NEW BIOINFORMATIC METHODS OF BACTERIOPHAGE PROTEIN STUDY

by

Emily Kerstiens

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

School of Agricultural and Biological Engineering

West Lafayette, Indiana

May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Kari Clase, Chair
School of Agricultural and Biological Engineering

Dr. Somali Chaterji
School of Agricultural and Biological Engineering

Dr. Stephen Byrn
School of Pharmacy

Approved by:
Dr. Nathan S. Mosier

To my family, for always supporting me and believing in me

To my best friends, for always encouraging me and making me smile

ACKNOWLEDGMENTS

I'd like to thank my committee chair, Dr. Kari Clase, for her guidance throughout my undergraduate and graduate studies. I appreciate her encouragement in exploring new research opportunities, even those outside the scope of my expertise. Dr. Clase's guidance led me to find my passion in computational life science and exposed me to many valuable research projects and experiences.

Thanks are also due to Dr. Stephen Byrn who advised on my pharmaceutical projects. His expertise was invaluable as I was learning in this new field of study.

I would also like to thank Dr. Somali Chaterji and Dr. Michael Gribskov for igniting my interest in computational projects and bioinformatics. Their perspective on different algorithms and coding methods helped shape my research.

Lastly, I'd like to thank my family and friends for always believing in me even when I doubted myself.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NKF	No Know Function, a hypothetical protein
PDB	Protein Data Bank
RMSD	Root-Mean-Square Distance, used to measure differences in protein structure
nR	Non Redundant Protein Database
QbD	Quality by Design, an approach to manufacturing pharmaceuticals that focuses on the quality of the manufacturing process
FDA	Food and Drug Administration
NBCD	Non-Biological Complex Drug
FAER	FDA Adverse Event Report
HHMI	Howard Hughes Medical Institute
GO	Gene Ontology
CDD	Conserved Domain Database
SCOP	Structure Classifications of Proteins
FAERS	FDA Adverse Event Reporting System
AERs	Adverse Event Reports
HMM	Hidden Markov Model

ABSTRACT

Bacteriophages are viruses that infect and kill bacteria. They are the most abundant organism on the planet and the largest source of untapped genetic information. Every year, more bacteriophages are isolated from the environment, purified, and sequenced. Once sequenced, their genomes are annotated to determine the location and putative function of each gene expressed by the phage. Phages have been used in the past for genetic engineering and new research is being done into how they can be used for the treatment of disease, water safety, agriculture, and food safety.

Despite the influx of sequenced bacteriophages, a majority of the genes annotated are hypothetical proteins, also known as No Known Function (NKF) proteins. They are expressed by the phages, but research has not identified a possible function. Wet lab research into the functions of the hundreds of NKF phages genes would be costly and could take years. Bioinformatics methods could be used to determine putative functions and functional categories for these hypothetical proteins. A new bioinformatics method using algorithms such as Domain Assignments, Hidden Markov Models, Structure Prediction, Sub-Cellular Localization, and iterative algorithms is proposed here. This new method was tested on the bacteriophage genome PotatoSplit and dropped the number of NKF genes from 57 to 40. A total of 17 new functions were found. The functional class was identified for an additional six proteins, though no specific functions were named. Structure Prediction and Simulations were tested with a focus on two NKF proteins within lytic phages and both returned possible functional categories with high confidence.

Additionally, this research focuses on the possibility of phage therapy and FDA regulation. A database of phage proteins was built and tested using R Statistical Analysis to determine proteins significant to phage infecting *M. tuberculosis* and to the lytic cycle of phages. The statistical methods were also tested on both pharmaceutical products recalled by the FDA between 2012 and 2018 to determine ingredients/manufacturing steps that could affect product quality and on the FDA Adverse Event Reporting System (FAERS) data to determine if AERs could be used to judge the quality of a product. Many significant excipients/manufacturing steps were identified and used to score products on their quality. The AERs were evaluated on two case studies with mixed results.

1. INTRODUCTION

This chapter is an overview of bacteriophages, protein bioinformatics, and the applications of this work. It includes a statement of purpose for the work done and provides the research questions asked in this thesis. Key terms are defined, and the scope and limitations of the work are given.

1.1 Statement of Purpose

Bacteriophages are viruses that infect and kill bacteria. They have applications in many industries included water treatment, agriculture, food biocontrol, and the treatment of bacterial infections. Bacteriophages are the most abundant organism on the planet, but there is still much unknown about them. To safely use them in these applications, they must be thoroughly studied to ensure no harm comes from them.

1.2 Research Questions

- Can statistical analysis be used to predict putative proteins significant to phage infection cycles?
- How can existing bioinformatics programs be used to better predict putative protein functions to decrease the number of proteins currently classified as “unknown function”?

1.3 Scope

This study uses the bacteriophage genomes annotated at Purdue University. This is 31 bacteriophage genomes covering eight Clusters and both bacteriophage life cycles. Phages were isolated at Purdue University using direct plating and two rounds of serial purification according to the SEA-PHAGES protocols[1]. *M. smegmatis* was grown on 7H9 media with albumin and dextrose (AD) at 37 °C. The phage DNA was extracted using a Wizard DNA prep kit from Promega. Transmission electron microscopy was used to characterize the morphology of phages. The genomes were sequenced by the Pittsburgh Bacteriophage Institute using Illumina MiSeq and assembled using Newbler (v2.5 for MrGordo, v2.8 for Afis, and v2.7 for others) and Consed (v20 for MrGordo, v28 for Afis, and vs22 for others).

Genomes were annotated in DNA Master v5.0 (http://cobamide2.bio.pitt.edu/computer.htm) and PECAAN v1 (https://pecaan.kbrinsgd.org/). The putative genes were predicted using Glimmer v3.02[2] and GeneMark v2.5[3]. Default settings were used for all softwares unless otherwise specified. The functions of genes were assigned using Phamerator[4], BLASTp[5], and HHPred[6].

R Studio was used for all statistical analysis using a logistical regression model and a Firth bias correction method. Significance of 0.05 was used.

1.4 Significance

It is estimated by the Center for Disease Control that more people will die from antibiotic resistance infections than from cancer by the year 2050[7]. A total of 10 million people are estimated to die in 2050 from these resistant microorganisms if nothing changes[8]. To combat this issue, research is being done into the effectiveness of bacteriophages, viruses that infect and kill bacteria, as treatments[9–12].

Mycobacteriophages are those that infect the host *M. smegmatis*, a nonpathogenic relative of *tuberculosis*[13–15]. Because of this similarity to tuberculosis, it is thought that if mycobacteriophages are isolated, they could infect tuberculosis as well. Based on this hypothesis and a desire to understand the genetic diversity of phages, the Howard Hughes Medical Institute launched the SEA-PHAGES Program[16,17].

Every year, more bacteriophages are isolated and sequenced. However, when annotating these genomes, the current method relies heavily on direct sequence comparison to a database. Using BLASTp and a local alignment within the Phages Database only provides matches that are in the database and many report back as “hypothetical proteins” or No Known Function proteins. If an NKF protein goes into the database as “hypothetical protein”, then every year more “hypothetical protein” matches are made, but this brings the community no closer to identifying the true function.

Around 69% of phage proteins are NKF proteins. With so much unknown about bacteriophages, their mechanics, and how they function, they are not yet a viable option for mass pharmaceutical production. They cannot be guaranteed to be safe treatments, and they cannot be engineered to be as efficient as possible. To use bacteriophages as a readily available treatment, more work needs to be done to understand their infection cycle and the proteins they create. A new method needs to be devised to fill this gap so phages can be safely regulated and used in the fight against antibiotic-resistant infections.

1.5 Assumptions and Limitations

Due to every phage being annotated as part of the Howard Hughes Medical Institute's (HHMI) SEA-PHAGEs program[17], the data here is viewed through their lens. The phage annotation process adheres to the quality guidelines set by HHMI and that quality evolves every year. Some phages annotated years ago do not meet today's quality standards. Because of this, the statistical analysis presented is limited by the quality of the annotations.

This work is limited because there was no wet-lab research done to confirm results. This is all bioinformatics and *in silico* methods that could lead to future projects in the wet lab.

1.6 Definition of Key Terms

- **lytic:** phage life cycle that immediately produces phage particles leading to the lysis of the host
- **lysogenic:** phage life cycle that involves DNA lying dormant in the host cell
- **temperate:** a phage that can enter the lytic or lysogenic life cycle
- **Cluster:** a grouping of similar phages based on overall genomic similarity
- **Structural Protein:** a protein relating to the physical structure of a phage particle such as capsid or tail
- **Phage Therapy:** the use of bacteriophages to treat bacterial infections
- **Phage Cocktail:** a phage therapy treatment consisting of multiple phages within one package
- **Titer:** the concentration of phage particles within a sample, typically measured in plaque-forming units per volume.
- **BLAST:** Basic local alignment tool, compares the sequences of two genes or proteins and outputs statistical values on the match

- **E-value:** the expected number hits one may see by random chance when searching a database of this size. Based on the p-value and size of a database.
- **P-value:** the probability value of an event occurring assuming two factors are unrelated
- **HMM, Hidden Markov Model:** a statistical model using one factor to learn about another

2. LITERATURE REVIEW

This chapter will go over specific existing work in the field of bacteriophage protein informatics and FDA regulations.

2.1 Bacteriophages

2.1.1 Bacteriophage Background

Bacteriophages are viruses that infect and kill bacteria. They are host-specific and the work here will focus on mycobacteriophages, phages that infect mycobacterium hosts. Specifically, the phages here were isolated using the host *M. smegmatis*. Figure 2.1 below shows the structure of a bacteriophage. Phages have a capsid/head that holds their DNA and then a contractile tail that provides motility and allows them to infect bacteria. The tail fibers recognize specific proteins on bacterial cells and attach there, then injecting their DNA[1]. Because of this, the infection process is specific to the bacterial host the phage infects and does not interact with other cells.

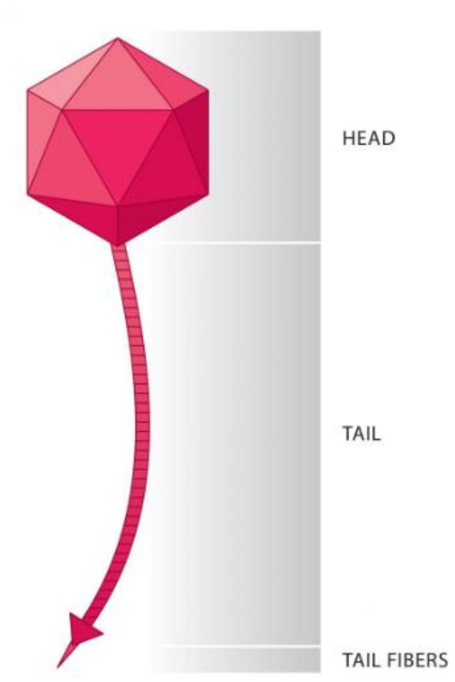


Figure 2.1. The structure of a bacteriophage with the capsid/head, tail, and tail fibers for attachment to bacteria[1].

Bacteriophages kill host cells by injecting their DNA and creating more phage particles until the host cell lyses. This is referred to as the lytic cycle and can be seen in Figure 2.2. In this cycle, first, the phage inserts its DNA into a bacterial cell and then integrates it with the host DNA. The host then begins to replicate and translate the viral DNA, producing more phage particles. Eventually, the cell lyses because of the phage created within it[13,18–20].

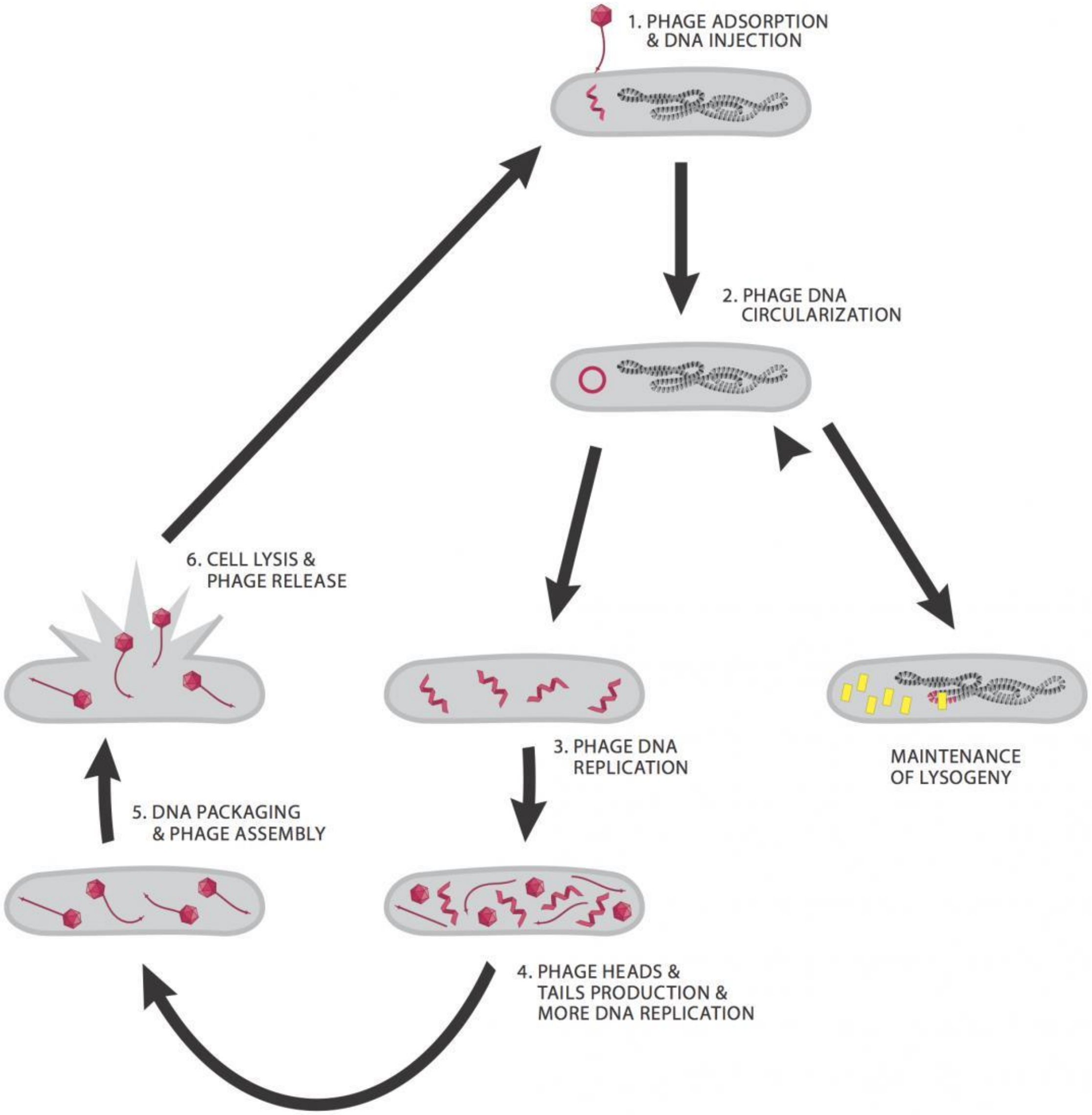

1. PHAGE ADSORPTION & DNA INJECTION
2. PHAGE DNA CIRCULARIZATION
3. PHAGE DNA REPLICATION
MAINTENANCE OF LYSOGENY
4. PHAGE HEADS & TAILS PRODUCTION & MORE DNA REPLICATION
5. DNA PACKAGING & PHAGE ASSEMBLY
6. CELL LYSIS & PHAGE RELEASE

Figure 2.2. (1) A phage attaches to a host bacterium using its tail fibers and injects its linear chromosome. (2) The phage chromosome circularizes and it is either maintained in the host as a prophage (shown here integrated into the bacterial chromosome in red) or enters the lytic cycle. The expression of the lytic genes is prevented through the constant expression of the immunity repressor protein (shown as yellow rectangles). When the lysogen is stressed, the prophage may excise and begin the lytic cycle. (3) The phage DNA is replicated. (4) New tail and capsids are produced. (5) New virions are assembled. (6) The bacterial cell is lysed and the new virions are released. Image and caption from the SEA-PHAGES Program[1].

Phage can be either lytic or temperate. Temperate phages can utilize the lytic cycle and the lysogenic cycle (Figure 2.2). Lytic phages cannot enter the lysogenic cycle. In the lysogenic cycle, phages can lie dormant within the host after inserting their DNA. The phage continues to live within the bacterial cell without lysing and killing the host. Phage DNA can cause increased bacterial infectiousness and worsen the severity of a disease[18,21].

Bacteriophages have highly mosaic genomes and high genetic diversity[13,18,20,22,23]. They are the most abundant microorganism on the planet, with an estimated 10^{32} bacteriophages on the planet.[24] Because of their high genetic diversity, phage genomes have been broken down into clusters. Clusters are groups of phages based on their overall genetic similarity and GC content[13,25].

GC content is the percent of the genome that is either a Guanine or Cytosine nucleotide base, and for phages, they typically have similar GC content to their bacterial hosts[13,20,25]. Phages are highly conserved within each cluster but bear little similarity to those outside the cluster. Phage clusters are either temperate or lytic and have a unique set of genes. Additionally, some clusters can be further broken down into subclusters, which contain phages of even stronger similarity. Mycobacteriophages have a total of 22 clusters[15]. The relationship of phage diversity for mycobacteriophage can be seen in Figure 2.3.

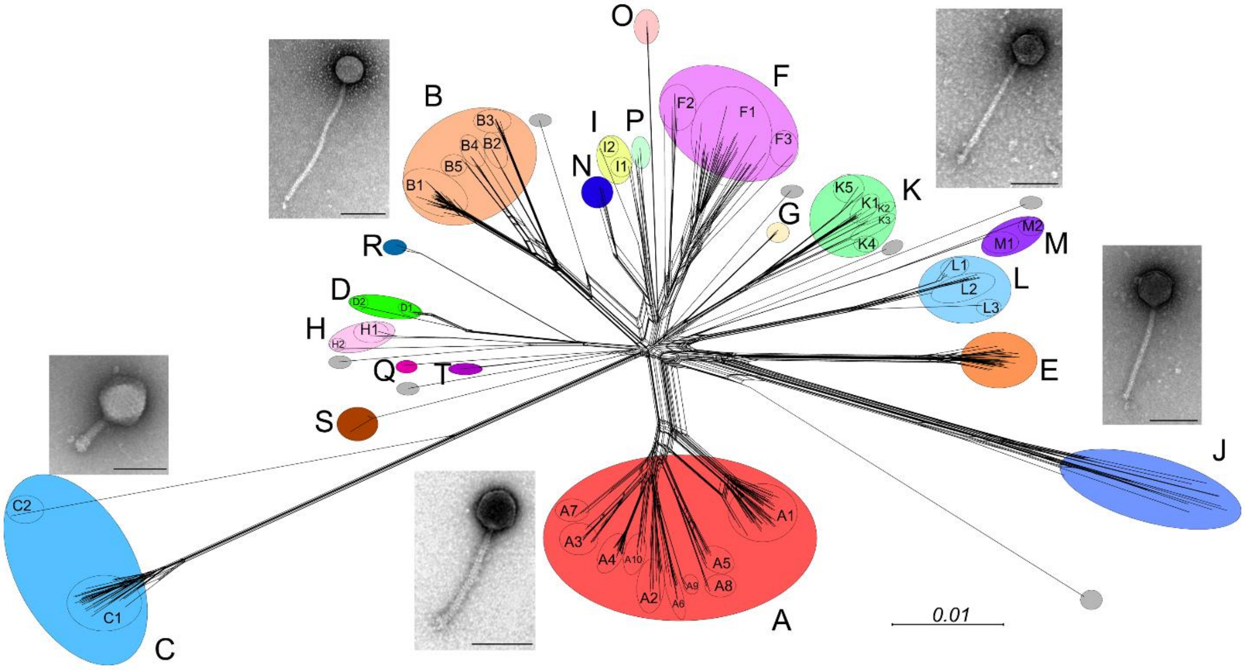


Figure 2.3. The diversity of mycobacteriophages within each Cluster from Hatful 2014[14].

Phage genomes are typically 100 to 300 genes with small, tightly packed genomes[15,20,23]. Their genomes are organized into cassettes, meaning genes of similar functions are often grouped on the genome[13,15,23]. Genomes can be annotated using bioinformatics software to determine gene location and putative protein function[1].

Protein functions are assigned based on homology. Homology refers to two proteins sharing a common ancestor. Based on statistics, one can rule that two proteins have high similarity and thus are homologs. If they share a common ancestor, one can assume they have the same function[6]. Using current methods, many proteins still return “hypothetical protein” or No Known Function (NKF) as their function. This means a protein is produced by the phage, but the function is unknown.

Protein function is predicted using BlastP, a sequence comparison to the Nonredundant Protein database (nR) and the Phages Database (PhagesDB)[5], and HHPred, a Hidden Markov Model (statistical algorithms) based on secondary structures compared to Protein Data Bank (PDB), Conserved Domain Database (CDD), pFam, and Structural Classifications of Proteins (SCOP)[6]. These tools are considered the “gold standard” of bioinformatics programs, but there are many more new and emerging methods that could yield results for phage proteins[1,20].

2.1.2 Previous Analysis with Hypothetical Proteins

Previous work into the annotation of hypothetical proteins has focused mainly on bacterial proteins[26–32]. While these methods have never been applied to phages, they provide a detailed background that can be used to inform phage research in this area.

Annotation of hypothetical proteins relies on exploratory methods and new programs. Many of these methods fall into the following categories: Physiochemical Properties, Sub-cellular Localization, Alignment Programs, Hidden Markov Models, Domain Assignment, Protein-Protein Interactions, and Structure Predictions[26–32].

The program ProtParam on ExPASy predicts the molecular weight, theoretical pI, amino acid composition, atomic compositions, extinction coefficient, estimated half-life, instability index, aliphatic index, and average hydropathicity[33]. ExPASy predicts based on the database Swiss-Prot and TrEMBL, both part of the UniProt databse[33]. This could be used for analysis into how the chemical properties of a protein may affect function but would take detailed work into the research of each protein. It could also provide information to be used as a feature for a machine learning algorithm[34,35].

Sub-cellular localization programs like TMHMM can predict the location of a protein inside the cell, outside the cell, or within a membrane by identifying transmembrane helices and the solubility of proteins[36]. TMHMM uses a Hidden Markov Model trained on PDB crystal structures[36]. Categorizing a protein as a membrane protein is a step forwards from a protein labeled hypothetical and can help narrow down more specific functions like holins that are found within the membrane.

Alignment programs such as PSI-BLAST are iterative programs that discover new homology matches the more times the program is iterated[37]. This is done by using the query and database hits to build a profile that is used as the query in the next iteration[37]. This type of profile building can result in hits that the typical BLAST algorithm may miss.

Hidden Markov Models are a type of statistical algorithm, sometimes considered a machine learning method, that can predict homology between two proteins. HHPred is the standard program used in phage annotation[1]. It predicts homologs based on structural predictions and pairwise comparisons. HMMER also predicts homologs but uses three different queries including sequence comparison, domain comparison, and profile building based on multiple sequence alignments[38]. HHBlits is an iterative algorithm that searches using both sequences and structure predictions[39]. Each algorithm has a variety of databases to choose from, including UnitRef, PDB, pFam, and CDD.

Domain Assignment is based on smaller amino acid motifs, or domains, found in specific proteins[27,29–32]. Rather than trying to find a match to the entire protein sequence, these are smaller sequences typically contained within larger proteins. These can be enzyme binding sites, activation sites, or another type of motif that gives a better understanding of function. There are databases of domains, including CDD and pFAM. The above algorithms can be used to search for domains by selecting these databases.

Protein-protein interaction networks model associations between genes and assume functional links[28,30–32]. In prokaryotes, genes of the same function tend to be close. This is also true in phages that contain cassettes of similar genes. There are databases such as STRING that analyze these associations and predict protein interactions[40]. However, these programs are analyzed based on the organism assigned to them and there are no phage entries yet in these databases. The host bacteria *M. smegmatis* is not yet within these databases either, so their usefulness to phages is low.

Structure Predictions are used in the HHMs, but can also be stand-alone programs such as I-TASSER. I-TASSER predicts structure by generating 3D atomic models using threading alignment programs and iterative assembly simulations. It matches the 3D models with known proteins in PDB and associates them with Gene Ontology (GO) terms[41]. This algorithm takes days to run on a server and is not a good option for searching entire genomes, but offers a more in-depth analysis.

Machine learning methods are also a new front for protein annotation[35,39,42–45]. Many different machine learning algorithms can be implemented to determine the function of phage proteins or their locations within host cells. Features used can range from the physicochemical properties to the functions of nearby genes to amino acid sequences divided into *k*-mers[42–44,46,47]. The *k*-mers of the amino acid sequence are typically two to three amino acid groupings that are each used as an input feature in order. Another issue to consider when using protein sequences is length. Machine learning algorithms must have the same number of input features for every data point, so longer genes need to be shortened and shorter genes padded to ensure the same number of features for every protein[34,42,47]. Feature selection is typically done with ANOVA ranking, and actual algorithms can include Support Vector Machine (SVM), Naïve Bayes, Deep Neural Networks, and Logistic Regression[26,34,42,44,47].

2.2 Food and Drug Administration

2.2.1 FDA Pharmaceutical Regulation Background

The current FDA process for drug approval relies on an innovating company undergoing drug discovery and then three phases of clinical trials to collect data. The innovating company then has exclusive rights to produce the product for a number of years. After that time, companies that want to produce a generic version can reproduce a slightly different drug without undergoing clinical trials[48–50]. To gain FDA approval, all generic drug makers have to do is prove that their product performs equivalently to the innovator drug[48,49]. Generic drugs currently submit Abbreviated New Drug Applications (ANDAs) and they do not have to undergo expensive clinical trials. An ANDA may be accepted by the FDA if the company has shown sufficient proof of bioequivalence between the two drugs, meaning that the new product has the same active ingredient and produces the same effect without being exactly the same[51,52]. Differences between a generic and an innovator drug must not have significant therapeutic effects[51,52]. They may make small changes to the drug in the interest of keeping costs down, trading certain inactive ingredients for others.

With biologics, however, this is not as simple. Biologics are specific protein products that perform one function, such as cell therapy medicinal products, STEM-cell engineered products, and tissue-engineered samples[51–54]. Unlike chemicals, proteins and amino acids cannot be easily swapped out and changed without changing the entire product. Some features of a protein may be key to its function, while others may be changeable for a generic drug. The FDA is currently handling this balancing act between creating a generic drug without having to undergo extensive clinical testing and the risk of ineffective or dangerous drugs[51–53].

Furthermore, some biologic products are crafted for each patient based on the patient's cells, and therefore every variation of the product cannot undergo clinical trials[55–57]. This is the issue with the new and upcoming personalized medicine trend. Even in cancer treatments, medicine is being personalized to fit the individual patient rather than a big-pharma mass-produced drug[57]. When the treatment is different for every patient, the issue of regulation becomes more complex. This applies to phage therapy products that may be screened for each infection[55,56,58].

On the other end of the spectrum, after a drug is approved, the FDA continues to monitor products and if needed, recall a product. Drug recalls are one of the most important actions the FDA can take to protect the public from potential adverse effects of pharmaceutical products[59–61]. When problems with a drug are discovered, either by the company or the FDA, the product is taken off the shelf. Drug recalls can happen for a variety of reasons, from labeling errors to drug product degradation. Some errors lead to small problems, and others could have a fatal outcome[59–61]. Formulation-based recalls are recalls for reasons related to the formulation of the product, such as:

- Contamination
- Defective Delivery
- Dissolution Specifications
- Failed Specifications
- Failed Tablet Specifications

- Foreign Substance
- Impurities and Degradation
- Presence of Particulate Matter
- Resuspension Problems
- Stability
- Sterility
- Subpotency
- Superpotency

Some pharmaceutical products are at higher risk for quality issues than others and a recall can have a bigger impact on some products. In the case of epilepsy drugs, there are 25 common epilepsy drugs in the United States. Losing even one of these products can put a segment of the population at risk for seizure and put more demand on the other drugs, ultimately leading to a double drug shortage.

Additionally, non-biological complex drugs (NBCDs) are a constant regulatory challenge[52,62,63]. These drugs are not biological products but contain complex active ingredients, formulations, routes of delivery, or dosage forms[64]. The rise in biotechnology innovation has led to an increase in these complex drug products. This scope of complex products spans from nanotechnology to topical to inhalation drugs. These products also pose a high risk because they are often manufactured by an involved process. For example, transdermal patch formulations and gels formulations have been implicated in several fatalities[65–68].

