Studies on Scaling Throughput in Protein Engineering

Thesis by

Lucas Jean Nicolas Schaus

In Partial Fulfillment of the Requirements for

the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2025

(Defended July 2nd 2024)

Lucas Schaus
ORCID: **0000-0002-6094-7402**

ACKNOWLEDGEMENTS

I would first like to thank my PI, Stephen L. Mayo, who has been an immensely supportive mentor over the past two and a half years. While working on a PhD is not always easy, working with Steve has always been a pleasure. He always had creative inputs to my problems and would always know what way to push a projects direction into was best. I thoroughly enjoyed my time in this lab and seeing it grow from just Sarah, Monica, Steve and I at some point to a flourishing lab was really great to experience.

I would also like to thank Monica Breckow for her support to all of us in the lab and always being a great person to talk to when I felt like just wandering around the lab.

I would like to thank all my coworkers in the Mayo lab: Adrian Bahn, Ali Nazi, Andres Orta, Blade Olsen, Jesse Gan, Mikhail Hameedi, Nick Friesenhahn, Olive Cheng, Sarah Gillespie, and former lab members that I overlapped with Arielle Tycko, Dr. Jingzhou Wang and Dr. Shang Huan. Special thanks to Adrian Bahn, Andres Orta, Dr. Anuvab Das, Dr. Anders Knight, Dr. Marc Garcia-Borras, Mikhail Hameedi, Jesse Gan, Arielle Tycko, Dr. Rebecca Warmack, Yuanbo Shen, Olive Cheng and Dr. Xiongyi Huang for being amazing collaborators. I would also like to thank Blade Olsen, whom I may not have worked with directly much but who always was a great helping me get things going in the lab and whom I had many interesting discussions with. Alex Cohen has also been an essential friend and colleague that has helped me understand antibodies and immunology to the point that I felt confident to start a project.

The people at Caltech that I owe a lot to that got me through my PhD, I cannot thank them enough for the support they have given me through sometimes very hard times. Prof. Bil Clemons, who showed me that I do belong at Caltech when I needed to hear it the most. Dr. Sabine Brinkman-Chen whom I believe is one of the most kind-hearted people I have ever met. When I was ready to quit, she convinced me to keep going and told me that everything is going to be alright. I have to admit she is also the toughest editor I have ever had, but I learned so much from her through her tough (editing) love. I want to thank the people in my cohort as well, Nicholas Sarai, Giovanni Tomaleri, Andrew Schacht and Anna Karen Orta. Even though I arrived at Caltech a term later than planned, they welcomed me with open arms and became great friends of mine.

This wouldn’t be a thesis written between 2019 and 2024 if I did not mention my pandemic pod: Carina Jette, Roscoe Lindstadt, George Mobbs, Taylor Stevens, Matt Levine, Jessica Slagle-Petrovic and Stefan Petrovic (was that too many people?). They have also become some of my best friends with whom I could have discussions about anything with, but mostly about food and wine (and that is why I love them).

I want to thank my family for being some of the most supportive people I know. Leila, Lino, Mama and Papa. I do not know how I got so lucky to have my family be who they are despite

the hard times we've been through. Our parents have always taught Leila, Lino and I that we should stick together no matter what. While we have managed that emotionally, we have somehow managed in the past decade to always live in different parts of the world than each other. Over the course of my PhD, I also got to welcome three fantastic people into my family, Alain Schumann, Charlotte DeWitte and Goldie June Schumann.

Lastly, I want to thank my adventure buddy, travel companion, climbing partner and love, Akie Jette. I truly got to meet my partner in the middle of the pandemic. I cannot believe how lucky I got to spend my time in the pandemic with her and even luckier to then get to call her my partner. Today I am lucky enough to almost every day wake up next to the smartest, prettiest, and most kind-hearted person I know. We could talk for hours about life, science, food, wine, climbing and anything that comes to our mind. We have gone on so many adventures, spent the nicest days at the beach, and hiked the prettiest mountains together. I love her with every part of my soul.

ABSTRACT

In this work we present three studies in protein engineering. While all three protein classes that have been targeted for engineering tasks are very different, the studies have a focus on scaling-up the throughput in protein engineering.

The first study concerns machine learning (ML) based antibody humanization techniques. Achieving a reduction of patient anti-drug antibody responses in clinical trials is the goal of antibody humanization. To measure this however, one needs to pass significant scientific, bureaucratic, and financial hurdles, which is very rarely done and especially never at scale. Most existing ML-based antibody humanization techniques claim that they work without providing any experimental evidence. We developed Mousify as an *in silico* antibody humanization platform to place existing models into one framework for wet-laboratory validation. We demonstrate that even the best models have a fundamental flaw in that they only generate a single antibody. We use Mousify and Markov chains to show that using ML-based antibody humanization models for library generation is not only feasible but produces both stable and functional variants. Learning the lessons from our wet-laboratory experiments, we then developed a variational autoencoder model with properties that hopefully improve the outcomes of antibody humanization experiments.

In the second study, we outline our plans and initial results to develop a bioelectrocatalytic system for the conversion of N_2 to ammonia using nitrogenase. Most of the world's ammonia is used for agricultural purposes and is produced via the environmentally damaging Haber-Bosch process. Engineering nitrogenase for the bioelectrocatalytic production of ammonia is not trivial and a high throughput is not guaranteed. We present preliminary results in how throughput can be increased through diazotrophic pre-selection of nitrogenase variants, as well as a quest to find the ideal starting point for engineering using a combination of ancestral sequence reconstruction and generative protein language models.

In the third and final study we present a directed evolution campaign to evolve protoglobins for the enantioselective catalytic formation of *cis*-trifluoromethyl substituted cyclopropanes, the first such reaction in both the chemical and biological world. Not only is the enzyme

*Ape*Pgb LQ capable of efficiently performing carbene insertions into double-bonds, but it also shows a much more diverse substrate scope than similar enantioselective formations of *trans*-trifluoromethyl substituted cyclopropanes. After demonstrating that *Ape*Pgb LQ reactions can be increased to a 1-mmol scale, we investigated the nature of protoglobin *cis*-selectivity using various computational methods.

PUBLISHED CONTENT AND CONTRIBUTIONS

L. Schaus, A. Das, A. M. Knight, G. Jimenez-Osés, K. N. Houk, M. Garcia-Borràs, F. H. Arnold, X. Huang, Protoglobin-Catalyzed Formation of *cis*-Trifluoromethyl-Substituted Cyclopropanes by Carbene Transfer, *Angew. Chem. Int. Ed.* 2023, 62, e202208936; *Angew. Chem.* **2023**, *135*, e202208936. doi: 10.1002/ange.202208936
[First Author, Determined reaction scope & scale-up]

S. Petrovic, D. Samanta, T. Perriches, C. J. Bley, K. Thierbach, B. Brown, S. Nie, G. W. Mobbs, T. A. Stevens, X. Liu, G. P. Tomaleri, L. Schaus, A. Hoelz, Architecture of the linker-scaffold in the nuclear pore, *Science* **2022**, 376, 6598. doi: 10.1126/science.abm9798
[Contributor, Solved the structure of Nup93 SOL with Nup53 R2]

Manuscripts in preparation (Working Titles):

L. Schaus, S. Mayo, Antibody dataset management with OAS C/S
[First Author, Manuscript in Preparation]

L. Schaus, A. Bahn, A. Tycko, S. Mayo, Generation of humanized antibody libraries using Mousify
[Co-first Author, Manuscript in Preparation]

L. Schaus, S. Mayo, AbVAE: A Variational Autoencoder for Antibody Humanization
[First Author, Manuscript in Preparation]

L. Schaus, S. Mayo, Mousify: A Software Package for Humanized Antibody Libraries
[First Author, Manuscript in Preparation]

J. Gan, L. Schaus, S. Mayo, Benchmarks for fast structure calculations of antibodies
[Co-first Author, Manuscript in Preparation]

TABLE OF CONTENTS

Chapter 1

SCALING THROUGHPUT IN PROTEIN ENGINEERING

Abstract

In this chapter we will introduce key concepts of molecular biology, biochemistry, and protein engineering to understand how protein evolution works and how protein engineers have hijacked the mechanisms of protein evolution to artificially create protein sequences with new functions. Two important decisions in protein engineering campaigns are the selection pressure a protein engineer uses as well as the mutagenesis method employed to generate variants. Both of these have an impact on throughput. One of the principal jobs of a protein engineer is to maximize throughput in an engineering campaign as it dictates how quickly one can engineer a protein and how diverse the set of sequences is one can test. Lastly, we take a look at a new wave of *in silico* techniques that can have a large impact on throughput in protein engineering. Over the past decade, machine learning models for protein engineering have had a large impact on how biochemists and protein engineers think about proteins and have provided us with new tools to form informed protein sequence libraries, making engineering more efficient.

Contributions

This review chapter was entirely written by Lucas Schaus.

Evolution of Proteins and Thermostability

For cells, the building blocks of life, information only flows in one direction. From DNA to RNA, and from RNA to proteins. A combination of only three polymers dictates the entire functioning of a cell and higher organisms.[1] Using these three tools, cells synthesize or recruit the help of other essential molecules such as lipids or metal ions.[2,3] While RNA biology is an incredibly fascinating subject,[4] full of unanswered questions,[5] for this work, we will focus on proteins and how DNA contains the information to produce them.

Deoxyribonucleic acid (DNA) is a polymer made up of four different nucleobases, adenine (A), cytosine (C), guanine (G), and thymine (T) (Figure 1.1).[1] When nucleobases are connected together, they form a strand of DNA that likes to pair up with another strand of DNA in a helical fashion. Moreover, there are specific rules to DNA strand pairing. The nucleobase A will preferably pair with nucleobase T, and nucleobase C will preferably pair with nucleobase G. While this is portrayed very simplistically and there is a lot more to the structure of DNA,[6,7] effectively a four-letter alphabet and the magic of physics and chemistry is all it takes to encode all of life on earth. This fact was first discovered in 1952 by Alfred Hershey and Martha Chase in a fascinating experiment.[8] While the concept of a gene existed for half a century at that point, [9] it took nearly a decade since the discovery by Hershey and Chase to understand how DNA conveys information to the cell. In 1961, work performed by Francis Crick, Sydney Brenner, Leslie Barnett, and R.J. Watts-Tobin demonstrated that DNA encodes information in sets of three nucleobases at a time, called a codon. Over the course of the next five years, Marshall Nirenberg, Heinrich Matthaei, Philip Leder, and Har Gobind Khorana designed experiments that would culminate at the 1966 Cold Spring Harbor Symposium with the presentation the genetic code.[10,11] This table is the translation from the language of DNA/RNA to proteins and with it ushered a new era of molecular biology where scientists could now start to manipulate DNA and get a desired protein sequence. It is worth noting that since that point, it was discovered that DNA encodes for more than just protein sequences.[12,13]

Similar to DNA, proteins are made up of an alphabet and of a few rules of assembly which encodes for a myriad of functions. Instead of nucleobases, proteins are made up of 20 amino acids (there are more than 20, but any amino acid outside those 20 is rare[14,15]) and amino acids connect together via peptide bonds (Figure 1.1). Each codon, a set of three DNA bases, encodes for an amino acid, with some amino acids being encoded by multiple codons and some codons encoding for the end of a protein sequence (Figure 1.1). To get from DNA to a protein sequence, the DNA first needs to be transcribed to a type of RNA molecule called messenger RNA (mRNA). The mRNA is then bound by the ribosome. This macromolecule is mainly made up of RNA and works as a catalyst that produces proteins. The ribosome reads the information contained on the mRNA strand and recruits another RNA molecule called a transfer RNA (tRNA). The tRNA carries the correct amino acid for a given codon to the ribosome, allowing it to catalyze the peptide bond formation and adding another member to the protein chain. The last three paragraphs describe what is known as the "Central Dogma" of molecular biology and provides us with a sufficient basis to understand the concepts of protein evolution and protein engineering.[1]

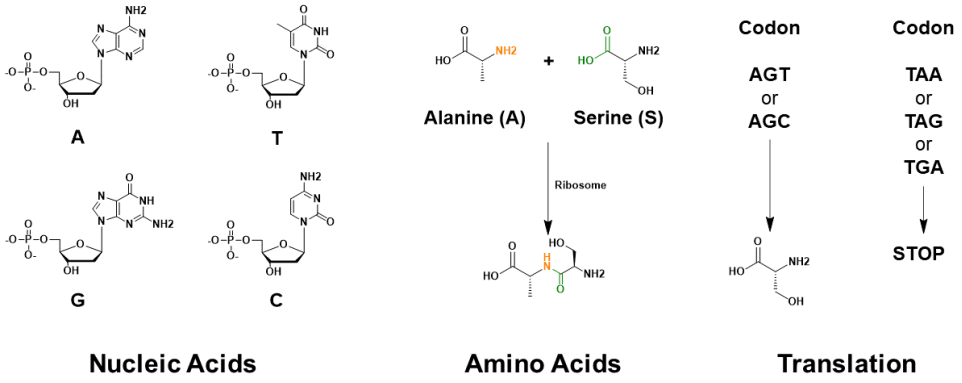


Figure 1.1: Nucleic acids and amino acids make up the building blocks of life, dictating almost all biological processes. (**Left**) Overview of the four deoxyribonucleic acids, adenine (A), cytosine (C), guanine (G), and thymine (T). (**Middle**) Example of two amino acids out of 20 (alanine and serine) and showing the formation of a peptide bond between them catalyzed by the ribosome. (**Right**) DNA codons, sets of three bases, encode the information to incorporate a defined amino acid into a peptide chain. In one example we show that the codons AGT and AGC translate to the amino acid serine. The other example shows that three codons TAA, TAG, and TGA are special in that they do not encode for an amino acid but for the action of stopping translation by the ribosome.

Proteins are fascinating objects. It is sometimes hard to believe that a chain of 20 molecules glued together can produce incredibly complex machinery. Proteins can be turbines that act as powerplants of the cell.[16] They can be two legs walking on a strand carrying cargo from one end of the cell to another.[17] They can read DNA and copy DNA.[18] When we breathe, a protein binds oxygen and transports it to cells that require it.[19] Some organisms have even evolved proteins that work like guns that shoot toxins into adjacent cells.[20] This is only a tiny fraction of the functionalities that nature has evolved proteins to do! So how did nature create such fascinating machinery over billions of years?

To understand protein evolution, we first need to understand how what mutations are. We have established above that DNA contains the genetic information of organisms, however that information is not static. Every time DNA gets copied by DNA polymerase, there is a small chance that the copy accumulated mistakes, e.g. an A was read as a G for example.[21] There are also non-biological factors that can change DNA, such as UV light that can damage DNA bases and lead to changes in the DNA composition.[22] All of these changes to a DNA sequence are considered mutations. When a mutation happens in a region of DNA that encodes a protein sequence, one of three things happens. The mutation is silent, which means that there is in the protein sequence. Consider the two codons for serine shown in Figure 1.1. If the last T is mistakenly interpreted as a C, the ribosome would still incorporate serine at that position. The mutation could also be deleterious. An example of such a change in a DNA sequence would be if a mutation produces one of the stop codons in the middle of the protein sequence, prematurely ending the sequence and often making the protein non-functional. The last two types of mutations are by far the most common types. In very rare cases is a mutation

beneficial. In fact, it is sometimes hard to define when a mutation is beneficial since it all depends on the task that one tests the mutation on. For example, the nicotinic acetylcholine receptor (nAChR) is a cell receptor that responds to the neurotransmitter acetylcholine but can also bind nicotine. In rats, that receptor is more sensitive to nicotine, lowering amount for a lethal dose of the drug 300-fold compared to humans. Researchers have traced back this difference in lethality to a single mutation of a threonine to an isoleucine at position 56 of the receptor.[23] Is this mutation beneficial? Not necessarily, but certainly is to the smoker. Therefore, beneficial mutations are most often context dependent. For protein engineers we define this context in terms of protein "fitness". In other words, if a protein has high fitness for a task, then it is good at performing that task.

In addition to context dependence of single mutations, there is also context to consider when multiple mutations are involved. While so far, we have described proteins as chains of amino acids, they do not look so in the cell. Proteins form highly complex 3-dimensional structures (Figure 1.2) and residues that would normally be distant from each other can be a lot closer in 3-dimensions. For example, residues A112 and F114 of homotetramer cystathionine β-synthase are in contact with each other even though the contact happens between two chains.[24] It turns out that these two residues are evolutionarily coupled, meaning in the course of evolution of the protein, if one of the residues mutated, then the other mutated as well.[25] This concept of evolutionary coupling is very important in the evolution of proteins and researchers have developed tools to identify evolutionarily coupled residues to help in protein engineering and in understanding protein evolution.[26]

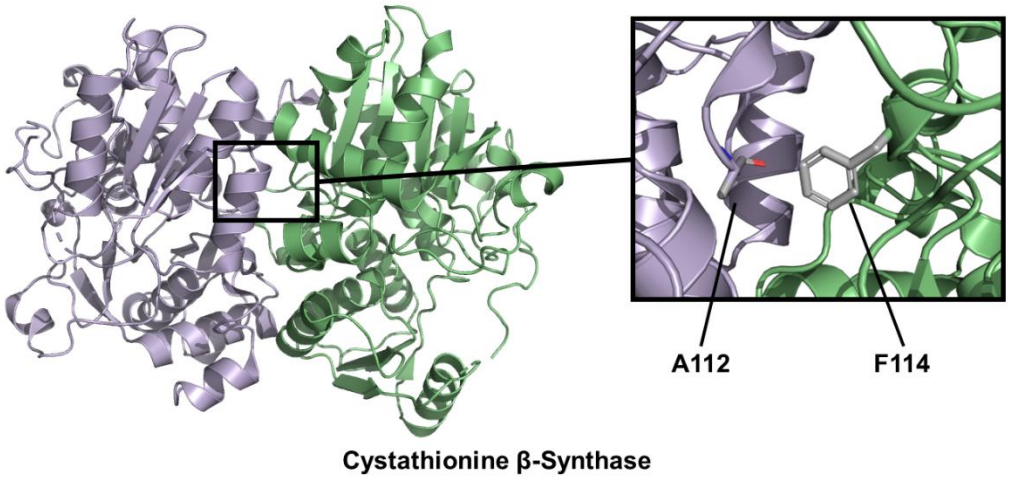


Figure 1.2: Structure of cystathionine β-synthase. The residues A112 and F114 are evolutionarily coupled, meaning if one mutates, then the other one needs to mutate too to not be deleterious.

This concept of evolutionary coupling is part of a broader phenomenon in protein evolution called epistasis.[27] A set of mutations is said to be epistatic if the summed contributions of the individual mutations to protein fitness is different than the fitness of all mutations taken together. To give an abstract example, lets imagine two mutations, $A \rightarrow A^*$ and $B \rightarrow B^*$, at two positions in a protein sequence. This gives us four possible sequences in this set of mutations: AB, A^*B, AB^*, and A^*B^*. Let f(AB) be the fitness of the sequence AB. An example of epistasis would be if $f(A^*B) = 1$, $f(AB^*) = -1$, and $f(A^*B^*) = 2 \neq f(A^*B) + f(AB^*)$. An example of non-epistasis would be if $f(A^*B^*) = 0 = f(A^*B) + f(AB^*)$. There are multiple mathematical formulations for this phenomenon,[28-30] but the fundamental consequence of epistasis is that protein fitness landscapes are rugged with epistasis and smooth without epistasis.[27,29,31] In other words, epistasis makes evolution hard and slow.

We have now seen that most mutations are silent or deleterious, that some protein residues are evolutionarily coupled to others and that epistasis makes exploring a protein fitness landscape hard. Since we observe fascinating protein machinery all around us, there must be at least one mechanism by which proteins manage to evolve towards new functions. While there are many such mechanisms,[32] one in particular is of high interest to protein engineering. In 2006, Jesse Bloom *et al*. suggested a mechanism for protein evolution that is closely related to a proteins' stability.[33] Protein stability is the property of proteins to withstand denaturation, whether thermal or chemical. For simplicity, we will only consider globular proteins to explain protein stability. When a protein is first produced by the ribosome, the amino acid chain immediately starts to fold-up into its 3-dimensional structure. Under normal conditions, this 3-dimensional state is the most stable conformation of the amino acid chain (it could be that it is not the most stable state but a kinetically stable local energy minimum[34]). When globular proteins fold, they usually do it such that hydrophobic amino acids inhabit the core of proteins, and hydrophilic amino acids populate the surface. That is because the hydrophilic residues can interact with water and the hydrophobic residues can interact with each other. When a protein denatures, the amino acid chain unravels and exposes hydrophobic residues to water which do not want to interact with each other (think of oil not dissolving in water). If another protein close-by is also unraveling, the hydrophobic residues of the two chains can interact with each other. At some point enough hydrophobic interactions between chains can happen that they fall into an energy minimum that cannot be easily reversed kinetically, i.e. the protein has denatured.[35] This unraveling of proteins can be induced by increasing the temperature or by exposing the protein to a denaturing chemical. A protein is considered more stable if it can withstand higher temperatures or higher concentrations of denaturing chemicals.

The seminal paper by Jesse Bloom *et al*. claims that thermostability of a protein and evolution are tightly connected to each other.[33] Since at the very least, a protein still needs to fold to a thermodynamically stable structure, any sequence of mutations that move a protein from one function to another must also make sure that the protein retains some minimal stability. As selection pressure only selects for the function of a protein and rarely its stability, a proteins potential extra stability is indifferent to evolution. Through simulations of protein evolution, the authors show that a thermostable protein is more evolvable though. I.e. the protein can

accumulate more mutations to improve its fitness for a task before becoming too unstable. That is because only vanishingly rarely is a mutation beneficial for both a new function and for thermostability. The more likely scenario is that a mutation that is beneficial for a function will reduce the stability of the protein. The authors conclude that functionally neutral mutations for a task accumulate silently in a protein over time. These mutations marginally increase the stability of a protein and allow the accumulation of a functionally positive mutation that is also destabilizing.[33] While very interesting from a protein evolution standpoint, this study has a much larger impact for protein engineers that want to artificially engineer proteins.

Protein Engineering from Low to High Throughput

Protein engineering is the scientific discipline of manipulating the system that nature evolved to produce proteins and hijack the tools of evolution to push a protein towards performing a desired task. Over the years, scientists have developed many techniques that help in engineering proteins. The most important technique in protein engineering is arguably directed evolution (DE). DE was invented by Frances Arnold in the late 1980s, showing that one can randomly induce mutations in the DNA encoding protein sequences to create a protein sequence library, screen them for improved function and picking the best performing variant for another round of mutagenesis.[36] The work was inspired by a paper from John Maynard Smith on evolutionary game theory, where Maynard Smith argues that in order for evolution to work a protein sequence needs to have other functional mutants nearby in sequence space, even if most sequences do not encode functional proteins.[37,38] The advent of DE evolution started a wave of protein engineering work. Protein engineers have evolved enzymes to catalyze reactions that have no equivalent in nature[39-45] DE has also been used to take a computationally designed protein with no intended function and engineered it to catalyze a wide range of reactions.[41,46]

The beauty of protein engineering compared to other engineering disciplines is that incremental improvements are trivial to make as they are inherent to the system itself. Nature evolved with incremental changes to DNA over billions of years, and protein engineers are now using the same system to simulate evolution in a test tube in a matter of days. The only thing that the engineer needs to provide is the selection pressure and a mutagenesis technique for a protein to be engineered. While this may sound simple on the surface, the choice of selection pressure and mutagenesis technique are extremely important and often not trivial. Between the two however, selection pressure and only selection pressure determines the direction in which a protein evolves. An important phrase was coined in a paper by Schmidt-Dannert and Arnold in 1999: “You get what you screen for”.[47] Another key insight in the paper by Bloom *et al*. illustrated this phrase neatly, is that evolution is lazy in a sense, a protein will only every be marginally stable because that is the minimum what it needs to be to survive.[33] “You get what you screen for” means that evolution will always try to find the shortest path to generate a minimum viable variant.

For directed evolution, there are two main methods to produce mutations in a DNA sequence. One makes use of error-prone DNA polymerases, which have a higher rate than natural polymerases to introduce mutations into the DNA sequence. These error-prone polymerases are used to amplify the DNA sequence of interest in an error-prone polymerase chain reaction (epPCR).[48,49] The resulting amplified DNA fragment has a few to many mutations, depending on the polymerase used. Another method to produce mutants in DE is to use site-saturation mutagenesis (SSM). This method makes use of degenerate codons, which are codons that can encode for many different amino acids instead of just one. By adding a mixture of bases during the artificial synthesis of DNA oligomers, a mixture of DNA oligomers is created where each oligomer encodes a different set of amino acids. SSM uses degenerate codons that encode for all 20 amino acids at a position of the protein sequence. SSM usually targets a specific site in the protein sequence based on a hypothesis which is

often times supported by structural data of the protein. This method can also be expanded to target multiple sites, however a DE campaign that targets N sites via site saturation, will produce 20^N different variants in the sequence library.[50] Generally, a protein engineer would prefer epPCR with a high mutation rate or SSM targeting many sites, however a major bottleneck in DE is the throughput with which sequences can be tested. If a generated library has 1000 members, but an engineer can only screen 10 of them in a reasonable timeframe, then the engineer is likely to only observe non-functional proteins since most mutations are deleterious. Broadly, an engineer should aim to only create libraries that are 1-2 orders of magnitude smaller than the throughput of the screen. This is not only a problem in DE, but any protein engineering technique also faces the same problem.

The ideal way to satisfy both the "You get what you screen for" problem and the throughput problem is via selection.[47] If the property one wants to engineer for can be tied to the survival of a microorganism, then one is only limited by the transformation efficiency in how many variants can be tested. Transforming that organism with a library and subjecting it to an experiment where only fit variants survive will automatically propagate the best variants. This technique has been used successfully many times to evolve proteins quickly and to extremely high efficiencies for the task at hand. In a paper showcasing selection well, Fahrig-Kamarauskait *et al*. added 4-fluorophenylalanine to the growth media of *E*. *coli* chorismate mutase knockouts which would normally kill the cells. Chorismate mutase is an essential gene in the pathway to produce phenylalanine and the incorporation of 4-fluorophenylalanine into a protein sequence is deleterious. The authors transformed *E. coli* with libraries of *M. tuberculosis* chorismate mutase and subjected the cells to ever-stringent selection pressure. Through multiple cycles of DE, they obtained a chorismite mutase variant that was 270-fold improved in k_{cat}/K_M.[51] An ingenious strategy was developed by Molina *et al*., where they used an orthogonal error-prone DNA replication system (OrthoRep) that would only amplify a defined gene. They used OrthoRep to target TrpB for mutagenesis while withholding indole from the media but adding indole derivatives. Through continuous evolution in the growth media, the TrpB variants evolved to efficiently produce non-canonical amino acids with great efficiency.[52,53]

Unfortunately, most tasks one wants to engineer proteins for cannot be tied to the survival of a microorganism. In that case, one needs to resort to screening for improved variants. Screening is a lot more tedious than selection, since one has to measure the fitness of each variant without knowing ahead of time if they are functional or not. When resorting to screening, throughput dictates the pace of the engineering, and it is an engineer's goal to maximize throughput as much as possible. Over the past few decades, many instruments and techniques have been developed to increase the throughput for protein engineering. The advent of using robotics in biochemistry such liquid handling robots have allowed researchers to accelerate the rate of assembling library plasmids and automate directed evolution.[54,55] Likely the most consequential methodology developed for increasing engineering throughput comes from the culmination of two technologies, cell surface display of proteins and fluorescence activated cell sorting (FACS). In cell surface display, the protein

of interest to be engineered is genetically fused to a cell-surface anchor protein, together with a signal

sequence to transport the protein to the surface of the cell. Proteins can be displayed on the surface of bacteria, insect, phage, mammalian, and yeast cells. Most often, protein engineers employ yeast surface display due to the simplicity of establishing the system and yeast, being eukaryotic cells, can perform various post-translational modifications that are essential for the production of certain proteins[56] During FACS, cells are placed into microfluidic lipid droplets containing only one cell. As the cells pass through lasers that excite fluorophores active on the surface of the cell, the emission of the fluorophores is measured and if the signal surpasses a threshold, the cells are either kept in media or discarded.[57] If the expression the protein of interest and its function can be coupled to the presence of fluorophores in the droplet, one can potentially screen 10^7 variants in a matter of hours.[58] While the throughput of cell surface display and FACS is immense, allowing for the screening of very large libraries, the advantages do not just end there. Usually, one of the fluorophores is used to measure presence of the protein on the surface of the cell and the other one is used to measure the function of the protein. It turns out that there is a correlation between the cell surface expression levels and a variants thermostability, effectively allowing an engineer to optimize two properties at once.[59,60]

So far, we have only considered the limitation of protein engineering in throughput and haven't put much thought into the limitations of library sizes. If we go back to the hypothesis by Maynard Jones that there are always sequences with similar function close-by in sequence space to a functional sequence and we think about what happens when we move further and further away from that functional variant.[38] Considering that the vast majority of mutations are deleterious, and that epistasis makes protein fitness landscaped hard to explore, means that most sequences screened would be non-functional. Anthony Keefe and Jack Szostak estimate that for a random protein sequence of length 80, around one in 10^{11} are functional.[61] Douglas Axe argues that a in a larger range of protein sequence lengths that ratio is around one in 10^{77}.[62] As libraries get bigger, the fraction of functional variants approaches these estimates of one in 10^{11} or one in 10^{77}. It might be advisable to not just look at random protein sequence libraries, but informed protein sequence libraries.

Machine Learning in Protein Engineering

In silico methods such as machine learning (ML) can vastly accelerate protein engineering campaigns. The general concept behind ML applied to protein engineering is to use the inferences made by an ML model to perform predictions on protein variants.[63,64] Sometimes called artificial intelligence, ML is a computer science discipline that focuses on using algorithms to learn from large corpuses of data. The learning algorithms can be very simple ones like multilinear regression or random forest classifiers,[65] to highly complex ones such as the transformer/graph neural network hybrid model behind Alphafold2.[66] Compared to other *in silico* methods, ML has a huge potential in generating informed libraries. Most non-ML methods make use of complex physical models that are computationally expensive, severely limiting the *in silico* throughput.[67,68] ML models however may take days, weeks or even months to train, but once trained, can generate variants on the scale of seconds (Figure 2.21) or even sub-seconds (Figure 2.12).

Informed library design can be very powerful even with simple models. In a series of papers, Wu *et al.*, Wittmann *et al.*, and Yang *et al*. demonstrated that an ensemble of simple machine learning models trained on a small set of GB1 sequences can generate libraries that are highly enriched in fit variants, through creative training set design. This was later experimentally verified by applying machine-learning assisted directed evolution to engineering TrpB. The concept of training ML models with low amounts of data is very important for protein engineering since labelled sequence datasets are rare,[69,70] and collecting sequence-function data is slow. This lack of data created a set of ML models for protein engineering called "low-N", for low numbers of datapoints.[71-76] One of the key concepts in "low-N" ML is the idea that more informative protein sequence embeddings are, the less data is required to train a powerful prediction model. A protein sequence embedding is the representation of the protein sequence in a way that a computer can understand it, often in the form of a vector or a matrix.[77] Arguably the simplest form of protein sequence embedding is one-hot encoding, where each position in a protein is represented by a 20-dimensional vector. Each dimension represents an amino acid, and the vector is 0 everywhere except at the dimension that represents the amino acid at a given residue position. While data availability is an issue to train "low-N" models, it is not an issue for learning highly informative protein sequence embeddings. Since the 2010s, next generation sequencing techniques were showering the literature with large, high quality, unlabeled sequence datasets and being summarized in databases such as UniRef,[78] GenBank,[79] and OAS.[80] The key to unlocking the potential behind this unlabeled sequence data, was to consider protein sequences like a language where amino acids act like words. At the time, a very large corpus of work had already been published on natural language processing (NLP), which also relied on large unlabeled datasets.[81-83] The application of NLP model architectures to the language of proteins created new model architectures collectively known as protein language models (PLM). A PLM often does not require labeled sequences but can infer properties from sequence data alone. Three important learning tasks have been applied to PLMs to achieve these inferences, masked language modelling,[84] autoregressive next-token prediction,[85] and diffusion (Figure 1.3).[86]

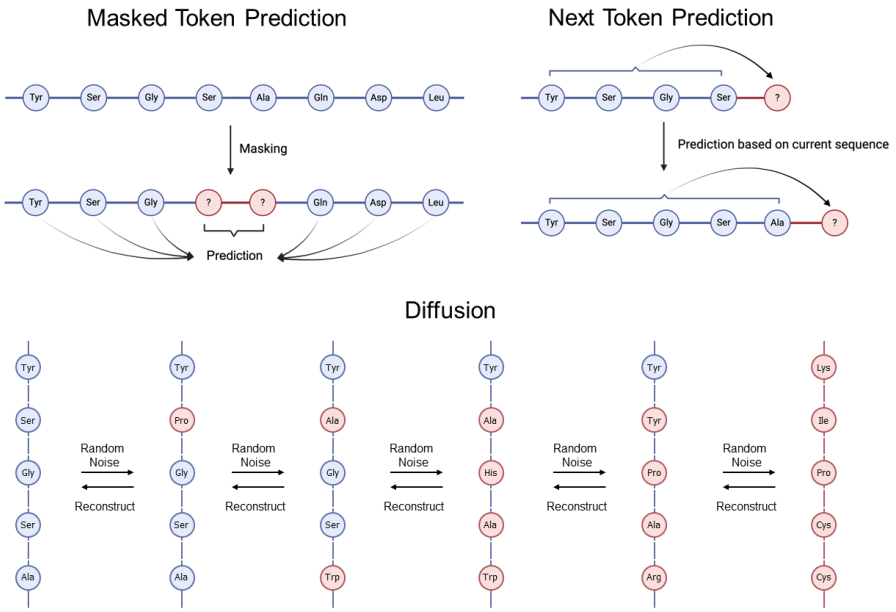


Figure 1.3: Examples of different learning tasks for protein language models. In masked token prediction, residues are chosen at random to be masked and the model needs to predict the masked residues given the context of other unmasked sequences. In next token prediction, the model needs to predict the nature of the next amino acid given the context of all the residues that come before it, iteratively reconstructing a sequence. In diffusion, random noise is added to a protein sequence in consecutive steps, with each step adding more noise. At each step that noise is added, the model is tasked with reconstructing the sequence of the previous step. The final model can generate functional sequences from random noise. Note: Diffusion is usually not performed in the way explained here, noise addition happens in a latent space where gaussian noise is added to a vector or matrix representing the protein.

Very powerful PLMs have been generated over the course of the past half-decade that alone can predict a wide range of protein properties. ESM2 for example is a masked language model trained with the transformer architecture.[87] Transformers were first introduced in the paper "Attention is all you need", where they presented a new model architecture that uses modules called "attention heads".[82] An "attention head" gathers global information on a sequence, i.e. it tries to answer the question how an amino acid at a certain position is defined by its context in the entire sequence. Models like ESM2 trained their transformer on 250 million sequence clusters from UniRef50,[78] using up to 15 billion parameters.[87] Even though the model has only been trained on masked token prediction of 1-dimensional sequences, the model is able to accurately predict the 3-dimensional structure of proteins as well as determine functional properties of the protein.[84,87-89] Another very powerful model relying on the transformer architecture is Alphafold2 (AF2), which is a structure prediction model.[66] Alphafold was born out of the CASP protein structure prediction competition and was the first ML model that surpassed the performance of physico-chemical models.[67] Where ESM2 is trained solely on protein sequences, AF2 is a lot more complicated. First of all, AF2 has as training objective the minimization of the predicted local-distance difference test (pLDDT) which is a measure of how accurate a structure prediction is. Second, AF2s

transformer does not work on single sequences but on multiple sequence alignments (MSA). The MSA transformer architecture is used in a variety of models where making predictions on the evolutionary context of a sequence is important.[90,91] Additionally, AF2 uses graph neural networks to infer 3-dimensional connections between residues. The importance of AF2 for biochemists is hard to overstate, spawning a variety of similar more specialized models for all sorts of structure prediction tasks.[92-94] AF2 even has an influence on experimental structure prediction as well, as some researchers have used AF2 structures to model crystallographic data onto, solving the "phase problem" for a large variety of situations.[95,96]

The true potential of large PLMs, however, lies in the specific models they can spawn that can help in scaling throughput in protein engineering. A powerful tool here is fine-tuning, which is the process where one takes a pre-trained model and re-trains it using a very specific dataset. Let's say we want to use a generative PLM such as RFDiffusion,[86] a diffusion model, or ProtGPT2,[85] a next-token prediction model (Figure 1.3). As originally trained, these models are generalists. Asking them to perform a specific task, such as creating a TIM-barrel in RFDiffusion or generating a nitrogenase-like sequence in ProtGPT2, would likely fail. Re-training the models using specific datasets like a dataset of TIM-barrel structures or of nitrogenase sequences nudges the models in a direction where they are capable of performing these tasks. In ProtGPT2, this fine tuning is part of the operating procedure of using it. In RFDiffusion, fine tuning was part of the training procedure where the developers used it to condition the model to certain situations. Such fine-tuned models can generate thousands of informed sequence libraries in a matter of hours, creating diverse sets of sequences for protein engineers to screen for.

The Three Studies of this Thesis

In this thesis, I will present three protein engineering studies that may be disjoint in their tasks that we are evolving for but have in common that they all attempt to maximize the throughput when trying to engineer for hard-to-measure tasks.

In Chapter 2, we tackle the problem of antibody humanization. Out of all the problems in this thesis, the actual property that we are engineering for can only be tested in what is best described as the ultimate ultra-low throughput. In order to check if humanization was successful an antibody candidate must pass all the scientific, bureaucratic, and financial hurdles to make it to phase 1 of clinical trials.[97-99] While many ML-based antibody humanization techniques claim that they work,[65,100] we demonstrated that the approach of most published methods is deeply flawed when the target is to bring an antibody to clinical trials. Through a combination of Markov chains, library generations and yeast display, we present a high-throughput technique to obtain as many antibodies as possible that conserve or improve properties of non-humanized antibodies, while decreasing their risk in potential future clinical trials.

In Chapter 3, we outline a project in collaboration with the Rees lab to engineer nitrogenase for the bioelectrocatalytic conversion of N_2 to ammonia. This project's significance stems from the fact that the process that produces most of the global ammonia, the Haber-Bosch process, is simultaneously one of the essential processes for modern agriculture and responsible for widespread environmental damage.[101] Our goal is to engineer nitrogenase using a bioelectrocatalytic system presented by Lee *et al*. using a multi-well system inspired by work from the Schwanenberg lab.[102,103] To further increase throughput, we are also developing a diazotrophic pre-selection method to allow us to test much larger libraries and selecting only for variants that produce a functional nitrogenase.

In Chapter 4, we use enzymes from a family of proteins of thermophilic organisms to engineer for a new-to-nature reaction to produce *cis*-trifluoromethylated cyclopropanes.[40] Compared to previous studies with chemical catalysts,[104-107] we were able to selectively synthesize a new conformation of these compounds which is not possible with published chemical catalysts. A key idea in this study is the usage of trapped carbene precursor in ethanol to run reactions in 96-well plates. We were able to evolve for this reaction and obtain a protein that catalyzes the reaction on the 1-mmol scale.

