Journal of Research High School

Volume 2023 (1) 37 - 45
ISSN 2688-3651 [online]

Analysis of Mutation Pathogenicity and the Viability of HSP Therapy for Mutated HEX-A in Tay-Sachs Disease

Aditya Shrinivasan[1*]

[1]Evergreen Valley High School, San Jose, CA, USA

Received June 22, 2022; Revised December 13, 2022; Accepted, December 28, 2022

Abstract

This work explored three different concepts. First, docking studies were performed with various mutant HexA structures and Arimoclomol (a Heat Shock Protein inducer), with statistical analysis to establish any correlation between noted binding affinity and either mutation pathogenicity or mutation type. This was followed by the review of Arimoclomol and Heat Shock Therapy as a potential therapeutic option for Tay Sachs disease, exploring the possibility for a future clinical trial. Finally, a gene mutation pathogenicity prediction model was developed using classification with the available dataset of HEXA gene mutations to experimentally determine the pathogenicity of any HEXA mutation. The statistical analysis found no correlations between either the mutation type or mutation pathogenicity and binding affinity. This leads to the conclusion that amino acid alterations don't play a role in causing pathogenicity and benignity in a mutation and that the mutation type doesn't affect the strength of interaction between a potential treatment and the mutant protein. The mutation pathogenicity prediction model study indicated that due to the lack of sufficient features and further compounded by the low correlation between the few features, the accuracy of the resulting model was not very high. Additionally, Arimoclomol was recommended for a clinical trial with Tay Sachs Disease.

Keywords: Tay Sachs, Arimoclomol, Heat Shock Therapy, Docking, Statistical Analysis

1. Introduction

1.1 Tray Sachs Disease

Tay Sachs disease is a genetic condition caused by a mutation to the HEX-A gene preventing it from creating functional β-Hexosaminidase A (Bergeron, n.d.). β-Hexosaminidase A is a protein responsible for breaking down gangliosides, molecules that accumulate on cell surfaces in the nervous system and are broken down by the lysosome (Solovyeva, et al., 2018). Tay Sachs disease is one of several Lysosomal Storage Disorders, conditions that involve the inability to break down toxic materials such as gangliosides (Lemieux, et al., 2006). Patients with Tay Sachs disease are unable to express β-Hexosaminidase A to at least 10% of the level of healthy humans, leading to gangliosidal buildup in the brain and spinal cord and the destruction of neurons as a part of GM2 Gangliosidosis (Suzuki, 2014).

This paper sought to explore both the effect of both mutation type and pathogenicity upon the binding affinity of a collection of mutant β-Hexosaminidase A strands with Arimoclomol, a commercially produced inducer of Heat Shock Protein– using Molecular Docking studies with Statistical Analysis (Correlation Tests). Additionally, a mutation pathogenicity prediction model was created using Machine Learning Algorithms to

* Corresponding Author
adityashrinivasan@outlook.com

Advisor: Rebecca Taylor
taylorre@esuhsd.org

predict the pathogenicity of future mutations, using the collection of available β-Hexosaminidase A mutations on the internet. In the writing of this paper, it was hypothesized that since existing Tay Sachs disease therapies do not differentiate in treatment methodology for mutation type, no correlation was expected between the type of mutation and binding affinity. Furthermore, due to the singular point of intervention present in most point mutations, no correlation was expected between pathogenicity and binding affinity.

Hypothetical results confirming the hypothesis would have enabled the consideration of Arimoclomol (or other Heat Shock Protein inducers) as a widespread potential treatment for Tay Sachs disease, due to perceived equal performance across a variety of different genotypical conditions. On the contrary, results negating the hypothesis would have called into question the methodology behind current treatment efforts, opening up possibilities for mutation-specific treatments developed independently and uniquely.