These existing pathways of drug approval and regulation can be adapted for bacteriophage products. The FDA process is important in maintaining the quality of pharmaceutical products, both new products, and existing products. However, there is a need for growth in the inclusion of personalized medicines and biological products[52,53,55,56].

2.2.2 History of Phages

In the 1940s, there was a race to see who could come up with a bacterial infection treatment first: antibiotics or bacteriophage[18,21,24]. The antibiotic community was focusing on growing larger quantities of antibiotics, and they eventually succeeded when the first mass production of penicillin was achieved[18,24]. With the success of antibiotics, the interest in bacteriophages as a treatment fell away[18,21,24].

Antibiotics are compounds produced by living organisms that inhibit the growth of bacteria. Bacteriophages, however, are viruses that infect and kill bacteria[13,19,20]. There are phages known for every common bacterial infection and the phages of each host are unique from each other[20]. Historically, they have been used and identified as treatments before antibiotics rose to fill the need[18,21,24]. This use, however, was not the commercialized pharmaceutical world that exists in the 21st century. Every case was anecdotal, with one scientist studying the results of phages[18,21,24]. When companies in the 1940s attempted to use bacteriophages as a regular treatment, reports came back with inactive phages and the death of patients[21,24].

There was not enough known about bacteriophages and the technology needed to study them thoroughly did not exist yet. Antibiotics were a far safer alternative and easily mass-produced for patient consumption, so the world moved on from phages. It was not until the 2000's that interest rose again in bacteriophages as a treatment option. With antibiotic resistance on the rise, new and alternative treatment methods are being studied.

In the wake of this new challenge, people began studying bacteriophages with renewed vigor. What the scientific community could not accomplish in the 1940s can be accomplished today. Bacteriophage genomes can be easily sequenced and studied, bacteriophages are isolated from the environment in mass, and synthetic biology techniques can even be employed to make phage therapy more effective[10,69–74]. Recently, clinical trials and cases of emergency treatment have been seen in the United States and elsewhere with positive patient outcomes[9,11,12,58,75]. Phages can be screened and selected for their therapeutic qualities, and even engineered to be better suited for treatment. The new techniques and methods from synthetic biology have opened the door for phage therapy to be a viable option to treat bacterial infections[10,58,69,70,74,76].

However, for everything the community has learned about phages, there are still gaps in knowledge that need to be filled. The pharmacodynamics, safety, and efficacy of phage as pharmaceuticals are not well studied[21,71]. Furthermore, every phage is genetically unique. A company may prove one phage is safe for treatment, but does that mean every phage is? Research shows cocktails of multiple phages are the most effective[9,58,70,75], but the money needed to run clinical trials for every phage by itself and as a group would be astronomical. The current regulatory standards for pharmaceuticals do not easily allow phage therapies to be developed and to be economically feasible.

Many problems need to be addressed before phage therapy becomes a common treatment for infections. One issue with using bacteriophages as treatment is that every phage is unique. Even among phages that infect the same host, there are large differences among them, such as the two life cycles lytic and temperate.

The lytic life cycle is more desirable for phage therapy options because lysogenic phages can increase bacterial infectiousness and fail to treat the disease if they enter the lysogenic cycle[9,77]. In some cases, synthetic biology can be used to remove proteins related to lysogeny and force phage to remain in the lytic life cycle[11,58,70]. The knowledge of these proteins is instrumental in using phages as treatment.

Screening and testing these phage Clusters revealed that cluster K and subclusters A1, A2, and A3 have proved more effective at infecting tuberculosis[78,79]. This is unexpected because these clusters can enter the lysogenic cycle and fail to kill bacterial cells. When they stayed in the lytic cycle, however, they yielded high rates of tuberculosis infection and eventually lysis of the host cells.

Studying these differences between proteins found in different clusters and different life cycles can help identify the best candidates not only for phage therapies but also for applications for food safety, agriculture, and water treatment. Once this knowledge is available, synthetic biology techniques could be used to edit the A and K cluster phages into being lytic only or to alter already lytic phages to mimic their characteristics[78,79].

However, before bacteriophages can become a safe and effective treatment, a regulatory model needs to be developed. There are a few ideas for adapting the FDA model for phage purposes, from using the standard pathway to the biologics pathway to the new "microbiome-based drug"[9,69,80,81].

Currently, there are 57 open clinical trials in the United States for a phage product, ranging from university studies to pharmaceutical companies[82]. It is more common in Europe, where regulations allowed the easier study of bacteriophage products[69,83]. The evolving nature of phages, being organisms that can replicate within host cells and change their genomes, makes them unlike any other pharmaceutical products[83]. Additionally, the use of phages as personalized medicine, screening every specific bacterial strain for highly infectious phages, opens the door to tailor-made products, which are currently too experimental and costly to be adapted to mass production[9,80,83]. In all emergency use scenarios in the United States, personalized therapies were created after screening the infecting bacteria[9].

2.2.3 Case Study: Statistical Analysis of Drug Recalls

This case study focuses on the application of statistical methods to determine significant factors in drug product manufacturing.

Publicly available data from the FDA enforcement report program, FDA labels, and patents can be combined to determine potential leading indicators of drug recalls. Previous research into all formulation-based recalls has shown significant risk factors ranging from the dosage form to release mechanism[84]. There have also been case studies investigating the formulation factors that may affect different products, but it was done on a drug-by-drug basis[85,86]. Knowing sources of error and risk factors allows the manufacturing company to control them and minimize that risk.

In this regard, there has been a movement in the FDA for Quality by Design (QbD), using a thorough understanding of a product and manufacturing process at the beginning of a product design to reduce recalls and quality issues later[87–91]. QbD uses the understanding of raw materials and manufacturing steps to design the best possible process for a product with the highest quality.

There is a large variety of pharmaceutical processing steps that can be used to manufacture a drug. Most dosage forms involve mixing or blending, and some use both[92–96]. Heating or filtration steps are used depending on the formulation design and dosage form. Granulation and the drying of granules are also typically used in pharmaceutical manufacturing[92–96]. Granulation is the process of turning dry powder with fine particles into a larger mass with multiple particles packed together. The process can be wet or dry and is often used in pharmaceutical manufacturing to transform and combine powder products[92–96].

More challenging processing steps require specialized equipment or more attention to the process[92–96]. Some of these steps are forming a gel matrix, forming an emulsion, applying a drug reservoir to a backing, and forming particles[92–96]. Applying a drug reservoir is a process used to create transdermal patch drugs by binding the drug product to the polymer backing. This process can involve two different methods: (1) laminating or knife coating, which involves applying pressure to combine two or more layers with a binding agent, and (2) spray coating, which involves spraying the drug product onto the adhesive and backing[67,97].

These steps involve complex processes that require exact calibrations and amounts of ingredients. Even the ingredients used can play a role in these steps, as different stabilizers, adhesives, preservatives, or binders can affect the results.

Excipients play a large role in all processing steps. Binders, stabilizers, and granulation agents can affect many dosage forms and recall reasons. The exact significance of different excipients and processing steps will be analyzed for the drugs in this study. Studying the significant steps and ingredients that have led to recalls in the past will better inform QbD of new products.

2.2.4 Case Study: FDA Adverse Event Report

The FDA Adverse Event Reporting System (FAERS) is part of the FDA's effort to maintain safety surveillance on approved products. Consumers, patients, and health professionals can voluntarily submit adverse event reports, while industry reports from manufacturers, distributors, and importers are mandatory[98]. These event reports ask users to report the primary suspect in the adverse event, the date occurred, the manufacturer, patient health details, and the outcome of the event[98].

3. METHODOLOGY

3.1 Purdue Phage Database & Statistical Analysis

3.1.1 Building a Database

Two hundred and eighteen bacteriophages were isolated at Purdue University by the year 2020 with the intention of purifying and sequencing phages. Using the guidance from Howard Hughes Medical Institute's SEA-Phages program, bacteriophages were isolated from environmental samples using the host *M. smegmatis*. After isolation, the bacteriophages were amplified and purified using serial dilutions and filtering with a 0.22 micrometer filter. Their DNA was extracted from high titer lysates and sequenced by Pittsburgh Bacteriophage Institute by using Illumina Sequencing[1].

Of those 218, 25 had their DNA sequenced and then annotated. An additional six genomes were also annotated at Purdue University, for a total of 31 genomes submitted to GenBank. Glimmer[2] and GeneMark[3] were used to predict gene locations, then functions were assigned using Phamerator[4], NCBI Blastp[99], HHPred[6], and BLASTp on PhagesDB[100].

These 31 phages were used to build a database of bacteriophage information in Excel. To collect information, a number of websites were used. Information taken from PhagesDB[100] includes: the year found, year annotated, isolation institute, location found, GenBank accession numbers, Cluster, life cycle, morphotype, number of genes, GC content, and archive lysate titer value. Based on the attached photo of phage plaques (if available), the morphology information on plaques was recorded. Plaques could be large, tiny, or medium and could be halo, clear, or cloudy. Capsid size was taken from the attached electron microscopy image (if available) and measured.

From NCBI's GenBank, the genome file was downloaded as a text file and a Python code was used to parse the information on each gene. For every phage genome, the gene sequence, start site, stop site, length, direction (reverse or forwards), and putative protein product were taken. A total of 3,814 genes were annotated and cataloged. Functional categories were assigned using the InterPro database and scholarly articles from the PubMed database. The categories used are:

- **Structural:** Proteins related to phage structure such as tail proteins, major capsid proteins, and membrane proteins.

- **DNA Replication/Translation:** Proteins related to the replication and translation of DNA such as DNA Polymerase, Helicase, and terminase.
- **DNA Integration:** Proteins related to the insertion of phage DNA into the host cell through the use of recombination such as immunity repressors, Integrase, and holiday junction proteins.
- **Lysis:** proteins relating to the destruction of the host cell such as holing, Lysin A, and Lysin B.
- **Other:** proteins that do not fall into the above categories and may span multiple categories such as biosynthesis proteins like O-methyltransferase or signaling pathways.

3.1.2 Statistical Analysis

The information from this database was parsed and used in statistical analysis to determine significant protein functions within categories. Independent variables were the presence of specific protein functions within a phage coded as binary values, 1 being "yes, the protein is present" and 0 being "no, the protein is not present". Dependent variables studied were: life cycle type and ability to infect *M. tuberculosis*. Statistical analysis was done using R with a logistical regression model and a Firth bias correction method. Significance of 0.05 was used.

3.2 Case Study: Pharmaceutical Quality Recalls

3.2.1 Building Database.

To begin, all drug recalls from June 2012 to December 2018 were downloaded from the FDA enforcement report program[84]. At the time of download, a cut-off date of December 2018 was chosen because it was the last full month available. The recalls were sorted to remove all drugs recalled for packaging or label issues, as these were not recalled for a product quality issue. Duplicate products with the same recall ID numbers were removed to avoid counting the same data more than once.

After that, a Python code was used to extract all the complex drugs and epilepsy drugs from the list. Epilepsy drugs were found by a specific drug name, based on a list of epilepsy drugs. Complex drugs were identified by a specific name and by the definition offered by the FDA, which states complex drugs are those with complex active ingredients, formulations, routes of delivery, and dosage forms[64]. The focus was kept on Epilepsy and Complex Drugs due to their prevalent regulatory issues. No products that were not recalled were used due to the scope of the experiment. The analysis was instead done by comparing the reasons for recall.

Based on these definitions, 178 complex drugs were identified, and 31 epilepsy drugs were identified, for a combined total of 209. Also, 113 drugs that were not included in the original 209 were selected to act as a control group of drugs that were not complex or epilepsy treatments, making the total now 323. The control group was selected randomly from the list of all drug recalls using a Python code.

The following procedure was used to determine the formulation and manufacturing process. First, the FDA labels for the specific drugs and manufacturers were found. No data for other manufacturers were added, as the focus is on recalled products. The active and inactive ingredients information was taken from these labels. If a label could not be found, the information was taken from DailyMed or RxList. Using SciFinder and Google Patents, patents were found matching each drug name and all inactive ingredients. Patents from the specific manufacturer could often not be found. Generic manufacturers typically do not hold patents, and thus finding patents for generic products was not manageable. Still, if a patent utilized the same inactive ingredients as the drug product, it was assumed the process and formulation of the drug were the same as the marketed product. The basis of this assumption is supported by the best mode requirement of a patent, which requires that the patent disclose the best method for the invention. Using this method, 83 matching patents were found for complex drugs, 27 for epilepsy drugs, and 23 for the control group, for a total of 113 patents found.

The patent information was condensed into a numbered list of steps. From this list of steps, the equipment used was cataloged, and “hard steps” defined as stated in the introduction. These hard steps are “applied to backing,” “formed emulsion,” “micronized,” and “formed a gel matrix.” Conversely, steps like “mixed,” “stirred,” and “dissolved” were not considered to be hard steps. Based on the number of total steps, hard steps, and excipients, a manufacturing “rank” was assigned to every drug. Drugs could either have a high rank, meaning a harder manufacturing process, or a low rank, meaning an easier manufacturing process.

From each patent, the equipment used was extracted and listed. This study focused on critical pieces of equipment: blender, sieve, mixer, fluidized bed dryer, and filters. Granulation was considered for tablets, while for transdermal drugs, the techniques of spray coating and knife coating/laminating were studied. Any steps involving heating or freezing were also studied.

3.2.2 Statistical Analysis.

Once this information was gathered from all patents, statistical analysis was done using RStudio to determine the significant factors that affected drug recalls. The recall reasons were the dependent variable, listed here:

- Contamination
- Current Good Manufacturing Practices (CGMP)
- Defective Delivery
- Dissolution Specifications
- Failed Specifications
- Failed Tablet Specifications
- Foreign Substance
- Impurities and Degradation
- Presence of Particulate Matter
- Resuspension Problems
- Stability
- Sterility
- Subpotency
- Superpotency
- Other

The independent variables studied for their effect on the above recall reasons were:

- Category (Complex or Epilepsy)
- Number of excipients
- Presence of a specific excipient
- Number of manufacturing steps used
- Specific manufacturing step used

The independent variables were coded as binary values, with one representing yes, this step/excipient was contained within this product, and 0 representing no, this step/excipient was not contained within this product. In the case of the number of steps, excipients, and hard steps, continuous data were used for the explanatory variable. The dependent variable of recall reason was also coded as a binary value. The tests done used a logistic regression model adjusted with the Firth bias correction method. A significance level of 0.05 was used.

Significant excipients and equipment from the statistical analysis were taken and parsed into an Excel file. Rather than work with each recall reason individually, the decision was made to group all significant factors due to the application of the score. Knowing the score for a drug's likelihood to be recalled for every recall reason is not as necessary as knowing it will be recalled. A company will not be aware if the recall will be for impurities and degradation or dissolution, so one score was created on the basis that any drug recall is not wanted.

Based on the statistical coefficient, all factors were given a "weight" of 1 or -1. A weight of 1 means this factor had a significant effect on preventing drug recall and was assumed to be a high-quality excipient/manufacturing step. A weight of -1 means this factor had a significant effect on the drug being recalled and was assumed to be a low-quality excipient/manufacturing step.

Based on the p-value of the factor, a ranking system was derived for the steps that are very significant to significant. This was done to maintain the power of the p-value rather than grouping all values.

- P-value below 0.05: 1
- P-value below 0.025: 2
- P-value below 0.01: 3
- P-value below 0.001: 4
- P-value below 0.0001: 5

To calculate the score of a factor, the weight was multiplied by the p-value ranking. If a factor was significant multiple times, the weights multiplied by the p-values were added together. For example, Citric Acid was significant three times for various recall reasons. Two were negative, one was positive, and the p-values ranged from rank 4 to 2. The scores were added together for the final score of the excipient. Note that they were not averaged to maintain the power of each ranking value.

After creating a score for every significant factor, a total score was calculated for each drug based on its excipients and manufacturing steps. Any step that was not significant weights 0, so they do not affect the score. R code was written to calculate the scores for every drug within the database.

3.3 Case Study 2: FAERs Data

Case studies were done on drug products that had a high number of recalls. This was to ensure there was enough data to conclude the FAERs reports and recalls. FAERs data was downloaded for each drug from the SafetyRx Database, which includes adverse enforcement reports up until December 2017. Only adverse event reports where the drug product was listed as the Primary Suspect drug were used.

Adverse event reports that specified “off label use” or “withdrawal syndrome”/“drug omission” were removed using a Python code, as these adverse events do not speak to a problem with the drug product. The first study done was characterizing the data using R, recording the frequency of each manufacturer, report source, outcome given, and outcome category. The outcome category was determined by the most serious outcome listed on the report. The data was also sorted by date, recording the frequency of adverse events by manufacturer each month.

Statistical tests were then done to attempt to correlate the adverse event frequency to a drug being recalled in a specific month. Both FDA Date and Event Date were tested. Tests were done using R to determine the significant factor correlating to a drug being recalled. The recall was listed as the dependent variable in binary code (1 = yes, 0 = no). The independent variables studied were:

- Total Frequency of Reports in the month recalled
- Total Frequency of Reports in the month before a recall
- Total Frequency of Reports two months before a recall
- Frequency of Reports from each Source: Physician, Pharmacist, RN, Health Professional, Lawyer, Consumer

Drug recalls from June 2012 to December 2017 were downloaded from the FDA enforcement report program, using December 2017 as the cut-off to match the FAERs data. All recalls for labeling errors, marketed without approval, and packaging issues were removed because these were not formulation quality recalls. Repeat drugs with the same Event ID were removed to avoid counting data twice.

3.4 New Bioinformatics Method

3.4.1 Testing Programs

To construct a new bioinformatics method, the first step was deciding which programs to include. To test programs, five known function proteins were selected from the bacteriophage genome AFIS. AFIS was chosen due to it being in Cluster A1, the most common cluster. Five proteins were chosen due to their prevalence in every phage genome: major tail protein (Structural protein), terminase small subunit (DNA Replication/Translation), portal protein (Structural), HNH endonuclease (DNA Replication/Translation), DNA Polymerase (DNA Replication/Translation), and one hypothetical protein (NKF).

Using these protein sequences, the following programs were tested to determine if they would yield informative results on phage proteins:

- BlastP (NCBI)
- HHPred
- Conserved Domain Database (on NCBI)
- HHBlits
- HMMER
- InterProScan
- TMHMM
- PSI-BLAST (NCBI)

Of these programs, BlastP and HHPred are the current programs used for phage genome annotation. BlastP yielded informative results for all proteins, including the NKF protein. HHPred was informative for 3/6 proteins; CDD was informative for 2/6; HHBlits was informative for 6/6; HMMER was informative for 6/6; InterProScan was informative for 3/6; TMHMM was informative for 6/6; PSI-BLAST was informative for 6/6.

To move forwards with testing hypothetical proteins, the programs chosen were: BlastP, HHPred, HHBlits, HMMER, InterProScan, TMHMM, and PSI-BLAST. CDD was not included because it is one of the databases searched during the HHPred scan and HHPred was a more informative algorithm.

3.4.2 Test Case: PotatoSplit

After using the genome AFIS to test programs due to AFIS being in the most common phage Cluster, a different phage was chosen to continue with the hypothetical analysis. PotatoSplit was chosen due to it being in Cluster A2, a cluster known to infect *M. Tuberculosis*. It is also closely related to AFIS, which was used to test programs. To begin, PotatoSplit had a total of 94 genes. Of these, 57 were hypothetical proteins and 33 had known functions. The 57 NKF genes were used in this test case. Figure 3.1 below shows the process followed in this work.

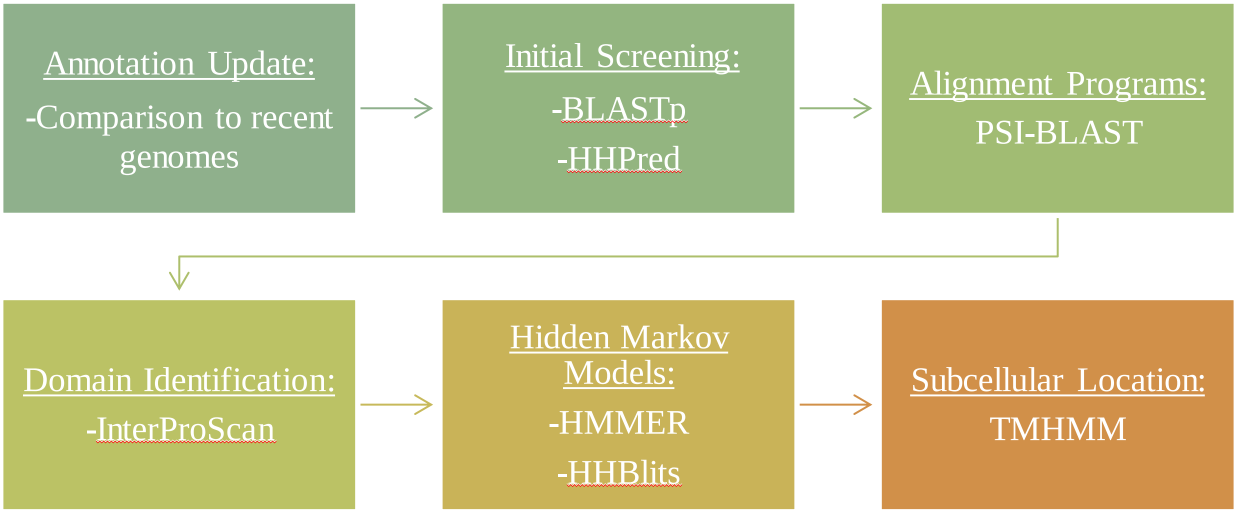


Figure 3.1. Annotation process used for the 57 hypothetical proteins within PotatoSplit, beginning with an Annotation Update, then covering an Initial Screening, Alignment Programs, Domain Identification, Hidden Markov Models, and Subcellular Location prediction.

To begin, an "Annotation Update" was added as the first step to adjust for genomes annotated more than one year ago. PotatoSplit was annotated in 2018, and this comparison to newer genomes allowed for a quality check. For the Annotation Update, the genome was locally blasted on PhagesDB to find similar genomes that were annotated more recently. The genome files were then loaded into an Excel spreadsheet that allowed comparisons across multiple genomes. Six similar genomes were used: Fernando, Sabinator, Penny1, MoneyMay, Beaurxregard13, and JenCasNa. Comparing to these files, any discrepancies in functional assignment were noted and recorded.

After the annotation update, the Initial Screening step uses the current programs in phage annotation: BlastP on the Non-Redundant Protein (nR) database and HHPred on databases pFam, CDD, and PDB. Next, the alignment program PSI-BLAST was run for three iterations, also on nR database to determine if any hits were missed by a basic local alignment. Domains were searched using InterProScan, which searches on CATH, CDD, HAMAP, PANTHER, pFAM, PIRSF, PRINTS, PROSITE, SFLD, SMART, SUPERFAMILY, and TIGRFAMs. Then Hidden Markov Models HMMER and HHBlits were searched. HMMER runs on nR and HHBlits runs on UnitPro. After this, TMHMM was run.

To analyze results, e-values were used. A high cut-off of 5 was used based on some databases not containing many phage proteins, thus inflating the e-values with proteins from other organisms. Any e-value below 5 was considered not informative and any probability below 80%.

To determine function, if a protein had an Annotation Update match and an NCBI BlastP result with an e-value below 10^{-4}, this function was automatically assumed to be correct. If it did not have results that informative, the results from other programs were considered. If all programs matched with informative hits on a specific function, that function was assumed to be correct so long as they did all match to the same database hit. Because the programs sometimes overlapped on databases, if every program returned the same informative hit, this function was not considered. The results had to come from various sources and be informative in their program for the function to be considered. If they did not match on a specific function but did match on functional classification, that classification was listed without a specific function assigned.

3.4.3 I-TASSER & Simulations

I-TASSER is a structural prediction program that matches crystal structures within Protein Data Bank (PDB)[41]. This program was downloaded to a computing network and ran locally. However, the process took a few days at best and was unrealistic to run for every protein. A subset of proteins from five phages within the B1 Cluster was chosen to predict structures for and run GROMACs simulations. Cluster B1 was chosen to due to its lytic life cycle being possibly significant for phage therapy.

The proteins chosen were hypothetical proteins within known cassettes in the phage genome. The first protein was chosen from the Lysis Cassette of B1 phage near genes 49/50/51 and the second from the Replication Cassette near genes 59/60/61. Conserved sequences were determined by analyzing the genomic sequence of each gene and taking the amino acid used in most genomes. The structure was predicted based on the conserved sequence.

Simulations were carried out using GROMACs software[101]. GROMACs were used to simulate each protein in a box of water at two sets of temperatures and pressures. The values picked were based on the highest temperature in an autoclave and the ideal temperature for growing *M. smegmatis*. These temperatures are 394.1 K, hereafter referred to as Autoclave, and 310.15, hereafter referred to as Cell. The accompanying pressures are 1.03421 bar and 1.01325 bar.

The process of the simulations began with the coordinate files produced by I-TASSER. With these files, a GROMACs function was used to create a box with the protein placed in the center at least 1.0 nm from the edge. The size of the boxes varies for each protein based on the protein size. The box was then solvated and filled with water molecules. To use this software, the net charge of the system had to be neutral. Ions were added to achieve a net charge of 0. Na+ ions were added to raise charge and Cl- ions were added to lower charge.

After a box was created and solvated, the energy was minimized to ensure the structure had no inappropriate geometry or steric clashes. If the final energy was negative and stable, the simulation was continued. The step size was 0.01, with a maximum number of steps at 50,000. After energy minimization, the system was equilibrated.

The first equilibration step used was the NVT ensemble, which holds constant the Number of Particles, Volume, and Temperature. The temperature coupling method used was a Berendsen thermostat with a heat bath at the temperature for Autoclave or Cell. The Particle Mesh Ewald method was used for electrostatics, and the Verlet cutoff scheme was used for buffered neighbor searching. The cut-off value for the radius was adjusted for each box size. This was run for 100 picoseconds for each protein, then the temperature was graphed to ensure it had reached a plateau. If it had not, this step was run for another 100 picoseconds with velocity generation turned off.

The second equilibration used was NPT, which holds constant the Number of Particles, Pressure, and Temperature. Again, the Berendsen thermostat was used, along with PME and Verlet. The barostat used was Parrinello-Rahman, and velocity generation was again off. This was run for 100 picoseconds, then the average pressure was checked to make sure the system was equilibrated properly. If it was not within a close range of the desired pressure, this step was repeated for another 100 picoseconds.

Once the system was equilibrated, a production run of the simulation was run for one nanosecond and data collected. If the protein was too large to run for 1 nanosecond, it was run for 0.5 ns or 0.1 ns. The trajectory files were loaded into PyMol[102] to capture images and videos of the simulations, and the root mean square distance data was used to examine how the proteins changed throughout the simulation. The Root Mean Square Distance (RMSD) was compared with the backbone of the molecule for the equilibrated structure after the NPT step and the crystal structure.

3.4.4 Machine Learning

A machine learning algorithm was written in Python. To begin, the Phages Database was downloaded as an SQL file and converted to SQLite to be read in Python. The database was read into Python and relevant sections were extracted and saved as a Database. The data was filtered to include only finalized genome files and only proteins on known functions.

The data was converted into a Pandas data frame. The amino acid sequences were converted into numerical values for each amino acid. These numbers were coded as binary values representing the properties of an amino acid in the following order: Small – Tiny – Negative – Positive – Polar- Aliphatic – Aromatic – Hydrophobic. "1" represents the amino acid having the quality and "0" represents the amino acid not having that quality. For example, amino acid A is "11000001" because it is small, is tiny, is not negative, positive, polar, aliphatic, or aromatic, but is hydrophobic. In this way, similar amino acids will have similar values. The median number of amino acids was used to gauge the number of amino acids to include as features, typically 350. Shorter amino acids were padded with "X", or "0000000" and longer sequences were shortened.

The function labels also had to be parsed into a workable format. The official function list from PhagesDB was downloaded as a CSV file and each function was assigned a number. "Fuzzy" string matching was used to determine which function names were the same despite being listed differently. For example, "LysB" is "Lysin B". Some function names from older phages are no longer used and were completely removed from the dataset; for example, "Pnk", "RDF protein", and "non-heme haloperoxidase". Roughly 80 functions were removed from the dataset.

This is structured data and could not be split randomly due to each protein sequence being within the data multiple times. To split the data, first genes were grouped into sets of similar sequences using BLAST analysis (phams). Each pham corresponds to a group of similar genes, typically with the same function.

To split into testing and training data sets, phams were split randomly between testing and training data sets. Before any pham is placed in the testing/training data, it is checked to see if the pham is already within a dataset. This will avoid repeating phams that may have multiple functions within them.

