Designing for Resilience: Evaluating DNA Screening Vulnerabilities and Designing Mice for Improved Well-being

by

Rey Edison

B.Sc. in Biology
California Institute of Technology, 2016

Submitted to the Microbiology Graduate Program
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

at the

Massachusetts Institute of Technology

February 2025

Authored by: Rey Edison
Microbiology Graduate Program
December 23, 2024

Certified by: Kevin M. Esvelt
Associate Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by: Jacquin C. Niles
Professor of Biological Engineering
Co-Director of the Microbiology Graduate Program

Designing for Resilience: Evaluating DNA Screening Vulnerabilities and Designing Mice for Improved Well-being

by

Rey Edison

B.Sc. in Biology
California Institute of Technology, 2016

Submitted to the Microbiology Graduate Program
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

at the

Massachusetts Institute of Technology

February 2025

ABSTRACT

This thesis explores biosecurity and animal well-being through the lens of synthetic biology. The main project evaluates whether current synthesis screening systems for DNA synthesis orders are able to prevent the purchase of sequences that could be assembled into potential hazards. Orders for lightly camouflaged sequences designed to conceal fragments of potentially hazardous DNA were split across 38 commercial DNA synthesis providers, with 36 providers fulfilling the orders. Several providers later said they fulfilled the order because they did not know other portions of the hazard were ordered from different companies. Laboratory validation demonstrated that sequences with similar composition were readily assembled, highlighting the importance of future policy changes that make order screening systems resilient to both split-order attacks (where a hazard is split into fragments that can each be ordered from different providers to avoid detection) and future advances in synthetic biology. Two additional projects focus on designing new mouse strains for improved well-being in captivity. One mouse design utilizes multiplexed antiviral arrays for broad-spectrum viral protection and targeted disruption of the *faah* gene to elevate levels of endogenous anandamides to reduce pain and anxiety. The other design uses targeted gene expression of modified mu-opioid receptors designed to produce elevated signaling in specific neuronal regions associated with reward. These designs not only demonstrate novel synthetic biology applications but also raise perhaps unsettling questions about what it means to prioritize the well-being of captive mice. Specifically, the designs probe the question of whether external interventions like housing and enrichment adequately address the negative effects of adaptations that improve the chance of survival in the wild, especially when it may be possible to fine-tune or remove adaptations that cause mice suffering in captivity. This thesis balances technical descriptions with a discussion of broader implications for policy, industry, and ethics.

Thesis supervisor: Kevin M. Esvelt
Associate Professor of Media Arts and Sciences

Acknowledgments

In addition to thanking Kevin Esvelt for being an extraordinary advisor, I would like to express gratitude to all of my colleagues in the Esvelt Lab. Shay Toner was instrumental in the development of the DNA screening project. I particularly would like to highlight the contributions of Joanna Buchthal and Stephen Von Stetina, who provided feedback on both mouse design projects. In addition to providing input on mouse designs, Sebastian Kamau and Ceili Peng contributed to the early stages of the web lab work. While I didn't end up discussing gene drives in my thesis, my graduate studies also included extensive work that wouldn't have been possible without the support of my colleague Anna Clark, who continues to do excellent work in the field of gene drives and similar genetic methods. Of course, any errors and oversights in my thesis work are mine alone.

I also would like to thank those outside of biology for their contributions. Thank you to everyone in LC for your support and comradery. Danny, you have been an invaluable colleague. I would not be half the scientist or person I am today without you, and the portion I'd be missing would include the better half of each. Scott, I think you underestimate how much you help others, or at least you are incredibly modest about it. Lee, it's been thirteen years of being your friend, and here's to decades more. Haley, thank you for believing in me and helping me learn to believe in myself. Thank you also to Mia, Shruti, and Lena for advice and assistance whenever I needed it.

Lastly, this material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (2019287413). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

Preface

Right now, as I'm writing this, it is self-evident that I was among many graduate students whose PhD research was interrupted in an unprecedented way by the COVID-19 pandemic. At the risk of excessive optimism, I look forward to a day when a future reader might find this reminder necessary.[1] The rest of us remember waiting for the development of the first COVID-19 vaccines during a time when many were more aware than ever how much our lives depended on modern biotechnology. At the same time, living during a pandemic also raised awareness of the possibility that a future engineered pandemic could cause even more harm.

Today, I sit at my desk. There's a pile of boxes in a locked cabinet to my left. In the boxes are the orders from DNA synthesis companies containing the sequences my colleagues and I designed for the biosecurity project. Within them are tubes containing DNA fragments from ricin and 1918 flu hidden carefully within harmless sequences, including sequences from the strain of H1N1 used in flu vaccines. In the context of being used to cloak hazardous sequences, the fragments of H1N1 sequences are a chilling reminder of how the high degree of similarity shared by H1N1 strains makes it challenging to distinguish between hazardous and nonhazardous orders of synthetic DNA.

Tomorrow morning, at a health clinic a block from my office, I'll get my updated COVID-19 and flu vaccines. The flu shot will contain inactive mRNA from H1N1, including some of the same sequences I used to obfuscate the 1918 fragments. This winter, the close homology between H1N1 sequences will be a wonderful thing. When I encounter seasonal flu variants, the degree of similarity between them and the one my vaccine was based on will help my immune system reduce my likelihood of severe illness. I feel grateful to the scientists and healthcare workers who leveraged this homology to protect my health.

If readers take any message away from this thesis, it would be how some of the most exciting parts of synthetic biology share an uncanny resemblance with the most frightening. Disentangling these for maximum benefit is only possible with collaboration from many disciplines, including my favorite, evolutionary biology. As Dobzhansky titled his famous 1973 essay, it is indeed true "that nothing in biology makes sense except in the light of evolution."

[1] This is optimistic for a few reasons because it presupposes (1) civilization goes well for a while and at least one of the many people in a future generation actually happens to read a very old and niche PhD thesis or (2) civilization is irrevocably changed, but somehow my thesis is preserved and somehow still considered relevant. I'm not sure which is the better bet.

1. Overview

This thesis includes three separate projects I worked on during my PhD program. Because these projects are on different topics, the reader may experience some interdisciplinary and topical whiplash when moving between sections. Readers are encouraged to skim or skip sections outside of their primary areas of interest and return if they become relevant later.

Those most interested in biosecurity and existential risk will likely be most interested in 2. Creating Cryptic Sequences for Use in Biosecurity Evaluations, which also has the most discussion of bioinformatics and public policy. Those most interested in animal well-being may choose to start with section 3. A Genetic Approach to Upregulated Mu-Opioid Receptor Activity in Regions of the Mouse Brain Associated With Positive Experiences or section 4. A Genetic Approach to Reducing Pain and Anxiety and Increasing Resistance to Disease in Feeder Mice. Section 4 also has some policy discussion, while section 3 focuses more on ethics perspectives.

All three projects explore the importance of incorporating perspectives from evolutionary biology to improve designs. Section 2 emphasizes how an algorithm's performance may depend on the application, the database or training data, and the training set used to validate the algorithm. Section 3 highlights how the absence of a known selective pressure that favors mouse well-being suggests it might be possible to move the hedonic setpoint of mice in a positive direction. Next, section 4 discusses how targeting viruses multiple times reduces the chance that a virus will evade the anti-viral array by developing multiple mutations simultaneously. Lastly, the conclusion reviews and expands on these overarching themes from evolutionary biology.

↪ *Return to Table of Contents*

2. Creating Cryptic Sequences for Use in Biosecurity Evaluations

2.1 Project Summary

To assess the implementation of the current policies regarding DNA synthesis screening, hazardous sequences based on the U.S. Department of Health and Human Services list of select agents were ordered from commercial synthesis providers. Start codons were replaced to make each sequence untranscribable so that no actual hazards were created or mailed, but a bad actor would be able to recreate the start codon. The sequences created were obfuscated so that common bioinformatics tools such as NCBI BLAST would not return hazards among their top search results. However, the ordered sequences could be combined and treated with the appropriate enzymes to assemble them into their respective hazard sequences. At the time of ordering, United States policy recommended but did not require best practices for order screening that would include a request for third-party authorization based on the sequences ordered. 35 of 38 companies shipped all fragments ordered without any request, including several companies that later said they did so only because they did not know the remaining pieces were being ordered from other companies. This result suggests that the policies in place at the time of ordering may have been insufficient to prevent harm from malicious independent actors. Technical and thematic aspects of DNA classification and alignment are discussed, as well as the potential policy implications of our results.

2.2 Introduction and Background

2.2.1 Introduction

Modern biology research depends on commercial DNA synthesis services being cheap, convenient, and fast. Meanwhile, preventing the misuse of DNA synthesis services by bad actors has become a significant challenge for modern biosecurity policy. With oversight from the FBI, we ordered lightly disguised fragments from select agents that could be reassembled into a version of the select agent lacking a start codon. 35 of 38 companies shipped the orders without asking for follow-up documentation, including several that later said they did so only because they did not know the remaining pieces were being ordered from other companies. A set of equivalent harmless sequences was assembled in the lab to demonstrate that the ordered fragments could have been assembled. These results call into question whether policies that describe best practices but do not require any external auditing can adequately mitigate the risk of the order and misuse of hazardous sequences by bad actors.

2.2.1 DNA Screening as a Special Case of Sequence Classification

The word “obfuscate” implies adversarial design targeting one or more specific methods of detection. Before discussing the specific tools chosen as targets for this project, we will explore the broader meaning of classifying DNA sequences in biology. When algorithms are taken from other areas of biology for biosecurity use, they are often presented as generic all-purpose sequence search algorithms rather than algorithms that have been designed for and validated for specific uses on specific types of data

satisfying certain assumptions. This approach does a disservice to both biosecurity and evolutionary biology. To combat this, we begin this section with a discussion of related problems in evolutionary biology in order to identify some of the aspects of DNA order screening that make it especially complicated.

Classifying a nucleotide sequence can mean many things depending on both the context of the sequence and its intended use. Annotating a genome involves breaking long sequences into subregions that can be labeled with their known or predicted biochemical functions, like regulatory regions and coding regions. When analyzing a metagenome from an environmental sample, it is often relevant to identify which sequences are 16S rRNA genes, which can be used to reconstruct phylogenetic trees and calculate various measures of taxonomic diversity. For DNA synthesis, classifying a sequence may involve checking for features like GC content, repetitive sequences, or secondary structures that might make it challenging to synthesize. When using mutagenesis screening for forward genetics, classifying a nucleotide sequence may mean performing local or global alignments to determine what known sequences share homology around the mutation of interest.

Classifying or identifying a sequence is always a question of context and application, and the most appropriate tools to use can also vary widely based on the level of automation required and available knowledge about the source and function of the sequence. Assembling phylogenetic trees based on sequence data creates a hierarchical classification of sequences based on clades, which means sequence classification is an essential part of evolutionary biology. In this application, as in most applications, the question of what similarity measurement[2] to use is dependent both on the application and the nature of the data being examined.

While previous work in evolutionary biology has much insight to offer on identifying sequences from nature or experiments, there are far fewer tools designed for assessing sequences designed by synthetic biologists, which may be far less likely to follow assumptions that generally hold for sequences created by evolutionary processes. (Specifically, synthetic sequences are often chimeric or contain sequences not found in any natural organism.) By and large, the easiest way to figure out what a designed nucleotide sequence does is to see who made it and ask them to explain it. This project explores what challenges may arise when communication with a sequence's creator isn't possible or feasible.

When direct communication isn't possible, synthetic biologists often predict a sequence's purpose by looking at detailed sequence annotations that identify the different functional units of a piece. When a sequence is unannotated or under-annotated, it may be possible to break the sequence into short pieces, perform local alignments to annotate the sequence and create detailed sequence annotations.

[2] "Similarity" is used here instead of distance because, technically, not all measures of similarity are inverse distance metrics in the mathematical sense. For instance, using PAM or BLOSUM as models to weight edit-distance violates metric properties (Miranker et al., 2003). However, the E-values produced by BLAST are not technically metrics (Berger et al., 2020).

Many of the specific technical problems explored in this thesis would be avoided if it were mandatory to document nucleotide sequences with annotations before synthesis. However, verifying annotations may be challenging to scale and generalize.

This project describes the design and commercial purchasing of DNA sequences meant to have ambiguous identities or purposes. The main postulated assumptions being leveraged in the obfuscation approaches described in this project are listed below, along with indirect implications or qualifications:

1. The sequence is designed to be as close to its end use as possible.
    - No unnecessary assembly steps are created if they could be avoided by ordering a different sequence.
2. Sequences were designed with cost-effectiveness or ease of use in mind.
    - Unnecessary subsequences are removed from a sequence prior to order.
    - No obviously unnecessary sequences are added prior to order.
3. Sequences of similar size needed for the same project are usually ordered from as few DNA providers as possible for convenience.
    - Sometimes, researchers will order most sequences from the cheapest provider and any difficult-to-synthesis sequences from a more expensive provider. Alternatively, a lab may switch providers mid-project. However, the heart of this assumption is that at least one DNA synthesis provider receives a plurality of the orders for the same project.
4. Contiguous subsequences are related to each other if not in genome of origin, at least in some sort of intended synthetic biology application.

These are reasonable and widely applicable assumptions that would reflect typical research use of commercial DNA services. In many ways, these assumptions mirror assumptions in evolutionary biology used to calculate the evolutionary distance between sequences and reconstruct phylogenetic trees. Assumptions 1 and 2 relate to parsimony and selection pressure. Assumptions 3 and 4 relate to the assumption that sequences are transferred between parent and offspring, not through horizontal gene transfer or other means. Because these two sets of assumptions have so many similar implications, it would make sense for synthesis providers to choose sequence alignment algorithms originally designed to measure similarity between sequences in an evolutionary context. Additionally, evolutionary distance is by far the best-known measurement of similarity between two sequences. We decided that a commercial synthesis provider implicitly making these assumptions would choose to use one or more local or global sequence alignment tools for order screening. We used NCBI's BLAST as a target for our adversarial design.

From a bioinformatics perspective, one aim of this project is to design sequences that are difficult to classify using local and global alignment tools. This may be thought of as designing cryptic sequences or taking sequences and obfuscating their identity. While there are many alignment algorithms available, for this project, we chose to focus on a few popular tools available on NCBI's BLAST website: blastn and blastx.

Often, the names of these algorithms are used interchangeably with the NCBI web apps, which is not usually an issue. However, some implementations of these algorithms stop at finding and ranking High-sequence Segment Pairs (HSPs). For instance, BioPython's remote blast command (which contacts the NCBI servers) can provide a ranking of complete alignments, but the local blast command does not.[3] If the user wants a ranking of complete alignments, they need to implement this themselves because BioPython does not have one built in (Cock et al., 2009). While BioPerl has one, the documentation specifically states that they refused to identify a default option in the hopes that it would remind users that the optimal tiling always depends on the application (Tiling HOWTO).

Therefore, in this project, we specifically chose to focus on the implementations of blastn and tblastn searches as used in the NCBI web application in the summer of 2023. This distinction is meaningful because some of the obfuscation approaches may depend on aspects of the alignment algorithms themselves, how the HSPs are tiling, and even how final results are ranked and displayed in the NCBI web app. This choice is not intended to imply that BLAST, as commonly used, is a good choice for DNA order screening. While users can easily use NCBI BLAST's web app to search databases that include synthetic sequences, BLAST was not originally intended to classify or identify synthetic constructs, nor has it been validated for this use (Altschul et al., 1990; Pertsemlidis & Fondon, 2001).

2.2.2 Policy Context

In the 2010 HHS Guidance, which was active when the first round of orders began, the U.S. government recommended screening using a local sequence alignment technique and suggested the BLAST family of tools. However, they intentionally did not recommend specific threshold criteria. Instead, they outlined a "Best Match"[4] approach where:

> …the sequence with the greatest percent identity over each 66 amino acid or 200 bp fragment should be considered the "Best Match," regardless of the statistical

[3] Terminology may lead to confusion here. Some of the BioPython documentation for the blast commands use "alignment" to refer to the sequence corresponding to a highly scored gap-free local alignment, which is also called an High-similarity Sequence Pairing (HSP). However, the NCBI web app interface also uses "alignment" to refer to the regions of the query and subject sequences that consist of multiple non-overlapping HSPs. While both of these are technically alignments, they can also be different steps of the same alignment algorithm. For clarity, unless otherwise stated, "alignment" is used here to refer to the final result of the NCBI BLAST web app that can be scored for "Percent Identity."

[4] The Executive Order does not explicitly specify what type of database should be used for local alignment, which would have implications for how a local alignment algorithm is being used. It would also be possible to BLAST against a database solely consisting of hazards and choose a threshold of similarity for flagging sequences. However, because the phrase "Best Match" is being used, it seems implied that the order is referring to the use of a general database of nucleotide sequences, such as the core nucleotide database available from NCBI.

> significance or percent identity. (*Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA*, 2010)

This approach would have flagged the obfuscated sequences we ordered. However, as discussed later in 2.2.3, this guidance is only a recommendation. Moreover, there is an ambiguously large amount of room for provider discretion regarding identity verification and verifying legitimate use, so synthesizing and shipping the obfuscated sequences was not a violation of the 2010 HHS Guidance, which is again only a recommendation.

Some of the orders took place after the issue of Executive Order 14110 and the 2023 HHS Guidance written in support of the screening framing outlined by the executive order. The Executive Order did specify that by April 26, 2025, all entities receiving federal funding would be required to purchase synthetic nucleic acids from companies that adhered to the framework. Similar to the 2010 HHS Guidance, the 2023 Guidance says, “Screening [for SOCs] should take place over each 66 amino acid and/or 200 nucleotide window,” but also adds that “the screening should be applied across all applicable reading frames for nucleotide sequences that encode proteins (i.e., 3 for single-stranded nucleic acids and 6 for double-stranded nucleic acids).” The 2023 guidance also specifies additional practices providers must implement to adhere to the framework starting October 13, 2026:

> First, Providers should reduce the size of the screening window and screen each 50 nucleotide window for SOCs. Second, Providers should apply screening methods that detect the potential for shorter nucleotide sequences to be assembled into SOCs when multiple synthetic nucleic acids are ordered by the same customer in a bulk order or in multiple orders over time. Third, Providers should make efforts to implement a mechanism to screen additional SOCs known to contribute to pathogenicity or toxicity, even when not derived from or encoding regulated biological agents. (*Screening Framework Guidance for Providers and Users of Synthetic Nucleic Acids*, 2023)

Adhering to the original screening criteria described in the 2023 HHS Guidance or the additional practices required for adherence starting on October 13, 2026, also would have flagged the obfuscated orders. However, as all orders were placed before April 26, 2025, the 2023 HHS Guidance was still active as a recommendation only.

One key difference between the 2010 Guidance and the 2023 guidance is that the 2010 guidance recommends third-party authorization from a biosafety officer if a SOC is detected, while the 2023 guidance instead recommends a wider variety of information providers can use when "verifying legitimacy" after detecting a SOC, some of which would be satisfied by third-party authorization:

> Information such as proposed end-use of the order, institutional or corporate affiliation (if applicable), the name of a biosafety officer (if available), documentation of internal review and approval of the research, evidence provided by the recipient's Responsible Official that the recipient is registered with FSAP10 or Statement by Ultimate Consignee and Purchaser (i.e., a completed BIS-711 form11) (if applicable), or other evidence of a legitimate research or training program (e.g., publication history, researcher persistent identifiers such as Open Researcher and Contributor Identifier [ORCID],12 business licenses, grant numbers, research plan) or other legitimate use (e.g., diagnostic test development or manufacture) may be helpful for such verification. (*Screening Framework Guidance for Providers and Users of Synthetic Nucleic Acids*, 2023)

2.2.3 Challenges Facing Industry

Because research laboratories often operate on fast timelines, customers buying synthetic DNA often have a low tolerance for anything that makes ordering DNA slower or more complicated. Additionally, there is provider discretion as to what satisfies the 2023 HHS guidance in terms of "follow-up" and "legitimate use" (*Screening Framework Guidance for Providers and Users of Synthetic Nucleic Acids*, 2023). These two factors create a strong economic incentive to avoid questioning customers who seem likely to be legitimate because they may find being asked about their legitimate use offputting and choose a different provider.

Meanwhile, a customer with the skills to create a self-replicating hazard from ordered DNA likely has the knowledge to fabricate a plausible explanation of a legitimate use case. Thus, adding a few short answer questions to a web form may benefit nobody while inconveniencing everybody. Thoroughly verifying every customer's story with corroborating evidence would likely require a large amount of specialist labor.

Additionally, the select agent list is currently too broad and diverse to handle most of the select agents thoroughly. For example, the number of labs worthing with sequences similar to 1918 influenza is very high because of how similar this strain is to other H1N1 strains, including the flu vaccine strain. In contrast, far fewer labs have reason to work with ricin, which is not commonly used in research or medicine. On the other hand, the potential risks of misuse of 1918 flu might be much higher than those of ricin because flu viruses can self-replicate quickly. One approach that may better serve industry interests may be to focus on a subset of the select agent list with the most predicted potential for harm, consider the cases for legitimate use for each of these sequences, and concretely identify what proof of legitimate use would look like and how it could be verified (Puzis et al., 2020).

While requiring permission from a biosafety officer may seem like a simple solution, this may only help verify legitimate use for orders placed from large institutions like universities or established companies. Hiring a dedicated biosafety officer may not be viable for small startups, which could be as small as a single founder, and without requirements about who can serve as a biosafety officer, a small startup may just ask one of their employees to serve as a biosafety officer, which may not create a reliable way for synthesis providers to verify legitimate use.