1.2 Heat Shock Protein

Heat Shock Proteins (HSPs) are a class of molecular chaperones that are responsible for regulating misfolded proteins, assisting in growth, and preventing degradation. They do this by binding with faulty proteins and assisting in their repair (Ingemann and Kirkegaard, 2014). Naturally produced by the bodies, HSPs contain disease-specific signaling processes, highlighting their ability to act as inflammation-suppressing agents exclusively (Dukay, et al., 2019). Regarding LSDs such as Tay Sachs Disease, Heat Shock Therapy is considered potentially promising due to the role of the lysosome in regulating homeostasis. With that organelle impaired, HSP Therapy and the chaperone capabilities of HSPs can maintain that homeostasis, additionally supporting deregulated proteins essential throughout the body (Miller and Fort, 2018).

This paper specifically explored the role of Arimoclomol, an HSP amplifier, in generating the Heat Shock Response throughout the body by triggering the production of specific Heat Shock Proteins (ALS News Today, n.d.). Currently being tested with Niemann-Pick disease type C (NPC), Arimoclomol shows promise as an instigator of protein re-folding and appreciation in lysosomal function (Susman, 2021).

The study used Molecular Docking and Machine Learning. Molecular Docking is a tool that enables researchers to predict the binding positions (modes) of a particular protein-ligand combination. In this experiment, a docking software called Autodock Vina (Trott and Olson, 2010) was used to predict the binding affinity and modes of that β-Hexosaminidase A and Arimoclomol complex, revealing potential correlations between binding affinity and either mutation type or pathogenicity. The machine learning model utilized features known from existing mutation data in public repositories to predict the classification of mutations as Pathogenetic or Benign. This allowed for the development of links between specific factors and the prevalence of pathogenetic mutations leading to Tay Sachs.

2. General Methods

2.1 Molecular Docking

Molecular Docking required several components, each of which was independently prepared. At its core, docking needs a protein, a ligand, and a Grid Box. The Grid Box exists to select a region of the protein file to dock with the ligand. In this case, the protein became the mutated copies of a HexA structure found online, while the ligand was the proposed treatment- an Arimoclomol structure.

To go about the procedure, various tools were used to prepare the components for the docking. A total of 64 simulations were performed, requiring each of the 32 protein files to be properly converted. The ligand used for each was the Arimoclomol file from the PDB (CID: 9568077), while the proteins all varied from the default HexA structure (code 2GJX).

First, each of the mutated proteins was created from the base copy of the HexA protein. A list of mutations was selected from the National Institute of Health's (NIH's) Clinvar database, (Search: tay sachs and HEXA). In total, 32 mutations were selected, 18 being missense (14 Pathogenetic, 4 Benign) and 14 being Pathogenetic deletions. Using Pymol's mutagenesis wizard, a tool to mutate downloaded structures, each of the mutations was performed. The deletions were also completed with Pymol, using the substitution feature to change each amino acid at the point of the deletion to the corresponding change. Given that there was no way to insert a stop codon, only deletions that had amino acid replacements were selected among the 14.

Several processes were used in the preparation of the 32 protein files. The downloaded .pdb files were cleaned up with standard procedures using AutoDock Tools, including deleting waters, removing non-amino acid chains, adding polar hydrogens, adding charges, and spreading the charge deficit. These processes allow for smoother docking. The ligand file, after being converted to a .pdb format, was then loaded into each docking simulation with each protein file.

The size of the Grid Box was limited, leading to the establishment of 2 docking simulations for each mutation (a total of 64). Each Gridbox covered roughly half the protein, explaining why 2 were used to test the entirety of the structure. Each simulation required the creation of a Config File, an Output File locating the contextual position of the Grid Box, and a folder containing the protein and ligand files. Utilizing AutoDock Vina, docking was executed for each mutation twice, with the Position and Binding Affinity results collected and consolidated.

Utilizing SPSS statistics, the collected data was sorted and tested for correlations between two sets of variables: Mutation Type v Binding Affinity and Pathogenicity v Binding Affinity. Again utilizing SPSS, a scatter plot was created to visualize these results.

2.2 Mutation Model

The purpose of the mutation model was to use established Tay Sachs mutation pathogenicity, type, and location data to predict the pathogenicity of new mutations or mutations of unknown pathogenicity. In order to do so, the list provided on ClinVar was downloaded, with feature consolidation including the establishment of a unique ID for model-making purposes and the addition of Mutation Type and Allele Change Classification for Single Nucleotide Polymorphisms from NIH information.

