase humans' ability to interpret mathematical models.

## 1. INTRODUCTION

The process that reveals a therapeutically beneficial compound for use in curing and treating diseases is called drug discovery. It takes approximately 12-15 years from initial drug discovery to marketing. Therefore, the identification, synthesis, characterization, optimization, and determination of therapeutic efficacy values of such compounds is a very long and challenging process (Deore et al., 2019). To overcome this complexity, molecular fingerprints represent and compare molecular structures. In this process, the input data encoded with the SMILES (Simplified Molecular Input Line Entry Specification) information of the compounds can be used for training. Therefore, the relationship between

structure and biological, chemical, and physicochemical activities are revealed Staszak et al. (2021).

In recent years, the increasing interest in computer-assisted drug discovery (CADD) has enabled drug discovery studies using artificial intelligence (AI).CADD applications include new compound design, de novo design, structure and ligand-based design, estimation of the physicochemical and pharmacokinetic properties of the drug, and machine learning algorithms for drug repurposing, Lavecchia (2015); Lo et al. (2018); Vamathevan et al. (2019).In addition to deep learning methods for predicting properties based on input data and determining nonlinear input-output relationships, machine learning approaches based on molecular descriptors contribute to the synergy of drug discovery and cheminformatics, Xue & Bajorath (2000). Explainable artificial intelligence (XAI) aims to increase the intelligibility of models derived from input data and transform them into more human-interpretable formats. As AI is increasingly used in drug discovery and related fields, there is a growing demand for using XAI to interpret fundamental models.

Additionally, to mathematical models, XAI can support the design of new drugs, the extraction of pharmacological activities from molecular structures, and the creation of new bioactive compounds with desired properties by providing ways to make the underlying 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. The use of XAI studies in drug discovery and pharmacological studies is thought to increase in the coming years, Jiménez-Luna et al. (2020).Friedreich's ataxia (FRDA/FA) is a progressive and neurodegenerative genetic disorder that usually occurs between 10-15 years old. At the onset of FA, the first symptoms include unbalanced gait, frequent falls, and impaired movement coordination ability, Cook & Giunti (2017); Lew et al. (2020).The most prevalent inherited disease occurs with the expansion of the GAA triplet located in intron 1 of the Frataxin gene on chromosome 9q13,Alper & Narayanan (2003).There is a linear correlation between the expanded repeats size and the phenotypic severity of FA. Frataxin (FXN) protein (218 aa., 18 kDa), the protein of the frataxin gene, is a mitochondrial protein involved in Fe metabolism. FXN deficiency in FA, a nuclear-encoded mitochondrial disease, causes Fe accumulation in the mitochondria, disruption of mitochondrial enzymes, and

sensitivity to oxidative stress. Due to this sensitivity, free radical-mediated cell death occurs (Gottesfeld, 2019; Seznec et al., 2005; Wilson, 2006).

Furthermore, FA patients may have variable clinical diseases such as heart disease, diabetes mellitus, and glucose intolerance. However, the pathogenesis of FA has not yet been resolved. Therefore, there is no current treatment method for FA. Bulteau et al. (2007) found that mitochondrial Fe has a primary key role in developing and advancing the disease in their studies on *S. cerevisiae* yeast.

The clinical signs and symptoms associated with FA result from degenerative changes in the dorsal root ganglia. Nerve fibers in the spinal cord degenerate and cause a lack of signals to the cerebellum, which is responsible for voluntary movement coordination, Pandolfo (1999).FA causes decreased iron-sulfur cluster and heme formation, leading to Fe accumulation in mitochondria. When Boddaert et al. evaluated the reduction of Fe accumulation with appropriate iron chelators, they showed that chelators reduced Fe accumulation ( Boddaert et al., 2007). It is also known that Fe chelator agents such as deferiprone, Idebenone, and Desferoxamine are also effective in moderate Fe accumulation, Goncalves et al. (2008); Pandolfo & Hausmann (2013); Soriano et al. (2013).Another treatment modality based on research and clinical evidence is gene silencing of FXN alleles with extended repeats. This method involves a heterochromatin-mediated mechanism that reverses FXN gene silencing by a histone deacetylase (HDAC) inhibitor. These molecules provide therapeutic benefits by acetylating lysine residues on histones H4 and H3 (Herman et al., 2006; Soragni et al., 2011).However, there is no current treatment method for FA since Fe chelator molecules are only effective in moderate Fe accumulation, and HDAC inhibitors (HDACi's) are either highly toxic or have low specificity. To discover new HDACi and Fe chelator molecules, we used the XAI method in this study to identify the critical features for discovering molecules and to address the main challenge for possible drug discovery for FA. Whether or not the molecules obtained from the databases carry molecular fingerprints from the bioactivity data were investigated. We explained the importance of XAI and molecular fingerprints for these molecules.

The research paper is designed as follows. Section 1 contains the introduction to FRDA and XAI. Section 2 presents an analysis of current methods and materials used for FA. Specifically, it analyzes data, descriptors, and models for FA datasets using XAI and SHAP techniques. Section 3 presents the results and discussion of the research and the statistical validation of the data. Finally, section 4 concludes that XAI can be a

valuable tool for pre-drug discovery in many diseases, including genetic diseases.

### *1.1. 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. However, since the function studies of proteins are a challenging and complex process, there is no defined treatment method yet. Therefore, selecting a faster and less costly computational procedure and using large datasets are considered 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, 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, Aranca et al. (2016). 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. Instead, it can be developed by synergetic research efforts with different scientific backgrounds. Modelling choices and predictions must be examined in highly costly and time-sensitive situations such as drug discovery, deep learning, and machine learning applications. With XAI, it makes sense to develop hybrid approaches that are easier to understand and computationally affordable without forgetting the limitations of drug discovery.