The algorithm used was a Random Forest algorithm. Features to be used as inputs into the model can be changed within the code. Features were tested for how they affect accuracy. Number of amino acids was tested for affect on the model: 300, 400, 500, and 600. Start site, stop site, and length were also tested for their affect. The number of trees to be used was also tested: 500, 1000, 2000, and 2500. The model was evaluated using mean weighted accuracy and mean weighted precision.

After the best version of the model is decided, it can be ran on any specific phage genome. The phage genome PotatoSplit is ran here to showcase the results. Outputted by the model is the existing function label, the gene number, the predicted function, and the model's confidence in that prediction. Predictions that match the existing function label are tagged with "#" and those with confidence levels above 90% are tagged with "**".

4. RESULTS

4.1 Phage Database & Statistical Analysis

Within the Purdue Phage Database, there is a total of 3,814 genes. Of these, 69% are hypothetical proteins, or No Known Function (NKF). Figure 4.1 below shows the breakdown of these genes split into functional categories.

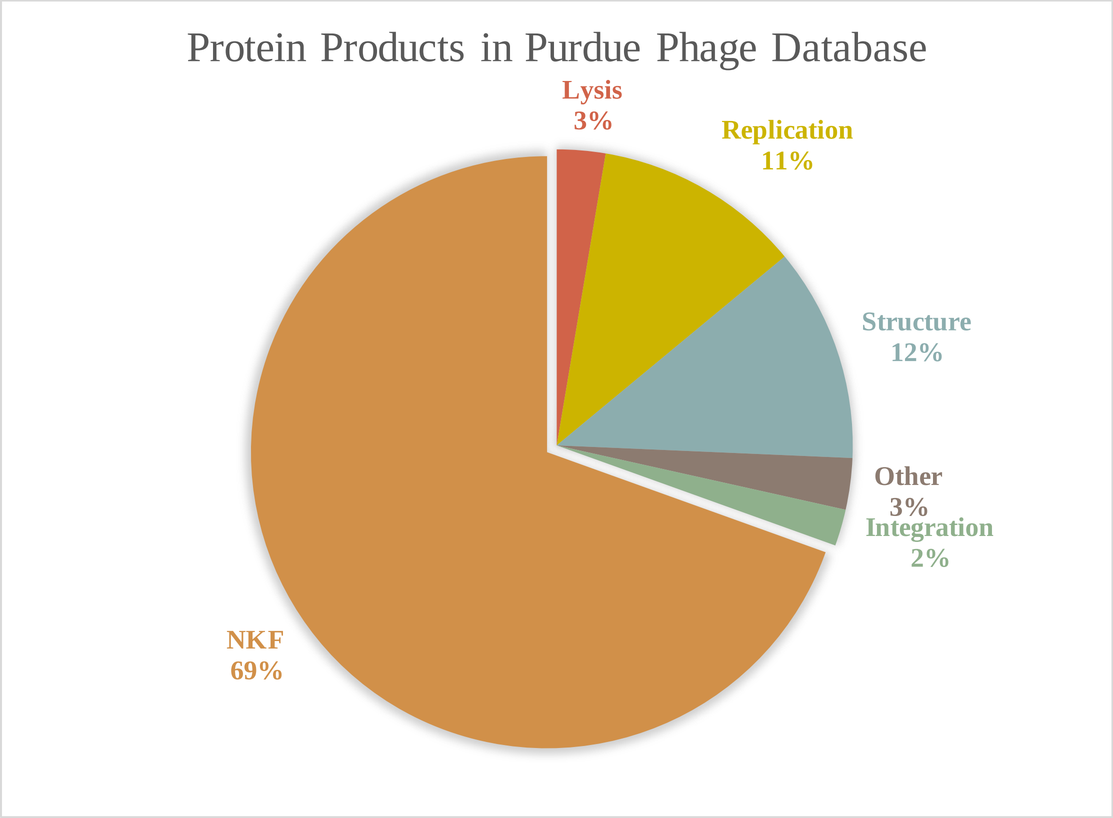


Figure 4.1. The breakdown of functional categories within the genes of the 31 phages included in the Purdue Phage Database.

Of the 3,814 genes studied, 69% had No Known Function. Proteins involved in the physical Structure of bacteriophages made up 12%; those involved in DNA Replication/Translation made up 11%. The rest were 3% involved in Lysis or the breakdown of cells, 2% were involved in DNA Integration, and 3% of proteins had multiple functions and are labeled Other.

The proteins involved in lysis and DNA integration are hypothesized to be important for phage therapies[9,10,77]. DNA integration proteins are necessary for the infection cycle, as it is the first step in producing phage particles and lysing a host, and proteins involved in the lysogenic cycle need to be identified and removed so the phages do not lie dormant rather than killing the host. Yet, without knowing the function of the other 69% of proteins, it is hard to determine those proteins most important in the infection cycle. Even if the known proteins involved in DNA integration are removed using engineering, there could be others that are still unknown.

Figure 4.2 shows the proteins found in every phage and how many were found. Due to the continued active study of bacteriophages, every year there are new standards of genome annotations. These phage genomes were published in GenBank after passing quality checks, but every year the quality of annotations rises. Because of this, some of the older phage genomes have gaps in their proteins that may be due to changes in annotation knowledge. Guidelines are set by the Howard Hughes Medical Institute and updated every year. For example, MrGordo, annotated in 2011, does not have a Lysin A protein that is now required to be in every single phage genome. VasuNzinga does not have a tape measure protein; Zalkecks does not have a head-to-tail adaptor protein, and EricMillard does not have a major capsid protein.

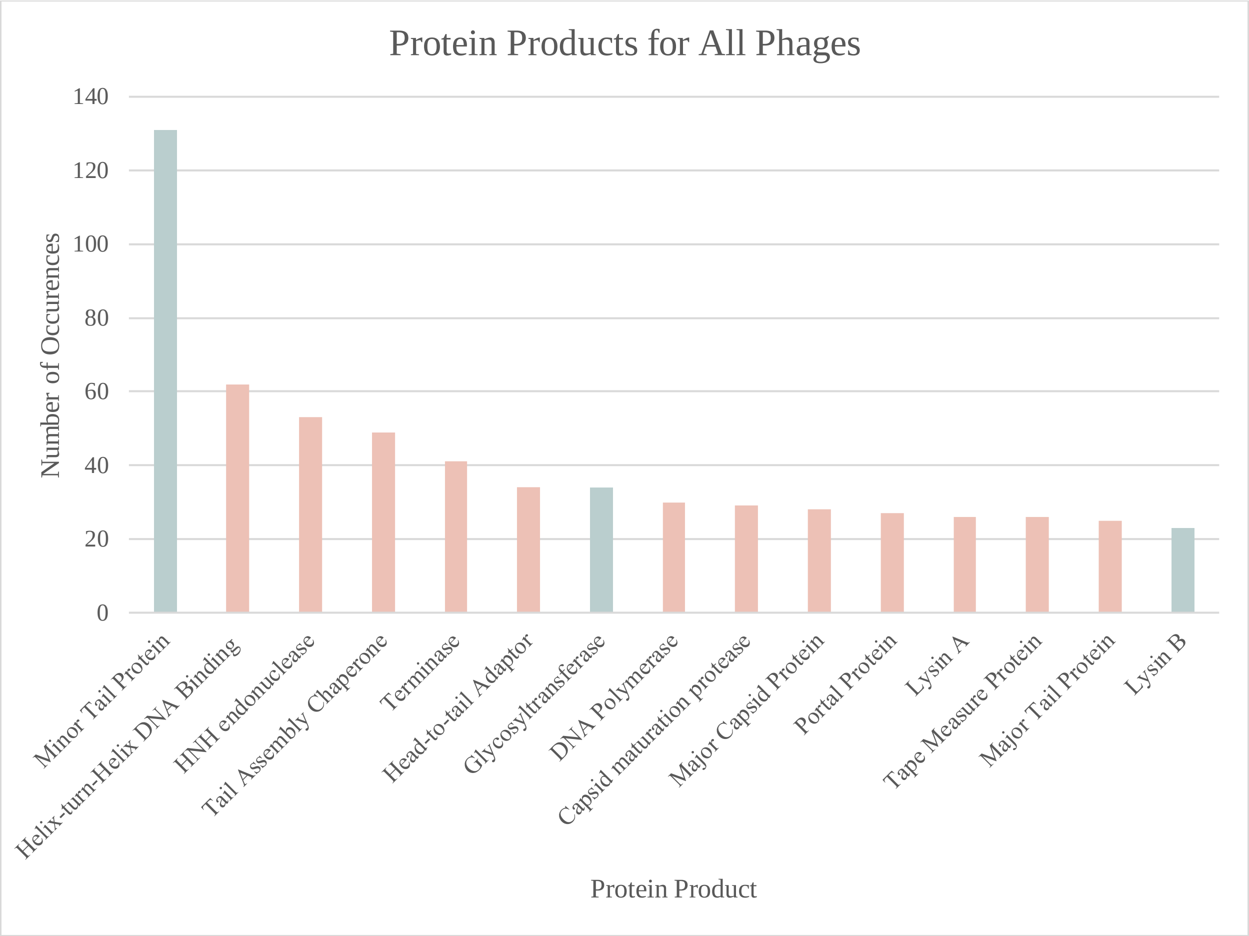


Figure 4.2. The top 15 protein products found in all genomes within the Purdue Phage Database. Those in gray are not found in every genome but occur in a large number within genomes. Those in pink are found within every phage genome.

The functions of these proteins within phage can be seen in Table 4.1. Many of these proteins are required to be within the phage genome following guidelines set by HHMI. The majority of these common proteins are required in phage structure, such as the building blocks for the phage tail or capsid.

Table 4.1. The functions of the proteins found most often within phage, including their category and specific function. Most of these proteins are required by HHMI's SEA-PHAGES program to be within every phage genome.

Protein	Classification	Function
Minor Tail Protein	Structure	Component of the phage tail[103]
Helix-Turn-Helix DNA Binding Protein	DNA Replication	Recognizes DNA protein and binds to stabilize proteins during DNA replication[104]
HNH Endonuclease IV	DNA replication	Used in DNA packing[105]
Tail Assembly Chaperone	Structure	Involved in the production of the phage tail[103]
Head-to-Tail Adaptor	Structure	Component of tail attachment to the phage capsid[103]
Glycosyltransferase	Other	Many functions, ranging from restriction modification to biosynthetic processes to energy utilization[106]
DNA Polymerase	DNA Replication	Synthesizes DNA during replication[107]
Capsid Maturation Protease	Structure	Digests the scaffold proteins from the capsid head after the protein is built[103]
Major Capsid Protein	Structure	Component of the phage capsid[103]
Portal Protein	Structure	Component of tail attachment to the phage capsid, DNA passes through portal protein before entering the bacteria[103]
Lysin A	Lysis	Involved in the breakdown of the bacteria cell membrane[108]
Tape Measure Protein	Structure	Guides the production of the phage tail by acting as a measure for the length of the phage tail[103]
Major Tail Protein	Structure	Component of the phage tail[103]
Lysin B	Lysis	Involved in the breakdown of the bacteria cell membrane[108]

After analyzing protein products in a qualitative sense, statistical analysis was run, beginning with determining which proteins are significant to the lytic life cycle. Table 4.2 below shows the significant proteins. The temperate phages have the same significant proteins, but with the opposite associations. The temperate phage proteins can be found in Table A-1 of the Appendix.

Proteins with negative estimates are negatively associated with being a lytic phage. For example, not having a terminase small subunit is a significant characteristic of being a lytic phage. On the other hand, having a terminase is a significant characteristic of being a lytic phage.

Table 4.2 Proteins that are significant to lytic phage, sorted by functional category. Those with positive coefficients are significantly present in lytic phage, while those with negative coefficients are significantly absent in lytic phage.

PROTEIN	COEFFICIENT	STD. ERROR	CHI^2	P-VALUE
DNA Replication/Translation				
Terminase small subunit	-3.615	1.5999	11.7222	0.0006
Terminase	3.3718	1.6036	9.8746	0.0016
ClP like protease	-2.3978	1.6733	4.0093	0.0452
DNA Polymerase III Subunit	-2.3978	1.6733	4.0093	0.0452
DNA Primase	-2.6559	1.6442	5.2806	0.0215
DNA helicase	2.8462	1.0962	9.8994	0.0016
DNAb like sDNA helicase	-3.1354	1.6117	8.1998	0.0041
Cas4 Family Endonuclease	-2.8991	1.6247	6.6740	0.0097
Queuine tRNA ribosyltransferase	3.6635	1.6633	10.990	0.0009
Structure				
Capsid maturation protease and MuF like fusion protein	2.5024	1.0937	7.4632	0.0063
Scaffold Protein	-2.6835	1.0544	9.4461	0.0021
Head to Tail Stopper	-3.6703	1.1579	16.067	6.11E-5
Tail Terminator	-4.3838	1.6573	16.925	3.88E-4
Other				
Metallophosphoesterase	-2.8991	1.6247	6.6740	0.0097
O-methyltransferase	2.9856	1.7263	6.2731	0.01225
Adenylate Kinase	3.3294	1.6859	8.2607	0.0035
ParB like nuclease domain protein	2.6092	1.7981	4.2702	0.0387
NrdH like gutaredoxin	-2.8991	1.6247	6.6740	0.0097
DNA Integration				
Integrase	-4.1571	1.6072	16.0853	6.05E-5
RuvC like resolvase	4.0073	1.6538	13.773	0.0002
Immunity Repressor	-4.8933	1.6556	21.813	3.00E-6

There are only a few proteins that have a positive association with being a lytic phage. These are terminase, queuine tRNA ribosyltransferase, DNA helicase, capsid maturation protease and MuF like fusion proteins, O-methyltransferase, adenylate kinase, RuvC like resolvase, and ParB like nuclease domain protein.

No temperate phages except for EricMillard have a DNA helicase, while only temperate phages possess the DNAB-like dsDNA helicase. This is shown in the data by the significance of lytic phage having DNA helicase and not having DnaB-like dsDNA helicase. Temperate phages, and all of those within the A cluster, contain a Cas4 family endonuclease. Temperate phages have those proteins related to DNA integration such as integrase and the immunity repressor. These are not present in the lytic phage.

The next variable analyzed was the proteins significant to clusters known to infect *M. tuberculosis*, shown in Table 4.3. The only protein that has a negative association is the holin protein, with all others being positively significant for the A1 and A3 subclusters. Looking into this data, along with the cluster K phage, is a method for identifying the key proteins needed to have therapeutic value.

Table 4.3. Proteins significant to the A1 and A3 clusters that were shown to be infectious to *M. tuberculosis.*

PROTEIN	**COEFFICIENT**	**STD. ERROR**	**CHI^2**	**P-VALUE**
Lysis				
Holin	-3.1876	1.6099	7.2091	0.0072
DNA Replication/Translation				
DNA primase	4.3489	1.6940	12.7648	3.53E-4
Endonuclease VII	3.1876	1.6099	7.2091	7.25E-3
Cas4 family endonuclease	3.9648	1.6527	10.971	9.25E-4
Structure				
Scaffold Protein	2.7979	1.6032	5.4413	.0196
DNA Integration				
Integrase	2.7979	1.6032	5.4413	0.0196
Immunity Repressor	2.4485	1.6042	7.3773	0.0449

This statistical significance is the first step in examining key differences between lytic and temperate phage and how that could affect phage therapy. More wet lab research needs to be done to examine the association of the proteins and how they affect the infection cycle.

Furthermore, there is an issue of NKF genes in every cluster that needs to be addressed. When pursuing phage therapy, those protein differences between temperate and lytic phages are important in determining how to stop the lysogenic cycle. Identifying those proteins in the temperate phages is necessary, as is identifying the proteins used in lysis in the lytic phages.

4.2 Case Study 1: Pharmaceutical Quality Recalls

The breakdown of the database can be seen in Figure 4.3. There are 323 drugs in the database initially, and 113 drugs with patents found. Complex drugs had the most patents found, representing 63% of all the patents found. Epilepsy drugs were the smallest group, but only four of the 31 did not have a patent found. Epilepsy drugs are 20% of all patents found, and the control group represents the remaining 17%.

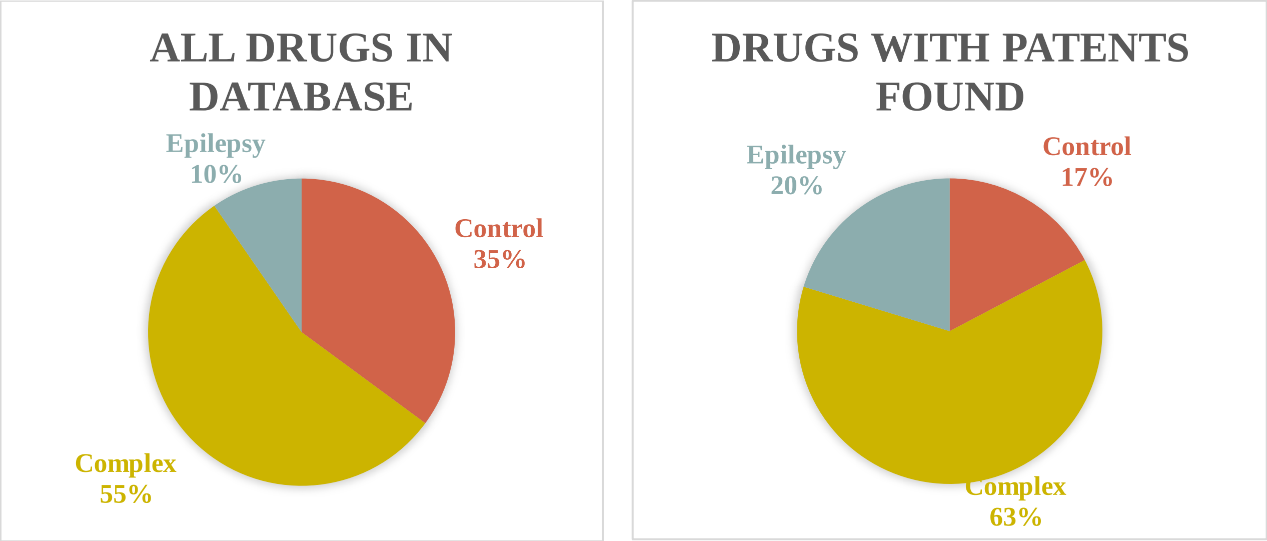


Figure 4.3. The contents of the database between the categories Epilepsy Drugs, Complex Drugs, and Control Group.

Of all the drugs found, the top recall reasons were: Defective Delivery, Impurities/Degradation, Dissolution Specifications, Sterility, and Presence of Particulate Matter. Figure 4.4 shows the percent of each recall reason within the database. The top reasons were studied for any leading indicators in this data.

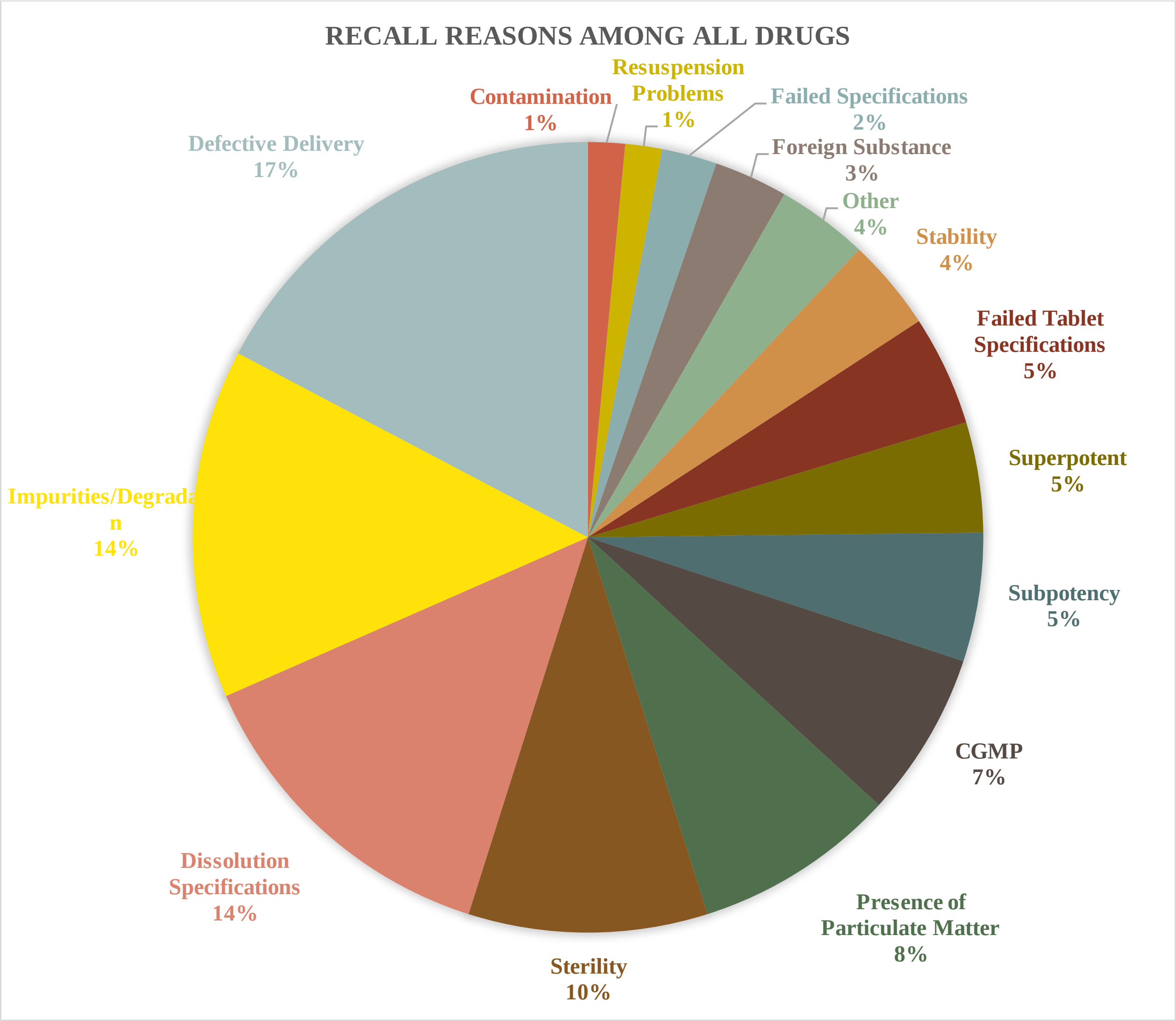


Figure 4.4. The recall reasons among the database of 323 drugs. The top reasons are further studied for leading indicators.

Table 4.4 shows the significance of drug type (complex, epilepsy, or control) on drug recall reasons. Only two recall reasons, defective delivery, and stability showed any significance. Being a complex drug has a positive association, while epilepsy drugs have a negative association. Complex drugs are more likely to be recalled for defective delivery, whereas epilepsy drugs are less likely to be recalled. Being a complex drug has a significant positive effect on being recalled for stability as well.

Table 4.4. The effect of drug type (complex or epilepsy) on drug recall reason. Being a complex drug has a significant positive effect on being recalled for Defective Delivery.

Defective Delivery		
Drug Type	**Coefficient**	**P-value**
Complex	3.055	2.216E-8
Epilepsy	-2.7394	0.0032
Stability		
Drug Type	**Coefficient**	**P-value**
Complex	1.0123	0.0496

Studying the excipients listed for the 323 drugs, Figure 4.5 shows the most common excipients. These were studied for any significance with the recall reasons. The most common excipient was water, appearing in 95 different drug formulations. It is followed by magnesium stearate at 63 drugs, microcrystalline cellulose at 47 drugs, citric acid at 43 drugs, and alcohol at 42 drugs.

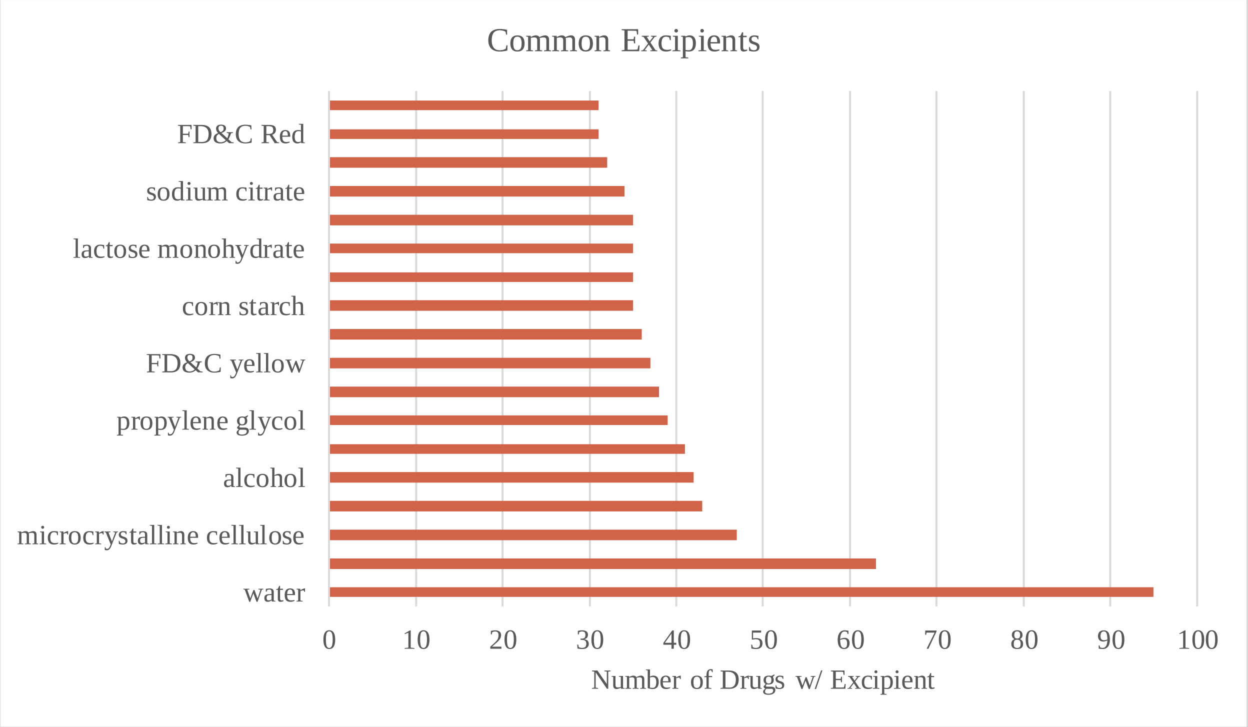


Figure 4.5. The common excipients of the 323 drugs in the database.

Using the most common excipients found, significance among the 178 complex drugs was tested for the top recall reasons. Results are shown in Table 4.5, showing every significant excipient found. There were no significant excipients for the recall reasons Sterility or CGMP, which is expected because the ingredients would not play a role in a product remaining sterile or in the manufacturing practices of a company. The excipients alcohol, silicon dioxide, and titanium dioxide were never significant.

Table 4.5. The effect of excipients on recall reason for 178 complex drugs. Positive coefficients reflect a positive correlation, while negative coefficients show a negative correlation.

Impurities and Degradation		
Excipient	Coefficient	P-value
Magnesium Stearate	1.2719	0.0361
Corn Starch	1.1935	0.0470
Sodium Hydroxide	-2.0657	0.0464
Sodium Citrate	1.0853	0.0388
Defective Delivery		
Excipient	Coefficient	P-value
Microcrystalline Cellulose	-2.1729	0.0332
Citric Acid	-2.7493	0.0024
Water	-1.9619	0.0657
Magnesium Stearate	-2.2362	0.0026
Corn Starch	-2.2966	0.0211
Glycerin	-2.6598	0.0040
Sodium Benzoate	-2.5151	0.0082
Sucrose	-2.6598	0.0040
Stability		
Excipient	Coefficient	P-value
Citric Acid	1.2497	0.0417
Superpotent		
Excipient	Coefficient	P-value
Microcrystalline Cellulose	1.6039	0.0373
Citric Acid	1.4068	0.0412
Glycerin	1.5057	0.0303
Sodium Benzoate	2.0962	0.0022
Presence of Particulate Matter		
Excipient	Coefficient	P-value
Sodium Hydroxide	4.6282	7.72E-12
Subpotent		
Excipient	Coefficient	P-value
Propylene Glycol	1.8633	0.0023
Glycerin	1.6845	0.0052
Dissolution Specifications		
Excipient	Coefficient	P-value
Microcrystalline Cellulose	2.6042	9.27E-5
Magnesium Stearate	1.7170	0.0141
Propylene Glycol	1.7726	0.0066
Povidone	2.9144	8.06E-6
FD&C Yellow	1.9271	0.0156
Silicon Dioxide	1.8096	0.0212
Corn Starch	2.4310	0.0002
Lactose Monohydrate	1.5746	0.0223
Sucrose	1.6019	0.0127

Magnesium stearate has a significant positive effect on being recalled for Impurities and Degradation. If a drug contains magnesium stearate, it is more likely to be recalled for Impurities and Degradation. On the other hand, if a drug contains sodium hydroxide, it is less likely to be recalled for Impurities and Degradation, as it shows a significant negative effect.