The dataset was then split into a test and training set, with basic statistical models (Log Reg, Nearest Neighbour, Support Vector Machines (SVMs), Kernel SVMs, Bayes, Decision Trees, and Random Forest) run against the data.

Several trials were run using different alterations to the dataset. Experimentally, predict2snp, a mutation pathogenicity prediction tool, was used to classify mutations that on ClinVar were listed with unknown pathogenicity in trial 5. Trials 2, 3, and 4, instead, classified all mutations of unknown pathogenicity as benign, while trial 2 only utilized missense mutations, eliminating the others from the dataset. Trial 1 was the only trial utilizing 3 distinct classifications (P, B, and Unknown). Trial 5 was referred to as the “Golden Trial” in the results.

3. Results

3.1 Docking Studies + Statistical Analysis

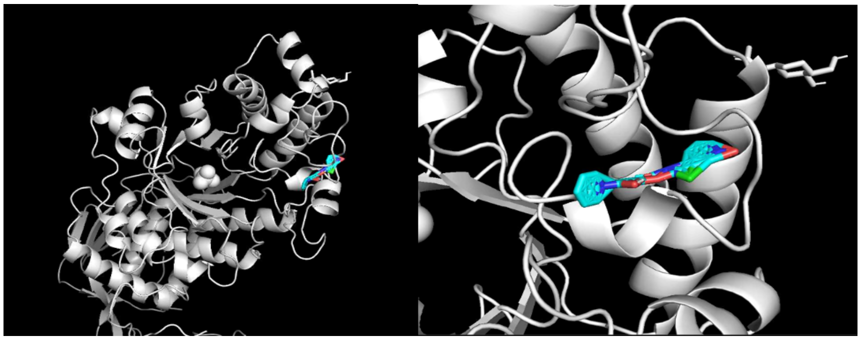
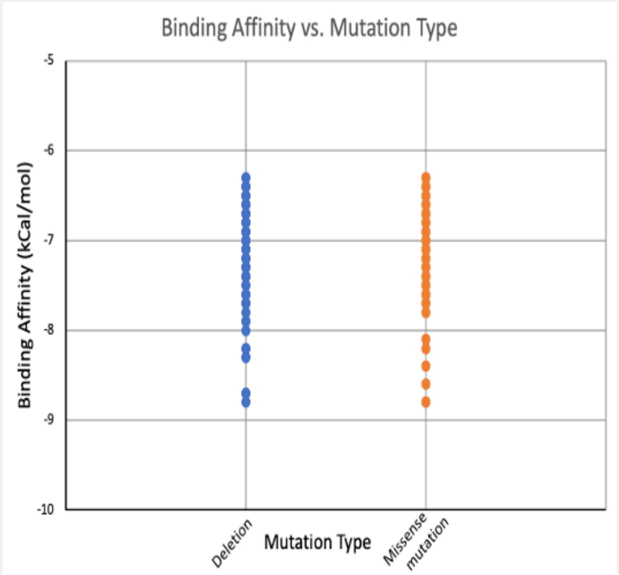


Fig. 1: Most optimal of 9 proposed binding positions (zoomed in on the right) for Arimoclomol with the mutated Hex-A (deletion of the 47th Amino Acid