References

1. Alberts, B. Molecular Biology of the Cell. (Garland Science, Taylor and Francis Group, New York, 2015).
2. Cockcroft, S. Mammalian lipids: structure, synthesis and function. Essays Biochem 65, 813–845 (2021).
3. Philpott, C. C. Pumping iron. eLife 3, e03997 (2014).
4. Jeong, S. RNA in Biology and Therapeutics. Mol Cells 46, 1–2 (2023).
5. Mattick, J. S. et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol 24, 430–447 (2023).
6. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature 171, 737–738 (1953).
7. Rosalind Franklin: A Crucial Contribution | Learn Science at Scitable. http://www.nature.com/scitable/topicpage/rosalind-franklin-a-crucial-contribution-6538012.
8. Hershey, A. D. & Chase, M. INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE. Journal of General Physiology 36, 39–56 (1952).
9. Roll-Hansen, N. Commentary: Wilhelm Johannsen and the problem of heredity at the turn of the 19th century. Int J Epidemiol 43, 1007–1013 (2014).
10. Nirenberg, M. & Leder, P. RNA Codewords and Protein Synthesis. Science 145, 1399–1407 (1964).
11. Nirenberg, M. Historical review: Deciphering the genetic code – a personal account. Trends in Biochemical Sciences 29, 46–54 (2004).
12. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology 3, 318–356 (1961).
13. Kruger, K. et al. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. Cell 31, 147–157 (1982).
14. Castro, T. G., Melle-Franco, M., Sousa, C. E. A., Cavaco-Paulo, A. & Marcos, J. C. Non-Canonical Amino Acids as Building Blocks for Peptidomimetics: Structure, Function, and Applications. Biomolecules 13, 981 (2023).
15. Leisle, L., Valiyaveetil, F., Mehl, R. A. & Ahern, C. A. Incorporation of non-canonical amino acids. Adv Exp Med Biol 869, 119–151 (2015).
16. Jonckheere, A. I., Smeitink, J. A. M. & Rodenburg, R. J. T. Mitochondrial ATP synthase: architecture, function and pathology. J Inherit Metab Dis 35, 211–225 (2012).
17. Roberts, A. J., Kon, T., Knight, P. J., Sutoh, K. & Burgess, S. A. Functions and mechanics of dynein motor proteins. Nat Rev Mol Cell Biol 14, 713–726 (2013).
18. Berdis, A. J. Mechanisms of DNA Polymerases. Chem. Rev. 109, 2862–2879 (2009).
19. Schechter, A. N. Hemoglobin research and the origins of molecular medicine. Blood 112, 3927–3938 (2008).
20. Coulthurst, S. The Type VI secretion system: a versatile bacterial weapon. Microbiology (Reading) 165, 503–515 (2019).

21. Errors in DNA Replication | Learn Science at Scitable. http://www.nature.com/scitable/topicpage/dna-replication-and-causes-of-mutation-409.
22. Pfeifer, G. P. Mechanisms of UV-induced mutations and skin cancer. Genome Instab Dis 1, 99–113 (2020).
23. Shorey-Kendrick, L. E. et al. Nicotinic receptors in non-human primates: analysis of genetic and functional conservation with humans. Neuropharmacology 96, 263–273 (2015).
24. Meier, M., Janosik, M., Kery, V., Kraus, J. P. & Burkhard, P. Structure of human cystathionine β-synthase: a unique pyridoxal 5′-phosphate-dependent heme protein. EMBO J 20, 3910–3916 (2001).
25. Kim, D. et al. Evolutionary coupling analysis identifies the impact of disease-associated variants at less-conserved sites. Nucleic Acids Res 47, e94 (2019).
26. Hopf, T. A. et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol 35, 128–135 (2017).
27. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. Protein Science : A Publication of the Protein Society 25, 1204 (2016).
28. Poelwijk, F. J., Socolich, M. & Ranganathan, R. Learning the pattern of epistasis linking genotype and phenotype in a protein. Nat Commun 10, 4213 (2019).
29. McCandlish, D. M., Rajon, E., Shah, P., Ding, Y. & Plotkin, J. B. The role of epistasis in protein evolution. Nature 497, E1–E2 (2013).
30. Otwinowski, J. Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function. Molecular Biology and Evolution 35, 2345–2354 (2018).
31. Meger, A. T. et al. Rugged fitness landscapes minimize promiscuity in the evolution of transcriptional repressors. Cell Systems 15, 374-387.e6 (2024).
32. Jayaraman, V., Toledo-Patiño, S., Noda-García, L. & Laurino, P. Mechanisms of protein evolution. Protein Sci 31, e4362 (2022).
33. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. Proceedings of the National Academy of Sciences 103, 5869–5874 (2006).
34. Sorokina, I., Mushegian, A. R. & Koonin, E. V. Is Protein Folding a Thermodynamically Unfavorable, Active, Energy-Dependent Process? Int J Mol Sci 23, 521 (2022).
35. Louros, N., Schymkowitz, J. & Rousseau, F. Mechanisms and pathology of protein misfolding and aggregation. Nat Rev Mol Cell Biol 24, 912–933 (2023).
36. Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. Angewandte Chemie International Edition 57, 4143–4148 (2018).
37. Maynard Smith, J. Natural Selection and the Concept of a Protein Space. Nature 225, 563–564 (1970).
38. Smith, J. M. Evolutionary game theory. Physica D: Nonlinear Phenomena 22, 43–49 (1986).
39. Knight, A. M. et al. Diverse Engineered Heme Proteins Enable Stereodivergent Cyclopropanation of Unactivated Alkenes. ACS Cent. Sci. 4, 372–377 (2018).

40. Schaus, L. et al. Protoglobin-Catalyzed Formation of cis-Trifluoromethyl-Substituted Cyclopropanes by Carbene Transfer. Angewandte Chemie International Edition 62, e202208936 (2023).
41. Basler, S. et al. Efficient Lewis acid catalysis of an abiological reaction in a de novo protein scaffold. Nat Chem 13, 231–235 (2021).
42. Blomberg, R. et al. Precision is essential for efficient catalysis in an evolved Kemp eliminase. Nature 503, 418–421 (2013).
43. Gao, S. et al. Enzymatic Nitrogen Incorporation Using Hydroxylamine. J. Am. Chem. Soc. 145, 20196–20201 (2023).
44. Athavale, S. V. et al. Enzymatic Nitrogen Insertion into Unactivated C–H Bonds. J. Am. Chem. Soc. 144, 19097–19105 (2022).
45. Miller, D. C., Athavale, S. V. & Arnold, F. H. Combining chemistry and protein engineering for new-to-nature biocatalysis. Nat Synth 1, 18–23 (2022).
46. Studer, S. et al. Evolution of a highly active and enantiospecific metalloenzyme from short peptides. Science 362, 1285–1288 (2018).
47. Schmidt-Dannert, C. & Arnold, F. H. Directed evolution of industrial enzymes. Trends in Biotechnology 17, 135–136 (1999).
48. Pritchard, L., Corne, D., Kell, D., Rowland, J. & Winson, M. A general model of error-prone PCR. Journal of Theoretical Biology 234, 497–509 (2005).
49. Mullis, K. B. The Polymerase Chain Reaction (Nobel Lecture). Angewandte Chemie International Edition in English 33, 1209–1213 (1994).
50. Kille, S. et al. Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. ACS Synth Biol 2, 83–92 (2013).
51. Fahrig-Kamarauskait÷, J. et al. Evolving the naturally compromised chorismate mutase from Mycobacterium tuberculosis to top performance. J Biol Chem 295, 17514–17534 (2020).
52. Rix, G. et al. Scalable continuous evolution for the generation of diverse enzyme variants encompassing promiscuous activities. Nat Commun 11, 5644 (2020).
53. Molina, R. S. et al. In vivo hypermutation and continuous evolution. Nat Rev Methods Primers 2, 37 (2022).
54. Bryant, J. A., Jr., Kellinger, M., Longmire, C., Miller, R. & Wright, R. C. AssemblyTron: flexible automation of DNA assembly with Opentrons OT-2 lab robots. Synthetic Biology 8, ysac032 (2023).
55. Handal-Marquez, P., Koch, M., Kestemont, D., Arangundy-Franklin, S. & Pinheiro, V. B. Antha-Guided Automation of Darwin Assembly for the Construction of Bespoke Gene Libraries. in Directed Evolution: Methods and Protocols (eds. Currin, A. & Swainston, N.) 43–66 (Springer US, New York, NY, 2022). doi:10.1007/978-1-0716-2152-3_4.
56. Cherf, G. M. & Cochran, J. R. Applications of yeast surface display for protein engineering. Methods Mol Biol 1319, 155–175 (2015).
57. Fei, C., Nie, L., Zhang, J. & Chen, J. Potential Applications of Fluorescence-Activated Cell Sorting (FACS) and Droplet-Based Microfluidics in Promoting the

Discovery of Specific Antibodies for Characterizations of Fish Immune Cells. Front. Immunol. 12, (2021).
58. Van Deventer, J. A., Kelly, R. L., Rajan, S., Wittrup, K. D. & Sidhu, S. S. A switchable yeast display/secretion system. Protein Eng Des Sel 28, 317–325 (2015).
59. Kowalski, J. M., Parekh, R. N. & Wittrup, K. D. Secretion efficiency in Saccharomyces cerevisiae of bovine pancreatic trypsin inhibitor mutants lacking disulfide bonds is correlated with thermodynamic stability. Biochemistry 37, 1264–1273 (1998).
60. Shusta, E. V., Kieke, M. C., Parke, E., Kranz, D. M. & Wittrup, K. D. Yeast polypeptide fusion surface display levels predict thermal stability and soluble secretion efficiency. J Mol Biol 292, 949–956 (1999).
61. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. Nature 410, 715–718 (2001).
62. Axe, D. D. Estimating the prevalence of protein sequences adopting functional enzyme folds. J Mol Biol 341, 1295–1315 (2004).
63. Johnston, K. E. et al. Machine Learning for Protein Engineering. ArXiv arXiv:2305.16634v1 (2023).
64. Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. Advances in machine learning for directed evolution. Curr Opin Struct Biol 69, 11–18 (2021).
65. Marks, C., Hummer, A. M., Chin, M. & Deane, C. M. Humanization of antibodies using a machine learning approach on large-scale repertoire data. Bioinformatics 37, 4041–4047 (2021).
66. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
67. Marcu, Ș.-B., Tăbîrcă, S. & Tangney, M. An Overview of Alphafold's Breakthrough. Front. Artif. Intell. 5, (2022).
68. Leman, J. K. et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat Methods 17, 665–680 (2020).
69. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. eLife 5, e16965 (2016).
70. Bryant, D. H. et al. Deep diversification of an AAV capsid protein by machine learning. Nat Biotechnol 39, 691–696 (2021).
71. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. Proceedings of the National Academy of Sciences 116, 8852–8858 (2019).
72. Wittmann, B. J., Yue, Y. & Arnold, F. H. Informed training set design enables efficient machine learning-assisted directed protein evolution. cels 12, 1026-1045.e7 (2021).
73. Yang, J. et al. DeCOIL: Optimization of Degenerate Codon Libraries for Machine Learning-Assisted Protein Engineering. ACS Synth. Biol. 12, 2444–2454 (2023).
74. Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. Nat Biotechnol (2022) doi:10.1038/s41587-021-01146-5.

75. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. Nat Methods 18, 389–396 (2021).
76. Chu, H. Y. et al. Accurate top protein variant discovery via low-N pick-and-validate machine learning. Cell Systems 15, 193-203.e6 (2024).
77. Yang, K. K., Wu, Z., Bedbrook, C. N. & Arnold, F. H. Learned protein embeddings for machine learning. Bioinformatics 34, 2642–2648 (2018).
78. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23, 1282–1288 (2007).
79. Benson, D. A. et al. GenBank. Nucleic Acids Res 41, D36-42 (2013).
80. Olsen, T. H., Boyles, F. & Deane, C. M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci 31, 141–146 (2022).
81. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. Neural Comput. 9, 1735–1780 (1997).
82. Vaswani, A. et al. Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).
83. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at https://doi.org/10.48550/arXiv.1810.04805 (2019).
84. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 118, e2016239118 (2021).
85. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun 13, 4348 (2022).
86. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100 (2023).
87. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023).
88. Hie, B. et al. A high-level programming language for generative protein design. 2022.12.21.521526 Preprint at https://doi.org/10.1101/2022.12.21.521526 (2022).
89. Verkuil, R. et al. Language models generalize beyond natural proteins. 2022.12.21.521521 Preprint at https://doi.org/10.1101/2022.12.21.521521 (2022).
90. Rao, R. M. et al. MSA Transformer. in Proceedings of the 38th International Conference on Machine Learning 8844–8856 (PMLR, 2021).
91. Lupo, U., Sgarbossa, D. & Bitbol, A.-F. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. Nat Commun 13, 6298 (2022).
92. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871–876 (2021).
93. Mirdita, M. et al. ColabFold: making protein folding accessible to all. Nat Methods 19, 679–682 (2022).

94. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun 13, 1265 (2022).
95. Barbarin-Bocahu, I. & Graille, M. The X-ray crystallography phase problem solved thanks to AlphaFold and RoseTTAFold models: a case-study report. Acta Crystallogr D Struct Biol 78, 517–531 (2022).
96. Stevens, T. A. et al. A nanobody-based strategy for rapid and scalable purification of native human protein complexes. 2023.03.09.531980 Preprint at https://doi.org/10.1101/2023.03.09.531980 (2023).
97. Lu, R.-M. et al. Development of therapeutic antibodies for the treatment of diseases. Journal of Biomedical Science 27, 1 (2020).
98. Kernstock, R., Sperinde, G., Finco, D., Davis, R. & Montgomery, D. Clinical Immunogenicity Risk Assessment Strategy for a Low Risk Monoclonal Antibody. AAPS J 22, 60 (2020).
99. Aubrey, N. & Billiald, P. Antibody Fragments Humanization: Beginning with the End in Mind. in Human Monoclonal Antibodies: Methods and Protocols (ed. Steinitz, M.) 231–252 (Springer, New York, NY, 2019). doi:10.1007/978-1-4939-8958-4_10.
100. Prihoda, D. et al. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. MAbs 14, 2020203.
101. Ritter, S. The Haber-Bosch Reaction: An Early Chemical Impact On Sustainability. Chemical & Engineering News https://cen.acs.org/articles/86/i33/Haber-Bosch-Reaction-Early-Chemical.html.
102. Lee, Y. S., Yuan, M., Cai, R., Lim, K. & Minteer, S. D. Nitrogenase Bioelectrocatalysis: ATP-Independent Ammonia Production Using a Redox Polymer/MoFe Protein System. ACS Catal. 10, 6854–6861 (2020).
103. Chen, H., Dong, F. & Minteer, S. D. The progress and outlook of bioelectrocatalysis for the production of chemicals, fuels and materials. Nat Catal 3, 225–244 (2020).
104. Morandi, B. & Carreira, E. M. Iron-catalyzed cyclopropanation with trifluoroethylamine hydrochloride and olefins in aqueous media: in situ generation of trifluoromethyl diazomethane. Angew Chem Int Ed Engl 49, 938–941 (2010).
105. Morandi, B., Mariampillai, B. & Carreira, E. M. Enantioselective cobalt-catalyzed preparation of trifluoromethyl-substituted cyclopropanes. Angew Chem Int Ed Engl 50, 1101–1104 (2011).
106. Artamonov, O. S., Mykhailiuk, P. K., Voievoda, N. M., Volochnyuk, D. M. & Komarov, I. V. Simple and Efficient Procedure for a Multigram Synthesis of Both trans- and cis-1-Amino-2-(trifluoromethyl)cyclopropane-1-carboxylic Acid. Synthesis 2010, 443–446 (2010).
107. Hock, K. J. et al. Corey-Chaykovsky Reactions of Nitro Styrenes Enable cis-Configured Trifluoromethyl Cyclopropanes. J Org Chem 82, 8220–8227 (2017).

Chapter 2

MACHINE LEARNING BASED ANTIBODY HUMANIZATION AT SCALE USING MOUSIFY

Abstract

In the past half-decade more monoclonal antibodies (mAb) have been approved for therapeutic use than in the 25 years before that. When mAbs are raised in non-human organisms and injected into humans, the patient will much more likely produce anti-drug antibodies than if the antibody is humanized. Since the late 1980's antibody humanization has been developed to reduce immunogenic risk of mAb therapeutics. Achieving a reduction of patient anti-drug antibody responses in clinical trials is the goal of antibody humanization. To measure this however, one needs to pass significant scientific, bureaucratic, and financial hurdles, which is very rarely done and especially never at scale. Most existing ML-based antibody humanization techniques claim that they work without providing any experimental evidence. We developed Mousify as an *in silico* antibody humanization platform to place existing models into one framework for wet-laboratory validation. We demonstrate that even the best models have a fundamental flaw in that they only generate a single antibody. We use Mousify and Markov chains to show that using ML-based antibody humanization models for library generation is not only feasible but produces both stable and functional variants. Learning the lessons from our wet-laboratory experiments, we then developed a variational autoencoder model with properties that hopefully improve the outcomes of antibody humanization experiments.

Contributions

The project was designed and conceived by Lucas Schaus and Prof. Steve Mayo. Wet lab experiments were performed by Lucas Schaus and Adrian Bahn. Lucas Schaus performed all the experiments outlined in “Benchmarks for Antibody Humanization”. Adrian Bahn performed the experiments outlined in “Generation of Humanized Antibody Libraries using Mousify”, with the help of Lucas Schaus. The code base was principally written by Lucas Schaus, with the help of Arielle Tycko and Jessie Gan. Arielle Tycko wrote parts of the random forest model described in “Mousify Antibody Humanization Software”. Jessie Gan wrote the code described in “Fast Structure Calculations for Mousify Libraries”, with help and guidance by Lucas Schaus. Writing was done by Lucas Schaus, with contributions from Jessie Gan in “Fast Structure Calculations for Mousify Libraries”. The project was performed in the labs of Prof. Steve Mayo and Prof. Pamela Bjorkman.

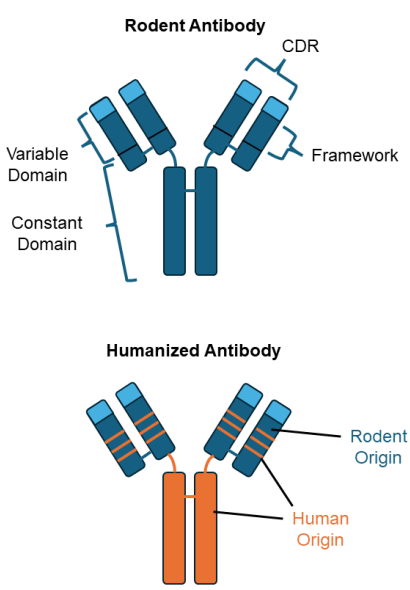
Monoclonal Antibodies and Humanization

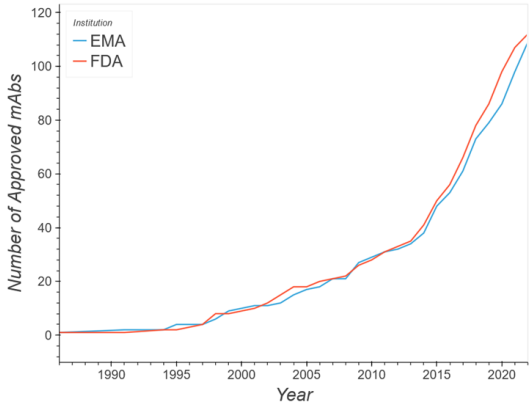
Antibodies are proteins that are produced by certain organisms as part of their adaptive immune system. When an organism with an adaptive immune system is exposed to an antigen, antibodies are responsible for identifying and neutralizing it. The immune system achieves this by producing a large corpus of highly diverse antibody sequences via B-cells. This diversity is achieved by the random recombination of 3 genes called V, D, and J,[1] followed by a process called somatic hypermutation. The diversified region of the antibody, made up of the complimentary determining region (CDR) and the framework (FWR), is called the variable domain.[2] Through this process it is estimated that a human can produce between 10^{16}-10^{18} different antibody variants.[3] When B-cell receptor antibody binds an antigen, the cell activates a mechanism by which it clones itself, producing many daughter cells and proliferating an antibody that can target the antigen that is to be neutralized. Shortly after the first monoclonal antibody therapeutic (mAb) was expressed in 1975 by Köhler and Milstein, researchers realized that one could employ the highly specific binding capacity of mAbs to neutralize targets in the human body. Monoclonal means that the antibody comes from a single B-cell clone. In 1986, the FDA approved the first mAb called muronomab-CD3 for the use in preventing kidney transplant rejection.[4] This drug worked by blocking the effects of the T-cell receptor CD3, a target that the body would not naturally build antibodies against, since the body has a mechanism to reject antibodies that bind to self-antigens. However, human CD3 is not a self-antigen in a mouse, therefore a mouse injected with human CD3 would produce antibodies against the desired target. However, the use of muronomab-CD3 was very limited due to the high number of patients developing antibodies against the drug (anti-drug antibodies, or ADA).[5]

While a successful milestone, muronomab-CD3s induction of ADA in patients made it clear that scientist needed to overcome some immunogenic and efficacy hurdles to make mAbs truly useful. Researchers started to take antibodies from rodents and attempted to make them look more like antibodies a human would make without affecting the positive properties that drew them to the rodent antibody in the first place, a.k.a. antibody humanization. The first successful attempt was abciximab, an anti-GPIIb/IIIa mAb whose variable domain was attached to the constant domain of a human antibody, producing the first chimeric antibody (Figure 2.1). A big leap in antibody humanization was made with the first demonstration of CDR-grafting.[6] The CDR is the region of an antibody and is the region of the antibody that is responsible for binding the antigen. In CDR grafting, non-human CDR sequences are grafted onto the framework of a human antibody. To achieve this one usually identifies the V-gene that the non-human antibody originated from, and one grafts it to the evolutionarily closest human version of that V-gene. CDR-grafting is still a very popular antibody humanization technique which still sees development today by combining it with state-of-the-art *in silico* techniques.[7] All these wet-laboratory humanization techniques have continuously been evolving by combining CDR-grafting or chimerization with back-mutations of "human" residues to "non-human",

conservation of Vernier zone residues and *in silico* stabilization of frameworks or CDRs.[7,8] Often times, these methods are nothing more than trial-and-error based and cannot be systematically evaluated due to the difficulty of bringing a mAb candidate to clinical trials.[9]

Since then, mAbs have been developed to treat a wide range of diseases and conditions. Notable examples include Xolair (Omalizumab), an anti-human IgE mAb that is used to treat severe asthma in adults and children, chronic urticaria, as well as severe allergies from aerosols and food allergies, or Keytruda (Pembrolizumab) a highly successful anti-human PD-1 mAb used in cancer therapy. In addition, the approval of monoclonal antibody therapeutics has exploded in the past decade with nearly two-thirds of all mAbs being approved in that time period (Figure 2.1).[10] mAbs are also highly valuable, for example the drug Cimizia (Certolizumab Pegol) is the only mAb in the portfolio of the Belgian company UCB, Cimizia alone makes up 40% of the companies yearly revenue (~EUR 2B) in 2022.[11] In total the mAb market had a valuation of USD 115.2B in 2018.[12]





Approved or Reviewed mAbs	
Human	35%
Humanized	45%
Chimeric†	10%
Other	13%

Figure 2.1: Overview of monoclonal antibodies and humanization. (**Left**) Illustration of an antibody showing the location of the CDR, framework, variable domain, and constant domain. At the bottom we are showing an example of a humanized antibody where certain segments of the mAb have been replaced with sequences of human origin. (**Right**) Number of approved antibody therapeutics by year and their percentage of origin. †Since chimeric antibodies can count both as human or non-human, the sum of all percentages exceeds 100%.

With modern biochemistry and synthetic biology techniques, many different antibody humanization techniques have been developed. For the purposes of this study, we will only briefly mention techniques that originate from non-hybridoma

techniques as they are not as often employed nor are they applicable to the main topic of this chapter which is machine learning based antibody humanization. Phages expressing human V, D, and J genes can be used together with phage display to screen for antibody libraries that originate from human antibody germlines. The main problem with this technique is that polyreactivity, the property of non-specifically binding antigens, is common. In addition, often times the pI and hydrophobicity of the resulting antibodies is unsuitable for usage in organisms, often times requiring further downstream engineering of the mAb.[12,13] Another commonly used technique is to express human germline genes in transgenic mice and raise antibodies in them. A common problem here is that the antibodies are not counter-selected for human antigens, but for mice antigens resulting in not-fully humanized antibodies.[13]

Most commonly, the mouse hybridoma technique is used to raise antibodies against an antigen. After immunizing a mouse, the B-cells of the mouse are harvested and screened for binding against the antigen.[12] Once a sufficiently good binder has been identified, the B-cell is sequenced, its antibody expressed and further characterized for desired properties. At this point the candidate mAbs are evaluated for their developability. Developability is a collection of properties of a drug that have been determined to be essential to meet in order to increase chances of making it to clinical trials with a candidate. Some of these key properties include the solubility, binding affinity, thermostability, polyreactivity, and propensity to aggregate, which are properties most academic biologists are interested in as well.[14] However, there are also industry specific properties such as the expression level of the antibody and its purity. If the antibody is deemed sufficiently developable, the aforementioned humanization techniques are used to engineer the antibody for the purposes of de-risking the drug. Immunogenicity risk assessment is then performed according to the guidelines provided by the FDA or the EMA, which includes evaluating the risk of immunogenic incidence such as the humanization of the antibody as well as indirect indicators of risk such as the pharmacokinetic properties of the drug.[15,16] Only after passing through the long process all of the scientific, bureaucratic, and financial hurdles associated with experiments and drug candidate assessments can a drug be considered for clinical trial and the chosen humanization technique evaluated.

Antibody Humanization and Machine Learning

Since the mid-2010s, along with the publication of many other machine learning models applied to protein engineering, researchers have looked at ML to create more reliable and systematized antibody humanization techniques. While the wet-laboratory techniques described in the previous section have been validated multiple times by the FDA and EMA approved drugs they contributed to producing, they are still very risky, with a company not knowing ahead of time if a given drug could cause a large fraction of patients to develop ADA.[9] With the greater availability of both labeled and unlabeled antibody sequence data in databases such as SAbDab,[17] OAS,[18] and Thera-SAbDab,[19] researchers were trying to replicate the success of machine learning models applied to protein engineering on antibody humanization. The main objective of ML-based antibody humanization is to train a model that is able to predict the "humanness" of antibody sequences and for that score to be correlated to %ADA data from Thera-SAbDab. Pre-ML such humanness scores existed in the form of pairwise sequence identity scores of a sequence to the closest human germline, however these scores poorly correlated with %ADA.[20]

At the core of the training is the large antibody sequence database called the observed antibody space (OAS). This database contains billions of antibody sequences that are gathered from next generation sequencing of human and non-human B-cells. OAS has since its inception supported many antibody-specific ML models, beyond just antibody humanization.[20–27] To our knowledge, the first antibody humanization model that made use of OAS to train a model was Hu-mAb which is a random forest classifier model. As the name suggests, the classifier takes the antibody sequence as an input and tries to predict the species of the antibody. In this case, Hu-mAb trains for binary classification of human vs non-human. Hu-mAb is actually a collection of V-gene specific classifier models, each model distinguishing between human and non-human sequences of every V-gene type. The classifier performed very well in the classification task with an AUROC score of 0.977 and a classification accuracy of ≥ 0.999 in the test set. In order to obtain a humanness score out of the random forest classifier model, they inquire the model for the probability it gives to the human vs the non-human class. Meaning the humanness score reflects the probability of a sequence that the model assigns to the human class. The reported Pearson's correlation coefficient of the Hu-mAb score to %ADA is -0.58. Marks *et al*. then used this score to perform *in silico* antibody humanization by plugging-in all possible single mutants from the sequence to be humanized into the model and ranking them by their increase in humanness. The top variant is then chosen for the next round and repeated until a score threshold is achieved.[20]

BioPhi is a model developed for the humanization of mAbs that makes use of two separate models, one ML-based and the other one not. OASis is the model of BioPhi that assigns a humanness score to a sequence by counting the occurrence of all 9-mer peptides in the sequence and referencing them against a database of all 9-mer

peptides that occur in human sequences found in OAS. As opposed to the humanness score of Hu-mAb, the OASis score can be tuned since the score is calculated as the fraction of 9-mer peptides that can be found in more than x% of humans. Where x is the tunable parameter. They find that the OASis score can be used to make a fantastic classifier with an AUROC score of 0.966. Additionally, the OASis score correlates with %ADA with a Pearson's correlation coefficient $\rho = -0.53$. [21] In order to humanize antibody sequences, they developed a RoBERTa-based transformer model that was trained on only human antibody sequences, called Sapiens[28]. The logic behind training the model only on human sequences is that they want to use the softmax layer at the end of the Sapiens model to obtain a probability distribution of each residue if the sequence was human. Reasoning that the model would identify the mutations that would maximize the human content. We suspect that this reasoning is flawed since now non-human sequences might be too far outside of the training set to make generalizable predictions on the mutations that would humanize a sequence. In order to humanize the sequence, BioPhi picks all the highest probability amino acids at each position according to the output from Sapiens, optimizing the entire sequence at once. This is performed up to five times to humanize the antibody sequence.[21] Two flaws become immediately apparent in this methodology. First, during training a transformer model on ever predicts one masked token at a time (Figure 1.3, See Chapter 1) and therefore the method of predicting all positions at once is illogical. The transformer gives a probability of an amino acid at a position, conditioned on the fact that all other amino acids remain the same. Second, the usage of the OASis score that correlates well with ADA is completely circumvented in this method and has no effect on humanization at all. They show however, that the mutations predicted by Sapiens increase the OASis score of a sequence, but there is no attempt to maximize the OASis score.

To validate the efficacy of both models the authors of Hu-mAb and BioPhi collected the data of 25 clinically approved mAbs for which the original murine sequence was known and re-humanize the sequences. They then evaluate the overlaps of humanizing mutations and concluded that their model works well since there is some mutational overlap with therapeutic sequences. Both Hu-mAb and BioPhi achieve humanization in fewer mutations from the original sequence. Neither group of authors provided any wet-laboratory experimental evidence to support their claims.[20,21]

Overview of the Mousify Software Package

The Mousify software package is made up of three independent parts that can be found on the authors GitHub repository (lschaus0408):

- Mousify: Contains the antibody humanization software of Mousify as well as all supporting ".py" files.
- OAS C/S: Contains the client/server module responsible for downloading and packaging antibody sequence datasets.
- Mousify Analysis: Contains python files (.py) as well as ipython notebook files (.ipynb) used to analyze data from the Mousify sub-package, as well as generate figures.

The software package was entirely written in python 3.10 and 3.11 (for notebooks) on a Windows 10 machine running the Windows Subsystem for Linux (WSL) Ubuntu 18.04. The WSL was only used for development of this software package and nothing else, to ensure that its environment is reproducible. The python environment was managed via Anaconda. It is worth noting that this software package cannot be used on Windows, only on Linux or MacOS, due to Mousify and OAS API dependence on HMMER,[29] which only works on the latter operating systems.

Antibody Dataset Management Using OAS C/S

Arguably the most useful sets of databases for antibody data are the resources provided by the Oxford Protein Informatics Group (OPIG).[30] Amongst many other tools, the ones that are of particular interest to machine learning (ML) tasks are the Structural Antibody Database (SAbDab)[17] and the Observed Antibody Space (OAS),[18] as demonstrated by the plethora of antibody-specific ML models that make use of them.[20–27]

	sequence	locus	stop_codon	vj_in_frame	v_frameshift	productive	rev_comp	complete_		v_call	d_call	j_call
0	CCTGTGCCATCTCCGGGGGACAGTGTCTCTGGCACCAATGTTACATG...	H	F	T	F	T	F		0	IGHV6-1*01	IGHD3-9*01	IGHJ4*02
1	CCTGTGCCATCTCCGGGGGACAGTGTCTCTAACAACCATGTTGCTTG...	H	F	T	F	T	T		1	IGHV6-1*01	IGHD2-15*01	IGHJ3*02
2	CCTCAGTGAAGGTCTCCTGCAAGGCTTCTGGATACACCTTCACCGG...	H	F	T	F	T	T		2	IGHV1-2*02	IGHD4-17*01	IGHJ4*02
3	CCTCAGTGAAGGTCTCCTGCAAGGCTTCTGGATACACCTTCACTAC...	H	F	T	F	T	F		3	IGHV1-3*01	IGHD2-2*01	IGHJ4*02
4	CCTCAGTGAAGGTCTCCTGCAAGGCTTCTGGATTCAACGTCAGCAA...	H	F	T	F	T	F		4	IGHV1-3*01	IGHD6-19*01	IGHJ3*01

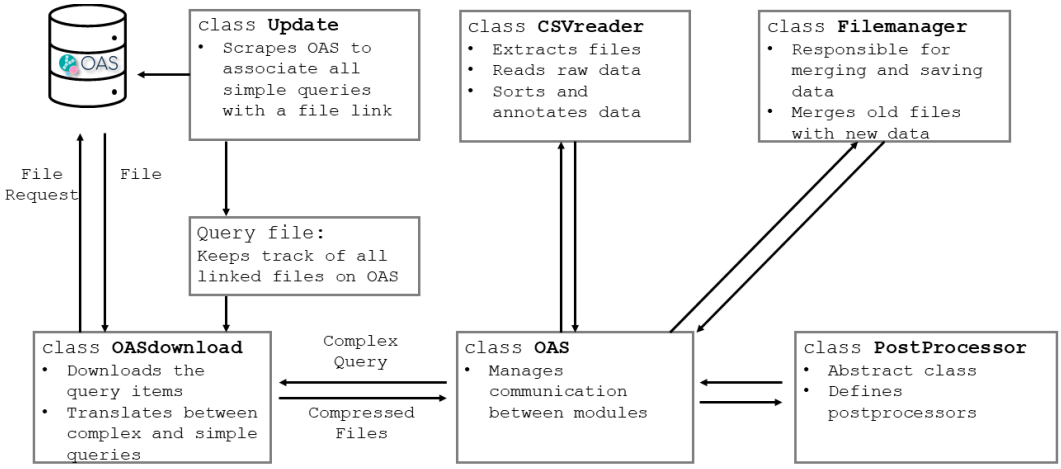
	fwr1	fwr1_aa	cdr1	cdr1_aa	v_cigar	d_cigar	j_cigar	v_support
0	CCTGTGCCATCTCC	CAIS	GGGGACAGTGTCTCTGGCACCAATGTTACA	GDSVSGTNVT	61N244M180S	246S11N6M172S14N	256S4N44M124S	4.519000e-78
1	CCTGTGCCATCTCC	CAIS	GGGGACAGTGTCTCTAACAACCATGTTGCT	GDSVSNNHVA	61N109M9D123M186S3N	238S24N7M173S	246S5N45M127S	3.149000e-68
2	CCTCAGTGAAGGTCTCCTGCAAGGCTTCT	SVKVSCKAS	GGATACACCTTCACCGGCTTCTAT	GYTFTGFY	46N247M198S3N	252S3N13M180S	280S10N38M127S	2.104000e-95
3	CCTCAGTGAAGGTCTCCTGCAAGGCTTCT	SVKVSCKAS	GGATACACCTTCACTACGTATGCC	GYTFTTYA	46N250M206S	254S5N19M183S7N	287S2N46M123S	5.514000e-94
4	CCTCAGTGAAGGTCTCCTGCAAGGCTTCT	SVKVSCKAS	GGATTCAACGTCAGCAATTACGCT	GFNVSNYA	46N246M205S4N	249S11N10M192S	278S4N46M127S	1.338000e-69

Figure 2.2: Example of data columns visualized in pandas. Full antibody sequence data can only be found in the form of the DNA sequence which contains often contains more sequence data than just the desired VH/VL sequence. The amino acid sequence is separated into framework (fwr) and complimentary determining region (CDR), and never represented as the full VH/VL sequence. Much of the useful annotation data for ML is in the JSON formatted header (not shown), e.g.: Organism of origin, exposed antigen, vaccination(s), B-cell type. Each row represents a sequence. Showing 19 columns out of 95 columns. Data downloaded from OAS, from the study of Galson *et al*. (2015).

While vast, with over 2 billion sequences, and thorough, with important sequence annotations, in the data represented, OAS is not tailored towards being an ML specific database. To start, not all of the 2 billion sequences are equal in quality. Some sequences are incomplete, represent non-viable antibodies, or are the result of misreads in the original next-generation sequencing data.[31] Depending on the learning task, there is also lot of unnecessary data (Figure 2.2). For example, the full sequence of the variable chain is only represented as its DNA sequence, however, this also contains more than just the variable chain. To obtain the amino acid sequence, one has to concatenate the sequences in the columns “fwr1_aa”, “cdr1_aa”, “fwr2_aa”, “cdr2_aa” etc. Then there is data that is specific to the alignment against the putative germline, such as the Compact Idiosyncratic Gapped Alignment Report (CIGAR) string and the support score for the chosen germline alignment. While useful, the use-cases for ML might be quite niche. In addition, the more useful data, e.g. species of origin, exposed antigens, B-cell type, is stored in a JavaScript Object Notation (JSON) formatted header, which needs to be parsed. All of this causes unnecessarily bloated file sizes that need to be processed in order to be useful for machine learning. Lastly, more complex antibody sequence queries can be tedious to make on the web client of OAS. For example, if one is interested in all IgG VH antibody sequences that originate from B-Cells that have undergone self-antigen counter-selection that come from all available species, except camels, one has to make 99 different requests, by clicking through drop-down tables in OAS. With the

tool described in this section, OAS C/S, such a request can be made with just a few lines of a query file or in a command line (Figure SC.1, 2).

The Observed Antibody Space Client/Server (OAS C/S) tool was developed because of the desire to explore the effects of dataset qualities on the performance of antibody-specific ML models. Being able to think of a hypothesis on how an antibody sequence dataset could perform, and rapidly gaining access to filtered, high-quality sequences, annotated with only the data necessary for the training task, can give ML scientists better flexibility and higher quality models than relying on OAS alone. One alternative is downloading and storing the entire database which would incur a download of approximately 1TB distributed across more than 15,000 compressed data files. These files need to be extracted, sorted, and processed each time one wants to form a new dataset. The other alternative is to download each file in a complex query from OAS separately, followed by extraction, sorting and processing. Our goal was to make OAS C/S the better than both alternatives, simplifying complex queries, resulting in download, and processing times not longer than needed, as well as helping uses extract only the high-quality sequences of the database.

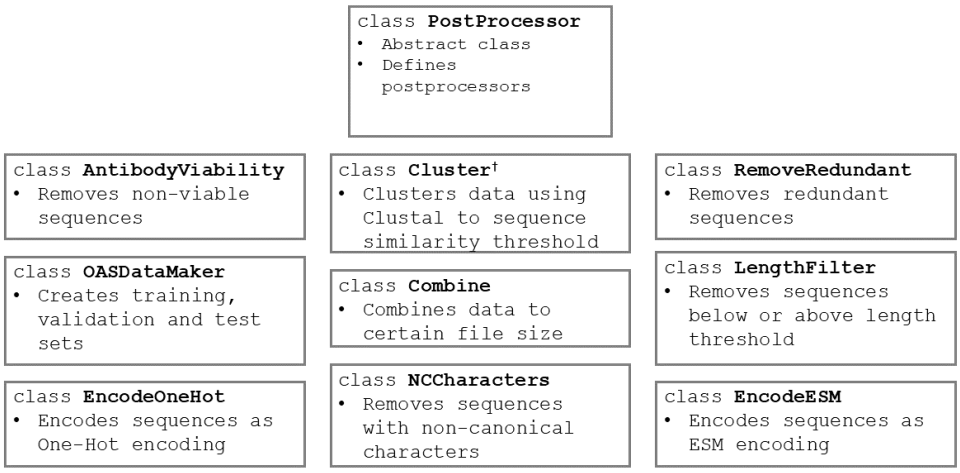


Scheme 2.1: Workflow and modules of OAS CS.

In Scheme 2.1, we outlined the workflow, and the modules of OAS C/S. Users directly interact with the "OAS" module, either through making queries in the command line interface (CLI) or by providing the program with a query file in the CLI (Figure SC.1, 2). Complex queries are translated to simpler queries by the module "OASdownload", which uses the query file to know which simple query is associated with which files on the OAS database. "OASdownload" sends the request to the OAS database and receives a file in return. The query file can be updated to reflect the actual database using the "Update" module if the user wishes to do so. Once all compressed files have been downloaded and placed into a temporary folder, the "OAS" module will ask the "CSVreader" module to extract the files and process

them. In the user's query request, it is required to provide what data one wants to keep from the downloaded files (e.g. amino acid sequence, antibody subtype, B-cell type etc.), the "CSVreader" extracts that data and only keeps the data requested. This module also filters out any sequences that are missing any framework or CDR. After receiving the temporary folder with the desired antibody data, the "OAS" module asks the "Filemanager" module to rename the files according to the desired systematic file naming scheme provided by the user. The "Filemanager" can also translate files to different formats such as CSV or JSON, as well as merge data from different OAS C/S requests.