The number of excipients had a significant effect on defective delivery, dissolution specifications, and presence of particulate matter, shown in Table 4.6. Having more excipients makes a drug less likely to be recalled for defective delivery and the presence of particulate matter. This could be because defective delivery is a recall reason used for transdermal patch drugs, and having more adhesives makes it less likely to be recalled. For the presence of particulate matter, this could be because more excipients lead to a more stable dosage form. Dissolution specifications, however, show a positive association. More excipients make a drug more likely to be recalled for dissolution specifications, perhaps because more excipients can cause problems in the active ingredient's dissolution and suggest the formulation was more difficult.

Table 4.6. The effect of the number of excipients on recall reason. Having more excipients has a negative effect on being recalled for defective delivery and the presence of particulate matter, but a positive effect on being recalled for dissolution.

Recall Reason	Coefficient	P-value
Defective Delivery	-0.2959	0.0001
Dissolution Specifications	0.2805	1.5E-5
Presence of Particulate Matter	-0.2595	0.0111

The manufacturing process itself plays a role in the recall of drugs as well. The number of steps in a process had a significant effect on superpotency, shown in Table 4.7. Having more steps makes a drug more likely to be recalled for superpotency and was not significant for any other recall reason.

Table 4.7. The effect of the number of steps on recall reason. Having more steps has a positive effect on being recalled for superpotency.

Recall Reason	Coefficient	P-value
Superpotency	0.3112	0.0249

The significant specific steps are shown in Table 4.8. The processing steps heating and mixing were never significant, most likely because they were common steps. The recall reasons Subpotency, CGMP, and Impurities/Degradation had no significant steps. Defective delivery had the most significant results, with negative associations with granulation, blender, sieve, and filter steps, and a positive association with the use of a fluidized bed dryer. The steps spray coating and laminating for transdermal products were also significant because these steps are associated with applying drug reservoir and adhesive to transdermal patch drugs.

Table 4.8. The effect of equipment on recall reason.

Defective Delivery		
Equipment	Coefficient	P-value
Filter	-1.174265	0.0341
Sieve	-1.6918	0.0167
Blender	-1.5484	0.0320
Granulation	-3.011	6.711E-4
Fluidized Bed Dryer	1.4395	0.0018
Spray coating	3.8538	1.153E-9
Laminating	1.9583	0.0109
Sterility		
Equipment	Coefficient	P-value
Granulation	-2.3357	0.0214
Superpotency		
Equipment	Coefficient	P-value
Granulation	1.793	0.0245
Presence of Particulate Matter		
Equipment	Coefficient	P-value
Filter	1.8293	0.00358
Granulation	-2.1528	0.0412
Fluidized Bed Dryer	-2.70659	0.0053
Dissolution Specifications		
Equipment	Coefficient	P-value
Fluidized Bed Dry	-1.4851	0.0154

Additionally, the defined hard steps had an association with recall reasons, shown in Table 4.9. Having a high manufacturing rank had a significant effect on the recall reasons for defective delivery and the presence of particulate matter. A high ranking has a positive effect on being recalled for defective delivery, and a negative effect on the presence of particulate matter.

Table 4.9. The effect of high manufacturing rank on recall reason.

Recall Reason	Coefficient	P-value
Defective Delivery	2.5647	1.7E-8
Presence of Particulate Matter	-1.4930	0.0025

Specific hard steps are shown in Table 4.10. Apply to backing shows positive significance for being recalled for Defective Delivery, which is expected because transdermal drugs are recalled for defective delivery. Formed gel matrix showed positive significance for superpotency and subpotency, likely because of the difficulties in topical drugs to maintain the correct dosage. The reason “formed particles” was also significant to superpotency, and formed emulsion was significant to dissolution specifications. These difficult manufacturing steps all had positive effects, further demonstrating that they can be challenging for manufacturers to perfect.

Table 4.10. The effect of hard steps on recall reason.

Defective Delivery		
Step	Coefficient	P-value
Apply to Backing	4.0913	3.44E-15
Superpotency		
Step	Coefficient	P-value
Formed Particles	1.8246	0.0401
Formed Gel Matrix	2.4878	0.0100
Subpotency		
Step	Coefficient	P-value
Formed Gel Matrix	1.9331	0.0343
Dissolution Specifications		
Step	Coefficient	P-value
Form Emulsion	4.3506	1.71E-4

Table 4.11 below shows the top ten highest risk judges based on their significant excipients and manufacturing steps. Also included is the score calculated using p-values and weights. Note that Minivelle is listed three times because it was recalled three separate times for patches failing to stick to skin. The others were recalled for Failed Impurities/Degradation, Superpotent, Failed Specifications, Failed Dissolution Specifications, and Subpotent.

Table 4.11. The ten drug products with the worst quality scores after statistical analysis.

Drug Product	Score
Minivelle (estradiol Transdermal System) 0.1 mg per day, pack of 8 systems per carton, Rx only, Dist. by: Noven Therapeutics, LLC. Miami, Florida 33186. NDC: 68968-6610-8	-21
Minivelle (estradiol transdermal system) 0.1 mg/day, 1 System per pouch (NDC 68968-6610-1), packaged in 8 pouches per box (NDC 68968-6610-8), Rx only, Mfd. by: Noven Pharmaceuticals, Inc., Miami, Florida 33186; Dist. by: Noven Therapeutics, LLC, Miami, Florida 33186.	-21
Minivelle (estradiol transdermal system) Patches Delivers 0.1 mg/day, a) 2 count and b) 8 count boxes, Rx only, , Mfd. by: Noven Pharmaceuticals, Inc. Miami, Florida 33186 Dist. By: Noven Therapeutics, LLC. Miami, Florida 33186 --- NDC 68968-6610-8	-21
Cetirizine HCl Chewable Tablet, 10 mg, 6-tablets in one blister, in 12 (2 blisters) and 24 (4 blisters) tablet count configurations. Manufactured by Sandoz Private Limited Village-Digham Opp. Thane-Belapur Road Navi Mumbai, 400 078, India, for Sandoz Private Limited 100 College Road West, Princeton, NJ 08540. NDC 66394-041-06	-20
ZyGenerics ATENOLOL Tablets, USP 25 mg 1000 count bottle, Rx Only Manufactured by: Cadila Healthcare Ltd. Ahmedabad, India Distributed by: Zydus Pharmaceuticals USA Inc. Pennington, NJ 08634 USA NDC 68382-022-01	-19
CHILDREN'S IBUPROFEN, ORAL SUSPENSION, 200 mg/10mL cup BERRY FLAVOR, ALCOHOL FREE, MFG: ACTAVIS, PGK BY SAFECOR Columbus, OH	-18
Unit Dose Valsartan Tablets, USP, 80 mg. Rx only, Distributed by: Major Pharmaceuticals, Livonia, MI 48152, NDC# 0904-6594-61.	-17
Valsartan Tablets USP 320 mg, 90-count, plastic child resistant bottle, Rx Only, Preferred Pharmaceuticals, Inc., 1250 N. Lakeview Ave., Suite O, Anaheim, CA 92807, NDC 68788-6882-9	-17
Fentanyl Transdermal System, 100 mcg/h, each transdermal system contains: 10 mg fentanyl and 0.4 mL alcohol USP, Rx only, supplied in single pouches (NDC 0591-3214-54 (pouch)), 5 pouches per carton (NDC 0591-3214-72 (Carton)), Manufactured by Watson laboratories Inc., Corona, CA, Distributed by: Watson, Pharma Inc.	-16
Fentanyl Transdermal System. 25 mcg/h, packaged in 5 pouch system cartons (NDC 0591-3198-72), Rx Only, Manufactured by: Actavis Laboratoies UT, Inc. Salt Lake City, UT 84108, Distributed by: Actavis Pharma, Inc. Parsippany, NJ 07054 USA. Individual pouch NDC 0591-3198-54.	-16

Table 4.12 below shows the top ten highest-scoring drugs, meaning the proposed best quality. These were recalled for Chemical Contamination, Empty Capsules, Failed Stability at 12-month mark, Superpotent, Failed Impurities/Degradation, CGMP Deviations: Inadvertent release of a drug product with unapproved active ingredient manufacturer, Discoloration, and Presence of Foreign Matter.

Table 4.12. The ten drug products with the best scores after statistical analysis.

Product	Score
Glenmark Gabapentin Tablets, a) 600mg, 500- count bottle (NDC 68462-126-05), b) 800 mg, 500- count bottle (NDC 68462-127-05), Rx only, Manufactured by Glenmark Generics Ltd.Colvale- Bardez Ltd 403513, India, Manufactured for : Glenmark Generics USA Mahwah, NJ 07430.	8
Gabapentin Capsules, USP, 400 mg, Rx Only, a) 100 capsules per bottle, NDC 14550-513-02, b) 500 Capsules per bottle, NDC 45963-557-50, Manufactured by: Actavis Pharma Manufacturing Pvt. Ltd., Plot No 101, 102, 107, & 108, SIDCO Pharmaceutical Complex, Alathurt, Kanchipuram Dist-603 110, Tamlinadu, India, Distributed by: Actavis Elizabeth LLC, 700 Elmora Ave, Elizabeth, NJ 07207 USA.	6
Alinia (nitazoxanide), powder for oral suspension, 100mg/5mL, 60 mL/bottle. Rx only, Manufactured for Lupin Pharmaceuticals Inc, Baltimore, Maryland 21202 for Romark Laboratories 3000 Bayport Dr. Suite 200, Tampa, FL 33607	6
Trokendi XR (topiramate) extended-release capsule, 50mg, 30-count blister pack, Rx only, Manufactured by: Catalent Pharma Solutions, Winchester, KY 40391, Manufactured for: Supernus Pharmaceuticals, Inc., Rockville, MD 20850, NDC 17772-102-15	6
candesartan cilexetil, tablets, 16 mg, 90-count bottles, Rx only, Manufactured for Sandoz Inc., Princeton, NJ 08540 by Mylan Laboratories Limited Hyderabad, 500 034, India, NDC 0781-5938-92	5
Gabapentin Oral Solution, 250 mg/5 mL (50 mg/mL) in a 470 mL amber-colored bottle, Rx Only. Manufactured by: Hi-Tech Pharmacal Co., Inc. Amityville, NY 11701. NDC: 50383-311-47	4
Suprax (cefixime for oral suspension) USP, 100 mg/5 mL, 50 mL bottles (when reconstituted), Manufactured for Lupin Pharmaceuticals, Inc. 111 South Calvert Street, Baltimore, MD 21202, Manufactured by Lupin Limited Mumbai 400 058 India, NDC 68180-202-03.	3
Suprax, Cefixime for Oral Suspension USP 500 mg/5 ml, 10mL (when reconstituted), Rx only, Manufactured for Lupin Pharmaceuticals, Inc. 111 South Calvert Street, Baltimore, MD 21202, Manufactured by Lupin Limited Mumbai 400 058 India, NDC 27437-207-02.	3
Synjardy (empagliflozin and metformin hydrochloride) Tablets. 5 mg/1000 mg. Rx only. 180-count bottle. Distributed by: Boehinger Ingelheim (BI) Pharmaceuticals, Inc. Ridgefield, CT 06877. Made in Germany. Marketed by: BI Pharmaceuticals, Inc. Ridgefield, CT 06877 and Eli Lilly and Company Indianapolis IN 46285 NDC 0597-0175-18	3
Amoxicillin and Clavulanate Potassium for Oral Suspension, USP, 250/62.5 mg per 5 mL, 100 mL (when reconstituted) bottle, Rx Only, Manufactured By: Cipla Ltd. at Medispray Laboratories Pvt. Ltd., Kundaim Goa, India; Manufactured For: Wockhardt USA, LLC, Parsippany, NJ 07054, NDC 60432-065-00.	2

4.3 Case Study 2: FAERs Data

Daytrana has 11 recalls from Noven Pharmaceuticals between 2011 and 2017. No other manufacturers have recalls for Daytrana. The FAERs data supplied had a total of 10,357 adverse events reported. Of those, 10,141 were for Noven. 10,130 had no event outcome reported, with the other outcome options (Disability, Hospitalization, Life-Threatening, Other Serious Event, Death) having between 1 and 50 event reports. Figure 4.6 shows how many were reported with a date given for when the adverse event happened.

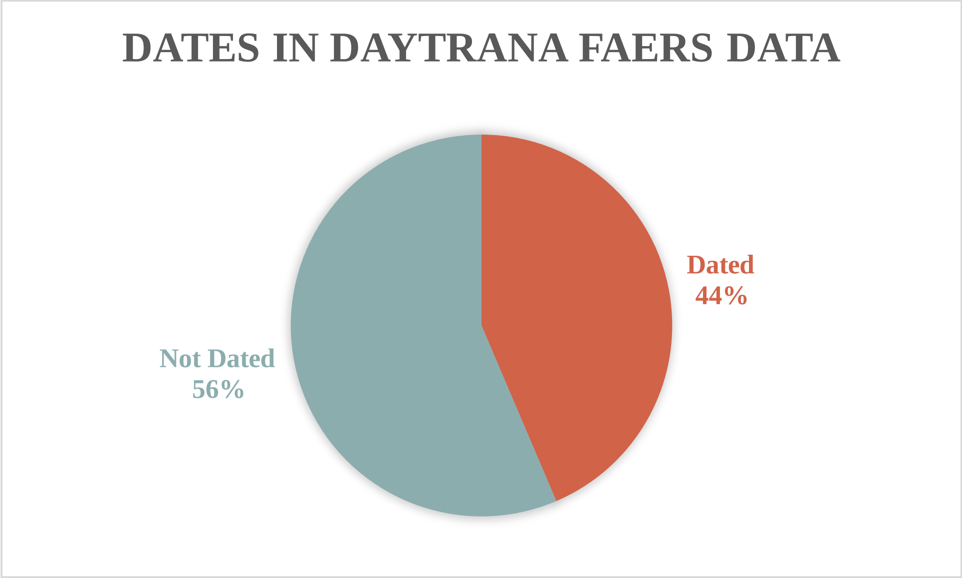


Figure 4.6. Breakdown of how many Datrana AERs were dated or not dated.

Without dates, the data cannot be used for analysis. The 4,512 points with dates were further studied. Figure 4.7 below shows a bar graph featuring the number of adverse reports by month in red and the dates recalled in yellow.

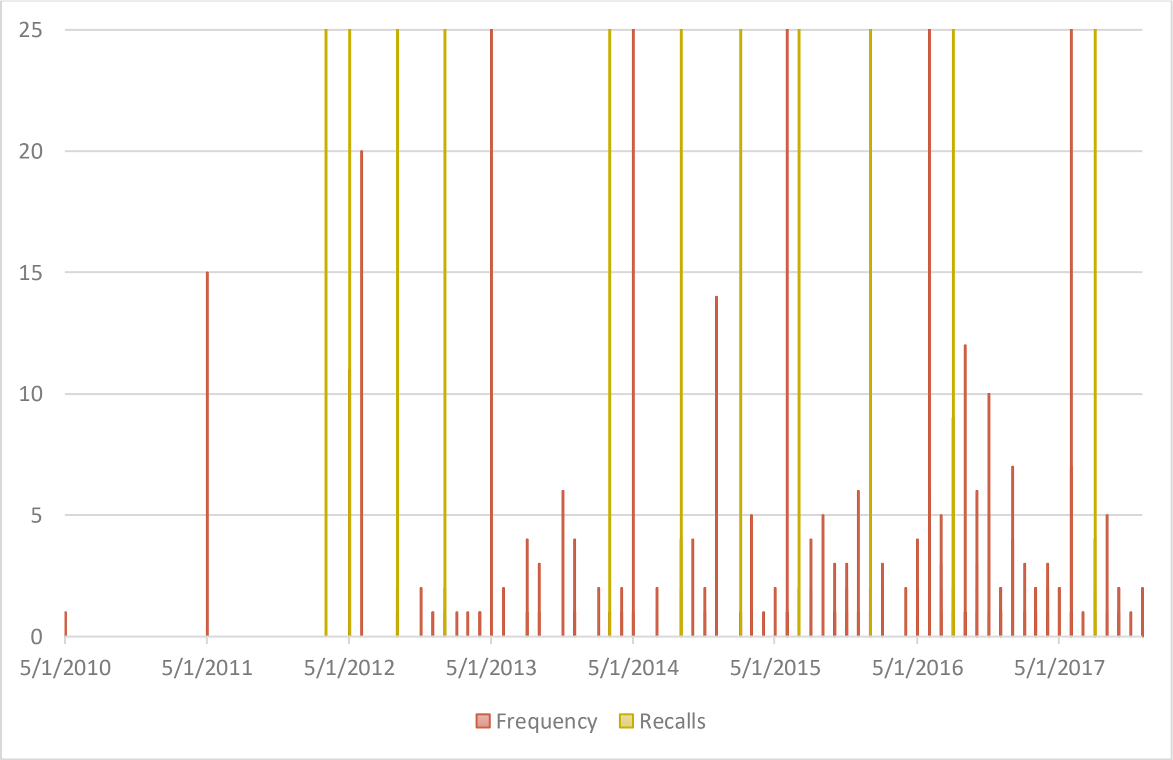


Figure 4.7. Daytrana adverse event reports by month, with recalls shown as yellow bars.

Running statistical analysis, Table 4.13. Shows the significance of frequency of adverse event reports on being recalled for any reason.

Table 4.13. Daytrana adverse event report frequency significance on being recalled for any reason. All time frames tested were significant.

Factor	**Coefficient**	**St. Error**	**Chi^2**	**P-value**
Frequency This Month	0.0137	0.0063	4.4506	0.0348
Frequency Last Month	0.0134	0.0062	4.3482	0.0370
Frequency Two Months Ago	0.0154	0.0062	5.8357	0.0157

These results show that each timeline tested was significant, meaning the number of adverse reports in a certain month, in the month before, or two months before have a positive correlation with the recall then happening. Further analysis on the source of each report showed that only consumer reports were significant. Reports from other sources (listed in Methodology) were not significant, so they could not be used to predict recalls. This could be due to a lawsuit coming against Daytrana in 2015, adding a confounding variable to this study.

The same procedure was tested on Gabapentin and Propofol. Gabapentin had only three recalls ranging from 2014 to 2017, by three different manufacturers: Aurobindo, Actavis, and Hi-Tech Pharmcal Co. The Aurobindo recall was due to empty capsules, and the Hi-Tech Pharmacal Co. recall was due to a Good Manufacturing Practices violation. Actavis failed tablet specifications, so it was tested for any significance in adverse event reports. Gabapentin had a total of 15,710 adverse event reports spanning 179 manufacturer companies show in Figure 4.8, including 595 with no manufacturer reported. There were also 7,408 without dates, leaving a total of 8,302 reports to study.

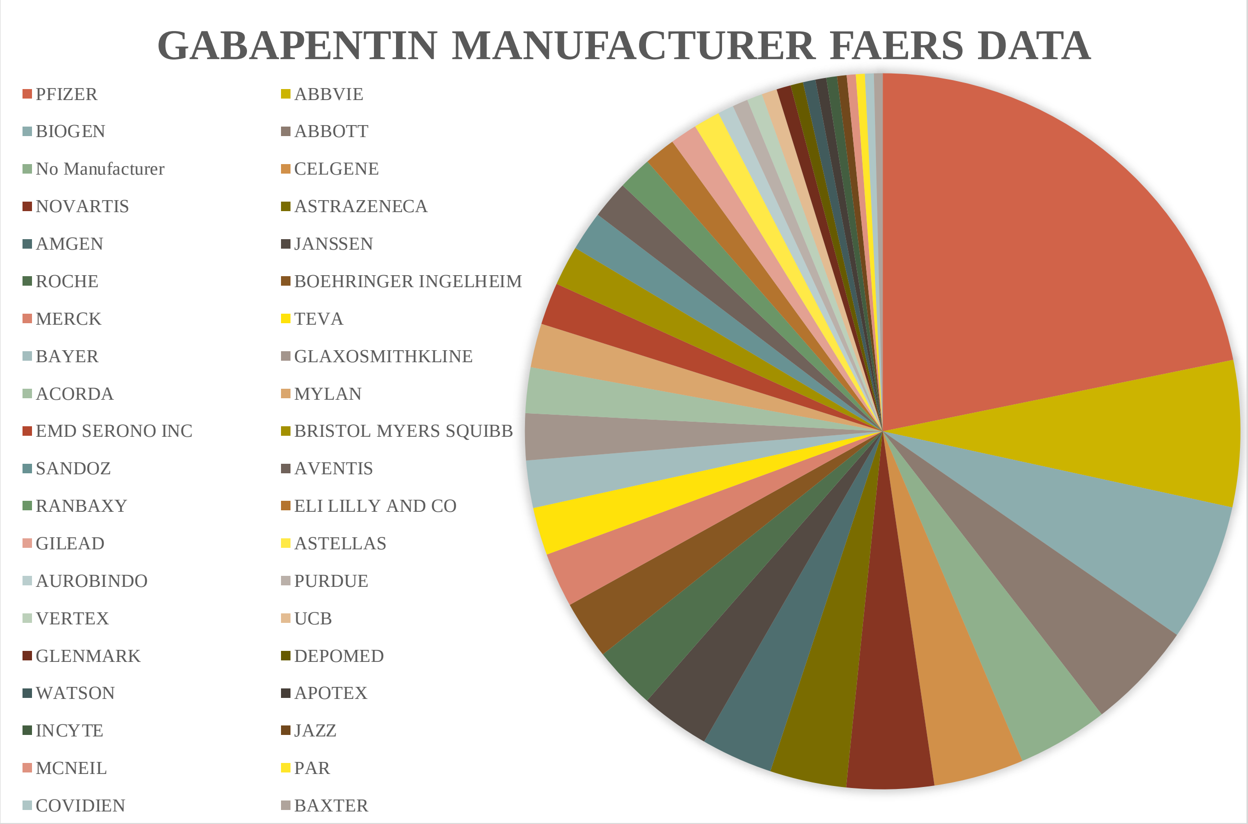


Figure 4.8. Manufacturers of Gabapentin with over 50 adverse event reports. Manufacturers with below 50 reports were left off due to limited spacing.

The statistical analysis of Gabapentin yielded no significant results. Figure 4.9 shows the bar graph of adverse event reports and recalls. In red are reports from all manufacturers, in gray are reports from Actavis, and in yellow is the Actavis recall.

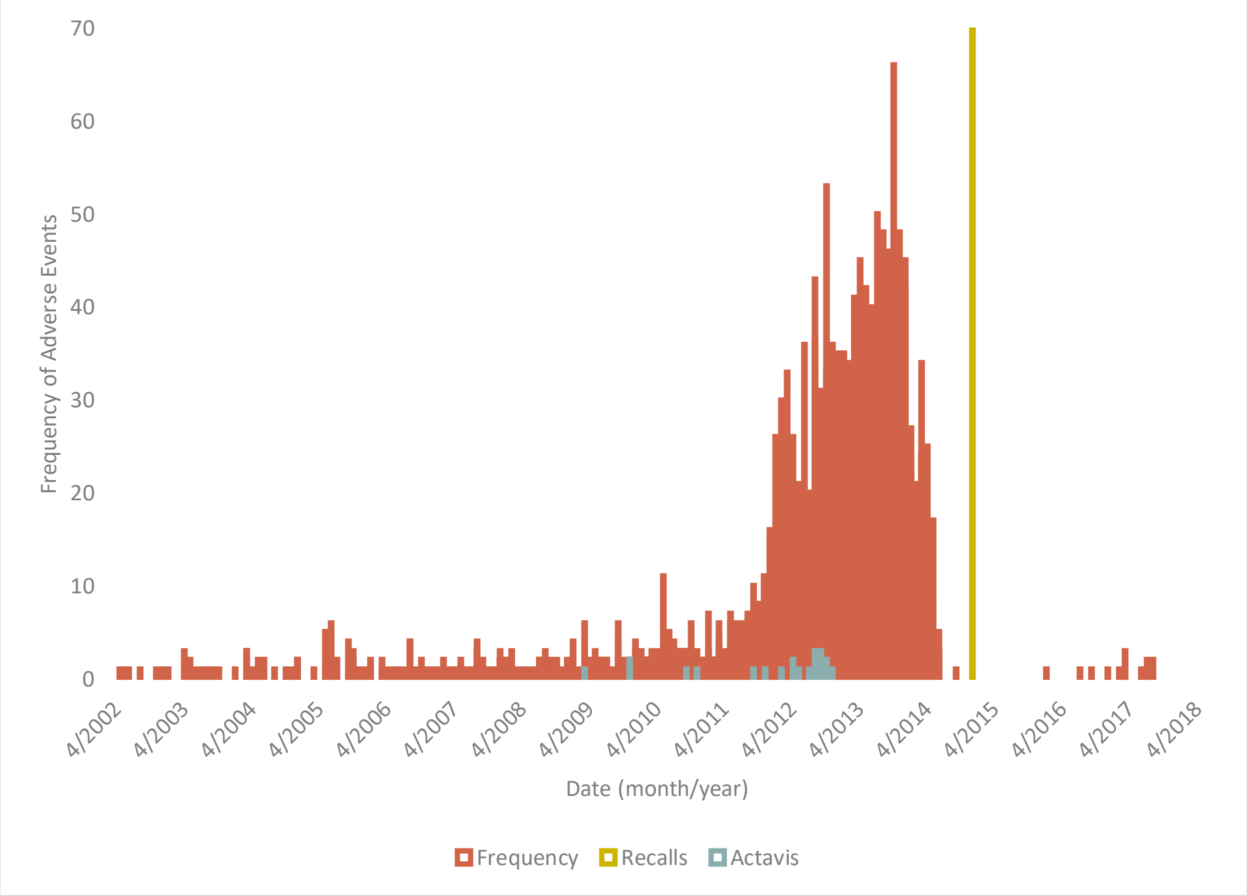


Figure 4.9. Bar graph showing the number of adverse events from 2002 to 2018 for Gabapentin. In orange is the overall number of reports from all manufacturers, with Actavis shown in gray and the recall event shown as a yellow bar.

For Propofol, there were 13 recalls between 2012 and 2016. 11 of them were “presence of particulate matter: visible particles embedded in the glass” and one was for “temperature abuse”. All of them were from Hospira Inc. Because of the lack of drug-quality recalls, the results are not expected to yield results.

4.4 New Bioinformatics Method

4.4.1 Bioinformatics Programs

After going through the new bioinformatics method, the number of NKF proteins within PotatoSplit dropped from 57 to 40. A total of 17 new functions were found. The functional class was identified for an additional six proteins, though no specific functions were named. Figure 4.10 shows the functional breakdown of PotatoSplit before and after the bioinformatics process was used. One new Lysis protein was identified, 5 new DNA Replication/Translation proteins, and 17 Structural proteins. The NKF percentage is no longer the majority of proteins for this phage, which is a large step forwards.