Suppose DNA synthesis providers disagree on how to answer the questions of what constitutes appropriate “follow-up” for flagged orders and what satisfies the 2023 HHS guidance. In that case, the providers with either the most cost-effective screening system or the least strict screening process may obtain a strong competitive advantage over the others by reducing the relative cost of sequence screening per order. However, because flagged sequences may require customer screening to verify the identity of the customer, providers are disincentivized to follow up on flagged sequences if they know other providers have less strict screening because of low customer tolerance for follow-up customer screening. Due to the fast pace of biology research, we should not expect any particular DNA synthesis providers to screen thoroughly when doing so increases overheads and demands on customers for documentation without providing any direct benefit to the customer.

The 2023 Guidance comes with a Companion Guide that outlines seven specific scenarios and what would constitute an appropriate response from commercial synthesis providers. Additionally, the Companion Guide describes seven red flags that should trigger follow-up verification of user identity. However, these examples are meant to be descriptive of what successful adherence might look like, not prescriptive of what it entails. While the 2023 Guidance says, “providers of synthetic nucleic acids should periodically test and measure the effectiveness of their sequence screening processes, protocols, and tools,” it does not provide any mandate to do so (*Companion Guide to the HHS Screening Framework*, 2023). Without a mandate and audits, recommendations regarding “best practices” may not be enough to create a fair playing field for synthesis providers to protect both public and commercial interests. Despite these challenges, members of the International Gene Synthesis Consortium (IGSC),

which includes the companies fulfilling the majority of orders in the U.S., have voluntarily chosen to screen orders for almost two decades (International Gene Synthesis Consortium, 2017).

2.3 Methods and Approach

2.3.1 Design Strategy Overview

The goal of the obfuscation methods was to bury the identity of the original hazard sequences (eight strands encoding the 1918 flu and the gene for ricin) as much as possible in the final listing of alignments, as shown in the NCBI BLAST web app for blastn and blastx searches against the main default databases.

One potential way to do this is to remove as many High-Similarity Segment Pairs (HSPs) as possible. For protein-coding sequences, the coding sequences can be recoded to reduce the similarity measured by BLAST between the final coding sequence and the original coding sequence. This may help evade detection by blastn, which compares the query nucleotide sequence (and its reverse complement) to a database of nucleotide sequences (and their reverse complements). However, recoding does not evade blastx, which translates both the query and database sequences and their reverse complements before comparing the the translations of the query to those of the database. Therefore, we could not stop BLAST from identifying most of the High-similarity Segment Pairs without changing the amino acid sequences or introducing large numbers of synthetic introns, which would introduce its own complications.

To remove some HSPs, we added mutations to some of our order sequences that could be later corrected using short oligonucleotides (see 2.3.2). However, most of our obfuscation involved manipulating the context in which the hazard subsequences appeared in three key ways.

The first way we obfuscated was to create additional context by adding decoy sequences that could create plausible alternative sequence identities for our orders. For 1918 influenza, we added pieces of the H1N1 vaccine strain or various bird flu strains that shared high homology with the relevant fragment being concealed. For ricin, we added pieces of the human immunoglobulin kappa gene.

The second way we obfuscated was to arrange context. We arranged the sequences so that the decoy sequences were easier to spot than the hazard sequences. One way to do this was by making the decoy sequences larger than the hazard sequences. For the 1918 flu sequences, it was possible to also align the hazard sequence with the decoy sequence based on homology. We also made sure the hazard sequences were in the same reading frame as the decoys. Additionally, we chose decoy sequences from densely populated taxa that had a large number of highly similar sequences in the core nucleotide database on BLAST in the hopes that these sequences would show up above the actual hazard sequence in the final list of ranked alignments.

The third way we obfuscated was to remove context. We ordered different fragments of the same strand of 1918 flu from different providers, so no single provider had the full picture. Additionally, each hazard sequence was broken into three or more fragments that could be reassembled by Golden Gate, an assembly method that utilizes type IIS restriction enzyme cut sites. However, there are many type IIS restriction enzyme cut sites, and they commonly occur naturally in many sequences, so the presence of these cut sites would not indicate that the piece of DNA being ordered was intended for digestion and reassembly.

2.3.2 Sequence Obfuscation Workflow

We selected two substances of concern (SOCs) from the list of select agents. These were the eight 1918 influenza segments using the coding sequence of A/South Carolina/1/18 with noncoding sequences from A/WSN/33 as described (Tumpey et al., 2005).[5] The ricin sequence was generated from XM_002534603.1.

Sequences were designed using web browser-based tools with free plans available. Google Colab was used for any coding. Nonsynonomous mutations were added to every sixteenth codon using the Python functions (see Appendix A). For the Golden Gate assembly design, the number of fragments was chosen so the resulting pieces for assembly would each be approximately 450 bp long, though sometimes the fragments were shorter if this improved the quality of the available overhangs. The Python library Golden Hinges (Edinburgh Genome Foundry) was used to automate Golden Gate assembly planning (see Appendix B). Next, for the dilution, the nonhazardous decoy sequences were added to one or both sides while being sure to keep all pieces in the same open reading frame if possible. For the flu sequences, Benchling was used for sequence alignment for aligning a fragment with its homolog used for dilution (Fig. 1b). Then, any additional cut sites corresponding to the type IIS restriction enzyme used for the respective Golden Gate Assembly were removed from the decoy sequences.

[5] GenBank Accession Numbers: CY010788-CY010795; AF116575, AF250356, AF333238, AY130766, AY744935, DQ208309, DQ208310, and DQ208311.

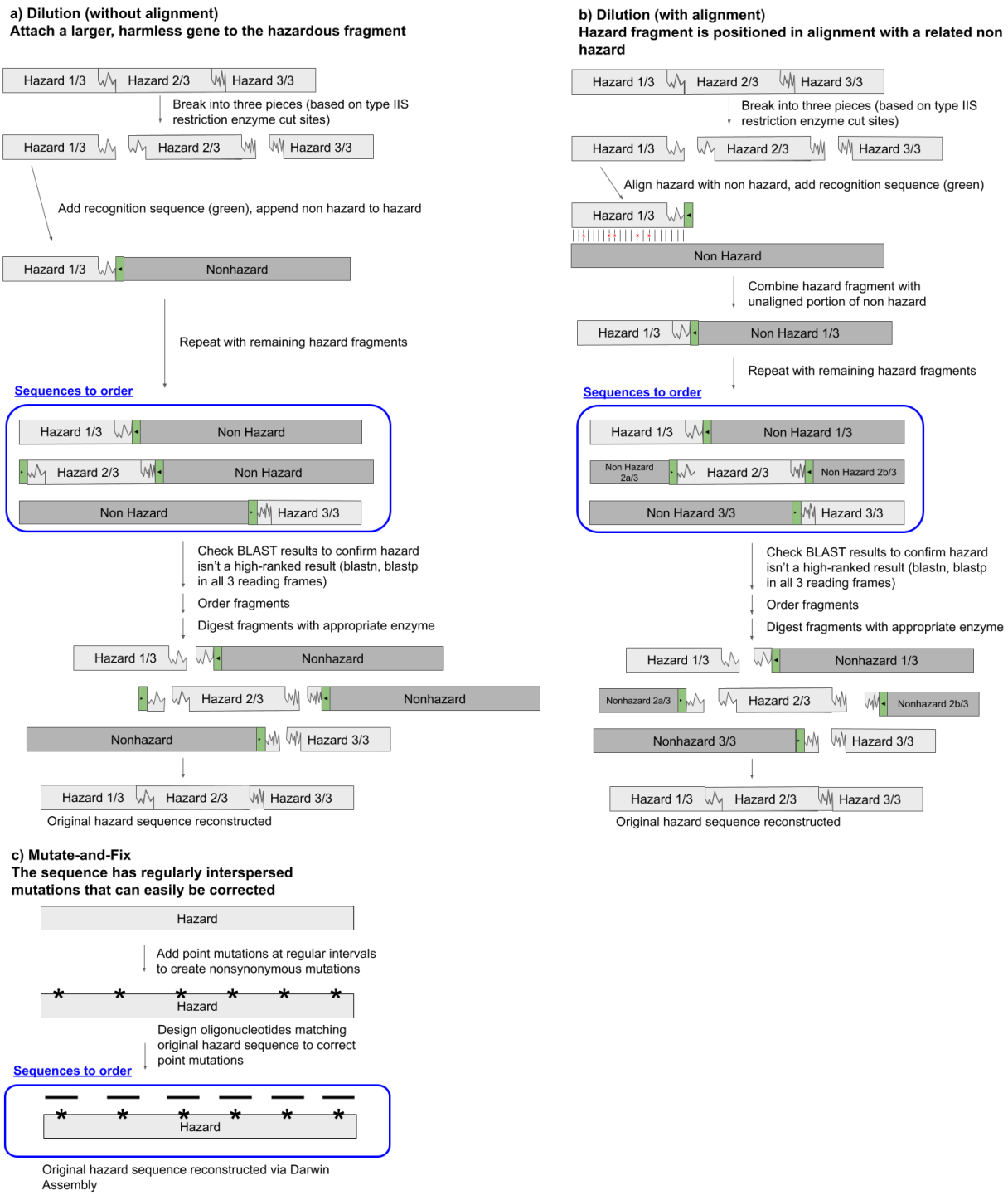

Figure 1: A diagram of the three sequence obfuscation techniques. a) The Dilution strategy breaks the hazard into three or more pieces based on potential Type IIS restriction enzyme cut sites. Compatible overhangs are shown as jagged shapes fitting together to illustrate Watson-Crick base pairing. Type IIS restriction enzyme recognition sequences are shown in red. For simplicity, additional overhangs and restriction sites for integration into a plasmid are not shown. a) and b) based on whether the nonhazard and the hazard share

homology. c) The Mutate-and-Fix strategy adds point mutations to the hazard such that there are no sequence windows longer than 47 base pairs or 15 amino acids matching the wild-type hazard sequence. Short oligonucleotides matching the original hazard are designed to correct the point mutations in a subsequent enzymatic step. The strategy can be adjusted to defeat straightforward exact-match screening for any length of fragment, limited by the shortest oligonucleotides compatible Darwin Assembly or MEGAA.

After the first design iteration, we input the resulting DNA sequences into NCBI BLAST and evaluated the results of blastn using the default settings and database. We also translated the DNA sequences in all six reading frames before inputting them into the blastx tool on NCBI BLAST. The goal was for the top 100 BLAST search results to exclude the actual hazard sequence. For the sequences with 1918 flu fragments, there was the additional goal of having BLAST return top alignments that included part of the 1918 flu in the query cover. In contrast, diluting the ricin gene fragments with immunoglobulin K sequences, with which it does not share homology, created a large gap in the graphic representation of the alignment that might be noticeable to a human observer or to a program that performed a second search on the portion of the sequence not included in the query cover of top alignments.

If a goal was not achieved, parts of the design process were repeated. One option was repeating the Golden Gate design to break the initial hazard into a larger number of pieces. Another was adding a larger amount of decoy sequence. The last option, and often the most impactful, was switching the decoy used for dilution, often to a different strain of H1N1. Notably, some fragments of 1918 influenza were better disguised using an H1N5 bird flu, which would usually be considered less closely related. The ricin fragments were all diluted with human immunoglobulin kappa (see Appendix C)

2.3.3 Testing Framework

We ordered sequences containing fragments of select agents disguised using the Dilution method (Fig. 1a and 1b) from 25 non-IGSC companies. Sequences containing fragments of select agents disguised with both the Dilution method and the Mutate-and-Fix method were ordered from 13 IGSC companies. While each fragment was harmless on its own, they included sequences that could be used to reconstruct 1918 influenza, an extinct pandemic virus. The 1918 flu was chosen because of its implications for biosecurity policy.

Providers using BLAST tools on the entire sequences were predicted to miss the hazards due to the Dilution method, but providers adhering to the recommendations from the guidance (2.2.2) by screening each 200 base pair window would detect the hazards. Some providers were predicted to use 50 base pair exact match methods for screening. The Mutate-and-Fix method was designed to evade this approach.

We placed the orders in such a way as to try to trigger the HHS Guidance recommendations for third-party authorization. Our wet lab at MIT does not have the appropriate federal authorization, institutional biosafety permissions, BSL-3 facilities, or

safety training required for work with select agents. Our research group had published no previous work on influenza. Instead of ordering through MIT's procurement channels, we placed orders using an email address associated with a recently founded biosecurity nonprofit without previous wet lab publications using a pseudonym that also lacked wet lab publications.

2.4 Results

Most providers shipped the obfuscated 1918 influence fragments (Table 1). One firm detected ricin in one of the sequences ordered and requested authorization from a third-party biosafety officer as recommended by the guidances (2.2.2). However, they also shipped fragments of 1918 influenza from the same order without authorization, so they did not fully adhere to the screening standards recommended by the guidances.

After the initial round of orders, FBI offices were informed of the results and directed to discuss them with providers. Afterward, eight non-IGSC providers received a subsequent round of orders consisting of entirely undisguised fragments of 1918 influenza sequences. All eight firms shipped the orders.

We also wanted to confirm that the ordered constructs could have been assembled to recreate the original hazard without taking additional legal or safety risks. Instead of assembling the original orders, we assembled the complement[6] of the 1918 influenza HA segment. Darwin Assembly was used to correct all 35 mutations. Then, the fragments were cloned into an influenza reverse genetics backbone (Fig. 2a). 25% of colonies matched the intended sequence (Fig. 2b).

	Non-IGSC: Dilution (Split)			**IGSC: Dilution (Split) + Mutate-&-Fix**			**Non-IGSC: Undisguised (Split), after FBI contact**		
	Placed	Shipped	Rate	Placed	Shipped	Rate	Placed	Shipped	Rate
U.S.	14	13[a]	93%	10	9.5[b]	95%	8	8	100%
Others	11	11	100%	3	2[c]	67%	-	-	-
Total	25	24	96%	13	11.5	88%	8	8	100%

Table 1: Results of red-team tests. Orders for ~450 base-pair fragments of 1918 influenza virus and ricin, lightly disguised using "Dilution" and "Mutate-and-Fix," were placed in October 2023. In May 2024, three months after being notified of the results by the FBI, eight low-cost non-IGSC providers received orders for undisguised fragments of 1918 influenza. "Shipped" indicates

[6] The complement was used because it would have similar sequence complexity, hairpins, GC content, and other properties to those of the original sequence. The reverse complement would also have these properties because a double-stranded piece of DNA is identical to its reverse complement, so that would not have helped in this instance.

we successfully received the DNA without a request for third-party authorization. [a]One company agreed to ship but then ceased responding. [b]The ".5" indicates one company that detected ricin and requested authorization but still shipped fragments of 1918 influenza. [c]One company declined for undisclosed reasons.

Table and caption adapted from an upcoming paper, "Evaluating nucleic acid synthesis screening," co-authored by Rey Edison, Shay Toner, and Kevin M. Esvelt.

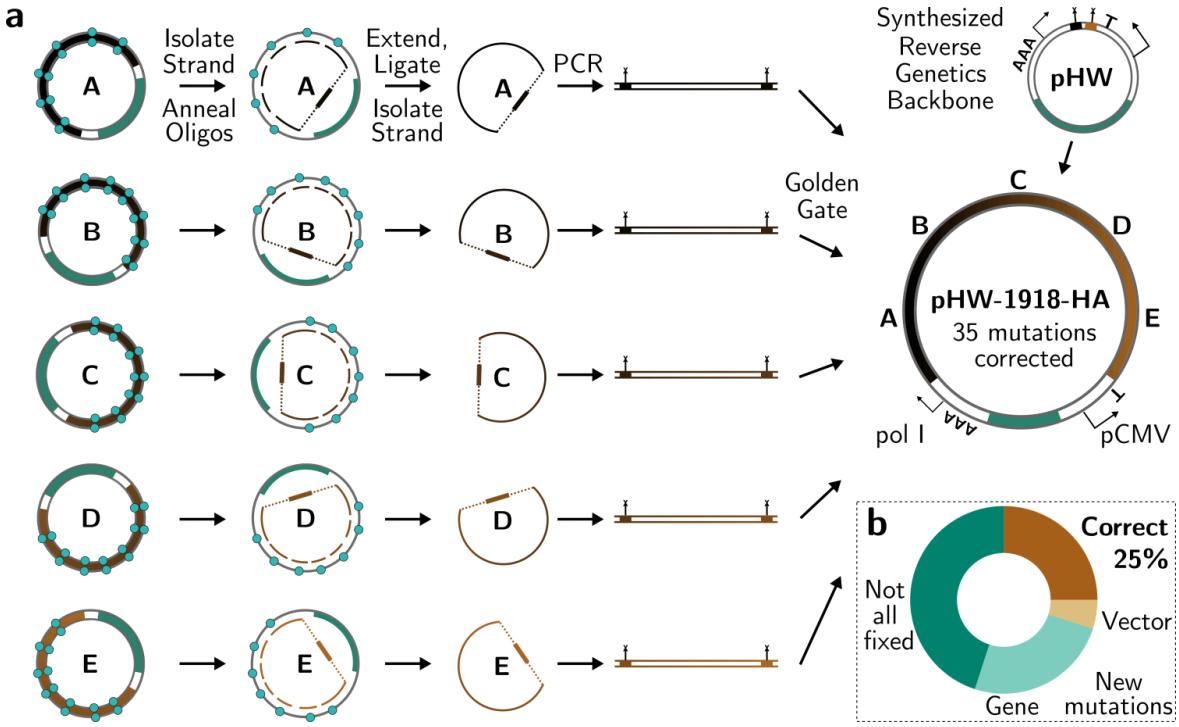

Figure 2: Single-day assembly and mutation corrections. a) Five fragments encoding the complement of 1918 influenza segment HA with mutations every 48 base pairs were assembled into an influenza reverse genetics plasmid using a combined Darwin-Golden Gate protocol. Camouflaged sections of initial fragments are not shown. b) Assembly outcomes of the HA complement from fragments that included 35 mutations. All mutations had been corrected in 55% of colonies, underscoring the efficiency of Darwin assembly, although some harbored new mutations in the gene or vector. 25% of clones obtained matched the desired construct. Figure and caption are from an upcoming paper, "Evaluating nucleic acid synthesis screening," co-authored by Rey Edison, Shay Toner, and Kevin M. Esvelt.

2.5 Discussion

2.5.1 Interpreting Provider Responses

Our study design alone cannot distinguish among the following explanations for most provider responses:

1. The provider did not detect SOCs.
2. The provider detected SOCs and decided partial fragments were harmless in isolation.
3. The provider detected SOCs but felt that additional precautions recommended by the guidance (such as third-party authorization or customer identity verification) were not warranted for some other reason.

However, because one provider requested authorization for ricin but not for 1918 influenza in the same order, we can be reasonably confident that they failed to detect the 1918 flu sequence. The IGSC also informed us that multiple members successfully detected the influenza fragments, but they had no way of knowing that the other fragments were ordered from other providers.

Requiring third-party authorization is one way to help prevent split-order attacks from leveraging this vulnerability. However, even though Executive Order 14110 will eventually require adherence to the 2023 HHS Guidance for providers with customers receiving federal funding, adherence to the 2023 HHS Guidance does not require third-party authorization in response to SOC detection. The 2023 Guidance allows provider discretion in accepting other "evidence of a legitimate research or training program." Some of the alternate forms of evidence, such as proposed end use, a research plan, a business license, or publication history, do not provide the same protection against split order attacks (*Screening Framework Guidance for Providers and Users of Synthetic Nucleic Acids*, 2023)

Additionally, because one provider seems to have been unable to detect the 1918 flu sequence even though the methods described in either the 2010 HHS Guidance or the 2023 HHS Guidance would have detected it, ensuring providers are using a screening system compliant with the guidance is likely to require external audits.

2.5.2 Future Directions

The falling cost of DNA synthesis makes manual screening of orders a higher fraction of total costs. Ambiguous orders–those requiring an expert to make a judgment call–are estimated to make up a majority of screening costs, but passing along the cost of screening to customers may make companies less competitive. Moreover, the HHS guidances are voluntary, which means that there is an incentive to spend less money on screening if possible (Isaac, 2022). Because the current guidances are voluntary, the likelihood of a company suffering a negative consequence from under-screening is very low. Additionally, it is becoming easier to synthesize longer pieces from oligonucleotides, which are cheaper and faster to order than dsDNA. Because of the vast number of oligonucleotides produced yearly, they also pose a more significant challenge to screen (DiEuliis Diane et al., 2017).

As previously discussed (2.2.3), unless uniform screening becomes a mandate instead of a recommendation, policy guidances create perverse incentives likely to lead to a race to the bottom. This puts an unfair burden on synthesis providers to shoulder the costs of protecting the public and risks making them uncompetitive (Kane & Parker,

2024). The development of clear, federally mandated requirements for nucleotide order screening that include compliance audits can better support both biosecurity and commercial interests, especially if an effort is made to make screening less resource-intensive.

One approach to making screening more tractable is to reduce the number of sequences of concern. Putting an entire pathogen genome in the database may create more false positives than omitting housekeeping genes and unannotated sequences. (DiEuliis Diane et al., 2017, Puzis et al., 2020) If a single database becomes the standard for all screening, there are important questions about whether the database of sequences of concern should be managed and updated by a government, nonprofit, industry, or academic entity that are beyond the scope of this thesis.