### *1.2. Related Works*

Recent advances in computational techniques and artificial intelligence have contributed significantly to medication discovery for various genetic and metabolic illnesses. One of the most considerable 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 developing 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. In addition, they demonstrated the importance of nitrogen and aromatic bonded compounds (Schaduangrat et al. 2021).

Rodriguez-Perez et al. explored a variant for precisely calculating Shapley values for decision tree methods. Further, they 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, Rodríguez-Pérez & Bajorath (2020).

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 critical descriptors identified, it was revealed that polarity resolution is required for membrane lytic antimicrobial activity (Li & Nantasenamat, 2019).

### *1.3. 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 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 a positive correlation between the SHAP values of each molecular fingerprint practical on the model and the data in the literature affecting the drug discovery and development process?

Therefore, the main task of this research article is to construct an interpretable regression model for iron inhibition and HDA inhibitors. This includes (i) generating a bioactivity dataset from a new kinase family inhibitor ChEMBL for the regression model, (ii) identifying 881 PubChem fingerprints as features for inhibitors, (iii) generating eight regression model families, one for molecules, and (iv) removing preferred portions of inhibitors for each inhibitor family and shared portions of inhibitors for all inhibitors family using SHAP. Finally, our approach can provide an effective strategy for identifying and designing selective inhibitors targeting the iron and HDA families. XAI is more effective at the pre-drug discovery stage than existing drugs.

## 2. Materials and Detailed Methods

### *2.1. Data compilation and curation*

Datasets were collected from the version 25 ChEMBL database for Fe chelation and HDAC, Gaulton et al. (2017).The IC50 selection for the bioactivity unit was carried out through a data improvement process. 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 μM 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.

### *2.2. Descriptors of Molecular*

Molecular fingerprint identifiers for the compounds in the obtained datasets were calculated using the PaDEL-Descriptor software, Yap (2011). 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 crucial 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 (Malik et al., 2020; Yap, 2011).

### *2.3. 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, which has a higher data percentage. The second group is the smaller dataset dedicated to testing the model. Typically, there are many variables in the creation of mathematical descriptors. On the other hand, the training set 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 (Saeys et al., 2007).The model is trained after determining the best subset of variables. Overtraining should be avoided to keep the model's validity when dealing with unknown data. Cross-validation (CV) techniques are commonly used in these situations. CV provides for measuring 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. Finally, 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, (Carracedo et al., 2021; Gramatica & Sangion, 2016).

### *2.4. 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 based on descriptors of compounds. The results were visualized with the Seaborn v0.8.1 Python package, Waskom et al.(2017).

### *2.5. Explainability with Shapley additive explanations (SHAP)*

In this work, we were finally interested in understanding the effect of molecular fingerprints on models. Several features are advantageous in our SHAP selection among many explainability techniques Guidotti et al. (2019);

Lundberg & Lee (2017). The most important thing is that these values are independent of the model. Like RandomForestRegressor (Hu et al., 2020). They are not tied to a particular model type. It will also suit the RandomForestRegressor model we chose for HDAC inhibitors. It also offers SHAP values, accuracy, consistency, and incompleteness (Hu et al., 2020; Lundberg & Lee, 2017; Lundberg et al., 2018). Finally, SHAP applications are straightforward and can be conveniently documented with images. Shapley's values were initially proposed as a game theory used to pay players somewhat based on their contribution to their total earnings, Shapley (1953). Estimating a model means assigning its 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. A SHAP value for a given molecular fingerprint value can be explained as the difference between the actual and average predictions 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, Molnar (2020).

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 (1).

$$\phi\,(\hat{f}\,(\,Xi\,)) \equiv \phi 0 + \sum K\ k=1\ \phi k\,(\,Xi\,)\,,\ \forall i = 1, \ldots, n \qquad (1)$$

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. $\phi \in RK$; $\phi k \in R$. $\phi k$ (Xi) is the Shapley values of local functions, Hu et al. (2020); Lundberg & Lee (2017); Lundberg et al. (2019); Lundberg et al. (2018); Shapley (1953); Shapley (2016).

Also, SHAP values use a specific index on individual features and among all binary features. In this way, SHAP values can easily explain the modelling of interactions that would go unnoticed. This feature is essential because it provides a clearer understanding of the variables in the model and the relationship between them, Li et al. (2020).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. Finally, we analyzed the critical interactions between these molecular fingerprints and their targets, Moncada-Torres et al. (2021).

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 decided on the SHAP findings and the effect of Fe chelation molecules and HDAC inhibitors on drug discovery.

### *2.6. 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 (R2), 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

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

$n$= number of data points
$Y_i$= observed values
$\hat{Y}_i$= predicted values

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

**MAE**= mean absolute error
$y_i$= prediction
$x_i$= true value
$n$= total number of data points

### *2.7. 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. This way, the estimator is adapted to the training data, and each feature's importance is calculated on the model. It ranks the meta-transformer properties according to their significance and selects the most suitable ones depending on the threshold value. In our study, the threshold values were 0.005 and 0.003 Fe chelation and HDACi, respectively, Stefanidou-Voziki et al. (2021).

A property significance analysis was performed on selected informative descriptors to understand better the mechanistic details ruling Fe chelation and HDACi compounds. 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 **Figure4**, 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). In addition, the SelectFromModel feature selection technique was included in the methodology to increase the prediction accuracy, and essential features were tested with machine learning models again (Shobana & Priya, 2021).