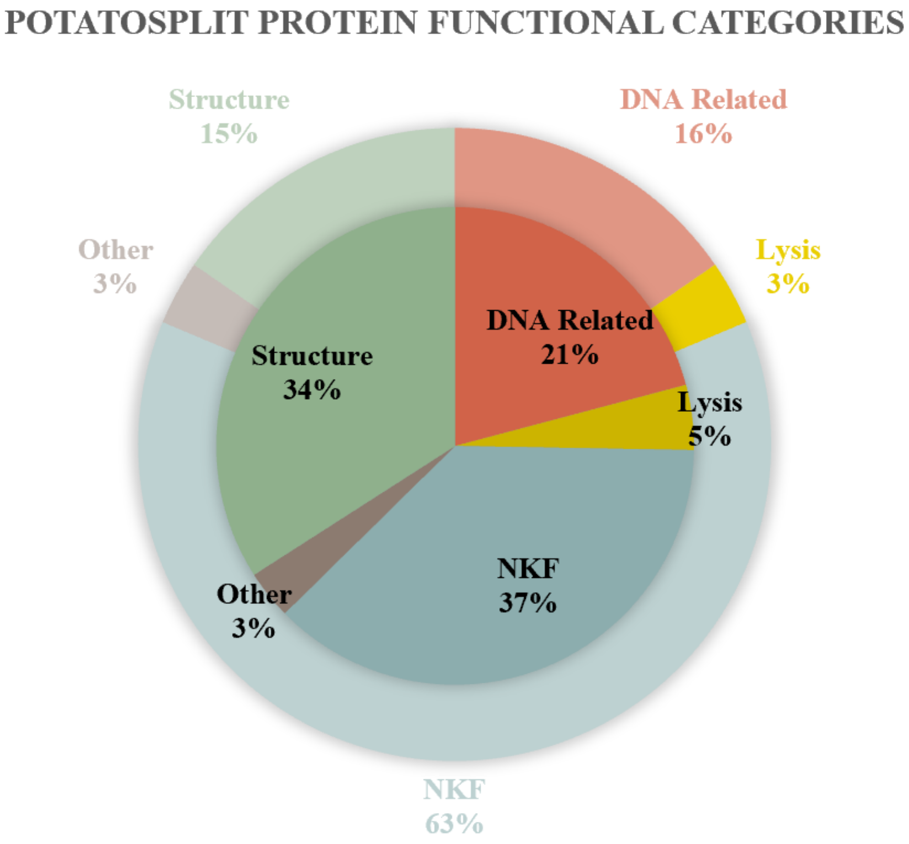


Figure 4.10. A graph showing the functional categories of PotatoSplit proteins before (outer chart) and after (inner chart) going through the new bioinformatics process.

Figure 4.11 below shows the percent informative hits from each program out of the 57 hypothetical proteins identified. The most informative program was HHBlits, with a mean e-value of 1.94 and a median e-value of 0.31. The mean probability for HHBlits results was 91.2%, and the median was 91.5%. The next most informative program was TMHMM, which has a mean probability of 0.50 and a median of 0.80. These programs are recommended to be added to the phage annotation process.

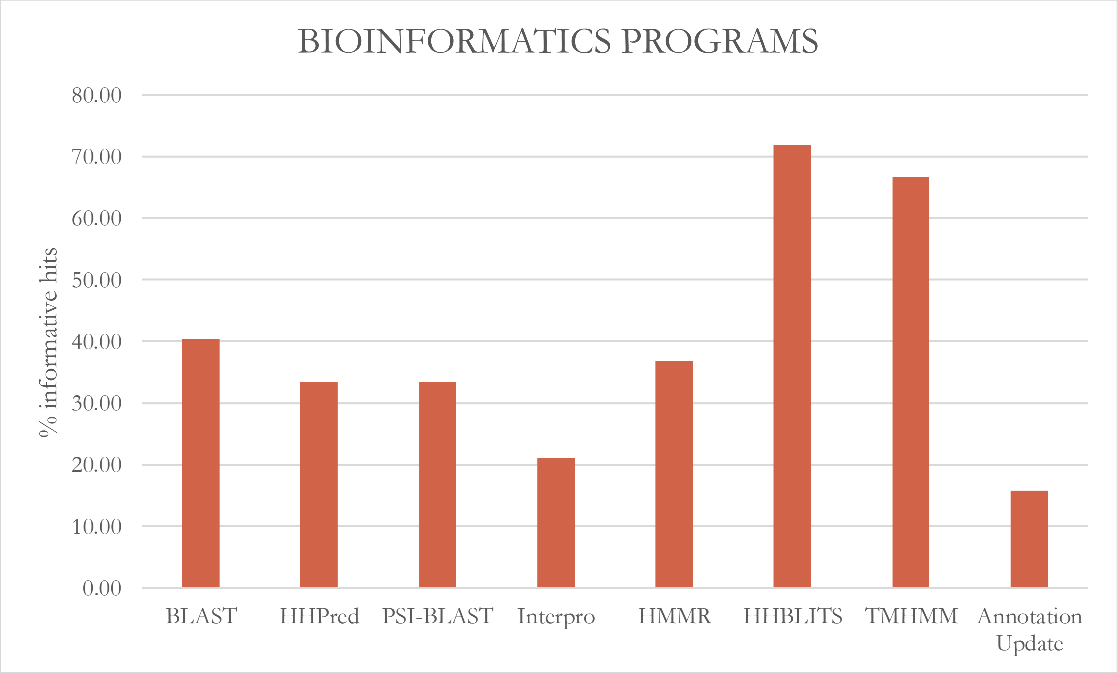


Figure 4.11. The percent of informative results from each bioinformatics program used on the 57 PotatoSplit hypothetical proteins.

4.4.2 I-TASSER and Simulations

I-TASSER compared the amino acid sequence for the Lysis Cassette protein with those in the Protein Data Bank and deduced protein function based on ligand binding sites and Gene Ontology (GO) terms. Table 4.14 shows the PDB results, with the top two ranking classifications being apoptosis functions. The next three are lyase, an enzyme that catalyzes the breaking of bonds. These results also match with what was hypothesized at the function of this protein.

In this table, RMSD is the root-mean-square distance and is the measure of the average distance between the atoms of superimposed proteins. Identity is the percentage sequence identity in the structurally aligned region. Coverage is the coverage of the alignment by TM-align and is equal to the number of structurally aligned residues divided by the length of the query protein.

Table 4.14. The Protein Data Bank classifications for functional matches to the lysis cassette NKF gene.

Rank	Classification	RMSD	Identity	Coverage
1	Apoptosis	3.92	0.101	0.806
2	Apoptosis	4.2	0.099	0.816
3	Lyase	5.01	0.069	0.83
4	Lyase	4.89	0.049	0.806
5	Lyase	4.7	0.039	0.791

Table 4.15 shows the GO term results, the highest of which suggests the protein has a primary metabolic function for biological processes and a molecular function of carbon-oxygen lyase activity. The GO-Score associated with each prediction is defined as the average weight of the GO term, where the weights are assigned based on CscoreGO of the template. The higher the number, the better the result is. The results here are not confident assignments.

The function reported by COFACTOR and COACH ligand binding site programs on the biological annotations of the target protein only had no results with a confidence interval higher than 0.8.

Table 4.15. The GO terms for the lysis cassette NKF gene.

Type	Function	GO Score
Molecular Function	carbon-oxygen lyase activity	0.37
Biological Process	primary metabolic process	0.48
Cellular Component	extracellular region	0.18
	nucleolus	0.07
	cytosol	0.07
	Golgi apparatus	0.07

The top structural results reported from I-TASSER had a C-score of -4.18, an estimated TM-score of 0.27±0.08, and an estimated RMSD of 5.6±3.3Å. This structure was then analyzed in PyMOL (Figure 4.12).

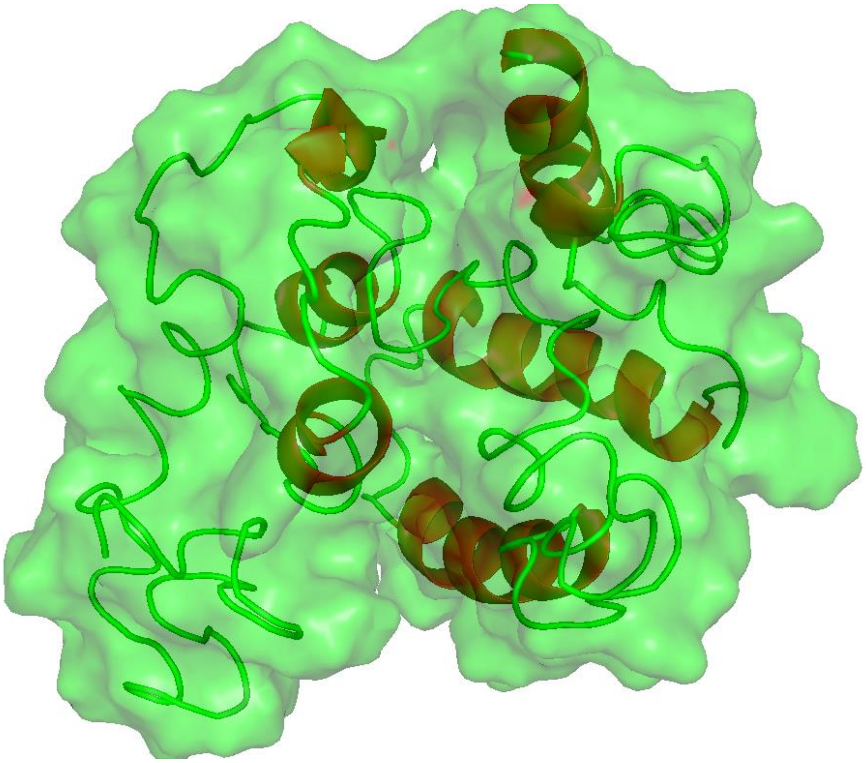


Figure 4.12. The predicted structure for the lysis cassette NKF gene.

The structure was then superimposed onto the secondary structure of a known bacteriophage hydrolase taken from the PDB database (Figure 4.13). The protein with PDB ID 3A9L was chosen as one of the few bacteriophage structures in PDB with the classification of hydrolase, and the structural similarity indicates they have some structure in common. They have a MatchAlign score of 64.509 and an RMSD of 1.62, indicating a medium level of structural similarity. However, the ITASSER model here does not have any beta sheets and is not as large. This could be due to the PDB structure being from a bacillus phage, which would not be as closely related to a mycobacteriophage.

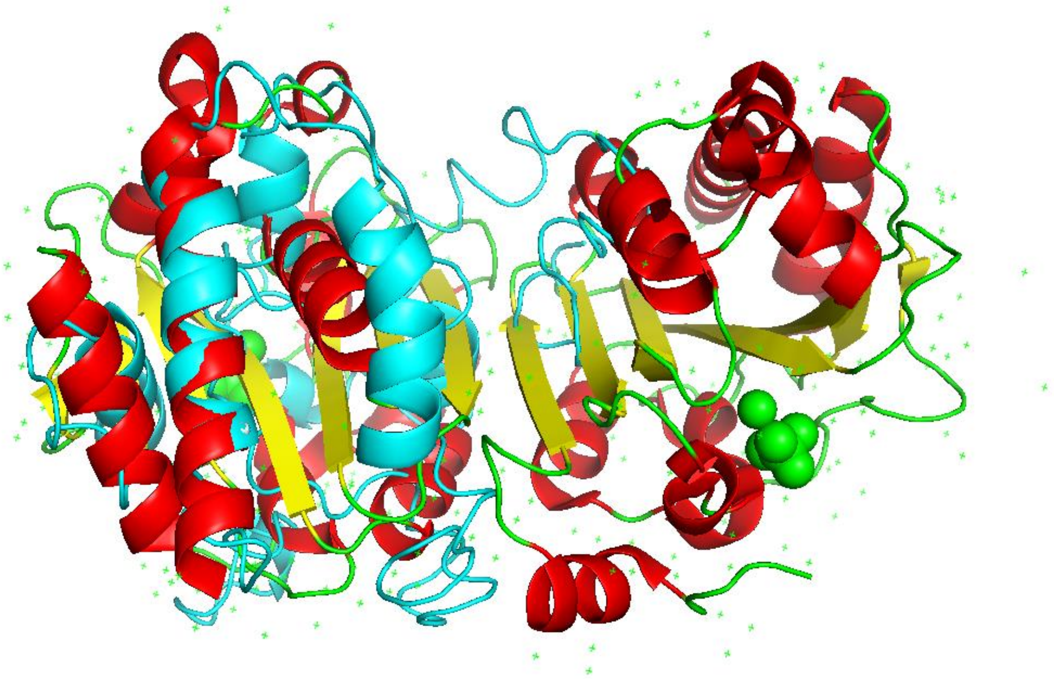


Figure 4.13. The secondary structure of the lysis cassette, shown in cyan, superimposed on the structure of the bacteriophage hydrolase (PDB ID: 3A9L), colored by secondary structure.

Table 4.16 below shows the top five PDB classifications for the Replication Cassette protein. All of the results are of the hydrolase classification, which is an enzyme that can break down bonds in proteins and polypeptides.

Table 4.16. Top 5 PDB results for the replication cassette.

Rank	Classification	RMSD	Identity	Coverage
1	Hydrolase	5.04	0.026	0.799
2	Hydrolase	5.02	0.039	0.799
3	Hydrolase	4.85	0.068	0.772
4	Hydrolase	5.03	0.079	0.788
5	Hydrolase	4.96	0.072	0.783

Table 4.17 has the top GO terms, with the highest scoring options being DNA binding, DNA-binding transcription, steroid hormone receptor, regulation of transcription, and steroid hormone-mediated signaling pathway. Most of those functions match what was expected of this protein, as it should be related to DNA replication. Transcription being a step in DNA replication, this is a promising result.

The function reported by COFACTOR and COACH ligand binding site programs on the biological annotations of the target protein only had no results with confidence above 0.80.

Table 4.17. The consensus prediction of GO terms for the replication cassette among the top scoring templates.

Type	Function	GO Score
Molecular Function	steroid hormone receptor activity	0.13
	DNA binding	0.13
	DNA-binding transcription factor activity	0.13
	oxidoreductase activity	0.07
	metal ion binding	0.07
Biological Process	regulation of transcription, DNA-templated	0.13
	steroid hormone mediated signaling pathway	0.13
	regulation of cell cycle	0.07
	oxidation-reduction process	0.07
Cellular Component	intracellular membrane-bounded organelle	0.37

The top structural results reported from I-TASSER had a C-score of -4.18, an estimated TM-score of 0.27±0.08, and an estimated RMSD of 15.6±3.3 Å. It is shown in Figure 4.14.

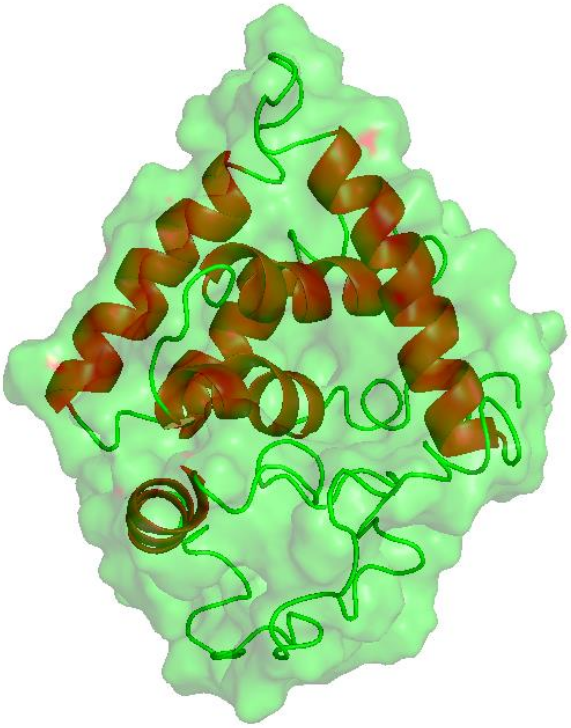


Figure 4.14. The structure of the replication cassette protein showing secondary structure and surface.

This structure was then analyzed in PyMOL (Figure 4.14) and superimposed onto the secondary structure of multiple known function proteins taken from PDB. The first is the structure of a dihydrofolate reductase from bacteriophage T4, a lyase classified protein (Figure 4.15). The MatchAlign score is low at 41.204, but the RMSD is 1.419, which shows a good match for part of the structure.

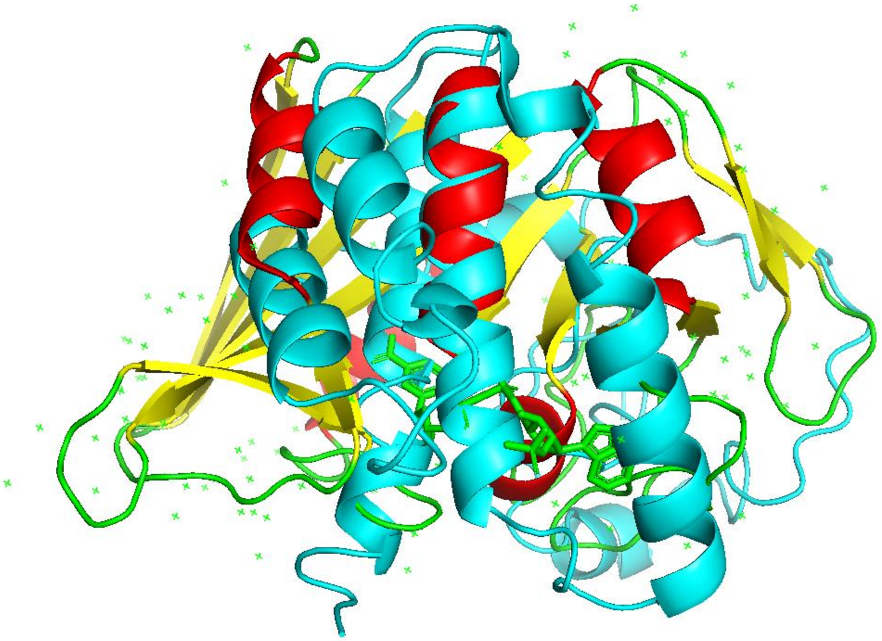


Figure 4.15. The secondary structure of the replication cassette, shown in cyan, superimposed on the structure of the bacteriophage reductase (PDB ID: 1JUV), colored by secondary structure.

After the structure prediction, simulations were run. First, the system was initialized to ensure standard temperatures and pressures, then the structure was simulated at Cell Temperature and Autoclave Temperature for 0.5 nanoseconds (ns). The lysis cassette had an average Root Mean Square Distance (RMSD) of 0.45 nm for the Autoclave simulations, and 0.32 nm for the Cell simulations over a time course of 0.5 ns. The results are shown in Figure 4.16. While the Autoclave simulation continues to deteriorate over time, the Cell simulation remains relatively stable after the first 0.1 ns. The RMSD for the lysis protein is slightly lower than the replication cassette, but the curve is very similar.

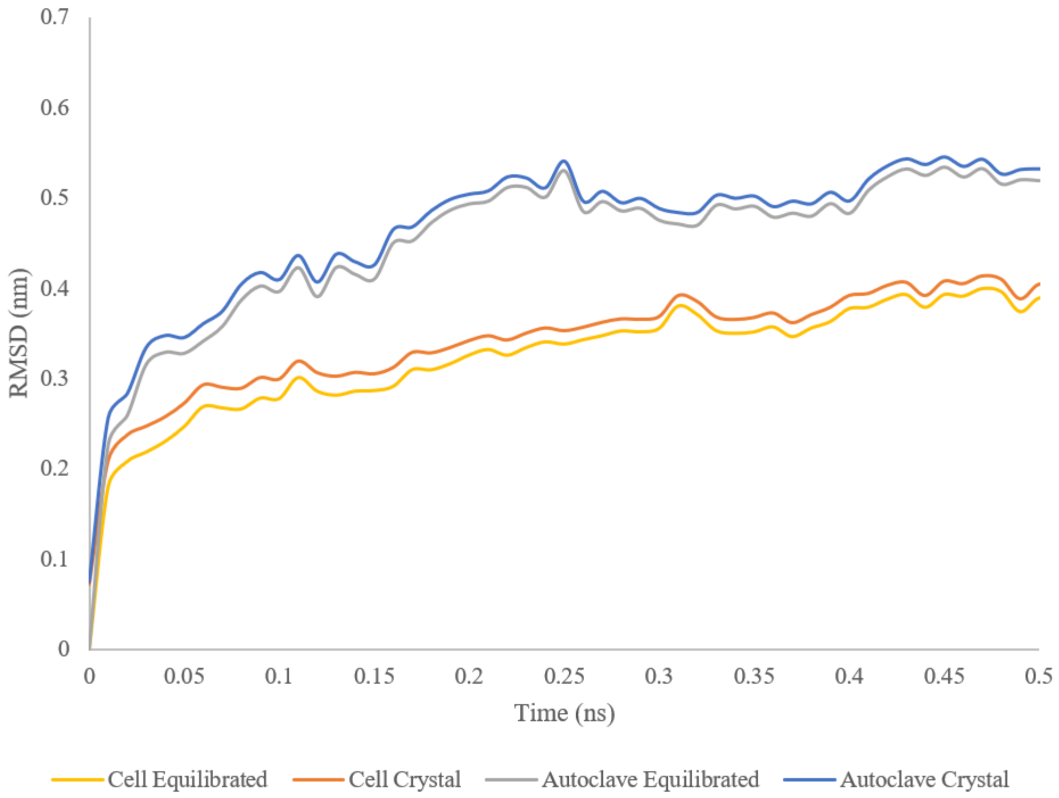


Figure 4.16. The RMSD of the Lysis Cassette protein over 0.5 ns. Both simulations, Autoclave and Cell, are compared with the original crystal structure and the equilibrated structure. The Autoclave simulation has higher RMSD because the protein deteriorates at this temperature.

The replication cassette had an average RMSD of 0.88 nm/fs for the Autoclave simulations, and 0. nm/fs for the Cell simulations over a time course of 0.5 ns. The results are shown in Figure 4.17. While the Autoclave simulation continues to deteriorate over time, the Cell simulation remains relatively stable after the first 0.1 ns. This is expected, as the natural production of the protein would be at this temperature. It would have to stay in the configuration to carry out its function.

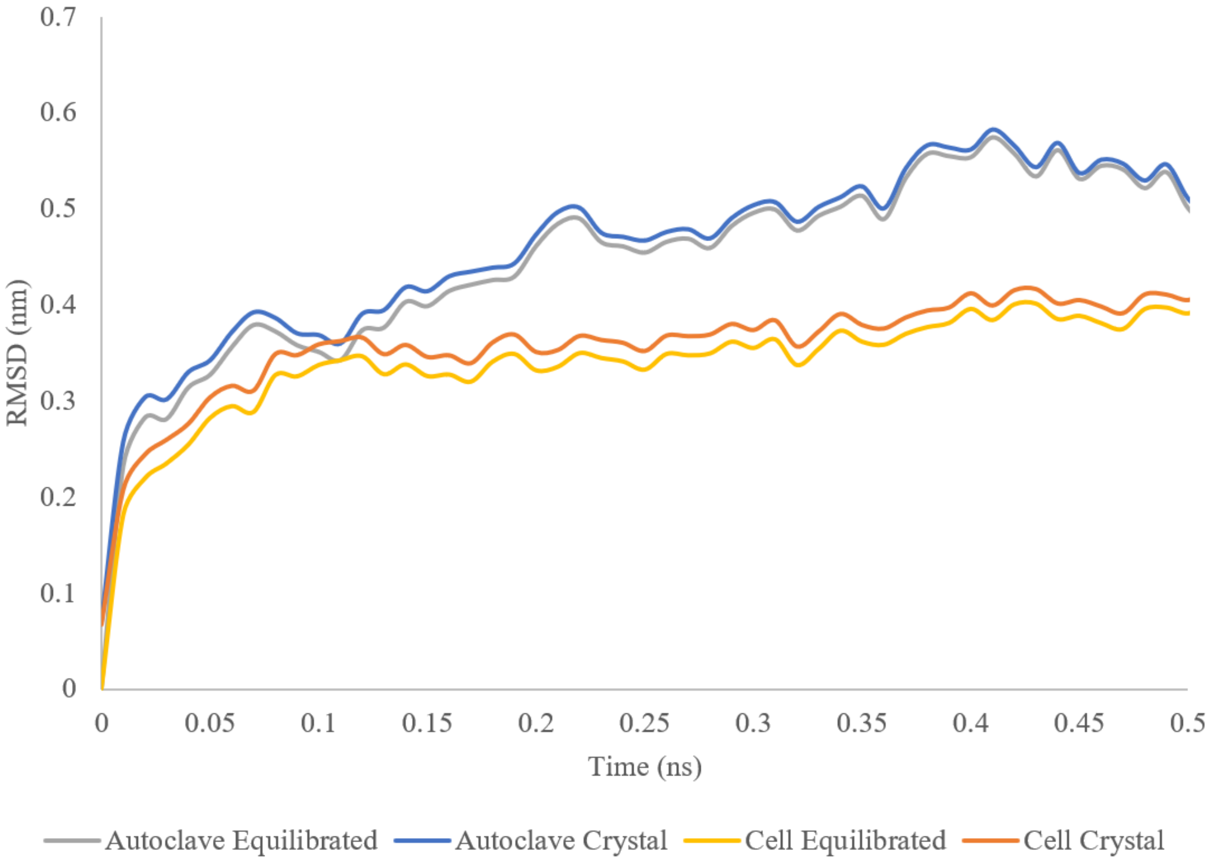


Figure 4.17. The RMSD of the Replication Cassette protein over 0.1 ns. Both simulations, Autoclave and Cell, are compared with the original crystal structure and the equilibrated structure. The Autoclave simulation has higher RMSD because the protein deteriorates at this temperature.

In Figure 4.18, the RMSD of two of the NKF proteins is plotted along with one unrelated protein a Tail Assembly Chaperone (TAC), and one known structure from Protein Data Bank, a bacteriophage dihydrofolate reductase that was structurally compared to the Replication cassette protein earlier. For this protein, 1JUV, the same GROMACs simulation was run at the Autoclave temperature setting in the interest of comparing results with a known protein structure. The known structure had a much lower average RMSD, with an average of 0.14 nm.

This is because of the mistakes with predicted protein structures. ITASSER, while the best program for predicting protein structure, is not always correct. The models used from ITASSER have larger conformational changes throughout the simulation because they are not experimentally derived.

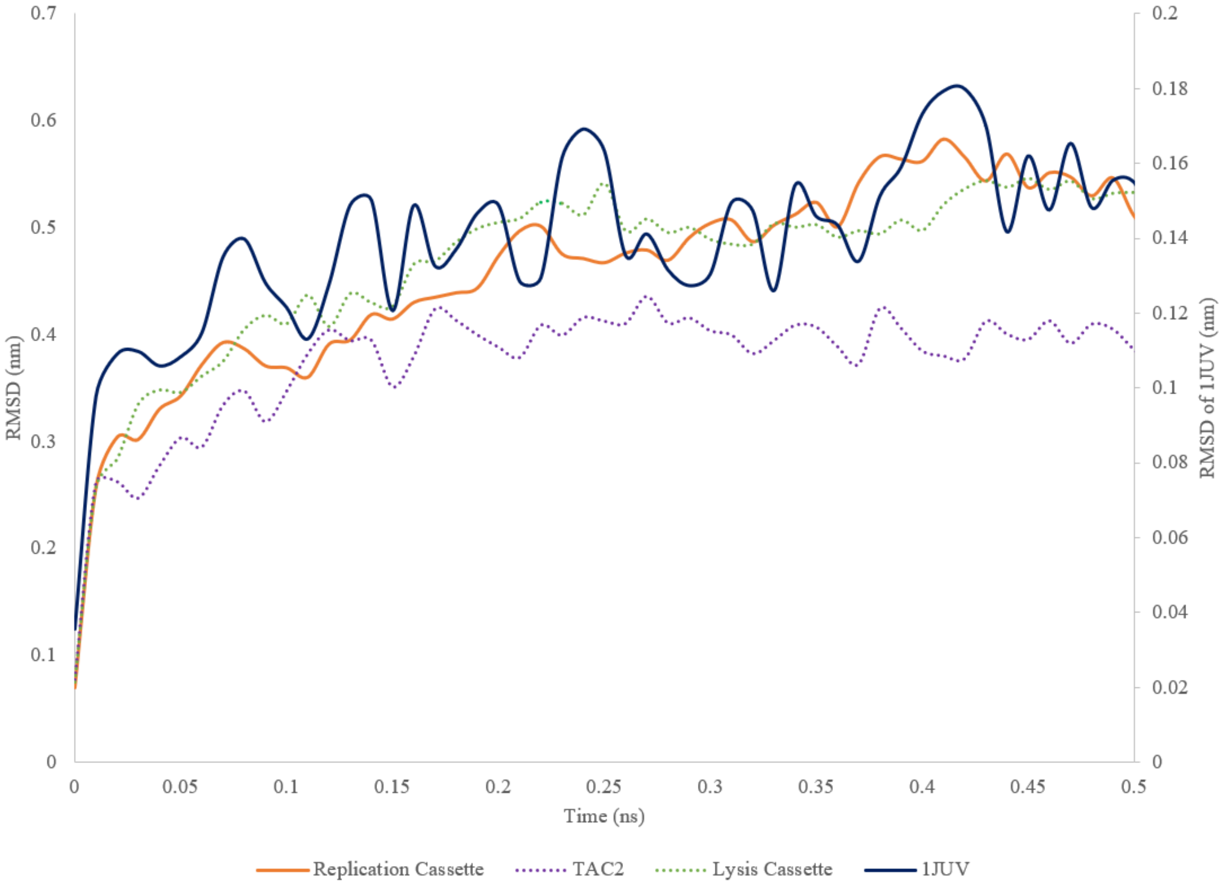


Figure 4.18. The RMSD functions during the Autoclave simulation for the Replication Cassette protein, the tail assembly chaperone two protein, the lysis cassette protein, and the known protein structure IJUV from PDB.