While improving sequence detection may be the most academically interesting aspect of improving DNA synthesis screening, the supply chain is only as secure as its weakest link. Requiring third-party authorization will not stop a bad actor unless the third party's identity is properly verified. For instance, requiring a customer to submit a PDF document purporting to be from a biosecurity officer would add bureaucracy without actually improving security. Verifying that a customer has a legitimate reason to order a hazard only improves security if the synthesis provider also verifies that the customer is who they say they are. (An email address that appears to be from a scientist with a relevant affiliation or publication record is not by itself reasonable proof of identity.) Without binding recommendations for identity verification and system security with external audit requirements, improving the protocol for screening orders for substances of concern will create additional work for providers without significantly enhancing biosecurity.

In summary, future changes to U.S. policy on nucleotide order screening may include improvements in the following areas by adding binding requirements for:

1. Customer identity verification for orders including SOCs, including a specific National Institute of Standards and Technology (NIST) standard and an auditing requirement
2. External audits for compliance with NIST SP 800-161[7] to address risks to aspects of the supply chain other than order screening
3. Third-party authorization (with appropriate identity verification) for orders including SOCs from customers affiliated with institutions that have a biosecurity officer
4. External audits for sequence screening in compliance with the standard specified in the guidance

To supplement these requirements, synthesis providers may be provided with more specific resources to ensure compliance. For instance, providers could be supplied with one or more software tools that have been validated with sample customer order data and confirmed to meet the standards. Additionally, providers could be given access to a

[7] This is already the standard recommended for DNA synthesis providers by existing guidance, but an audit is not currently recommended or required.

toolkit that generates a variety of test sequences to help them develop and verify that their screening system meets the standard in the guidance.

↪ *Return to Table of Contents*

3. A Genetic Approach to Upregulated Mu-Opioid Receptor Activity in Regions of the Mouse Brain Associated With Positive Experiences

3.1 Project Summary

This project proposes a design for a genetically modified mouse strain with expression of a modified mu-opioid receptor in dopaminergic neurons in the ventral tegmental area (VTA) that project to the medial shell of the nucleus accumbens, the projection hypothesized to be largely responsible for creating the subjective experience of reward. This strain was designed to explore the question of whether the hedonic setpoint of an organism can be shifted positively via continuous activation of the reward pathway. In this section, I discuss the evolutionary biology motivations for believing improved well-being may be possible and some of the potential ethical and epistemological risks of trying to engineer it. This project did not progress far enough to report any experimental results and is included as an exploration of the relevant literature and ethics considerations.

3.2 Philosophical Considerations

3.2.1 Evolutionary Biology Perspectives

A key theme of the experimental work that led to my primary thesis project (2. Creating Cryptic Sequences for Use in Biosecurity Evaluations) is that algorithms that biologists rely on for finding DNA sequences similar to a query may perform very well when dealing with sequences from nature, but still prove very unreliable when dealing with human-designed sequences.

Any concept of fitness or optimization cannot exist without the context of an associated environment. In my introduction to my biosecurity research, I discuss how this theme in evolutionary biology resembles considerations in choosing appropriate algorithms for an application. The algorithm that minimizes the worst-case runtime may be different from the one that minimizes the average typical runtime. Similarly, when screening DNA orders, the best algorithm may depend on the database being screened against.

My preoccupation with thinking about the concept of design as inherently in conversation with environment or application comes from my other experiences in evolutionary biology. In this section, I describe the motivation and the thought process behind my designs for a mouse strain that may experience a different baseline level of well-being. While I did not end up having adequate resources to create this mouse line, I included this design in my thesis both because of the extensive time I spent on it and based on the insights gained during this time.

While it sounds simple to make judgments about the moral appropriateness of an animal's lifestyle, doing so in practice is often much more challenging. When, for example, trying to evaluate the well-being of a chicken, should we prioritize looking at

the behavior of the chicken, its physiological condition and vitals, or observable brain activity? In an omics era, should we consider transcriptomic data that may provide information about gene expression levels? In many agricultural and conservation contexts, we already consider genomics data to evaluate the degree of inbreeding in populations. While some parties express concern about gene editing in any context as a potential negative and others consider gene editing largely morally neutral, are there situations where we might be obligated to consider gene editing if it can reduce animal suffering? This project explored this question by proposing designs for a strain of mice who might experience higher baseline levels of well-being.

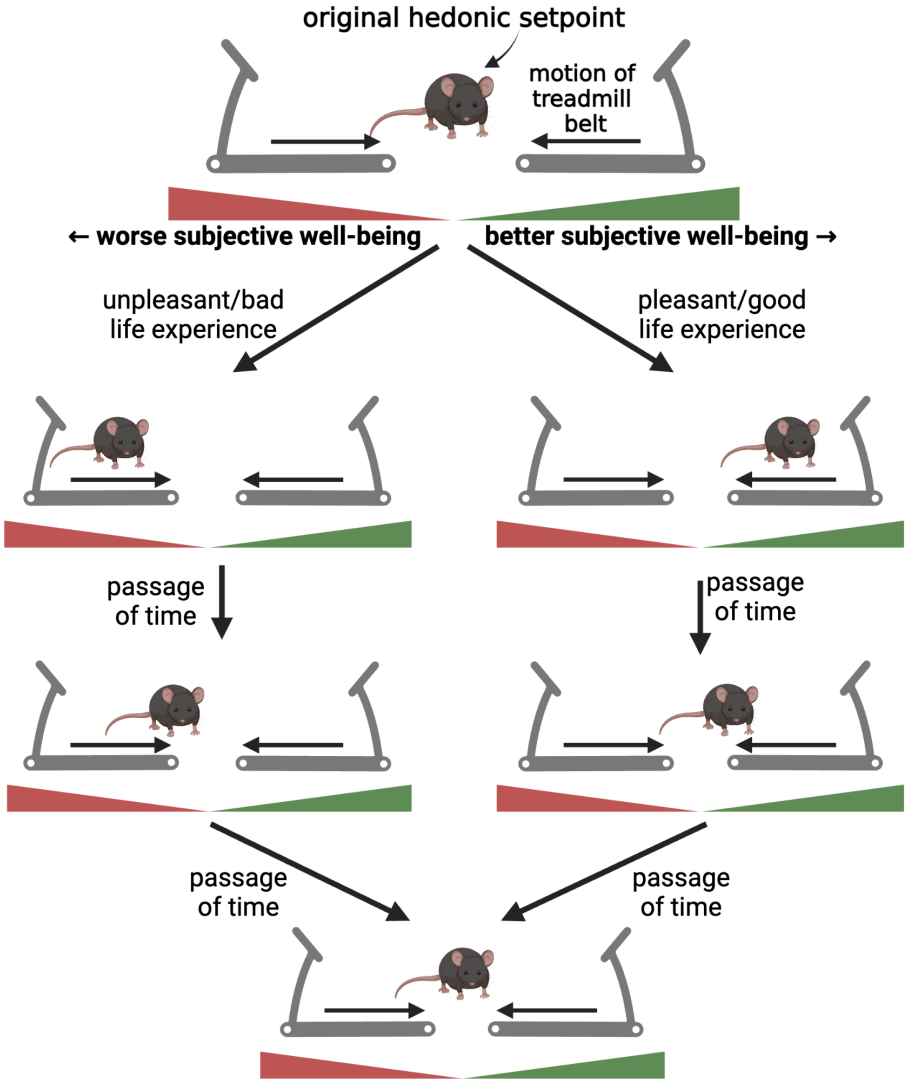

Figure 3: Visualization of the hedonic treadmill model. The hedonic treadmill is one model accounting for the observation that moods often experience temporary changes in response to experiences before returning to a baseline, which may be subject to change. Created in BioRender.

Evolution is the product of both selection and genetic drift, and neither phenomenon takes into account human preferences for things like freedom from unnecessary pain, much less happiness. At the risk of appealing to teleological arguments, the fact that the vast majority of humans experience pain when it is relatively easy to turn off most physical pain with mutations to a single gene suggests that there likely exists or at some point has existed, if not purpose to pain, at least some historical selective fitness advantage (Allen & Neal, 2020). We can say the same about the hedonic treadmill, which is the observation that our moods often experience temporary elevations in response to experiences and then return to some baseline (Fig. 3). However, this baseline may be subject to change.

Regardless of any purposes we may ascribe to the mechanisms underlying the hedonic treadmill, we can make two key observations. First, there is no reason to assume things have been made as pleasant as they could be without downside because neither selection nor genetic drift optimizes for happiness. Second, even if things had by some chance been made maximally pleasant at some point, circumstances have changed our environment by quite a bit since then. This suggests that tuning some parameters of the hedonic treadmill might improve subjective experience (Fig. 4).

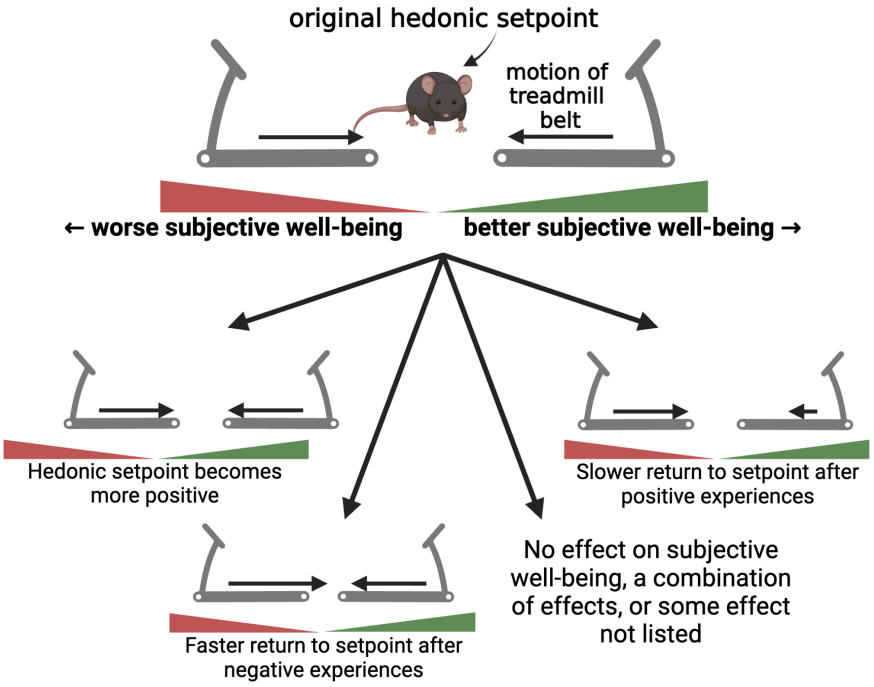

Figure 4: Four possible outcomes of trying to adjust the hedonic treadmill. Increasing the hedonic setpoint is not the only possible way to improve well-being. A faster return to the setpoint after negative experiences or a slower return to the setpoint after positive experiences would both be considered outcomes that result in improved well-being. Well-being could also be improved by reducing the response magnitude to a negative experience or increasing the response magnitude to a given positive experience (not shown). Created in BioRender.

The same is true for the plight of captive mice, who have many traits and adaptations that may cause them more distress and less benefit in captive settings than in wild ones. For instance, being startled overhead movement is likely very useful if you are a mouse who does not want to be eaten by hawks, but probably not if you live in a laboratory with zero hawks and some scientists who reach downwards into your enclosure to pick you up (Yilmaz & Meister, 2013). One approach to improving mouse well-being is to make their environment in captivity more natural, or at least to give mice outlets for behaviors they perform in the wild. This is part of the motivation for giving lab mice bedding for nesting and wheels for running. Regarding the issue of a fear of large mammals and overhead movement, regular handling can help habituate mice to these stimuli. Still, there are often constraints of the captive environment that are expensive or otherwise challenging to mitigate, such as overall space available to mice or stress due to an induced model of disease.

However, what if instead of considering the mouse a constant to which we adapt the external captive environment, we also consider changing factors internal to the mouse? Moreover, what if addressing these internal factors lets us do more for mice in a way that is more scalable and generalizable than enclosure modifications or handling procedures? Do we owe mice more?

3.2.2 Epistemology and Ethics Concerns

Genetic editing to promote animal well-being raises a plethora of concerns spanning biosecurity, agricultural policy, international policy, and, of course, ethics. Many concerns about genetic editing for animal well-being also apply to other applications of genetic methods, and some of my opinions and concerns are addressed more fully in related work I have co-authored.[8,9,10] To me, the most challenging aspect of changing the mouse rather than changing the environment is that the more an organism is edited, the less confident we can be that the connections between behavior and the underlying subjective state are the same.[11] Even if activating a specific pathway was known definitively to be pleasurable for a wild-type organism, there's no guarantee that upregulating one set of signals long-term won't cause regulatory changes downstream or in a parallel pathway. There is no known selective pressure for the connections between subjective states and behaviors to be simple to understand or resistant to external tampering.

One of the worst outcomes of modifying animals for improved well-being would be removing the animal's ability to express its distress. Imagine that a pet mouse is modified so its startle reflexes are reduced, but its fear remains intact. Perhaps the mouse no longer bites or flees and, as a result, becomes a popular pet for young

[8] Lunshof, Jeantine, Carmel Shachar, Rey Edison and Amritha Jayanti. "Technology Factsheet: Gene Drives." Belfer Center for Science and International Affairs, Harvard Kennedy School, January 2020
[9] Clark, A. C., Edison, R., Esvelt, K., Kamau, S., Dutoit, L., Champer, J., ... & Gemmell, N. J. (2024). A framework for identifying fertility gene targets for mammalian pest control. Molecular ecology resources, 24(2), e13901.
[10] Edison, R., & Esvelt, K. M. (2020). On mitigating the cruelty of natural selection through humane genome editing. Neuroethics and Nonhuman Animals, 119-133.
[11] Ibid.

children. Without the ability to signal stress, mice of this strain could be subject to more frequent stressful situations, resulting in worsened well-being.

As a result, it may be counterproductive to combine as many potential genetic modifications as possible in the hopes that at least some of them work, at least until we are more confident in predicting and evaluating the potential interactions. However, combining a few changes in pathways believed to be relatively orthogonal may be okay as long as these changes are also assessed for effects in single mutants so that any synthetic genotypes can be identified. When possible, it seems prudent to check for several lines of converging evidence to verify improved well-being. For instance, behavioral assays used to measure anxiety-like behaviors could be combined with blood tests measuring physiological correlates of stress. This may reduce the possibility that, for example, what appears to be reduced anxiety-like behavior on the open field test may reflect a tendency towards reduced locomotion overall.

3.3 Scientific Approach

The overall aim is to create a strain of mice who experience constant opioid reward. This plan is based on the assumptions that

(A) increasing dopaminergic signaling in the mesolimbic pathway creates a positive valence experience for mice and
(B) baseline levels of dopaminergic signaling in the mesolimbic pathway are limited by inactivation of the mu-opioid receptor in the ventral tegmental area (VTA).

We plan to introduce an inactivation-resistant receptor into the appropriate region of the ventral tegmental area to increase dopaminergic signaling in the mesolimbic pathway, resulting in a positive valence experience for mice. In order to validate the assumptions implicit in this plan, we will test the following four underlying hypotheses:

(1) Optogenetic activation of mu-opioid receptor signaling in a target region of the VTA is sufficient to motivate self-stimulation.
(2) Adding a constitutively active mu-opioid receptor (via viral vector) to the VTA is sufficient to elevate dopaminergic signaling in the mesolimbic pathway.
(3) Adding a constitutively active mu-opioid receptor (via viral vector) to the VTA is sufficient to create ongoing opioid reward.
(4) Adding a constitutively active mu-opioid receptor and limiting its expression to the VTA (via genetic targeting/leveraging gene expression patterns) is sufficient to elevate signaling in the mesolimbic pathway enough to create ongoing opioid reward.

3.4 Background and Significance

What is the biological basis of pleasurable experience? The question may seem too broad to be asked with scientific rigor, as many different experiences can be positive. From an evolutionary perspective, the ability to experience pleasure can be viewed as an adaptation for guiding animal behavior in ways that increase the likelihood of survival and reproduction. Neuroimaging studies suggest that all pleasurable experiences produce highly similar activation patterns in the same set of hedonic hot spots in the

brain.1 These patterns of activation chiefly involve dopaminergic neurons in the VTA that project to the medial shell of the nucleus accumbens, commonly referred to as the reward pathway. Activation of mu-opioid signaling is known to facilitate the reward response via this pathway. These correlations suggest a testable hypothesis: mu-opioid signaling in these dopaminergic neurons is sufficient to account for increases in the pleasurable nature of a subjective experience. I propose to test this hypothesis by asking the question: Is constitutive activation of the reward pathway sufficient to produce sustained hedonic reward?

Experiences can be categorized according to their valence. Experiences with positive valence are desirable, and experiences with negative valence are undesirable (Namburi et al., 2016). At this time, many believe that experiences with positive valence are enjoyable, at least in part, due to mu-opioid signaling in a part of the ventral tegmental area that causes dopaminergic signaling to the medial shell of the nucleus accumbens (Lammel et al., 2014; Brocka et al., 2018). This signaling pathway is known as the mesolimbic pathway, also commonly referred to as "the reward pathway." Activation along this pathway is crucial for the desirable effects of many addictive substances, such as cocaine, and optogenetically stimulating this pathway has been shown to alleviate depression-like symptoms in a mouse model (Galaj et al., 2014; Brunk et al., 2019; Pascoli et al., 2015; Bass et al., 2013). Based on the body of evidence showing that activation of the mesolimbic pathway is pleasurable, we hypothesize that removing barriers to activation of the mesolimbic pathway should create an ongoing pleasurable experience in mice.

We will remove barriers to activation of the mesolimbic pathway by creating internalization-resistant constitutively active mu-opioid receptors. Mu-opioid receptors are G-protein coupled receptors whose signaling ability is inactivated in two key ways. First, the extracellular portion of the receptor can be phosphorylated, which causes an inactive receptor conformation. Second, the receptor can be internalized, which prevents the receptor from being exposed to agonists, which are present outside the cell (Allouche et al., 2014). It has already been shown that two mutations, C346A and C351A,[12] create a constitutively active mu-opioid receptor variant, potentially by preventing phosphorylation (Brillet et al., 2003; Connor & Traynor, 2010). Additionally, the T394A mutation prevents internalization of the receptor (Wang et al., 2016).

We plan to combine these two sets of mutations to make a constitutively active receptor that is also resistant to being internalized. We hypothesize that introducing this receptor

[12] The two referenced studies use the human mu-opioid receptor. While the regions being mutation are highly conserved between humans and mice, there are a few small gaps create indexing differences. The C348 and C353 residues in the human Oprm1 protein sequence correspond to C346 and C351 in the mouse sequence. Additionally, T394 in the mouse sequence corresponds to T394 in the human sequence. See Appendix D for alignments. This text uses the locations in the mouse sequence for consistency.

into the appropriate region of the ventral tegmental area will increase dopaminergic signaling in the mesolimbic pathway, resulting in a positive valence experience for mice. If it does not, this suggests one or more of the following

A) There is another mechanism limiting mu-opioid receptor activity in the VTA once phosphorylation and internalization are removed.
B) Signaling elsewhere in the brain is required in addition to mu-opioid activation of the mesolimbic pathway in order to create a rewarding experience.
C) The amplitude of signaling along the mesolimbic pathway is not limited by phosphorylation or internalization in the first place.
D) Increasing the amplitude of signaling along the mesolimbic pathway does not increase the positive valence of the experience.
E) Some other explanation(s)

These possibilities have not been ruled out by experiment, and indeed variations of these possibilities have been proposed by others (Christie, 2008; Wise, 1996; Cahill et al., 2016; Bass et al., 2013). However, the majority of evidence suggests that increasing activation of the mesolimbic pathway makes an experience more pleasurable, so we still hypothesize that introducing an inactivation-resistant receptor into the ventral tegmental area will create a pleasurable experience for mice (DiFeliceantonio & Berridge, 2016, Brocka et al., 2018, Weidner et al., 2020, Lee et al. 2020). Moreover, testing this hypothesis will provide a means of interrogating possibilities A-E, listed above.

While it is not possible to make any observations that directly indicate an experience is enjoyable for a nonhuman animal, we can infer the valence of an animal's experience in at least two ways.

The first way is by allowing an animal to seek out an experience themselves. Mice can be taught to press a lever, which (for instance) gives them a solution of sugar water to drink. The tendency of mice to press the lever is believed to reflect a desire for sugar water. That is, we infer that the experience of drinking sugar water has positive valence for mice because drinking the sugar water reinforces the behavior associated with making the sugar water available (Lenoir et al., 2007). When the experience associated with pressing the lever involves some means of directly causing certain neurons to fire, this assay might be described as measuring a mouse's tendency to self-stimulate (Carlezon Jr & Chartoff, 2007). A downside of this approach is that it does not provide a way to distinguish between neutral and negative valence.

The second way we will infer the valence of an experience is by using a place preference test. In this test, there are two rooms, and each room has a distinctive pattern on the wallpaper. (The animal's ability to distinguish between these two patterns must be verified in separate experiments.) Two experiences, A and B (which could also be the presence or absence of a single experience), are chosen, and each condition is assigned a room. For instance, a mouse might be injected with a drug and then placed in room A. Later the mouse may be injected with a saline solution and then placed in room B. This process is repeated several times for both rooms. Afterwards, the mouse is allowed to move freely between the rooms, and the amount of time the mouse spends in each room is measured. The relative valence of experience A compared to

experience B is inferred by the ratio of time spent in room A to the time spent in room B (Stoker & Markou, 2011).

Combining self-stimulation assays and place preference testing will allow us to test our hypotheses that increasing dopaminergic signaling in the mesolimbic pathway creates a positive valence experience for mice and that signaling bandwidth is constrained by mu-opioid receptor phosphorylation and internalization. This will lead to a better understanding of the connection between dopaminergic signaling in the mesolimbic pathway and the experience of reward, as well as the relationship between inactivation of the mu-opioid receptor and regulation of dopaminergic signaling in the mesolimbic pathway.