## 3. RESULTS and DISCUSSION

This study evaluates the activity and bioactivity of Fe chelator molecules and HDA inhibitors for FA genetic disease. This study also determines which molecular fingerprints dominate the model with XAI in developing molecules. As shown in **Figure1**, 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 contained molecular fingerprints. Finally, molecular fingerprints that contributed the most to the model were determined by visualizing with XAI.

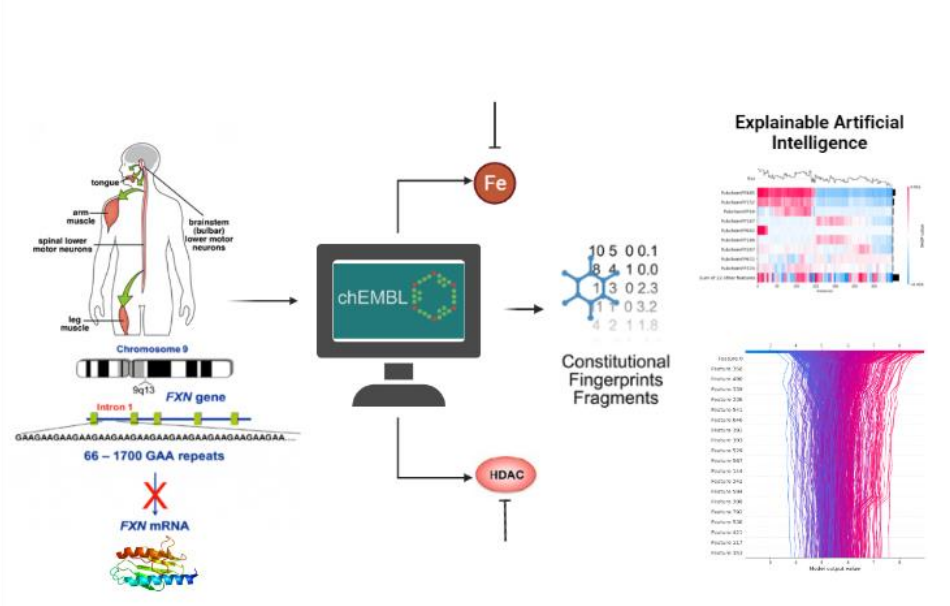
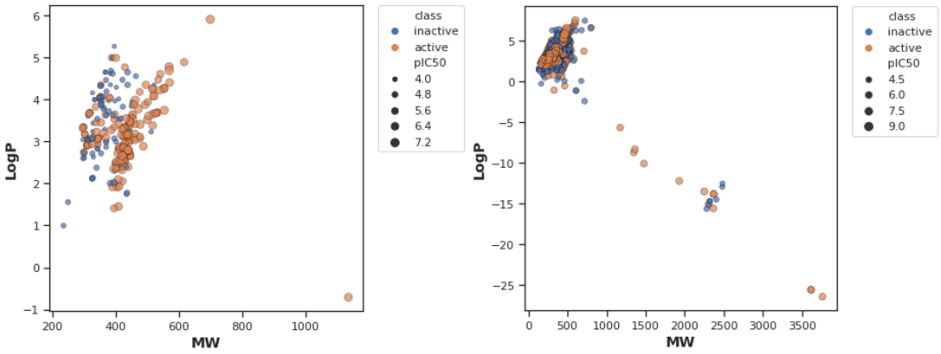



Figure 1. This study's methodological path is depicted as a diagram.

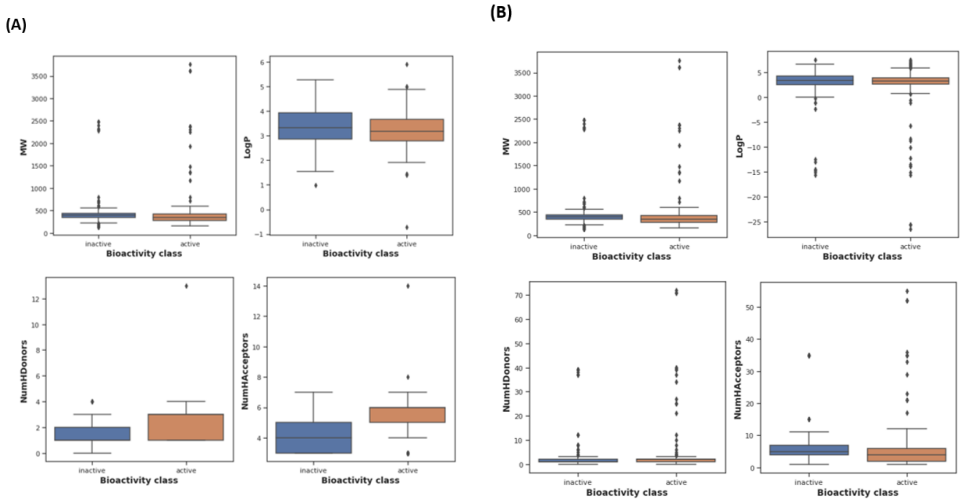
### *3.1. Chemical Space Analysis*

Chemical space analysis is used to find typical characteristics between active and inactive chemicals in the classification of substances. The 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, Lipinski et al. (2001).The chemical space analysis of MW-LogP is visualized in **Figure 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.




Author's own conception

**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). 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. **Figure 3**, according to the Ro5 descriptors, illustrates the distribution of active and inactive substances.