When viewing the graph, certain aspects of the curves are similar across all proteins. The protein structure IJUV was a structural match for the replication cassette of the function hydrolase. This makes it the same classification as the replication cassette and the lysis cassette. These three curves follow a similar pattern of peaks and valleys that is distinct from the pattern displayed by the tail assembly chaperone. This could speak to the proteins behaving in a similar manner, which can be used to infer that they may have similar functions as well.

4.4.3 Machine Learning

To determine the optimal number of amino acids to use as features, the length of each sequence was graphed as a histogram in Figure 4.19 below. The median amount was 569 amino acids.

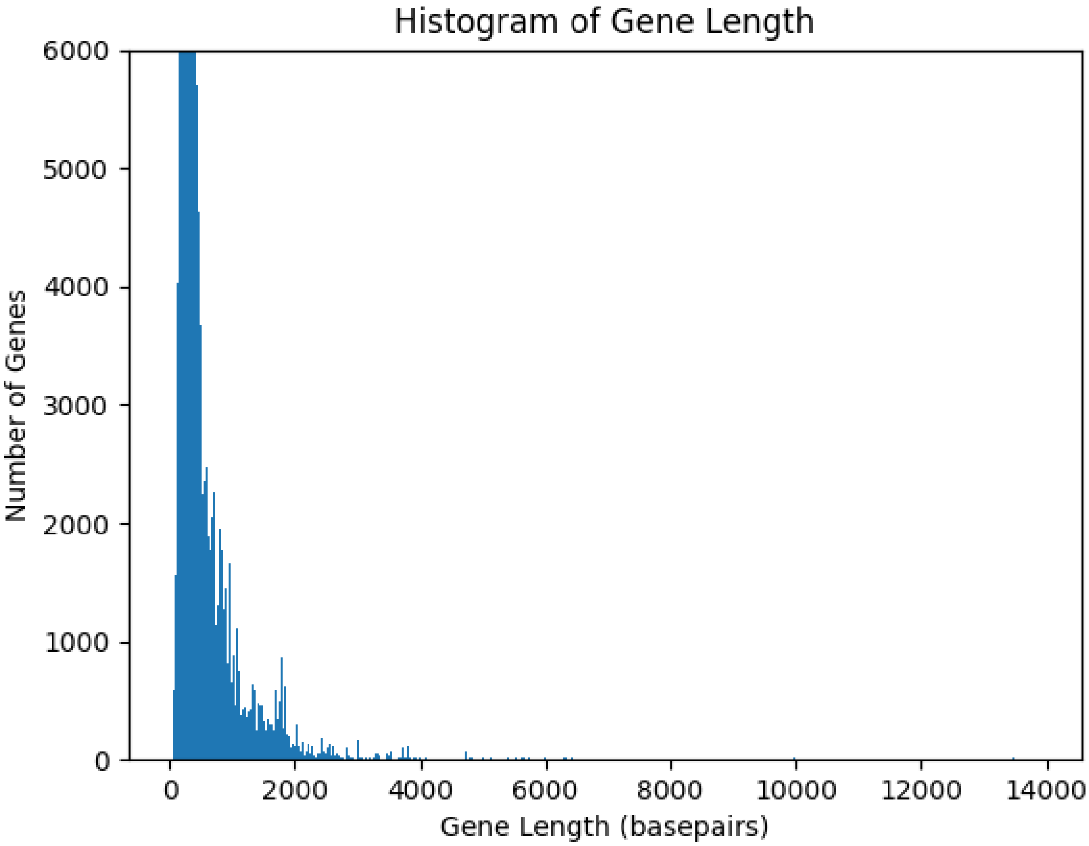


Figure 4.19. A histogram of gene length showing the number of genes at each length. The distribution is right-tailed, with the majority of genes falling between 0 and 2000 base pairs.

To test the optimal number of features, the model was trained on 300, 400, 500, and 600 amino acids. Figure 4.20 below shows the testing data's precision and accuracy values for various amino acid inputs. Start site and length were always inputted. Three statistical replicates were tested for each amino acid value and error bars show the standard deviations. There was high variability.

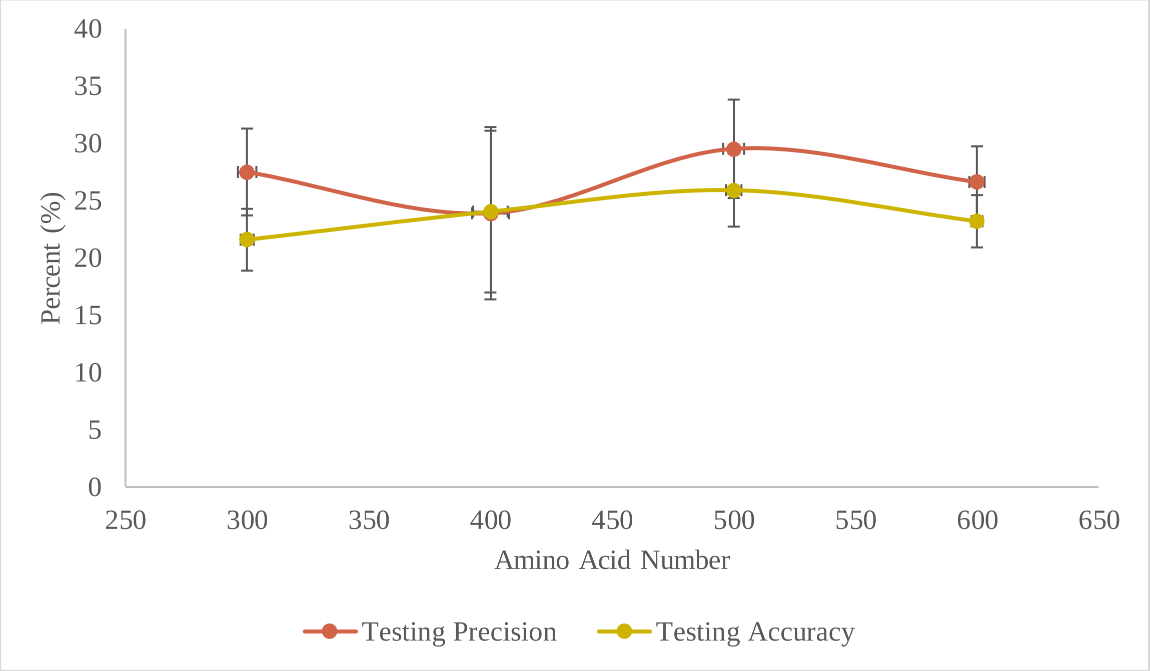


Figure 4.20. The precision and accuracy of the testing data based on amino acids inputted. The error bars show the standard deviations; three replicates were tested for each amino acid value.

The amino acid value of 500 was chosen to continue with testing. After this, there was a decrease in precision and accuracy, possibly due to the median number of amino acids being 569, so adding more amino acids resulted in larger filler values.

After picking this amino acid value, the number of trees or estimators was tested. Values tested were 500, 1000, 2000, and 2500.

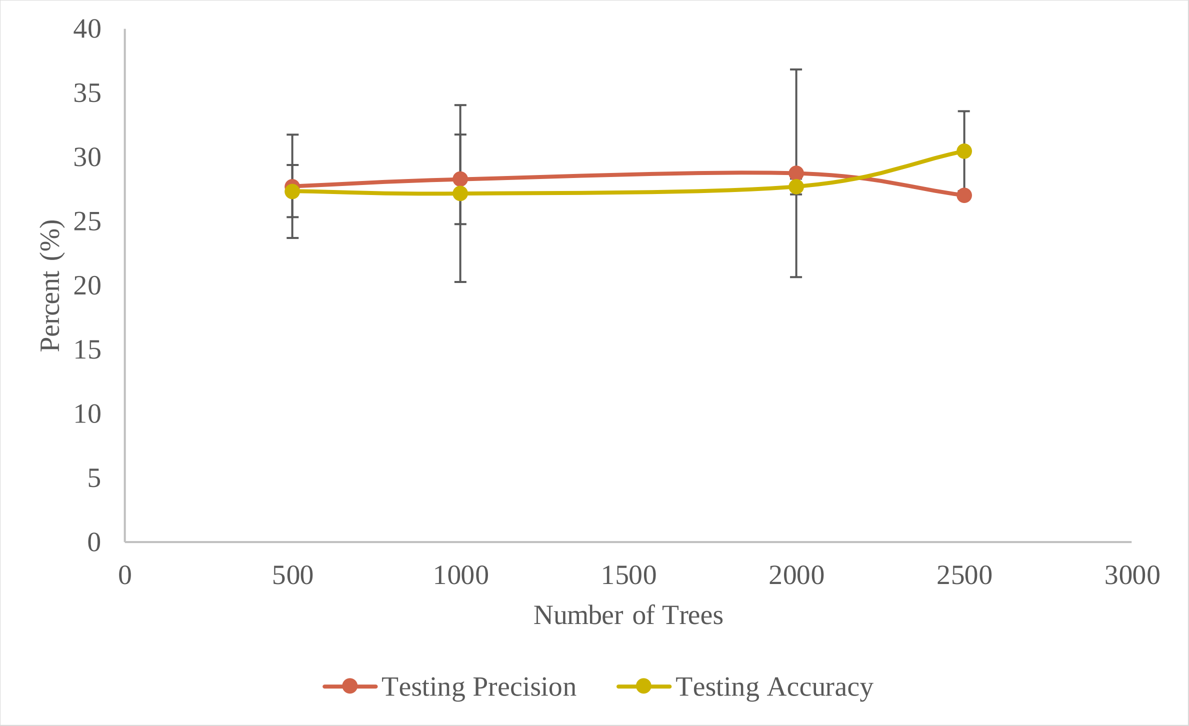
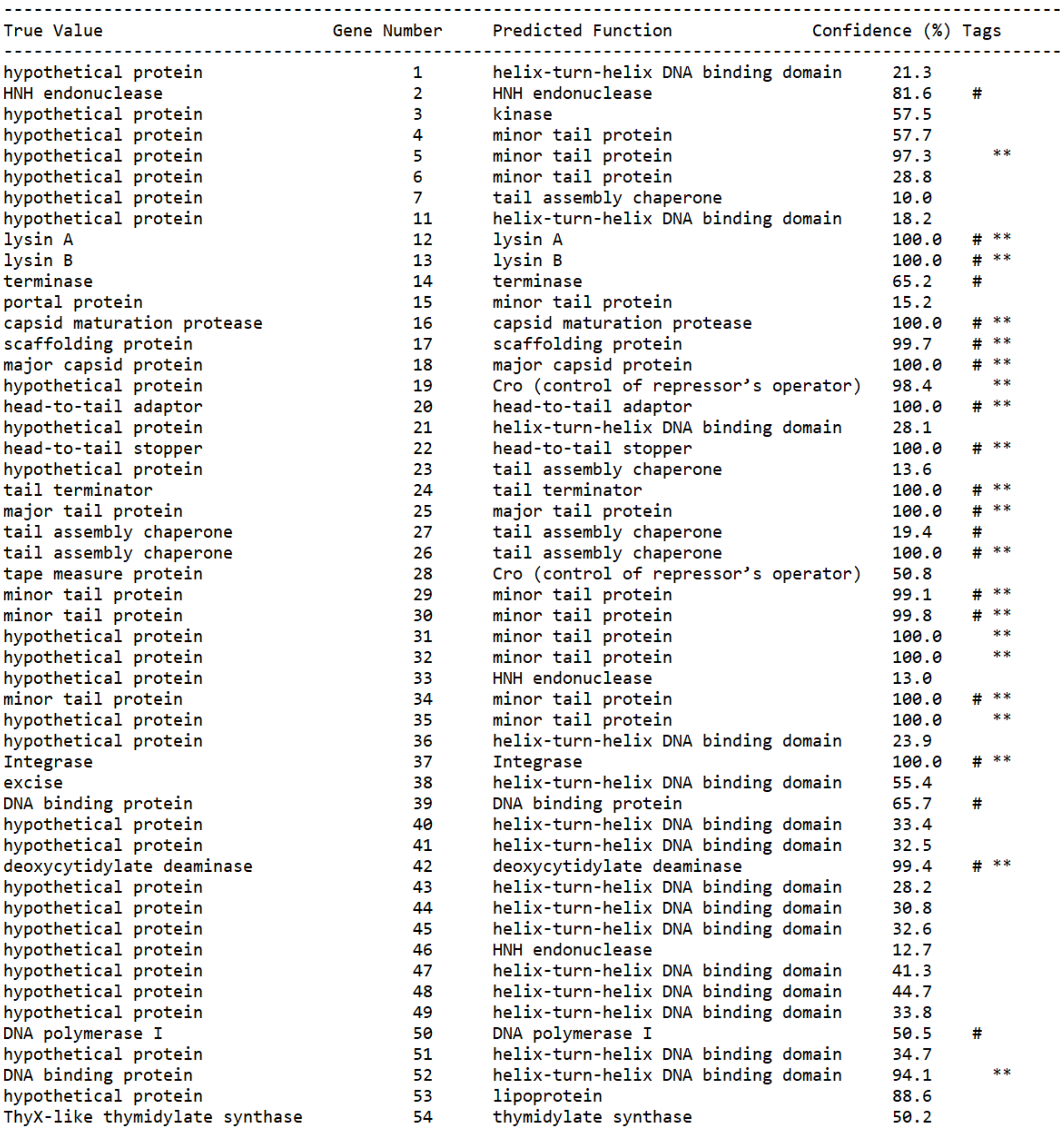


Figure 4.21. The precision and accuracy of the testing data when using 500 amino acids, start site, and length as inputs. Three statistical replicates were preformed for each estimator value; error bars show the standard deviations.

The value of 2000 was chosen as the optimal value. After this, accuracy increased but precision decreased.

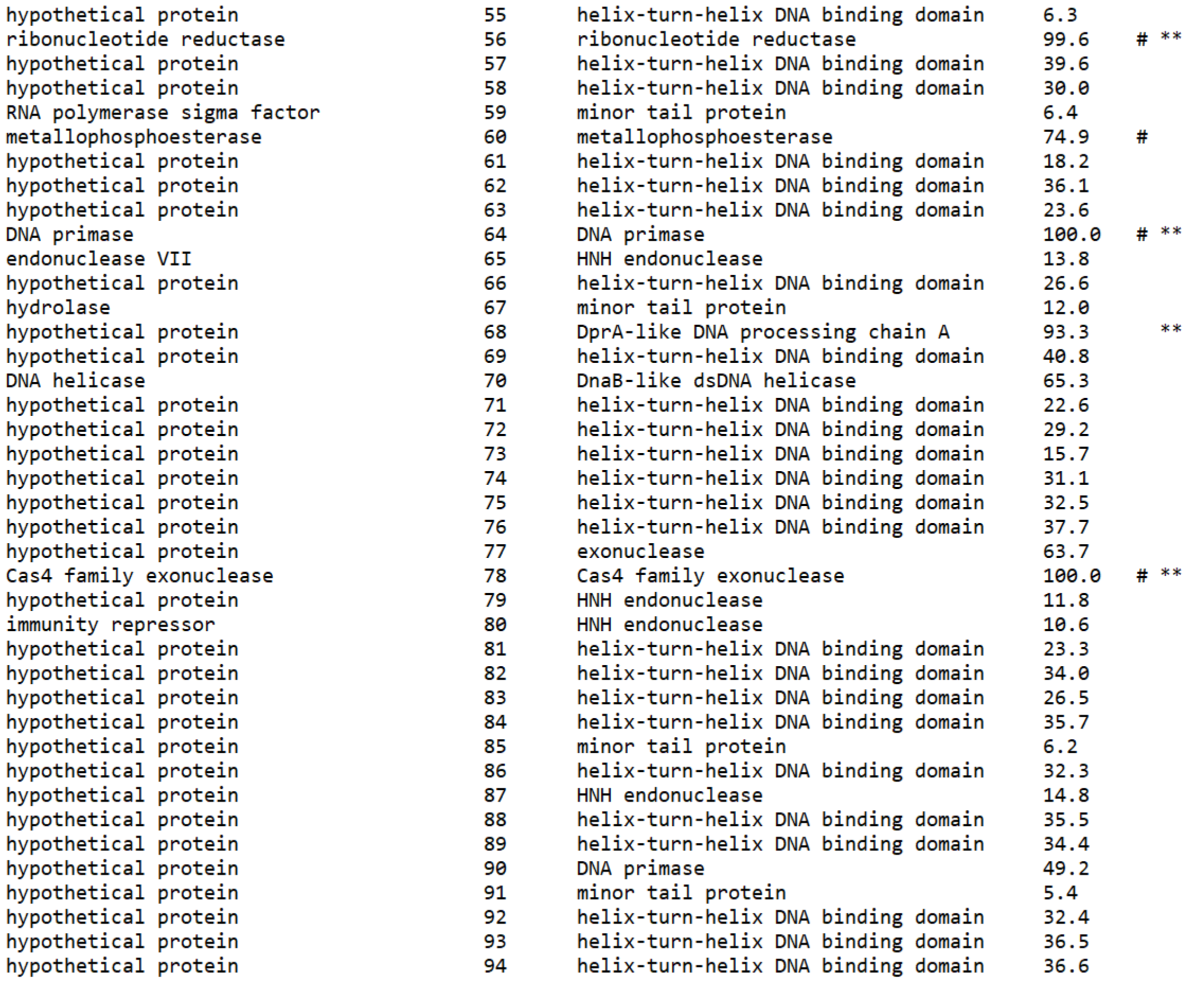
For this model version with 500 amino acids, start site, and length inputs and 2000 trees used, the average training accuracy was 98.99%, testing accuracy was 27.71%, testing precision was 28.74%. The results for PotatoSplit can be viewed in Figure 4.22

Figure 4.22. Results for PotatoSplit from the best version of the Random Forest algorithm. The “#” tag shows when the prediction matched the true value and the “**” tag shows that the confidence is above 90%.



```
----------------------------------------------------------------------------------------------------------
True Value                       Gene Number     Predicted Function            Confidence (%) Tags
----------------------------------------------------------------------------------------------------------
hypothetical protein                     1       helix-turn-helix DNA binding domain     21.3
HNH endonuclease                         2       HNH endonuclease                        81.6    #
hypothetical protein                     3       kinase                                  57.5
hypothetical protein                     4       minor tail protein                      57.7
hypothetical protein                     5       minor tail protein                      97.3      **
hypothetical protein                     6       minor tail protein                      28.8
hypothetical protein                     7       tail assembly chaperone                 10.0
hypothetical protein                     11      helix-turn-helix DNA binding domain     18.2
lysin A                                  12      lysin A                                 100.0   # **
lysin B                                  13      lysin B                                 100.0   # **
terminase                                14      terminase                               65.2    #
portal protein                           15      minor tail protein                      15.2
capsid maturation protease               16      capsid maturation protease              100.0   # **
scaffolding protein                      17      scaffolding protein                     99.7    # **
major capsid protein                     18      major capsid protein                    100.0   # **
hypothetical protein                     19      Cro (control of repressor’s operator)   98.4      **
head-to-tail adaptor                     20      head-to-tail adaptor                    100.0   # **
hypothetical protein                     21      helix-turn-helix DNA binding domain     28.1
head-to-tail stopper                     22      head-to-tail stopper                    100.0   # **
hypothetical protein                     23      tail assembly chaperone                 13.6
tail terminator                          24      tail terminator                         100.0   # **
major tail protein                       25      major tail protein                      100.0   # **
tail assembly chaperone                  27      tail assembly chaperone                 19.4    #
tail assembly chaperone                  26      tail assembly chaperone                 100.0   # **
tape measure protein                     28      Cro (control of repressor’s operator)   50.8
minor tail protein                       29      minor tail protein                      99.1    # **
minor tail protein                       30      minor tail protein                      99.8    # **
hypothetical protein                     31      minor tail protein                      100.0     **
hypothetical protein                     32      minor tail protein                      100.0     **
hypothetical protein                     33      HNH endonuclease                        13.0
minor tail protein                       34      minor tail protein                      100.0   # **
hypothetical protein                     35      minor tail protein                      100.0     **
hypothetical protein                     36      helix-turn-helix DNA binding domain     23.9
Integrase                                37      Integrase                               100.0   # **
excise                                   38      helix-turn-helix DNA binding domain     55.4
DNA binding protein                      39      DNA binding protein                     65.7    #
hypothetical protein                     40      helix-turn-helix DNA binding domain     33.4
hypothetical protein                     41      helix-turn-helix DNA binding domain     32.5
deoxycytidylate deaminase                42      deoxycytidylate deaminase               99.4    # **
hypothetical protein                     43      helix-turn-helix DNA binding domain     28.2
hypothetical protein                     44      helix-turn-helix DNA binding domain     30.8
hypothetical protein                     45      helix-turn-helix DNA binding domain     32.6
hypothetical protein                     46      HNH endonuclease                        12.7
hypothetical protein                     47      helix-turn-helix DNA binding domain     41.3
hypothetical protein                     48      helix-turn-helix DNA binding domain     44.7
hypothetical protein                     49      helix-turn-helix DNA binding domain     33.8
DNA polymerase I                         50      DNA polymerase I                        50.5    #
hypothetical protein                     51      helix-turn-helix DNA binding domain     34.7
DNA binding protein                      52      helix-turn-helix DNA binding domain     94.1      **
hypothetical protein                     53      lipoprotein                             88.6
ThyX-like thymidylate synthase           54      thymidylate synthase                    50.2
```

Figure 4.22 Continued.



```
hypothetical protein                     55      helix-turn-helix DNA binding domain     6.3
ribonucleotide reductase                 56      ribonucleotide reductase                99.6    # **
hypothetical protein                     57      helix-turn-helix DNA binding domain     39.6
hypothetical protein                     58      helix-turn-helix DNA binding domain     30.0
RNA polymerase sigma factor              59      minor tail protein                      6.4
metallophosphoesterase                   60      metallophosphoesterase                  74.9    #
hypothetical protein                     61      helix-turn-helix DNA binding domain     18.2
hypothetical protein                     62      helix-turn-helix DNA binding domain     36.1
hypothetical protein                     63      helix-turn-helix DNA binding domain     23.6
DNA primase                              64      DNA primase                             100.0   # **
endonuclease VII                         65      HNH endonuclease                        13.8
hypothetical protein                     66      helix-turn-helix DNA binding domain     26.6
hydrolase                                67      minor tail protein                      12.0
hypothetical protein                     68      DprA-like DNA processing chain A        93.3      **
hypothetical protein                     69      helix-turn-helix DNA binding domain     40.8
DNA helicase                             70      DnaB-like dsDNA helicase                65.3
hypothetical protein                     71      helix-turn-helix DNA binding domain     22.6
hypothetical protein                     72      helix-turn-helix DNA binding domain     29.2
hypothetical protein                     73      helix-turn-helix DNA binding domain     15.7
hypothetical protein                     74      helix-turn-helix DNA binding domain     31.1
hypothetical protein                     75      helix-turn-helix DNA binding domain     32.5
hypothetical protein                     76      helix-turn-helix DNA binding domain     37.7
hypothetical protein                     77      exonuclease                             63.7
Cas4 family exonuclease                  78      Cas4 family exonuclease                 100.0   # **
hypothetical protein                     79      HNH endonuclease                        11.8
immunity repressor                       80      HNH endonuclease                        10.6
hypothetical protein                     81      helix-turn-helix DNA binding domain     23.3
hypothetical protein                     82      helix-turn-helix DNA binding domain     34.0
hypothetical protein                     83      helix-turn-helix DNA binding domain     26.5
hypothetical protein                     84      helix-turn-helix DNA binding domain     35.7
hypothetical protein                     85      minor tail protein                      6.2
hypothetical protein                     86      helix-turn-helix DNA binding domain     32.3
hypothetical protein                     87      HNH endonuclease                        14.8
hypothetical protein                     88      helix-turn-helix DNA binding domain     35.5
hypothetical protein                     89      helix-turn-helix DNA binding domain     34.4
hypothetical protein                     90      DNA primase                             49.2
hypothetical protein                     91      minor tail protein                      5.4
hypothetical protein                     92      helix-turn-helix DNA binding domain     32.4
hypothetical protein                     93      helix-turn-helix DNA binding domain     36.5
hypothetical protein                     94      helix-turn-helix DNA binding domain     36.6
```

Of the 34 PotatoSplit genes with functional tags, the algorithm identified 24 of them correctly, for an accuracy of 70.59%. One additional protein was called as helix-turn-helix DNA binding protein by the algorithm but had a label “DNA Binding Protein”, so this may also be considered a match. Genes 3, 4, 5, and 6 were identified as minor tail proteins by the new bioinformatics method, and 4, 5, and 6 show this function from the algorithm as well. Genes 31-35 were also predicted to be minor tail proteins, which the algorithm also matched other than Gene 33. Gene 68 was called by both methods, as well gene 73.

I compared my Random Forest model to the weak leaner K-nearest neighbor method. The k-nearest neighbor method has an average testing accuracy of 14.16% and the Random Forest model has an average testing accuracy of 27.71%. This shows the random forest model is a better algorithm, but the data set may be too small for the model to learn more. Figure 4.23 shows the accuracy and precision of the k-nearest neighbor method, the random forest model, and a version of the Random Forest model that predicts only functional classification. I used eight functional classes: Structure, DNA Replication & Translation, DNA Integration, Lysis, Biosynthesis & Energy, Gene Regulation, Defense, and Other.