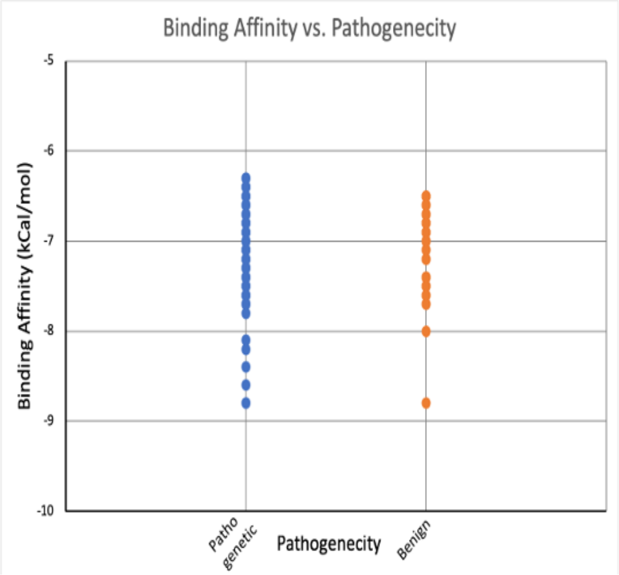


Fig. 2: Binding Affinity plotted vs Mutation type

Fig. 3: Binding Affinity plotted vs Pathogenicity

Table. 1: Correlation between Binding Affinity, Mutation type, Pathogenicity

Correlation between Binding Affinity and Mutation Type			
		Mutation Type	Affinity
Mutation Type	Pearson Correlation	1	0.021
	Sig (2-tailed)		0.634
	N	504	504
Affinity	Pearson Correlation	0.021	1
	Sig (2-tailed)	0.634	
	N	504	504
Correlation between Binding Affinity and Pathogenicity			
		Mutation Type	Affinity
Pathogenicity	Pearson Correlation	1	-0.074
	Sig (2-tailed)		0.183
	N	324	324
Affinity	Pearson Correlation	-0.074	1
	Sig (2-tailed)	0.183	
	N	324	324

3.2 Mutation Pathogenicity Prediction Model

Table 2. Performance of different Machine Learning Algorithms in predicting Pathogenicity of mutations

Parameter		Golden Trial	Trial 4	Trial 3	Trial 2	Trial 1
Dataset size		441 entries, 5 features	439 entries, 5 features	441 entries, 5 features	384 entries, 4 features	384 entries, 4 features
Characteristics of dataset		All mutations of uncertain pathogenicity were classified using predicti2snp prediction tool, included insertion and inversion mutation types.	All mutations of unknown uncertain were classified as benign, no inversions or insertion mutation types were included in the dataset.	All mutations of unknown uncertain were classified as benign, indel, duplication, and deletion mutation types included in the dataset.	All mutations of unknown uncertain were classified as benign, only missense mutations were in the dataset. Mutation Type feature not included.	3 categories of diagnosis: Pathogenetic, Benign, or Uncertain. Only missense mutations were in the dataset, without Mutation Type as a feature.
Algorithm Accuracy	Logistic Regression	63%	65%	65%	75%	43%
	Nearest Neighbor	56%	68%	64%	73%	46%
	Support Vector Machines	56%	64%	65%	75%	44%
	Kernel SVM	63%	61%	68%	73%	46%
	Naïve Bayes	45%	65%	58%	75%	41%
	Decision Tree Algorithm	59%	63%	62%	73%	49%
	Random Forest Classification	56%	63%	64%	71%	47%

4. Discussion

4.1 Docking Studies + Statistical Analysis

The goal of the statistical analysis was to reveal correlations- represented by the Pearson Correlation Coefficient, a number from (-1) to (1) constituting the strength and direction of the relationship between two variables. Paired with the significance- another number that reveals the probability of observing a specific correlation due to chance alone (lower significance is preferable), the resulting numbers can be used to create observable conclusions regarding the presence of accurate and impactful correlations.

With the observed results, the most pressing takeaway would be the noted lack of correlation between either the mutation type and binding affinity or the mutation pathogenicity and the binding affinity. These discoveries line up with the hypothesis. Figures 1, 2, and 3 display the arimoclomol-mutated Hex-A interaction and the graph of the binding affinity values first versus pathogenicity of missense mutations and second versus distinct pathogenetic mutation types.