To ensure that the data used for ML applications is of the highest quality, we developed a suite of post-processing tools. All post-processing modules inherit from the abstract class "PostProcessor", which can help in the open-source development of new post-processing modules. As of writing this, we have implemented eight post-processing modules and are planning on implementing a ninth one (Scheme 2.2).



Scheme 2.2: Post-processing modules of OAS C/S that are implemented / being implemented. Each class inherits from the abstract class "PostProcessor", which provides a framework for future post-processing modules. †Planned implementation

Post-processor: Remove Redundant Sequences

This post-processing module iterates through all the sequences and removes any duplicates. In order to ensure that this is performed in a fast fashion, the program iterates through each sequence in every file and uses the SipHash hashing function on each sequence to store in a lookup table. If a sequence produces a hash that already exists, that sequence is removed from the dataset.

Post-processor: Sequence Length Filter

This module iterates through all sequences and removes any longer or shorter than the provided threshold.

Post-processor: Non-Canonical Characters Filter

This module iterates through all sequences and removes any that contain non-canonical amino acid single letter codes or any characters that are not amino acids.

Post-processor: Data Maker

This module creates training, validation, and test sets for ML training purposes. The user can determine how many sequences to process, which are then randomly sampled from the folder containing OAS C/S files. The user can specify the training/validation/test set ratios or directly specify the number of sequences in each set. If desired, a label for each sequence can be kept for classification tasks. The user can also determine the ratio between each class in the dataset or specify the specific number of sequences for each class.

Post-processor: Combine Files

This module combines files to create new files that are as close as possible to a defined file size. This module is especially useful to use before other post-processing modules that can use up a lot of memory per file processed such as "antibody_viability.py" and "encode_esm.py".

Post-processor: Antibody Viability Filter

This module is responsible for retaining high-quality antibody sequences while filtering out non-viable sequences. This module is a re-implementation of ABOSS, a deprecated software written in python 2.7.[31] ABOSS itself no longer works as its dependencies have been updated to the point that ABOSS cannot process sequences and the source code is only available upon request from OPIG. Briefly, the module first aligns each sequence to its closest germline using ANARCI[33] and IMGT numbering.[34] If no alignment could be found or the alignment score to the closest V or J gene is less than 0.5, the sequence is filtered out of the dataset. Next, the sequences are checked if certain conserved residues are present, most importantly Cys23 and Cys104, which are important for antibody stability. The check for conserved residues is species and light/heavy chain specific. Lastly, for each batch of 1000 sequences that is processed through this module, the probability of Cys23 and Cys104 mutations is calculated. This mutation rate can be considered the upper bound of the frequency of false reads in the next-generation sequencing read. Therefore, if a residue at a certain position occurs less frequently than the Cys23 or Cys104 mutations, then we assume that this residue is present due to a false read. A sequence containing a residue with that low of a probability of occurring is filtered-out.

Post-processor: Encode One Hot

Encodes sequences in the dataset as a one-hot encoded N x 21 matrix, where N is the padding length and 21 is the number of amino acids including one empty character. Sequences can also be returned as a flattened vector. The module returns the encoded sequences as a Numpy ".npy" file,[35] if a classification label is provided, then another file is returned with the labels.

Post-processor: Encode ESM

Encodes sequences in the dataset as an ESM encoded N x D matrix, where N is the padding length and D is the embedding dimension.[36] D is dependent on the ESM model used, which can be specified. Sequences can also be returned as a flattened vector. The module returns the encoded sequences as a Numpy ".npy" file,[35] if a classification label is provided, then another file is returned with the labels.

	Author	Species	Vaccine	Isotype	BType	v_call
1638	Jiang et al., 2013	human	Flu	IGHG	Naive-B-Cell/Plasmablast	IGHV5-51*01
3515	Jiang et al., 2013	human	Flu	IGHG	Naive-B-Cell/Plasmablast	IGHV3-30*02
1638	Turner et al., 2021	human	SARS-COV-2	IGHG	Germlinal-Center-B-Cells	IGHV3-21*01
3515	Turner et al., 2021	human	SARS-COV-2	IGHG	Germlinal-Center-B-Cells	IGHV3-23*01
12747	Galson_2015 et al., 2015	human	MenACWY-polysaccharide	IGHG	Plasma-B-Cells	IGHV4-4*07

	Sequence_aa	cdr1_aa	cdr2_aa	cdr3_aa
1638	ISCQGSGYSFTNYWIGWVRQLPGKGLEYMGIVYPGDSDTRYSPSFQ...	GYSFTNYW	VYPGDSDT	ARRQGHSFGYGPGGTHDTFDI
3515	GGSLRLSCAASGFTFTNYGMHWVRQAPGKGLEWVALIGFDGQNKHY...	GFTFTNYG	IGFDGQNK	ATLRGSSYDTYVMDS
1638	EVQLVESGGGLVKPGGSLRLSCAASGFTFSSYTINWVRQAPGKGLE...	GFTFSSYT	ISSSSSYI	ARERYDSSGSESYYFDY
3515	EVQLLESGGGLVQPGGSLRLSCAASGFTFSSYAMSWVRQAPGKGLE...	GFTFSSYA	ISSSGGDT	AKGVRGAMIVVVIPYFDY
12747	SETLSLTCTVSHDSISSSYWSWIRQSADKGLEYIGRIHAAGSTAYN...	HDSISSSY	IHAAGST	ARRSLDNWYFDR

Figure 2.3: Example of a file downloaded and processed via OAS C/S, viasualized using "Pandas". Showing five sequences, due to the width of the "Pandas" data frame, the data was split in two and stitched together. Showing all columns.

Together, this suite of post-processing tools provides a practical and reliable way to obtain high-quality data to train antibody-specific ML models. The output can be specified to be CSV or JSON format, an example of a CSV file can be seen in Figure 2.3.

Mousify Antibody Humanization Software

A curious observation can be made once one looks at recent ML-based antibody humanization tools. Fundamentally, they all share the same structure to translate a non-human sequence (query sequence) to a humanized one (final sequence).[9,20,21] For example, Hu-mAb first produces every possible single mutant from the query sequence *in silico*, then scores each sequence based on a random forest classifier and then picks the highest scoring sequence. If the sequence has a higher score than the humanization threshold, then it is returned as the final sequence. If the score is less than the threshold, the program starts anew with the intermediate sequence replacing the query sequence (Figure 2.4)[20]. In another recent example, BioPhi, the program performs a masked token prediction on each possible position of the sequence. Then at every sequence position, it picks the highest probability residue and generates that sequence. The intermediate sequence is scored using a 9-mer peptide counting scheme. If the If the sequence has a higher score than the humanization threshold, then it is returned as the final sequence. If the score is less than the threshold, the program starts anew with the intermediate sequence replacing the query sequence (Figure 2.4).[21] Looking at these examples in a more abstract fashion: First, a query sequence is used to define a matrix of transition probabilities. Then an intermediate sequence is generated by picking the maximum likelihood transition over the whole matrix or over per-position vectors. Then a score is given to this intermediate sequence and a decision is made if the sequence is humanized enough or if the cycle should continue (Figure 2.4).

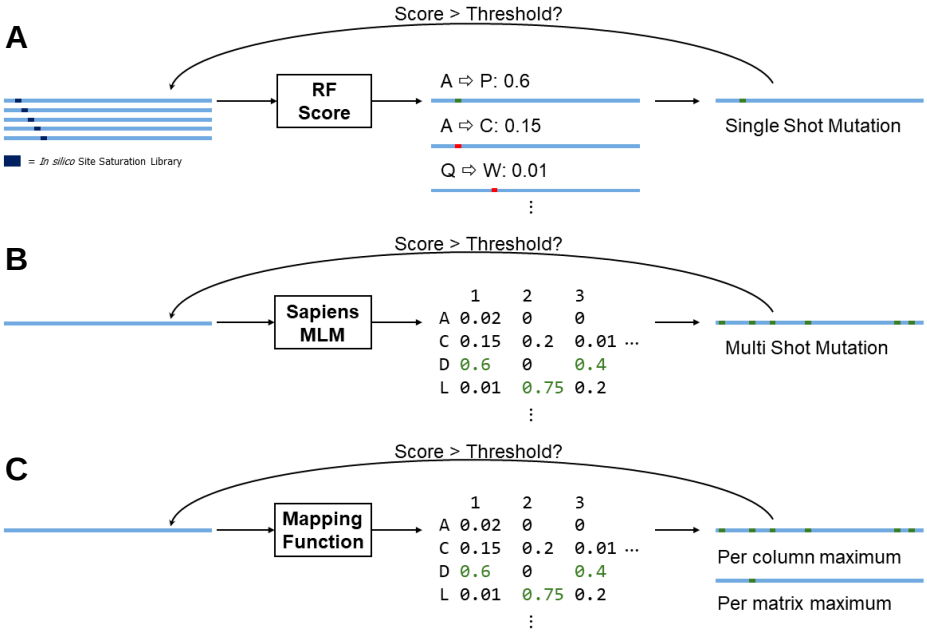


Figure 2.4: Examples of the general model structure for antibody humanization. (A) General structure of the Hu-mAb model. The model generates all possible mutants from the query sequence in silico. Then each possible next sequence is scored according to a random forest classifier model. The highest scoring sequence is selected, and if the score of that sequence is below a set threshold,

the cycle continues. (B) General structure of the BioPhi model. The model calculates a (LxN) probability matrix of the query sequence via the Sapiens Masked Language Model (MLM). The Lth column represents a position in the antibody sequence and the Nth row an amino acid. The probabilities in each entry are the probabilities that the Ni amino acid appears at position Lk, conditioned on all other positions are occupied by the sequence of the query. The model selects the maximum likelihood amino acid at each position of the sequence. If the score of that sequence is below a set threshold, the cycle continues. (C) Abstraction of antibody humanization models. Fundamentally, the models are made up of a mapping function (calculating the transition probability matrix), a scoring function and a mutation algorithm. In this case one can choose between a multi shot or a single shot "Greedy Walk" algorithm.

The advantage of thinking about antibody humanization in this abstract fashion is that it allows us to start combining the components that make up individual models. For example, we could run BioPhi with a single shot greedy walk, instead of the multi shot defined in the web application,[37] which makes more sense anyway (See: Antibody Humanization and Machine Learning). We could also use the transition probability matrix calculation scheme from Hu-mAb and apply it to the scoring function of BioPhi (OASis).[21] More importantly for this thesis, this abstraction allows us to easily replace the "Greedy Walk" algorithm shared by all of the models so far,[20,21] with a Markov chain or a diffusion algorithm (See sections: Generation of Humanized Antibody Libraries using Mousify & Antibody Humanization).

We developed Mousify as a ML-based antibody humanization model scaffold, that takes advantage of the aforementioned abstracted structure. Mousify is made up of four major components, represented by their respective abstract classes (Figure 2.5). The main interface of Mousify manages the entire system by keeping track of the state of each class, as well as which sequences have been accepted by the mutator in a sequence registry. At the end of humanization, Mousify returns this sequence registry as a file which contains the sequence, the score of the sequence and what mutation(s) generated this sequence from the previous one. The main function is also managing multiprocessing in case a user wants to accelerate humanization by using multiple cores. The "Discriminator" abstract class has the task of scoring a sequence with a "humanness" score. Since most scoring functions are ML-based, a notable exception being OASis, this class also requires an "Encoding" class to represent the amino acid string as a numerical object that an ML algorithm can understand. The "Map" abstract class translates an amino acid sequence into a Lx21 matrix of transition probabilities. Where L is the length of the antibody sequence and 21 represents the 20 canonical amino acids, plus one character representing an "empty" position. Lastly, the "Mutator" abstract class implements a "walking" algorithm. Metaphorically one can imagine walking in a mountain range, where taking a step in any direction represents one of the L·21 possible mutations and where the elevation gain or drop associated with that step represents the increase or decrease in the "humanness" score.

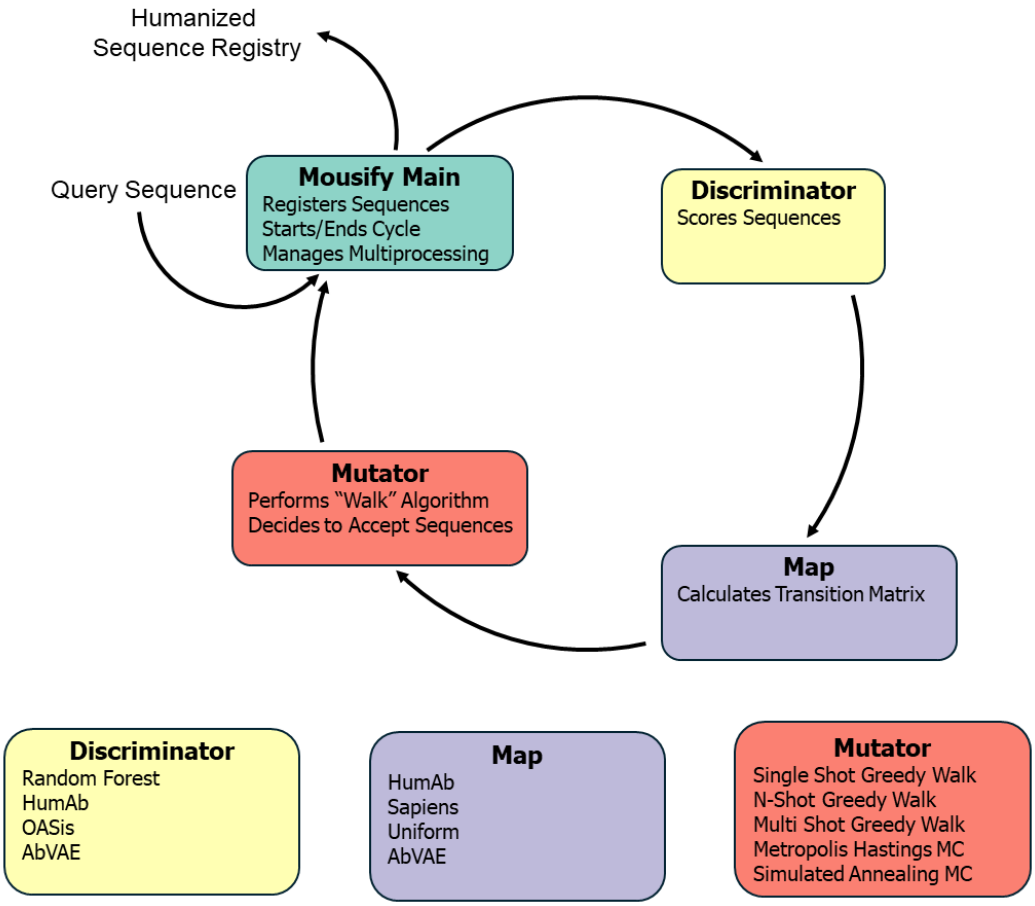


Figure 2.5: Mousify model architecture overview (top) and an overview of specific implementations (bottom). The main class of Mousify is the part that users interface with. It keeps track of all sequences that have been accepted and decides when the cycle of generating sequences stops. The discriminator class scores a given sequence (i.e. it discriminates between a human and a non-human sequence). The map class calculates a transition probability matrix from a sequence to all possible mutants of that sequence. The mutation class performs the "walking" algorithm specified. Specifically, it decides which sequence(s) is/are selected to pass on, based on the results of the discriminator and map functions.

Briefly, the Mousify architecture humanizes antibodies in the following way: A user defines a query sequence and chooses one implementation of the Discriminator, Map and Mutator class. Mousify registers the query sequence as the first sequence in the sequence registry and calculates its score using the Discriminator. Then Mousify uses the Map class to calculate the transition probability matrix and passes it on to the Mutator class. The map is also provided an antibody numbering scheme (IMGT, Chothia etc.) to define CDR residues, since we do not want those mutated. The map sets the probability of mutating CDR residues defined by the numbering scheme to 0 and normalizes the map. Depending on the Mutator chosen, the class generates a sequence that is either a single mutation away or multiple mutations away from the query sequence, using the information from the transition probability matrix to decide

on the sequence to generate. The Mutator then needs to accept the sequence as a valid next step. If the sequence is rejected, the Mutator will continue to generate sequences until one is accepted. If the sequence is accepted, Mousify will register the sequence in the sequence registry along with its score, and what mutation(s) from the query sequence generated this entry. Mousify will then decide whether to stop the cycle and return the sequence registry, or whether to continue the cycle while using the last accepted sequence as the input for the cycle. Stopping the cycle can happen under multiple user-defined conditions. The cycle can stop when a threshold of humanness has been reached, when a certain number of sequences have been generated, or, in case of a “Greedy” algorithm if no improvement of the humanness score can be achieved.

Benchmarks for Antibody Humanization

To the best of our knowledge, all currently published ML models lack experimental validation, as described in more detail above (See section: Antibody Humanization and Machine Learning). This poses an interesting problem on how to compare the different models available in the literature. Models like Hu-mAb or BioPhi, judged the effectiveness of their models based on three metrics: The models' classification performance, the correlation of the score to experimentally available ADA values, and the number of mutations it takes to mutate a non-human sequence to a human sequence.[20,21] With the exception of the correlation coefficient, we do not believe that the other provided metrics make a good comparison between models. First, classification performance is a metric that is not related to the problem one wants to solve when humanizing an antibody, which is to reduce ADA. For example, ~25% of mouse antibodies have an ADA that falls within the 95% of human ADAs. It is worth noting that the dataset that we have access to is heavily biased towards antibodies that at least made it to the first phase of clinical trials (Figure SC.3). Second, the number of mutations it takes to humanize, or the amount of overlap to therapeutic antibodies, is an irrelevant metric in a system as complex as antibodies. Single mutations can vastly change the properties of an antibody,[38–40] and as demonstrated by the plethora of protein language models (PLM),[27,36,41–44] mutations are also heavily dependent on context.

Given this context dependance of mutations, we have to ask how well do ML humanized antibodies perform experimentally compared to the original antibodies? There are a few sets of experiments that pharmaceutical companies subject antibodies to in order to evaluate the effectiveness of experimental humanization. These experiments are essential to know which antibodies are moved forward to be tested in more complex experiments, such as pharmacokinetic studies, animal studies or clinical trials.[14–16] In addition, our goal is to set benchmarks that other groups can easily replicate after developing an ML-based humanization model. We chose to set benchmarks for humanization models using four simple, easily replicable experiments:

- Expression of antibodies in HEK293F or Expi293F cells.
- Thermofluor assay to determine the apparent melting temperature.
- Quantitative ELISA to determine the EC_{50}.
- Polyreactivity ELISA.

It is also important to consider which antibodies to subject to humanization to set benchmarks. The starting antibody cannot be too unstable, otherwise already a small set of mutations is very likely to be deleterious.[45] The antibody should also be able to bind its antigen well and shouldn't be polyreactive. We believe that an ideal antibody humanization model should be able to conserve antibody properties while improving immunogenicity and not necessarily improve other properties, even

though that is desirable. Therefore, the benchmarking antibodies should have a good developability profile.[14] We hypothesize that engineered mAbs that have FDA approval likely originated from antibodies that had a good developability profile. Therefore, using the original antibodies of FDA approved drugs not only provides us with a good benchmarking candidate, but also provides us with two reference points to compare against. One reference point being the original antibody, the other one being the therapeutic antibody that is FDA approved. For the purpose of this study, we chose three FDA approved mAbs to re-humanize and one very promising anti-SARS-CoV-2-RBD mAb from the laboratory of Pamela Bjorkman in order to set benchmarks for antibody humanization models (Table 2.1).

Antibody	Antigen	Indication(s)	ADA (%)
Certolizumab	Human TNFα	Crohn's Disease Rheumatoid Arthritis Psoriatic Arthritis Axial Spondyloarthritis	4.4-11.7†
Omalizumab	Human IgE	Severe Allergic Asthma Chronic Spontaneous Urticaria	0‡
Palivizumab	RSV Fusion Protein	Prophylactic against RSV infections	1.1-4.8*
M8a-3	SARS-CoV-2 RBD	N/A	N/A

Table 2.1: Antibodies used in this study, their targets and, if applicable, their indications and ADA. † Indication and ADA data from Deeks (2016); ‡Indication data from Okayama *et al.* (2020) and ADA data from Chen *et al*. (2023); * Indication data from O'Hagan *et al*. (2023), ADA lower bound from Marks *et al*. (2021) and upper bound from Robbie *et al*. (2012).

Certolizumab is an anti-human-TNFα mAb that is indicated for Crohn's disease, rheumatoid arthritis, and other inflammatory autoimmune diseases.[46] Also known as Cimizia®, it is produced by the Belgian pharmaceutical company UCB as a PEGylated recombinant Fab` fragment. Generally, the drug is well tolerated and has low ADA with between 4.4-11.7% of patients developing antibodies against the drug. Most antibodies developed by the patient are against the Fab`, but antibodies against the PEG moiety itself are possible. Certolizumab pegol is on the World Health Organization's list of essential medicines.

Omalizumab is an anti-human-IgE mAb that is indicated for the treatment of a severe asthma in adults and children, chronic urticaria, as well as allergies from aerosols and food allergies.[47,48] Omalizumab is marketed as Xolair® by the US based company

Genentech. The mAb can have a significant impact on patients' quality of life, with improved symptoms and up to 96% reduction in hospitalizations of patients after 24 weeks of treatment. The drugs ADA is impeccable with multiple studies reporting 0% of patients developing anti-drug antibodies.

Palivizumab is an anti-RSV-F-Protein mAb that can be used prophylactically in populations in high-risk of contracting RSV to mitigate serious RSV disease.[49] Palivizumab is marketed as Solair® by the Swedish pharmaceutical company AstraZeneca. While effective prophylactically, it has been demonstrated that treatment with palivizumab does not benefit the patient in symptomatic RSV disease. Depending on the source, the percent of patients that develop anti-palivizumab antibodies is between 1.1-4.8%.

M8a-3 is an anti-SARS-CoV-2-RBD mAb co-developed by research groups at Caltech, Rockefeller University, and the University of Oxford. This antibody is highly interesting as it is broadly neutralizing amongst Omicron SARS-CoV-2 variants due to it being able to target the conserved class 1/4 epitope of RBD. It was developed by immunizing mice with a mosaic nanoparticle displaying RBDs from different sarbecoviruses. Contrary to the other antibodies in this study, it is not approved for clinical use and no patient ADA response is known.

In this study, we compared six different antibody humanization models applied to four different antibodies. We first used CDR grafting as a baseline for benchmarking since it is the simplest form of antibody humanization that does not require further engineering steps. Usually, CDR grafting is followed by introducing stabilizing mutations or back-mutating Vernier zone residues.[8,13] These engineering steps can be subjective however and therefore we decided not to introduce any further mutations. We performed CDR grafting by using the web application hosted by BioPhi and applied it to all four mAbs presented above (Table SC2.1).

Next, we generated humanized antibodies using our own Mousify software. In order to integrate OASis and Sapiens into Mousify, we installed the BioPhi package in the Mousify environment. To implement OASis as a "Discriminator" we replicated the code that OASis uses inside of a new class that inherits from "Discriminator".[21] The Sapiens model was available as a pretrained model from fairseq,[50] therefore only a call to the Sapiens' prediction method was necessary to implement a replica as a "Map".[21,28] The Hu-mAb model source code is not available on any public repository, therefore we had to fully replicate the model as best we could from the description of the model architecture and training scheme.[20] One major difference to the original Hu-mAb model is that we trained the random forest models only on antibody sequences from B-cell types that would have under gone self-antigen counter-selection. We hypothesized that such a dataset would perform better in generating humanized sequences, however it did not result in a model that classifies sequences better or an improved correlation coefficient of score vs ADA. The module called

"Hu-mAb Discriminator" represents the random forest probability score implemented in a class inheriting from "Discriminator". The module called "Hu-mAb Map" describes the class inheriting from "Map" that assigns a probability to a mutation depending on its increase in humanization score compared to all other possible mutations. The humanness scoring function used by "Hu-mAb Map" is the same as defined as the "Discriminator" in Mousify. All Mousify generated sequences used the "Greedy Walk" algorithm in single-shot mode All models using OASis as a discriminator were run with min_fraction_subjects = 0.15.[21] We also used the web applications of Hu-mAb and BioPhi to humanize mAbs to generate sequences the way the developers intended to.

In addition to the humanized antibodies above, we sub-cloned the original mouse antibodies as well as their therapeutic versions into a p3BNC vector. Heavy chains were sub-cloned with human IgG1 CH1 and Fc, light chains were sub-cloned with human CL kappa. We decided to clone mouse VH/VL chains with human CH/CL because otherwise ELISA experiments would have to be performed with different secondary antibodies, making direct comparisons harder.

After sub-cloning we transfected paired heavy and light chain vectors into Expi293F cells. We attempted expression of antibodies at least four times before labelling the variant as not being able to express. After expression and purification via MabSelect column, we ran a thermofluor assay of each purified antibody (See Materials and Methods: Thermofluor Assay) at pH7.2 and pH5. The latter assay was performed solely to help in assigning unfolding transitions to their antibody domains as described previously (Figure 2.2C-D, Figure SC.4-11).[51,52] The thermofluor assay was performed in six replicates. The results were first derived using the Numpy[35] gradient function (Figure 2.2A-B, Figure SC.12-19) and we then generated 1000 unfolding curves using statistical bootstrapping. The median apparent melting temperature (T_M(App)) of the Fab is reported in Table 2.1.

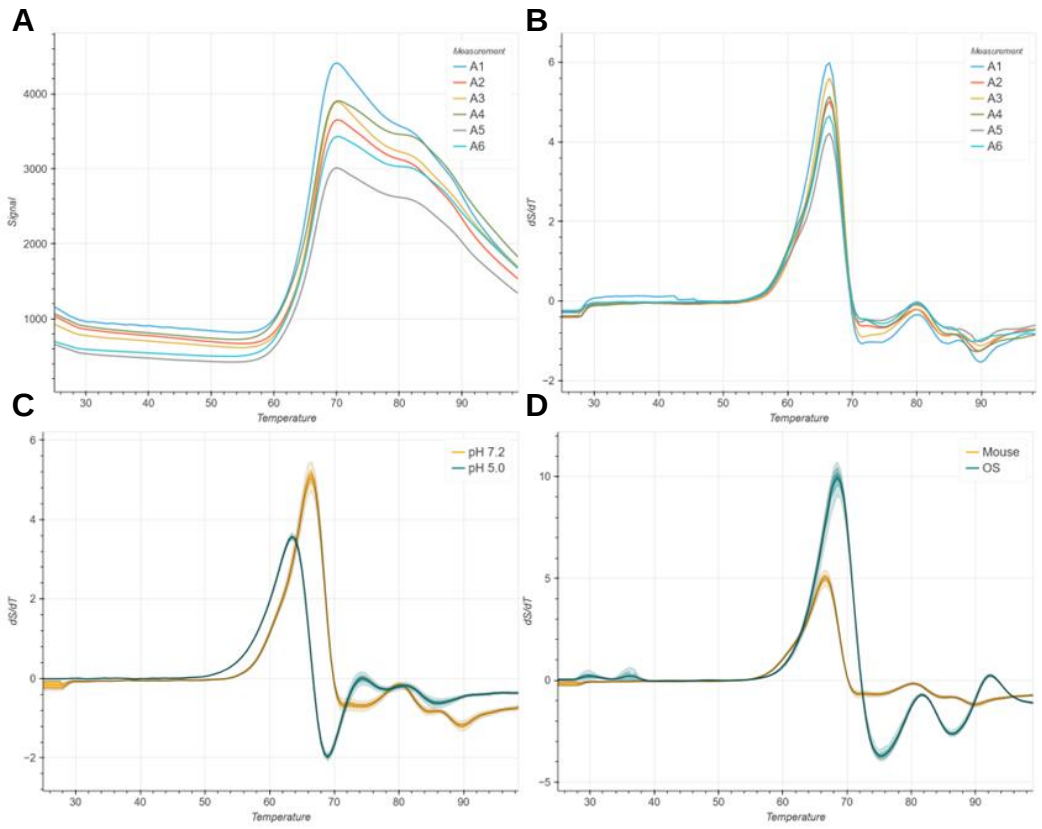


Figure 2.6: Overview of thermofluor data analysis and peak assignment. (**A**) Plot of raw thermofluor data is derived using the numpy gradient function (**B**) to find the inflection points (i.e. apparent melting temperatures) at the maxima. Then we perform statistical bootstrapping on these datapoints, as well as a measurement of the same antibody at pH5, to obtain the derived melting curves at different confidence levels (**C, D**). Both of these plots help in assigning the peaks to their respective domains. For example, it has been reported that IgG CH1 and the Fab shift the most when exposed to lower pH buffers. Considering literature values of IgG CH1, we can determine that the peak at around 65°C is IgG CH1. The peak at around 80°C is IgG Fc,[51] as it does not change at lower pH. Lastly, we use the plot in panel (D) to solidify our peak assignment hypothesis. Since we are comparing a humanized antibody to its original mouse antibody, we would only expect the peak of the Fab to change significantly. This is used as further evidence that the T_M(App) of M8a-3 is 86°C at pH 7.2.

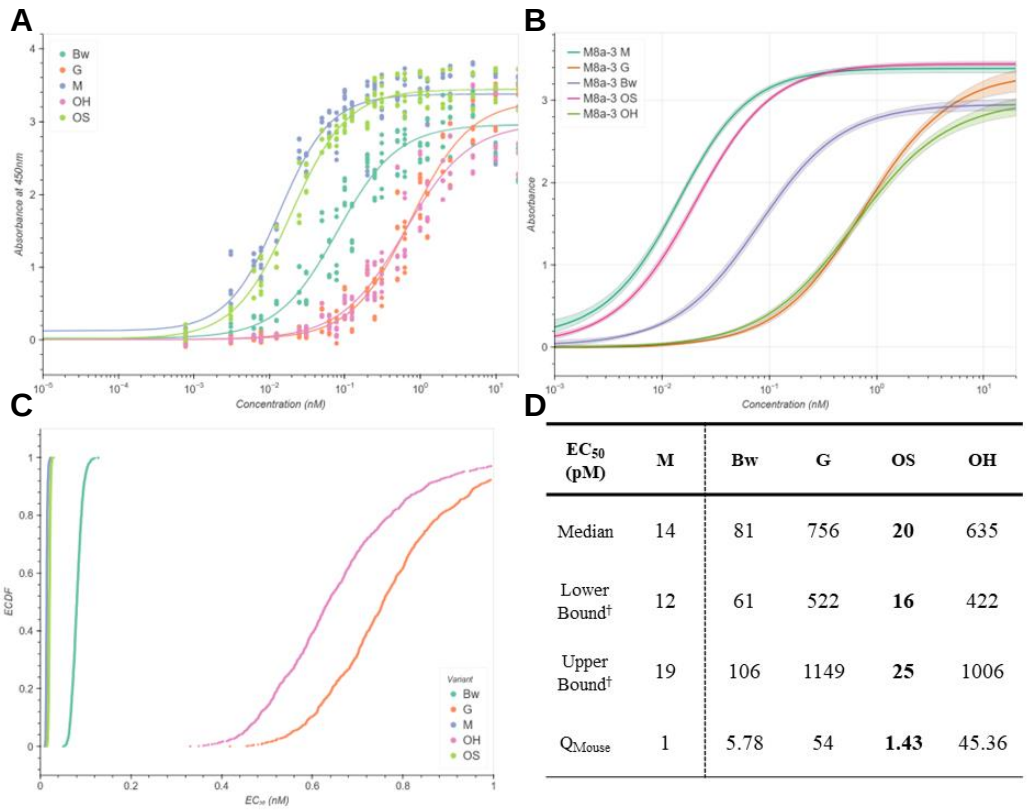
Unsurprisingly, the therapeutic antibodies all expressed and have a T_M(App) of at least the value of the original mouse antibody. CDR grafted antibodies only reduced the T_M(App) for the M8a-3 antibody, however on average CDR grafting reduced the T_M(App) by 2°C (Table 2.2). The BioPhi web application generated one antibody that we were not able to express, but generally performs well, increasing the mean T_M(App) by 1.2°C. This value is similar to its most closely related Mousify variant (OS: OASis Discriminator & Sapiens Map), for which every antibody expressed and increased the T_M(App) by 1.5°C on average. Both Hu-mAb models did not perform well, most antibodies did not express and those that did had a significantly lowered T_M(App). Lastly, the hybrid model (OH: OASis Discriminator & Hu-mAb Map) performed very well on antibodies that did express, on average increasing the T_M(App) by 5.8°C. However, given that we couldn't count non-expressed antibodies

in the calculation of the average temperature difference, gives a false impression that the OH model performs better than OS. In fact, if we omitted the Omalizumab OS humanized antibody, the average difference would rise to +5.4°C, on par with the OH model.

$T_{M(App)}$ (°C) †	M	T	G	Bw	Hw	OS	OH	HH
Certolizumab	75.5	80	77	*d.n.e.*	*d.n.e.*	75.5	**86.5**	79.5
Omalizumab	75.5	86	77.5	**78***	*d.n.e.*	65	*d.n.e.*	*d.n.e.*
Palivizumab	79.5	79.5	80.5	87.5	43	**90**	78	53
M8a-3	86	N/A	75.5	78.5	*d.n.e.*	92	**94**	*d.n.e.*
Mean $\Delta T_{M(App)}$	0	5	-2	1.2	-36.5	1.5	**5.8**	-11.2
$\sigma \Delta T_{M(App)}$	0	4.3	4.9	6.5	0	7.9	5.3	15.3

Table 2.2: Results of the expression and thermofluor experiments on humanized antibodies. At the bottom of the table we are showing the mean difference to the mouse antibody and the biased standard deviation of the mean difference. **M**: Mouse; **T**: Therapeutic; **G**: Grafted; **Bw**: BioPhi Web Application; **Hw**: Hu-mAb Web Application; **OS**: Mousify OASis Discriminator, Sapiens Map; **OH**: Mousify OASis Discriminator, Hu-mAb Map; **HH**: Mousify Hu-mAb Discriminator, Hu-mAb Map; d.n.e.: did not express; †Apparent melting temperatures are accurate up to ±0.5°C, temperature percentiles generated during bootstrapping produced confidence intervals less than ±0.5°C, which represents the accuracy of the assay. *Omalizumab "Bw" produced two peaks that could be identified as the melting temperature of the Fab, the second peak is at 74°C±0.5°C.

After comparing their T_M(App), we wanted to investigate the change in binding to an antibody's antigen via quantitative ELISA (See Materials and Methods: ELISA). Every ELISA for each antibody/antigen pair was performed in triplicates using the same antigen concentration (0.2μg/ml) and the same range of antibody concentrations (20nM-0.078pM). After measuring the ELISA absorption data, we first check whether the raw data looks like it could be modelled by a four-parameter logistic curve (Figure 2.7A). Then we generated 200 logistic curves fitted using statistical bootstrapping on the raw data, from which we evaluated the median EC_{50} of each antibody and its 95% confidence interval (Figure 2.2B-D). For clarity, we plotted the entire distribution of EC_{50} data as ECDFs (Figure 2.2B, Figure SC.20). For M8a-3 the best performing model was the OS model, which slightly worsens the EC_{50} of the antibody. Its most closely related model Bw, again performed similarly, but a bit worse than OS. Both OH and the CDR grafted antibodies performed poorly, worsening the EC_{50} over 45-fold.

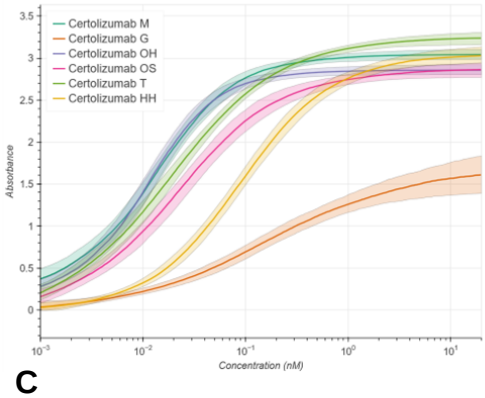


EC_{50} (pM)	M	Bw	G	OS	OH
Median	14	81	756	**20**	635
Lower Bound†	12	61	522	**16**	422
Upper Bound†	19	106	1149	**25**	1006
Q_{Mouse}	1	5.78	54	**1.43**	45.36

Figure 2.7: ELISA results of M8a-3 binding to WA1 SARS-CoV-2 RBD. (**A**) Plot of the raw absorbance data from the ELISA vs log-concentration. Each color represents a different humanization model. Curve in the plot represents the mean logistic curve fit. (**B**) Plot of the logistic curve fits of different humanization models at the median, the 25th percentile and the 75th percentile. Curves generated via statistical bootstrapping from the raw data. (**C**) Empirical cumulative density function (ECDF) plot of the EC_{50} values calculated from data in panel (B). (**D**) Table of the EC_{50} values calculated, reporting the median, bounds of the 95% confidence level and the ratio between the mouse EC_{50} and the humanized variant EC_{50}. **M**: Mouse; **G**: Grafted; **Bw**: BioPhi Web Application; **Hw**: Hu-mAb Web Application; **OS**: Mousify OASis Discriminator, Sapiens Map; **OH**: Mousify OASis Discriminator, Hu-mAb Map; **HH**: Mousify Hu-mAb Discriminator, Hu-mAb Map; †Lower bound represents the 2.5th percentile of data and the upper bound represents the 97.5th percentile of data.

Certolizumab humanized variants (Figure 2.8A-B) generally performed well with OH showing the only improvement of EC_{50} of any of the humanized antibodies in this study. For Omalizumab humanized variants (Figure 2.8C-D), the OS antibody performed the best, barely changing the EC_{50} compared to the mouse antibody. Combining all the data (Figure 2.8E), we can observe that the OS model performs the best overall, with the BioPhi web application model coming in in second place. The OH model has too high of a variance to draw any meaningful conclusions, but its change in EC_{50} on the M8a-3 antibody is significant, especially when considering the high degree of thermostabilization that the model improved on M8a-3.

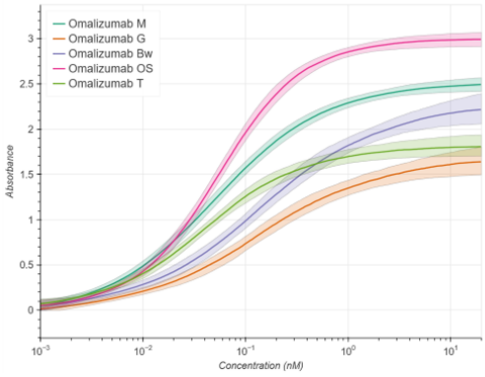
A



B

EC_{50} (pM)	M	T	G	HH	OS	OH
Median	13	18	166	90	21	**11**
Lower Bound†	4	3	31	52	4	7
Upper Bound†	30	42	$2.5 \cdot 10^5$	169	61	**26**
Q_{Mouse}	1	1.38	12.77	6.92	1.61	**0.85**

C



D

EC_{50} (pM)	M	T	G	Bw	OS
Median	51	40	139	144	**55**
Lower Bound†	25	13	34	56	**36**
Upper Bound†	99	94	2647	1790	**86**
Q_{Mouse}	1	0.78	2.73	2.82	**1.08***

E

Q_{Mouse}	M	T	G	Bw	HH	OS	OH
Certolizumab	1	1.38	12.77	*d.n.e.*	6.92	1.61	**0.85**
Omalizumab	1	0.78	2.73	2.82	*d.n.e*	**1.08***	*d.n.e*
M8a-3	1	N/A	54	5.78	*d.n.e*	**1.43**	45.36
Mean	1	1.08	23.17	4.3	6.92	**1.37**	23.11
σ	0	0.42	27.17	2.09	0	**0.27**	31.47

Figure 2.8: ELISA results of Certolizumab and Omalizumab. (**A**) Logistic curve fits at 50% confidence intervals for each expressed humanized variant of Certolizumab. (**B**) EC_{50} data summary of each humanized variant of Certolizumab. (**C**) Logistic curve fits at 50% confidence intervals for each expressed humanized variant of Omalizumab. (**D**) EC_{50} data summary of each humanized variant of Omalizumab. (**E**) Summary of ELISA EC_{50} data shown as the ratio between a humanized variant EC_{50} and the value for the mouse antibody. Showing the mean and the biased standard deviation. **M**: Mouse; **G**: Grafted; **Bw**: BioPhi Web Application; **Hw**: Hu-mAb Web Application; **OS**: Mousify OASis Discriminator, Sapiens Map; **OH**: Mousify OASis Discriminator, Hu-mAb Map; **HH**: Mousify Hu-mAb Discriminator, Hu-mAb Map; †Lower bound represents the 2.5th percentile of data and the upper bound represents the 97.5th percentile of data. *Antibody is polyreactive.