Author's own conception

**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 of 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.

Besides, it was observed that Fe chelation compounds have higher nHBDon and nHBAcc values in active compounds. In contrast, nHBDon and nHBAcc values do not present a significant difference for active and inactive compounds in HDACi.

### *3.2. QSAR modelling*

We followed the Organization for Economic Co-operation and Development (OECD) guidelines to develop an interpretable QSAR model, OECD(2007).These guidelines consist of the following main points: (i) datasets have a defined endpoint, (ii) the algorithm for learning is straightforward, (iii) the area to which the QSAR model will be applied is well defined, and (iv) measures of predictability, mechanistic interpretations, and robustness are available, Malik et al. (2020).For a robust QSAR model, molecular fingerprint descriptors were applied in this study using PaDEL-Descriptor software. Yap (2011).The interpretable properties from the 12 fingerprints (i.e., PubChem, Substructure, and Klekota–Roth) are listed on our GitHub page(https://github.com/tissueandcells/XAI).

### *3.3. 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 calculated. 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 the 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 |

The table was developed by the author

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 |

The table was developed by the author

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 **Figure4**.

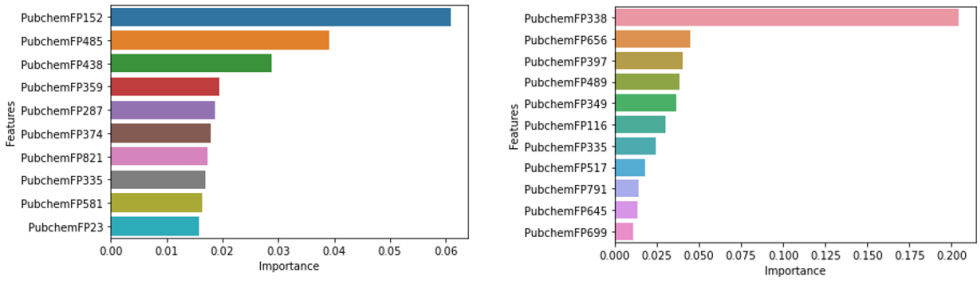



Figure 4 Feature importance plot from Fe chelation(left) and HDAC (right) models.
Author's own conception

### *3.4. Explainability with Shapley additive explanations (SHAP)*

The effect of the selected features on the model's functioning was determined with the SHAP method. The contribution of each attribute to the model is visualized with SHAP values. As shown in **Figure 5** and **Figure 6**, bar and force graphs were created locally according to each sample's degree of importance. Beeswarm, heatmap, and decision graphics are general graphics. Molecular fingerprints on all compounds are ranked according to their importance (**Figure7, 8, 9**).

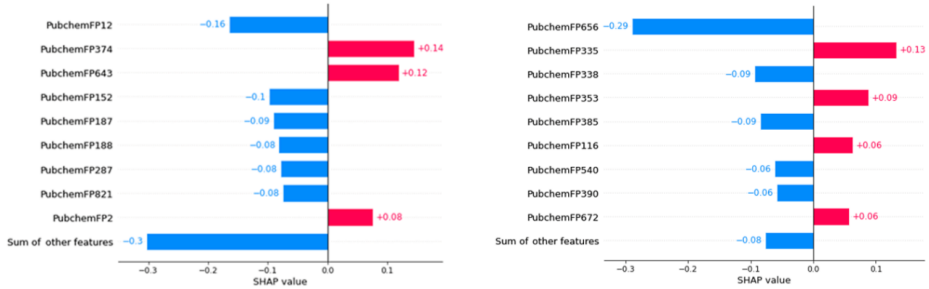



Figure 5 Bar graphs created according to SHAP Values for Fe chelation and HDAC inhibition, respectively (from left to right)
Author's own conception

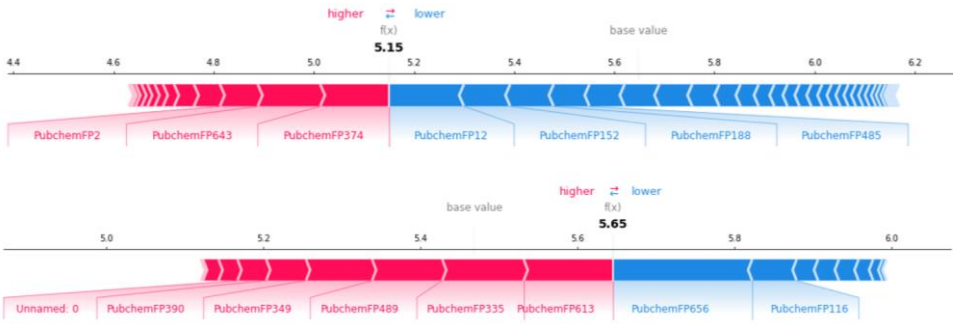



Figure 6 Force graph created according to SHAP Values for Fe chelation and HDAC inhibition, respectively (from top to bottom).
Author's own conception