In the first case, mutations were sorted into only missense mutations and deletions, given that those were the two most commonly occurring types of mutations, and performing insertion mutations with Pymol was difficult. The second case sorts missense mutations of benign and malignant pathogenicity against the tested binding affinity, only testing missense mutations for internal control. Table 1 (IBM, n.d.) displays results from the Pearson's Correlation Coefficient Test, comparing Binding Affinity with both Mutation Type and Pathogenicity. The 2 Tailed Significance, being over the accepted 0.001, implicates a lack of correlation between either test, suggesting that neither mutation type nor mutation pathogenicity plays a key role in the binding affinity calculation. As visualized in Figure 2, the distribution of binding affinities across both mutation types indicates that mutations vary in strength of attraction with the treatment. Given that binding affinity correlates with the strength in response to the treatment, these results indicate that neither mutation type tested has a stronger response to the Arimoclomol, and that response to the Arimoclomol varies across patients with the same mutation type.

In the second case (Figure 3), pathogenicity was tracked for binding affinity, holding the type of mutation as constant: all missense mutations. The lack of correlation observed emphasizes the finding that amino acid changes that constitute a pathogenetic variant are no less likely to have a stronger or weaker interaction than amino acids that constitute a benign mutation. It can thus be concluded that in addition to pathogenicity, amino acid changes are not correlated with the binding affinity of the Arimoclomol.

4.2 Mutation Model

Table 2 displays the trials and their corresponding results. The mutation pathogenicity predictive model suffers from a lack of features and entries. Working with an incredibly small dataset- under 500 total entries- and without many or uniquely categorized features (just position data, pathogenicity, mutation type, and factual classifications), any results must be scrutinized with the understanding of the nature of classification problems and the small dataset (Goel, 2018).

This means that although certain trials gave a higher accuracy rating, the manipulation of variables to get that accuracy rating affects the reliability of the methodology of the model. In theory, by reducing the number of features, the classification becomes increasingly binary, reliant upon correlations between a few features. If there exists a strong correlation between those features, then the accuracy rating would be high. However, if there doesn't exist a strong correlation, then the accuracy rating is much lower (ProjectPro, 2022). In either case, the model doesn't consider the holistic picture, focusing on only a few factors that individually may or may not have a high correlation with the pathogenicity of the mutation. Ideally, the model should consider a wide range of factors, analyzing all possible variables affecting mutation pathogenicity.

Due to the limited data, the model lacks the features required to make those distinctions, instead of searching for a correlation between a few individual features. Since none of the few noted features had a strong correlation, the accuracy was quite low. Still, the developed methodology, despite the lower predictive score, is indicative of the process carried out with larger data sets, outlining the processes that should be followed for the long-term development of the tool as more mutations are classified into databases and more features are categorized by researchers over time.

In the trials provided, the most complete trial would be the golden trial, which incorporated the usage of predict2snp (Bendl, et al., 2014) to classify mutations of uncertain pathogenicity as either benign or pathogenetic. Within this trial, either the Logistic Regression or Kernel Support Vector Machine yields the highest percentage- at 63% accuracy. However, the highest pure accuracy would be in the 2nd trial, with mutations of uncertain pathogenicity classified as benign, non-missense mutations excluded completely, and mutation type not yet added as a feature. The 75% reflected with the Logistic Regression, Support Vector Machine, and Naïve Bayes method reflects upon the increasingly binary nature of the classification in that problem, as there were fewer features for the model to consider- illustrating the point being made about small datasets and a limited feature set per mutation.

The methodology in the golden trial, when applied to a larger dataset with more features, would reflect the sought procedure for the development of a mutation prediction model. The conclusions to draw from these results are based on the need for a more comprehensive categorization of data. With the inclusion of a few physical features, such as the dimensions and size of the noted cherry-red spot characteristic of Tay Sachs, the model would have more factors to analyze and would become an effective tool in the immediate classification of a HEXA gene mutation as either benign or pathogenetic. Still, at its current effectiveness, the model does not successfully predict mutation pathogenicity.

4.3 Heat Shock Proteins and HSP Therapy

Heat Shock Proteins have emerged as a potential target for therapeutic intervention with several Lysosomal Storage Disorders, such as Niemann-Pick Type C (NPC). Specifically, Arimoclomol, a promoter of HSP Expression, has been explored with NPC, with promising results. HSPs are desirable for their induced responses- among which include the Heat Shock Response, a nerve response that is triggered by the body to combat various stressors such as protein misfolding (ALS News Today, n.d.).