The final benchmarking experiment that we performed was a baculovirus prostate-specific membrane antigen (PSMA) polyreactivity assay (Figure 2.9).[53–55] The assay is often performed as a quantitative ELISA to obtain a BV score that correlates well with pharmacokinetic data.[55] However in our case, we were only interested in the affinity of humanized antibodies to baculovirus PSMA versus well-known positive and negative controls to determine polyreactivity. One humanized antibody variant of Omalizumab (OS variant) is striking for its very high polyreactivity. It compares to the Fleish antibody, that is known to be highly polyreactive. It is possible that the polyreactivity of Omalizumab OS is related to the 10°C decrease in thermostability, possibly indicating a change in the structure of the Fab that causes the high polyreactivity. The CDR grafted variant of M8a-3 also shows slight polyreactivity, being on par in signal with HCI Sap10 antibody, a slightly polyreactive antibody. Another M8a-3 antibody (OH variant) is possibly slightly polyreactive as some of the data near 1.5-times the inter-quartile range falls within the distribution of the HCI Sap10 antibody. No other antibodies variants showed signs of polyreactivity, meaning that ML-based antibody humanization models are generally good at conserving non-polyreactive properties of antibodies.

ML-based antibody humanization methods seem to struggle the most with generating antibodies that express, with seven out of twenty antibodies not expressing after four attempts. Notably, the Mousify OASis Discriminator/Sapiens Map model is the only model that performs as well as the CDR grafting model in terms of antibody expression. However, the Mousify OS model performs better in terms of thermostabilization of generated antibodies vs the mouse reference. Interestingly, it seems that models using the OASis discriminator perform better than models using the Hu-mAb discriminator. Possibly indicating that the 9-mer peptide counting metric performs well in correlating with patient ADA and generating stable antibodies. This also supports the hypothesis that models good at predicting a metric are not necessarily good at generating data that fit the metric. In other words, to make an ML-based discriminator that performs well, one needs to have antibody sequence generation as part of the training scheme. Due to insufficient data, we cannot yet conclude that this trend holds for EC_{50} data. We have attempted to express two different constructs from the literature of RSV Fusion Protein pre-fusion stabilized variants with their fusion peptides removed in our own laboratory and with the Caltech Protein Expression Center. So far, we have been unable to express this protein to use in the Palivizumab variant ELISA, which would complete the benchmarking data and allow us to draw conclusions more confidently on the Hu-mAb discriminator and Hu-mAb map performance. Nevertheless, we believe that the Mousify OS model is the best performing model so far when accounting for all four metrics. It is the only model that expresses as well as CDR grafted antibodies, while performing better in thermostabilization of antibodies. It is also the only model that is on par with the change in EC_{50} with the therapeutic antibodies reference points. However, one needs to consider that this model can generate polyreactive antibodies. Lastly, we want to discuss the difference in the BioPhi web application model and

the Mousify OS model, which are closely related to each other, but show different performances in the benchmarks. We believe that at fault is the wrong assumption by the developers of BioPhi that transformers trained as a masked language model (MLM) can predict high-order epistatic interactions, through the multi-shot greedy walk implemented.[21] While most MLMs mask multiple tokens during training, the prediction is performed on a single masked token at a time conditioned on all other tokens.[56] Therefore, only first order epistatic interactions can be reasonably expected to be accounted for in an MLM. In order to account for higher order interactions, one needs to perform multi-token prediction,[57] which can vastly increase the computational complexity due to combinatorial explosion,[58] or specifically embed higher order interactions in the model.[59] Given that Sapiens is based off of RoBERTa, means that no multi-token prediction was performed.[21,28,60]

In this study we have presented new benchmarking assays for the evaluation of ML-based humanization models. These benchmarks finally allow comparison of models beyond their score correlations to the percentage of patients that form ADAs. We have shown that the biggest hurdle for ML-based humanization methods is to produce antibodies that express, as well as negatively affecting binding to the target antigen. From the data we could conclude that the Mousify model using OASis as a discriminator, Sapiens as a map, and a single-shot greedy walk as a mutator module, is the best performing model in this set of comparisons. In the future we will be completing the data with ELISA data from Palivizumab variants, as well as comparing against the AbVAE model that we developed (See Antibody Humanization using AbVAE).

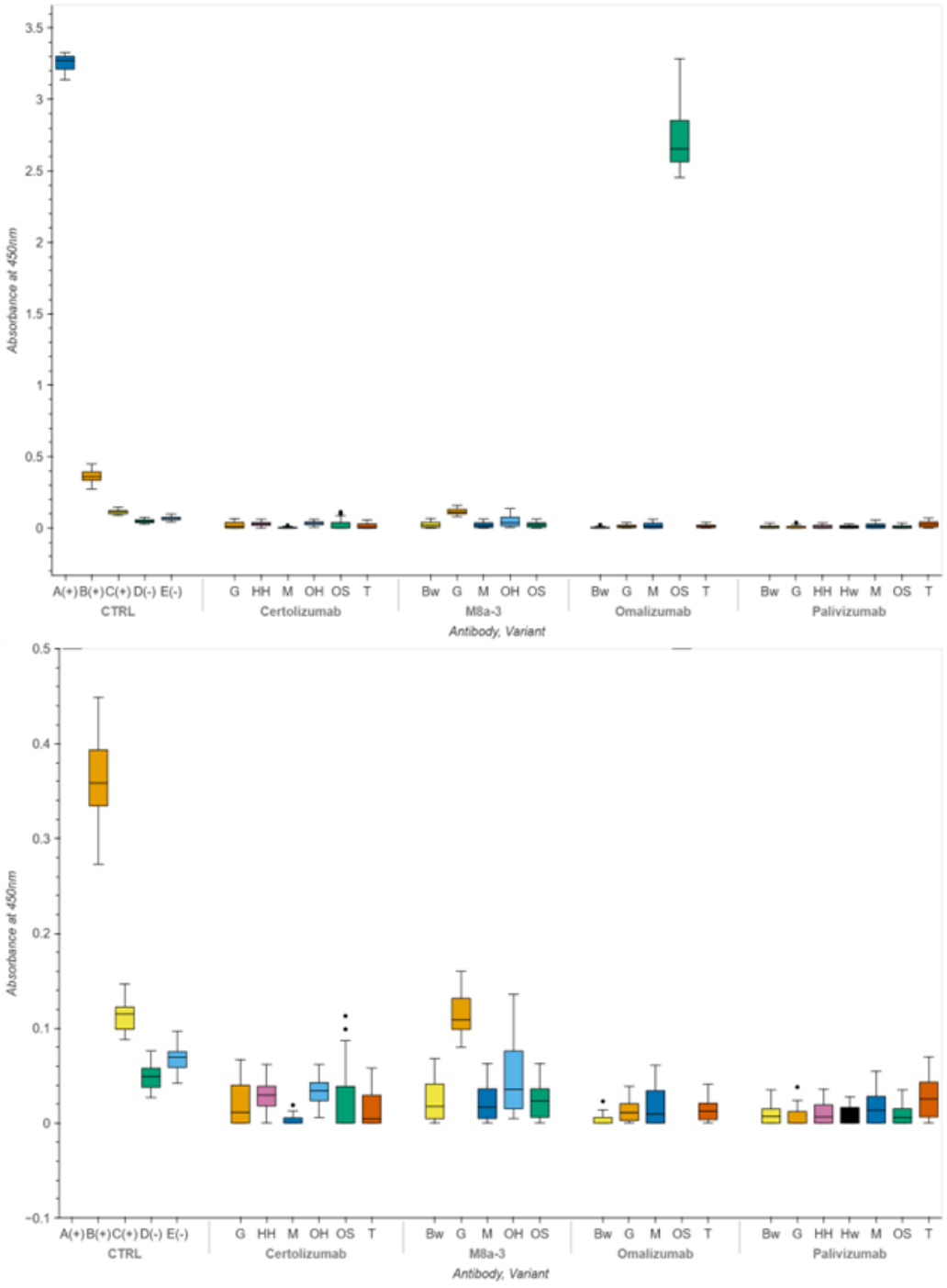


Figure 2.9: Baculovirus PSMA polyreactivity ELISA results for all expressed humanized antibodies. (**Top**) Full range of absorbance axis. (**Bottom**) Zoom-in on the absorbance axis to show the lower end of absorbances for clarity. **M**: Mouse; **G**: Grafted; **Bw**: BioPhi Web Application; **Hw**: Hu-mAb Web Application; **OS**: Mousify OASis Discriminator, Sapiens Map; **OH**: Mousify OASis Discriminator, Hu-mAb Map; **HH**: Mousify Hu-mAb Discriminator, Hu-mAb Map;

CTRL: Control; **A(+)**: Fleish antibody positive control; **B(+)**: 45-46m2 positive control; **C(+)**: HCI Sap10 positive control; **D(-)**: N6 negative control; **E(-)**: 10-1074 negative control.

Generation of Humanized Antibody Libraries using Mousify

A fundamental attribute of current ML-based antibody humanization models is that a query sequence is given as the input and a single "humanized" sequence is returned as the output. The user for such a system would be some entity that has a functioning non-human monoclonal antibody, and they want to engineer it to give it the highest chances of passing through clinical trials and be approved for medical use. Each antibody in the pipeline is engineered and then tested for essential properties, such as binding, thermostability, propensity to aggregation, and polyreactivity.[12] Afterwards, a risk assessment is performed on each antibody to figure out if it is worth to push a candidate to the pre-clinical development stage. This risk assessment includes the immunogenic risk,[15,16,61] as well as developability of the candidate.[14,62] We would argue that the probability of a single engineered antibody that passes the initial property tests to also pass all the risk assessments is very low. Now let us consider the probability that an ML-based antibody humanization model produces an antibody that passes all the essential properties tests. Assuming that the results from our experiments are all independently and identically distributed (iid.), we sampled 10,000 values from the distribution of values from our benchmarking experiments. Out of all the generated antibodies, only 9.6% would pass all the tests (Figure SC.21). Note: the data sampled is likely not iid. For example, a very unstable antibody is also very unlikely to bind to the target antigen specifically.

Having considered that fact, we have designed the Mousify architecture in such a way that we can generate libraries form any set of Discriminator and Map modules implemented (Figure 2.5). Since the map represents a transition probability distribution and the discriminator score a sort of "fitness" of the protein, it seems natural to generate libraries via Markov chain monte carlo (MCMC). A Markov chain is a process in which a step that is taken in the process is completely independent of past and future steps. Similar to the "Greedy Walk" described above, a Markov chain is a "walker", but with the caveat that the probability at arriving at a point θ is proportional to the fitness of the point $F(\theta)$ (Note: Normally MCMC is discussed as a walk over probabilities $P(\theta)$, but we adapted this vocabulary to protein fitness landscaped).[63] In a greedy walk, certain points are not accessible from a given starting point. It cannot reach the global maximum if it starts near a local maximum, as the greedy-ness would make the walker get stuck at the local maximum. Since in an MCMC the walker's probability of reaching a point is proportional to the fitness of the point, the walker can move downhill from a local maximum. The probability of doing so is just smaller than remaining in place. Therefore, given enough time, MCMC processes will explore the *entire* fitness landscape, but spending most of its time in areas of *high* fitness.[64] This is important when considering applying MCMC to ML-based antibody humanization models, the walker will eventually explore all possible variants, but will spend most of its time near variants with a high humanness score.

We have implemented two MCMC modules as a Mutator class, one being a Metropolis-Hastings MCMC (MHMC)[65] and the other being Simulated Annealing MCMC (SAMC).[66] Briefly, the implementations work as follows: Mousify proposes a mutation by sampling a residue to mutate to another specific residue. The probability to sample this position/amino acid pair is taken from the map module's transition probability matrix. This mutant is now the proposed next sequence. Next, we calculate the humanness score of the proposed next sequence. We then calculate the acceptance ratio $A(\delta, \theta) \in [0,1]$, where θ is the current sequence and δ is the proposed next sequence. We then accept this proposed next sequence as the next sequence with probability $A(\delta, \theta)$. This entire implementation works with multiprocessing, which means that we simultaneously run a chain for each CPU provided by the user. This allows us to explore the protein fitness landscape faster. The two modules, MHMC and SAMC, differ by the implementation of the acceptance distribution $A(\delta, \theta)$:

Metropolis Ratio (MHMC) : $$A(\delta,\theta) = Min\left(1, \frac{P(\delta)g(\theta|\delta)}{P(\theta)g(\delta|\theta)}\right)$$

Kirkpatrick Function (SAMC): $$A(\delta,\theta) = Min\left(1,\ e^{\frac{-(P(\delta)-P(\theta))}{T}}\right)$$

Where P(x) is the humanness score of the sequence x, g(x|y) is the probability of choosing the mutation from x→y, and T is the temperature parameter.

One of the primary methods to optimize MCMC algorithms is through getting the mean acceptance ratio to be around 1/3. Empirically, that is the value at which the algorithm has an optimal exploration/exploitation ratio to find high-humanness sequences efficiently. In this case, efficiently means with the fewest mutations. This is a problem for the MHMC method since the mean of the acceptance ratio proportional to the square of the standard deviation of the transition probability function ($A(\delta, \theta) \sim \sigma^2$). The only way to do this in Mousify is to restrict the transition probability matrix from the map module. We implemented a parameter that sets all probabilities below a user-defined quantile as 0 and re-normalizes the matrix. I.e. if the quantile is set to 0.9, then only the top 10% of transitions are considered. Besides the target acceptance ratio, we are also looking at getting the average number of mutations of the MCMC algorithm to the window of therapeutic antibodies. We analyzed the number of mutations that therapeutic antibodies are away from their original mouse antibodies using data provided in the Hu-mAb paper.[20] We found that for heavy chains, the number of mutations in the framework of VH are between 14 and 39, with a median of 24. Increasing the quantile value in the MHMC algorithm, does not greatly change the average number of mutations. In fact, increasing the quantile only slightly lowers the average number of mutations, but tightens the distribution of mutations around the mean (Figure 2.2), which does not provide us with any flexibility on the number of mutations desired. The SAMC algorithm has a much more elegant way of reducing the acceptance probability as well as the average

mutations, it is controlled by the temperature parameter T (Figure 2.10). We have also implemented a temperature annealing algorithm to manage the temperature over the course of the run. This allows us to have a high temperature in the warmup of the SAMC algorithm, and then lower the temperature once each chain has found its optimal exploration area of the fitness landscape. For this reason, we will only be using the SAMC algorithm going forward.

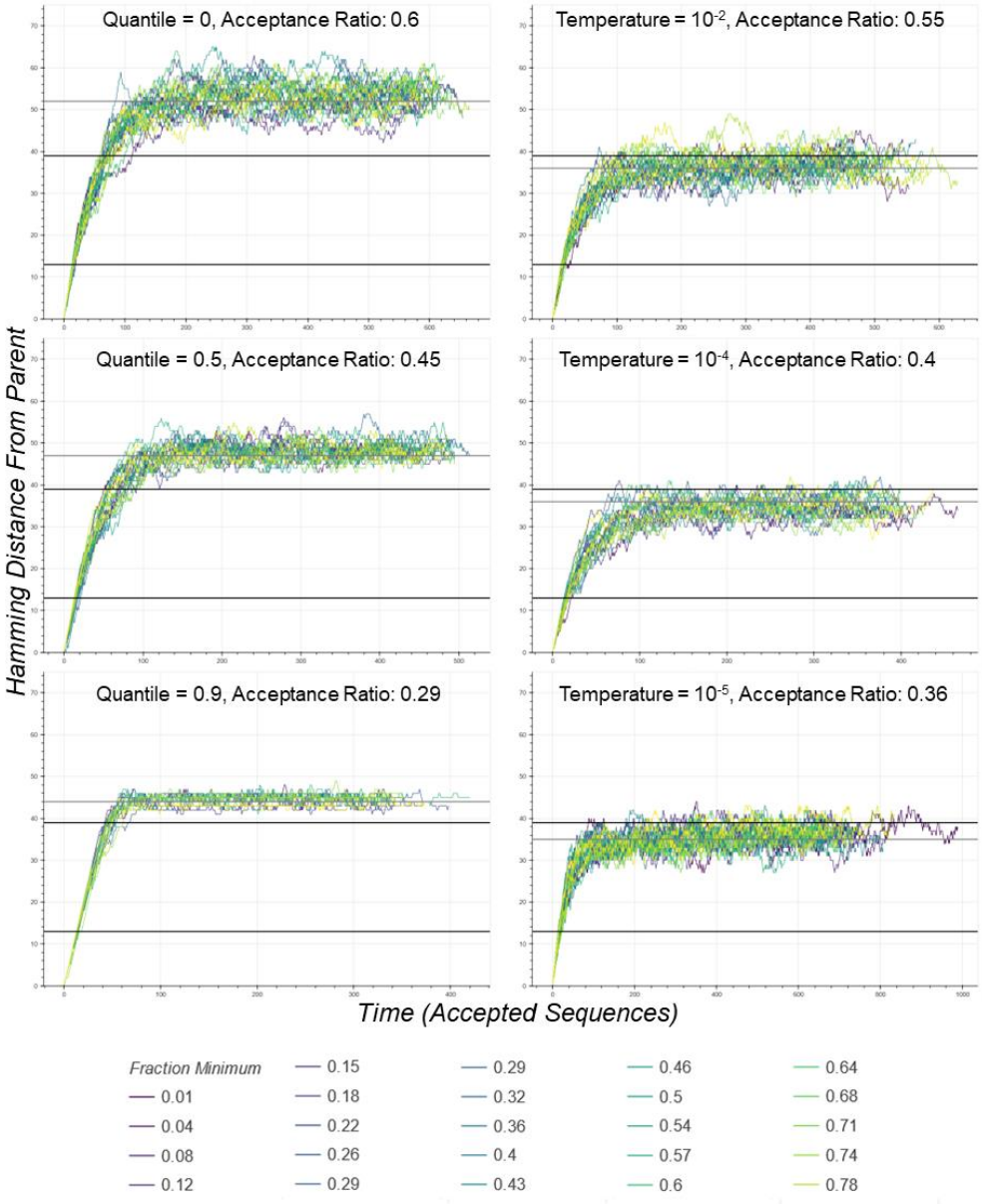


Figure 2.10: Chain-time vs Hamming distance from parent antibody plot comparisons between different MHMC and SAMC configurations. Left column: MHMC Plots. Regardless of the value of the quantile parameter, MHMC cannot manage to get the Hamming distance into the desired range for a random OAS non-human antibody. Right column: SAMC Plots. The temperature

parameter of the Kirkpatrick acceptance distribution allows for fine tuning of the acceptance ratio and the Hamming distance to the desired ranges, for a random OAS non-human antibody. Gray line: Average Hamming distance from parent for MCMC run. Bottom black line: Lower bound of mutations for therapeutic mAb VH. Top black line: Upper bound of mutations for therapeutic mAb VH.

In Figure 2.10, we explored each chain in the SAMC and MCMC runs with a different value for the OASis parameter "min_fraction_subjects" ("fraction minimum" in Mousify). This parameter controls the strictness of what 9-mer peptide is considered to be human. I.e. at a "fraction minimum" of 0.5, only 9-mer peptides that occur in at least 50% of human subjects are counted towards an antibody's humanness. While no discernible trend can be seen in Figure 2.10, we were curious to see what the distribution of OASis scores for each value of "fraction minimum" looks like. In Figure 2.11, we plotted the OASis scores achieved in an SAMC run with a temperature of 10^{-4} and compared it to the OASis score of therapeutic antibodies at the same "fraction minimum" threshold. From this we deduced that a "fraction minimum" score of 0.15 for an SAMC run best matches the distribution of therapeutic mAb OASis scores.

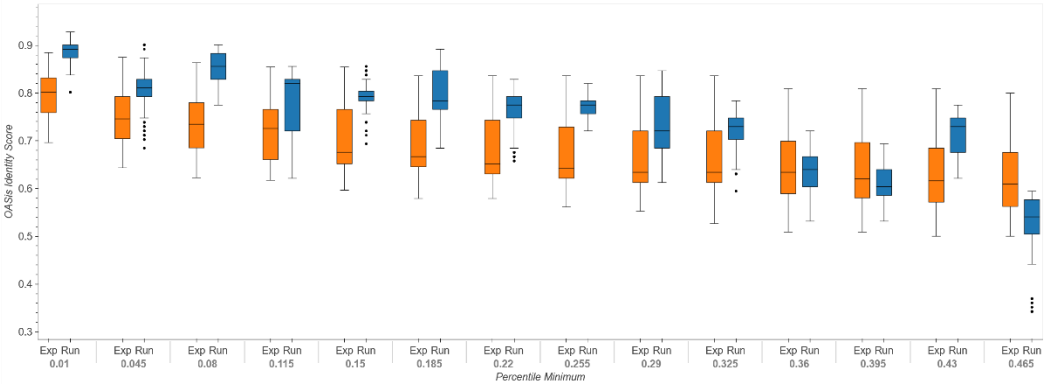


Figure 2.11: Boxplot of OASis scores vs Percentile Minimum values in an SAMC Mousify run. Percentile minimum was cut-off at < 0.5 for clarity. For plot below the 0.5 cutoff see Figure SC.25. Each box represents around 300 data points.

While our goal is to generate as many antibodies as possible using Mousify, one has to consider the cost of a library as well. For example, ordering a library of size 10^5 would cost around $30,000 for just the library of one of two chains, if ordered as an oligo pool from Twist Biosciences.[67] For many academic laboratories, this is prohibitively expensive. To overcome this cost constraint, we need to generate a sub-library that can be synthesized on a single DNA strand using degenerate codons. Sub-library generation tools such as DeCoDe take a set of sequences and generate N sub-libraries, where each sub-library can be synthesized in a single process. We attempted using DeCoDe but it failed to converge, as reported before.[68] In our case, this was likely due to the generated libraries being too diverse. A sub-library can only be

generated around sequence nodes that the Markov chains pass through multiple times.

Our task to bring down the cost of a Mousify library is to reduce the diversity of the libraries. The first implementation to reduce diversity is to analyze the data from the warm-up run of Mousify Markov chains. After finishing the warm-up, Mousify will analyze the distribution of residues at each position and calculate the consensus sequence. To make sure the starting sequence is a sequence that "makes sense" according to the model, we then find the closest sequence to the consensus sequence as the starting point of the second phase of MCMC. The second implementation is to reduce the number of sites that can be mutated. Mousify analyzes warmup data to rank the positions by the number of mutations accepted. A user can then specify, how many of these hotspots to consider for mutation in the second phase of MCMC. These library diversity reduction methods are optional for Mousify users, and both diversity reduction implementations can be used independently.

Even with these implementations of diversity reduction, DeCoDe still does not converge on a solution. In addition, DeCoDe is unpredictable in the time it takes to generate a library, which is impractical to use on an HPC where node time is reserved ahead of time. To circumvent the time and convergence issues of DeCoDe, we wrote a brute-force algorithm to generate libraries where users can define how much time to invest into finding a solution. Mousify LibraryMaker takes a sequence registry file from Mousify and always returns the best sub-library it could find in the time allocated. LibraryMaker first analyzes the sequence registry file for the residue distribution at each hotspot. The software then randomly picks a starting point from the set of sequences in the library, as well as a random permutation of residues from the hotspot distribution. Then in the order sampled, the software checks if adding a residue to the sub-library would generate sequences that are elements of the library. A slack parameter can allow the user to define how strict the membership of the sub-library needs to be compared to the library. For example, a slack of 0.1 would mean that in each step, 10% of the generated sequences can be outside the library. In our experience, this brute-force method can generate libraries of size ~10,000 in 4 hours, when run with a 20% slack and Mousify was run with 6 mutational hotspots. Note that 10'000 library members per chain, results in a 100,000,000-library size when combining light and heavy chains.

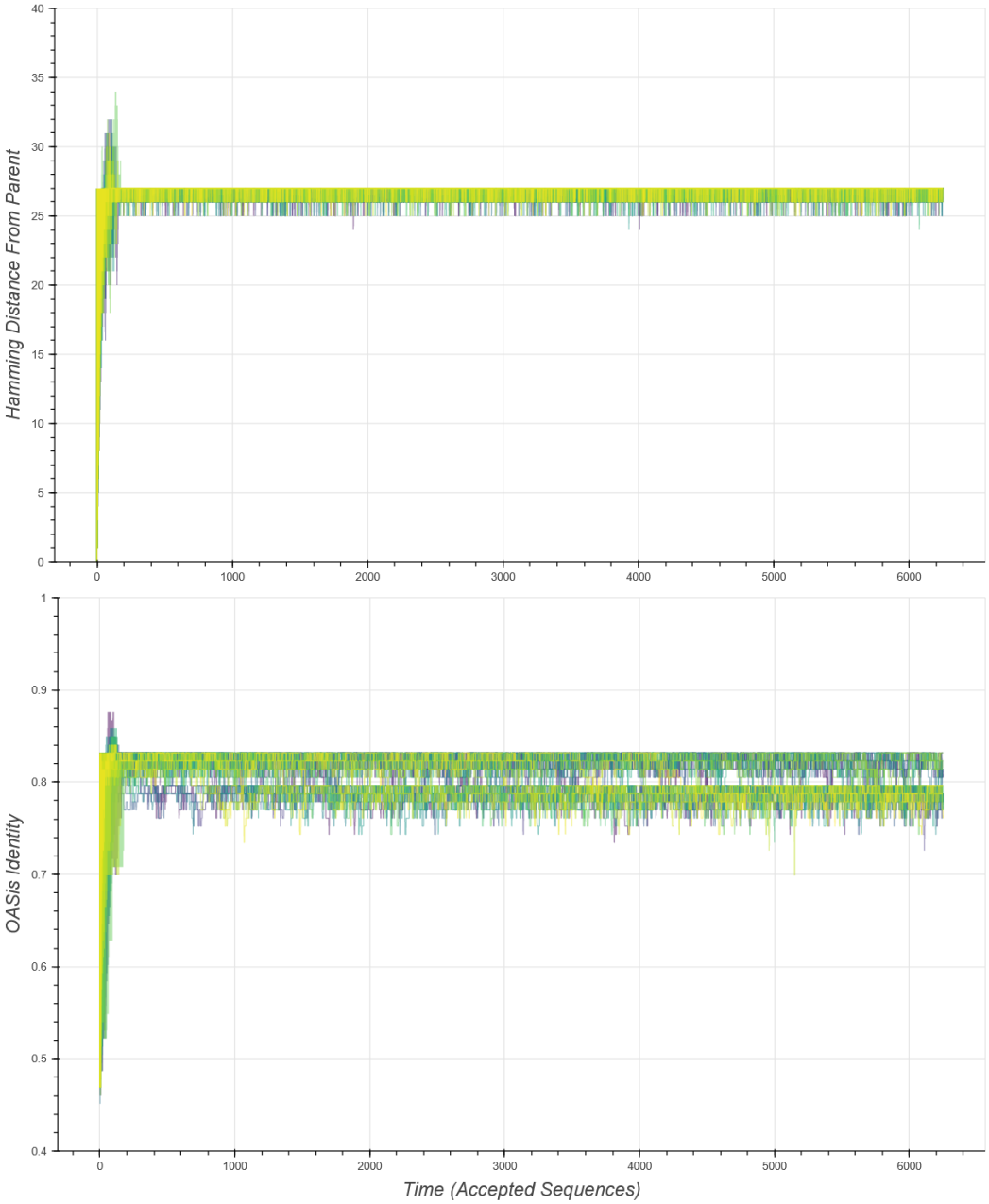


Figure 2.12: Markov chain time plots showing each chain in a different color versus a library metric. Only showing accepted sequences, 32 chains in total. (**Top**) Chain time versus Hamming distance to parent. (**Bottom**) Chain time versus OASis Identity score. One can observe the end of the warmup period at around 200 accepted sequences per chain, after which the diversity restraints are added. While diversity severely reduced the Hamming distance, diversity in humanness scores is till observed. In total running Mousify in this configuration, generated 750,000 sequences in 72 hours, of which around 200,000 were accepted.

We used the SAMC algorithm to produce a library of humanized antibodies based on the M8a-3 antibody. From the results of the section “Benchmarks for Antibody Humanization”, we concluded that the Mousify OS model is the best performing ML-

based antibody humanization model so far. Therefore, we used the OASis discriminator and the Sapiens map to generate a library of VH of M8a-3. We generated 10,000 warmup sequences after which we generated 750,000 sequences, where 200,032 sequences were accepted. We used an annealing schedule of T=0.005 during warmup, T=0.0025 for the first 10,000 sequences and T=0.000001 for the rest of the sequences. In total, 81,810 sequences were unique (40.9%) and showed a wide range of diversity in OASis scores (Figure 2.12, Figure SC.1). After using LibraryMaker with a slack of 20%, we obtained a library of 10,944 humanized M8a-3 VH domains. Note that a roughly 10,000-member library was our initial target since when combined with a VL library of the same size, that would result in a total library size of 10^8, which as large than what can reasonably be transformed[69] or explored[70] using yeast surface display. The library had a mutational Hamming distance from parent (25-27) just above the humanized therapeutic mAb median (24) and a humanization score distribution within the distribution of humanized therapeutic mAb scores (Figure 2.12).

To experimentally validate Mousify libraries, we use yeast display of Mousify libraries expressed as a Fab, together with fluorescence activated cell sorting (FACS).[70–72] The latter allows us to roughly evaluate the fitness of a library member in a very high-throughput fashion, before validating their fitness using ELISA, thermofluor and a polyreactivity assay. Performing yeast display in a conventional fashion, that is anchoring a Fab to the surface of a yeast cell via Aga2-Aga1,[73] would provide us with a non-scalable validation system. Since every sorted cell needs to be sequenced, the gene(s) of interest need(s) to be synthesized, followed by subcloning into p3BNC, and expression in mammalian cells. These steps are not only very time-consuming when scaled up to 10^2 sorted cells, but prohibitively time-consuming and expensive when sorting more than 10^3 cells. Therefore, we adapted a switchable yeast display/expression system developed by Van Deventer *et al*. (Figure 2.13).[74] The key idea behind the switchable system is the inclusion of an amber stop codon[75] before the gene of Aga2. Together with a plasmid that includes a constitutively expressed suppressor tRNA, as well as an Ome-Tyr-tRNA synthetase, the system can incorporate the non-canonical amino acid O-methyltyrosine (OmeY) and lead to yeast surface display of the Fab. When OmeY is not added to the yeast media, the cells cannot express Aga2 and due to the export signal peptide expressed with the Fab, the antibody is secreted by the cells. This allows us to add OmeY to the library transformants, perform cell sorting and sort the cells into media that does not contain OmeY. Growing the sorted single cells takes a week to express enough antibody for downstream validation experiments.

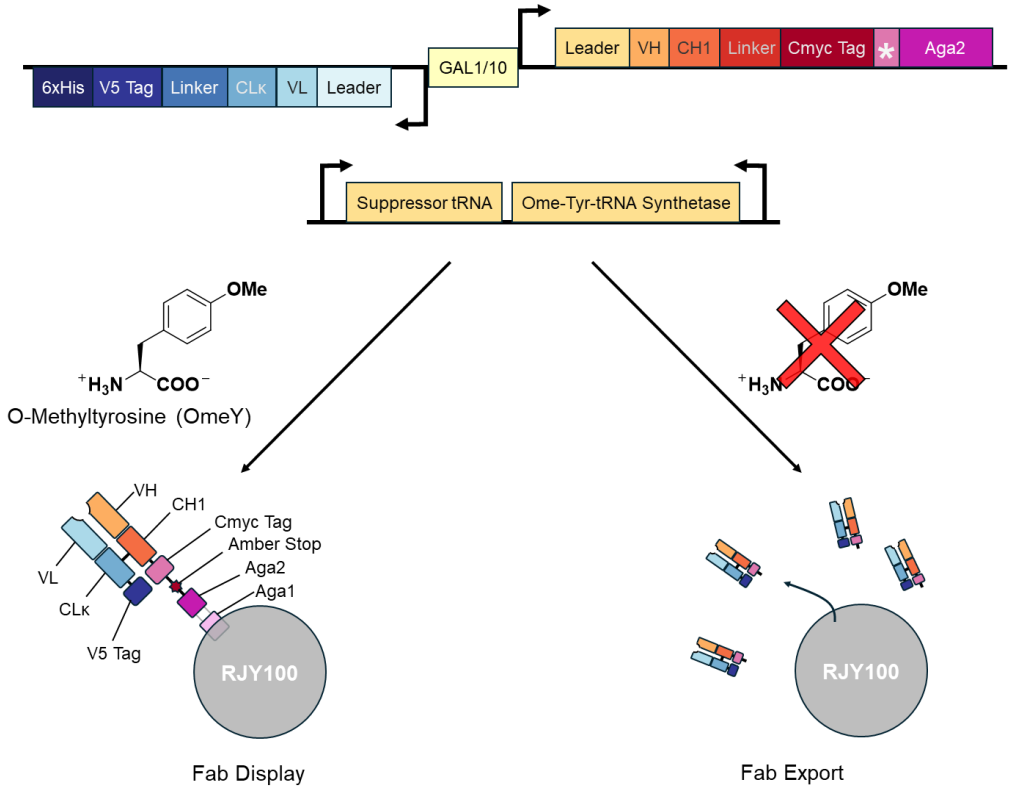


Figure 2.13: Yeast display system used in this study to experimentally evaluate Mousify libraries. This system is based on the yeast display method described by Van Deventer *et al*. (2015). The Fab is expressed on a dual-promoter (GAL1/10) and each chain of the Fab has its own tag for cell sorting. On the strand expressing the heavy chain of the antibody, we added an amber stop codon (TAG) before Aga2. A second plasmid contains the constitutively expressed suppressor tRNA and the Ome-Tyr-tRNA synthetase, which enables the OmeY incorporation. If OmeY is added to the RJY100 media (left), the Fab is displayed on the surface of the yeast cell, due to the successful expression of Aga2. If no OmeY is added (right), the Fab is exported and the system can be used for antibody expression. Leader: Export signal peptide; *****: Amber stop codon.

After transforming RJY100 yeast cells with the M8a-3 library incorporated into the genetic system outlined in Figure 2.13, the cells were grown in media containing 1mM OmeY to induce yeast display (See Materials and Methods). The cells were incubated with biotinylated SARS-CoV-2 RBD , before being stained with dye for one hour at room temperature. The original M8a-3 antibody was used as a positive control. Since the switchable yeast display/expression system is not entirely efficient, only 11.06% of cells had signal in the AF488 (V5-Tag) and AF647 (RBD) channels above background. This is slightly lower than the 18% that is reported in the literature.[76] However, the literature values were measured with scFv's, as opposed to the Fab's that were displayed in this experiment. Overall, the library display was successful, with 1.65% of cells in the double-positive region. Roughly, this would

indicate that 15% of the library express and bind RBD, while around 25% of the library express but do not show RBD binding above background.

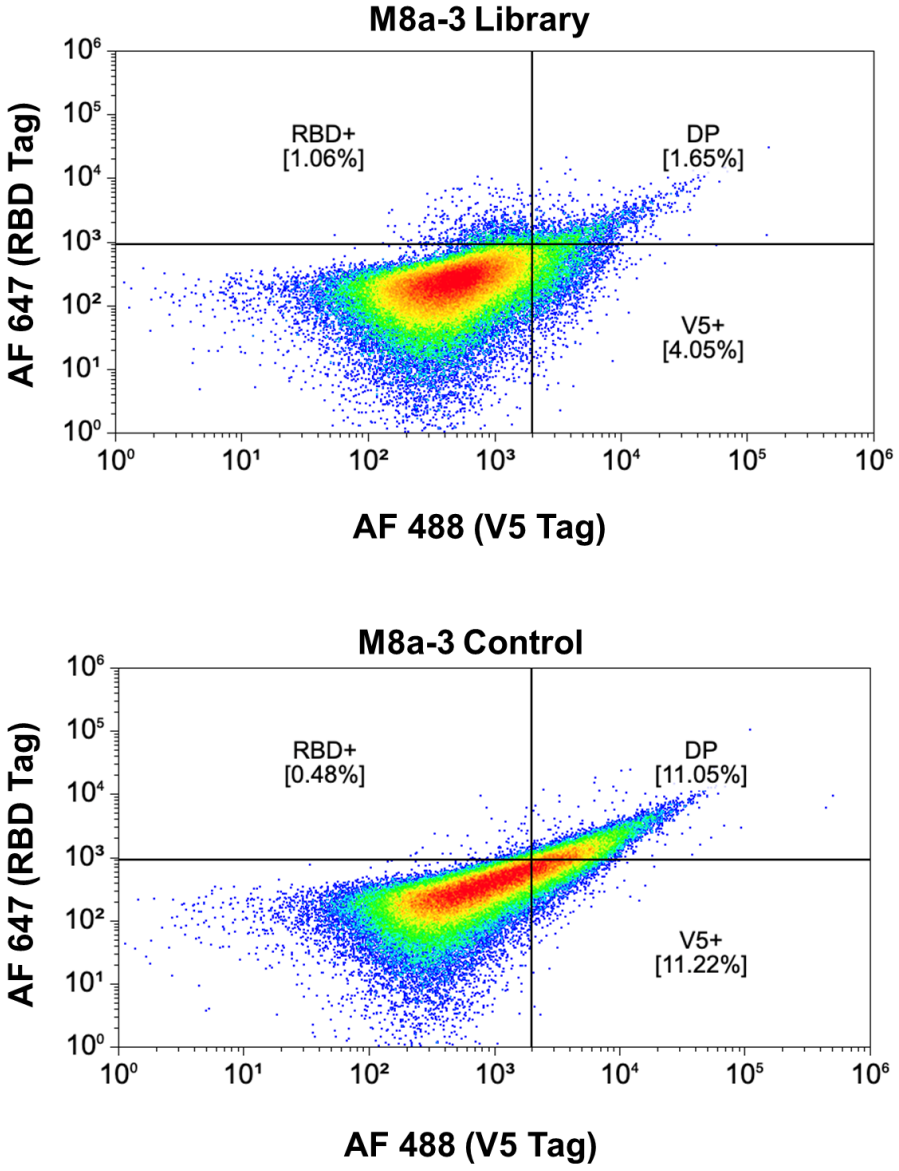


Figure 2.14: Yeast display results of the Mousify M8a-3 library. (**Top**) Cell sorting results of the Mousify M8a-3 library. In the cell sorting run, 5.7% of cells presented Fabs on their surface above background, with 1.65% of cells also binding RBD above background. (**Bottom**) Cell sorting results of the original M8a-3 antibody. Due to inefficiencies in the switchable system, only 11.05% of cells were double-positive above background.

As of writing, the sorted cells are growing at 30°C without OmeY to obtain enough cells for sequencing and are expressing the Fabs. Once we obtain sufficient amounts of Fabs, we will validate the sorted cells via Thermofluor, ELISA and the polyreactivity assay outlined in the previous section.

We have developed a generalized method to transform any antibody humanization model into a model that can generate a library of humanized antibodies, an approach can vastly improve the chances of a positive outcome for antibody humanization compared to the low chances estimated from our benchmarking experiments (Figure SC.21-23). This was achieved by abstracting the way ML-based antibody humanization models work and incorporating them into the Mousify framework. Mousify is then able to run a Markov chain over the sequence space of humanized antibody sequences defined by a discriminator and a map module. Due to the high cost of ordering an oligo pool of all generated library members, we sampled a sub-library of ~10,000 sequences that could be ordered at a reduced cost. If cost is not an issue, Mousify can be run without the constraints that aid in generating sub-libraries and could potentially result in a higher quality library. We speculate that the method outlined in this section would result in multiple humanized and functional antibodies.