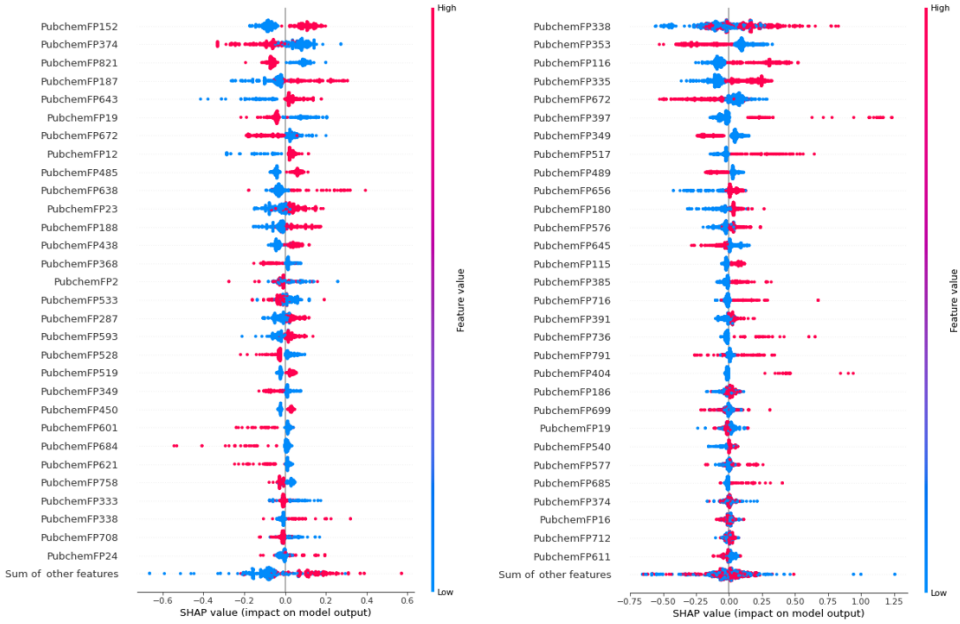



Figure 7 Beeswarm graphs created according to SHAP Values for Fe chelation and HDAC inhibition, respectively (from left to right).
Author's own conception

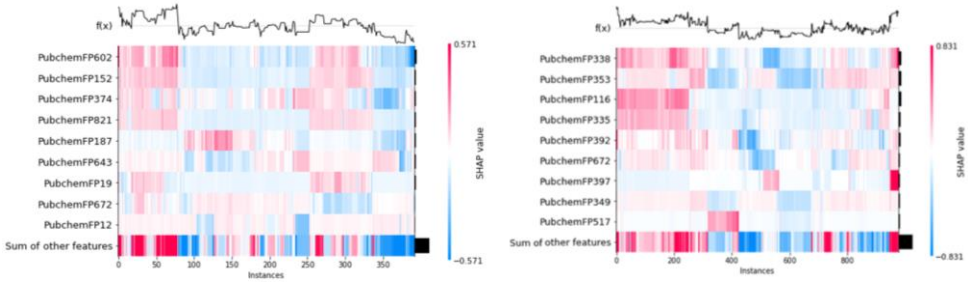



Figure 8 Heatmap created according to SHAP Values for Fe chelation and HDAC inhibition, respectively (from left to right)
Author's own conception

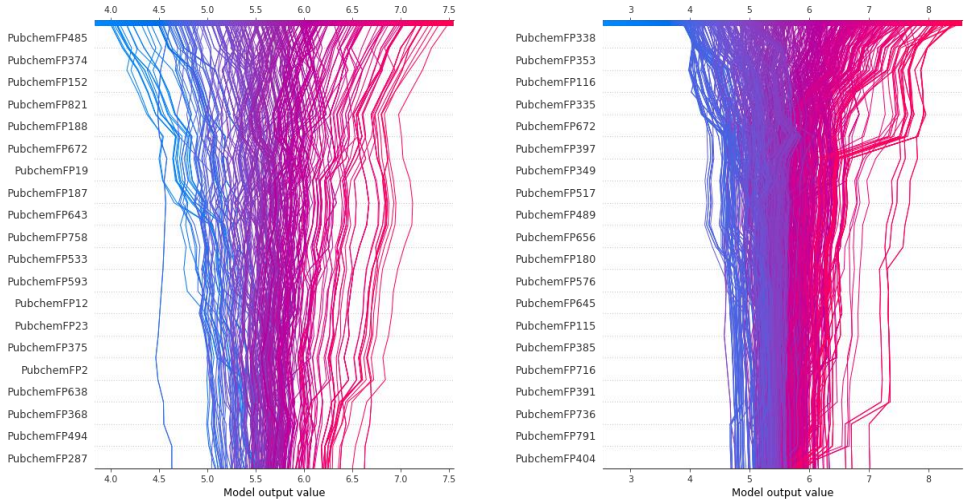



Figure 9 Decision plot created according to SHAP Values for Fe chelation and HDAC inhibition, respectively (from left to right).
Author's own conception

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).

### *3.5. 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, and PubChemFP187 fingerprints were associated with the aromaticity and ring structures as seen in the explanations. Deferiprone (Ferriprox), Beutler (2007); Molina-Holgado et al. (2008); Reeder & Wilson (2005) and Clioquinol, Cherny et al. (2001); Hider et al. (2008); Kaur et al. (2003) 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 explain the importance of single and double bonds made by cyclic structures and also reflect the content of many existing drugs. For example, alpha-tocopherol quinone, a synthetic molecule comparable to Coenzyme Q 10 but with 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, Rodríguez et al. (2020). 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 indirectly targeting oxidative stress.

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, Bendheim et al. (2002). However, studies were conducted by including 46 FA participants. While tolerance was observed at all doses, great benefits were not observed.

### *3.6. 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. Furthermore, it is known that the 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. Kumar et al. (2011).