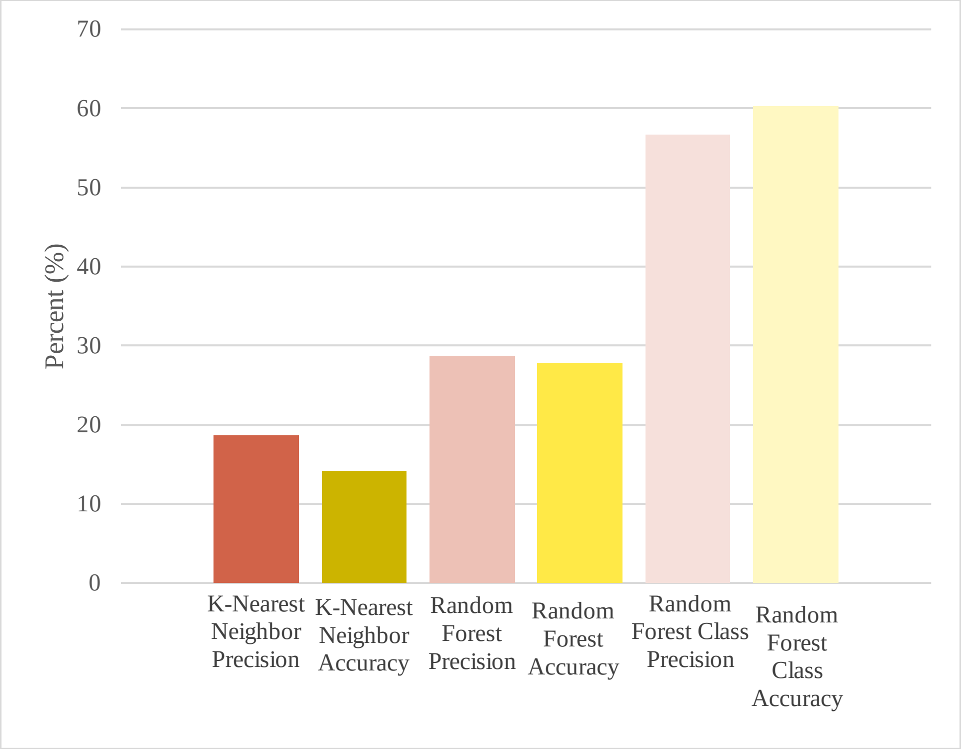
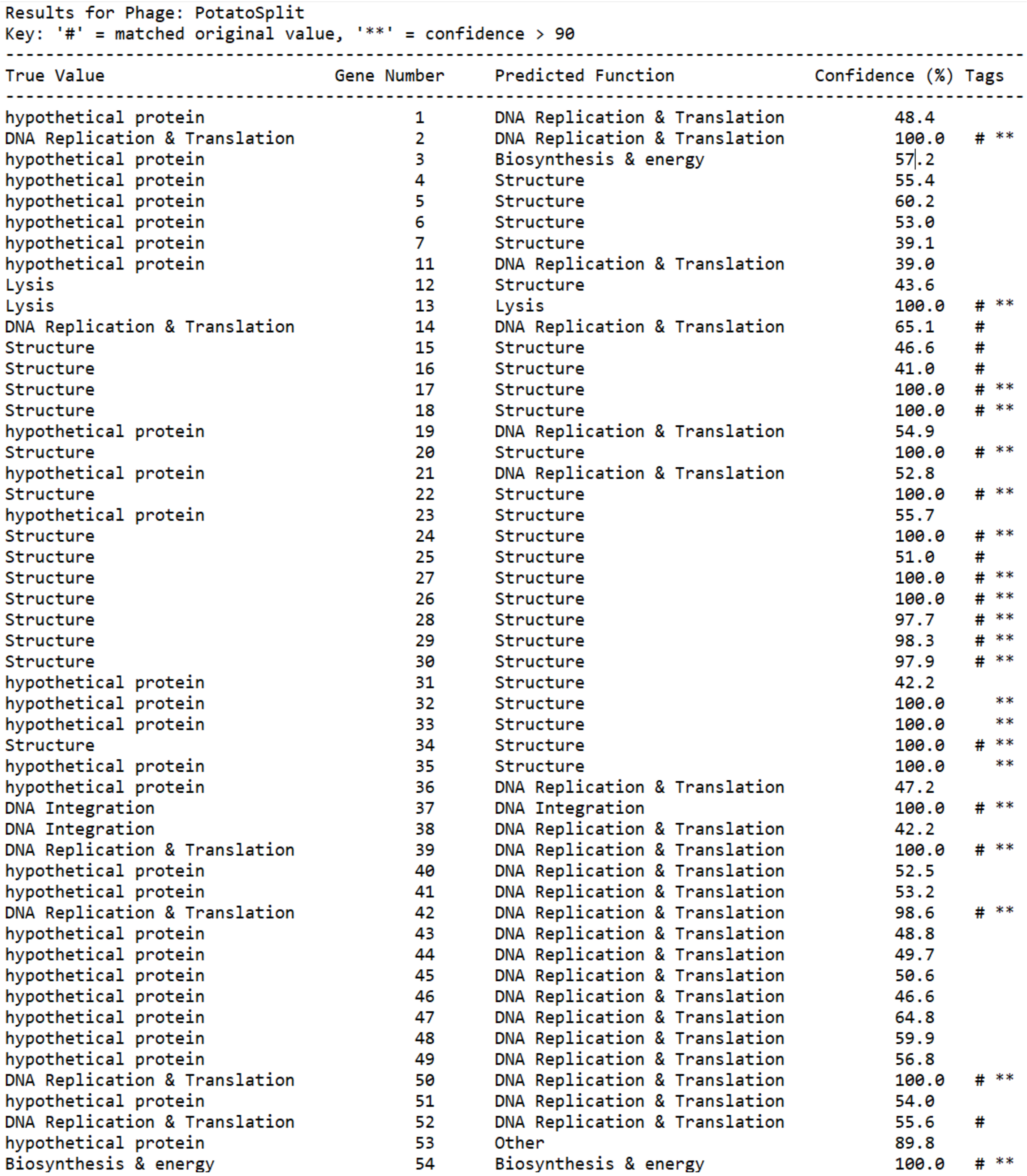


Figure 4.23. Precisions and accuracies of three model types, with precision shown in red and accuracy shown in yellow. Random Forest is almost twice as accurate as the weak learner.

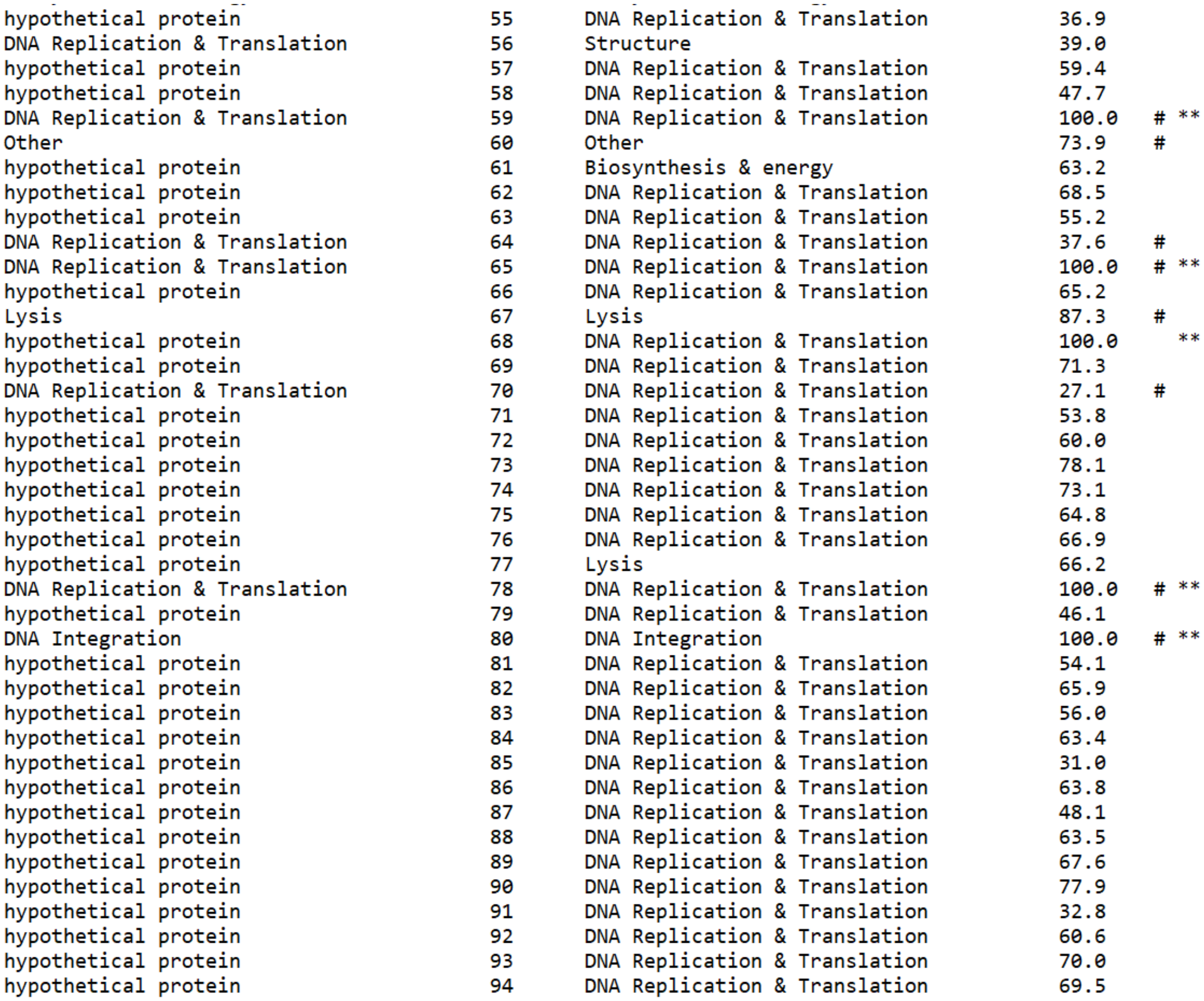
The class predictions of PotatoSplit's genome can be seen in Figure 4.24. It matched the predictions from the new bioinformatics method and offers more confidence into the possible functional classes of genes.

Figure 4.24. The functional classes of genes within PotatoSplit.



```
Results for Phage: PotatoSplit
Key: '#' = matched original value, '**' = confidence > 90
------------------------------------------------------------------------------------------------------
True Value                       Gene Number     Predicted Function              Confidence (%) Tags
------------------------------------------------------------------------------------------------------
hypothetical protein                     1       DNA Replication & Translation           48.4
DNA Replication & Translation            2       DNA Replication & Translation           100.0    # **
hypothetical protein                     3       Biosynthesis & energy                   57.2
hypothetical protein                     4       Structure                               55.4
hypothetical protein                     5       Structure                               60.2
hypothetical protein                     6       Structure                               53.0
hypothetical protein                     7       Structure                               39.1
hypothetical protein                     11      DNA Replication & Translation           39.0
Lysis                                    12      Structure                               43.6
Lysis                                    13      Lysis                                   100.0    # **
DNA Replication & Translation            14      DNA Replication & Translation           65.1     #
Structure                                15      Structure                               46.6     #
Structure                                16      Structure                               41.0     #
Structure                                17      Structure                               100.0    # **
Structure                                18      Structure                               100.0    # **
hypothetical protein                     19      DNA Replication & Translation           54.9
Structure                                20      Structure                               100.0    # **
hypothetical protein                     21      DNA Replication & Translation           52.8
Structure                                22      Structure                               100.0    # **
hypothetical protein                     23      Structure                               55.7
Structure                                24      Structure                               100.0    # **
Structure                                25      Structure                               51.0     #
Structure                                27      Structure                               100.0    # **
Structure                                26      Structure                               100.0    # **
Structure                                28      Structure                               97.7     # **
Structure                                29      Structure                               98.3     # **
Structure                                30      Structure                               97.9     # **
hypothetical protein                     31      Structure                               42.2
hypothetical protein                     32      Structure                               100.0      **
hypothetical protein                     33      Structure                               100.0      **
Structure                                34      Structure                               100.0    # **
hypothetical protein                     35      Structure                               100.0      **
hypothetical protein                     36      DNA Replication & Translation           47.2
DNA Integration                          37      DNA Integration                         100.0    # **
DNA Integration                          38      DNA Replication & Translation           42.2
DNA Replication & Translation            39      DNA Replication & Translation           100.0    # **
hypothetical protein                     40      DNA Replication & Translation           52.5
hypothetical protein                     41      DNA Replication & Translation           53.2
DNA Replication & Translation            42      DNA Replication & Translation           98.6     # **
hypothetical protein                     43      DNA Replication & Translation           48.8
hypothetical protein                     44      DNA Replication & Translation           49.7
hypothetical protein                     45      DNA Replication & Translation           50.6
hypothetical protein                     46      DNA Replication & Translation           46.6
hypothetical protein                     47      DNA Replication & Translation           64.8
hypothetical protein                     48      DNA Replication & Translation           59.9
hypothetical protein                     49      DNA Replication & Translation           56.8
DNA Replication & Translation            50      DNA Replication & Translation           100.0    # **
hypothetical protein                     51      DNA Replication & Translation           54.0
DNA Replication & Translation            52      DNA Replication & Translation           55.6     #
hypothetical protein                     53      Other                                   89.8
Biosynthesis & energy                    54      Biosynthesis & energy                   100.0    # **
```

Figure 4.24 Continued.



```
hypothetical protein              55    DNA Replication & Translation       36.9
DNA Replication & Translation     56    Structure                           39.0
hypothetical protein              57    DNA Replication & Translation       59.4
hypothetical protein              58    DNA Replication & Translation       47.7
DNA Replication & Translation     59    DNA Replication & Translation       100.0   # **
Other                             60    Other                               73.9    #
hypothetical protein              61    Biosynthesis & energy               63.2
hypothetical protein              62    DNA Replication & Translation       68.5
hypothetical protein              63    DNA Replication & Translation       55.2
DNA Replication & Translation     64    DNA Replication & Translation       37.6    #
DNA Replication & Translation     65    DNA Replication & Translation       100.0   # **
hypothetical protein              66    DNA Replication & Translation       65.2
Lysis                             67    Lysis                               87.3    #
hypothetical protein              68    DNA Replication & Translation       100.0     **
hypothetical protein              69    DNA Replication & Translation       71.3
DNA Replication & Translation     70    DNA Replication & Translation       27.1    #
hypothetical protein              71    DNA Replication & Translation       53.8
hypothetical protein              72    DNA Replication & Translation       60.0
hypothetical protein              73    DNA Replication & Translation       78.1
hypothetical protein              74    DNA Replication & Translation       73.1
hypothetical protein              75    DNA Replication & Translation       64.8
hypothetical protein              76    DNA Replication & Translation       66.9
hypothetical protein              77    Lysis                               66.2
DNA Replication & Translation     78    DNA Replication & Translation       100.0   # **
hypothetical protein              79    DNA Replication & Translation       46.1
DNA Integration                   80    DNA Integration                     100.0   # **
hypothetical protein              81    DNA Replication & Translation       54.1
hypothetical protein              82    DNA Replication & Translation       65.9
hypothetical protein              83    DNA Replication & Translation       56.0
hypothetical protein              84    DNA Replication & Translation       63.4
hypothetical protein              85    DNA Replication & Translation       31.0
hypothetical protein              86    DNA Replication & Translation       63.8
hypothetical protein              87    DNA Replication & Translation       48.1
hypothetical protein              88    DNA Replication & Translation       63.5
hypothetical protein              89    DNA Replication & Translation       67.6
hypothetical protein              90    DNA Replication & Translation       77.9
hypothetical protein              91    DNA Replication & Translation       32.8
hypothetical protein              92    DNA Replication & Translation       60.6
hypothetical protein              93    DNA Replication & Translation       70.0
hypothetical protein              94    DNA Replication & Translation       69.5
```

5. DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

5.1 Discussion

5.1.1 Bacteriophage Research

Analysis of phage proteins can lead to a better understanding of key aspects of phage infection. The differences outlined in the Purdue University phages already offer key insights into phage differences and how they could be used in the creation of a phage therapy product. This research offers insight involving *in silico* methods with future work being needed in wet lab to confirm any findings.

Just by examining the proteins found in lytic versus temperate phages, differences arise with key proteins like DNA polymerase and terminase. These phages have evolved to not need DNA Polymerase I, which has a function of hydrolyzing the RNA primer during DNA replication to fill in the gaps with complementary DNA bases at the end of the DNA replication[107,109]. DNA Polymerase III selects and adds bases to the DNA template strand, catalyzes the bonds between bases, and "proofreads" the bases against the template to remove any mismatches[110–112].

The differences in terminase proteins have been previously studied to show the specific functions within a T4-like bacteriophage. Terminases are used to package viral DNA, with the small terminase initiating the packaging and the large terminase helping with the ATP-powered translocation of DNA[113]. This difference could be what makes some clusters more effective than others and utilizing that knowledge could help create an effective phage therapy tool.

Identification of phage protein function is instrumental in creating a safe effective phage therapy treatment. However, there is still a present gap in our knowledge of phage proteins, with 69% of proteins having no identifiable function. Without the knowledge of what these proteins are doing, researchers cannot guarantee that they are safe or necessary for phage infection.

Building a database of phage proteins and examining them for trends and statistical differences will help determine what aspects of phage are necessary for infection. Once key differences such as the DNA polymerase and terminase are identified, they can be further verified and tested with wet lab techniques. Mass spectrometry has been used before to identify proteins during the infection cycle and view other phage proteins in vitro[114–116]. It could confirm the proteins that are most important and prevalent during infection.

Genetic engineering techniques could be used to test the efficacy of phage when proteins are deleted, up-regulated, or down-regulated. Techniques involving knocking out genes to determine the least number of genes required for function have been employed in the past to better understand bacteria[117,118]. The same principles can be used on phage genomes to test any key proteins identified using statistical analysis. Phage promoter sequences are being identified through wet lab research and this knowledge could be used in the creation of a phage therapy product with the key proteins produced more and those unnecessary produced less[119–122].

Figure 5.1 below shows a proposed process for filling the gap in phage knowledge and building safe and effective phage treatments.

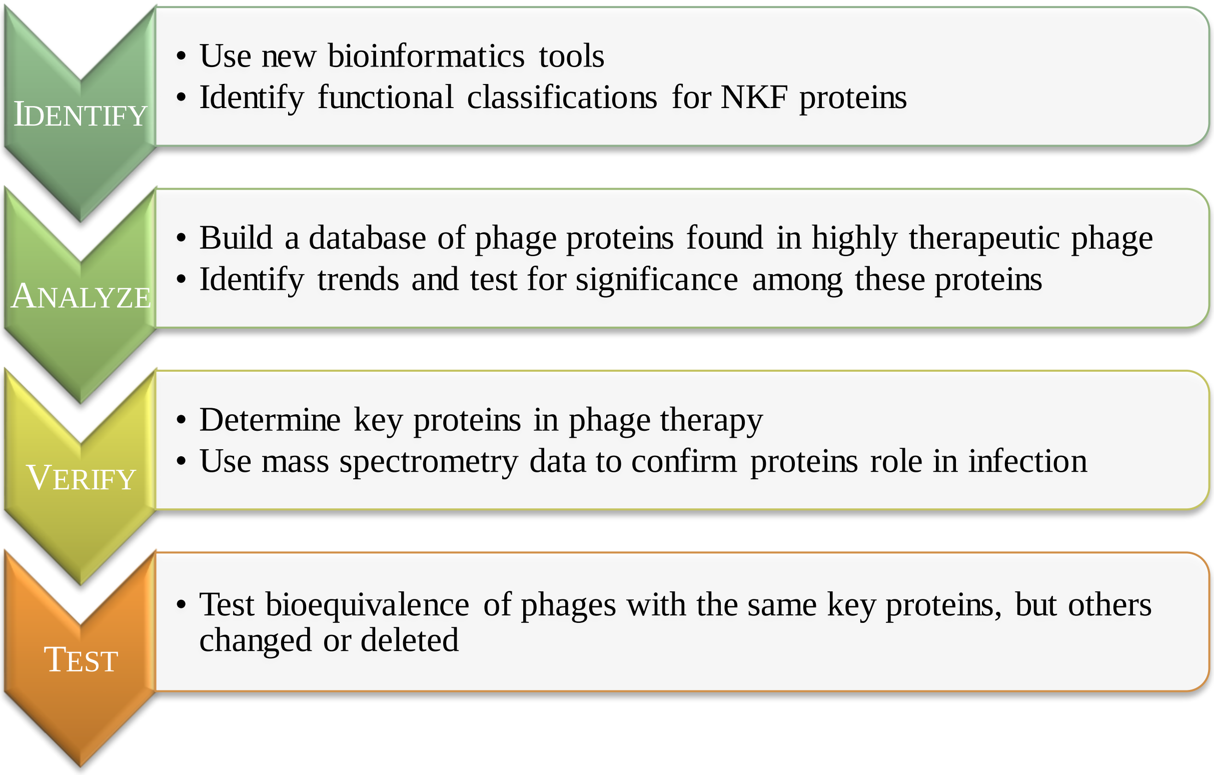

IDENTIFY
• Use new bioinformatics tools
• Identify functional classifications for NKF proteins
ANALYZE
• Build a database of phage proteins found in highly therapeutic phage
• Identify trends and test for significance among these proteins
VERIFY
• Determine key proteins in phage therapy
• Use mass spectrometry data to confirm proteins role in infection
TEST
• Test bioequivalence of phages with the same key proteins, but others changed or deleted

Figure 5.1. A possible process for creating safe phage therapies that could be regulated as biologics by the FDA.

The method proposed here can offer new insights into phage function, taking the number of NKF proteins down by 30%. The newly identified proteins were mainly structural proteins, but also included were proteins related to DNA Replication/Translation and one protein related to

Lysis (a holin). This information adds to the database of knowledge available on phages and outlines a future process that others could use. The amount of NKF proteins is no longer the majority of PotatoSplit's genome, which is a step forwards in phage functional annotation.

The two proteins studied with structure predictions and simulations also show promising results on possible functions. The lysis cassette protein has strong results marking it as a lyase. The specifics of the function may still be unknown, but even a family classification is more than was previously known about this protein. The structural comparison results showed similarity to a reductase and a hydrolase, both members of the lyase classification. The simulation results also shaped similar behavior to a hydrolase. Based on this, the function of this protein is a lyase.

The replication cassette protein GO results prove that it is related to DNA replication in some way, probably through transcription. The classifications listed are hydrolase options. Hydrolysis plays a key role in the transcription of DNA, which matches these results. The structural comparisons also show some similarity to a hydrolase, while the simulation results show some similar behavior to the known hydrolase. It can be said that this protein is in the hydrolase family.

This validates the belief that proteins near each other have similar functions. The replication and lysis cassette proteins selected did have families related to those functions. Examining proteins near those of known function can lead to more information about them and the function of proteins.

Based on these results, molecular dynamics and structural comparisons can be a strong method of identifying the functions of mycobacteriophage proteins. In the future, more known proteins could be simulated and their behaviors compared with the NKF proteins. Monte Carlo simulations could be added as a first step to minimize the energy of the protein structures and guarantee they are in the most statistically likely form. Additionally, the RMSD can be used to determine the stability of a protein at various temperatures, which can be used to help identify the protein based on thermodynamic stability.

Overall, this analysis shows that statistical methods can be used to analyze significant phage proteins and the newly proposed bioinformatics method can help determine functions for NKF proteins.

5.1.2 Case Studies

The case studies are done to show the validity of the statistical process on existing drug products.

The study done here is only on complex and epilepsy drugs, meaning the data here does not correlate to every product on the market. The risk factors outlined here in terms of excipients, the complexity of recipes, and process steps would need further wet-lab research to confirm any direct causation. Still, they show a significant correlation with being recalled. Additionally, this research compares drug products recalled for one reason against those recalled for other reasons. There is no data within this dataset for products that were never recalled.

Complex drugs show a significant positive association with being recalled for defective delivery and stability. Defective delivery can be explained by the recall of transdermal drugs for this reason. Transdermal patch drugs are at risk for formulation errors when applying the drug and adhesive, shown by the significance of the steps “spray coating” and “laminating.” Previously it was shown that transdermal and gel products are less likely to be recalled for formulation issues, but this study shows that it depends instead on the recall reason.

Many ingredients had effects on drug recalls. Impurities and degradation are more likely in products containing magnesium stearate, corn starch, sodium citrate, and less likely products containing sodium hydroxide. Several ingredients are less likely to be found in products recalled for defective delivery, and much of this can be explained by dosage form. Products recalled for Defective Delivery are often transdermal drugs and thus do not contain water or the other excipients listed in Table 4.5.

Citric acid is the only significant ingredient for those products recalled for stability and shows a positive correlation. Citric acid is also positively significant for superpotency drug recalls, as is Microcrystalline cellulose, glycerin, and sodium benzoate. Sodium hydroxide is the only significant ingredient for the presence of particulate matter, also showing a positive correlation. Glycerin, positively significant for superpotency, also positively affects subpotency, which could be due to glycerin being used as a solvent in gel products. Propylene glycol was also positively significant for subpotency. Dissolution specifications recalls have nine positive significant excipients shown in Table 4.5. Many excipients can cause disruptions in the dissolution of the active ingredient.

The number of steps was only significant for superpotency, showing that having more steps increases the likelihood to be recalled for superpotency. The number of steps did not affect any of the other recall reasons, though it was expected that formulations with fewer steps were less likely to be recalled. This could be because this study compared drugs recalled for different reasons against each other, and future work could be done by adding more data to the set.

The effect of the number of excipients was unexpected, with a high number of excipients correlating with a reduced likelihood for recall for defective delivery and presence of particulate matter. For defective delivery, this could be because transdermal drugs must use many adhesives and excipients to guarantee the drug reservoir adheres to the backing. Similarly, the likelihood of particulate matter recalls may be reduced if the product has more binding excipients. Being recalled for dissolution specifications is directly related to the number of excipients, which follows logically as the active ingredient's dissolution could be interrupted by excipients.

The use of a filter, sieve, blender, or granulation step during formulation is less likely in products being recalled for Defective Delivery, again perhaps due to the transdermal products that dominant the Defective Delivery recall category. The use of a fluidized bed dryer, spray coating, or laminating is more likely in products being recalled for Defective Delivery. It is more likely to be recalled for Superpotency if granulation is used. For the presence of particulate matter, the use of a filter shows a significant positive correlation to being recalled. This is likely due to the dosage form; most products recalled for the presence of particulate matter are liquid and thus use filtering in their formulations. Granulation and fluidized bed drying have a negative significant effect on the presence of particulate matter. Using a fluidized bed dryer also has a negative effect on products recalled for dissolution specifications.

The hard step "Formed a Gel Matrix" is more likely for drugs recalled for Superpotency and Subpotency, most likely because gels can become super or subpotent more easily than other dosage forms. "Formed particles" is also significant for Superpotency, due to processes that use microparticles within the product. "Applied to Backing" is more likely to be recalled for Defective Delivery, and "Formed Emulsion" is more likely to be recalled for Dissolution Specifications.

The results of the statistical analysis are the quality ranking applied to the drug products. The highest risk products should be closely monitored for safety issues and their formulations can be evaluated. Any high-risk excipients could be substituted for a lower-risk excipient.

Additionally, the two case studies of FAERS had conflicting results. Daytrana returned significant results, possibly due to the confounding variable of a lawsuit being in progress since 2015[123]. Consumers had a motivation to go and file reports on their adverse event effects, which could be why consumers were the only significant report source and not physicians or other health care professionals. The fact that Noven was the only manufacturer for Daytrana due to it being a new product could also affect results. The reports were not split between manufacturers, and the reports for Noven vastly outnumbered any other company.

In contrast, Gabapentin had over 100 manufacturers. The reports were split between each manufacturer, and the accuracy of the manufacturer reported cannot be guaranteed. The points that had no manufacturer assigned can also cause inaccurate results.

For the FAERS data to yield significant results and be a powerful tool, the manufacturers need to be recorded on each report and the dates the events took place need to be included. The points without dates are of no use in the safety surveillance. They could have taken place years ago by the time they reached the FDA reporting system, which was a trend seen in the data. The FDA Reporting Date often differed from the Event Date by six months or more, if the Event Date was included at all. Without careful recording of all details on the event reports, the data will continue to yield no significant results.

5.2 Conclusions

5.2.1 Bacteriophage Research

With the newly proposed method, the results show that HHBlits and TMHMM are highly informative programs that should be included in phage annotation processes. The use of this new process can cut down on the number of NKF proteins by 30%. Every piece of knowledge helps when determining the safety and efficacy of phages. The current process of phage annotation has been stagnant and still outputs phages with ~70% NKF genes. While more phages are sequenced every year, there is still a need for more methods of annotation and new bioinformatics tools.

The structure prediction and simulation analysis poses a new front for phage protein research. While these steps are too time-consuming to be used on every phage protein, they can be used to provide more details into a select number of proteins. Modeling the structure of an amino acid sequence offers new insight into the possible functions, and the information given by I-TASSER in the GO terms and PDB models matched the expected results based on protein cassette.

Once more proteins are identified, the statistical analysis carried out will be more informative on identifying proteins significant to therapeutically equivalent phages.

5.2.2 Case Studies

The results and discussion above show the risk factors for excipients and manufacturing information for non-biological complex drugs and epilepsy drugs. The data shown in this paper is the beginning of a shift to focus on the quality of products. Using this data, the industry can begin its formulation design with knowledge of the risk factors for different quality issues.

The FAERs data shows the importance of data quality. The lack of consistent reporting within the FAERs system leads to inadequate data to conclude. Applying these principles to phage research will allow for the collection of high-quality data and the ability to draw significant conclusions on the therapeutic equivalence of phages.

The high-risk products identified here should be monitored for future quality issues. Their formulations may need to be changed to ensure higher quality products with low-risk excipients.

These case studies show the application of statistics to identify key factors in drug manufacturing and provide a strong foundation for how this may be applied to phage proteins in future applications.

5.3 Future Recommendations

5.3.1 Bacteriophage Research

Phage therapy becomes a viable option for treating bacterial infections with a clear understanding of phage proteins and how they relate to the safety, efficacy, and regulations of a pharmaceutical product. The answer to moving phage therapy forwards does not lie solely in one discipline, but in a combination of computational biology, synthetic biology, and regulatory science. The proposal outlined here requires future work to be done in conjunction with wet lab research to confirm findings and finalize a path forward.

New bioinformatics methods need to be implemented to produce high-quality phage data that can be used for safe phage therapy treatments. Reliance on annotation methods and databases for bacteria and eukaryote data lead to many phage proteins of No Known Function. Phage-specific methods need to be developed to lead to a better understanding of phage research. Structure prediction, simulations, and machine learning algorithms are promising fields for phage protein bioinformatics.

5.3.2 Case Studies

Future work in this area will explore the correlations with products that were never recalled. For example, a comparison of transdermal products recalled and not recalled could allow a better understanding of what adhesive excipients are better to use. With more knowledge on leading indicators and root causes of drug quality recalls, more products can be manufactured to have high quality that will keep consumers safe and save manufacturers future FDA recalls. Additionally, suggestions are being made on the improvement of the FAERS in hopes future data could be used for safety surveillance. Specifically, the dates and manufacturers are the most important pieces of information that need to be logged.