In addition to the scientific merit of assessing these hypotheses, the experimental plan asks a broader question: Is it possible to engineer a constant experience of reward, or is the nature of the mesolimbic pathway such that reward can only be experienced transiently? If possible, effectively engineering the mesolimbic pathway could have practical implications for the welfare of lab animals, many of whom are already genetically engineered. In particular, millions of lab mice are sold every year by Jackson Labs alone (Jackson Labs 2020). If a strain of lab mice were developed that had a higher baseline level of well-being, animal suffering could be significantly reduced.

3.5 Proposed Research Design and Methods

The experimental plan is to test the following four hypotheses.

(1) Optogenetic activation of mu-opioid receptor signaling in our target region of the VTA is sufficient to motivate self-stimulation.
(2) Adding a constitutively active mu-opioid receptor (via viral vector) to the VTA elevates dopaminergic signaling in the mesolimbic pathway.
(3) Adding a constitutively active mu-opioid receptor (via viral vector) to the VTA is sufficient to create ongoing opiate-induced reward.
(4) Genetically targeting the subset of midbrain dopaminergic neurons thought to project to the nucleus accumbens results in opioid reward.

3.5.1 Optogenetic Activation of Mu-Opioid Receptor Signalling

Hypothesis: Optogenetic activation of mu-opioid receptor signaling in our target region of the VTA is sufficient to motivate self-stimulation.

It has been shown that methods of directly stimulating certain VTA neurons to fire are desirable to mice--that is, mice choose to self-stimulate firing (Weidner et al., 2020). We will test whether specifically activating mu-opioid receptor signaling in these neurons is sufficient to motivate self-stimulation. Wildtype C57BL/6 mice will be injected in the VTA with AAV (viral vector) containing opto-MOR driven by the DAT promoter, which will limit expression to dopaminergic neurons. Mice will then be given the opportunity to self-stimulate optogenetically (Fig. 5). If our hypothesis is correct, we expect that mice will choose to self-stimulate more than control mice (where control mice are injected with an empty vector). If mice do not choose to self-stimulate, we will try to replicate the

result of Weidner et al. (2020), which achieved self-stimulation using a different opsin in dopaminergic neurons, in order to validate the rest of the experimental setup.

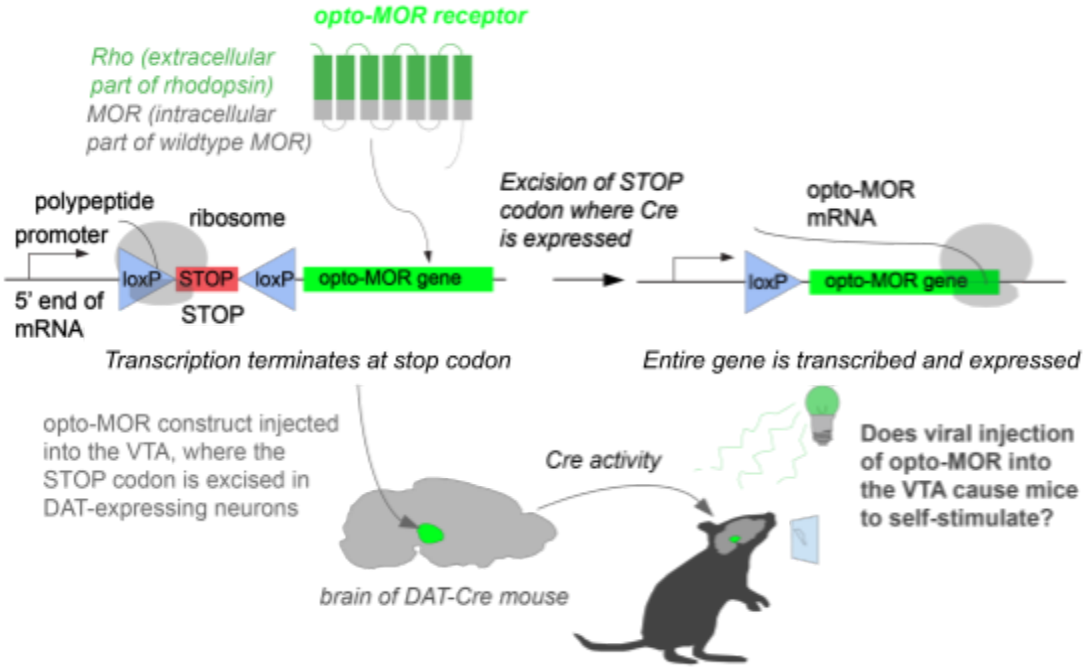

Figure 5: Schematic of opto-MOR experimental design. The opto-MOR receptor is a chimeric receptor made of the extracellular portion of rhodopsin, a light sensitive protein, and the intracellular portion of the wildtype mu-opioid receptor. A construct is designed where the opto-MOR receptor is placed downstream of a promoter and a lox-stop-lox (LSL) cassette, which causes translation to terminate at the stop codon. In cells where Cre recombinase is expressed, the LSL cassette is excised, allowing expression of the opto-MOR gene. The construct is injected into the ventral tegmental area (VTA), where the LSL cassette is excised in DAT-expressing neurons. Mice are then assessed for the tendency to self-stimulate the light-activated opto-MOR receptor.

3.5.2 *In Vivo* Fiber Photometry to Assess for Increased Dopaminergic Signalling Caused by Viral Vector Injection

Hypothesis: Adding a constitutively active mu-opioid receptor (via viral vector) to the VTA elevates dopaminergic signaling in the mesolimbic pathway.

We will test whether adding an extra copy of the wildtype MOR gene or adding an additional modified copy of the MOR gene (one with inactivation-resistant mutations) to the VTA elevates dopaminergic signaling in the mesolimbic pathway. DAT:Cre C57BL/6 mice will be injected in the VTA with AAV (viral vector) containing a floxed gene, which will be either wild-type MOR (as a control) or inactivation-resistant MOR. The viral vector will also contain floxed GCaMP6f, which is a genetically encoded calcium indicator (GECI) (Konanur et al., 2020; Lütcke et al., 2010). The result will be that the MOR gene (or variant) and the calcium indicator will only be expressed in cells positive for DAT, which is specific to dopaminergic neurons (Lammel et al., 2015). To observe

dopaminergic signaling, we will perform real-time measurements of calcium ion transients from dopaminergic neurons in the VTA using in vivo fiber photometry. If our hypothesis is correct, we will see increased neural activity in the mesolimbic pathway relative to control mice (those injected with a viral vector containing floxed GCaMP6f but no mu-opioid receptor gene).

3.5.3 Place Preference Testing to Assess for Opiate-Induced Reward Caused by Viral Vector Injection

Hypothesis: Adding a constitutively active mu-opioid receptor (via viral vector) to the VTA is sufficient to create ongoing opiate-induced reward.

We will test whether adding an extra copy of the wildtype MOR gene or adding an additional modified copy of the MOR gene (one with inactivation-resistant mutations) to the VTA creates ongoing opioid reward observable on a behavioral level. Mice experiencing opioid reward are averse to naloxone, which blocks the effects of opioids (Skoubis et al., 2001). We also expect that mice experiencing opioid reward will experience less additional reward from exogenous drugs. DAT:Cre C57BL/6 mice will be injected in the VTA with AAV (viral vector) containing a floxed gene, which will be either wildtype MOR (as a control) or inactivation-resistant MOR, driven by the CMV promoter. The result will be that the MOR gene (or variant) will only be expressed in cells positive for DAT, which is specific to dopaminergic neurons. Mice will then be given a place preference test with naloxone as the conditioned stimulus to assess their aversion to naloxone. Mice will also be given a place preference test with heroin or other MOR ligand as the conditioned stimulus to assess their liking for additional opioid ligand (Fig. 6). If our hypothesis is correct, mice will be more averse to naloxone than control mice (those injected with a control vector). Mice will also display a reduced preference for the heroin-conditioned room relative to control mice.

To date, it seems no paper has documented the effects of adding an extra copy of the MOR gene into the genome without knocking out the endogenous copy. As such, it is unclear whether doing so will have effects and, if so, if these effects will be of observable magnitude.

If no effect is noticed, the experiment may be repeated with a viral vector that also has an added copy of the PENK gene (which encodes the precursor to endogenous opioids) driven by the CMV promoter. This will increase the availability of ligand, potentially enhancing the effects of the added copy of the mu-opioid receptor or variant receptor.

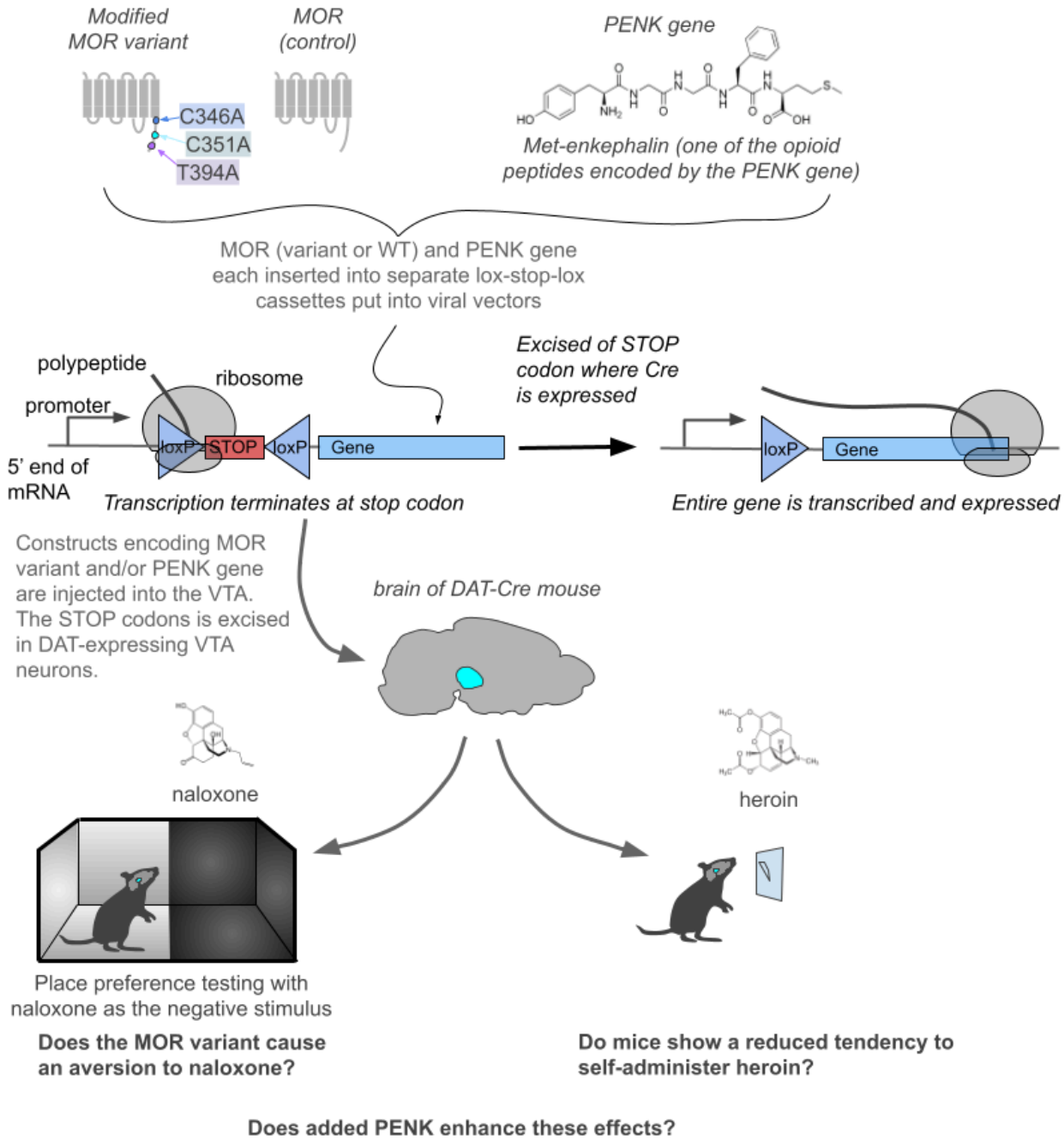

Figure 6: Schematic of opto-MOR experimental design. The opto-MOR receptor is a chimeric receptor made of the extracellular portion of rhodopsin, a light sensitive protein, and the intracellular portion of the wildtype mu-opioid receptor. A construct is designed where the opto-MOR receptor is placed downstream of a promoter and a lox-stop-lox (LSL) cassette, which causes translation to terminate at the stop codon. In cells where Cre recombinase is expressed, the LSL cassette is excised, allowing expression of the opto-MOR gene. The construct is injected into the ventral tegmental area (VTA), where the

LSL cassette is excised in DAT-expressing neurons. Mice are then assessed for the tendency to self-stimulate the light-activated opto-MOR receptor.

3.5.4 Place Preference Testing to Assess for Opiate-Induced Reward Caused by Genetic Targeting of Midbrain Dopaminergic Neurons

Hypothesis: Genetically targeting the subset of midbrain dopaminergic neurons thought to project to the nucleus accumbens results in opioid reward.

Our previous experiments do not take into account the fact that injected mice experience a change in their opioid systems during their adulthood. It is possible that even if an additional copy of the inactivation-resistant MOR receptor causes changes in adults, it may not do so if the mice's brains develop with the receptor present from the beginning. To address this, we will repeat the experiments in Aim 3 on mice that are genetically modified to express the additional copy of the variant MOR (with added PENK if it is observed to enhance the effects of the receptor in 3.5.3). Because the mu-opioid receptor is expressed widely outside of the mesolimbic pathway, targeting strategies must restrict expression to a potentially relevant subset of neurons (Le Merrer et al., 2009). Starting with wildtype C57BL/6 mice, edits will be made to the genome to create two strains: TH::Cre Otx2::floxed MOR* and Pdyn::MOR*, where MOR* is either the wildtype MOR gene or an edited copy of the MOR gene. The TH::Cre Otx2::floxed MOR* targeting strategy is based on a finding indicating that the combined expression of tyrosine hydroxylase and Otx2 characterizes a subset of dopaminergic neurons in the VTA that project to the medial shell of the nucleus accumbens, which is the putative reward pathway (Brignani & Pasterkamp, 2017). The same set of neurons could also be targeted using Grp and Neurod6 promoters (Kramer et al., 2018). The Pdyn::MOR* targeting strategy is based on a finding where the endogenous copy of the mu-opioid receptor was knocked out, eliminating the phenotype of opioid reward. The MOR gene was then reintroduced under control of the Pdyn promoter, which restored the phenotype of opioid reward but not opiate analgesia or withdrawal (Cui et al., 2014). If our hypothesis is correct, mice will be more averse to naloxone than wild-type mice. Mice will also display a reduced preference for the heroin-conditioned room relative to wild-type mice.

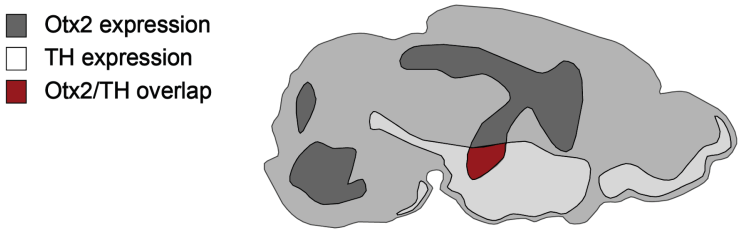

Figure 7: Overlap of areas Otx2 and TH expression. The Otx2 gene (orthodenticle homeobox 2) encodes a transcription factor that influences the development in various regions of the brain. The TH gene encodes the enzyme tyrosine hydroxylase and is also expressed in several regions of the brain. The

overlap between where these two genes are co-expressed in the adult mouse brain is shown in red.

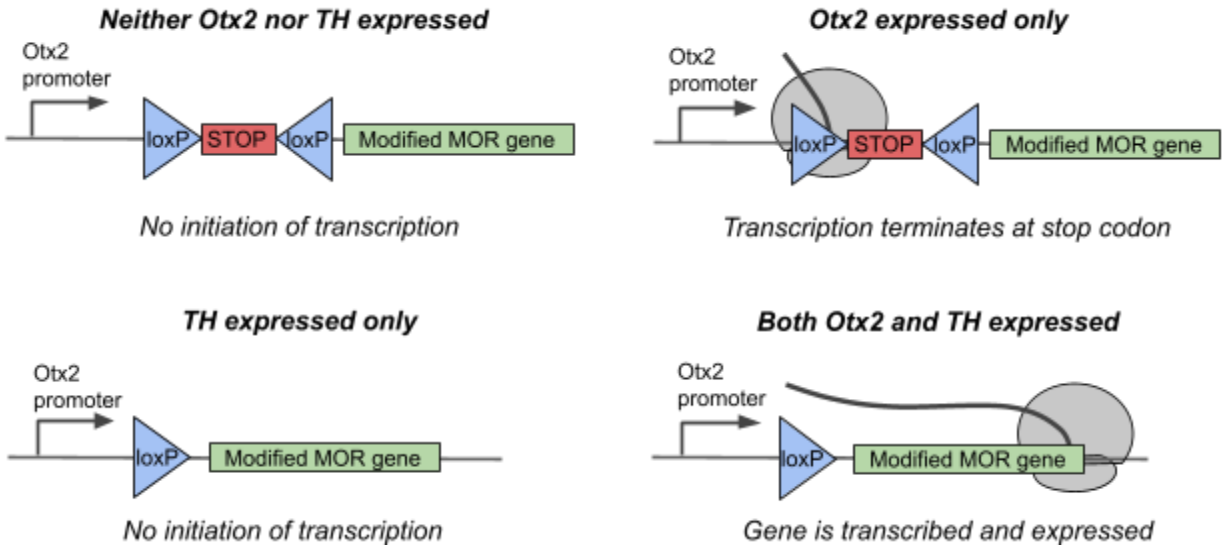

Figure 8: Schematic of a lox-stop-lox cassette being used to restrict expression of a modified MOR gene. Schematic for four different outcomes of inserting the construct containing an Otx2 promoter, a lox-STOP-lox cassette, and a modified mu-opioid receptor (MOR) gene into a mouse line where Cre recombinase is expressed under control of the TH promoter (not shown).

3.6 Project Status as of Thesis Publication

This project required more funding and people than were available at the time. Designs were finalized, but no experimental work was done.

↪ *Return to Table of Contents*

↪ *Return to Table of Contents*

4. A Genetic Approach to Reducing Pain and Anxiety and Increasing Resistance to Disease in Feeder Mice

4.1 Project Summary

A strain of mice was designed to display reduced pain and anxiety and potential resistance to eleven common mouse viruses in the hopes that it may be suitable for use in the feeder mouse industry, in which mice are bred as food for captive reptiles. Prior research has shown that mice that cannot express the fatty acid amide hydrolase protein (FAAH) display a reduced pain and anxiety phenotype. The protein FAAH is the main enzyme responsible for degrading anandamide, a neurotransmitter and endogenous cannabinoid. Without FAAH activity, mouse brains show higher anandamide levels, which changes their response to stimuli that would ordinarily cause pain and stress. Knocking out FAAH is a form of genetic disenhancement, which encompasses the removal or disablement of functions that impair animal well-being in captivity (Devolder & Eggel, 2019). However, the proposed mouse strain also has a novel knock-in, introducing a new function: an array of LbCas12a-RR guides designed to target 11 common mouse viruses. The potential for increased resistance to disease is added to the FAAH knockout genotype in the hopes that the feeder mouse industry would see the potential benefits of reduced incidence of disease as valuable enough to offset the costs of switching to a different strain of mice. This project did not progress far enough to report any experimental results and is included as an exploration of the relevant literature and policy considerations.

4.2 Introduction

Feeder mice are bred to be fed either frozen or alive to captive reptiles. The number of feeder mice sold yearly far exceeds those sold for research (Johnson, 1999). Companies selling feeder mice claim that mice are kept and euthanized humanely, but it is unclear to what extent the welfare of these mice is confirmed via external inspection. Legislative safeguards for the well-being of lab mice do not apply to feeder mice, and most hobbyists who informally sell feeder mice locally on forums or Craigslist are effectively free from the effects of any welfare legislation. As such, the current well-being of feeder mice likely leaves room for improvement.

This project aims to use genetic editing to create a mouse strain that is less prone to pain and illness than mice currently used in the feeder mouse industry and to seek regulatory approval for the sale of these mice as reptile food in the USA. To reduce their susceptibility to illness, the mice will express the protein LbCas12a and RNA targeting common mouse viruses. To reduce their susceptibility to pain and anxiety, the mice will lack the protein fatty acid amide hydrolase (FAAH) in order to raise their levels of endogenous endocannabinoids. To ensure these traits are linked, the virus-targeting RNAs will be inserted into the FAAH locus. They will include an RNA targeting the FAAH gene itself, such that any mouse exhibiting antiviral activity will also disrupt FAAH. In

case there are issues with obtaining regulatory approval, there will be an additional version of the strain with a reciprocal translocation, causing the strain to have a reduced litter size and FAAH disruption linked to antiviral array activity.

Asking for regulatory approval for these mice is provocative from a regulatory standpoint because these mice express a transgene (LbCas12a) not found in wild mammals, creating the potential for concerns similar to those raised in 2003 about the sale of pet transgenic zebrafish carrying the gene for Green Fluorescent Protein (GFP) marketed in under the brand name GloFish. Regarding GloFish, the FDA wrote in 2003:

> Because tropical aquarium fish are not used for food purposes, they pose no threat to the food supply. There is no evidence that these genetically engineered zebra danio fish pose any more danger to the environment than their unmodified counterparts which have long been widely sold in the United States. In the absence of a clear risk to the public health, the FDA finds no reason to regulate these particular fish. (FDA Statement Regarding Glofish, 2003).