The fact that 10 10 feature selections are nitrogen-themed features also highlighted this importance. 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 nitrogen and a sulfur atom in their ring structure. It has been supported by various studies that thiazoles can be used in antiepileptic drugs, Işık et al. (2015). Since the muscle symptoms in the mechanism of epilepsy are similar to FA, azole groups may be promising targets for drug discovery. The sulfur content of PubChemFP489 and 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 of molecular fingerprints where nitrogen-based fingerprints are intense. N atoms and cyclic structures in thiamine are dominant, and their deficiency causes severe neurological changes in the central and peripheral nervous systems. Thiamine deficiency's molecular and clinical changes, such as impaired oxidative stress metabolism, increased oxidative stress, and selective neuronal loss, islike 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 the ventricular septum. These results also showed it could be therapeutic, Costantini et al. (2016).

## 4. LIMITATIONS

This study, although a guiding study for discovering new drug candidates for FA, which is an orphan genetic disease, has some limitations:

The study used only two data sets (ChEMBL and PubChem) to predict the bioactivity values of molecules related to FA. Therefore, it is not clear whether the results are valid for molecules obtained from different data sources or different diseases.

The study used only one of the XAI methods, namely the SHAP (Shapley Additive exPlanations) technique, to explain the bioactivity values of the molecules. Therefore, it did not provide information on the performance and comparison of other XAI methods.

The study, being a theoretical study, requires the validation of the predicted bioactivity values and explanations experimentally. Therefore, the

practical applications and contributions of the study have not been proven yet.

## 5. CONCLUSIONS

FA is the most wide spread in herited ataxia, accounting for around half of all ataxia cases and seventy percent of individuals under 25, Aranca et al. (2016).However, current inhibitors and gene therapy, on the other hand, are insufficient since they cannot 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 and inactive compounds. The activity prediction of these molecules was appraised through machine learning algorism and various fingerprint identifiers classes. The results showed that combining the RF technique with PubChem fingerprints produced the most interpretable identifiers and the highest-performing model. Aromaticity and amine groups are significant for active compounds, according to the property analysis of the vital infrastructure contributions from the SHAP values. However, 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 disorders. 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 valuable tool for predicting pharmaceutical manufacturing and development, but its broad applicability must be verified in various inherited disorders. Explainable algorithms like SHAP are critical for increasing trust and transparency in machine learning models, so 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 using the SHAP annotator.

- Future studies should focus on analyzing SHAP plots for long-term prediction and 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, has negative aspects such as low model accuracy, conflicting molecular fingerprints, or choosing the suitable threshold. Nevertheless, expanding computational studies, increasing their applicability, in vitro and in vivo experiments, and XAI and drug discovery studies will support each other and are considered promising for future drug studies.

## Acknowledgement

The preprint version of our article (https://www.researchgate.net/publication/359090417_Ignition_of_Small_Molecule_Inhibitors_in_Friedreich's_Ataxia_with_Explainable_Artificial_Intelligence) is the only submitted version of our article and no other preprint versions are under evaluation in other journals.

## References

Alper, G. & Narayanan, V. (2003). Friedreich's ataxia. *Pediatr Neurol*, *28*(5), 335-341. https://doi.org/10.1016/s0887-8994(03)00004-3

Aranca, T. V., Jones, T. M., Shaw, J. D., Staffetti, J. S., Ashizawa, T., Kuo, S.-H., Fogel, B. L., Wilmot, G. R., Perlman, S. L., Onyike, C. U., Ying, S. H. & Zesiewicz, T. A. (2016). Emerging therapies in Friedreich's ataxia. *Neurodegenerative Disease Management*, *6*(1), 49-65. https://doi.org/10.2217/nmt.15.73

Bendheim, P. E., Poeggeler, B., Neria, E., Ziv, V., Pappolla, M. A. & Chain, D. G. (2002). Development of indole-3-propionic acid (OXIGON™) for alzheimer's disease. *Journal of molecular neuroscience*, *19*(1), 213-217.

Beutler, E. (2007). Iron storage disease: Facts, fiction and progress. *Blood Cells, Molecules, and Diseases*, *39*(2), 140-147. https://doi.org/10.1016/j.bcmd.2007.03.009

Boddaert, N., Le Quan Sang, K. H., Rötig, A., Leroy-Willig, A., Gallet, S., Brunelle, F., Sidi, D., Thalabard, J.-C., Munnich, A. & Cabantchik, Z. I. (2007). Selective iron chelation in Friedreich ataxia: biologic and clinical implications. *Blood*, *110*(1), 401-408. https://doi.org/10.1182/blood-2006-12-065433

Bulteau, A. L., Dancis, A., Gareil, M., Montagne, J. J., Camadro, J. M. & Lesuisse, E. (2007). Oxidative stress and protease dysfunction in the yeast model of Friedreich ataxia. *Free Radic Biol Med*, *42*(10), 1561-1570. https://doi.org/10.1016/j.freeradbiomed.2007.02.014

Carracedo, P., Blanco, J., Rodriguez-Fernandez, N., Cedrón-Santaeufemia, F., Nóvoa, F., Carballal, A., Maojo, V., Pazos, A. & Fernandez-Lozano, C. (2021). A review on machine learning approaches and trends in drug discovery. *Computational and Structural Biotechnology Journal*, *19*, 4538-4558. https://doi.org/10.1016/j.csbj.2021.08.011

Cherny, R. A., Atwood, C. S., Xilinas, M. E., Gray, D. N., Jones, W. D., Mclean, C. A., Barnham, K. J., Volitakis, I., Fraser, F. W., Kim, Y.-S., Huang, X., Goldstein, L. E., Moir, R. D., Lim, J. T., Beyreuther, K., Zheng, H., Tanzi, R. E., Masters, C. L. & Bush, A. I. (2001). Treatment with a Copper-Zinc chelator markedly and rapidly inhibits β-Amyloid accumulation in Alzheimer's disease transgenic mice. *Neuron*, *30*(3), 665-676. https://doi.org/10.1016/s0896-6273(01)00317-8

Cook, A., & Giunti, P. (2017). Friedreich's ataxia: clinical features, pathogenesis and management. *British Medical Bulletin*, *124*(1), 19-30. https://doi.org/10.1093/bmb/ldx034

Costantini, A., Laureti, T., Pala, M. I., Colangeli, M., Cavalieri, S., Pozzi, E., Brusco, A., Salvarani, S., Serrati, C. & Fancellu, R. (2016). Long-term treatment with thiamine as possible medical therapy for Friedreich ataxia. *Journal of neurology*, *263*(11), 2170-2178.