Furthermore, the products identified as the highest risk for recall issues can be further studied in how their formulations can be varied to lower their risk level. These quality rankings should be applied to products that were never recalled to evaluate their risk and determine if they can be used to accurately show the quality of a drug.

APPENDIX A

Table A-1. Proteins that significant to temperate phage, sorted by functional category. Those with positive estimates are significantly present in temperate phages, while those with negative estimates are significantly absent in temperate phages.

PROTEIN	ESTIMATE	STD. ERROR	CHI^2	P-VALUE
DNA Replication				
Terminase small subunit	3.615	1.5999	11.7222	0.0006
Terminase	-3.3718	1.6036	9.8746	0.0016
ClP like protease	2.3978	1.6733	4.0093	0.0452
DNA Polymerase III Subunit	2.3978	1.6733	4.0093	0.0452
DNA Primase	2.6559	1.6442	5.2806	0.0215
DNA helicase	-2.8462	1.0962	9.8994	0.0016
DNAb like sDNA helicase	3.1354	1.6117	8.1998	0.0041
Cas4 Family Endonuclease	2.8991	1.6247	6.6740	0.0097
Queuine tRNA ribosyltransferase	-3.6635	1.6633	10.990	0.0009
Structure				
Capsid maturation protease and MuF like fusion protein	-2.5024	1.0937	7.4632	0.0063
Scaffold Protein	2.6835	1.0544	9.4461	0.0021
Head to Tail Stopper	3.6703	1.1579	16.067	6.11E-5
Tail Terminator	4.3838	1.6573	16.925	3.88E-4
Other				
Metallophosphoesterase	2.8991	1.6247	6.6740	0.0097
O-methyltransferase	-2.9856	1.7263	6.2731	0.01225
Adenylate Kinase	-3.3294	1.6859	8.2607	0.0035
RuvC like resolvase	-4.0073	1.6538	13.773	0.0002
ParB like nuclease domain protein	-2.6092	1.7981	4.2702	0.0387
NrdH like gutaredoxin	2.8991	1.6247	6.6740	0.0097
DNA Integration				
Integrase	4.1571	1.6072	16.0853	6.05E-5
Immunity Repressor	4.8933	1.6556	21.813	3.00E-6

REFERENCES

1. Poxleitner M, Pope W, Jacobs-Sera D, Sivanathan V, Hatful G. *Phage Discovery Guide*. Howard Hughes Medical Institute Accessed January 24, 2020. https://seaphagesphagediscoveryguide.helpdocsonline.com/3-0-overview

2. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinforma Oxf Engl*. 2007;23(6):673-679. doi:10.1093/bioinformatics/btm009

3. Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res*. 2005;33(Web Server issue):W451-454. doi:10.1093/nar/gki487

4. Cresawn SG, Bogel M, Day N, Jacobs-Sera D, Hendrix RW, Hatfull GF. Phamerator: a bioinformatic tool for comparative bacteriophage genomics. *BMC Bioinformatics*. 2011;12(1):395. doi:10.1186/1471-2105-12-395

5. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*. 2004;32(Web Server issue):W20-W25. doi:10.1093/nar/gkh435

6. The HHpred interactive server for protein homology detection and structure prediction. Accessed June 8, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1160169/

7. More People in the United States Dying from Antibiotic-Resistant Infections than Previously Estimated | CDC Online Newsroom | CDC. Published December 4, 2019. Accessed July 1, 2020. https://www.cdc.gov/media/releases/2019/p1113-antibiotic-resistant.html

8. Carlet J. The world alliance against antibiotic resistance: consensus for a declaration. *Clin Infect Dis Off Publ Infect Dis Soc Am*. 2015;60(12):1837-1841. doi:10.1093/cid/civ196

9. Duplessis CA, Stockelman M, Hamilton T, et al. A Case Series of Emergency Investigational New Drug Applications for Bacteriophages Treating Recalcitrant Multi-drug Resistant Bacterial Infections: Confirmed Safety and a Signal of Efficacy. *J Intensive Crit Care*. 2019;5(2). Accessed April 5, 2020. https://criticalcare.imedpub.com/abstract/a-case-series-of-emergency-investigational-new-drug-applications-for-bacteriophages-treating-recalcitrant-multidrug-resistant-bacterial-infections-confirmed-safety-and-a-signal-of-efficacy-24591.html

10. Golkar Z, Bagasra O, Pace DG. Bacteriophage therapy: a potential solution for the antibiotic resistance crisis. *J Infect Dev Ctries*. 2014;8(02):129-136. doi:10.3855/jidc.3573

11. Nale JY, Redgwell TA, Millard A, Clokie MRJ. Efficacy of an Optimised Bacteriophage Cocktail to Clear Clostridium difficile in a Batch Fermentation Model. *Antibiotics*. 2018;7(1):13. doi:10.3390/antibiotics7010013

12. Fong SA, Drilling AJ, Ooi ML, et al. Safety and efficacy of a bacteriophage cocktail in an in vivo model of Pseudomonas aeruginosa sinusitis. *Transl Res*. 2019;206:41-56. doi:10.1016/j.trsl.2018.12.002

13. Hatfull GF. The Secret Lives of Mycobacteriophages. In: Łobocka M, Szybalski WT, eds. *Advances in Virus Research*. Vol 82. Bacteriophages, Part A. Academic Press; 2012:179-288. doi:10.1016/B978-0-12-394621-8.00015-7

14. Hatfull GF. Mycobacteriophages: Windows into Tuberculosis. *PLOS Pathog*. 2014;10(3):e1003953. doi:10.1371/journal.ppat.1003953

15. Hatfull GF, the Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science Program, the KwaZulu-Natal Research Institute for Tuberculosis and HIV Mycobacterial Genetics Course Students, the Phage Hunters Integrating Research and Education Program. Complete Genome Sequences of 138 Mycobacteriophages. *J Virol*. 2012;86(4):2382-2384. doi:10.1128/JVI.06870-11

16. Caruso SM, Sandoz J, Kelsey J. Non-STEM Undergraduates Become Enthusiastic Phage-Hunters. *CBE—Life Sci Educ*. 2009;8(4):278-282. doi:10.1187/cbe.09-07-0052

17. Hanauer DI, Graham MJ, Sea-Phages, et al. An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning. *Proc Natl Acad Sci*. 2017;114(51):13531-13536. doi:10.1073/pnas.1718188115

18. Sharp R. Bacteriophages: biology and history. *J Chem Technol Biotechnol*. 2001;76(7):667-672. doi:10.1002/jctb.434

19. Stern A, Sorek R. The phage-host arms-race: Shaping the evolution of microbes. *Bioessays*. 2011;33(1):43-51. doi:10.1002/bies.201000071

20. Hatfull GF, Hendrix RW. Bacteriophages and their genomes. *Curr Opin Virol*. 2011;1(4):298-303. doi:10.1016/j.coviro.2011.06.009

21. Duckworth DH, Gulig PA. Bacteriophages: Potential Treatment for Bacterial Infections. *BioDrugs*. 2002;16(1):57-62. doi:10.2165/00063030-200216010-00006

22. Duckworth DH, Gulig PA. Bacteriophages. *BioDrugs*. 2002;16(1):57-62. doi:10.2165/00063030-200216010-00006

23. Pedulla ML, Ford ME, Houtz JM, et al. Origins of Highly Mosaic Mycobacteriophage Genomes. *Cell*. 2003;113(2):171-182. doi:10.1016/S0092-8674(03)00233-2

24. Wittebole X, De Roock S, Opal SM. A historical overview of bacteriophage therapy as an alternative to antibiotics for the treatment of bacterial pathogens. *Virulence*. 2014;5(1):226-235. doi:10.4161/viru.25991

25. Pope WH, Bowman CA, Russell DA, et al. Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. Kolter R, ed. *eLife*. 2015;4:e06416. doi:10.7554/eLife.06416

26. Ijaq J, Chandrasekharan M, Poddar R, Bethi N, Sundararajan VS. Annotation and curation of uncharacterized proteins- challenges. *Front Genet*. 2015;6. doi:10.3389/fgene.2015.00119

27. McKay T, Hart K, Horn A, et al. Annotation of proteins of unknown function: initial enzyme results. *J Struct Funct Genomics*. 2015;16(1):43-54. doi:10.1007/s10969-015-9194-5

28. Eisenstein E, Gilliland GL, Herzberg O, et al. Biological function made crystal clear — annotation of hypothetical proteins via structural genomics. *Curr Opin Biotechnol*. 2000;11(1):25-30. doi:10.1016/S0958-1669(99)00063-4

29. Islam MdS, Shahik SMd, Sohel Md, Patwary NIA, Hasan MdA. In Silico Structural and Functional Annotation of Hypothetical Proteins of Vibrio cholerae O139. *Genomics Inform*. 2015;13(2):53-59. doi:10.5808/GI.2015.13.2.53

30. Kumar K, Prakash A, Anjum F, Islam A, Ahmad F, Hassan MdI. Structure-based functional annotation of hypothetical proteins from Candida dubliniensis: a quest for potential drug targets. *3 Biotech*. 2015;5(4):561-576. doi:10.1007/s13205-014-0256-3

31. Naqvi AAT, Shahbaaz M, Ahmad F, Hassan MI. Identification of Functional Candidates amongst Hypothetical Proteins of Treponema pallidum ssp. pallidum. *PLOS ONE*. 2015;10(4):e0124177. doi:10.1371/journal.pone.0124177

32. Naveed M, Tehreem S, Usman M, Chaudhry Z, Abbas G. Structural and functional annotation of hypothetical proteins of human adenovirus: prioritizing the novel drug targets. *BMC Res Notes*. 2017;10(1):706. doi:10.1186/s13104-017-2992-z

33. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*. 2003;31(13):3784-3788. doi:10.1093/nar/gkg563

34. Bernardes J, Pedreira C. A Review of Protein Function Prediction Under Machine Learning Perspective. *Recent Pat Biotechnol*. 2013;7. doi:10.2174/18722083113079990006

35. Bonetta R, Valentino G. Machine learning techniques for protein function prediction. *Proteins Struct Funct Bioinforma*. 2020;88(3):397-413. doi:https://doi.org/10.1002/prot.25832

36. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 2001;305(3):567-580. doi:10.1006/jmbi.2000.4315

37. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem Sci*. 1998;23(11):444-447. doi:10.1016/S0968-0004(98)01298-5

38. HMMER web server: interactive sequence similarity searching. Accessed July 1, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125773/

39. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment | Nature Methods. *Nature*. Published online 2012. Accessed June 8, 2020. https://www.nature.com/articles/nmeth.1818

40. STRING: a database of predicted functional associations between proteins | Nucleic Acids Research | Oxford Academic. Accessed March 26, 2021. https://academic.oup.com/nar/article/31/1/258/2401231?login=true

41. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res*. 2015;43(W1):W174-W181. doi:10.1093/nar/gkv342

42. Sampaio M, Rocha M, Oliveira H, Dias O. Predicting Promoters in Phage Genomes Using Machine Learning Models. In: Fdez-Riverola F, Rocha M, Mohamad MS, Zaki N, Castellanos-Garzón JA, eds. *Practical Applications of Computational Biology and Bioinformatics, 13th International Conference*. Advances in Intelligent Systems and Computing. Springer International Publishing; 2020:105-112.

43. Ali M, Taniza FA, Niloy AR, Saha S, Shatabda S. Prediction of Bacteriophage Protein Locations Using Deep Neural Networks. In: Abraham A, Dutta P, Mandal JK, Bhattacharya A, Dutta S, eds. *Emerging Technologies in Data Mining and Information Security*. Advances in Intelligent Systems and Computing. Springer; 2019:29-38. doi:10.1007/978-981-13-1951-8_4

44. Cheng J-H, Yang H, Liu M-L, et al. Prediction of bacteriophage proteins located in the host cell using hybrid features. *Chemom Intell Lab Syst*. 2018;180:64-69. doi:10.1016/j.chemolab.2018.07.006

45. Bileschi ML, Belanger D, Bryant D, et al. Using Deep Learning to Annotate the Protein Universe. *bioRxiv*. Published online July 15, 2019:626507. doi:10.1101/626507

46. Nguyen CD, Gardiner KJ, Cios KJ. Protein annotation from protein interaction networks and Gene Ontology. *J Biomed Inform*. 2011;44(5):824-829. doi:10.1016/j.jbi.2011.04.010

47. Bonetta R, Valentino G. Machine learning techniques for protein function prediction. *Proteins Struct Funct Bioinforma*. 2020;88(3):397-413. doi:10.1002/prot.25832

48. Generic Drugs Undergo Rigorous FDA Scrutiny | FDA. Accessed July 1, 2020. https://www.fda.gov/consumers/consumer-updates/generic-drugs-undergo-rigorous-fda-scrutiny

49. Meredith PA. Generic Drugs. *Drug Saf*. 1996;15(4):233-242. doi:10.2165/00002018-199615040-00001

50. Frank RG. The Ongoing Regulation of Generic Drugs. *N Engl J Med*. 2007;357(20):1993-1996. doi:10.1056/NEJMp078193

51. Kelleher KR. FDA Approval of Generic Biologics: Finding a Regulatory Pathway. Published online 2007:21.

52. Nicholas JM. Complex Drugs and Biologics: Scientific and Regulatory Challenges for Follow-on Products. *Drug Inf J DIJ Drug Inf Assoc*. 2012;46(2):197-206. doi:10.1177/0092861512437759

53. Schwerin A von, Stoff H, Wahrig B, eds. *Biologics: A History of Agents Made from Living Organisms in the Twentieth Century*. Pickering & Chatto; 2013.

54. Morgan MR. Regulation of Innovation under Follow-On Biologics Legislation: FDA Exclusivity as an Efficient Incentive Mechanisms. *Columbia Sci Technol Law Rev*. 2010;11:93. https://heinonline.org/HOL/Page?handle=hein.journals/cstlr11&id=93&div=&collection=

55. Sachs RE. Innovation Law and Policy: Preserving the Future of Personalized Medicine. 49:60.

56. Hamburg MA, Collins FS. The Path to Personalized Medicine. *N Engl J Med*. 2010;363(4):301-304. doi:10.1056/NEJMp1006304

57. Fauconnier A. Phage Therapy Regulation: From Night to Dawn. *Viruses*. 2019;11(4):352. doi:10.3390/v11040352

58. Schooley RT, Biswas B, Gill JJ, et al. Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection. *Antimicrob Agents Chemother*. 2017;61(10):e00954-17, e00954-17. doi:10.1128/AAC.00954-17

59. Hall K, Stewart T, Chang J, Freeman MK. Characteristics of FDA drug recalls: A 30-month analysis. *Am J Health Syst Pharm*. 2016;73(4):235-240. doi:10.2146/ajhp150277

60. Wang B, Gagne JJ, Choudhry NK. The Epidemiology of Drug Recalls in the United States. *Arch Intern Med*. 2012;172(14):1110-1111. doi:10.1001/archinternmed.2012.2013

61. Nagaich U, Sadhna D. Drug recall: An incubus for pharmaceutical companies and most serious drug recall of history. *Int J Pharm Investig*. 2015;5(1):13-19. doi:10.4103/2230-973X.147222

62. The therapeutic equivalence of complex drugs - ScienceDirect. Accessed July 7, 2020. https://www-sciencedirect-com.ezproxy.lib.purdue.edu/science/article/pii/S0273230010001893?casa_token=3CBvEjkKJoEAAAAA:_BWTZoMNobQAAjtaFFkSdWe4J5SYZdFoZnsfCwykN5CieNpiEhrHgzDzzdx-Tp9kmoAjEB3IcQ

63. Schellekens H, Stegemann S, Weinstein V, et al. How to Regulate Nonbiological Complex Drugs (NBCD) and Their Follow-on Versions: Points to Consider. *AAPS J*. 2014;16(1):15-21. doi:10.1208/s12248-013-9533-z

64. Jiang X. Introduction to Complex Products and FDA Considerations. Presented at the: August 6, 2017.

65. Wichman K, U D. Overdose a Risk of Transdermal Patch in Diverse Settings: Problems Occur Even with Discarded Patch. *Can Pharm J Rev Pharm Can*. 2005;138(7):65-66. doi:10.1177/171516350513800709

66. Lampert A, Seiberth J, Haefeli WE, Seidling HM. A systematic review of medication administration errors with transdermal patches. *Expert Opin Drug Saf*. 2014;13(8):1101-1114. doi:10.1517/14740338.2014.926888

67. Wokovich AM, Prodduturi S, Doub WH, Hussain AS, Buhse LF. Transdermal drug delivery system (TDDS) adhesion as a critical safety, efficacy and quality attribute. *Eur J Pharm Biopharm*. 2006;64(1):1-8. doi:10.1016/j.ejpb.2006.03.009

68. Greenall J, Koczmara C, Cheng R, Hyland S. Safety Issues with Fentanyl Patches Require Pharmaceutical Care. 2008;61(1):3.

69. Hesse S, Adhya S. Phage Therapy in the Twenty-First Century: Facing the Decline of the Antibiotic Era; Is It Finally Time for the Age of the Phage? *Annu Rev Microbiol*. 2019;73(1):155-174. doi:10.1146/annurev-micro-090817-062535

70. Kincaid R. Treatment and Prevention of Bacterial Infections Using Bacteriophages: Perspectives on the Renewed Interest in the United States. In: Górski A, Międzybrodzki R, Borysowski J, eds. *Phage Therapy: A Practical Approach*. Springer International Publishing; 2019:169-187. doi:10.1007/978-3-030-26736-0_7

71. Hauser AR, Mecsas J, Moir DT. Beyond Antibiotics: New Therapeutic Approaches for Bacterial Infections. Weinstein RA, ed. *Clin Infect Dis*. 2016;63(1):89-95. doi:10.1093/cid/ciw200

72. Russ ZN. Synthetic biology: enormous possibility, exaggerated perils. *J Biol Eng*. 2008;2(1):7. doi:10.1186/1754-1611-2-7

73. Gladstone EG, Molineux IJ, Bull JJ. Evolutionary principles and synthetic biology: avoiding a molecular tragedy of the commons with an engineered phage. *J Biol Eng*. 2012;6(1):13. doi:10.1186/1754-1611-6-13

74. Liu Y, Huang H, Wang H, Zhang Y. A novel approach for T7 bacteriophage genome integration of exogenous DNA. *J Biol Eng*. 2020;14(1):2. doi:10.1186/s13036-019-0224-x

75. Chan B, Turner P, Kim S, Mojibian H, Elefteriades J, Narayan D. Phage treatment of an aortic graft infected with Pseudomonas aeruginosa. *Evol Med Public Health*. 2018;2018(1):60-66. Accessed July 1, 2020. https://academic.oup.com/emph/article/2018/1/60/4923328

76. Hajimorad M, Gray PR, Keasling JD. A framework and model system to investigate linear system behavior in Escherichia coli. *J Biol Eng*. 2011;5(1):3. doi:10.1186/1754-1611-5-3

77. Chamakura K, Young R. Phage single-gene lysis: Finding the weak spot in the bacterial cell wall. *J Biol Chem*. 2019;294(10):3350-3358. doi:10.1074/jbc.TM118.001773

78. Hatfull GF. Mycobacteriophages: Windows into Tuberculosis. *PLoS Pathog*. 2014;10(3). doi:10.1371/journal.ppat.1003953

79. Jacobs-Sera D, Marinelli LJ, Bowman C, et al. On the nature of mycobacteriophage diversity and host preference. *Virology*. 2012;434(2):187-201. doi:10.1016/j.virol.2012.09.026

80. Cooper CJ, Khan Mirzaei M, Nilsson AS. Adapting Drug Approval Pathways for Bacteriophage-Based Therapeutics. *Front Microbiol*. 2016;7. doi:10.3389/fmicb.2016.01209

81. Garber K. First microbiome-based drug clears phase III, in clinical trial turnaround. *Nature*. Published online September 11, 2020. Accessed March 26, 2021. https://www.nature.com/articles/d41573-020-00163-4

82. U.S. National Library of Medicine Clincal Trials. Clinical Trials. Accessed March 26, 2021. https://clinicaltrials.gov/ct2/home

83. Debarbieux L, Pirnay J-P, Verbeken G, et al. A bacteriophage journey at the European Medicines Agency. Millard A, ed. *FEMS Microbiol Lett*. 2016;363(2):fnv225. doi:10.1093/femsle/fnv225

84. Yang M, Byrn SR, Clase KL. An Analytic Investigation of the Drug Formulation-Based Recalls in the USA: See More Beyond the Literal. *AAPS PharmSciTech*. 2020;21(5):198. doi:10.1208/s12249-020-01726-9

85. Patil P, Joshi P, Paradkar A. Effect of formulation variables on preparation and evaluation of gelled self-emulsifying drug delivery system (SEDDS) of ketoprofen. *AAPS PharmSciTech*. 2004;5(3):43-50. doi:10.1208/pt050342

86. Subramanian N, Yajnik A, Murthy RSR. Artificial neural network as an alternative to multiple regression analysis in optimizing formulation parmaeters of cytarabine liposomes. *AAPS PharmSciTech*. 2009;5(1):11. doi:10.1208/pt050104

87. Rathore AS, Winkle H. Quality by design for biopharmaceuticals. *Nat Biotechnol*. 2009;27(1):26-34. doi:10.1038/nbt0109-26

88. Woodcock J. The Concept of Pharmaceutical Quality. *Am Pharm Rev*. 2004;7(6):10-15. Accessed June 30, 2020. https://www.researchgate.net/profile/Janet_Woodcock2/publication/279577343_The_concept_of_pharmaceutical_quality/links/5a69e01eaca2728d0f5f27f3/The-concept-of-pharmaceutical-quality.pdf

89. Yu LX. Pharmaceutical Quality by Design: Product and Process Development, Understanding, and Control. *Pharm Res*. 2008;25(4):781-791. doi:10.1007/s11095-007-9511-1

90. Riley BS, Li X. Quality by Design and Process Analytical Technology for Sterile Products—Where Are We Now? *AAPS PharmSciTech*. 2011;12(1):114-118. doi:10.1208/s12249-010-9566-x

91. Waterman KC. The Application of the Accelerated Stability Assessment Program (ASAP) to Quality by Design (QbD) for Drug Product Stability. *AAPS PharmSciTech*. 2011;12(3):932. doi:10.1208/s12249-011-9657-3

92. Singh R, Yuan Z. *Process Systems Engineering for Pharmaceutical Manufacturing*. Elsevier; 2018.

93. Gad SC. *Pharmaceutical Manufacturing Handbook: Production and Processes*. John Wiley & Sons; 2008.

94. Sittig M. *Pharmaceutical Manufacturing Encyclopedia*. Second Edition. Noyes Publications; 1988.

95. Qiu Y, Chen Y, Zhang GGZ, Liu L, Porter W. *Developing Solid Oral Dosage Forms: Pharmaceutical Theory and Practice*. Academic Press; 2009.

96. Lachman L, Lieberman HA, Kanig JL. *The Theory and Practice of Industrial Pharmacy*. Lea & Febiger; 1986.

97. Bose S, Bogner RH. Solventless Pharmaceutical Coating Processes: A Review. *Pharm Dev Technol*. 2007;12(2):115-131. doi:10.1080/10837450701212479

98. Commissioner O of the. Reporting Serious Problems to FDA. FDA. Published September 9, 2020. Accessed March 26, 2021. https://www.fda.gov/safety/medwatch-fda-safety-information-and-adverse-event-reporting-program/reporting-serious-problems-fda

99. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

100. Russell DA, Hatfull GF. PhagesDB: the actinobacteriophage database. *Bioinformatics*. 2017;33(5):784-786. doi:10.1093/bioinformatics/btw711

101. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers - ScienceDirect. Accessed March 23, 2021. https://www.sciencedirect.com/science/article/pii/S2352711015000059

102. *The PyMOL Molecular Graphics System*. Schrödinger, LLC

103. Aksyuk AA, Rossmann MG. Bacteriophage Assembly. *Viruses*. 2011;3(3):172-203. doi:10.3390/v3030172

104. Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. The many faces of the helix-turn-helix domain: Transcription regulation and beyond⋆. *FEMS Microbiol Rev*. 2005;29(2):231-262. doi:10.1016/j.fmrre.2004.12.008

105. Kala S, Cumby N, Sadowski PD, et al. HNH proteins are a widespread component of phage DNA packaging machines. *Proc Natl Acad Sci U S A*. 2014;111(16):6022-6027. doi:10.1073/pnas.1320952111

106. Structures and mechanisms of glycosyltransferases | Glycobiology | Oxford Academic. Accessed March 31, 2021. https://academic.oup.com/glycob/article/16/2/29R/592330

107. Lemon KP, Grossman AD. Localization of Bacterial DNA Polymerase: Evidence for a Factory Model of Replication. *Science*. 1998;282(5393):1516-1519. doi:10.1126/science.282.5393.1516

108. Fischetti VA. Development of Phage Lysins as Novel Therapeutics: A Historical Perspective. *Viruses*. 2018;10(6). doi:10.3390/v10060310

109. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. Accessed July 15, 2020. https://www.jbc.org/content/264/15/8935.short

110. Kelman Z, O'Donnell M. DNA POLYMERASE III HOLOENZYME: Structure and Function of a Chromosomal Replicating Machine. *Annu Rev Biochem*. 1995;64(1):171-200. doi:10.1146/annurev.bi.64.070195.001131

111. Kong X-P, Onrust R, O'Donnell M, Kuriyan J. Three-dimensional structure of the β subunit of E. coli DNA polymerase III holoenzyme: A sliding DNA clamp. *Cell*. 1992;69(3):425-437. doi:10.1016/0092-8674(92)90445-I

112. Wang J, Sattar AKMA, Wang CC, Karam JD, Konigsberg WH, Steitz TA. Crystal Structure of a pol α Family Replication DNA Polymerase from Bacteriophage RB69. *Cell*. 1997;89(7):1087-1099. doi:10.1016/S0092-8674(00)80296-2

113. Sun S, Gao S, Kondabagil K, Xiang Y, Rossmann MG, Rao VB. Structure and function of the small terminase component of the DNA packaging machine in T4-like bacteriophages. *Proc Natl Acad Sci U S A*. 2012;109(3):817-822. doi:10.1073/pnas.1110224109

114. Kang S, Hawkridge AM, Johnson KL, Muddiman DC, Prevelige PE. Identification of Subunit−Subunit Interactions in Bacteriophage P22 Procapsids by Chemical Cross-linking and Mass Spectrometry. *J Proteome Res*. 2006;5(2):370-377. doi:10.1021/pr050356f

115. Mouradian S, Rank DR, Smith LM. Analyzing Sequencing Reactions from Bacteriophage M13 by Matrix-assisted Laser Desorption/Ionization Mass Spectrometry. *Rapid Commun Mass Spectrom*. 1996;10(12):1475-1478. doi:10.1002/(SICI)1097-0231(199609)10:12<1475::AID-RCM696>3.0.CO;2-C

116. Poliakov A, Duijn E van, Lander G, et al. Macromolecular mass spectrometry and electron microscopy as complementary tools for investigation of the heterogeneity of bacteriophage portal assemblies. *J Struct Biol*. 2007;157(2):371-383. doi:10.1016/j.jsb.2006.09.003

117. Ostrov N, Landon M, Guell M, et al. Design, synthesis, and testing toward a 57-codon genome. *Science*. 2016;353(6301):819-822. doi:10.1126/science.aaf3639

118. Deutscher D, Meilijson I, Kupiec M, Ruppin E. Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet*. 2006;38(9):993-998. doi:10.1038/ng1856

119. Nesbit CE, Levin ME, Donnelly-Wu MK, Hatfull GF. Transcriptional regulation of repressor synthesis in mycobacteriophage L5. *Mol Microbiol*. 1995;17(6):1045-1056. doi:10.1111/j.1365-2958.1995.mmi_17061045.x

120. Jain S, Hatfull GF. Transcriptional regulation and immunity in mycobacteriophage Bxb1. *Mol Microbiol*. 2000;38(5):971-985. doi:10.1046/j.1365-2958.2000.02184.x

121. Garcia M, Pimentel M, Moniz-Pereira J. Expression of Mycobacteriophage Ms6 Lysis Genes Is Driven by Two σ70-Like Promoters and Is Dependent on a Transcription Termination Signal Present in the Leader RNA. *J Bacteriol*. 2002;184(11):3034-3043. doi:10.1128/JB.184.11.3034-3043.2002

122. Klucar L, Stano M, Hajduk M. phiSITE: database of gene regulation in bacteriophages. *Nucleic Acids Res*. 2010;38(suppl_1):D366-D370. doi:10.1093/nar/gkp911

123. Daytrana Lawsuits. Accessed March 26, 2021. https://www.aboutlawsuits.com/daytrana/