The therapeutic capabilities of the HSR response can protect against degenerative disorders, such as Lysosomal Storage Diseases. LSDs involve the buildup of toxic substances due to a lack of specific enzymes (Orphazyme, 2021). Aggregate conditions caused by LSDs include protein misfolding- such being the case in both NPC and Tay Sachs. The capabilities of various HSPs and their capacity for controlling protein misfolding in addition to the HSR's ability to regulate cell death, autophagy, membrane permeabilization, and aid in controlling various other cellular processes make them an attractive choice for LSDs.

In mice, Recombinant HSPs have been shown to mitigate the effects of NPC, providing conclusive evidence that HSPs can aid with disorders affecting the lysosome (Susman, 2021). However, in humans, Recombinant HSPs struggle to cross the Blood-Brain Barrier, meaning that the next best solution is Arimoclomol- an HSP regulation enhancer. In mice, Arimoclomol causes co-localization of HSPs in the cerebellum with their target enzyme, allowing the "fixing" process to initiate protein re-folding. In humans, Arimoclomol was shown to mitigate disease progression in patients transferred from the placebo arm of the trial to the Arimoclomol treatment. In general, the associated NPCCSS score between the two treatments favored Arimoclomol as the more effective option, with a "65% relative reduction in annual disease progression." (Mengel, et al., 2021).

Researchers speculate that Arimoclomol in tandem with a drug that slows disease progression (such as miglustat- an inhibitor of glucosylceramide synthase- in NPC) could be the best approach for targeting NPC in humans (Susman, 2021).

Tay Sachs Disease shared several similarities with NPC, from the fact that both are autosomal recessive Lysosomal Storage Diseases to the misfolding of proteins that cause symptoms in both (National Organization of Rare Disorders, n.d.). The promise of HSPs as molecular chaperones inducing refolding efforts, autophagy, and cell permeabilization extends to Tay Sachs, where Arimoclomol would be an intriguing option to test with a clinical trial.

In the absence of a reliable gene therapy and given the doubts over the efficacy and deliverability of chaperones such as Pyrimethamine (Parenti, et al., 2015), a chaperone inducer such as Arimoclomol could help deliver the same desired effects as Recombinant HSPs without the issues in deliverability.

5. Conclusion

In the docking study and ensuing statistical analysis, neither mutation type nor pathogenicity correlated with binding affinity, meaning that the effectiveness of Arimoclomol as a treatment for Tay Sachs disease would be expected to remain the same regardless of disease type. Additionally, the study with pathogenicity confirmed that the factors affecting binding affinity do not include the amino acid changes, as unique affinities were not observed for benign or pathogenetic variants.

With regards to the mutation pathogenicity prediction ML model, the methodology has been set for future feature addition and the addition of further entries over time. An interesting exploration would be the derivation of physical features explaining physically noted traits (for example, the size of the famous red spot in the back of the eye indicative of Tay Sachs). The accuracy is understandably low, in contrast to the Hypothesis, given that all public datasets lacked data beyond the 5 features derived, of which one was the unique ID. The size of the dataset-under 500 entries for each trial, also limits the size of the test and training sets. The potential with such a model, to immediately categorize Tay Sachs mutations and provide additional time for advancement treatment, is immense, and the classification of additional features in the mutation diagnosis stage will only help it improve.

In conclusion, Arimoclomol remains an exciting option to be tested for a clinical trial for Tay Sachs Disease. The existence of testable animal models means that the search for a cure for Tay Sachs is accelerating, and the end product is closer than ever. Given Arimoclomol's success with Niemann Pick Type C, another Lysosomal Storage Disease, it is an ideal candidate for a clinical trial with Tay Sachs.

Acknowledgment

I would like to thank Mrs. Rebecca Taylor of Evergreen Valley High School for her help and mentorship during this research work.