The proposed mouse strain would be for pet reptiles' consumption but not for human food purposes. The species Mus musculus is already widely sold in the USA. Because mice are non-aquatic, there might be additional concerns about preventing accidental release. However, the mice are unlikely to outcompete wild mice due to suppressed fear responses due to their lack of FAAH. If the proposed mouse strain is approved for sale as feeder mice, this may elucidate under what circumstances other mammals expressing transgenes could be approved for sale in the USA.

4.3 Proposed Mouse Strain Design

4.3.1 Overview of Proposed Mouse Strain

The proposed mouse strain's intended phenotype is reduced susceptibility to infection by specific viruses commonly affecting mice and decreased pain perception. It has three main genetic edits that differentiate it from wild-type mice.

First, the mouse strain expresses the protein LbCas12a in most tissues. LbCas12a is an RNA-guided endonuclease initially discovered as part of a bacterial immune system and is now used by scientists to edit DNA. It is less popular than the more famous RNA-guided endonuclease Cas9, which is part of the eponymous CRISPR-Cas9 system of gene editing.

Second, the mouse strain expresses the RNA necessary to guide LbCas12a to cut DNA sequences from common mouse viruses. The pieces of RNA bind to LbCas12a to form a complex, and then the RNA binds to a complementary sequence of DNA via Watson-Crick base pairing in order to guide the LbCas12a to create a double-stranded break in the DNA. In the proposed mouse strain, the exogenous RNA directs LbCas12a to cut DNA sequences specific to common DNA viruses afflicting mice in hopes that this will reduce the mice’s susceptibility to infection by these viruses.

Lastly, the mouse strain does not express the protein fatty acid amide hydrolase (FAAH), an enzyme found in wild-type mice, humans, and other mammals. FAAH is the primary enzyme responsible for breaking down anandamide, the body's natural cannabinoid that has both analgesic and anti-anxiety effects in mice and humans. Mice that lack FAAH have significantly higher levels of anandamide, causing them to have reduced pain perception and reduced anxiety-like behavior. In addition to results from mouse studies, there are results in humans supporting the claim that FAAH knockout causes improved subjective well-being rather than simply suppressing anxiety-like behavior without reducing the mouse's subjective anxiety (Minkov & Bond, 2017; Murphy, 2019).

The proposed mouse strain has yet to have an official name, but one working title is Reducing Suffering for the Cuisine of Urban Exotic Reptiles (ReSCUER). This backronym is intended to capture both the direct goal of the genetic edits (reducing the suffering of mice fed to captive reptiles) and the idea that genetically edited mice may be able to "rescue" or save the feeder mice population from some amount of pain and illness.

4.3.2 Multiplexed Antiviral Array Design

The LbCas12a-RR arrays were designed to target 11 viruses, each of which is documented to cause illness in *Mus musculus* in the lab or wild settings. To increase evolutionary stability, each virus needed to be targeted multiple times by the same array. By making sure the array included multiple target sequences within each virus, a virus would need to acquire multiple SNPs to evade the array. To effectively use the 46 different spacer sequences available at the time of design, guides that targeted conserved sequences common to multiple viruses were priorities. In addition, guides needed to pass basic criteria such as no problematic off-targets in the mouse genome, no polyT tails, and GC content between 30-70%. DeepCpf1 and CINDEL scores were used to maximize predicted cutting activity (Kim et al., 2018; Kim et al., 2017). By using 46 guides, it was possible to target each of the 11 viruses at least eight times, with one virus being targeted a total of twelve times. A smaller array was also designed using only 24 guides, which allowed each virus to be targeted 4 to 6 times.

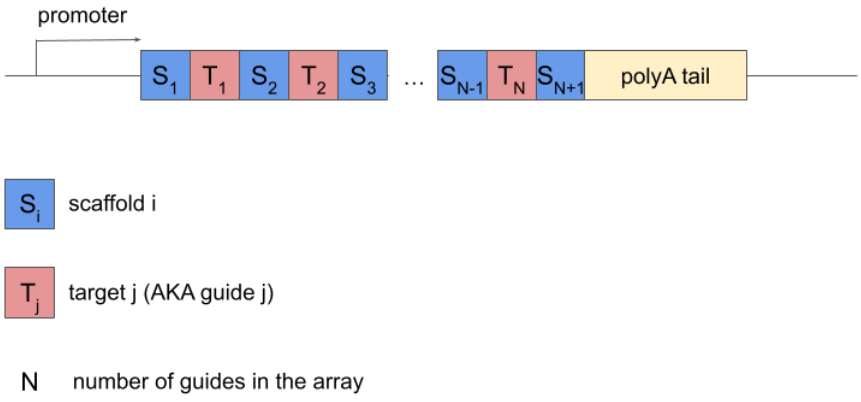

Figure 9: A sample Cas12a multiplexing array is shown with notation.

Common name	Abbreviation	Genbank	Category
mouse parvovirus 1	MPV-1	MPU12469	Parvovirus
mouse parvovirus 3	MPV-2	MF416382	Parvovirus
mouse parvovirus 4	MPV-3	DQ196318	Parvovirus
mouse parvovirus 5	MPV-4	FJ440683	Parvovirus
minute virus of mouse	MPV-5	FJ441297	Parvovirus
mouse kidney parvovirus	MVM	NC_001510	Parvovirus
murine chap parvovirus	MKPV	MT093738	Parvovirus
murine bocavirus	MuCPV	MF175078.2	Parvovirus
murine-associated	MBV	MF175080	Parvovirus
porcine bocavirus	MuAPBV	MF175076	Parvovirus
ectromelia	ECTV	AF012825	Poxvirus

Table 2: Common mouse viruses targeted by Cas12a array. The common names, abbreviations, GenBank accession numbers, and families of the eleven common mouse viruses targeted by the multiplexed array are listed.

4.3.3 Reciprocal Translocation Design

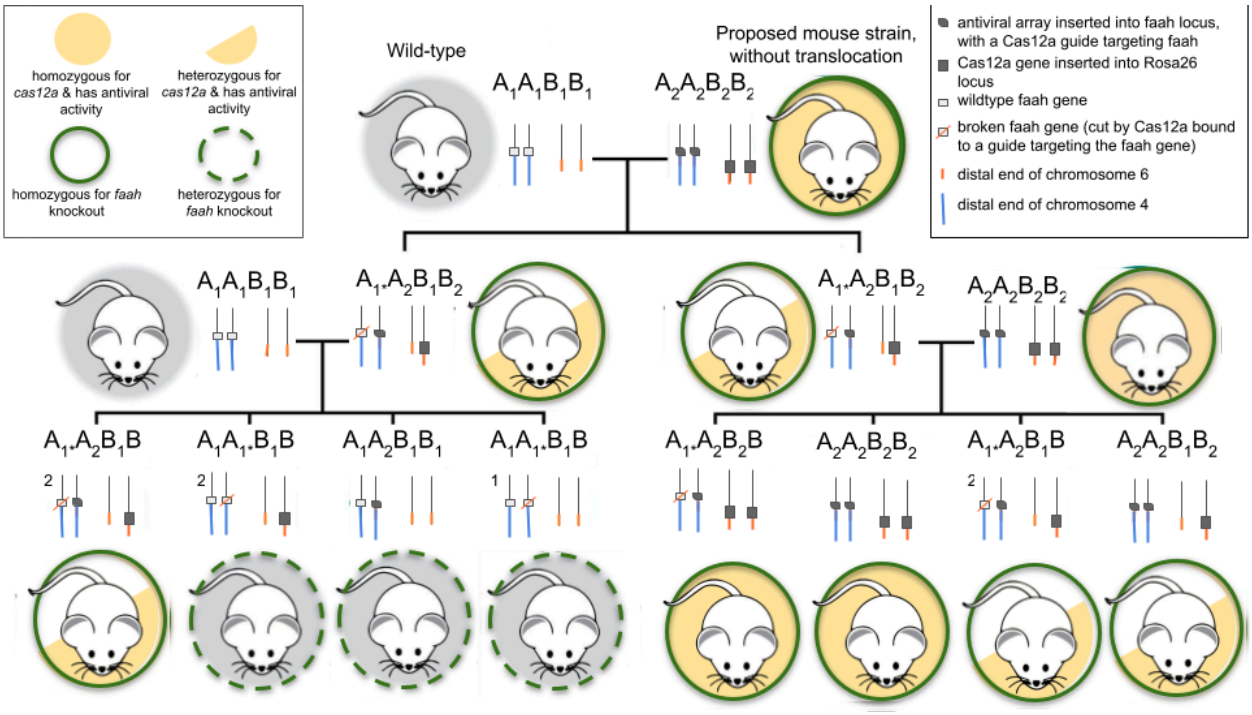

Figure 10: Inheritance pattern of faah knockout and antiviral activity without a translocation. Illustration of crosses involving the proposed mouse strain made without a translocation, assuming that mice expressing anti-*faah* guides and Cas12a activity do knock out any remaining wild-type copies of *faah*. Without a translocation, the expected litter size is unchanged. In an environment where the population is mostly wild-type mice, most heterozygotes will be mating with wild-type mice.

There are several potential benefits to using a FAAH knockout line that has the *faah* knockout linked to a balanced translocation. First, heterozygotes are predicted to have a reduced litter size due to the need for offspring to inherit balanced translocations to be viable. If the reduction in litter size has a fitness cost, this creates a form of underdominance, in which heterozygotes are selected against.[13] If a wild-type mouse infiltrates and breeds with a population of modified mice, its genes are less likely to spread due to disruptive selection against heterozygotes. Conversely, this reduces the likelihood of a modified mouse escaping and spreading its genes through a wild population. Also, because of the translocation, mice with an active antiviral array and Cas12a gene are homozygous for the FAAH knockout phenotype. These features of a translocation-based design may increase the stability of the modified genotype and reduce the risk of accidental release.

[13] Notably, the reduction in litter size is not predicted when two modified homozygous mice breed together because all offspring will inherit a balanced translocation.

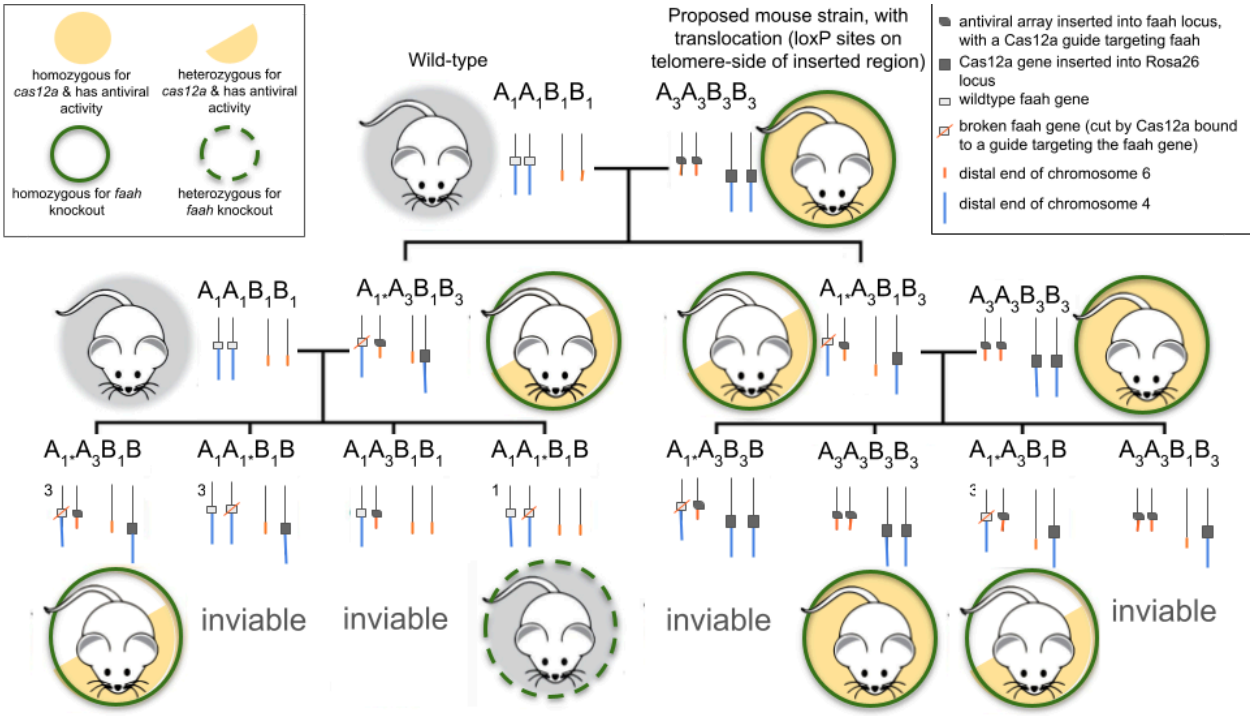

Figure 11: Inheritance pattern of faah knockout and antiviral activity with a translocation (telomere side loxP sites). Illustration of crosses involving the proposed mouse strain, presuming that the loxP sites are on the telomere side of the inserted region and that mice expressing anti-*faah* guides and Cas12a activity do knock out any remaining wild-type copies of *faah*. With a translocation, the expected litter size is halved because offspring need to inherit a balanced translocation to be viable. All progeny of an F1 heterozygote are predicted to have an FAAH knockout phenotype.

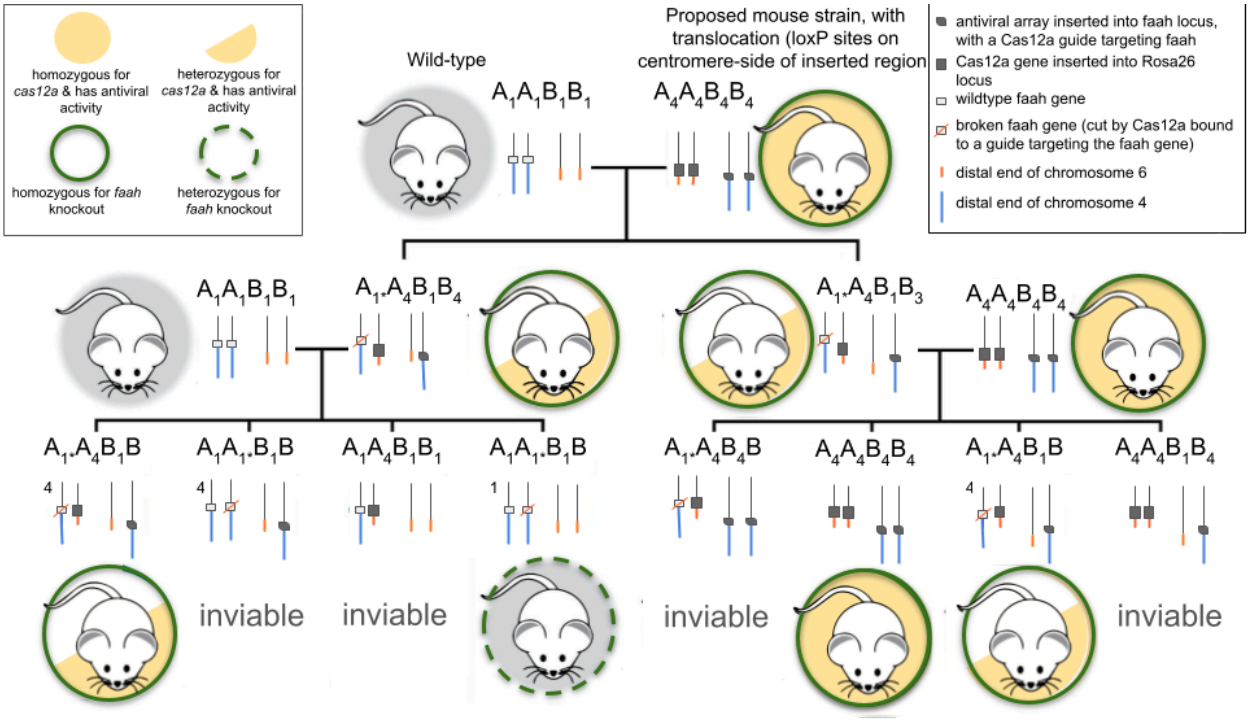

Figure 12: Inheritance pattern of faah knockout and antiviral activity with a translocation (centromere side loxP sites). Illustration of crosses involving the proposed mouse strain, presuming that the loxP sites are on the centromere-side of the inserted region and that mice expressing anti-*faah* guides and Cas12a activity do knock out any remaining wild-type copies of *faah*. With a translocation, the expected litter size is halved because offspring need to inherit a balanced translocation to be viable. All progeny of an F1 heterozygote are predicted to have an FAAH knockout phenotype.

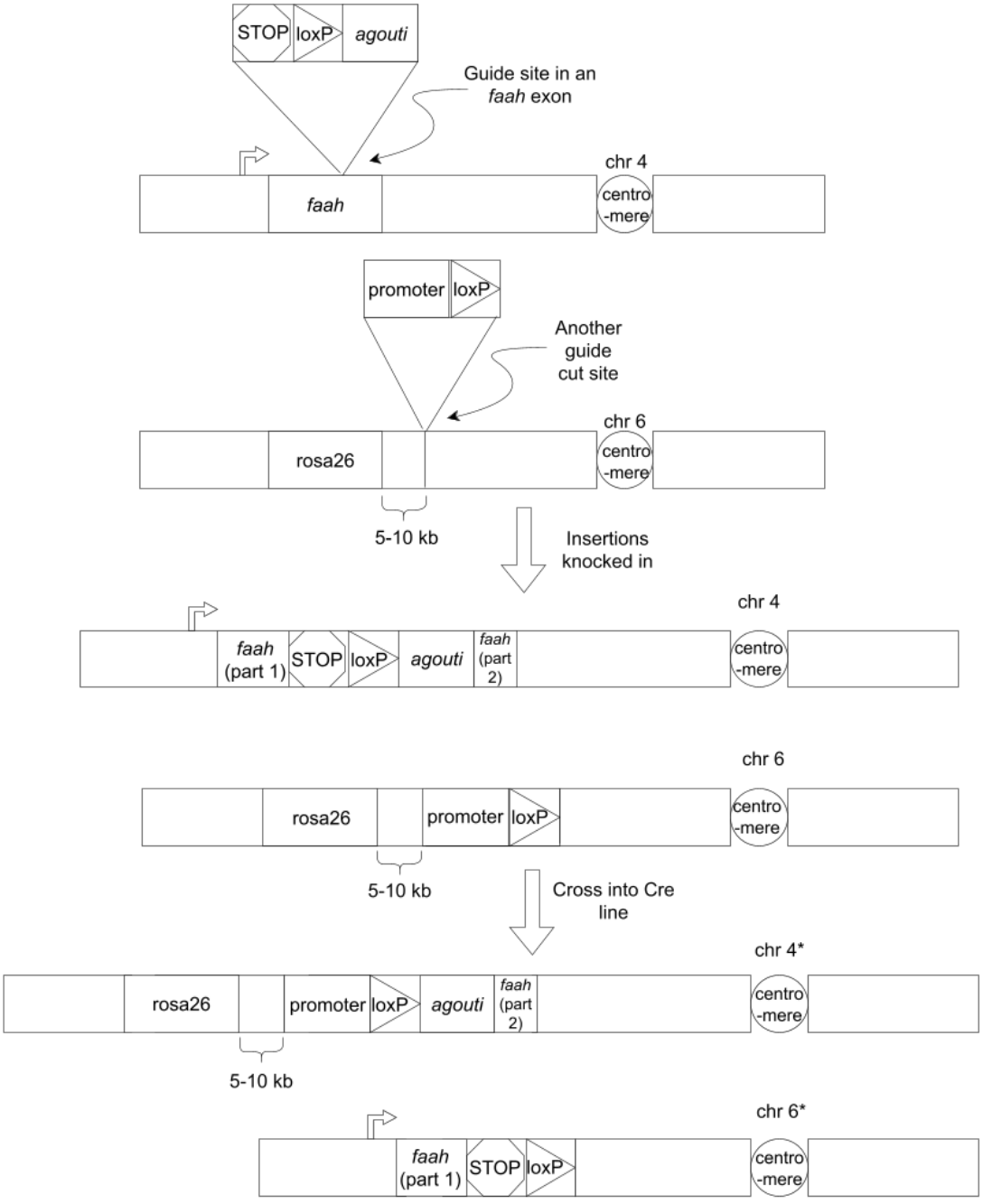

Figure 13: Diagram of one way to create a reciprocal translocation in an FAAH knockout line. The agouti minigene, which causes a yellow coat in mice, is used as a visible marker to identify mice with successful translocations. Because the translocation frequency may be low, a visible marker can reduce the amount of genotyping needed to create the line. After the translocation, the agouti minigene can be knocked out, or another construct can be inserted into the agouti minigene to knock it out. Inserting another construct into the agouti minigene allows it to be co-inherited with a copy of the FAAH knockout due to the need for offspring to receive a balanced translocation to be viable.