Deore, A., Dhumane, J., Wagh, R. & Sonawane, R. (2019). The stages of drug discovery and development process. *Asian Journal of Pharmaceutical Research and Development*, *7*, 62-67. https://doi.org/10.22270/ajprd.v7i6.616

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I. & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, *45*(D1), D945-D954. https://doi.org/10.1093/nar/gkw1074

Goncalves, S., Paupe, V., Dassa, E. P. & Rustin, P. (2008). Deferiprone targets aconitase: Implication for Friedreich's ataxia treatment. *BMC Neurology*, *8*(1), 20. https://doi.org/10.1186/1471-2377-8-20

Gottesfeld, J. M. (2019). Molecular mechanisms and therapeutics for the GAA·TTC expansion disease Friedreich Ataxia. *Neurotherapeutics*, *16*(4), 1032-1049. https://doi.org/10.1007/s13311-019-00764-x

Gramatica, P. & Sangion, A. (2016). A Historical excursus on the statistical validation parameters for QSAR models: A clarification concerning metrics

and terminology. *Journal of Chemical Information and Modeling*, *56*(6), 1127-1131. https://doi.org/10.1021/acs.jcim.6b00088

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, *34*(6), 14-23.

Herman, D., Jenssen, K., Burnett, R., Soragni, E., Perlman, S. L., & Gottesfeld, J. M. (2006). Histone deacetylase inhibitors reverse gene silencing in Friedreich's ataxia. *Nat Chem Biol*, *2*(10), 551-558. https://doi.org/10.1038/nchembio815

Hider, Robert C., Ma, Y., Molina-Holgado, F., Gaeta, A. & Roy, S. (2008). Iron chelation as a potential therapy for neurodegenerative disease. *Biochemical Society Transactions*, *36*(6), 1304-1308. https://doi.org/10.1042/bst0361304

Hu, L., Liu, B., Ji, J. & Li, Y. (2020). Tree-based machine learning to identify and understand major determinants for stroke at the neighborhood level. *Journal of the American Heart Association*, *9*(22). https://doi.org/10.1161/jaha.120.016745

Işık, M., Demir, Y., Kırıcı, M., Demir, R., Şimşek, F. & Beydemir, Ş. (2015). Changes in the antioxidant system in adult epilepsy patients receiving antiepileptic drugs. *Archives of physiology and biochemistry*, *121*(3), 97-102. https://doi.org/10.3109/13813455.2015.1026912

Jiménez-Luna, J., Grisoni, F. & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, *2*(10), 573-584. https://doi.org/10.1038/s42256-020-00236-4

Kaur, D., Yantiri, F., Rajagopalan, S., Kumar, J., Mo, J. Q., Boonplueang, R., Viswanath, V., Jacobs, R., Yang, L., Beal, M. F., Dimonte, D., Volitaskis, I., Ellerby, L., Cherny, R. A., Bush, A. I. & Andersen, J. K. (2003). Genetic or pharmacological Iron chelation prevents MPTP-induced neurotoxicity in Vivo. *Neuron*, *37*(6), 899-909. https://doi.org/10.1016/s0896-6273(03)00126-0

Kumar, R., Zakharov, M. N., Khan, S. H., Miki, R., Jang, H., Toraldo, G., Singh, R., Bhasin, S. & Jasuja, R. (2011). The dynamic structure of the estrogen receptor. *Journal of Amino Acids*, *2011*, 812540. https://doi.org/10.4061/2011/812540

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today*, *20*(3), 318-331. https://doi.org/10.1016/j.drudis.2014.10.012

Lew, S.-Y., Yow, Y.-Y., Lim, L.-W. & Wong, K.-H. (2020). Antioxidant-mediated protective role of Hericium erinaceus (Bull.: Fr.) Pers. against oxidative damage in fibroblasts from Friedreich's ataxia patient. *Food Science and Technology*, *40*(suppl 1), 264-272. https://doi.org/10.1590/fst.09919

Li, H. & Nantasenamat, C. (2019). Toward insights on determining factors for high activity in antimicrobial peptides via machine learning. *PeerJ*, *7*, https://doi.org/10.7717/peerj.8265

Li, R., Shinde, A., Liu, A., Glaser, S., Lyou, Y., Yuh, B., Wong, J. & Amini, A. (2020). Machine learning–based interpretation and visualization of nonlinear interactions in prostate cancer survival. *JCO Clinical Cancer Informatics*(4), 637-646. https://doi.org/10.1200/cci.20.00002

Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. *Advanced Drug Delivery Reviews*, *46*(1), 3-26. https://doi.org/https://doi.org/10.1016/S0169-409X(00)00129-0

Lo, Y. C., Rensi, S. E., Torng, W. & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discov Today*, *23*(8), 1538-1546. https://doi.org/10.1016/j.drudis.2018.05.010

Lundberg, S. & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*.