References

ALS News Today. (n.d.). *Arimoclomol*. https://alsnewstoday.com/arimoclomol-orph-001/

Bendl, J., et al. (2014). PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS computational biology*, *10*(1), e1003440. https://doi.org/10.1371/journal.pcbi.1003440

Bergeron, S. (n.d.). *Tay-Sachs*. http://www-personal.umd.umich.edu/~jcthomas/JCTHOMAS/1997 Case Studies/S. Bergeron.html

Dukay, B., Csoboz, B., & Tóth, M. E. (2019). Heat-Shock Proteins in Neuroinflammation. *Frontiers in pharmacology*, *10*, 920. https://doi.org/10.3389/fphar.2019.00920

Goel, V. (2018, September 29). *Building a Simple Machine Learning Model on Breast Cancer Data*. Towards Data Science. https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3

IBM Corp. (n.d.). *IBM SPSS Software*. IBM Corp. https://www.ibm.com/analytics/spss-statistics-software?utm_content=SRCWW&p1=Search&p4=43700068092265694&p5=p&gclid=CjwKCAjwtcCVBhA0EiwAT1fY7xIpmbUD0Wqo98dYleTtElf5NJWjCwfOLhj7Z6rGQc_Vhl_93CMzpBoCZs0QAvD_BwE&gclsrc=aw.ds

Ingemann, L. & Kirkegaard, T. (2014). Lysosomal storage diseases and the heat shock response: convergences and therapeutic opportunities. *The Journal of Lipid Research, 55*(11), 2198–2210. https://doi.org/10.1194/jlr.R048090

Lemieux, M. J., et al. (2006). Crystallographic structure of human beta-hexosaminidase A: interpretation of Tay-Sachs mutations and loss of GM2 ganglioside hydrolysis. *Journal of molecular biology*, *359*(4), 913–929. https://doi.org/10.1016/j.jmb.2006.04.004

Mengel, E. et al. (2021). Efficacy and safety of arimoclomol in Niemann-Pick disease type C: Results from a double-blind, randomised, placebo-controlled, multinational phase 2/3 trial of a novel treatment. *Journal of inherited metabolic disease*, *44*(6), 1463–1480. https://doi.org/10.1002/jimd.12428

Miller, D. J., & Fort, P. E. (2018). Heat Shock Proteins Regulatory Role in Neurodevelopment. *Frontiers in neuroscience*, *12*, 821. https://doi.org/10.3389/fnins.2018.00821

National Organization of Rare Disorders. (n.d.). *Rare Disease Database – Tay Sachs Disease*. https://rarediseases.org/rare-diseases/tay-sachs-disease/

Orphazyme. (2021, March 3). *About Heat Shock Proteins (HSPs)*. https://www.orphazyme.com/our-science/

Parenti, G., Andria, G., & Valenzano, K. J. (2015) Pharmacological Chaperone Therapy: Preclinical Development, Clinical Translation, and Prospects for the Treatment of Lysosomal Storage Disorders. *Molecular therapy : the journal of the American Society of Gene Therapy*, *23*(7), 1138–1148. https://doi.org/10.1038/mt.2015.62

ProjectPro. (2022, May 24). *Classification vs. Regression Algorithms in Machine Learning*. https://www.projectpro.io/article/classification-vs-regression-in-machine-learning/545

Solovyeva, V. V., et al. (2018). New Approaches to Tay-Sachs Disease Therapy. *Frontiers in physiology*, *9*, 1663. https://doi.org/10.3389/fphys.2018.01663

Susman, E. (2021, October 6). Arimoclomol *Appears to Slow Progression of Niemann-Pick Type C Disease*. *Neurology Today*. https://journals.lww.com/neurotodayonline/blog/neurologytodayconferencereporterscnsannualmeeting/pages/post.aspx?PostID=48#:~:text=October%206%2C%202021&text=%E2%80%8BTreatment%20with%20the%20investigative,Child%20Neurology%20Society%20annual%20meeting

Suzuki, Y. (2014). Emerging novel concept of chaperone therapies for protein misfolding diseases. *Proceedings of the Japan Academy. Series B, Physical and biological sciences*, *90*(5), 145–162. https://doi.org/10.2183/pjab.90.145

Trott, O., & Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, *31*(2), 455–461. https://doi.org/10.1002/jcc.21334