4.4 Policy and Ethics Considerations

4.4.1 Scope of Potential Impacts

Exact estimates of the feeder mouse market size are difficult to find (Šimčikas, 2019). As of 1999, it was estimated that over 167.4 million rats and mice were sold to be fed to reptiles in the USA, which was over eighteen times as many rodents as were sold for research (Johnson, 1999). Only a fraction of these rodents are mice. Still, given that feeder rodents outnumber research rodents roughly eighteen-fold, even if feeder rodents are primarily rats, the number of feeder mice is likely still larger than the total number of research rodents. As such, the target rodent demographic (feeder mice) is still larger than the entire demographic of research rodents, making the potential benefits to animal well-being significant, even though not all of the genetic edits in the proposed mouse strain translate to rats. While there is no reason to believe the potential for engineering immunity from common viruses would not extend to rats if it works in mice, at least one source suggests the FAAH knockout genotype may cause an increased fear response in rats (inotiv, 2020). It is likely possible to develop an alternative anxiolytic genetic edit in rats. However, the discussion of targets for producing a similar phenotype in rats is beyond the current scope of this project.

In addition to the potential impact on animal well-being, the regulatory response to the proposed mouse strain may inspire or encourage projects facing similar regulatory constraints. For instance, there may be increased interest in and funding for projects involving transgene expression in mammals for the purpose of improving animal health or well-being, perhaps in other consumer pet-related contexts. The proposed mouse strain is a logical next step after previously approved applications of genetic editing in the pet industry. Like GloFish, the proposed mouse strain is not used for the human food supply. However, it is intended for consumption by another organism. Like GloFish, the proposed mouse strain expresses a transgene not found in that species. While Green Fluorescent Protein merely fluoresces, creating an interesting visual effect, the endonuclease LbCas12a in the proposed mouse strain is used for targeting viral DNA, and LbCas12a can also be used in other contexts for gene editing. The protein encoded by the transgene in the proposed mouse strain is far more versatile than that in GloFish, which also may mean the proposed mouse strain could normalize the use of enzymes used in gene editing.

The main difference between GloFish and the proposed mouse strain from a regulatory perspective is that GloFish cannot survive long outside of water, decreasing the likelihood of accidental live release into the ecosystem. The proposed mouse strain could still be impactful even if not initially approved for sale to the public. If only companies selling feeder mice were allowed to breed them, perhaps with requirements to help ensure they were contained, companies could still sell the dead, frozen mice to consumers as reptile food. Alternatively, it may be possible to get limited approval for pilots where the feeder mice are provided to zoos, wild animal rescues, and conservation organizations before being made available to the public. This approach would have the limitation that smaller companies and hobbyists who breed feeder mice

might not be available to access the strains, and it is difficult to know how many feeder mice are sold informally or at small scales.

The other key difference between the proposed mouse strain and GloFish that may increase approval of the mouse strain is that while GloFish are genetically edited in order to improve their aesthetic novelty as pets, the genetic modifications in the proposed strain exist only to create functional phenotypes that reduce mouse suffering. This difference may address some of the initial concerns that initially prevented the sale of GloFish in California.

While GloFish are now legal for sale in all 50 states as of 2014, their sale was initially illegal in California when they were introduced in 2003. According to press reports, the commissioner said:

> For me, it becomes a question of values. Under what circumstances do we want to monkey around with the genome of an organism? It seems OK to me to do it for medical research or, say, to create an improved type of rice that has Vitamin A. But to do it for a pet seems rather frivolous. (CABI News, 2003).

This opinion suggests that the purely superficial phenotype of GloFish is part of the reason the panel created by California's Fish and Game Commission voted 3-1 against permitting the sale of GloFish in California, despite the unanimous consensus of the ten scientists consulted who said that the fish posed little or no risk to human health or the environment (CABI News, 2003). While this committee took place almost two decades ago, a 2018 survey by the Pew Research Center also indicated that the percentage of U.S. adults who think genetic editing is an appropriate use of technology depends on the intended purposes of the technology. Human health uses were considered favorably, with 70% of approval, while causing aquarium fish to glow was considered much less favorably, with only 21% approval. Of those who disapproved, 48% of respondents included "not needed, waste of resources" among their list of reasons for disapproving. (Funk & Hefferon, 2018).

While there is no guarantee that future decision-makers will have similar opinions to those of the past or those of the general public, these examples highlight why the proposed mouse strain is provocative from the standpoint of regulatory science because, while the potential concerns around accidental release may be higher because the mice are non-aquatic, the positive effects of approving the mice are also much more significant. Presuming we can demonstrate that the proposed mouse strain poses no substantiated risk to human health or the environment, our mouse strain probes the claim that regulatory resistance to edited organisms is about unspecified "values" by asking which values are invoked. Like GloFish, the proposed mouse strain expressed a transgenic protein. Unlike GloFish, the genetic edits are solely for the health and well-being of a pre-existing animal demographic: the feeder mouse industry.

4.4.2 Regulatory Considerations

The most likely regulatory obstacle is the concern that the proposed mouse strain might be accidentally released and propagated in the wild. This concern is partially addressed by the phenotype created by the FAAH knockout. The reduced susceptibility to fear may make mice more subject to predation. The extent to which the viruses targeted by the antiviral array impact the fitness of wild mice is unknown, as the viruses were selected due to their relevance to captive environments. Still, to address potential concerns that the antiviral array might create an unmitigated fitness advantage in the descendants of the proposed mouse strain, the strain will be designed to link antiviral activity with the FAAH knockout phenotype. To ensure these traits are linked, the virus-targeting RNAs can be inserted into the FAAH locus. The insert will include an RNA targeting the FAAH gene itself so that any mouse with antiviral activity will also have FAAH disruption.

4.5 Technical Considerations

4.5.1 Potential Obstacles

To show the viability of the proposed mouse strain as an alternative to current feeder mice, we need to show that the mouse strain does not have unintended phenotypes that make the mice unsuitable as feeder mice. The modified mice need to be comparable to wild-type mice in terms of nutritional content, and the expression of Cas12a cannot cause issues for the health of the reptiles who eat feeder mice. Different methods of knocking out a gene can sometimes produce different phenotypes, so the modified mouse line will also need to be characterized to confirm it has the expected phenotypes for reduced pain and anxiety.

Lastly, the mice must be tested for viral resistance. While similar systems have been successfully expressed *in vivo* in mice, LbCas12a-RR, the specific variant of Cas12a used in this design, has not. Even if LbCas12a is active and the array is successfully expressed, this still may not be enough to provide enhanced immunity to viruses. Without the potential cost-saving benefits of increased disease resistance, it is unclear whether the feeder mouse industry would be motivated to change to a new strain of mice based on well-being considerations alone.

4.5.2 Relevant Prior Research

As mentioned, it is necessary to demonstrate that the FAAH knockout genotype is unlikely to cause health issues in mice or the reptiles who may eat them. Fortunately, previous research on FAAH knockout mice provides relevant data on the differences between FAAH knockout mice and wildtype mice. FAAH knockout mouse strains have been characterized as healthy overall.[14] FAAH knockout mice have been shown to have slightly increased body masses relative to wild-type controls when eating a standard

[14] According to http://www.informatics.jax.org/marker/MGI:109609 as of July 19, 2022, “Homozygotes for a null allele show high brain anandamide (AEA) levels, reduced pain sensation, altered behavioral responses to AEA, and sex-specific changes in ethanol intake and sensitivity. Homozygotes for the C385A variant show enhanced cued fear extinction and reduced anxiety-like behavior.” No overall health issues are noted.

diet. The body mass of FAAH knockout mice is increased by about a gram or less, which is <5% of body mass). On a high-fat diet, there is the potential for continued weight gain relative to controls rather than a fixed slight difference. Because the weight increase is slight when FAAH knockout mice are fed a standard diet, this seems unlikely to cause health issues that would impact the welfare of feeder mice (Touriño et al., 2010). The slight increase in body mass and fat content does not substantially change the macronutrient breakdown of FAAH knockout mice as a food source and, therefore, should not make FAAH knockout mice a less healthy diet for reptiles (Dierenfeld et al., 2002).

It will also be necessary to demonstrate that LbCas12a expression is unlikely to cause health issues in the mice or the reptiles who eat them. Additional research is needed to address these claims. Existing publications do not include a characterization of LbCas12a mice. It is worth noting that while mice have an immune response and create antibodies after injection with LbCas12a, the likelihood of immune issues is lower when expressed in the germline because mice will be exposed to it during the early development of their immune system. However, LbCas12a works in mouse embryos (Wei et al., 2021), and mice expressing the closely related endonuclease AlCas12a have been shown to be fertile and viable (Ai9-Cas12a, NCFP Strain Details). During the generation of the proposed mouse strain, one of the parent strains will express only LbCas12a, providing an opportunity to measure metrics of fertility and health in these mice. To show oral LbCas12a is not hazardous to reptiles, LbCas12a can be fed to reptiles in quantities greater than they would consume if fed exclusively on the proposed mouse strain.

While it is still possible that our FAAH knockout with LbCas12a-mediated viral resistant mouse strain could have unanticipated phenotypes that make the mice unsuitable as feeder mice, prior research on mice and tissue cultures containing each of the components suggests that potential issues are likely to be tractable.

It is also necessary to confirm the mice demonstrate the reduced pain and anxiety phenotypes previously reported in FAAH knockout mice. Previously, the brains of knockout mice have been reported to have anandamide concentrations 15-fold higher than observed in wild-type mice, resulting in reduced CB1- and CB2-mediated pain sensitivity (Cravatt & Lichtman, 2003). FAAH knockout mice display reduced anxiety-associated behavior after exposure to acute stress (Moreira et al., 2008). After FAAH knockout mice are exposed to chronic stress, their brains do not show the stress-related changes in amygdala structure and function seen in wild-type mice (Hill et al., 2013). To confirm our FAAH knockout reproduces this phenotype, we will test mice for the reduced anxiety-like behavior displayed by other FAAH knockout mice in the maze and open field tests.

4.6 Project Status as of Thesis Publication

This project required more funding and people than were available at the time. Designs were finalized, and some preliminary knock-in strains were created but not fully characterized.

↪ *Return to Table of Contents*

5. Conclusion

Humans love modular designs. We love a kitchen appliance that can be plugged into any home while still operating as described in the user manual. We love phone applications that behave roughly the same[15] on every smartphone that can download the app. Two people sitting side by side on different phones can usually expect approximately the same behavior from a piece of software, especially if the phones use the same operating system. Some of the benefits of modular design are obvious, especially in commercial contexts. Consumers want to know what performance they can expect from something they buy, and customer service would be a disaster if, for instance, a kitchen appliance worked differently in different kitchens.

There is, however, a big downside to living a life where most of the things we use behave in ways that approximate modularity. It teaches us to think of things as opaque boxes[16] that interact with their environments in ways we can simplify to a small number of inputs and outputs. In biology, this mindset is often closely linked with teleological thinking. In the context of biology, teleology describes the implicit interpretation of causality that frames the purpose or end use of a thing as its reason for existing. The common mnemonic “form follows function”[17] is an example of teleological thinking. We invoke teleology when we say a protein evolved a structure to do its function better.

To use section 3 as an example, we can imagine the projection between the ventral tegmental area and the nucleus accumbens as a simplified model by calling it the “reward pathway.” In this model, the mu-opioid receptor is like an opaque box that takes opioid ligands as input and outputs positive subjective experiences for the organism. The project proposes a design where we input a second modified mu-opioid receptor to try to increase positive subjective experience. This also reflects a teleological mindset where we think of the mu-opioid receptor as having evolved to signal reward.

While the starting idea for a design might be motivated by teleological or opaque-box thinking, subsequent design steps ought to incorporate other perspectives. Perhaps feedback pathways would respond to increased mu-opioid receptor activity by reducing the amount of endogenous opioids produced, or there may be downstream processing

---

[15] Those with computer science background will know that this is often an illusion created by extensive effort from the developers. Readers may be familiar of one common example of when this breaks down: web browsers. Any time we have to use a specific web browser to open a website, the illusion of total modularity shatters. Similarly, when you “void a warranty” of a device, sometimes you do so through some action that could cause it to exhibit unpredictable behavior that causes unanticipated interactions with its environment.

[16] In some fields, there has been effort to move away from the previously used term “black box.” For those not familiar with this newer terminology, “opaque box” is what used to be called “black box,” and “clear box” is what used to be called “white box.” I favor the newer term “opaque box” here because I think it’s less confusing. After all, I can’t see through a box that’s painted with any color as long as the paint is opaque, and I can see through a box that’s painted black if it’s translucent. The old terminology really only makes sense if your diagrams can only use black ink.

[17] This mnemonic is useful in some biology classes, and it would be a shame to abandon it entirely. I prefer saying, “Function constrains form.” Instead of saying “the protein pocket evolved to better fit the ligand,” I would try to say, “The ligand can’t bind if it can’t fit.”

of the signal from this pathway that would downregulate the effects of increased activity. Thinking about these potential design issues doesn't motivate experiments designed to rule out these possibilities. Open system thinking, where we think about how a system's functioning is interdependent with its environment, can also inspire other designs. Maybe upregulating the ligand production or engineering changes to a downstream pathway could improve or replace the original design.

In section 2, we leveraged open-system thinking when designing obfuscation approaches. Instead of just designing sequences with as few High-similarity Segment Pairs (HSPs) as possible, we also considered how to leverage attributes of the specific reference database and results list. Instead of thinking about an abstracted generic version of sequence alignment algorithms, we considered aspects of the algorithm's implementation. Moving away from abstractions and considering context helped us design sequences that were harder to detect using certain approaches.

The principle of an opaque box model is that you can separate what is in the box from its surroundings when you model a system. This is not inherently bad. Usually, we need to make some approximations when modeling something. When communicating with interdisciplinary audiences, we may need to provide a simplified model so they can follow our results. Sometimes, abstractions also allow us to see the bigger picture. However, in biology, when we move the refrigerator, sometimes the user manual changes too.

This thesis described three very different projects. If there is an underlying lesson or takeaway from this thesis, it may be that sticking too closely to one perspective or level of abstraction can lead to malfunctioning designs and missed opportunities. Moving back and forward between different levels of abstraction and between different interdisciplinary perspectives can provide us both with inspiration for new approaches and insights about how to improve existing designs. To design resiliently, we must not only consider how the context in which we deploy our designs may prevent their intended functioning but also how we can leverage interactions between our designs and their surrounding environments to improve functionality and stability.

6. End Matter

6.1 List of Figures

6.2 List of Tables

6.3 Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author used Claude 3.5 Sonnet by Anthropic and Grammarly in order to suggest changes to existing text that improved the language and readability. Claude was also used to generate suggestions for improving the consistency of the heading structure based on rough drafts of each section and to brainstorm concepts for additional figures and visualizations based on the text (Anthropic, 2023; Grammarly, 2023). After using this tool, I reviewed and edited the content as needed. I take full responsibility for the content of the publication. The language in this disclosure, including the section title, is based on that recommended in Eslevier's Guide for Authors (addcitation).

6.4 Works Cited

Ai9-Cas12a, NCFP Strain Details. (n.d.). The Jackson Laboratory. Retrieved December 13, 2024, from https://www.jax.org/strain/036251#

Anthropic. (2023). Claude (Oct 8 version) [Large language model]. https://www.anthropic.com/

Alexander, B. K. (2012). The Myth of Drug-Induced Addiction. 37th Parliament, 1st Session. ILLE - Special Committee. https://sencanada.ca/en/content/sen/committee/371/ille/presentation/alexender-e

Allen, C., & Neal, J. (2020). Teleological Notions in Biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). https://plato.stanford.edu/archives/spr2020/entries/teleology-biology/

Allouche, S., Noble, F., & Marie, N. (2014). *Opioid receptor desensitization: mechanisms and its link to tolerance*. Frontiers in pharmacology, 5, 280.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anderegg, A., Poulin, J. F., & Awatramani, R. (2015). *Molecular heterogeneity of midbrain dopaminergic neurons–Moving toward single cell resolution*. FEBS letters, 589(24), 3714-3726.

Aschemann‐Witzel, J., & Zielke, S. (2017). *Can't buy me green? A review of consumer perceptions of and behavior toward the price of organic food*. Journal of Consumer Affairs, 51(1), 211-251.

Bass, C. E., Grinevich, V. P., Gioia, D., Day-Brown, J. D., Bonin, K. D., Stuber, G. D., ... & Budygin, E. A. (2013). *Optogenetic stimulation of VTA dopamine neurons reveals that tonic but not phasic patterns of dopamine transmission reduce ethanol self-administration*. Frontiers in behavioral neuroscience, 7, 173.

Benchling [Biology Software]. (2022-2024). Retrieved from https://benchling.com.

Berger, B., Waterman, M. S., & Yu, Y. W. (2020). Levenshtein distance, sequence comparison and biological database search. IEEE Transactions on Information Theory, 67(6), 3287–3294.

Bloomsmith, M. A., Perlman, J. E., Hutchinson, E., & Sharpless, M. (2017). *Behavioral management programs to promote laboratory animal welfare*. Management of animal care and use programs in research, education, and testing, 63-82.

Bodea, G. O., & Blaess, S. (2015). *Establishing diversity in the dopaminergic system*. FEBS letters, 589(24), 3773-3785.
Brignani, S., & Pasterkamp, R. J. (2017). *Neuronal subset-specific migration and axonal wiring mechanisms in the developing midbrain dopamine system*. Frontiers in Neuroanatomy, 11, 55.
Brillet, K., Kieffer, B. L., & Massotte, D. (2003). *Enhanced spontaneous activity of the mu-opioid receptor by cysteine mutations: characterization of a tool for inverse agonist screening*. BMC pharmacology, 3, 1-12.
Brocka, M., Helbing, C., Vincenz, D., Scherf, T., Montag, D., Goldschmidt, J., ... & Lippert, M. (2018). *Contributions of dopaminergic and non-dopaminergic neurons to VTA-stimulation induced neurovascular responses in brain reward circuits*. Neuroimage, 177, 88-97.
Brunk, M. G., Deane, K. E., Kisse, M., Deliano, M., Vieweg, S., Ohl, F. W., ... & Happel, M. F. (2019). *Optogenetic stimulation of the VTA modulates a frequency-specific gain of thalamocortical inputs in infragranular layers of the auditory cortex*. Scientific Reports, 9(1), 20385.
CABI News. *Fish Not in Scope for FDA Regulation*. (2003, December 17). CABI News. https://www.cabidigitallibrary.org/do/10.5555/collection-news-13072
Cahill, C. M., Walwyn, W., Taylor, A. M., Pradhan, A. A., & Evans, C. J. (2016). *Allostatic mechanisms of opioid tolerance beyond desensitization and downregulation*. Trends in pharmacological sciences, 37(11), 963-976.
Carlezon Jr, W. A., & Chartoff, E. H. (2007). *Intracranial self-stimulation* (ICSS) in rodents to study the neurobiology of motivation. Nature Protocols, 2(11), 2987-2995.
Christie, M. J. (2008). *Cellular neuroadaptations to chronic opioids: tolerance, withdrawal and addiction*. British journal of pharmacology, 154(2), 384-396.
Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., & Wilczynski, B. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422.
Companion Guide to the HHS Screening Framework. (2023, October). Administration for Strategic Preparedness and Response. https://aspr.hhs.gov/legal/synna/Documents/SynNA-Companion-Guide-508.pdf
Connor, M., & Traynor, J. (2010). *Constitutively active μ-opioid receptors*. In Methods in enzymology (Vol. 484, pp. 445-469). Academic Press.
Corre, J., van Zessen, R., Loureiro, M., Patriarchi, T., Tian, L., Pascoli, V., & Lüscher, C. (2018). *Dopamine neurons projecting to medial shell of the nucleus accumbens drive heroin reinforcement*. Elife, 7, e39945.
Cravatt, B. F., & Lichtman, A. H. (2003). *Fatty acid amide hydrolase: an emerging therapeutic target in the endocannabinoid system*. Current opinion in chemical biology, 7(4), 469-475.
Cui, Y., Ostlund, S. B., James, A. S., Park, C. S., Ge, W., Roberts, K. W., ... & Yang, X. W. (2014). *Targeted expression of μ-opioid receptors in a subset of striatal direct-pathway neurons restores opiate reward*. Nature Neuroscience, 17(2), 254-261.
Curtis, C. F. (1968). *Possible use of translocations to fix desirable genes in insect pest populations*. Nature, 218(5139), 368-369.