Antibody Humanization using AbVAE

The results from the humanized antibody benchmarks have demonstrated that the best performing discriminator model is not ML-based, but that adding a ML-based map module does help in generating better antibodies (Figure 2.6-7, Figure SC.21-23. Our main criticism behind the discriminator model developed for Hu-mAb was that the model was not trained to generate mAb sequences, only to classify them. This makes a model great at distinguishing sequences similar to the ones in the dataset, but when asked to generate sequences of a certain class, such a model can quickly overextend its capabilities. We were interested in developing a model that is on-par in score to ADA correlation with previously published classifiers, but that also uses sequence generation as an essential part of training.

We decided that a variational autoencoder model (VAE) would best fit our needs to develop a new model for Mousify. A VAE is an unsupervised learning model first presented in 2013 by Diederik Kingma and Max Welling for the estimation of intractable posteriors in variational Bayes.[77] In order to understand VAE's, we have to give a short introduction on autoencoders (AE). AEs are models that take a datapoint from a dataset as an input of an encoder and the task is to compress the data to a latent space, i.e. an N-dimensional vector space, where N is known as the latent dimension. This vector is then passed on to a decoder to recover the information of the datapoint from that compressed state (Figure 2.15A). Fundamentally, it is a way to compress the information in any arbitrary dataset and be able to recover the data within a certain margin of error. The problem with AEs is that the latent space created is "unstructured". For example, we could hypothetically train an AE with images of animals and look at the latent vector produced by the image of a dog. If we then sample vectors very close-by to the latent vector of the image of a dog, we would very likely not get an image of a dog back, but rather an image that is mostly noise. In fact, images of different dogs would likely not be represented by vectors in proximity to each other, that is what we mean by "unstructured". The VAE creates a structured latent space. In close proximity to images of a dog, we will find other images of dogs and even images of dogs that were not in the dataset that the VAE generated. VAEs creates this structure in the latent space by adding two new properties to the AE (Figure 2.15B). (1) Instead of passing the calculated latent space vector, $\boldsymbol{v}$, to be decoded, we sample a new vector using the formula $z = \boldsymbol{v} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The addition of the vector ε sampled from a normal distribution ensures that the model is not only capable of reconstructing $\boldsymbol{v}$, but all vectors in proximity of $\boldsymbol{v}$ as the same datapoint. (2) VAEs use two loss functions (See Explanation of Common (ML) Software Engineering Terms), a reconstruction loss and a Kullback-Leibler divergence (KL divergence) loss. The reconstruction loss tells the model how accurately it reconstructed the datapoint and the KL divergence loss tells the model how close the sampled datapoint is to a normal distribution (See **Error! Reference source not found.**).[78] The seminal paper by Kingma and Welling was the beginning of a large wave of generative models from image generation to 3D object rendering and more.[79–82]

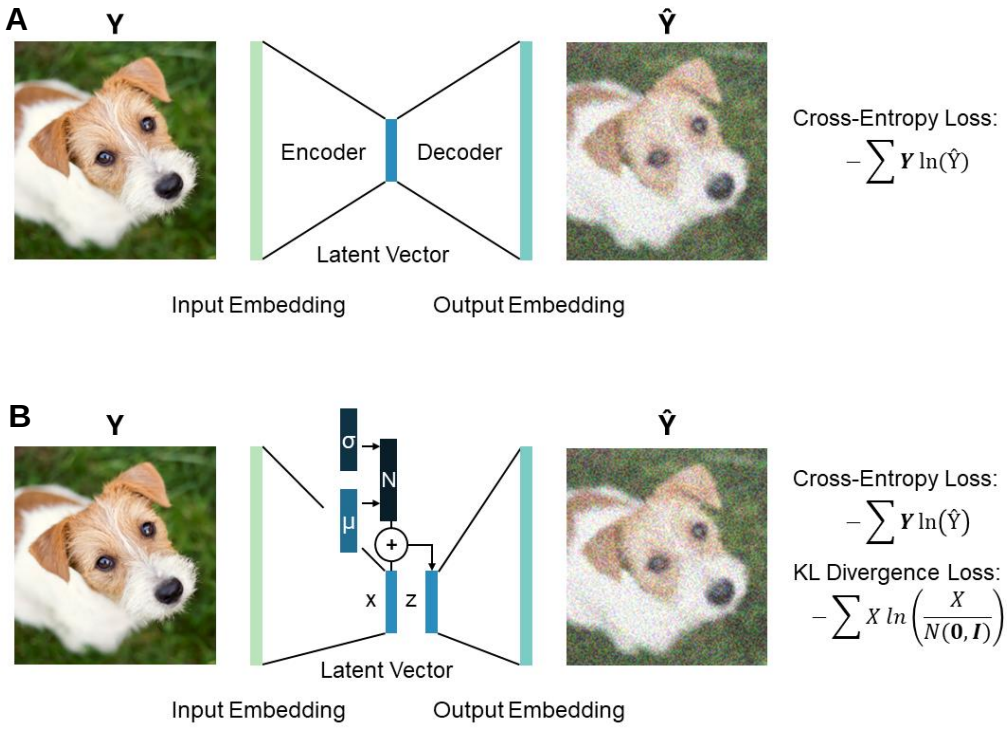


Figure 2.15: Abstract structure and functioning of an Autoencoder (AE) and a Variational Autoencoder (VAE). (**A**) An AE encoder takes a datapoint, in this case an image of a dog, and maps it to a latent space. Usually, the dimension of the latent vector is smaller than the dimension of the input embedding. Then the AE decoder decompresses the vector from the latent space and reconstructs the image of the dog. The performance of the process is judged by the cross-entropy loss between the input image **Y** and the reconstructed image **Ŷ**. (**B**) Similarly, the encoder of a VAE compresses the input embedding to a latent vector. However, then the latent vector **x** is added to a sample from a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The new latent vector **z** is decompressed by the decoder to reconstruct the image. The performance is calculated by a linear combination of the cross-entropy loss and the KL divergence loss.

VAEs have had a large amount of success in protein ML as well,[83–86] with models mainly being used for the generation of functional protein sequences. One notable recent example comes from Ziegler *et al.*, where the authors used an MSA input for a direct coupling analysis (DCA) model[87] and a VAE. Together, the latent information of the VAE and the DCA were used to create a new latent space which captures phylogenetic, function and fitness information of proteins.[85] Another notable example from Lyu *et al.*,[88] where novel AAVs were generated from a small 711-sequence dataset, with the help of multihead self-attention models in the

decoder. The latter's model structure formed the basis of AbVAE (Figure 2.16).

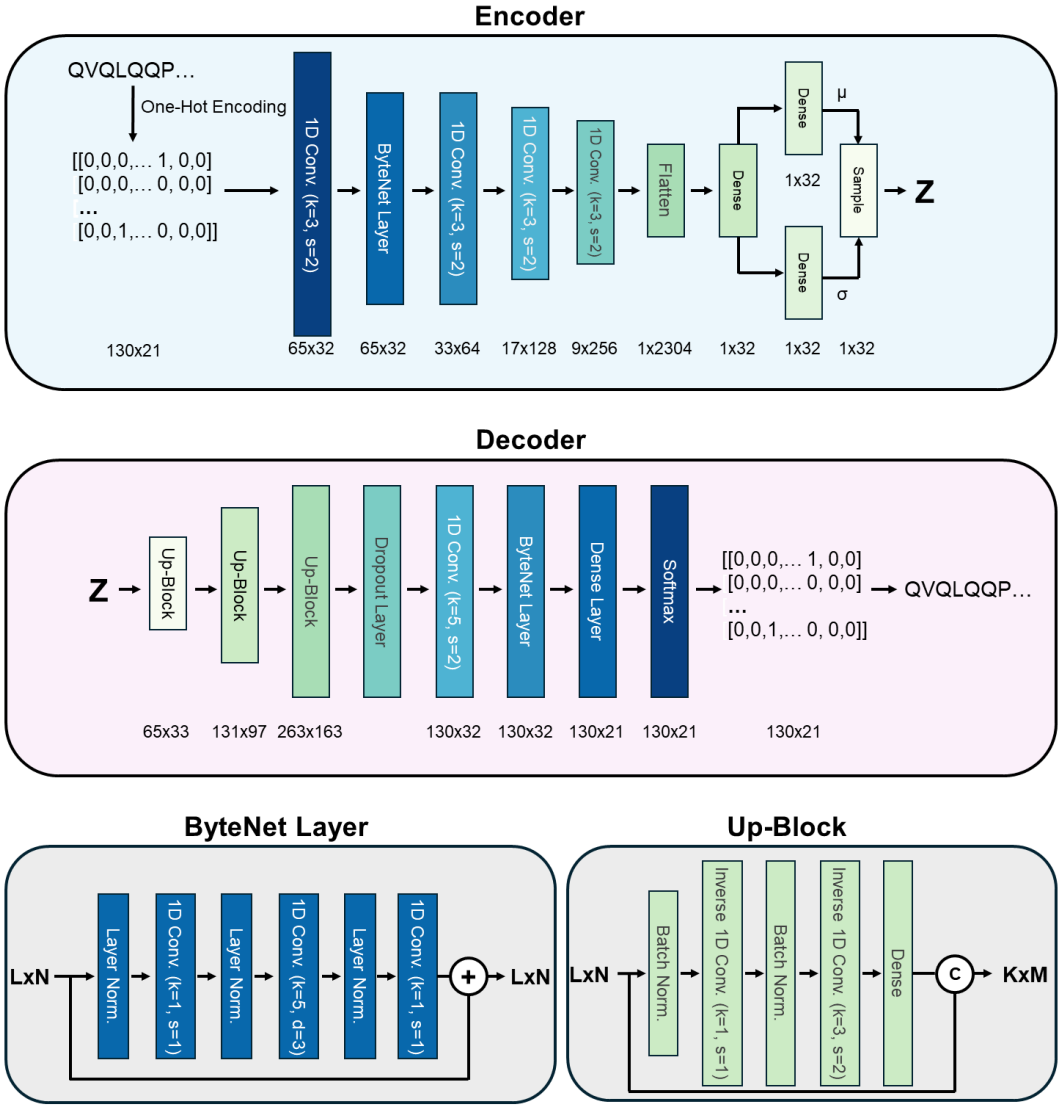


Figure 2.16: AbVAE model structure. Protein sequences are embedded as a one hot encoded matrix before being passed to the encoder. The 130x21 matrix is downsampled using a series of 1D convolutions with kernel size 3, a stride of 2, and "ELU" as the activation function. The latter 1D convolutions have the same parameters except that the filter size doubles each time. To extract global sequence context, we implemented a ByteNet layer, which consists of a series of 1D convolutions and layer normalizations, with one of the convolutions using a higher kernel and dilation size (k=5, d=3), with a "GeLU" activation function. The final step in the ByteNet layer is to add the input to the layer to the output. Once we obtain the sample from the latent representation **Z**, we pass the vector to the decoder which consists of a series of Up-Block layers. An Up-Block layer first performs an inverse 1D convolution with a kernel size of 1 and a stride of 1, which conserves the shape of the input. This is followed by an inverse 1D convolution with a kernel size of 3 and a stride of 2, which upsamples the data. To maintain the gradient, the input of the Up-Block is concatenated to the output.

Sequences for training were obtained from OAS via OAS C/S, using the following post-processing modules: Remove Redundant Sequences, Sequence Length Filter (Max. length 130), Non-Canonical Characters Filter, Combine Files, Antibody Viability Filter (Filter: "loose"). AbVAE was trained on 592,259 sequences with a training/test/validation split of 8:1:1 using the OAS C/S module Data Maker, unless specified otherwise. Each sequence was pre-encoded using either a one-hot encoding scheme or using ESM2 model t6_8M_UR50D. The model was trained by minimizing the objective function using a batch size of 10 and 300 epochs:

$$L_t(\boldsymbol{S}, \boldsymbol{X}) = -\sum \boldsymbol{S} ln(\hat{\mathrm{S}}) - \beta_t \sum \boldsymbol{X}\, ln\left(\frac{\boldsymbol{X}}{\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})}\right)$$

Where **S** is a batch of sequences, $\hat{\mathbf{S}}$ is the reconstructed batch of sequences, **X** is the latent representation of the batch and β_t is the value of the PID controller algorithm at time point t, defined by the function:

$$\beta_t = Max\left(K_p\, e(t) + K_i \sum_{\tau=0}^{t} e(\tau) + K_d\, \boldsymbol{\Delta} e(t) + \beta_{min}, \beta_{max}\right)$$

Where $e(t) = KL_{desired} - KL(\boldsymbol{X}_t, \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}))$, is the difference between the desired KL divergence value and the KL divergence value at time point t, and the discrete derivative $\Delta e(t) = e(t) - e(t\text{-}1)$.

The implementation of a PID algorithm to control the contribution of the KL divergence error is necessary to prevent KL vanishing during training, which creates very poor reconstructions of the data (Figure SC.26). A PID controller algorithm applied to VAEs was first reported by Shao *et al*. to dynamically control the contribution of KL divergence to training and achieve a desired KL divergence value at the end of training. This is in contrast to previous strategies that set β as a constant, which allows no control over the final KL divergence, or to train a separate model to train β, which is computationally expensive. Unless specified otherwise, we used the following parameters for the PID algorithm:

$\beta_{max} = 1$, $\beta_{min} = 10^{-5}$, $K_p = 10^{-2}$, $K_i = 10^{-4}$, $K_d = 10^{-3}$ and $KL_{desired} = 0.25$

The model was trained using the Adam optimizer with a learning rate $5 \cdot 10^{-4}$ and a weight decay of 10^{-4}. The learning rate was reduced by a factor of 10 when $\Delta L_{reconstruction} < 0.01$.

We evaluated models via four performance metrics: learning curve, reconstruction errors, classification performance, and correlation to %ADA. We obtained the learning curves from tensorboard by using the "Callbacks" tools from Tensorflow (Figure 2.17A, Figure SC.27-28).[89] The reconstruction errors were evaluated by sampling 25 therapeutic antibodies and plotting the distribution of Hamming

distances, as well as BLOSUM62 distances of the original sequences to the reconstructed sequences (Figure 2.17B). The inclusion of both the Hamming distance and the BLOSUM62 distance allows us to evaluate the difference between the exact reconstruction and a similar reconstruction. AbVAE is decent at reconstructing antibody sequences with reconstructions being on average 70% accurate, evaluated with BLOSUM62. However, AbVAE does not need to be perfectly accurate on the entire variable region of an antibody, since during antibody humanization the CDR remains conserved. AbVAEs accuracy on framework reconstructions is 76%, while only 44% accurate on CDR reconstructions (Figure 2.17B).

As an unsupervised model, we were curious if the model learned any differences between species that could help us in creating better humanized antibody sequences. This was evaluated by first obtaining the latent representation of test set sequences and fitting a principal component analysis (PCA) model with between 2- and 32-component dimensions (Figure 2.17C). For each species in the test set, we calculated the parameters of a normal distribution within every PCA model representation of the latent space. We then calculated the Mahalanobis distance of each point in the dataset to the center of each species' normal distribution. The Mahalanobis distance of a point **x** to a multivariate normal distribution $\mathcal{N}(\mu, S)$ is defined as:

$$d_M(x, \mathcal{N}) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)}$$

Where μ is the mean and S the covariance matrix of the distribution. Then we calculated the probability of each latent vector belonging to each species' normal distribution by using the fact that the Mahalanobis distance d_M is chi-squared (χ^2) distributed:

$$p_{species}\left(x, \mathcal{N}_{species}\right) = 1 - \chi_D^2\left(d_M(x, \mathcal{N})\right)$$

Where D is the latent dimension size. Lastly, the humanness score is then calculated by using Bayes theorem:

$$P_{human}(x) = \frac{p_{human}(x, \mathcal{N}_{human})}{\sum_{species} p_{species}\left(x, \mathcal{N}_{species}\right)}$$

We used this score to calculate the classification score (AUROC) for each PCA model (Figure SC.1). AbVAE outperforms all other antibody humanization models in the classification task of human vs non-human (Table SC.1). However, as stated above, classification tasks are not an important metric for antibody humanization. Therefore, we calculated the Pearson's correlation coefficient to the %ADA in patients of antibody sequences from the therapeutic antibody sequence dataset (Figure 2.17D). For this, the ideal number of PCA dimensions was 22, compared to 4 dimensions for the classification task. Indicating that %ADA can only be modelled with more complex data structures. It is worth noting that modelling each species as

a single multivariate normal distribution might not be the ideal way to generate a humanness score. We have contemplated using a linear combination of multivariate normal distributions to model each species, but we are not sure how to implement a distance measurement to the center of a linear combination of normal distributions.

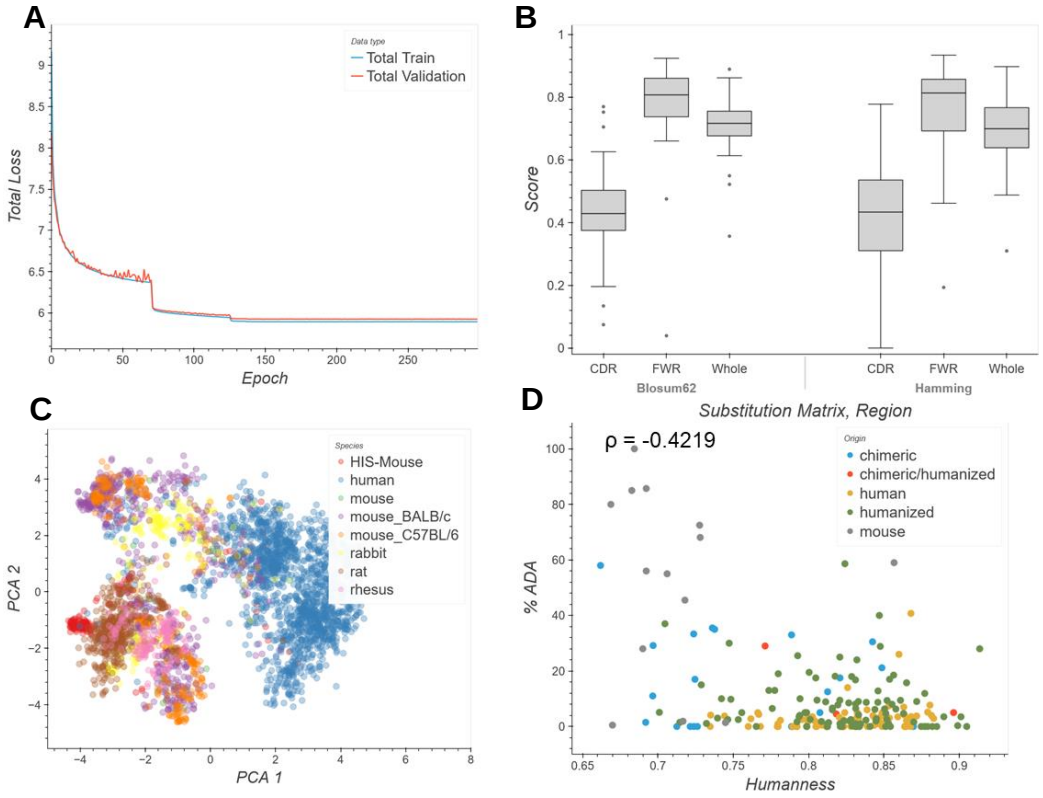


Figure 2.17: Performance metrics of AbVAE. (**A**) Learning curve of the total loss per epoch. Large jumps in the loss correspond to changes in the learning rate. (**B**) Reconstruction performance of the model calculated via alignment of the input sequence with the reconstructed sequence. Since the model can learn that similar amino acids are a valid substitution, we calculated the reconstruction performance with a Hamming distance and a BLOSUM62 substitution matrix. (**C**) Plot of the latent dimension, using a PCA model with 2 components, colored by species. Although separation of species is visible in this 2D-plot, in higher dimensions these clusters can be connected, or closer to each other than they may seem. (**D**) Correlation of "humanness" score to %ADA, using a PCA model with 4 components. The Pearson's correlation coefficient has a p-value of $9.8 \cdot 10^{-11}$.

We attempted to improve the AbVAE model by using a model that makes use of ByteNet layers. ByteNet was developed as a model to obtain long-distance correlation in convolutional neural networks.[90,91] This AbVAE model (AbVAE-ByteNet) performed similarly on the classification task, with an AUROC score of 0.982 versus 0.981 for AbVAE. Additionally, this model fared better in the sequence reconstruction task with an average reconstruction accuracy of only 74% (SI FIGURE), with most of the improvement seen in framework reconstruction (82%). However, the correlation of humanness to %ADA was worse as well, with a

Pearson's correlation coefficient ρ = -0.33 (Figure 2.18). Interestingly, the best correlation coefficient was found using 22 PCA components. At such high dimensions, the curse of dimensionality becomes apparent with all scores lying between 0.998 and 1 (Figure 2.18D).

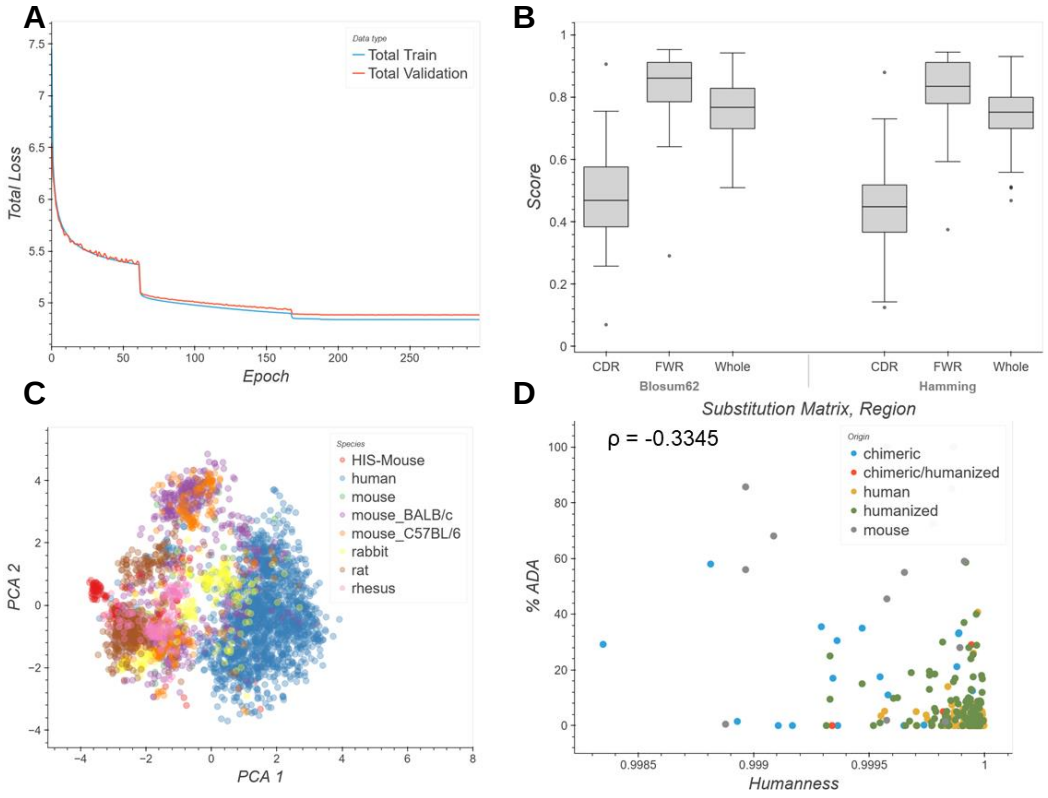


Figure 2.18: Performance of AbVAE-ByteNet. While performing slightly better in reconstruction and in classification, it performed worse in correlating humanness score to %ADA. (**A**) Learning curve of the total loss per epoch. (**B**) Reconstruction performance of the model calculated via alignment of the input sequence with the reconstructed sequence. (**C**) Plot of the latent dimension, using a PCA model with 2 components, colored by species. (**D**) Correlation of "humanness" score to %ADA, using a PCA model with 22 components. The Pearson's correlation coefficient has a p-value of $4.7 \cdot 10^{-4}$.

This poor performance in the key metric of %ADA correlation spurred us to explore more informative embeddings to improve the model performance. Due to the wide success of ESM in providing more informative protein embeddings,[36,41,92] we first explored a model trained on ESM embeddings. This new model resulted in a worse performance in both reconstruction and correlation, with the correlation coefficient being particularly poor (Figure SC.31-32).

While not outperforming other antibody humanization models in terms of humanness score correlation to %ADA, we were also interested in developing a model that

understands the structure of antibody sequences in order to improve the success rate of humanized antibody sequences. Additionally, AbVAE forms a good pre-trained model, on top of which, more specific models can be trained. One example would be to add a regression task to the pre-trained model and directly predicting %ADA or a sequence by using therapeutic antibody data and their %ADA as the dataset. While this dataset is small (217 sequences), the pre-trained model could add a lot of sequence context, making training such a model feasible (Figure 2.19A).

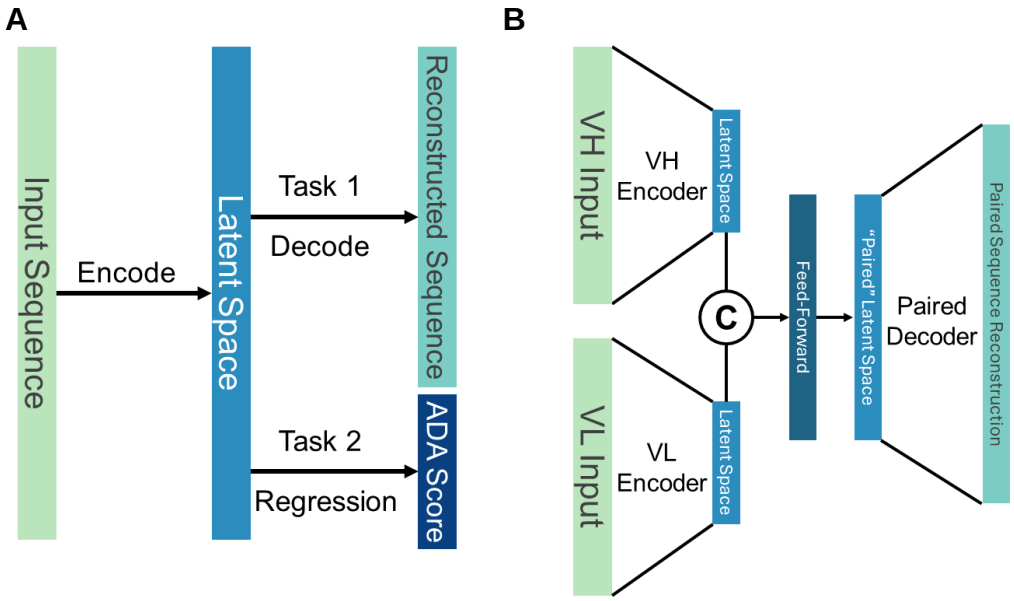


Figure 2.19: Proposed future models making use of the pre-trained AbVAE model presented in this thesis. (**A**) Proposed model to improve humanness correlation to %ADA. In this model we would use AbVAE to perform transfer learning on the ADA score prediction task. (**B**) Proposed model to predict scores and reconstruct sequences based on the information of paired antibody sequences.

Some antibody humanization models independently predict scores for the light chain and the heavy chain of an antibody.[20,21] This is due to the report in a previous publication that claims that VH and VL behave independently when it comes to humanization scores, however in the paper this claim is followed by the statement "data not shown".[9] Even if the two chains are independent in terms of humanization scores, they certainly are not in terms of compatibility when forming a complex.[93] We propose to expand the structure of AbVAE to account for paired antibody chain pairing during training. Due to less data availability on OAS for paired antibody sequences, one could use the light chain and heavy chain pre-trained AbVAE model encoders and concatenate their latent dimension outputs as the input for a paired antibody sequence VAE (Figure 2.19B).

Lastly, we want to discuss the potential new avenues that AbVAE could present for antibody humanization. While so far mostly presented for its potential as a discriminator module, one can imagine AbVAE as a map module as well. In that case, a Markov chain would explore the latent space of AbVAE, aiming to minimize

the Mahalanobis distance to human sequences. The transition probability matrix at any point in the latent space is then extracted at the softmax step in the VAE (Figure 2.16), which represents the probability distribution of a sequence at that coordinate in the latent space.

Fast Structure Calculations for Mousify Libraries

With our focus in Mousify to perform library generations from any machine learning model, we were wondering if the performance of large 10^5-10^6 libraries could be improved by increasing the likelihood of library members producing valid and stable antibodies. It is important that this method is very fast. Consider if it only takes one second per antibody in a 10^5-size library, it will add 27 CPU-hours and in a 10^6-size library it will add 11 CPU-days to the workflow. We envisioned two methods to tie this validation into the Mousify framework, depending on the time it takes to validate each generated antibody sequence. If the structure/thermostability validation works on the order of $\leq 10^0$ seconds, the validation could happen on each worker as soon as it accepts a sequence in the Markov Chain. If the validation step takes longer than is reasonable to perform on each sequence of the library (i.e. $\geq 10^1$ seconds), we would implement an "evolutionary bottleneck" algorithm. In this case, it is useful to think of the individual chains of the Markov Chain as different evolutionary lineages. One concurrent process in Mousify would be responsible to perform the validation on the evolutionary lineage, instead of the individual antibody sequences. Once warmup is complete, the validation process would occasionally probe the predicted structure/thermostability of the last sequence in the lineage. If the chain passes the test, the chain continues. If it fails, then the chain is forced to remove all sequences from its registry until it gets to a point in its lineage where a sequence has passed the test and starts over. This method is far more flexible for validation methods that take up more compute resources as one can now decide how often the validator probes evolutionary lineages depending on how fast it can process single sequences. In this part of the Mousify project, I describe the work performed together with Jessie Gan, to investigate different protein structure prediction methods on antibodies. We decided to study the structural accuracy and computational time of our Thermosurf (TS) method against homology modeling and machine learning-based methods. Developing a fast and reliable method for computing antibody structures will enable us to screen proposed humanized sequences for thermostability and may improve the quality of predicted antibody libraries.

In recent years, in silico methods of structure prediction have dramatically increased in accuracy and viability to investigate protein properties in the absence of experimental structures. Notably, AlphaFold2 (AF2), a machine learning approach to structure prediction, has distinguished itself as a near atomic resolution predictor given ideal conditions.[12,44] However, to structurally validate in silico protein libraries, AF2 is not computationally viable for library sizes of 10^5-10^6. In the past, many structure prediction methods involved searching for the nearest homologous sequence for which an experimental structure exists and using it as a template for homology modelling.[94] The template searching step, calculating a multiple sequence alignment (MSA), can be computationally expensive, while the structure prediction step is low-cost. This method may be efficient but is inaccurate when the available homologs are evolutionarily too distant from the query. The advantage in the case of mAb engineering is that the set of homologous sequences to consider is limited to

known antibody structures, vastly reducing the computational cost. We aim to accelerate the calculation of reliable antibody structures by combining AF2 with the classic homology modeling methods for unknown, but closely related, structures. Our hypothesis is that homology modeling based on an AF2 template of a highly related structure is very similar to an AlphaFold2 structure of the query sequence and is orders of magnitudes faster while being indiscernible in thermostability calculations.

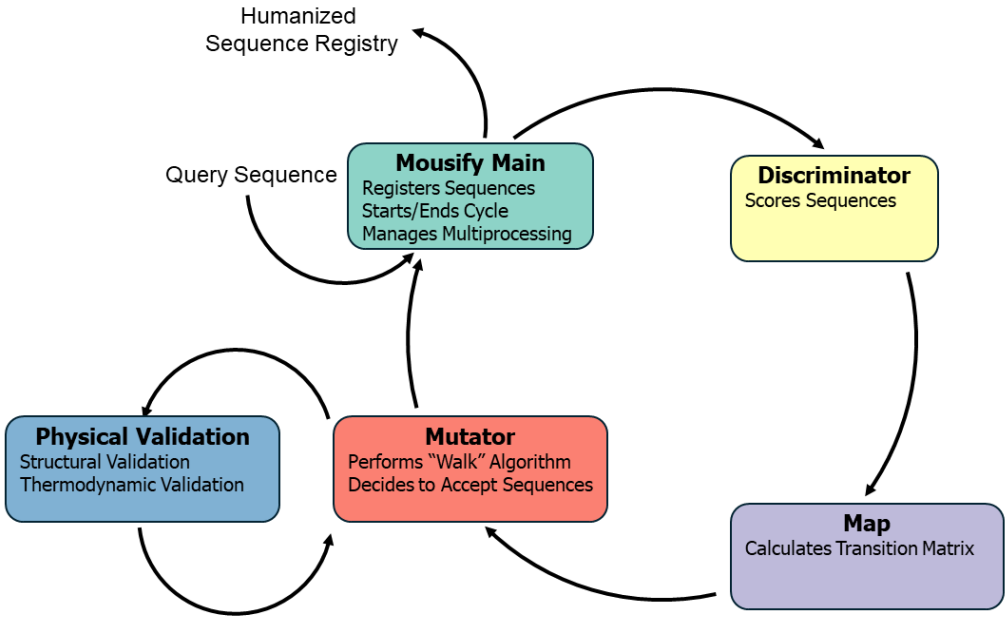


Figure 2.20: Potential implementation of a physical sequence validation model into the Mousify model architecture. The physical validation could either act as a second filter to the acceptance step of the mutator, or the score of the physical validation could be implemented into the mutators algorithm as a linear combination of the discriminator score and the physical score.

To investigate possible structure prediction methods, we first considered the performance of homology modeling (HM), the standard of protein structure prediction before machine learning based models, such as AlphaFold2 (AF)[44] and ESMFold (ESM).[36] First, we ran HM structure predictions, in which an MSA against sequences of an antibody structure database determines the template structure for a given query sequence based on the highest sequence similarity. In HM prediction, the major bottleneck in calculation speed originates from the MSA. We calculated the root mean squared deviation (RMSD) of a predicted structure's backbone atoms to the experimentally determined crystal structure (Xtal) from the Structural Antibody Database (SAbDab).[17] HM structure prediction performs with a median RMSD of 2.13 Å, compared to the median RMSD of AF of 1.41 Å (Figure 2.21A). To compare the relative performance between different models beyond the distribution of RMSDs, we calculated the probabilities of each method outperforming another (Figure 2.21C, Fig S4, Fig S5) .We hypothesized that to

improve on the speed of AF, but retain its accuracy, we developed a homology modeling-based method that uses AF calculated structures as templates and uses a classifier to rapidly find the homology modelling template (Supplemental 1). We started by clustering all paired sequences in the Observed Antibody Space (OAS) database using MMseqs2 down to 421 cluster representatives.[95] We then calculated AF structures from those representative sequences to provide us with templates. To determine the best template for a given structure, we trained simple classifiers to identify the best templates for a given query sequence. However, these classifiers were not able to capture the complexity of our dataset and could not effectively pair query sequences to their best template (Figure SC.33). For the duration of the project, we implemented a method based on Clustal Omega to match query sequences to their closest template sequence by sequence similarity.[96]

We developed a random (RD) method to set a baseline benchmark against TS, where a random antibody structure from SAbDab was used as the template for any given sequence. We were interested in the fitness of a random crystal structure against an AF structure for any given sequence. The least accurate method was TS with a median RMSD of 2.76 Å (Figure 2.21A). While the RD method has a median RMSD of 2.32 Å. While some predictions of RD are comparable to the HM method, in terms of probability that one is better than the other, RD only performs better than HM 27% of the time (Figure 2.21C).

To explore the qualities of high-performing templates in homology modeling methods, we further investigated template sequence similarity compared to method accuracy in homology-based methods. We observed that the homology modeling based methods demonstrated an inverse correlation between RMSD and percent identity to the sequence of the reference structure (Figure 2.21B). We also observed that the highest performing method, HM, also had most of its templates between 50-100% sequence similarity, compared to TS which cannot offer templates over 60% sequence similarity as a consequence of the sequence clustering threshold. We sought to test machine learning-based models against the homology model-based methods. As shown in figure 1, we observed that the machine learning-based models, AF, and ESM, had lower median RMSDs compared to homology modeling based methods, HM, RD, and TS (Fig. 1A). The lowest median RMSD was 1.41 Å for AF, making it the most accurate method while ESM performed only slightly worse at 1.47 Å. From our probabilistic analysis to determine how often one model outperforms the other, it became apparent that AF was consistently the most accurate method, while TS was the lowest performing method.

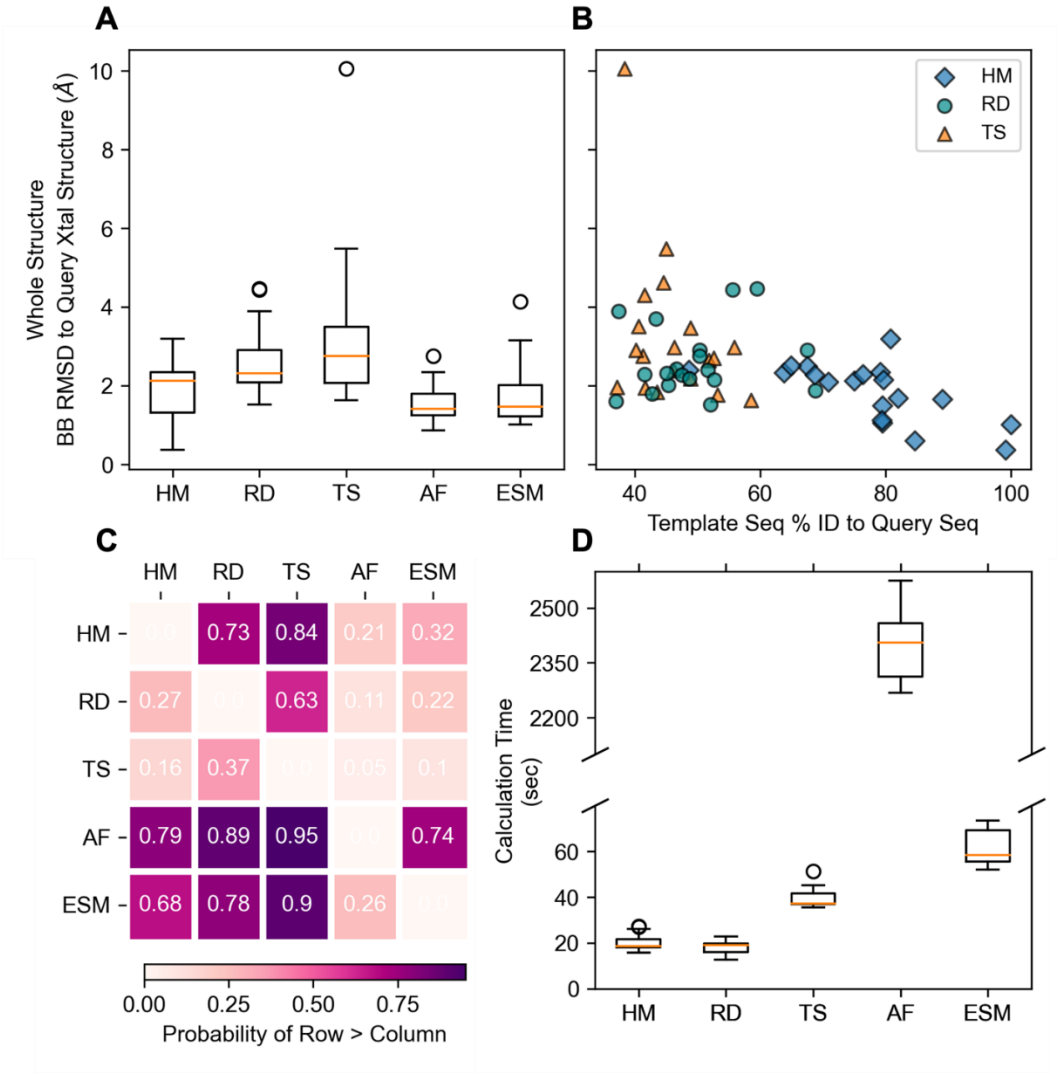


Figure 2.21: Performance of different antibody structure prediction models. (**A**) Boxplot summary of the RMSD of the entire backbone (BB) atoms of the predicted structure to the original query crystal (Xtal) structure for each method, with a sample size of 19 query sequences. (**B**) RMSD of the entire backbone (BB) atoms of the predicted structure to the original query crystal (Xtal) structure vs. template sequence percent identity to the query sequence. (**C**) Probability map of every method, showing the probability of one method being more accurate than the other. (**D**) Boxplot summary of the timing benchmarks performed for each viable method. Due to random sampling the sample sizes for each method differed over the 100 calculations. HM: Homology Modelling; RD: Random Sampling; TS: Thermosurf; AF: Alphafold2; ESM: ESMFold; Sample sizes per method is as follows: HM:27; TS:25; AF:26; ESM;22.