Lundberg, S. M., Erion, G. G. & Lee, S.-I. (2019). Consistent individualized feature attribution for tree ensembles.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J. & Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, *2*(10), 749-760. https://doi.org/10.1038/s41551-018-0304-0

Malik, A. A., Phanus-Umporn, C., Schaduangrat, N., Shoombuatong, W., Isarankura-Na-Ayudhya, C. & Nantasenamat, C. (2020). HCVpred: A web server for predicting the bioactivity of hepatitis C virus NS5B inhibitors. *Journal of Computational Chemistry*, *41*(20), 1820-1834. https://doi.org/10.1002/jcc.26223

Molina-Holgado, F., Gaeta, A., Francis, P. T., Williams, R. J. & Hider, R. C. (2008). Neuroprotective actions of deferiprone in cultured cortical neurones and SHSY-5Y cells. *Journal of Neurochemistry*, *105*(6), 2466-2476. https://doi.org/10.1111/j.1471-4159.2008.05332.x

Molnar, C. (2020). Interpretable machine learning. *Self published.* https://christophm.github.io/interpretable-ml-book/

Moncada-Torres, A., Van Maaren, M. C., Hendriks, M. P., Siesling, S. & Geleijnse, G. (2021). Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-86327-7

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. In: OECD, Paris, France.

Pandolfo, M. (1999). Friedreich's ataxia: clinical aspects and pathogenesis. *Semin Neurol*, *19*(3), 311-321. https://doi.org/10.1055/s-2008-1040847

Pandolfo, M. & Hausmann, L. (2013). Deferiprone for the treatment of Friedreich's ataxia. *Journal of Neurochemistry*, *126*, 142-146. https://doi.org/10.1111/jnc.12300

Reeder, B. J. & Wilson, M. T. (2005). Desferrioxamine inhibits production of cytotoxic heme to protein cross-linked myoglobin: A mechanism to protect against oxidative stress without iron chelation. *Chemical Research in Toxicology*, *18*(6), 1004-1011. https://doi.org/10.1021/tx049660y

Rodríguez-Pérez, R. & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, *34*(10), 1013-1026. https://doi.org/10.1007/s10822-020-00314-0

Rodríguez, L. R., Lapeña, T., Calap-Quintana, P., Moltó, M. D., Gonzalez-Cabo, P. & Navarro Langa, J. A. (2020). Antioxidant therapies and oxidative stress in Friedreich's Ataxia: The right path or just a diversion? *Antioxidants*, *9*(8), 664. https://doi.org/10.3390/antiox9080664

Saeys, Y., Inza, I. & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517. https://doi.org/10.1093/bioinformatics/btm344

Schaduangrat, N., Malik, A. A. & Nantasenamat, C. (2021). ERpred: a web server for the prediction of subtype-specific estrogen receptor antagonists. *PeerJ*, *9*, https://doi.org/10.7717/peerj.11716

Seznec, H., Simon, D., Bouton, C., Reutenauer, L., Hertzog, A., Golik, P., Procaccio, V., Patel, M., Drapier, J. C., Koenig, M. & Puccio, H. (2005). Friedreich ataxia: the oxidative stress paradox. *Hum Mol Genet*, *14*(4), 463-474. https://doi.org/10.1093/hmg/ddi042

Shapley, L. (1953). A Value for n-Person Games. *Princeton University Press, Princeton*, 307-317. https://doi.org/https://doi.org/10.1515/9781400881970-018

Shapley, L. S. (2016). 17. A value for n-person games. *Contributions to the Theory of Games (AM-28), II,* 307-318. Princeton University Press. https://doi.org/doi:10.1515/9781400881970-018

Shobana, G. & Priya, D. N. (2021). Cancer drug classification using artificial neural network with feature selection. 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV),

Soragni, E., Xu, C., Cooper, A., Plasterer, H. L., Rusche, J. R. & Gottesfeld, J. M. (2011). Evaluation of histone deacetylase inhibitors as therapeutics for neurodegenerative diseases. *Methods in molecular biology*, *793*, 495-508.

Soriano, S., Llorens, J. V., Blanco-Sobero, L., Gutiérrez, L., Calap-Quintana, P., Morales, M. P., Moltó, M. D. & Martínez-Sebastián, M. J. (2013). Deferiprone and idebenone rescue frataxin depletion phenotypes in a Drosophila model of Friedreich's ataxia. *Gene*, *521*(2), 274-281. https://doi.org/10.1016/j.gene.2013.02.049

Staszak, M., Staszak, K., Wieszczycka, K., Bajek, A., Roszkowski, K. & Tylkowski, B. (2021). Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. https://doi.org/10.1002/wcms.1568

Stefanidou-Voziki, P., Cardoner-Valbuena, D., Villafafila-Robles, R. & Dominguez-Garcia, J. L. (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),

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M. & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov*, *18*(6), 463-477. https://doi.org/10.1038/s41573-019-0024-5

Waskom, M., Botvinnik, Olga, Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, Qalieh, Adel. (2017). mwaskom/seaborn: v0.8.1. *Zenodo*. https://doi.org/https://doi.org/10.5281/zenodo.883859

Wilson, R. B. (2006). Iron dysregulation in Friedreich ataxia. *Semin Pediatr Neurol*, *13*(3), 166-175. https://doi.org/10.1016/j.spen.2006.08.005

Xue, L. & Bajorath, J. (2000). Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb Chem High Throughput Screen*, *3*(5), 363-372. https://doi.org/10.2174/1386207003331454

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, *32*(7), 1466-1474. https://doi.org/10.1002/jcc.21707