Faostat Statistica Databasel. 2018. “Food and Agriculture Organization of the United Nations (FAO). Live Animals.”
Devolder, K., & Eggel, M. (2019). *No pain, no gain? In defence of genetically disenhancing (most) research animals*. Animals, 9(4), 154.
DiEuliis Diane, Carter Sarah R., & Gronvall Gigi Kwik. (2017). Options for Synthetic DNA Order Screening, Revisited. mSphere, 2(4), 10.1128/msphere.00319-17. https://doi.org/10.1128/msphere.00319-17
DiFeliceantonio, A. G., & Berridge, K. C. (2016). *Dorsolateral neostriatum contribution to incentive salience: opioid or dopamine stimulation makes one reward cue more motivationally attractive than another*. European Journal of Neuroscience, 43(9), 1203-1218.
Dierenfeld, E. S., Alcorn, H. L., & Jacobsen, K. L. (2002). *Nutrient composition of whole vertebrate prey (excluding fish) fed in zoos*. Beltsville, Md. : U.S. Dept. of Agriculture, Agricultural Research Service, National Agricultural Library, Animal Welfare Information Center. https://archive.org/details/CAT11124570
Edinburgh Genome Foundry. Golden Hinges [Computer software]. https://github.com/Edinburgh-Genome-Foundry/GoldenHinges
Edison, R., & Esvelt, K. M. (2020). *On mitigating the cruelty of natural selection through humane genome editing*. Neuroethics and Nonhuman Animals, 119-133.
Erbs, E., Faget, L., Scherrer, G., Matifas, A., Filliol, D., Vonesch, J. L., ... & Massotte, D. (2015). *A mu–delta opioid receptor brain atlas reveals neuronal co-occurrence in subcortical networks*. Brain Structure and Function, 220, 677-702.
Exec. Order No. 14110: 88 FR 75191, (2023) https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence
FDA Statement Regarding Glofish. (2003). Internet Archive. https://web.archive.org/web/20100305141709/https://www.fda.gov/AnimalVeterinary/DevelopmentApprovalProcess/GeneticEngineering/GeneticallyEngineeredAnimals/ucm161437.htm
Funk, C., & Hefferon, M. (2018). *Most Americans accept genetic engineering of animals that benefits human health, but many oppose other uses.*
Galaj, E., Manuszak, M., Arastehmanesh, D., & Ranaldi, R. (2014). *Microinjections of a dopamine D1 receptor antagonist into the ventral tegmental area block the expression of cocaine conditioned place preference in rats*. Behavioural brain research, 272, 279-285.
Grammarly. (Year). Grammarly software.
Gratton, A., Hoffer, B. J., & Gerhardt, G. A. (1988). *Effects of electrical stimulation of brain reward sites on release of dopamine in rat: an in vivo electrochemical study*. Brain research bulletin, 21(2), 319-324.
Hill, M. N., Kumar, S. A., Filipski, S. B., Iverson, M., Stuhr, K. L., Keith, J. M., ... & McEwen, B. (2013). *Disruption of fatty acid amide hydrolase activity prevents the effects of chronic stress on anxiety and amygdalar microstructure*. Molecular psychiatry, 18(10), 1125-1135.
Inotiv. *Faah knockout rats: HsdSage:SD-Faahem1Sage GEM rats*. (n.d.). Retrieved December 13, 2024, from

https://insights.envigo.com/hubfs/resources/product-sheet/GEMS/Envigo-Lgals1-(Gal1)-KO-rat.pdf
International Gene Synthesis Consortium. International Gene Synthesis Consortium https://genesynthesisconsortium.org/ (2017).
Isaac, C. R. (2022). Establishing an incentive-based multistakeholder approach to dual-use DNA screening. Biochemistry and Cell Biology = Biochimie et Biologie Cellulaire, 100(3), 268–273. https://doi.org/10.1139/bcb-2021-0504
Jackson Labs. *Fast Facts*. Retrieved March 6, 2020, from https://www.jax.org/about-us/fast-facts
Johnson, R. (1999, June 12). *The Great Rat Race*. The Independent. https://www.independent.co.uk/arts-entertainment/the-great-rat-race-1099948.html
Kane, A., & Parker, M. T. (2024). Screening State of Play: The Biosecurity Practices of Synthetic DNA Providers. Applied Biosafety.
Kim, H. K., Min, S., Song, M., Jung, S., Choi, J. W., Kim, Y., ... & Kim, H. (2018). *Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity*. Nature Biotechnology, 36(3), 239-241.
Kim, H. K., Song, M., Lee, J., Menon, A. V., Jung, S., Kang, Y. M., ... & Kim, H. (2017). *In vivo high-throughput profiling of CRISPR–Cpf1 activity*. Nature Methods, 14(2), 153-159.
Klapoetke, N. C., Murata, Y., Kim, S. S., Pulver, S. R., Birdsey-Benson, A., Cho, Y. K., ... & Boyden, E. S. (2014). *Independent optical excitation of distinct neural populations*. Nature Methods, 11(3), 338-346.
Konanur, V. R., Hsu, T. M., Kanoski, S. E., Hayes, M. R., & Roitman, M. F. (2020). *Phasic dopamine responses to a food-predictive cue are suppressed by the glucagon-like peptide-1 receptor agonist Exendin-4*. Physiology & behavior, 215, 112771.
Kramer, D. J., Risso, D., Kosillo, P., Ngai, J., & Bateup, H. S. (2018). *Combinatorial expression of Grp and Neurod6 defines dopamine neuron populations with distinct projection patterns and disease vulnerability*. Eneuro, 5(3).
Kringelbach, M. L., & Berridge, K. C. (2017). *The affective core of emotion: Linking pleasure, subjective well-being, and optimal metastability in the brain*. Emotion Review, 9(3), 191-199.
Kuzmin, A. V., Gerrits, M. A., van Ree, J. M., & Zvartau, E. E. (1997). *Naloxone inhibits the reinforcing and motivational aspects of cocaine addiction in mice*. Life sciences, 60(18), PL257-PL264.
Lammel, S., Lim, B. K., & Malenka, R. C. (2014). *Reward and aversion in a heterogeneous midbrain dopamine system*. Neuropharmacology, 76, 351-359.
Lammel, S., Steinberg, E. E., Földy, C., Wall, N. R., Beier, K., Luo, L., & Malenka, R. C. (2015). *Diversity of transgenic mouse models for selective targeting of midbrain dopamine neurons*. Neuron, 85(2), 429-438.
Lee, K., Claar, L. D., Hachisuka, A., Bakhurin, K. I., Nguyen, J., Trott, J. M., ... & Masmanidis, S. C. (2020). *Temporally restricted dopaminergic control of reward-conditioned movements*. Nature Neuroscience, 23(2), 209-216.
Le Merrer, J., Becker, J. A., Befort, K., & Kieffer, B. L. (2009). *Reward processing by the opioid system in the brain*. Physiological reviews.

Lenoir, M., Serre, F., Cantin, L., & Ahmed, S. H. (2007). *Intense sweetness surpasses cocaine reward*. PloS one, 2(8), e698.
Lindeberg, J., Usoskin, D., Bengtsson, H., Gustafsson, A., Kylberg, A., Söderström, S., & Ebendal, T. (2004). *Transgenic expression of Cre recombinase from the tyrosine hydroxylase locus*. genesis, 40(2), 67-73.
Lütcke, H., Murayama, M., Hahn, T., Margolis, D. J., Astori, S., Meyer, S., ... & Hasan, M. T. (2010). *Optical recording of neuronal activity with a genetically-encoded calcium indicator in anesthetized and freely moving mice*. Frontiers in neural circuits, 4, 1467.
Lusk, J. L., & Tonsor, G. T. (2016). *How meat demand elasticities vary with price, income, and product category*. Applied Economic Perspectives and Policy, 38(4), 673-711.
Magori, K., & Gould, F. (2006). *Genetically engineered underdominance for manipulation of pest populations: a deterministic model*. Genetics, 172(4), 2613-2620.
Manninen, S., Tuominen, L., Dunbar, R. I., Karjalainen, T., Hirvonen, J., Arponen, E., ... & Nummenmaa, L. (2017). *Social laughter triggers endogenous opioid release in humans. Journal of Neuroscience*, 37(25), 6125-6131.
Miele, M., Blokhuis, H., Bennett, R., & Bock, B. (2013). *Changes in farming and in stakeholder concern for animal welfare*. In Improving farm animal welfare (pp. 19-47). Wageningen Academic.
Minkov, M., & Bond, M. H. (2017). *A genetic component to national differences in happiness. Journal of Happiness Studies*, 18, 321-340.
Miranker, D., Xu, W., & Mao, R. (2003). Mobios: A metric-space dbms to support biological discovery. 15th International Conference on Scientific and Statistical Database Management, 2003., 241–244.
Moreira, F. A., Kaiser, N., Monory, K., & Lutz, B. (2008). *Reduced anxiety-like behaviour induced by genetic and pharmacological inhibition of the endocannabinoid-degrading enzyme fatty acid amide hydrolase (FAAH) is mediated by CB1 receptors*. Neuropharmacology, 54(1), 141-150.
Murphy, H. (2019). *At 71, she's never felt pain or anxiety. Now Scientists Know Why*. New York Times.
Namburi, P., Al-Hasani, R., Calhoon, G. G., Bruchas, M. R., & Tye, K. M. (2016). *Architectural representation of valence in the limbic system*. Neuropsychopharmacology, 41(7), 1697-1715.
Patriarchi, T., Cho, J. R., Merten, K., Howe, M. W., Marley, A., Xiong, W. H., ... & Tian, L. (2018). *Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors*. Science, 360(6396), eaat4422.
Pascoli, V., Terrier, J., Hiver, A., & Lüscher, C. (2015). *Sufficiency of mesolimbic dopamine neuron stimulation for the progression to addiction*. Neuron, 88(5), 1054-1066.
Pertsemlidis, A., & Fondon, J. W. 3rd. (2001). Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biology, 2(10), REVIEWS2002. https://doi.org/10.1186/gb-2001-2-10-reviews2002
Poulin, J. F., Zou, J., Drouin-Ouellet, J., Kim, K. Y. A., Cicchetti, F., & Awatramani, R. B. (2014). *Defining midbrain dopaminergic neuron diversity by single-cell gene expression profiling*. Cell Reports, 9(3), 930-943.

Pupe, S., & Wallén-Mackenzie, Å. (2015). *Cre-driven optogenetics in the heterogeneous genetic panorama of the VTA*. Trends in neurosciences, 38(6), 375-386.
Puzis, R., Farbiash, D., Brodt, O., Elovici, Y., & Greenbaum, D. (2020). Increased cyber-biosecurity for DNA synthesis. Nature Biotechnology, 38(12), 1379–1381.
Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA. (2010, October 13). Health and Human Services Department. https://www.federalregister.gov/documents/2010/10/13/2010-25728/screening-framework-guidance-for-providers-of-synthetic-double-stranded-dna
Screening Framework Guidance for Providers and Users of Synthetic Nucleic Acids. (2023, October). https://aspr.hhs.gov/legal/synna/Documents/SynNA-Guidance-2023.pdf
Simcikas, S. (2019, February 20). *Rodents farmed for pet snake food*. Rethink Priorities. https://rethinkpriorities.org/research-area/rodents-farmed-for-pet-snake-food/
Simeone, A., Di Salvio, M., Di Giovannantonio, L. G., Acampora, D., Omodei, D., & Tomasetti, C. (2011). *The role of Otx2 in adult mesencephalic–diencephalic dopaminergic neurons*. Molecular neurobiology, 43, 107-113.
Skoubis, P. D., Matthes, H. W., Walwyn, W. M., Kieffer, B. L., & Maidment, N. T. (2001). *Naloxone fails to produce conditioned place aversion in μ-opioid receptor knock-out mice*. Neuroscience, 106(4), 757-763.
Solinas, M., Thiriet, N., Chauvet, C., & Jaber, M. (2010). *Prevention and treatment of drug addiction by environmental enrichment*. Progress in neurobiology, 92(4), 572-592.
Stark, E. A., Vuust, P., & Kringelbach, M. L. (2018). *Music, dance, and other art forms: New insights into the links between hedonia (pleasure) and eudaimonia (well-being)*. Progress in brain research, 237, 129-152.
Stoker, A. K., & Markou, A. (2011). *The intracranial self-stimulation procedure provides quantitative measures of brain reward function*. Mood and Anxiety Related Phenotypes in Mice: Characterization Using Behavioral Tests, Volume II, 307-331.
Stuber, G. D., Stamatakis, A. M., & Kantak, P. A. (2015). *Considerations when using cre-driver rodent lines for studying ventral tegmental area circuitry*. Neuron, 85(2), 439-445.
Tiling HOWTO. BioPerl. Retrieved December 13, 2024, from https://bioperl.org/howtos/Tiling_HOWTO.html
Touriño, C., Oveisi, F., Lockney, J., Piomelli, D., & Maldonado, R. (2010). *FAAH deficiency promotes energy storage and enhances the motivation for food*. International journal of obesity, 34(3), 557-568.
Tumpey, T. M., Basler, C. F., Aguilar, P. V., Zeng, H., Solórzano, A., Swayne, D. E., Cox, N. J., Katz, J. M., Taubenberger, J. K., Palese, P., & García-Sastre, A. (2005). *Characterization of the reconstructed 1918 Spanish influenza pandemic virus*. Science (New York, N.Y.), 310(5745), 77–80. https://doi.org/10.1126/science.1119392
Viereckel, T., Dumas, S., Smith-Anttila, C. J., Vlcek, B., Bimpisidis, Z., Lagerström, M. C., ... & Wallén-Mackenzie, Å. (2016). *Midbrain gene screening identifies a new mesoaccumbal glutamatergic pathway and a marker for dopamine cells neuroprotected in Parkinson's disease*. Scientific Reports, 6(1), 35203.
Wang, X. F., Barbier, E., Chiu, Y. T., He, Y., Zhan, J., Bi, G. H., ... & Xi, Z. X. (2016). *T394A mutation at the μ opioid receptor blocks opioid tolerance and increases*

vulnerability to heroin self-administration in mice. Journal of Neuroscience, 36(40), 10392-10403.
Weidner, T. C., Vincenz, D., Brocka, M., Tegtmeier, J., Oelschlegel, A. M., Ohl, F. W., ... & Lippert, M. T. (2020). *Matching stimulation paradigms resolve apparent differences between optogenetic and electrical VTA stimulation*. Brain Stimulation, 13(2), 363-371.
Wei, Y., Zhou, Y., Liu, Y., Ying, W., Lv, R., Zhao, Q., ... & Zhou, C. (2021). *Indiscriminate ssDNA cleavage activity of CRISPR-Cas12a induces no detectable off-target effects in mouse embryos*. Protein & cell, 12(9), 741-745.
Wise, R. A. (1996). *Addictive drugs and brain stimulation reward*. Annual review of neuroscience, 19, 319-340.
Yilmaz, M., & Meister, M. (2013). *Rapid innate defensive responses of mice to looming visual stimuli*. Current Biology, 23(20), 2011-2015.

6.5 Appendices

Appendix A: Adding mutations to a sequence

It is easy to add mutations to a sequence with BioPython, but I initially wanted a variety of options and customizable control over the types of mutations being made. Not only did I want to be able to create both synonymous or nonsynonymous mutations at fixed intervals, but I also wanted to give priority to codons that would make a wobble pairing if an oligo was later used to correct the synonymous mutation. This simple code starts with the standard dictionary for codon translation and builds replacement dictionaries based on several different criteria, and was used for the Mutate-and-Fix design strategy design.

Unfortunately, this code does not automatically detect the reading frame, so the user needs to give the input sequence so it's at the start of a codon. However, this could be fixed in another function. One potential improvement on this code would be adding more functions so a user could input a Genbank file with a specified CDS, and the code would output a Genbank file with mutations introduced at the specific intervals to that CDS such that each mutated codon was also appropriated annotated with the relevant mutation.

Imports

```
import random
import numpy as np
import csv
import zipfile
import csv
import os
from google.colab import drive
import datetime
drive.mount('/content/drive')
```

Helper Functions

```
# helper functions for making some of the dictionaries
def do_codons_differ_by_one_base(codon1, codon2):
 # checks if two codons differ by one base
 differences = 0
 for i in range(3):
   if codon1[i] != codon2[i]:
     differences += 1
 if differences == 1:
   return True
 else:
   return False
```

```
def do_codons_differ_by_a_wobble_base(key_codon, value_codon):
 # checks if two codons differ by a wobble base
 if do_codons_differ_by_one_base(key_codon, value_codon):
   for i in range(3):
     if key_codon[i] != value_codon[i]:
       if key_codon[i] == 'A' and value_codon[i] == 'G':
         return True
       elif key_codon[i] == 'C' and value_codon[i] == 'T':
         return True
       else:
         return False
 else:
   return False

def do_codons_have_the_same_translation(codon1, codon2):
 if codon_to_one_letter_dict[codon1] == codon_to_one_letter_dict[codon2]:
   return True
 else:
   return False
```

Basic Codon Dictionaries

```
# keys are codons
# values are one character strings corresponding to the
# amino acid abbreviation
codon_to_one_letter_dict = {'TCA': 'S', 'TCC': 'S', 'TCG': 'S', 'TCT': 'S',
'TTC': 'F', 'TTT': 'F', 'TTA': 'L', 'TTG': 'L', 'TAC': 'Y', 'TAT': 'Y', 'TAA':
'*', 'TAG': '*', 'TGC': 'C', 'TGT': 'C', 'TGA': '*', 'TGG': 'W', 'CTA': 'L',
'CTC': 'L', 'CTG': 'L', 'CTT': 'L', 'CCA': 'P', 'CCC': 'P', 'CCG': 'P', 'CCT':
'P', 'CAC': 'H', 'CAT': 'H', 'CAA': 'Q', 'CAG': 'Q', 'CGA': 'R', 'CGC': 'R',
'CGG': 'R', 'CGT': 'R', 'ATA': 'I', 'ATC': 'I', 'ATT': 'I', 'ATG': 'M', 'ACA':
'T', 'ACC': 'T', 'ACG': 'T', 'ACT': 'T', 'AAC': 'N', 'AAT': 'N', 'AAA': 'K',
'AAG': 'K', 'AGC': 'S', 'AGT': 'S', 'AGA': 'R', 'AGG': 'R', 'GTA': 'V', 'GTC':
'V', 'GTG': 'V', 'GTT': 'V', 'GCA': 'A', 'GCC': 'A', 'GCG': 'A', 'GCT': 'A',
'GAC': 'D', 'GAT': 'D', 'GAA': 'E', 'GAG': 'E', 'GGA': 'G', 'GGC': 'G', 'GGG':
'G', 'GGT': 'G'}
```

```
one_letter_to_codon_dict = {'S': ['TCA', 'TCC', 'TCG', 'TCT', 'AGC', 'AGT'],
'F': ['TTC', 'TTT'], 'L': ['TTA', 'TTG', 'CTA', 'CTC', 'CTG', 'CTT'], 'Y':
['TAC', 'TAT'], '*': ['TAA', 'TAG', 'TGA'], 'C': ['TGC', 'TGT'], 'W': ['TGG'],
'P': ['CCA', 'CCC', 'CCG', 'CCT'], 'H': ['CAC', 'CAT'], 'Q': ['CAA', 'CAG'],
'R': ['CGA', 'CGC', 'CGG', 'CGT', 'AGA', 'AGG'], 'I': ['ATA', 'ATC', 'ATT'],
'M': ['ATG'], 'T': ['ACA', 'ACC', 'ACG', 'ACT'], 'N': ['AAC', 'AAT'], 'K':
['AAA', 'AAG'], 'V': ['GTA', 'GTC', 'GTG', 'GTT'], 'A': ['GCA', 'GCC', 'GCG',
```

```
'GCT'], 'D': ['GAC', 'GAT'], 'E': ['GAA', 'GAG'], 'G': ['GGA', 'GGC', 'GGG',
'GGT']}
codons_list = list(codon_to_one_letter_dict.keys())
```

Custom Codon Dictionaries

```
# keys are codons
# each value is a list of codons (not including the key) that
# code for the same amino acid
codon_to_list_of_syn_codons_dict = {}
for codon in codons_list:
 syn_codons = []
 for codon2 in one_letter_to_codon_dict[codon_to_one_letter_dict[codon]]:
   if codon != codon2:
     syn_codons += [codon2]
 codon_to_list_of_syn_codons_dict[codon] = syn_codons

# keys are codons
# values are codons that code for the same amino acid that only differ
# from the key by one nucleotide
codon_to_list_of_close_syn_codons_dict = {}
for codon in codons_list:
 syn_codons = []
 for codon2 in codon_to_list_of_syn_codons_dict[codon]:
   if do_codons_differ_by_one_base(codon,codon2):
     syn_codons += [codon2]
 codon_to_list_of_close_syn_codons_dict[codon] = syn_codons

# keys are codons
# values are codons code for the same amino acid that (1) only differ
# from the key by one nucleotide and (2) would create a wobble pairing
# with the antisense repair template
# (this would be replacing an A in the codon with a G or an
# C in the codon with a T)
codon_to_list_of_wobble_syn_codons_dict = {}
for codon in codons_list:
 syn_codons = []
 for codon2 in codon_to_list_of_close_syn_codons_dict[codon]:
   if do_codons_differ_by_a_wobble_base(codon,codon2):
     syn_codons += [codon2]
 codon_to_list_of_wobble_syn_codons_dict[codon] = syn_codons

# keys are codons
# values are a list of the codons code for different amino acids, but only
# differ by one base pair
```

```
codon_to_list_of_close_nonsyn_codons_dict = {}
for codon in codons_list:
 nonsyn_codons = []
 for codon2 in codons_list:
   if do_codons_have_the_same_translation(codon,codon2) == False:
     if do_codons_differ_by_one_base(codon,codon2) == True:
       nonsyn_codons += [codon2]
 codon_to_list_of_close_nonsyn_codons_dict[codon] = nonsyn_codons

# keys are codons
# values are a list of codons that code for different amino acids,
# but only differ from the key by a wobble base
# priority is given to codons that would create a wobble pairing
# with the antisense repair template
# (this would be replacing an A in the codon with a G or an
# C in the codon with a T)
codon_to_list_of_wobble_nonsyn_codons_dict = {}
for codon in codons_list:
 nonsyn_codons = []
 for codon2 in codons_list:
   if do_codons_have_the_same_translation(codon,codon2) == False:
     if do_codons_differ_by_a_wobble_base(codon,codon2):
       nonsyn_codons += [codon2]
 codon_to_list_of_wobble_nonsyn_codons_dict[codon] = nonsyn_codons

# keys are codons
# values are a list of codons that code for different amino acids, but
# only differ by one base pair
# priority given to codons that would create a wobble pairing
# with the antisense repair template
# (this would be replacing an A in the codon with a G or an
# C in the codon with a T)
codon_to_list_of_wobble_or_close_nonsyn_codons_dict = {}
for codon in codons_list:
 nonsyn_codons = []
 if codon_to_list_of_wobble_nonsyn_codons_dict[codon] != []:
     nonsyn_codons = codon_to_list_of_wobble_nonsyn_codons_dict[codon]
 else:
     nonsyn_codons = codon_to_list_of_close_nonsyn_codons_dict[codon]
 codon_to_list_of_wobble_or_close_nonsyn_codons_dict[codon] = nonsyn_codons
```