Since predicting structures in a reasonable timeframe is important to determine the type of implementation in Mousify, we are not only interested in accuracy, but also time. Therefore, we performed timing benchmarks to compare their structure prediction time. We found HM to be the fastest method, with a median time of 13.50 seconds in real time, and AF to be the slowest with 2250 seconds, or approximately

37 minutes. TS was the second fastest method with a median time of 38.09 seconds, and ESM was next with a median time of 65.28 seconds (Figure 2.21D).

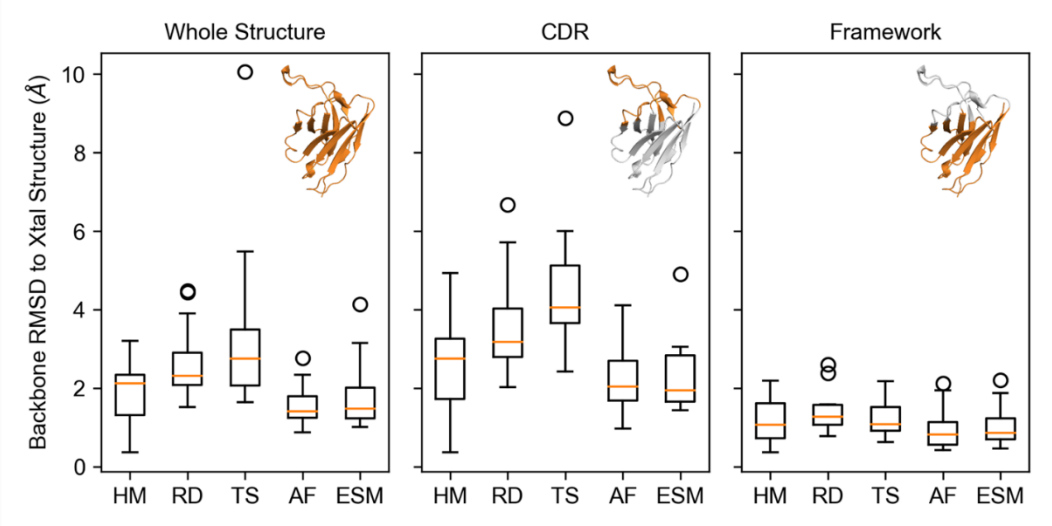


Figure 2.22: Model performance by section of the variable region of an antibody. CDR regions are more difficult to predict, while framework regions are predicted with similar accuracy across all methods. Method RMSDs of predicted structure backbone atoms to query crystal structures are shown in boxplots. Antibody structure shown for illustrative purposes only. PDB: 7TRH.

Since some antibody engineering methods can be more interested in the accuracy of the complementarity-determining region (CDR) or the framework region, we investigated the structural accuracy of predictions of all methods on different sections of the antibody (Figure 2.22). We computed RMSDs of the antibody CDR and the framework. The CDR region is of great interest due to its essential role in antibody function when interacting with the targeted antigen. We observed that on average across all methods the CDR regions have less accuracy in structure prediction than the whole structure, showing how the variability in known CDRs affects the reliability of all methods. We also observed that the methods perform similarly on the framework regions.

Considering the dependence of these structure prediction models on current dataset, we investigated potential areas for bias in the tested RD and TS methods by exploring our template database. The template database of crystal structures from SAbDab contained metadata about antigen species and antibody types. We observed that Homo Sapiens was the most common type of antigen species, as well as antibody origin species (Figure 2.23). We also observed many deposited antibodies that were specific to Sars-CoV-2 virus and HIV.

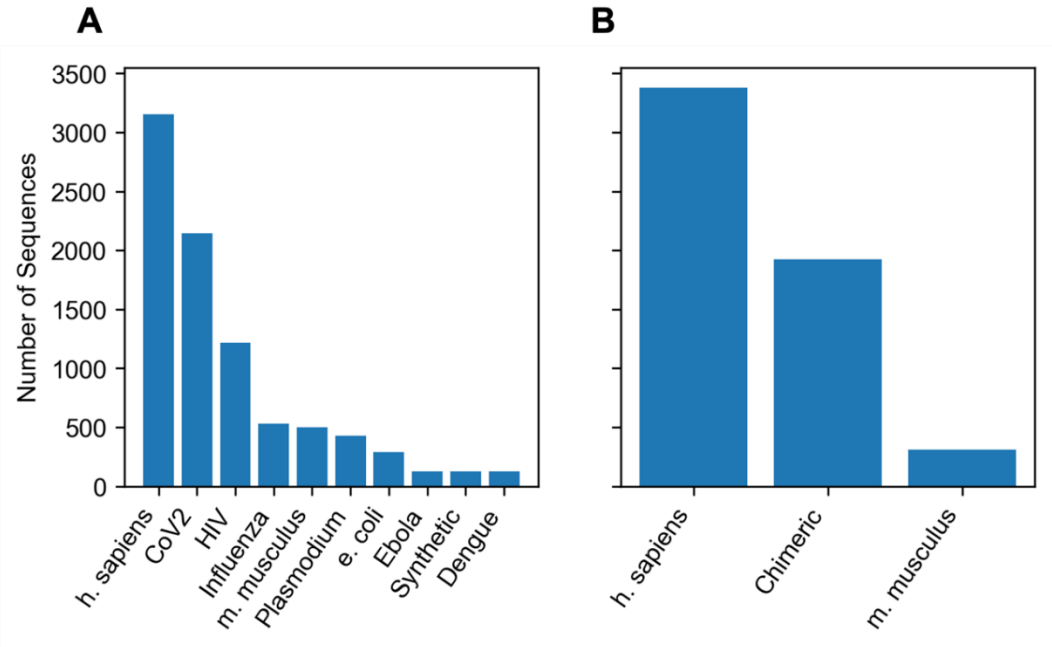


Figure 2.23: Bias in the antibody dataset used in this study, that needed to be accounted for during benchmarking. The template dataset for homology modeling is skewed towards commonly studied antibody targets and types. (**A**) Sequence distribution of antigen species which antibodies in the database target. Antigen species shown were the top 10 most common species. (**B**) Sequence distribution of antibody species of origin. Any antibody that was a fusion of more than one antibody species, for example, synthetic constructs, Homo Sapiens and Mus Musculus, or Homo Sapiens and Lama Glama, were considered "Chimeric".

To account for these biases in our dataset, we ensured our test sequences did not reflect the same skew towards commonly studied antibody types. We randomly sampled within each antigen species category such that our final test set was less likely to have sequences similar to the templates. As suggested in Figure 2.21B, high similarity between the test sequences and template sequences generally results in higher performances.

In this part of the Mousify project, we compared various structure prediction methods to validate large sets of antibody libraries in a high-throughput fashion. We were interested in a method that is not only as accurate as possible, but also fast enough to be able to validate millions of antibody sequences in an acceptable timeframe. Our findings suggest that homology modelling is the most viable method for large libraries due to its trade-off between speed and accuracy. If speed is not as important, or large amounts of compute capability is available, ESMFold achieves accuracy nearly on-par with Alphafold2 while finishing calculations two orders of magnitude faster. It is worth noting that in all methods except random sampling and ESMFold, the slow step is the calculation of an MSA. Combining that with the rather obvious observation that the higher the sequence identity of a template is to its query, the more accurate the homology modelling prediction. Therefore, it is not unreasonable

to imagine a deep learning method that could quickly pair a query sequence with a reference sequence from SAbDab and perform homology modelling to optimize the method even further. We also performed further analyses to investigate structural accuracy over different antibody components and biases in homology modeling-based methods. We found that the CDR regions were the most difficult to predict across all methods, but methods performed similarly across framework regions. We assume that CDRs are often the most important component to predict in antibody libraries, therefore it is important that this large discrepancy is considered when validating libraries.

Computational Details: OAS C/S

Main.py:

The user interacts with the main function of OAS C/S ("main.py"), which contains the argument parser implemented through python's argparse module. The parser expects that the query is defined either in the command line or through a query file (Figure SC.1). Based on the query provided, the program instantiates the following objects:

- DownloadOAS
- CSVReader
- FileManager
- List of PostProcessor

These objects are then passed to another object, instantiated from the API class, via dependency injection. The API class manages the communication between all modules by calling methods of the objects passed to it.

The first method called is "get_OAS_files" which first checks whether the query is a valid query and then tells DownloadOAS to download the files.

The next method called is "process_folder" which follows the factory pattern, which chooses the processing mode. The processing mode can either process each file with the CSVReader individually, in bulk, or split by file size. Individually means that each downloaded file is processed and then saved in a separate file. In bulk means that each downloaded file is processed and saved in a single large file. Splitting the processing by file size, processes each downloaded file and merges the results into a file until a set file size is achieved. Within this method, files are given a descriptive name by the FileManager to better keep track of files.

Then the API calls another factory pattern method to decide what to do with the raw files downloaded from OAS. By default, the files are deleted, but a user can choose to keep them or to move them to another folder.

Lastly, the API iterates through the list of postprocessors and executes them in the order provided.

Oasdownload.py:

This file contains the DownloadOAS class responsible for downloading files from OAS. An object instantiated from this class, requires a file ("OAS_files_dictionary_paired.json" or "OAS_files_dictionary_unpaired.json") that keeps track of which queries on OAS return which result in the web application. When a DownloadOAS object is called, it iterates through the download queries and

grabs every file name associated with it. At this step, complex user queries are also translated to their simpler equivalents. Then the make_union method, takes the union of all sets of file names of the individual queries. Next, the create_url method iterates through the file names associated with the full query and creates a list of URLs that can later be used to download the files. The list of URLs is packaged into a shell file that can be executed via Linux's "wget". The shell file is then executed to make a "wget" request to OAS and download the files using the download_files method. Lastly, the downloaded, compressed files are unpacked by the unpack method and given a descriptive name.

Csvreader.py

The CSVReader class contained in this program reads the files downloaded by DownloadOAS and processes them according to the user's specifications. The program opens each file and extracts the metadata stored in the downloaded files header as well as the desired column data. CSVReader stores all the data in python dictionaries. After processing a file, the dictionary is converted to a pandas dataframe. If the processing mode set in the API module is set to "bulk" or "split", the CSVReader will process more files to add to the dictionary. After meeting the conditions of the processing mode, the module will pass the dataframe to the Filemanager.

Filemanager.py

The FileManager class is responsible for saving and loading dataframes. The methods save_file, load_file can save and load files. We also implemented a python magic method to override the addition of Filemanager object to each other to return a concatenated dataframe. Another useful method in the Filemanager is the file_size method, which returns the size of a file. This is used in "split" mode to tell the API when to stop adding files to the CSVReader's dictionary.

Post_processing.py

This program defines the PostProcessor abstract class, defining the methods load_file, save_file and process.

Computational Details: Fast Structure Calculations

Structure Prediction Methods

All calculations were run on the Caltech Resnick High Performance Computing Center resources, including compute nodes with 32 16-core Intel Skylake CPUs, and GPU nodes with 4 Nvidia P100 GPUs each for AlphaFold2 (AF)[44] and ESMFold (ESM)[36] calculations. Dataset clustering was performed with MMseqs2.[95] Homology modelling was performed with OpenStructure[97] and ProMod3.[98] ProMod3 can superimpose the query sequence on a given template sequence and also construct missing loops and minimize structures for the ideal prediction. Machine learning classification was performed using scikit-learn python library[99] and XGBoost.[100] Structure analyses were performed with MDTraj.[101]

We tested five different models for accuracy and time. Out of five models, three were based on homology modeling. We first performed homology modeling where a template for a given query was determined by Clustal Omega via sequence similarity (HM). Templates were taken from the Structural Antibody Database, known as SAbDab,[17] which contains metadata and structures of all antibodies deposited in the PDB. We filtered this database for non-redundant antibody structures with both VH and VL being represented in the structure. We also performed a variant of this method where templates were randomly determined from the same dataset, instead of by sequence similarity (RD).

The TS method implemented a database of AF calculated templates for homology modeling. To determine the templates for structure prediction, we filtered the Observed Antibody Space database (OAS).[18] These were downloaded and processed using OAS API. We used redundancy removal and antibody viability post-processors to filter sequences. We limited our predictions to heavy chains. We next clustered this dataset by sequence similarity using the MMseqs2 software suite.[95] We then predicted the structure of cluster representatives using Alphafold2 using a model trained on data before June 2023.[44] Next, given a query sequence, we used Clustal Omega to identify the cluster representative that has the highest sequence similarity to the query sequence[96]. The cluster representative's Alphafold2 structure becomes the homology modelling template for the query sequence. We apply the homology modeling software ProMod3 to perform the homology modeling step.

We used AF version 2.2.0 installed on the Caltech HPC. We installed ESM from source, on pytorch version 1.12 with CUDA enabled on CUDA Toolkit 11.3. Apptainer environments for homology modelling, Thermosurf, and Alphafold2 are provided on this project's GitHub repository. All analysis scripts are also included for reproducibility.

Timing Benchmarks

We ran timing benchmarks to compare the speed of viable methods for structure prediction in thermostability calculations. We wrote a bash script that calls one of the structure prediction methods to test, at random, 120 times. This way each method is run on the same compute node with the same resources using SLURM. The random call to each structure prediction method is important to account for the variability in runtime running one function after another. The times were noted using the stdout of python which are then written to the SLURM output. Between each call of a structure prediction method, the environment was reset or the Apptainer was changed to the appropriate one. The output file was parsed after completion of all timing benchmarks via a python script.

Materials and Methods

Materials

Solvents and buffer salts were purchased from Sigma Aldrich, Koptec, Fisher Bioreagents and Merck. Media for bacterial expression were purchased from Merck Millipore. We used Invitrogen Mix&Go DH5α for plasmid expression and BL21-DE3 for protein expression. Reagents and media for mammalian cell expression were purchased from Gibco. We used Gibco Expi293F cells (donated by the Caltech Protein Expression Center) for all mammalian cell transfections. Expi293F cells were counted on a Roche Cedex® HiRes Analyzer. Expi293F cells were maintained in an Eppendorf CellXpert C170i. Taq DNA polymerase, Dpn1 and NEBuilder® HiFi DNA Assembly was obtained from New England BioLabs. Oligonucleotides were synthesized by Integrated DNA Technologies. Genes we synthesized by Integrated DNA Technologies and Twist Biosciences. DNA libraries were purchased from Twist Biosciences. PCR reactions were performed on an Eppendorf Mastercycler Gradient. Plasmids were purified using the Zymo Research Zyppy Miniprep Kit or the Macherey-Nagel Xtra MidiPlus kit. Agarose gel electrophoresis for DNA fragment purification was performed at 90V using TAE as buffer and 1% agarose gels (Merck Millipore). Gel excisions from agarose gel DNA electrophoresis were processed using the Zymoclean Gel DNA recovery kit. DNA gels were referenced against the Goldbio 1kb DNA ladder. Protein and DNA concentrations were measured with a Nanodrop ND-1000, for IgGs the Nanodrop was used in the IgG mode. Protein expression was verified via SDS-PAGE using the Invitrogen protein gel electrophoresis system. We used NuPAGE 4-12% Bis-Tris gels at 200V with MES as buffer and SeeBlue Plus2 prestained protein standard as the ladder. SDS-PAGE gels were stained using InstantBlue® Coomassie Protein Stain (Abcam). Proteins were purified using a Cytiva ÄKTA Start protein purification system, using either Cytiva MabSelect columns, Cytiva HisTrap HP columns and Cytiva HiTrapQ columns. Thermofluor assay was performed on a BioRad CFX 96 Real-Time PCR Detection System in BioRad Individual PCR Tubes 8-strip clear, using Biotium SYPRO® Orange as dye. ELISA assays were performed in Corning® Costar High Binding 96-well plates. Secondary antibody was purchased from Invitrogen. Development of ELISA plates were done with 1-Step Ultra TMB ELISA substrate solution (Thermo Scientific). Plates were analyzed on an Agilent Biotek Epoch 2 plate reader. Human recombinant IgE Fc (His-Tag) was purchased from Sino Biological. SARS-CoV-2 RBD (WA 1) was a gift from the Bjorkman lab (Caltech). Baculovirus prostate-specific membrane antigen (PSMA) was purchased from the Caltech Protein Expression Center.

Data Analysis and Visualization

All data was analyzed using python 3.10 or python 3.11 (depending on the Anaconda environment used). Two Anaconda environments were created for data analysis and data visualization, “Mousify” (python 3.10) and “Mousify-Holoviews” (python 3.11), the latter was created due to incompatibility of the newest version of

Holoviews with other packages used in the "Mousify" environment. Generally, we used Scipy, Scikit-Learn, Pandas and Numpy for data analysis and Holoviews with a Bokeh backend for data visuzalization.

Molecular Cloning

Genes encoding immunoglobulin G (IgG) VH and VL were ordered as gBlocks from Integrated DNA Technologies or as a gene fragment from Twist Biosciences. IgG CH1, Fc and CL_κ were amplified via polymerase chain reaction (PCR) from a p3BNC vector, also known as pAbVec2.1, containing an IgG heavy chain or an IgG light chain. PCR was performed with the following protocol: 50 μL total volume, 50 ng template DNA, 0.5 μM each primer, 0.2 mM dNTPs (NEB), 1 μL Taq Polymerase, NEB). Annealing temperatures, determined with Geneious software, were held for 15 s, while maintaining a ramp of 2 °C/s from the melting temperature (92 °C for 30 s) to the annealing temperature, followed by elongation and 30 cycles were run to obtain product. Fragments were separated via agarose gel electrophoresis and the relevant bands were extracted and purified with the Zymoclean Gel DNA Recovery Kit (Zymo Research Corp.). For heavy chain assembly, the genes were assembled using HiFi DNA Assembly Mix (NEB) into a p3BNC vector, together with a fragment containing IgG CH1 and IgG Fc. For light chain assembly, the genes were assembled using HiFi DNA Assembly Mix (NEB) into a p3BNC vector, together with IgG CL_κ. For the assembly of soluble TNF expression plasmid, the gene was ordered as a gBlock from Integrated DNA Technologies and subcloned into a pET-21a vector. Gibson assembly products were then mixed together with CutSmart Buffer (NEB) and diluted according to the NEB Dpn1 digest protocol, then 1µL Dpn1 restriction enzyme was added and incubated at 37°C for 30 minutes, followed by 15 minutes at 80°C. *E. coli* DH5α Mix&Go Competent Cells (Zymo Research Corp.) were transformed with the Dpn1 digested plasmids via the Mix&Go protocol. *E. coli* BL21(DE3) chemically competent cells were transformed with the Dpn1 digested Gibson assembly products via heat shock. An aliquot of SOC medium (750µL) was added, and the cells were incubated at 37°C and 220 rpm for 45 minutes. Both strains of transformed cells were plated on LB carbenicillin (LB_{Carb}, 100µg/mL) agar plates. Overnight cultures (5-mL LB_{Carb}) were grown at 37°C and 220rpm in culture tubes. Plasmids were subsequently isolated using the Zyppy Miniprep Kit (Zymo Research Corp.). The assembled plasmids were sequence verified via whole-plasmid sequencing by Primodium Labs.

Bacterial Protein Expression & Purification of Soluble TNF

His-tagged, soluble tumor necrosis factor (TNF) was expressed and purified as described previously.[102] Luria broth (LB, Merck Millipore) with 100µg/mL ampicillin in unbaffled Erlenmeyer flasks was inoculated 1% (v/v) with stationary-phase overnight cultures of *E. coli* BL21(DE3) cultures and shaken in an Innova 428 shaker at 180 rpm and 37 °C. At an optical density of 600 (OD_{600}) = 0.8, the cultures were chilled on ice for 20 minutes. Protein expression was induced with 1 mM

isopropyl β-d-1-thiogalactopyranoside (IPTG). The cultures were shaken at 180 rpm and 18°C overnight (18–24 hours). Cells were pelleted via centrifugation at 4000x*g* for 30 minutes at 4°C and the supernatant was discarded. The cell pellet was resuspended in Lysis buffer (20mM Tris pH 8.0, 300mM NaCl, 1mM DTT) and lysed via sonication on ice. The lysate was centrifuged for 1h at 4000xg at 4°C and applied to a HisTrap HP column (Cytiva) equilibrated with wash buffer (20mM Tris pH 8.0, 300mM NaCl, 1mM DTT, 50mM imidazole), using an Äkta Start. The column was washed with five column volumes of wash buffer. The protein was eluted with elution buffer (20mM Tris pH 8.0, 300mM NaCl, 1mM DTT, 400mM imidazole). The protein was then buffer exchanged into storage buffer (20mM Tris pH 8.0, 300mM NaCl, 1mM DTT) and concentrated using a Merck Amicon Ultra 10MWCO filter.

Mammalian Protein Expression & Purification

DNA for mammalian cell expression was obtained by growing *E. coli* DH5α cells in 50mL of LB_{Amp} and extracted using the Macherey-Nagel Xtra MidiPlus Kit. Expi293F cells were counted and only used if the viable cell count was between 3.5-$6 \cdot 10^6$ with a viability of ≥97%. Cells were diluted with Expi293 expression media, warmed to 37°C, to a cell count of $3 \cdot 10^6$ to total volumes of 10-, 25-, or 50mL. 50mL conical tubes (for 10- and 25mL transfections) or 125mL baffled Erlenmeyer flasks containing the cell suspensions were capped with AirOTop Sterile Flask Seals and shaken at 350rpm/125rpm respectively until the DNA complexation mixture was ready to be added. Transfections were performed using 1µg or DNA per mL of culture, with 1/3 of the DNA mass coming from the heavy chain plasmid and 2/3 of the DNA mass coming from the light chain plasmid. DNA was added to 500µL of OPTI-MEM in a 0.22µm Costar Spin-X centrifuge tube filter and spun at 21,000xg for 5 minutes. After sterile-filtering the DNA, the mixture was added to a volume of OPTI-MEM (See table below). Then Expifectamine was mixed with a volume of room temperature OPTI-MEM in a 15mL conical tube, gently inverted three times and incubated for three minutes. The DNA-OPTI-MEM mixture was added to the Expifectamine-OPTI-MEM mixture, gently inverted seven times and left to incubate to complete the complexation reaction. The cells were taken out of the incubator after 10 minutes have elapsed in the complexation reaction. The DNA was added to the cultures after a minimum of 11 minutes and a maximum of 15 minutes, resealed and protein was expressed for 96 hours at 37°C, 8% CO_2. After expression was completed, a 100µL was aliquoted and filtered through a 0.22µm filter and analyzed on SDS-PAGE to check for successful expression. If expression was successful, the culture was centrifuged at 3000xg for 30 minutes at 4°C and the supernatant was subsequently sterile filtered through a 0.22µm filter. The filtered supernatant was kept on ice. A Cytiva MabSelect column was equilibrated with wash buffer (20mM Sodium Phosphate pH7.2, 150mM NaCl), then the supernatant was loaded onto the column. The column was washed with 10 column volumes of wash buffer, then the protein was eluted with 5 column volumes of elution buffer (50mM Sodium Citrate pH3.2). To the eluted fraction was added 60µL of neutralization buffer (1M Tris

pH8.0) per mL of eluted protein volume. Antibodies were concentrated in a Merck Amicon Ultra 50MWCO filter.

Transfection Volume	DNA Volume	OPTI-MEM (DNA)*	Expifectamine	OPTI-MEM (Expifectamine)
10mL	10μL	500μL	28μL	500μL
25mL	25μL	1.5mL	80μL	1.4mL
50mL	50μL	3mL	160μL	2.8mL

*Note: If the DNA volume was more than the listed DNA volume, the amount of OPTI-MEM was reduced to account for the excess

Preparation of Electrocompetent Yeast Cells

Preparation modified from a previously described protocol.[103] Yeast cells were struck out on YPD plates (24g/L Agar, 20g/L bacto peptone, 10g/L yeast extract, 2% glucose) from a cryogenic stock of RJY100 cells and left to grow at 30°C for two days. A single colony was picked to inoculate 5mL of YPD (20g/L bacto peptone, 10g/L yeast extract, 2% glucose) in a culture tube and grown in a culture tube at 30°C and 220 rpm. The following day, the cells were passaged by adding 100μL of the previous days culture to a fresh aliquot of 5mL YPD in a culture tube. Cells were left to grow at 30°C and 220 rpm, overnight. The next day, the culture was diluted to an OD_{600} of 0.2 in 50mL YPD. Cells were grown until an OD_{600} of 1.5 was reached and subsequently pelleted at 2000xg for 5 minutes. The supernatant was discarded, and the cell pellet was resuspended in 25mL of sterile filtered 100mM lithium acetate by vortexing. Then 250μL of sterile filtered 1M DTT was added to the tube. The cap on the 50mL conical tube cap was loosened to ensure oxygenation and the cells were incubated at 30°C at 220 rpm for 10 minutes. After this step, the cells will always be kept on ice for the remainder of the preparation. The cells were pelleted at 2000xg for 5 minutes at 4°C and resuspended with 25mL of cold, sterile ddH_2O via vortexing. The cells were then again pelleted at 2000xg for 5 minutes at 4°C and resuspended with 10mL cold, sterile ddH_2O via vortexing. The electrocompetent cells were aliquoted to 0.5mL and immediately used.

Transformation of RJY100 Cells via Electroporation

250μL of electrocompetent cells were mixed with 10μL library DNA (30ng of library, 123ng of SP-GAL cassette, 133ng of pAB01 vector), 10μL of M8a-3 DNA (30ng of M8a-3 VH fragment, 123ng of SP-GAL cassette, 133ng of pAB01 vector) or 10μL of empty vector DNA (123ng of SP-GAL cassette, 133ng of pAB01 vector). Note: pAB01 vector is a derivative of pJC014 vector. Cells were then transferred to a prechilled 2mm electroporation cuvette using a pipette. Before electroporating the cells, the outside of the cuvette was wiped dry. The cuvette was placed in a BioRad Gene Pulser XCell and shocked using a square wave protocol with 500V, one 15ms pulse. Immediately afterwards, 1mL YPD was added to the electroporated cells and

mixed by pipetting up and down. The contents of the cuvette were added to a culture tube and incubated at 30°C without shaking for 1h. Aliquoted 10µL of the electroporated cells and diluted them to 1000x, 10'000x, 100'000x, 1'000'000x, 10'000'000x, 100'000'000x to determine the transformation efficiency. The diluted mixtures were plated on warm SDCAA plates (5.4g/L Na_2HPO_4, 8.56g/L $NaH_2PO_4 \cdot H_2O$, 182g/L sorbitol, 20g/L dextrose, 6.7g/L Difco yeast nitrogen base, 5g bacto casamino acids, 15g/L agar) and incubated for three days at 30°C.[104] The rest of the cells were pelted at 900xg for 5 minutes and resuspended in 5mL selection media and grown overnight at 30°C, 220rpm. The electroporated cells were passaged by diluting the overnight culture to OD_{600} of 1.0 and grown again overnight.

Thermofluor Assay

The assay was performed as described previously.[105,106] Each antibody was diluted to 5µM in PBS pH 7.2 and 45µL were added to six wells of an 8-well optically clear PCR strip. The two remaining wells were filled with 45µL of PBS pH 7.2. SYPRO orange master stock was diluted from 5000x to 200x and 5µL was added to each sample. Each tube was mixed by flicking and briefly centrifuged to ensure all the liquid is at the bottom of the tube. The strips were loaded on a BioRad CFX 96 Real-Time PCR Detection System. The protocol ran in scan mode "FRET" with an "Unknown" sample defined for each well. A run is started at 25°C and every 30 seconds the temperature is incremented by 0.5°C until a temperature of 99°C is reached, then the sample is cooled down to 25°C again. Data was analyzed and visualized using a custom python script (See GitHub repository for details).

ELISA

To each well of a 96-well clear flat-bottom high-binding microplate was added 75µL of 0.2µg/mL antigen in PBS pH 7.2 (M8a-3 derivatives: SARS-CoV-2 WA1 RBD; Certolizumab derivatives: Soluble Human TNF; Omalizumab derivatives: IgE Fc) and was incubated overnight at 4°C. The next day, the antigen solution was discarded, and the plate was washed three times with 200µL of PT solution (0.1% Tween-20 in PBS pH 7.2), thoroughly discarding the solution between each wash. The plate was blocked by adding 150µL blocking solution (3% BSA in PT solution) to each well and incubated at room temperature, covered. After 1h, the blocking solution was discarded thoroughly. Purified antibodies were diluted in blocking solution to the following concentrations: 20nM, 10nM, 5nM, 2.5nM, 1.25nM, 0.625nM, 0.313nM, 0.078nM, then diluting 10-fold, 2nM, 1nM, 0.5nM, 0.25nM, 0.125nM, 0.063nM, 0.031nM, 0.008nM, then diluting 10-fold again, 20pM, 10pM, 5pM, 2.5pM, 1.25pM, 0.625pM, 0.313pM, 0.078pM. Then 75µL of each antibody dilution was added to a well and incubated for 90 minutes at room temperature, covered. During this incubation period, Ultra TMB ELISA substrate solution was transferred to a 50mL conical tube wrapped in aluminum foil to shield from light and left to warm up to room temperature in a closed drawer to further shield from light. The antibody solutions were discarded, and the plate was washed three times with 200µL of PT

solution, thoroughly discarding the solution between each wash. Secondary antibody (Goat anti-human IgG Fc HRP fusion protein) was diluted 1000-fold in blocking solution and 75µL was added to each well. The plate was again incubated for 1h at room temperature, covered. The secondary antibody solution was discarded, and the plate was washed three times with 200µL of PT solution (0.1% Tween-20 in PBS pH 7.2), thoroughly discarding the solution between each wash. To each well, 75µL of Ultra TMB ELISA substrate solution was added and the plate was left to develop for 7 minutes and 30 seconds. The reaction was quenched by adding 75µL of 1M HCl in each well. The ELISA was analyzed using an Agilent Biotek Epoch 2 plate reader by measuring the absorbance at 450nm. Data was analyzed and visualized using a custom python script (See GitHub repository for details).

Baculovirus PSMA Polyreactivity Assay

100µL of baculovirus prostate-specific membrane antigen (PSMA)[107] was diluted in 9.9mL of 100mM sodium bicarbonate solution pH 9.6. The mixture was gently mixed by inversion in a 15mL conical tube until the PSMA was completely suspended in solution. 75µL of the PSMA suspension was added to each well of a 96-well clear flat-bottom high-binding microplate, covered with a pierceable seal foil and stored at 4°C over night. The next day, plates were washed three times with 200µL phosphate buffered saline (PBS) pH 7.2. Plates were blocked by adding 75µL of blocking solution (0.5% by weight of bovine serum albumin in PBS) and incubated at room temperature, covered, for one hour. Humanized monoclonal antibodies and controls were diluted to 1µg/mL in blocking solution. After thoroughly removing the blocking solution from the microplates, 75µL of antibodies were added to a respective well with four replicates per antibody, plus eight wells with only blocking solution as a background measurement. The solution was incubated in the plates at room temperature, covered, for 90 minutes. During this incubation period, Ultra TMB ELISA substrate solution was transferred to a 50mL conical tube wrapped in aluminum foil to shield from light and left to warm up to room temperature in a closed drawer to further shield from light. After thoroughly removing the solution, the plates were washed three times with 200µL PBS pH 7.2. Secondary antibody (Goat anti-human IgG Fc antibody HRP fusion protein) was diluted to 1µg/mL in blocking solution. 75µL of secondary antibody solution was added to each well and incubated in the plates at room temperature, covered, for one hour. After thoroughly removing the solution, the plates were washed three times with 200µL PBS pH 7.2. To develop the polyreactivity assay, 75µL of Ultra TMB ELISA substrate solution was added to each well and left to develop for 7 minutes and 30 seconds, after which 75µL of 1M HCl was added to each well to quench the reaction. Plates were analyzed with an Agilent Biotek Epoch 2 plate reader. The resulting CSV files were then analyzed with python scripts detailed in the section “Data Analysis and Visualization”.
Positive Controls: Fleish antibody, 45-46m2,[108] HCI Sap10.
Negative Controls: N6,[109] 10-1074.[110]

References

1. Roth, D. B. V(D)J Recombination: Mechanism, Errors, and Fidelity. Microbiol Spectr 2, 10.1128/microbiolspec.MDNA3-0041–2014 (2014).
2. Odegard, V. H. & Schatz, D. G. Targeting of somatic hypermutation. Nat Rev Immunol 6, 573–583 (2006).
3. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature 566, 393–397 (2019).
4. Liu, J. K. H. The history of monoclonal antibody development – Progress, remaining challenges and future innovations. Ann Med Surg (Lond) 3, 113–116 (2014).
5. Sgro, C. Side-effects of a monoclonal antibody, muromonab CD3/orthoclone OKT3: bibliographic review. Toxicology 105, 23–29 (1995).
6. Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S. & Winter, G. Replacing the complementarity-determining regions in a human antibody with those from a mouse. Nature 321, 522–525 (1986).
7. Tennenhouse, A. et al. Computational optimization of antibody humanness and stability by systematic energy-based ranking. Nat. Biomed. Eng 8, 30–44 (2024).
8. Arslan, M., Karadag, D. & Kalyoncu, S. Conformational changes in a Vernier zone region: Implications for antibody dual specificity. Proteins 88, 1447–1457 (2020).
9. Clavero-Álvarez, A., Di Mambro, T., Perez-Gaviro, S., Magnani, M. & Bruscolini, P. Humanization of Antibodies using a Statistical Inference Approach. Sci Rep 8, 14820 (2018).
10. Antibody therapeutics approved or in regulatory review in the EU or US. The Antibody Society https://www.antibodysociety.org/resources/approved-antibodies/.
11. UCB Financials | UCB. https://www.ucb.com/investors/UCB-financials.
12. Lu, R.-M. et al. Development of therapeutic antibodies for the treatment of diseases. Journal of Biomedical Science 27, 1 (2020).
13. Aubrey, N. & Billiald, P. Antibody Fragments Humanization: Beginning with the End in Mind. in Human Monoclonal Antibodies: Methods and Protocols (ed. Steinitz, M.) 231–252 (Springer, New York, NY, 2019). doi:10.1007/978-1-4939-8958-4_10.
14. Zhang, W. et al. Developability assessment at early-stage discovery to enable development of antibody-derived therapeutics. Antib Ther 6, 13–29 (2022).
15. Kernstock, R., Sperinde, G., Finco, D., Davis, R. & Montgomery, D. Clinical Immunogenicity Risk Assessment Strategy for a Low Risk Monoclonal Antibody. AAPS J 22, 60 (2020).

16. Mora, J. R. & Richards, S. M. The AAPS Journal Theme Issue: Compendium of Immunogenicity Risk Assessments: an Industry Guidance Built on Experience and Published Work. AAPS J 25, 43 (2023).
17. SAbDab: the structural antibody database | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/42/D1/D1140/1044118.
18. Olsen, T. H., Boyles, F. & Deane, C. M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci 31, 141–146 (2022).
19. Raybould, M. I. J. et al. Thera-SAbDab: the Therapeutic Structural Antibody Database. Nucleic Acids Research 48, D383–D388 (2020).
20. Marks, C., Hummer, A. M., Chin, M. & Deane, C. M. Humanization of antibodies using a machine learning approach on large-scale repertoire data. Bioinformatics 37, 4041–4047 (2021).
21. Prihoda, D. et al. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. MAbs 14, 2020203.
22. ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/38/7/1877/6517780.
23. Ruffolo, J. A., Chu, L.-S., Mahajan, S. P. & Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Nat Commun 14, 2389 (2023).
24. Luo, X. et al. BERT2DAb: a pre-trained model for antibody representation based on amino acid sequences and 2D-structure. MAbs 15, 2285904.
25. AbLang: an antibody language model for completing antibody sequences | Bioinformatics Advances | Oxford Academic. https://academic.oup.com/bioinformaticsadvances/article/2/1/vbac046/6609807.
26. Martinkus, K. et al. AbDiffuser: Full-Atom Generation of in vitro Functioning Antibodies.
27. Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N. & Madani, A. ProGen2: Exploring the boundaries of protein language models. Cell Systems 14, 968-978.e3 (2023).
28. Liu, Y. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at https://doi.org/10.48550/arXiv.1907.11692 (2019).
29. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. Genome Inform 23, 205–211 (2009).
30. OPIG. https://opig.stats.ox.ac.uk/.
31. Kovaltsuk, A., Krawczyk, K., Kelm, S., Snowden, J. & Deane, C. M. Filtering Next-Generation Sequencing of the Ig Gene Repertoire Data Using Antibody Structural Information. J Immunol 201, 3694–3704 (2018).
32. Aumasson, J.-P. & Bernstein, D. J. SipHash: A Fast Short-Input PRF. in Progress in Cryptology - INDOCRYPT 2012 (eds. Galbraith, S. & Nandi,

M.) vol. 7668 489–508 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012).
33. ANARCI: antigen receptor numbering and receptor classification | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/32/2/298/1743894.
34. Manso, T. et al. IMGT® databases, related tools and web resources through three main axes of research and development. Nucleic Acids Research 50, D1262–D1272 (2022).
35. Harris, C. R. et al. Array programming with NumPy. Nature 585, 357–362 (2020).
36. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023).
37. BioPhi Antibody design platform. https://biophi.dichlab.org/.
38. Stutz, C. & Blein, S. A single mutation increases heavy-chain heterodimer assembly of bispecific antibodies by inducing structural disorder in one homodimer species. J Biol Chem 295, 9392–9408 (2020).
39. Single Mutation on Trastuzumab Modulates the Stability of Antibody–Drug Conjugates Built Using Acetal-Based Linkers and Thiol-Maleimide Chemistry | Journal of the American Chemical Society. https://pubs.acs.org/doi/10.1021/jacs.1c07675.
40. Changing the Antigen Binding Specificity by Single Point Mutations of an Anti-p24 (HIV-1) Antibody1 | The Journal of Immunology | American Association of Immunologists. https://journals.aai.org/jimmunol/article/165/8/4505/7491/Changing-the-Antigen-Binding-Specificity-by-Single.
41. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 118, e2016239118 (2021).
42. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun 13, 4348 (2022).
43. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100 (2023).
44. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
45. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. Proceedings of the National Academy of Sciences 103, 5869–5874 (2006).
46. Deeks, E. D. Certolizumab Pegol: A Review in Inflammatory Autoimmune Diseases. BioDrugs 30, 607–617 (2016).
47. Okayama, Y. et al. Roles of omalizumab in various allergic diseases. Allergology International 69, 167–177 (2020).
48. Chen, M.-L., Nopsopon, T. & Akenroye, A. Incidence of Anti-Drug Antibodies to Monoclonal Antibodies in Asthma: A Systematic Review and Meta-Analysis. J Allergy Clin Immunol Pract 11, 1475-1484.e20 (2023).

49. O'Hagan, S., Galway, N., Shields, M. D., Mallett, P. & Groves, H. E. Review of the Safety, Efficacy and Tolerability of Palivizumab in the Prevention of Severe Respiratory Syncytial Virus (RSV) Disease. Drug Healthc Patient Saf 15, 103–112 (2023).
50. Ott, M. et al. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. Preprint at https://doi.org/10.48550/arXiv.1904.01038 (2019).
51. Heads, J. T. et al. Relative stabilities of IgG1 and IgG4 Fab domains: Influence of the light–heavy interchain disulfide bond architecture. Protein Sci 21, 1315–1322 (2012).
52. Bai, N., Roder, H., Dickson, A. & Karanicolas, J. Isothermal Analysis of ThermoFluor Data can readily provide Quantitative Binding Affinities. Sci Rep 9, 2650 (2019).
53. Cunningham, O., Scott, M., Zhou, Z. S. & Finlay, W. J. J. Polyreactivity and polyspecificity in therapeutic antibody development: risk factors for failure in preclinical and clinical development campaigns. MAbs 13, 1999195.
54. Dobson, C. L. et al. Engineering the surface properties of a human monoclonal antibody prevents self-association and rapid clearance in vivo. Sci Rep 6, 38644 (2016).
55. Hötzel, I. et al. A strategy for risk mitigation of antibodies with fast clearance. MAbs 4, 753–760 (2012).
56. Bepler, T. & Berger, B. Learning the Protein Language: Evolution, Structure and Function. Cell Syst 12, 654-669.e3 (2021).
57. Gloeckle, F., Idrissi, B. Y., Rozière, B., Lopez-Paz, D. & Synnaeve, G. Better & Faster Large Language Models via Multi-token Prediction. Preprint at https://doi.org/10.48550/arXiv.2404.19737 (2024).
58. Wang, B., Shen, T., Long, G., Zhou, T. & Chang, Y. Structure-Augmented Text Representation Learning for Efficient Knowledge Graph Completion. in Proceedings of the Web Conference 2021 1737–1748 (2021). doi:10.1145/3442381.3450043.
59. Armah-Sekum, R. E., Szedmak, S. & Rousu, J. Protein function prediction through multi-view multi-label latent tensor reconstruction. BMC Bioinformatics 25, 174 (2024).
60. Artamonov, O. S., Mykhailiuk, P. K., Voievoda, N. M., Volochnyuk, D. M. & Komarov, I. V. Simple and Efficient Procedure for a Multigram Synthesis of Both trans- and cis-1-Amino-2-(trifluoromethyl)cyclopropane-1-carboxylic Acid. Synthesis 2010, 443–446 (2010).
61. Gokemeijer, J. et al. Survey Outcome on Immunogenicity Risk Assessment Tools for Biotherapeutics: an Insight into Consensus on Methods, Application, and Utility in Drug Development. AAPS J 25, 55 (2023).
62. Bailly, M. et al. Predicting Antibody Developability Profiles Through Early Stage Discovery Screening. MAbs 12, 1743053 (2020).
63. 7. Introduction to Markov chain Monte Carlo — BE/Bi 103 b documentation. https://bebi103b.github.io/lessons/07/index.html.

64. Margossian, C. C. & Gelman, A. For how many iterations should we run Markov chain Monte Carlo? Preprint at https://doi.org/10.48550/arXiv.2311.02726 (2024).
65. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109 (1970).
66. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by Simulated Annealing. Science 220, 671–680 (1983).
67. Twist Bioscience | We lead innovation in DNA synthesis. https://www.twistbioscience.com/.
68. Yang, J. et al. DeCOIL: Optimization of Degenerate Codon Libraries for Machine Learning-Assisted Protein Engineering. ACS Synth. Biol. 12, 2444–2454 (2023).
69. Benatuil, L., Perez, J. M., Belk, J. & Hsieh, C.-M. An improved yeast transformation method for the generation of very large human antibody libraries. Protein Engineering, Design and Selection 23, 155–159 (2010).
70. Cherf, G. M. & Cochran, J. R. Applications of yeast surface display for protein engineering. Methods Mol Biol 1319, 155–175 (2015).
71. Telford, W. G. Flow cytometry and cell sorting. Front. Med. 10, (2023).
72. Fei, C., Nie, L., Zhang, J. & Chen, J. Potential Applications of Fluorescence-Activated Cell Sorting (FACS) and Droplet-Based Microfluidics in Promoting the Discovery of Specific Antibodies for Characterizations of Fish Immune Cells. Front. Immunol. 12, (2021).
73. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. Nat Biotechnol 15, 553–557 (1997).
74. Van Deventer, J. A., Kelly, R. L., Rajan, S., Wittrup, K. D. & Sidhu, S. S. A switchable yeast display/secretion system. Protein Eng Des Sel 28, 317–325 (2015).
75. Ovaa, H. & Wals, K. Unnatural amino acid incorporation in E. coli: current and future applications in the design of therapeutic proteins. Front. Chem. 2, (2014).
76. Sun, Y. et al. A fine-tuned yeast surface-display/secretion platform enables the rapid discovery of neutralizing antibodies against Clostridioides difficile toxins. Microb Cell Fact 22, 194 (2023).
77. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. Preprint at http://arxiv.org/abs/1312.6114 (2022).
78. Doersch, C. Tutorial on Variational Autoencoders. Preprint at http://arxiv.org/abs/1606.05908 (2021).
79. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. Preprint at http://arxiv.org/abs/1401.4082 (2014).
80. Kingma, D. P., Mohamed, S., Rezende, D. J. & Welling, M. Semi-supervised Learning with Deep Generative Models.

81. Gregor, K., Danihelka, I., Graves, A., Rezende, D. J. & Wierstra, D. DRAW: A Recurrent Neural Network For Image Generation. Preprint at http://arxiv.org/abs/1502.04623 (2015).
82. Kulkarni, T. D., Whitney, W. F., Kohli, P. & Tenenbaum, J. Deep Convolutional Inverse Graphics Network.
83. Wu, Z., Johnston, K. E., Arnold, F. H. & Yang, K. K. Protein sequence design with deep generative models. Current Opinion in Chemical Biology 65, 18–27 (2021).
84. Hawkins-Hooker, A. et al. Generating functional protein variants with variational autoencoders. PLOS Computational Biology 17, e1008736 (2021).
85. Ziegler, C., Martin, J., Sinner, C. & Morcos, F. Latent generative landscapes as maps of functional diversity in protein sequence space. Nat Commun 14, 2222 (2023).
86. Mansoor, S., Baek, M., Park, H., Lee, G. R. & Baker, D. Protein Ensemble Generation Through Variational Autoencoder Latent Space Sampling. J. Chem. Theory Comput. 20, 2689–2695 (2024).
87. Hopf, T. A. et al. The EVcouplings Python framework for coevolutionary sequence analysis. Bioinformatics 35, 1582–1584 (2019).
88. Lyu, S., Sowlati-Hashjin, S. & Garton, M. ProteinVAE: Variational AutoEncoder for Translational Protein Design. Preprint at https://doi.org/10.1101/2023.03.04.531110 (2023).
89. Abadi, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.
90. Yang, K. K., Fusi, N. & Lu, A. X. Convolutions are competitive with transformers for protein sequence pretraining. Cell Systems 15, 286-294.e2 (2024).
91. Kalchbrenner, N. et al. Neural Machine Translation in Linear Time. Preprint at https://doi.org/10.48550/arXiv.1610.10099 (2017).
92. Hie, B. et al. A high-level programming language for generative protein design. 2022.12.21.521526 Preprint at https://doi.org/10.1101/2022.12.21.521526 (2022).
93. Chailyan, A., Marcatili, P. & Tramontano, A. The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. FEBS J 278, 2858–2866 (2011).
94. Bordoli, L. et al. Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc 4, 1–13 (2009).
95. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35, 1026–1028 (2017).
96. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7, 539 (2011).
97. Biasini, M. et al. OpenStructure: an integrated software framework for computational structural biology. Acta Cryst D 69, 701–709 (2013).

98. Studer, G. et al. ProMod3—A versatile homology modelling toolbox. PLOS Computational Biology 17, e1008667 (2021).
99. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011).
100. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (2016). doi:10.1145/2939672.2939785.
101. McGibbon, R. T. et al. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys J 109, 1528–1532 (2015).
102. Lee, J. U., Shin, W., Son, J. Y., Yoo, K.-Y. & Heo, Y.-S. Molecular Basis for the Neutralization of Tumor Necrosis Factor α by Certolizumab Pegol in the Treatment of Inflammatory Autoimmune Diseases. International Journal of Molecular Sciences 18, 228 (2017).
103. Van Deventer, J. A. & Wittrup, K. D. Yeast Surface Display for Antibody Isolation: Library Construction, Library Screening, and Affinity Maturation. in Monoclonal Antibodies: Methods and Protocols (eds. Ossipow, V. & Fischer, N.) 151–181 (Humana Press, Totowa, NJ, 2014). doi:10.1007/978-1-62703-992-5_10.
104. Zhao, Q., Zhu, Z. & Dimitrov, D. S. Yeast Display of Engineered Antibody Domains. Methods Mol Biol 899, 73–84 (2012).
105. Lavinder, J. J., Hari, S. B., Sullivan, B. J. & Magliery, T. J. High-Throughput Thermal Scanning: A General, Rapid Dye-Binding Thermal Shift Screen for Protein Engineering. J. Am. Chem. Soc. 131, 3794–3795 (2009).
106. Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N. & Nordlund, P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. Analytical Biochemistry 357, 289–298 (2006).
107. Lodge, P. A. et al. Expression and purification of prostate-specific membrane antigen in the baculovirus expression system and recognition by prostate-specific membrane antigen-specific T cells. J Immunother 22, 346–355 (1999).
108. Diskin, R. et al. Restricting HIV-1 pathways for escape using rationally designed anti–HIV-1 antibodies. J Exp Med 210, 1235–1249 (2013).
109. Huang, J. et al. Identification of a CD4-Binding-Site Antibody to HIV that Evolved Near-Pan Neutralization Breadth. Immunity 45, 1108–1121 (2016).
110. Caskey, M. et al. Antibody 10-1074 suppresses viremia in HIV-1-infected individuals. Nat Med 23, 185–191 (2017).

Chapter 3

DEVELOPING A SCALABLE ENGINEERING PLATFROM FOR THE BIOELECTROCATALYTIC REDUCTION OF N_2 TO AMMONIA USING NITROGENASE

Abstract

Nature balances ammonia production and consumption in a highly efficient manner. Industrial nitrogen fixation processes, however, dramatically disrupt this balance, putting a heavy burden on the environment and rendering its long-term usage unsustainable. With the growing urgency of achieving sustainable and environmentally friendly industrial processes, the flaws of industrial nitrogen fixation, the process of reacting N_2 with H_2 to produce ammonia (NH_3), become alarmingly apparent: It consumes ~5% of the global supply of natural gas, makes up ~3% of global CO_2 emissions and consumes ~1% of the global power supply. Nevertheless, the importance of industrial nitrogen fixation is hard to overstate. Before the 20th century humans relied on the enzyme nitrogenase, native to some bacteria and archaea, animal waste or natural saltpeter deposits to fertilize soil, none of which could keep up with global population growth. In 1913, the Haber-Bosch process revolutionized agriculture and caused an unprecedented population boom by providing a way of synthesizing the fertilizer NH_3 at industrial scale. Today, around 230 megatons of NH_3 is produced per year via the Haber-Bosch process. Consequently, nearly 50% of the nitrogen found in human tissues originates from an energy-intensive process that requires the use of fossil fuels.

In this chapter, we discuss establishing an interdisciplinary research project with a translational focus to possibly supplant the Haber-Bosch process. We propose engineering nitrogenase, a protein complex that catalyzes the reduction of H^+ and N_2 to produce NH_3, for its use in a bioelectrocatalytic system, that is, a system where enzymes catalyze a redox reaction at an electrode. We present the early efforts in establishing the project, describing the engineering goals and the screening methods. Engineering nitrogenase to produce ammonia in an electrified reactor could provide an economically and ecologically viable alternative to the Haber-Bosch process, taking humanity a step closer to sustainable agricultural practices.

Contributions

The project was conceived by Prof. Steve Mayo and designed by Lucas Schaus with the help of Prof. Steve Mayo, Prof. Doug Rees, Dr. Rebeccah Warmack and Andres Orta. Techno-economic analysis of the project was performed by Lucas Schaus. The ancestral sequences were generated by Lucas Schaus and were tested by Dr. Rebeccah Warmack and Yuanbo Shen. ProtGPT2 and ESM2 sequence generation was performed by Olive Cheng and Lucas Schaus. Growth- and fluorescence assays were performed by Andres Orta, Mikhail Hameedi and Olive Cheng with guidance from Lucas Schaus. Writing was performed by Lucas Schaus. Experiments were performed in the lab of Prof. Steve Mayo and the lab of Prof. Doug Rees.

The Human and Environmental Impact of the Haber-Bosch Process

With the growing urgency of achieving sustainable and environmentally friendly industrial processes, the flaws of industrial nitrogen fixation, the process of reacting N_2 with H_2 to produce ammonia (NH_3), become alarmingly apparent:[1] It consumes ~5% of the global supply of natural gas,[2] makes up ~3% of global CO_2 emissions[1,3] and consumes ~1% of the global power supply.[2-4] The ammonia produced from this process is largely used to fertilize soil for food production, which makes industrial nitrogen fixation a process of global importance as 40-60% of global food production depends on fertilization with ammonia. Effectively, a single chemical process, the Haber-Bosch process, is responsible for 90% of global ammonia production.[5,6] Before the 20th century humans relied on the enzyme nitrogenase, native to some bacteria and archaea, animal waste or natural saltpeter deposits to fertilize soil, none of which could keep up with the global population growth.[4] In 1913, the Haber-Bosch process revolutionized agriculture and caused an unprecedented population boom by providing a way of synthesizing the fertilizer NH_3 at industrial scale.[2,4,7] Today, around 230 megatons of NH_3 is produced per year via the Haber-Bosch process.[8,9] Consequently, nearly 50% of the nitrogen found in human tissues is tied to an energy-intensive process that requires the use of fossil fuels.[2,4,9]

Even after more than 100 years, the Haber-Bosch process is still the most widely used synthesis method for ammonia.[9] In this process, N_2 gas is reduced to NH_3 using H_2 gas:

$$N_2(g) + 3H_2(g) \rightarrow 2NH_3(l), \quad E_0 = 0.057$$

The high energy consumption is necessary to maintain the high temperature and pressure of ~450°C and 200 atm.[10] One of the principal reasons for the high CO_2 emissions of the Haber-Bosch process is due to the hydrogen production method. The most economical pathway for H_2 production uses natural gas via steam methane reforming (SMR), followed by the water-gas shift reaction, which under ideal conditions produces one molecule of CO_2 per four molecules of H_2.[9] Much of the focus in developing a greener ammonia production process is to decrease the energy consumption by moving to less extreme conditions and to find an alternative method for producing hydrogen. Sustainably producing ammonia using a hydrogen gas dependent approach requires finding a green approach to hydrogen gas production that is low-cost in order to be competitive with SMR or coal gasification. However, there is currently no cost competitive hydrogen synthesis approach that does not use natural gas. The next best alternative, water electrolysis, consumes large amounts of energy and requires 9 tons of preprocessed high-purity water per ton of hydrogen, which makes this process too costly in all but very niche applications.[9] Producing ammonia using hydrogen gas directly adds an extra dimension of problems to solve

in order to make it sustainable before considering green ammonia synthesis processes themselves.

Roughly 80% of the ammonia produced in the Haber-Bosch process is used for crop fertilization.[11] As agriculture is largely decentralized, a high transportation cost is incurred when ammonia is produced in a centralized fashion as in the Haber-Bosch process.[7,12] The fertilizer is liquified at high pressure in its anhydrous form for transportation. This highly concentrated ammonia is toxic to humans and the environment. It is then either diluted for application on fields or injected at high pressure into the soil. Ammonia overuse for crop cultivation caused approximately 25 thousand deaths globally in 2012.[13] An alternative, low-concentration ammonia synthesis process where production is performed at a farm or highly distributed throughout agricultural regions can reduce fertilizer overuse and mitigate the environmental impact of ammonia toxicity as well as reduce transportation cost both economically and environmentally.[14] While such a process would be dependent on the local price of electricity, the current cost of ammonia is mostly dependent on the price of natural gas, a fossil fuel whose supply should drop by 65% by 2050 if the goals of the Paris Agreement are achieved.[15,16] Since centralized anhydrous ammonia production is not strictly necessary for agricultural applications, we believe that a small to medium scale aqueous process can have a significant impact towards achieving sustainable agricultural practices.

A Brief History of Nitrogenase

Before the dawn of the 20th century, humans were heavily restricted in the supply of nitrogen, mainly for the production of fertilizer and gun powder.[17,18] There were only a few ways of obtaining nitrogen for these purposes, collection of manure and night soil,[18] mining mineral nitrogen deposits or guano,[19] and creative usage of companion planting.[20] With the exception of some mineral deposits, all of the nitrogen sources above can be traced back to a single enzyme. Evidence suggests that organic, nitrogen-rich matter was deposited 2.5 billion years ago, around 50 million years before the earliest signs of rising atmospheric O_2 levels,[21] suggesting that the enzyme nitrogenase evolved before the advent of photosynthesis. While photosynthetic protein complexes eventually found themselves in eukaryotic organisms, as far as we know, nitrogenase only evolved in archaea and bacteria.

The scientific advances since the start of the industrial revolution have been responsible for an unprecedented boom in population and has increased the quality of life throughout the world. The economic growth that these technologies brought with them came to an impasse at the turn of the 20th century, as feeding the growing world population would only be possible by fertilizing crops with scarce sources of nitrogen, while also needing nitrogen to produce explosives. Between the 1820s and 1860s, all 12.5 million tons of guano were mined off the Chincha island in Peru. In 1879, the "Nitrate War" between Bolivia, Chile and Peru was fought over the geo-strategically important saltpeter deposits in the Atacama desert.[22] From the late 18th century onwards, chemists such as Georg Friedrich Hildebrandt attempted to synthesize ammonia, ushering in a century-long quest to achieve industrial-scale ammonia production.[23] The Haber-Bosch process solved the nitrogen scarcity problem and ushering a new era of food safety. Today, we find ourselves again at an impasse created by humanity's progress. This time the very technology that have carried humanity towards an era of prosperity is partially responsible for a climate crisis that risks undoing much of the progress made in the past 250 years.[17] For decades, researchers have attempted to dethrone the Haber-Bosch process[9] and replace it with a sustainable way to produce ammonia but have so far been unsuccessful.

One often proposed solution to the ammonia fertilizer production problem, is to go back to the roots of nitrogen fixation and circumventing synthetic ammonia completely. Inspired by the root nodules of legumes in which a symbiotic relationship with nitrogen fixing bacteria provides the plant with fixated nitrogen, plant biotechnologists have attempted to either replicate the root nodules in different plants or to express nitrogenase in plant cells. Starting in the latter half of the 20th century, this unachieved quest is considered the holy grail of plant biotechnology.[24] At the same time, biochemist and biophysicist became interested in the unique properties of nitrogenase, starting a journey of nitrogenase discoveries that lasts to this day.
The prokaryotic enzyme nitrogenase is a two-protein component system (Figure 3.1) that catalyzes the reduction of N_2 to ammonia coupled to the hydrolysis of ATP:[25]

$$N_2 + 8H^+ + 8e^- + 16ATP \rightarrow 2NH_3 + H_2 + 16ADP + 16P_i$$

This equation holds under the "standard" model, but other ratios of NH_3 to H_2 production and ATP consumption have been observed.[26] The protein complex catalyzing the reaction is made up of the Fe-protein, an ATPase responsible transferring electrons to the second component, the MoFe-protein, which contains the active site that reduces N_2 to ammonia. The full and detailed mechanism of nitrogenase is still unknown despite decades of effort to deduce it, but broadly the two proteins interact in four steps to perform the reduction (Figure 2A).[26] The active site of ammonia production contains a highly unique cofactor called FeMoco (Figure 3.1), whose structure was only recently elucidated to have the elemental composition [Mo:7Fe:9S:C] and which contains a spectroscopically elusive central C^{4-} carbide.[27] The MoFe-protein and its cofactor also suffer from two known mechanisms of permanent deactivation, one oxygen dependent[25] and one pH- and turnover dependent.[28] Additionally, nitrogenase is dependent on ferredoxin/flavodoxin to reduce the Fe-protein and it is dependent on ATP hydrolysis to transfer electrons to FeMoco.[26]

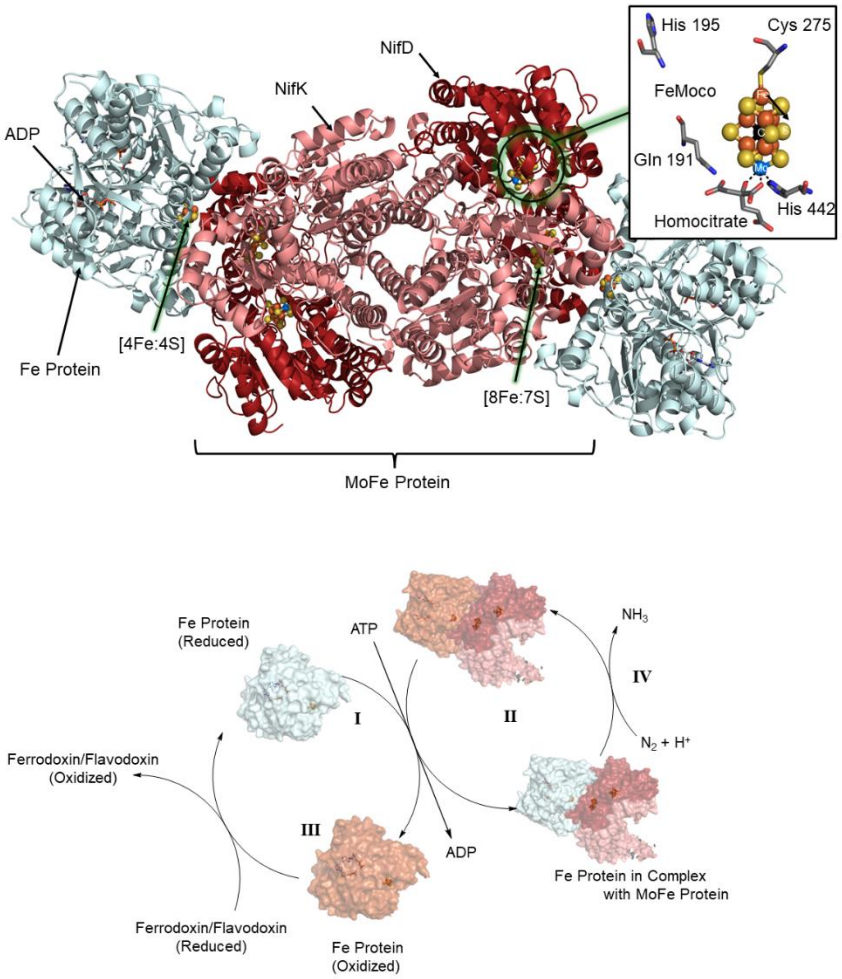


Figure 3.1: Structure of the nitrogenase complex (PDB ID: 1N2C) and overview of mechanism. (**Top**) The complex is composed of two main protein chains, the Fe-protein and the MoFe-protein. The Fe-protein is an ATPase that contains a [4Fe:4S] cluster. Through ATP hydrolysis, the protein is responsible for transferring electrons to the MoFe-protein. The MoFe-protein is a heterotetramer made up of two NifK and two NifD subunits and contains the active site of N2 reduction. In the interface between NifK and NifD lies the [8Fe:7S] cluster which acts as an intermediate to transfer electrons from the Fe-protein to FeMoco. FeMoco (top right box, PDB ID: 3U7Q) is a unique cofactor containing a molybdenum atom and a central carbide (formally C4-). The cofactor is the active site of N2 reduction. (**Bottom**) *In vivo* nitrogenase mechanism. I: Fe-protein with two ATP bound, forms a complex with the MoFe-protein. II: ATP is hydrolyzed by the Fe-protein and initiates the electron transfer from the [4Fe:4S] cluster to the FeMoco through the [8Fe:7S] cluster. III: Fe-protein dissociates from the MoFe-protein, releases ADP and is re-reduced by ferrodoxin/flavodoxin. IV: The previous steps repeat until sufficient electrons have been transferred to FeMoco to reduce N2 to NH3. (Figures adapted from Einsle and Rees 2020)

Electrochemical Approaches to Supplant the Haber-Bosch Process

Even after more than 100 years, the Haber-Bosch process is still the most widely used synthesis method for ammonia.[9] In this process, N_2 gas is reduced to NH_3 using H_2 gas:

$$N_2(g) + 3H_2(g) \rightarrow 2NH_3(l), \quad E_0 = 0.057V$$

The high energy consumption is necessary to maintain the high temperature and pressure of ~450°C and 200 atm.[10] One of the principal reasons for the high CO_2 emissions of the Haber-Bosch process is due to the hydrogen production method. The most economical pathway for H_2 production uses natural gas via steam methane reforming (SMR), followed by the water-gas shift reaction, which under ideal conditions produces one molecule of CO_2 per four molecules of H_2.[9] Much of the focus in developing a greener ammonia production process is to decrease the energy consumption by moving to less extreme conditions and to find an alternative method for producing hydrogen. Sustainably producing ammonia using a hydrogen gas dependent approach requires finding a green approach to hydrogen gas production that is low-cost in order to be competitive with SMR or coal gasification. However, there is currently no cost competitive hydrogen synthesis approach that does not use natural gas. The next best alternative, water electrolysis, consumes large amounts of energy and requires 9 tons of preprocessed high-purity water per ton of hydrogen, which makes this process too costly in all but very niche applications.[9] Producing ammonia using hydrogen gas directly adds an extra dimension of problems to solve in order to make it sustainable before considering green ammonia synthesis processes themselves.

A more direct approach to solving the green ammonia problem is to produce it electrochemically, which produces hydrogen *in situ* on a platinum anode from water:[9]

$$N_2(g) + 3H_2O(l) \rightarrow 2NH_3(l) + \frac{3}{2}O_2(g), \quad E_0 = -1.164V$$

Although a variety of temperature, pressure, and electrical conditions exist in this approach, the electrochemical synthesis methods can broadly be categorized into low-, intermediate-, and high-temperature processes. Among these, the low-temperature processes usually bring lower capital expenditures making them more competitive, while also reducing the environmental impact by reducing its energy intensity. Compared to the Haber-Bosch process, the electrochemical approach can boast higher efficiency and is more suitable at small and medium scale production,[9] ideal for a decentralized approach to ammonia synthesis. Although using inorganic

electrocatalysts for ammonia production can be efficient, they are not necessarily truly green technologies. Many of the proposed catalysts use metals such as ruthenium, vanadium, iridium, and lithium among others which are environmentally toxic either in their nature and/or through their production processes.[29-32] There are promising electrochemical approaches that utilize non-toxic metals such as iron, but their per-electron efficiency (faraday efficiency) is under 2% under normal conditions.[33,34] Such low faraday efficiencies lead to higher energy requirements per unit of ammonia produced. Ideally, an approach to making ammonia should have high faraday efficiency, a low carbon footprint by utilizing mild reaction conditions and be non-toxic to the environment.

While many studies aiming to find a green catalyst for ammonia production have taken inspiration from nature's solution to the problem, the Minteer lab took inspiration from electrocatalytic ammonia cells to achieve the first bioelectrocatalytic ammonia synthesis using nitrogenase.[3] In 2010, Roth *et al*., showed through covalently linking a ruthenium catalyst to the MoFe-protein that substrate reduction could occur independent of the Fe-protein and ATP hydrolysis.[35] The first attempt at bioelectrocatalytic reduction of N_2 by Milton *et al*. used ATP and H_2 as the energy source which generated low amounts of electricity while producing ~0.7 nmol NH_3 per nmol MoFe-protein per hour.[3] In 2020, Lee *et al*. decoupled the reaction from ATP and H_2 by using a cobaltocene modified polymer to transfer electrons to the MoFe-protein directly as has been demonstrated by Roth *et al*.[36] This setup improved the yield to ~140 nmol NH_3 per nmol MoFe-protein per hour and has a faraday efficiency of ~45%. While a remarkable feat, the lifetime of the catalyst in the reactor is at best a day[10] (and at worst one hour[36]) which makes this technology not competitive with traditional chemical catalysts. While the Minteer lab focused on the engineering of the reactor, the protein employed is the wild type nitrogenase of *Azotobacter vinelandii.* Since nitrogenase has been optimized via evolution to work inside its host cell, the enzyme is not optimized for a bioelectrocatalytic system.[12] However, we can use nature's mechanism to incrementally optimize properties of protein sequences via directed evolution. This places biocatalysts in a unique position compared to other catalysts because there is a direct iterative process with which enzyme sequences can be optimized.[37] Using a biocatalyst would not only be able to achieve the goal of green ammonia production, it also has a variety of other advantages that cannot be easily achieved using inorganic catalysts. Being produced by bacteria found in most soils, in water and on certain plant roots, nitrogenase is inherently benign to the environment. The individual components used to produce nitrogenase are recyclable; culture media, cell pellets, and the enzyme itself can be recycled to produce more nitrogenase or be used in other biological processes.[38]

Technoeconomic Analysis and Engineering Goals

In this section we will outline the engineering goals of this project by performing a technoeconomic analysis. We are not only seeking to engineer nitrogenase for the sake of scientific advancement, but we are fundamentally interested in performing translational research and producing a nitrogenase that can benefit the world at large. To do this, we need to set realistic engineering goals which reflect at what point the technology could be used in a non-laboratory setting. First, we would want to focus on the competitiveness of the technology purely in terms of engineering goals and not factoring in production cost and logistics. To achieve some sort of competitiveness, the catalyst needs to produce ammonia quickly enough to provide a farm with enough fertilizer for a year within that same year. According to the USDA,[39] the average farm in the US has around 445 acres of land, of which a third is cropland harvested in a year. Certain crops such as soybeans (~25% of crop land) do not need to be fertilized or only need to be fertilized in certain conditions such as corn (~30% of crop land). The median farm has around 100 acres of land. We will use both the average and the median farm for our calculations (100 and 22 fertilized acres respectively).[40] The amount of ammonia applied to a field can vary depending on the yield of the crop, with high-yielding crops needing more fertilizer. For simplicity, we will divide the necessary anhydrous ammonia per acre into three categories: Low yield with 64 kg per acre, medium yield with 91 kg per acre, and high yield with 136 kg per acre.[41] In our calculation, we multiple this number with the respective acreage of the average and median farm. Since our plan is to use solar energy to power the reactor and on average the US sees 205 days of sunshine, we assume need to produce the ammonia for the entire year within these 205 days (4920 hours).[42]

According to our own experience expressing nitrogenase, *A. vinelandii* yields 20 mg of nitrogenase per liter of culture. Assuming that we can keep this yield constant when scaling up, a 50,000-liter reactor can produce 1 kg of enzyme per batch.[43] For now, we assume that we provide the farm every 20 days with 100 g of enzyme for the reactor (I.e., the catalysts lifetime is 20 days). The final size of the bioelectrocatalytic reactor is not defined yet, so the 100g enzyme reactor size is subject to change and would influence the numbers below.

$$k = \frac{m(Ammonia) \cdot M(nitrogenasecomplex)}{t \cdot M(Ammonia) \cdot m(nitrogenasecomplex)}$$

Median US Farm Size	Low Yield	Medium Yield	High Yield
k(h^{-1})	4048	5782	8676
Fold improvement over Lee *et al*. (35 h^{-1})	115	165	248
Fold improvement over natural nitrogenase activity (14400 h^{-1})	0.28	0.4	0.6

Table 3.1: Reaction rate needed to produce enough ammonia for a given US farm size and crop yield. Yields are divided into low yield (64 kg NH3 per acre), medium yield (91 kg NH3 per acre), and high yield (136 kg NH3 per acre). Calculation of the reaction rate is performed via the formula above. As a comparison, we added the fold-improvement necessary against two references: The original bioelectrocatalytic reaction rate of nitrogenase by Lee et al.,[44] and the natural reaction rate of nitrogenase in Azotobacter vinelandii.[25]

We believe that improving the bioelectrocatalytic conversion of nitrogenase to around half the rate of the natural reactivity will make this approach competitive with the Haber-Bosch process, at least technologically, across most farms in the US. However, that would still require a 165-fold improvement in rate over the baseline presented in Lee *et al*.,[44] which is an ambitious goal (Table 3.1). Therefore, we propose that a reasonable goal for a translational research project is to achieve around 10% of the way to a technologically competitive catalyst, which would be a ~20-fold improvement in reaction rate. Of course, this model assumes that we can replace the catalyst every 20 days, which would require a catalyst that lasts as long.

For catalyst lifetime, we would like to set our first goal based on replacing the nitrogen in the enzyme, which makes it reaction rate and catalyst lifetime dependent. The range of reaction rates are chosen from the Lee *et al*. baseline (35 h^{-1})[44] and a 100-fold improvement of the reaction rate, which is where the technology could be used for low-yield crops on a median-sized farm. One molecule of the nitrogenase complex contains 4130 nitrogen atoms (PDB: 1N2C).[45] The time it takes to replace that amount of nitrogen is simply this number divided by the rate per unit time. Using these numbers, the original system from the Minteer lab could replace the nitrogen in the enzyme within ~5 days. At 20-fold improvement of the baseline rate, the same amount of nitrogen would be produced in 6 hours. Our initial goal is to improve the catalyst lifetime and reaction rate such that we surpass the blue line in (Figure 3.2). Next, we would want to tackle producing enough nitrogenase to replace the nitrogen in the culture *A. vinelandii* grows in. Typically, nitrogenase is expressed in Burk's Medium which contains 1.5 g/liter of ammonium acetate. At a 20-fold improvement of the rate over the baseline, we would need to have the catalyst last 20 days in order to produce as much nitrogen as has been used in the culture. Although this might

make it seem at first sight that our catalyst would need to last longer than 20 days, the medium can at least partially be recycled, either to produce more nitrogenase or to use as a feedstock for other biological applications.

$$t(Nitrogen\ in\ Culture) = \frac{1.5\ {}^{g}/_{l}}{k\ \cdot 2 \cdot 10^{-2}\ {}^{g}/_{l}}$$

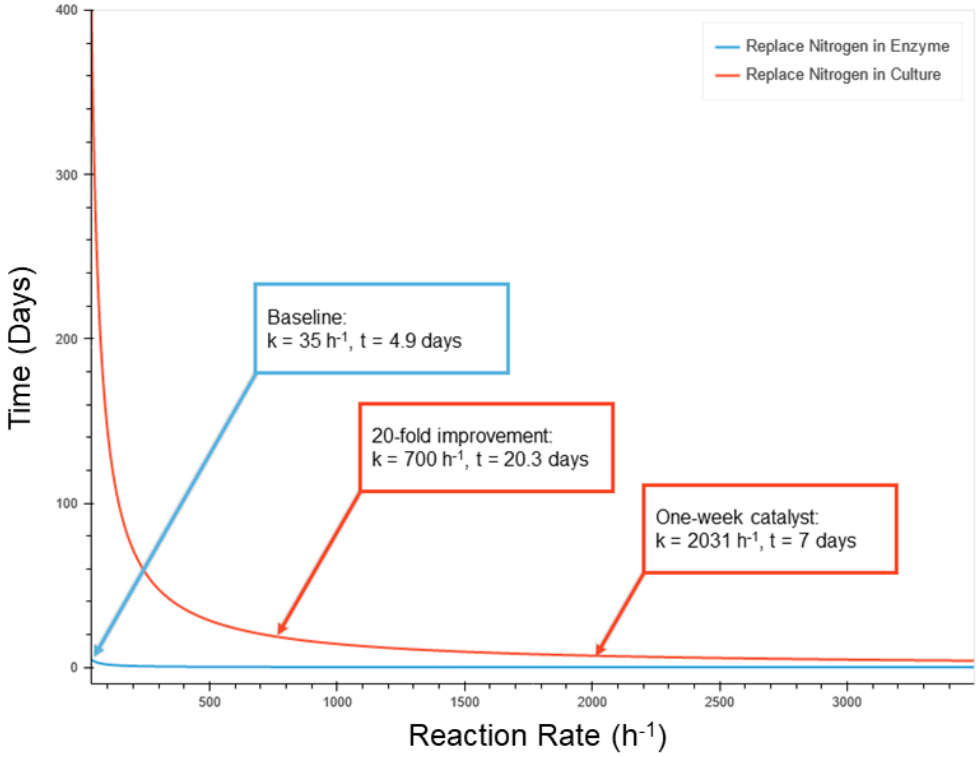


Figure 3.2: Graph representing the reaction rate – catalyst lifetime relationship necessary to replace either the nitrogen contained in the enzyme, or the nitrogen replaced in the culture used to express the enzyme.

Lastly, we want to focus on the price of ammonia produced via the Haber-Bosch process and how that could potentially influence the engineering the proposed technology. Currently, 15% of the price of ammonia is dictated by the price of natural gas. However, usually that share is larger since the supply chain issues and tariffs dictate a larger part of the price (19%).[46] The price of ammonia is often as volatile as the price of natural gas during the same time periods. As many nations are attempting to at least partially phase-out fossil fuels from their economy, one could expect this price volatility to get worse with time. Although the price of these commodities remains speculation, the price of ammonia that the farmer pays is highly competitive, even at the current all-time-high price of ammonia (USD 850 per ton).[47] In the context of translational research, it is important to keep the price of ammonia in mind

since it is difficult to sell a technology to farmers that ends up increasing their operational costs. Some of the factors that will affect the cost of a bioelectrocatalytic method of producing ammonia is the cost of electricity, the cost of water and the cost of expression, purification, and transportation of the enzyme. The latter is especially critical in our case since all of those steps likely need to be performed anaerobically.

In conclusion, the bioelectrocatalytic production of ammonia could from an engineering perspective become competitive with the Haber-Bosch process. If the price of natural gas continues to rise and if the carbon offsets from producing ammonia bioelectrocatalytically would reduce production costs, there is a case to be made for our method to be financially competitive as well.

Developing a Bioelectrocatalytic Screening Method for Nitrogenase

Natural selection induced by environmental pressures has been nature's means of pushing evolution forward. Through many generations of random mutations and selection that increase fitness, organisms overcome the challenges nature presents. Directed evolution has allowed humans to artificially simulate evolution, accelerating natural selection resulting in the generating proteins with new or improved functions.[37] While spontaneous mutations take far too long to occur in a laboratory environment, mutagenesis methods allow for the creation of genetic libraries that cover a vast array of gene variants which can be reintroduced into expression systems and screened for desired activity. The engineering of proteins through either *in silico* methods or directed evolution, requires the testing of protein fitness for each round putative improvements are made.[48] Through multiple generations of mutagenesis, a protein may be engineered to have enhanced characteristics such as solubility, thermostability, catalytic turnover and/or substrate affinity. Directed evolution is a well-established laboratory process that has proven to be successful in the engineering of proteins, such as evolving cytochromes P450 to perform new-to-nature reactions[49] or repurposing a bacterial protein to create a virus-like capsid.[50] The method chosen to screen new variants is one of the most important steps to successfully engineer a protein.[51] The screening assay needs to be as close to the final application as possible to avoid selecting for undesired properties and to improve upon the desired ones, commonly known by the phrase "You get what you screen for".[52] The throughput of the chosen screening method is also important in defining the library size, that is, how much diversity to generate in each round of evolution.[48] Ideally, the throughput should be as high as possible, but there are many factors that can determine the maximum throughput for a given protein engineering campaign.

Azotobacter vinelandii (*A. vinelandii*) is a gram-negative diazotroph bacterium commonly used as a model organism to study nitrogen fixation. Similar to *Escherichia coli* (*E. coli*), it is easily cultured and grown as it is an obligate aerobe. Under diazotrophic conditions, the organism strongly regulates nitrogen fixation to fulfill the requirements for cell growth due to the energetically expensive nature of biological nitrogen fixation.[53] Although the organism is an obligate aerobe, nitrogenase itself is highly oxygen sensitive.[26] It is currently not known how *A. vinelandii* achieves this feat. Likely due to the energetically expensive nature to fix nitrogen, the organism is very good at scavenging nitrogen from its environment and minor contaminants, such as cleaning products, can have an influence on the organism's growth. From a protein engineering perspective, the organism poses a challenge as it contains multiple copies of the nitrogenase producing NifD and NifK genes on its chromosomes.[54] To study the effect of mutations in the nitrogenase sequence, an *A. vinelandii* nitrogenase knock-out is used that is then rescued by knocking-in a mutated copy of the gene of interest.[55] In order to test mutants of nitrogenase in a bioelectrocatalytic system, we need to devise an appropriate

workflow to best assess improvements of the enzyme in the system where it will be used. As we are planning on engineering nitrogenase's ability to produce ammonia in a bioelectrocatalytic system, where the enzyme is isolated from other cellular components, we cannot use an *in vivo* selection assay. Besides generating mutants, screening for ammonia production requires a minimum of three main steps: (1) Aerobically grow nitrogenase mutants transformed into *A. vinelandii* in a 96-well format. (2) Isolate all 96 mutants and transfer them to a 96-deepwell plate equipped with an electrode under anaerobic conditions. (3) Measure ammonia production after a fixed period of time to assess the fitness of different variants. The aim is to create a low-to-medium throughput assay to engineer nitrogenase for bioelectrocatalytic production of ammonia. Using directed evolution, we can then improve the faraday efficiency, ammonia yield and catalyst lifetime of nitrogenase. This will provide a first step to engineering a sustainable approach to produce ammonia to help in fertilizing our crops and reduce the negative impact of modern agriculture on our planet.