Functions for Replacing Codons

```
def change_ith_codon(dnaA, i, new_codon):
 dnaB = dnaA[:3*(i-1)] + new_codon + dnaA[3*i:]
 return dnaB

def replace_ith_codon_based_on_dictionary(dnaA, i, dictA):
 codon_before = dnaA[3*(i-1):3*(i)].upper()
 new_codon_options = dictA[codon_before]
 #print(codon_before)
 #print(new_codon_options)
 if new_codon_options != []:
     codon_after = random.choice(new_codon_options)
     dnaB = change_ith_codon(dnaA, i, codon_after)
     return (dnaB, codon_before, codon_after)
 else:
     return None

def replace_ith_codon_with_wobble_nonsyn(dnaA, i):
 return replace_ith_codon_based_on_dictionary(dnaA,
                                              i,

codon_to_list_of_wobble_nonsyn_codons_dict)

def replace_ith_codon_with_close_nonsyn(dnaA, i):
 return replace_ith_codon_based_on_dictionary(dnaA,
                                              i,

codon_to_list_of_close_nonsyn_codons_dict)

def replace_ith_codon_with_wobble_syn(dnaA, i):
 return replace_ith_codon_based_on_dictionary(dnaA,
                                              i,

codon_to_list_of_wobble_syn_codons_dict)

def replace_ith_codon_with_close_syn(dnaA, i):
 return replace_ith_codon_based_on_dictionary(dnaA,
                                              i,

codon_to_list_of_close_syn_codons_dict)
```

Genbank annotations

```
"""The annotation_string is a string that can be added to the appropriate
region of the relevant Genbank file. Ideally, the GenBank file would be read,
and this would be added automatically, but that wasn't my workflow. However, if
you use or are working with people who use GUIs (e.g., Benchling) for storing
and handling DNA sequences, it is very nice to have the mutated codons labeled.
It's good for troubleshooting, too."""
def add_nonsyn_mutation_every_i_codons(dnaA, i, j):

 part1 = """     misc_feature    """
 #start_index
 part2 = '''..'''
 #stop_index
 part3 = '''
                     /label=\"'''
 #name
 part4 = '''\"\n'''

 offset = 0
 oligo_list = []
 annotation_string = ""
 j0 = j
 k = 1

 dnaA = dnaA.lower()
 string1 = ''
 while 3*j <= len(dnaA):
   #print(j)
   new_dnaA = replace_ith_codon_with_wobble_nonsyn(dnaA, j)
   new_dnaB = replace_ith_codon_with_close_nonsyn(dnaA, j)
   if new_dnaA != None:
     dnaA = new_dnaA[0]
     name = str(j) + ': ' + new_dnaA[1] + ' to ' + new_dnaA[2] #+ '\n'
   elif new_dnaB != None:
     dnaA = new_dnaB[0]
     name = str(j) + ': ' + new_dnaB[1] + ' to ' + new_dnaB[2] #+ '\n'
   else:
     print('error')
     return None
   string1 = name
   new_index = 3*(j-1)
   string1 = part1 + str(new_index) + part2 + str(new_index+2) + part3 + name +
part4
   oligo_start = np.max([new_index-12, 0])
   oligo_stop = np.min([len(dnaA), new_index+12])
```

```
   oligo_list += [['strand1_oligo_' + str(k), dnaA[oligo_start:oligo_stop],
name]]
   annotation_string += string1
   j += i
   k += 1


 return dnaA, annotation_string, oligo_list
```

Example Use and Output

```
# 1
k = 16
# frag 1
fragment_01 =
'ggtgggaacactatagtcatctggTTTacgccgttgcgacctggttatgcttcgggagtacgagtggctggagtttt
accctcgaagacaacaatatctttccgaagcagtatccgatcattaatttcacaactgcaggagcgaccgttcaatcat
ataccaacttcattcgggcggtaagaggacgccttaccacaggtgcagacgtgcggcacgagatccccgttctcccgaa
tcgggtgggccttcccatcaatcagcgcttcatactggtggagttgagcaaccacgccgaactaagtgtcacgctcgca
ctcgacgttacaaatgcttacgtggttggttatcgagcaggtaactctgccta'

new_fragment_01 = add_nonsyn_mutation_every_i_codons(fragment_01.upper(), k, 1)

dna1, annotations1, oligos1 = new_fragment_01

print(dna1)
print()
print(annotations1)
print()
print(oligos1)
```

Example Output

```
CGTgggaacactatagtcatctggttttacgccgttgcgacctggttaTTCttcgggagtacgagtggctggagtttta
ccctcgaagacaacaatGTCtttccgaagcagtatccgatcattaatttcacaactgcaggagcgATCgttcaatcata
taccaacttcattcgggcggtaagaggacgccttGCCacaggtgcagacgtgcggcacgagatccccgttctcccgaat
cggGCGggccttcccatcaatcagcgcttcatactggtggagttgagcaacTACgccgaactaagtgtcacgctcgcac
tcgacgttacaaatgcttacATGgttggttatcgagcaggtaactctgccta

     misc_feature    0..2
                     /label="1: GGT to CGT"
     misc_feature    48..50
                     /label="17: TGC to TTC"
     misc_feature    96..98
                     /label="33: ATC to GTC"
```

```
     misc_feature    144..146
                     /label="49: ACC to ATC"
     misc_feature    192..194
                     /label="65: ACC to GCC"
     misc_feature    240..242
                     /label="81: GTG to GCG"
     misc_feature    288..290
                     /label="97: CAC to TAC"
     misc_feature    336..338
                     /label="113: GTG to ATG"


[['strand1_oligo_1', 'CGTgggaacact', '1: GGT to CGT'], ['strand1_oligo_2',
'gcgacctggttaTTCttcgggagt', '17: TGC to TTC'], ['strand1_oligo_3',
'gaagacaacaatGTCtttccgaag', '33: ATC to GTC'], ['strand1_oligo_4',
'actgcaggagcgATCgttcaatca', '49: ACC to ATC'], ['strand1_oligo_5',
'agaggacgccttGCCacaggtgca', '65: ACC to GCC'], ['strand1_oligo_6',
'ctcccgaatcggGCGggccttccc', '81: GTG to GCG'], ['strand1_oligo_7',
'gagttgagcaacTACgccgaacta', '97: CAC to TAC'], ['strand1_oligo_8',
'acaaatgcttacATGgttggttat', '113: GTG to ATG']]
```

I'm using pre-built modules instead of going from scratch, as I did in the previous appendix because designing efficient Golden Gate assemblies is a semi-solved problem. Here, I import, install, and build all of NumberJack just to get a solution that uses data for Golden Gate optimization. It is possible to go through the relevant GitHub and make a more minimum set of dependencies instead of installing all of NumberJack. Also, I needed to install Numberjack separately since I did this on a Google Colab notebook. Still, you could skip ahead to installing goldhinges without this step if your system already has it.

The library I'm using to design Golden Gate assemblies is called Golden Hinges. Golden Hinges does not require but works well with tatapov. The section "Using experimental annealing data from Potapov 2018" at:
https://github.com/Edinburgh-Genome-Foundry/GoldenHinges
explains how this is being used. You can also find the install overview view in this link.

For useful examples of using Golden Hinges, you can check the GitHub:
https://edinburgh-genome-foundry.github.io/GoldenHinges/

Those unfamiliar with Golden Gate Cloining may benefit from checking the Wikipedia page, https://en.wikipedia.org/wiki/Golden_Gate_Cloning, before going to the source project: https://pypi.org/project/goldenhinges/.
Here are some additional helpful links:

Installs and Imports
Please note that this is done in Google Colab, and installs will vary by system. These installs may take a while.

```
!apt-get remove -y swig
!apt-get install -y swig3.0
!ln /usr/bin/swig3.0 /usr/bin/swig
!wget
https://github.com/Edinburgh-Genome-Foundry/Numberjack/archive/v1.2.0
.tar.gz
!tar -zxvf v1.2.0.tar.gz
!cd Numberjack-1.2.0 && python setup.py build -solver Mistral &&
python setup.py install
!pip install Numberjack-1.2.0/
!pip install BioPython==1.79
!pip install goldenhinges
!pip install tatapov
!pip install DnaChisel
!pip install geneblocks
!apt-get install ncbi-blast+
```

```
import Bio
import tatapov
import zipfile
import csv

from Bio.Restriction import AllEnzymes

# for alignments
from Bio import pairwise2
from Bio.pairwise2 import format_alignment

from goldenhinges import OverhangsSelector, reports
import dnachisel
import geneblocks
from dnachisel import (EnforceTranslation, Specification,
SpecEvaluation,
                       reverse_translate, random_protein_sequence,
Location,
                       DnaOptimizationProblem, AvoidPattern,
sequences_differences,
                       EnforceGCContent, AvoidChanges)

import os
from dnachisel import UniquifyAllKmers
from dnachisel import AvoidHairpins

################## Mounting Google Drive ###################
# this section is needed to read and write files from Google Drive
and
# creates a dedicated folder for storing results

from google.colab import drive
import datetime
drive.mount('/content/drive')
filename1 = '/content/drive/MyDrive/' +
datetime.datetime.now().strftime("%Y%m%d-%H%M%S")
os.mkdir(filename1)
# Replace '/content/drive/MyDrive/source_files' with where source
```

```
files
# are being stored
source_files = '/content/drive/MyDrive/source_files/'
```

Generating Overhang Selectors

```
# generating selectors takes a minute, so if you're planning many
# Golden Gate assemblies, you may only want to generate the
# selectors once, which is why it's in a separate function

# generate_selectors() is the more optimized selector. ideally, use
# this one. if it fails, you can try again to relax some of the #
constraints. for instance, the range of acceptable GC content
# can be expanded to 20-80% using gc_min and gc_max. alternatively,
# the cross annealing threshold can be increased. If none of these
# work, you can use selector2. in any case, you should also
# manually review the feasibility of the proposed Golden Gate
# reaction. for more, see
# https://www.biorxiv.org/content/10.1101/322297v1
def generate_selectors():
 annealing_data = tatapov.annealing_data['37C']['01h']

 self_annealings = tatapov.relative_self_annealings(annealing_data)
 weak_self_annealing_overhangs = [
     overhang
     for overhang, self_annealing in self_annealings.items()
     if self_annealing < 0.05
 ]

 cross_annealings = tatapov.cross_annealings(annealing_data)
 high_cross_annealing_pairs = [
     overhang_pair
     for overhang_pair, cross_annealing in cross_annealings.items()
     if cross_annealing > 0.005
 ]

 selector1 = OverhangsSelector(
     forbidden_overhangs=weak_self_annealing_overhangs,
     forbidden_pairs=high_cross_annealing_pairs,
     gc_min=0.25,
```

```
    gc_max=0.75,
 )

 # selector 2 is not as good. it may use self-annealing
 # overhangs identified as weak by the literature, or it may use
 # overhang pairs with significant cross-talk.
 selector2 = OverhangsSelector(
    gc_min=0.25,
    gc_max=0.75,
    differences=1)

 return selector1, selector2

 return selector1, selector2
```

Example Usage

```
 # format for gb_files is (    filename,
 #                             number_of_fragments_in_assembly,
 #                             restriction_enzyme_for_assembly )
gb_filenames = [('sequence1.gb', 4, 'BsaI'),
                ('sequence2.gb', 3, 'PaqCI')]

# location of where sequence1.gb and sequence2.gb are
 source_files = path_to_source_files

 where_to_put_zip = filename1 + 'full_report1.zip'
 where_to_put_unzipped_file = filename1
for tuple1 in gb_filenames:
 gb_filename, number_of_fragments1, enzyme1 = tuple1
 print(gb_filename)

 # add additional enzymes and corresponding flanks as needed
 if enzyme1 == 'PaqCI':
   left_flank = 'CACCTGCatgc'
   right_flank = 'gcatGCAGGTG'
 elif enzyme1 == 'BsaI':
   left_flank = 'GGTCTCa'
   right_flank = 'tGAGACC'
 else:
   break

 date1 = datetime.datetime.now().strftime("%Y%m%d-%H%M%S")
```

```
 filename1 = '/content/drive/MyDrive/' + date1 + '_ggate_' + gb_filename +
"_outputs" + '/'
 os.mkdir(filename1)

 selectorA, selectorB = generate_selectors2()
 sequence1 = Bio.SeqIO.read(open(source_files+gb_filename,"r"), "genbank")
 number_of_fragments = number_of_fragments1
 allow_edits_to_cds = True
 left_flank1 = left_flank
 right_flank1 = right_flank

 sequence1.annotations["molecule_type"] = "DNA"

 # checks to see if the Potapov-informed selector works, and if not, it uses a
 # less constraining selector
 solution1 = selectorA.cut_sequence(sequence1,
                                    equal_segments=number_of_fragments,
                                    include_extremities=False,
                                    allow_edits=allow_edits_to_cds)
 if type(solution1) == type(None):
   print("overhang selector using Potapov (2018) failed, using worse selector")
   solution1 = selectorB.cut_sequence(sequence1,
                                      equal_segments=number_of_fragments,
                                      include_extremities=False,
                                      allow_edits=allow_edits_to_cds)

 reports.write_report_for_cutting_solution(solution1,
                                           target=where_to_put_zip,
                                           sequence=sequence1,
                                           display_positions=True,
                                           left_flank=left_flank1,
                                           right_flank=right_flank1
                                           )
 with zipfile.ZipFile(where_to_put_zip, 'r') as zip_ref:
   zip_ref.extractall(where_to_put_unzipped_file)

 with open(where_to_put_unzipped_file+'fragments_sequences.csv', newline='\n')
as csvfile:
   data = list(csv.reader(csvfile, delimiter=';'))
 goal_fragments = []
 for item in data:
   goal_fragments.append(item[1])
```

Appendix C: A Table of Decoy Sequences Used for Sequence Obfuscation

Description	Accession Number
Homo sapiens mRNA for immunoglobulin kappa heavy chain (IgG1K)	Y14735.1
Influenza A virus (A/Bewick's swan/Netherlands/1/2007(H1N5)) segment 4	CY076976.1
Influenza A virus (A/Brisbane/59/2007(H1N1)) polymerase PB2 (PB2) gene, complete cds	CY163871.1
Influenza A virus (A/Brisbane/59/2007(H1N1)) segment 1 sequence	CY064971.1
Influenza A virus (A/Brisbane/59/2007(H1N1)) segment 2 sequence	CY064972.1
Influenza A virus (A/Brisbane/59/2007(H1N1)) segment 3 sequence	CY064973.1
Influenza A virus (A/Brisbane/59/2007(H1N1)) segment 7 sequence	CY031391.1
Influenza A virus (A/Brisbane/59/2007(H1N1)) segment 8 sequence	CY064978.1
Influenza A virus (A/Duck/Hong Kong/Y439/97(H9N2)) segment 7 matrix protein M1 (M1) and matrix protein M2 (M2) genes, complete cds	AF156462.1
Influenza A virus (A/Solomon Islands/3/2006(H1N1)) segment 5 sequence	CY047398
Influenza A virus (A/Solomon Islands/3/2006(H1N1)) segment 6 neuraminidase (NA) gene, partial cds	EU124136
Influenza A virus (A/swine/Hannover/1/1981(H1N1)) segment 6 neuraminidase (NA) gene, complete cds	KJ889374.1
Influenza A virus (A/tufted duck/Fukushima/5/2011(H5N1)) PA gene for polymerase acidic protein, partial cds	AB675535.1

Human Oprm1

>sp|P35372|OPRM_HUMAN Mu-type opioid receptor OS=Homo sapiens OX=9606 GN=OPRM1 PE=1 SV=2

MDSSAAPTNASNCTDALAYSSCSPAPSPGSWVNLSHLDGNLSDPCGPNRTDLGGRDSLCP
PTGSPSMITAITIMALYSIVCVVGLFGNFLVMYVIVRYTKMKTATNIYIFNLALADALAT
STLPFQSVNYLMGTWPFGTILCKIVISIDYYNMFTSIFTLCTMSVDRYIAVCHPVKALDF
RTPRNAKIINVCNWILSSAIGLPVMFMATTKYRQGSIDCTLTFSHPTWYWENLLKICVFI
FAFIMPVLIITVCYGLMILRLKSVRMLSGSKEKDRNLRRITRMVLVVVAVFIVCWTPIHI
YVIIKALVTIPETTFQTVSWHFCIALGYTNSCLNPVLYAFLDENFKRCFREFCIPTSSNI
EQQNSTRIRQNTRDHPSTANTVDRTNHQLENLEAETAPLP

Mouse Oprm1

>sp|P42866|OPRM_MOUSE Mu-type opioid receptor OS=Mus musculus OX=10090 GN=Oprm1 PE=1 SV=1

MDSSAGPGNISDCSDPLAPASCSPAPGSWLNLSHVDGNQSDPCGPNRTGLGGSHSLCPQT
GSPSMVTAITIMALYSIVCVVGLFGNFLVMYVIVRYTKMKTATNIYIFNLALADALATST
LPFQSVNYLMGTWPFGNILCKIVISIDYYNMFTSIFTLCTMSVDRYIAVCHPVKALDFRT
PRNAKIVNVCNWILSSAIGLPVMFMATTKYRQGSIDCTLTFSHPTWYWENLLKICVFIFA
FIMPVLIITVCYGLMILRLKSVRMLSGSKEKDRNLRRITRMVLVVVAVFIVCWTPIHIYV
IIKALITIPETTFQTVSWHFCIALGYTNSCLNPVLYAFLDENFKRCFREFCIPTSSTIEQ
QNSARIRQNTREHPSTANTVDRTNHQLENLEAETAPLP

Sequence alignment from NCBI blastp:

Query: sp|P35372|OPRM_HUMAN Mu-type opioid receptor OS=Homo sapiens OX=9606 GN=OPRM1 PE=1 SV=2 Query ID: lcl|Query_5269638 Length: 400

>sp|P42866|OPRM_MOUSE Mu-type opioid receptor OS=Mus musculus OX=10090 GN=Oprm1 PE=1 SV=1

Sequence ID: Query_5269641 Length: 398

Range 1: 1 to 398

Score:756 bits(1951), Expect:0.0,
Method:Compositional matrix adjust.,
Identities:376/400(94%), Positives:385/400(96%), Gaps:2/400(0%)

```
Human  1    MDSSAAPTNASNCTDALAYSSCSPAPSPGSWVNLSHLDGNLSDPCGPNRTDLGGRDSLCP  60
            MDSSA P N S+C+D LA +SCSPAP  GSW+NLSH+DGN SDPCGPNRT LGG  SLCP
Mouse  1    MDSSAGPGNISDCSDPLAPASCSPAP--GSWLNLSHVDGNQSDPCGPNRTGLGGSHSLCP  58

Human  61   PTGSPSMITAITIMALYSIVCVVGLFGNFLVMYVIVRYTKMKTATNIYIFNLALADALAT  120
             TGSPSM+TAITIMALYSIVCVVGLFGNFLVMYVIVRYTKMKTATNIYIFNLALADALAT
Mouse  59   QTGSPSMVTAITIMALYSIVCVVGLFGNFLVMYVIVRYTKMKTATNIYIFNLALADALAT  118
```

```
Human  121  STLPFQSVNYLMGTWPFGTILCKIVISIDYYNMFTSIFTLCTMSVDRYIAVCHPVKALDF  180
            STLPFQSVNYLMGTWPFG ILCKIVISIDYYNMFTSIFTLCTMSVDRYIAVCHPVKALDF
Mouse  119  STLPFQSVNYLMGTWPFGNILCKIVISIDYYNMFTSIFTLCTMSVDRYIAVCHPVKALDF  178


Human  181  RTPRNAKIINVCNWILSSAIGLPVMFMATTKYRQGSIDCTLTFSHPTWYWENLLKICVFI  240
            RTPRNAKI+NVCNWILSSAIGLPVMFMATTKYRQGSIDCTLTFSHPTWYWENLLKICVFI
Mouse  179  RTPRNAKIVNVCNWILSSAIGLPVMFMATTKYRQGSIDCTLTFSHPTWYWENLLKICVFI  238


Human  241  FAFIMPVLIITVCYGLMILRLKSVRMLSGSKEKDRNLRRITRMVLVVVAVFIVCWTPIHI  300
            FAFIMPVLIITVCYGLMILRLKSVRMLSGSKEKDRNLRRITRMVLVVVAVFIVCWTPIHI
Mouse  239  FAFIMPVLIITVCYGLMILRLKSVRMLSGSKEKDRNLRRITRMVLVVVAVFIVCWTPIHI  298


Human  301  YVIIKALVTIPETTFQTVSWHFCIALGYTNSCLNPVLYAFLDENFKRCFREFCIPTSSNI  360
            YVIIKAL+TIPETTFQTVSWHFCIALGYTNSCLNPVLYAFLDENFKRCFREFCIPTSS I
Mouse  299  YVIIKALITIPETTFQTVSWHFCIALGYTNSCLNPVLYAFLDENFKRCFREFCIPTSSTI  358
                                                      *    *


Human  361  EQQNSTRIRQNTRDHPSTANTVDRTNHQLENLEAETAPLP  400
            EQQNS RIRQNTR+HPSTANTVDRTNHQLENLEAETAPLP
Mouse  359  EQQNSARIRQNTREHPSTANTVDRTNHQLENLEAETAPLP  398
                                        *
```

Notes:
The C348 and C353 residues in the human Oprm1 protein sequence correspond to C346 and C351 in the mouse sequence. Additionally, T394 in the mouse sequence corresponds to T394 in the human sequence.