Recent scientific advances will help us in our endeavor to engineer nitrogenase. The Kacar Research Group at the University of Wisconsin-Madison has established expression of the *A. vinelandii* model organism in a 96-well format, which together with knocked-in nitrogenase mutants shows that step (1) is feasible.[56] Work from the lab of Uli Schwanenberg has demonstrated that it is possible to test bioelectrocatalytic reactions in 96-well microtiter plates by adding an electrode array to the plates on which enzymes have been immobilized. Using their setup, they were able to engineer a bacterial laccase for use in an enzymatic fuel cell, improving the power output of the system 1.72-fold.[57] We propose a similar setup that immobilizes wild type nitrogenase on an electrode array to produce ammonia at small scale. Lastly, the Minteer lab has shown that it is possible to coat an electrode with the MoFe-protein of nitrogenase along with a polymer modified with a redox mediator to achieve ATP-free bioelectrocatalytic conversion of N_2 to ammonia.[36,44,58] Additionally, the study describes a method to measure NH_3 production via a fluorescence assay. Combining the above methods from the Kacar, Minteer and Schwanenberg lab will form the basis of our screening assay (Figure 3.3).

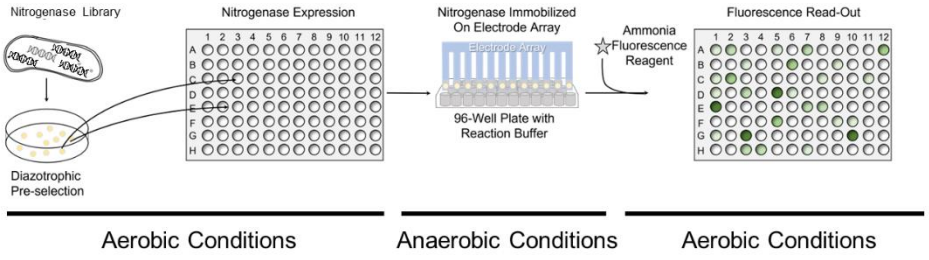


Figure 3.3: Workflow overview of bioelectrochemical screening assay. After generating nitrogenase libraries, we select for variants that have a functional nitrogenase through diazotrophic pre-selection on agar plates containing no nitrogen source. We then pick 24,48, or 96 colonies, depending on the possible throughput, to express nitrogenase. After sufficient nitrogenase has been expressed in each well, the plate is moved to an anaerobic chamber for purification and

electrochemical screening. Once the reaction has finished, the plate can be removed from the anaerobic chamber to measure the ammonia production through a fluorescence-based read-out.

The first step for our assay is to generate nitrogenase mutants and transform them into ΔNifD/NifK knock-out strains of *A. vinelandii*. While we will attempt to setup the bioelectrocatalytic screening setup in 96-well plates, the work from Chen *et al*. only demonstrated their system screening 8 reactions at a time. Since we want to screen for catalyst lifetime as well, that setup may reduce our throughput significantly. Therefore, after streaking out the bacteria onto growth media containing petri dishes, we will use the diazotrophic properties of the bacterium to our advantage to create a pre-selection of deleterious mutants (Figure 3.3) to greatly increase throughput. Since no nitrogenase gene copy other than a mutated one should be present within a cell, deleterious mutations in the protein will either cause the bacteria to grow very slowly or not grow at all under diazotrophic conditions. Due to the excellent scavenging skills of *A. vinelandii*, it is important to make sure no nitrogen containing contaminants influence the pre-screen.

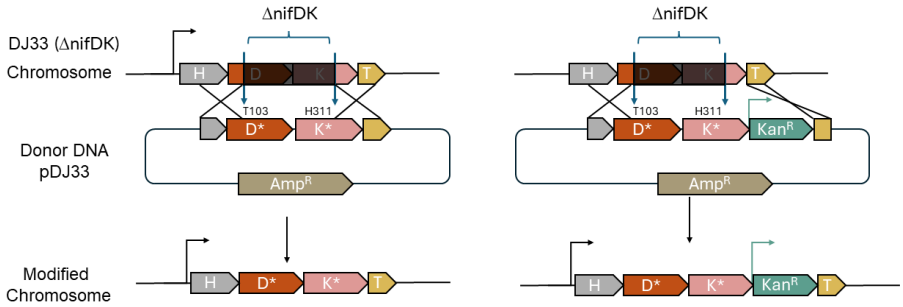


Figure 3.4: Explanation of NifD/NifK variant insertion into the chromosome of *A. vinelandii*. (**Left**) Insertion of NifD/NifK into the chromosome via homologies on the donor plasmid pDJ33. (**Right**) For this study we modified the pDJ33 plasmid to contain a kanamycin resistance gene between the NifK and NifT genes. While all Nif genes are regulated under one promoter, we added a second promoter into the Nif cluster to increase the expression of Kan^R.

We received a NifD/NifK knock-out strain from Dennis Dean and developed a method to insert NifD/NifK variants into the genome of *A. vinelandii*. Initial experiments of knocking-in wild-type nitrogenase rescued growth under diazotrophic conditions but cells grew very slowly and not to full density. Since *A. vinelandii* can accumulate up to 40 copies of its chromosome, we suspected that growth to full density can only be achieved if nearly all copies of the chromosome contain the knock-in. This is usually achieved by passaging cells through nitrogen deficient media multiple times.[59] This poses a problem in our experimental setup. We want to include negative controls during engineering to simulate non-functioning variants. Since negative controls are by definition non-functional, we cannot use passaging on them. Therefore, we developed a genetic system that contains a kanamycin resistance gene which is inserted to the chromosome with NifD/NifK (Figure 3.4). Only

variants containing the antibiotic resistance gene can grow on kanamycin containing plates. This selection pressure can then help us in accumulating multiple knock-in copies even if the variant is non-functional.

We confirmed the function of our genetic system by transforming *A. vinelandii* with the Kan^R containing version of pDJ33 as described in Figure 3.4. The cells were struck out on Burk's medium agar plates containing nitrogen as well as a defined kanamycin concentration (Figure 3.5A). Cells were grown for 60 hours at 30°C before imaging. The positive control contained no kanamycin and cells grew as a lawn. Increasing the kanamycin concentration from 0.5 μg/mL to 5 μg/mL thinned out the number of cells growing on the plates, indicating that the increase in selection pressure is allowing fewer cells to grow on the plates. This confirms that our genetic system works. Next, we will use the 96-well growth assay of *A. vinelandii* published by the Kacar lab to grow transformed *A. vinelandii* cells aerobically.[56] After transformation with pDJ33, we grew the cells to OD_{600}=1 in Burk's media (N+), followed by washing off the nitrogen containing media from the cells by spinning them down and resuspending in Burk's media (N-) twice. The cells were then diluted in Burk's media (N-) or (N+) containing 0.5μg/mL kanamycin to the starting OD_{600}=0.05 and transferred to a 96-well plate. Cells were then grown in a shaker-incubator at 30°C aerobically and diazotrophically, and their growth was measured in real time over 50 hours (Figure 3.5B). We were able to replicate the results from the Kacar lab, showing that our knock-in grows to an OD_{600} of around 0.3 in (N-) medium versus 0.4 in (N+) medium.

Before isolating nitrogenase mutants from the cultures it is important to transfer the cultures to an anaerobic environment to prevent the permanent deactivation of the enzyme by oxygen. Here, the Rees lab's expertise in working with nitrogenase under anaerobic conditions will allow us to isolate His-tagged enzymes via 96-well Ni-NTA columns. We will then coat the electrode array with nitrogenase mutants and apply a current to induce ammonia production, similar to the method proposed by the Schwanenberg lab and the Minteer lab. The electronic setup will also allow us to analyze important reaction parameters such as the potential, the current density, and the power input. Once the reaction is completed, the setup can be reintroduced to aerobic conditions to perform yield analysis. In a similar fashion to the Minteer lab, ammonia production by the system can be measured in a fluorescence detection assay, where an NH_3 detection buffer is added, and fluorescence is measured (Figure 3.3).[36,44,58]

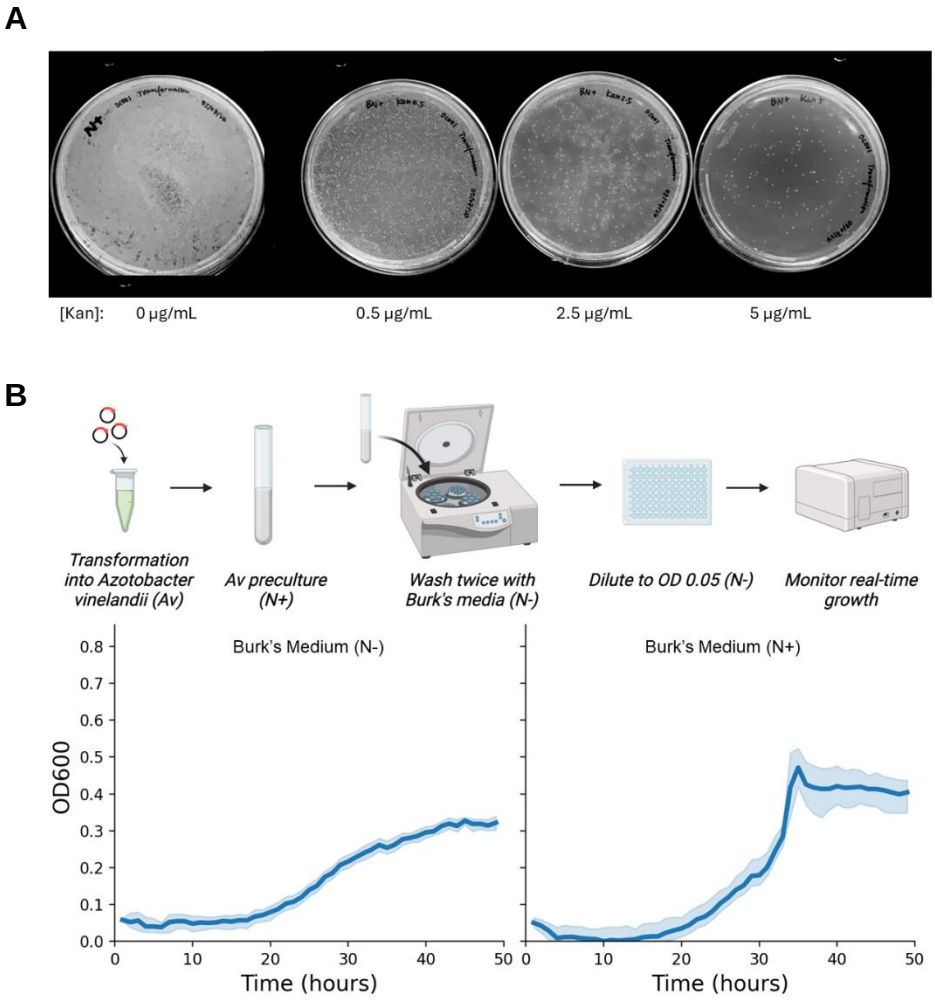


Figure 3.5: Preliminary results to establish a bioelectrocatalytic screen for nitrogenase engineering. (**A**) Demonstration of the function of the genetic system described in Figure 3.4. Knock-in of NifD/NifK into the genome of *A. vinelandii* were struck out on Burk's medium agar plates with added nitrogen (N+) and increasing concentrations of kanamycin. Plates were imaged 60 hours after transformation with pDJ33. (**B**) Illustration of growth assay and growth assay results. Cells were transformed using the pDJ33 containing wild type NifD/NifK and grown in Burk's media (N+) to OD_{600}=1. To remove nitrogen from the media, the cells were spun down and washed twice with Burk's media (N-) before diluting to the starting OD_{600}=0.05 for the start of the growth assay in nitrogen deficient or nitrogen supplemented media.

As of writing, we have not finished reproducing the bioelectrocatalytic system from the Minteer lab. Therefore, we have decided to attempt an engineering round of nitrogenase without screening for bioelectrocatalytic activity. We prepared an alanine scan of NifD to validate the proof-of-concept. This alanine scan will provide us with useful information on what sites of nitrogenase could potentially be targeted for site-saturation mutagenesis (Figure 3.6). As of writing, we have transformed the

alanine scan variants but have not yet subjected them to diazotrophic pre-selection or a real-time growth assay.

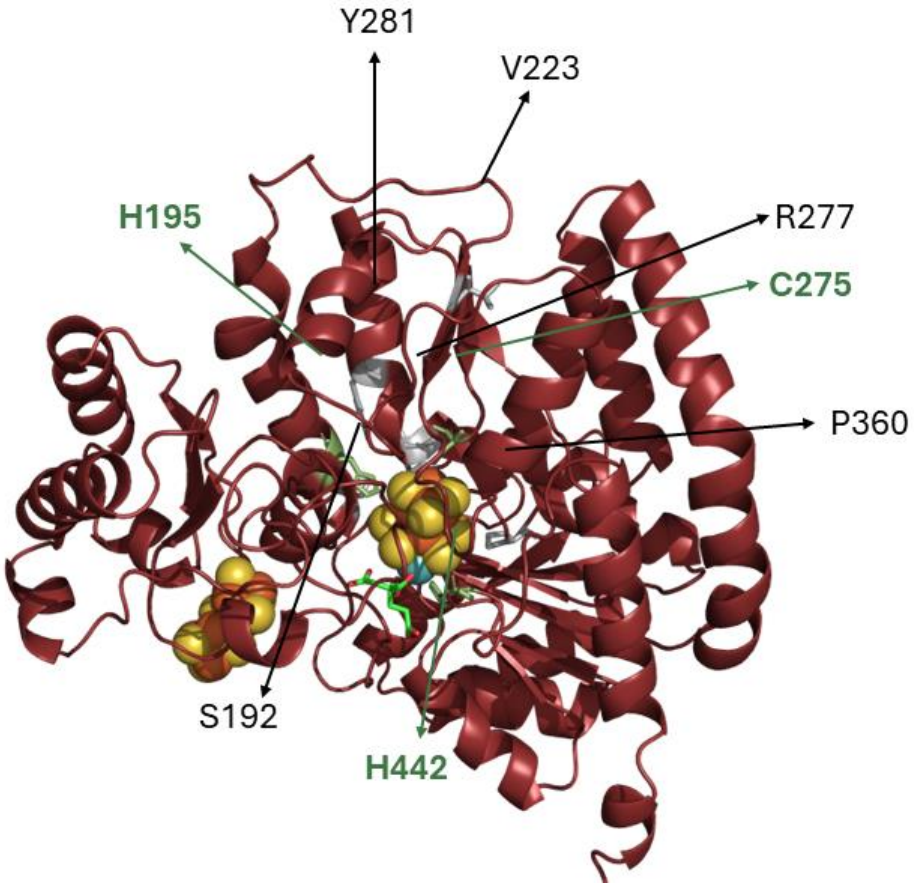


Figure 3.6: Structure of *A. vinelandii* NifD, highlighting positions of residues targeted in an alanine scan. Residues in green are negative controls as those positions are catalytically important.

The engineering platform outlined in this section will create an invaluable tool not only for our goal of bioelectrocatalytic conversion, but also for mechanistic studies of nitrogenase. Individual pieces of our setup can be modified and used by the scientific community to analyze the effects of mutations on nitrogenase at a much larger scale than previously possible. This has the potential to have a significant effect on our understanding of nitrogenase and can be revolutionary for applications in plant biotechnology where there is an ongoing quest to express nitrogenase in plant roots.

Computational Methods to Find Engineering Starting Points

Next-generation sequencing technologies have been revolutionary to the field of biology, providing high quantities of high-quality sequence data for various applications ranging from determining the B-cell profiles of COVID-19 patients for generating monoclonal antibody treatments,[60] to creating large sequence databases that helped in the development of Alphafold2.[61,62] In fact, the availability of protein sequence data on Uniprot has increased by an order-of-magnitude in the past decade. The development of highly efficient protein homology detection algorithms, with modern models being able to detect distant homologs with sequence similarities below 30%,[63] together with the ever-growing protein sequence databases has resulted in the construction of highly informative phylogenetic trees through a process called multiple sequence alignment (MSA).[64] The culmination of these technologies is found in Ancestral Sequence Reconstruction (ASR) a new and unique protein engineering tool with promising prospects.[65]

ASR is performed by using sequence data to find modern homologs of a protein of interest and then inferring the phylogenetic relationship between those homologs.[66-68] The inference uses statistical methods to calculate the probability distribution of sequences at each node of the phylogenetic tree. The inference is usually performed via Bayesian inference or by creating maximum likelihood distributions,[64] but recent advances in machine learning, in particular natural language processing, have allowed the usage of more complex inference models.[69] Regardless of the method of sequence distribution construction at the nodes, ancestral sequences are reconstructed by sampling from those distributions. This sampling can be performed with restraints by, among others, validating sequences thermodynamically or by forcing a certain consistency with a consensus sequence (Figure 3.7). ASR is fundamentally different from mutating proteins to the consensus sequence. The latter has consistently failed to produce functional proteins while ASR has produced enzymes that are more thermostable[66] and have catalyst lifetimes >100 times longer than their modern forms.[67] Because reconstructed sequences are a representation of historical sequences, the fraction of functional proteins in an ASR library is high while also providing a highly diverse library in terms of properties. Additionally, a reconstructed ancestral sequence is likely not as dependent on the cellular processes of the host and often have properties described as an ancestral generalist versus the modern specialist enzyme.[65] One notable recent example of applying ASR to protein engineering was a study by Lin *et al.* where they engineered Rubisco, a notoriously hard enzyme to engineer. Among 98 predicted ancestors in their library, 34 had superior catalytic efficiency to wild type modern Rubisco, showcasing the power of ASR.[68]

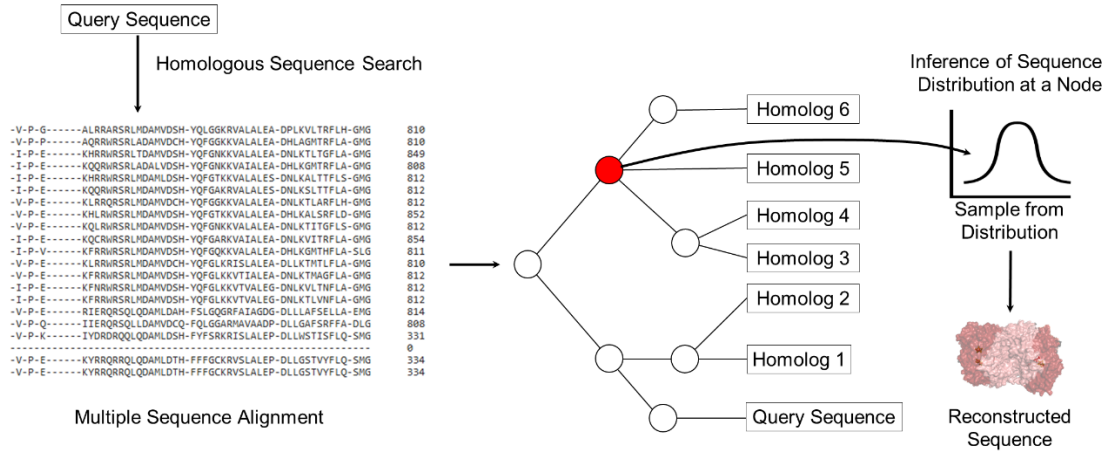


Figure 3.7: Overview of ancestral sequence reconstruction (ASR). Ancestral nodes are inferred through statistical methods and reconstructed ancestors can be created by sampling from the distribution in those nodes. (Adapted from Spence et al.[67])

As mentioned above, ASR has produced catalysts with lifetimes >100 times longer than their modern forms[67] (likely required due to the harsher environmental conditions found on a younger planet), and produced more efficient and more "generalized" enzyme sequences.[65] We believe the properties that ASR is able to extract will be well suited to engineering nitrogenase for a bioelectrocatalytic application. Nitrogenase is an ancient enzyme, dating from before the advent of photosynthesis,[21] for which many sequences have been deposited in databases (~90,000 on Uniprot),[61] making this an ideal candidate for the application of ASR. Our goal is to screen an ASR-based library for improved thermostability and catalyst lifetime. Additionally, a reconstructed ancestral nitrogenase is likely not as dependent on the cellular processes of the host and could be more amenable to engineering for higher faraday efficiency in a bioelectrocatalytic reactor.

We first used FireProtASR to perform ASR on NifD alone and selected two ancestors of nitrogenase to be experimentally validated.[70] Since FireProtASR does cannot manage a concatenated sequence to generate ancestors, we decided to generate NifD/NifK pairs manually. We assembled a dataset of concatenated NifD/NifK sequence homologs using jackhmmer and then formed an MSA using ClustalW.[71,72] The MSA created in ClustalW was then passed to RAxML for ancestral sequence reconstruction.[73] Unfortunately, the publication by Garcia *et al*. in 2023 outlines a very similar methodology for ASR of NifD/NifK pairs.[74] In addition to our own ancestors, we will also test the sequences from Garcia *et al.* as potential engineering starting points.

In addition to ASR, we were interested in the performance of generative protein language models on generating a diverse set of nitrogenase homologs to use as potential engineering starting points. Protein language models (PLMs) are a subset of natural language processing (NLP) models, where tokens (representations of either

words or a collection of words) are represented by amino acids. The aim of a PLM is to learn the relationship between amino acids when they form proteins.[75] The learning of this relationship is then used to infer certain properties of proteins such as its function,[76] phylogeny,[77] or structure.[78] Another common application of PLMs is the generation of protein sequences, in fact generation of protein sequences is built into the training of the model itself. While there are myriad different model architectures, two main training objectives exist to infer the sequence of a protein, masked token prediction and next token prediction (Figure 3.8, See Machine Learning in Protein Engineering). These models have been trained on large protein sequence datasets and are not directly capable of generating specific proteins without nudging them into the right direction. This nudging is performed via fine-tuning, where the model is re-trained with a very low learning rate on a family or protein sequence one wishes to generate homologous members of.[79]

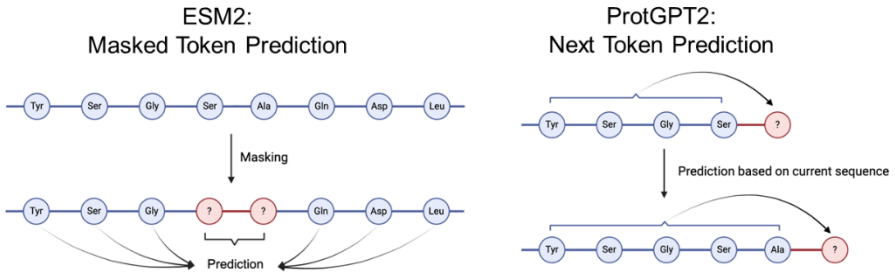


Figure 3.8: Main differences in training objectives for ESM2 and ProtGPT2 sequence generation tasks. ESM2 sequence generation works via masked token prediction, where random tokens (i.e. amino acid residues) are masked for the model to predict. The model takes the surrounding context into account when making a prediction on a token. ProtGPT2 sequence generation works via next token prediction where a protein is constructed sequentially and only the context of previous residues are considered for the prediction of the next one.

Evolutionary Scale Modelling 2 (ESM2) is a PLM based on the transformer architecture trained on 250 million sequence clusters from UniRef50 with up to15 billion parameters.[78] The first iteration of ESM was published in 2019 with the objective of learning about protein sequences by scaling-up models in terms of parameter size to capture the evolutionary relationships between protein sequences.[80] It is trained as a masked language model where amino acid residues are masked from the model during training and the model has to predict the amino acid at a position given its sequence context. ESM2 is a very powerful model, capable of capturing the function of proteins, the way they fold, and is capable of generating a sequence based on a particular fold.[78]

Similarly, ProtGPT2 is also a transformer model trained on 45 million sequence clusters from UniRef50, using 738 million parameters.[81] While ESM2 was developed for the general task of understanding proteins, ProtGPT2 was trained for the purpose of protein sequence design. This is also reflected in the next token prediction tasked

performed during training, which is similar to the task that ChatGPT is trained on for example.[82] In next token prediction, the model always predicts the next amino acid in the protein sequence, using only the context of the previously generated residues to determine the nature of the next one. Compared to ESM2, ProtGPT2 lacks the global context on protein sequence families, but is very good at generating sequences that are similar but still distant from the sequences in the dataset.

We fine-tuned both ESM2 and ProtGPT2 with two different datasets, one obtained via a BLAST search and the other via a jackhmmer search of NifD, resulting in four separate NifD sequence generation models. The jackhmmer search resulted in 23315 sequences for fine-tuning, where sequences can be from any homolog of NifD.[71] The BLAST search was very specific toward only NifD sequences from different organisms, resulting in a more uniform dataset in terms of the nature of the sequences, but a lot smaller with only 85 sequences.[83] After fine-tuning the models, we generated 5000 sequences in each model to be processed further. We then performed a pairwise alignment of each generated sequence vs NifD to determine how similar they are to the seed sequence. Next, we filtered the sequences for the existence of certain residues. The first filter determines whether all catalytic residues of NifD are present. For the second filter we analyzed the residues on the interface of NifD and NifK in PyMol and picked-out the residues that are within 3Å of each other. Sequences that missed those interface residues were removed from the generated sequence list. Lastly, the sequences' structures were predicted via AlphaFold2 and structurally aligned against *A. vinelandii* NifD (Figure 3.9A).[62]

None of the sequences generated by ESM2 had any of the catalytic residues and could not be considered homologs of NifD. ProtGPT2 fine-tuned using the jackhmmer dataset generated 287 sequences that passed the catalytic residue filter, but none that passed the NifD/K interface filter. ProtGPT2 fine-tunes using the BLAST dataset was a lot more successful, with 4483 sequences passing the first filter and 6 passing the second filter. We show the structural alignment of the 3 of those sequence in Figure 3.9B.

With these sequences generated and the structure prediction being similar to *A. vinelandii* NifD, we will be transforming these variants into *A. vinelandii* using the genetic system described in Figure 3.4 to determine whether the variants are functional NifD variants. We would then further explore the properties of these variants, especially for their bioelectrocatalytic efficiency.

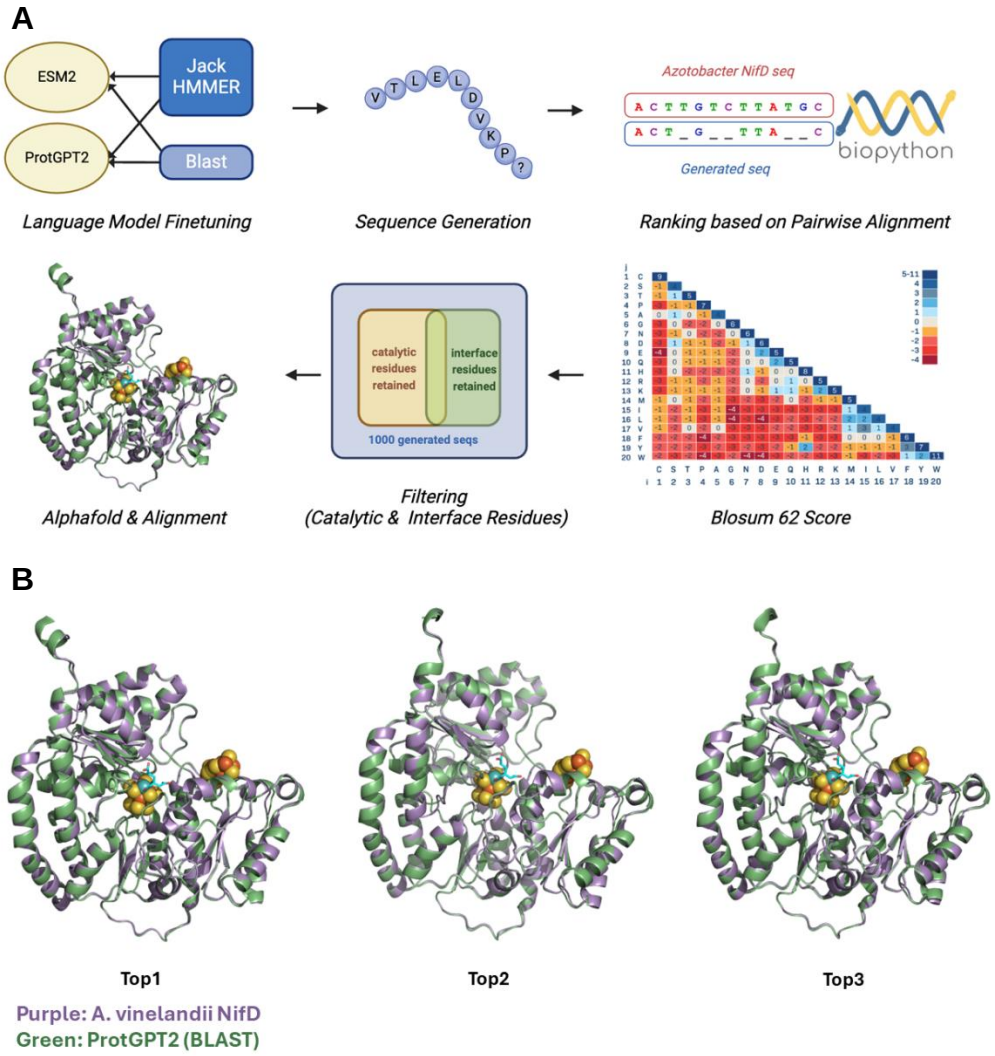


Figure 3.9: NifD PLM generation workflow and results. (**A**) Workflow of NifD PLM sequence generation. We first fine-tuned ESM2 and ProtGPT2 using two different datasets from jackhmmer and BLAST. We generated 5000 sequences using the fine-tuned models and ranked them based on their pairwise sequence alignment using Blosum62 as the substitution matrix. We then filtered the sequences based on the existence of important catalytic and interface residues, before finally performing a structure prediction and aligning against *A. vinelandii* NifD. (**B**) Results of the top 3 variants generated via ProtGPT2 fine-tunes on the BLAST dataset.

Producing a Longer-Lasting Catalyst

A main hurdle to overcome to make any green ammonia production process economically viable is increasing the catalyst lifetime. This metric often remains unreported in initial studies of sustainable ammonia production but is nevertheless a crucial property of the catalyst.[84] The Minteer lab has not reported the lifetime of nitrogenase in their reactor, but it is at best one day[10] (and at worst one hour[36]). From a practical perspective, increasing the catalyst lifetime will decrease the capital expenditure by reducing the amount of maintenance on a functional reactor as well as reduce the logistical issues that would come with regularly having to supply highly distributed reactor sites with new catalyst. One problem that many biocatalysts face is that nature's solution to short enzyme lifetime is often to manage it by continuous replacement.[85] Understanding the mechanisms of catalyst inactivation can be used as the basis to mitigate the problem. Nitrogenase has a few known mechanisms of permanent inactivation: permanent deactivation by contact with oxygen,[26] a pH- and turnover-dependent permanent deactivation mechanism,[28] and thermally induced denaturation. Oxygen dependent deactivation is not very well understood, but researchers have prevented this mechanism by manipulating nitrogenase using anaerobic techniques. Similar to other electrocatalytic ammonia production methods, hydrogen can be used to remove oxygen from the reactor.[9] One can imagine using the hydrogen produced by nitrogenase to keep the reactor atmosphere anaerobic, minimizing or eliminating usage of the gas as an additional consumable.

Increasing the thermostability of enzymes is often seen as an imperative to making a biocatalyst successful for two main reasons: (1) According to the Arrhenius equation, the reaction rate increases exponentially with a rise in temperature which is often summarized as the Q_{10} rule for enzymatic processes – the reaction rate doubles for every 10°C increase in temperature.[86] (2) Throughout evolution, proteins are usually only marginally more thermostable than is required by their environmental conditions. Naturally evolving proteins often accumulate functionally neutral but thermostabilizing mutations in order to tolerate a functionally beneficial mutation that is destabilizing.[87] For protein engineering this means that more thermostable proteins are often more evolvable because they can tolerate more mutations to produce a protein with higher fitness. Jesse Bloom summarized this in the concept of "protein stability promotes evolvability".[88] Therefore, many methods have been developed to recognize thermostabilizing mutations in a given protein, especially computational methods that reduce the cost and time expenditure of the process. Using metrics such as crystallographic B-factors, force-fields from MD simulations, machine learning based approaches, and/or protein design software such as TRIAD or Rosetta, leaves us with many options to consider for this aim.[87] As nitrogenase's structure has been well-studied in multiple organisms and in a wide range of conditions, the protein is well suited for thermostabilization using *in silico* methods.[26]

We aim to utilize these methods to increase the stability of *A. vinelandii* nitrogenase. In addition to providing us with a more stable catalyst with likely a longer lifetime, a thermostabilized nitrogenase might also allow us to increase the space-time yield by increasing the power input of the bioelectrocatalytic reactor, as increased power usually increases reactor temperatures. Thermostabilization software can provide us with a library of variants that can be measured for using differential scanning fluorimetry (DSF). DSF assays such as Thermofluor have been used in a high-throughput fashion to obtain approximate apparent protein melting temperatures (T_M) by using fluorophores that are activated when binding to the hydrophobic cores of unfolded proteins.[89] The main factor limiting throughput is often protein purification, but the need for only a small amount of protein in these assays enables the use of high-throughput protein purification methods utilizing, for example, nickel-NTA binding in a 96-well plate format. Increasing the thermostability of a protein could reduce the flexibility of the protein and possibly prevent catalytically required conformational changes. Given this, we'll need to make sure that thermostabilization does not impede nitrogenase reactivity. Data from thermostability and enzymatic activity could also provide insights into the necessary flexibility/rigidity of the nitrogenase complex to further advance our understanding of the enzyme.[87]

Lastly, we will tackle the irreversible pH- (pH > 8.6) and turnover-dependent deactivation of nitrogenase first reported by Pham and Burgess in 1993.[90] This deactivation mechanism prevented efforts to generate highly reduced MoFe-protein for structural and biochemical studies by lowering the presence of protons in the buffer. A subsequent study by the Rees lab has shown that this was not a simple denaturation of the protein due to high pH as is often observed with proteins, but rather a consequence of a complex turnover dependent mechanism. Even at the catalytically optimal pH of 7.8, this inactivation is observable, albeit slowly.[28] The study further revealed that the hydrodynamic radius of the protein increased, however they could not spectroscopically observe any changes to the metal cofactor environments. Recently, the Rees lab has solved the structure of the pH-inactivated nitrogenase complex, revealing a striking similarity to *apo*-nitrogenase, with the C-terminus of the protein being unstructured.[91]

We aim to use computational protein design tools to stabilize nitrogenase against the unfolding of the C-terminus in an effort to prevent turnover and pH dependent deactivation of nitrogenase. As opposed to purely thermostabilizing nitrogenase, in this approach we need to measure the deactivation of the protein under turnover conditions. A pH-stabilized variant would also allow us to determine the ideal pH of the system, which to our knowledge has not been performed by the Minteer lab and their system pH varies from study to study.

This part of the project, with its focus on deactivation mechanisms of nitrogenase, will serve a dual purpose. By providing stabilized variants of nitrogenase, we will not only likely increase the catalyst lifetime making our ammonia production approach more economically competitive but could also provide the scientific community with

new tools to probe the mechanisms of nitrogenase catalysis. Our initial goal is to improve catalyst lifetime to one week. The later benefit of this aim will have wide reaching consequences for the future of nitrogenase. A better understanding of nitrogenase's inner workings and a stabilized enzyme will also help plant biotechnology in the quest of expressing nitrogenase in the roots of plants in order to avoid nitrogen fertilization altogether for certain crops.

Materials and Methods

Materials

Solvents and buffer salts were purchased from Sigma Aldrich, Koptec, Fisher Bioreagents and Merck. Media for bacterial expression were purchased from Merck Millipore. We used Invitrogen Mix&Go DH5α for plasmid. *A. vinelandii* wild type was given to us as a gift from the lab of Doug Rees. *A. vinelandii* ΔNifD/K and the pDJ33 plasmid was given to us as a gift from the lab of Dennis Dean. Oligonucleotides and genes were synthesized by Integrated DNA Technologies. PCR reactions were performed on an Eppendorf Mastercycler Gradient. Plasmids were purified using the Zymo Research Zyppy Miniprep Kit. Agarose gel electrophoresis for DNA fragment purification was performed at 90V using TAE as buffer and 1% agarose gels (Merck Millipore). Gel excisions from agarose gel DNA electrophoresis were processed using the Zymoclean Gel DNA recovery kit. DNA gels were referenced against the Goldbio 1kb DNA ladder. DNA concentrations were measured with a Nanodrop ND-1000. 96-well growth assays were performed in Corning® 96-well plates. Growth was monitored on an Agilent Biotek Epoch 2 plate reader.

Burk's Medium Recipe

Burk's medium is prepared by making two separate, autoclaved solutions that are only combined when ready to grow cells. The first solution is a phosphate buffer made up of 0.2g/L KH_2PO_4, 0.8g/L K_2HPO_4. The second solution is made up of 200g/L sucrose, 2g/L $MgSO_4 \cdot 7H_2O$, 0.9g/L $CaCl_2 \cdot 2H_2O$, 10μM $Na_2MoO_4 \cdot H_2O$, 50mg/L $FeSO_4 \cdot 7H_2O$. The two solutions are mixed in a 9:1 ratio of the first solution to the second solution.

Iron deficient Burk's medium is prepared the same way, but without 10μM $Na_2MoO_4 \cdot H_2O$, 50mg/L $FeSO_4 \cdot 7H_2O$ in the second solution.

Burk's medium agar plates are prepared by mixing the first solution with 16g/L agar prior to sterilization.

Burk's medium (N+) is made by adding 1.5g/L NH_4OAc to the second solution prior to sterilization.

Transformation of A. vinelandii

From a glycerol stock, *A. vinelandii* is struck out on Burk's medium (N+) agar plates and grown at 30°C for 2-3 days until colonies are visible. Then a colony is struck out on iron deficient Burk's medium agar plates and passaged three times every 2-3 days. Then a colony is used to inoculate a liquid culture of iron deficient Burk's medium and grown at 30°C, 170rpm overnight. Cultures should be visibly green at this point, indicating that the cells are now competent. *A. vinelandii* is then transformed by adding 1μg of pDJ33 plasmid containing NifD/NifK and leaving shaking for 5

minutes. The culture is struck out on selection plates being either Burk's medium (N-) agar plates or Burk's medium agar plates with added kanamycin.

Diazotrophic Preselection Assay

Freshly transformed *A. vinelandii* cells or cells previously passaged through kanamycin plates are struck out on Burk's medium (N-) agar plates and left to grow for 60 hours before counting cells and imaging.

96-well Growth Assay

Freshly transformed *A. vinelandii* cells are first grown in Burk's medium (N+) overnight, or until reaching an OD_{600} of at least 1, at 30°C and 200rpm. Then cells are spun down at 500xg for 5 minutes and resuspended in Burk's medium (N-) twice to get rid of excess ammonium in the medium. Cells are then diluted to an OD_{600}=0.05 before adding to the wells of a 96-well plate. Cell growth is then monitored on an Agilent Biotek Epoch 2 plate reader in incubator mode at 30°C for 72 hours.

References

1. Jewess, M. & Crabtree, R. H. Electrocatalytic Nitrogen Fixation for Distributed Fertilizer Production? ACS Sustainable Chem. Eng. 4, 5855–5858 (2016).
2. Ritter, S. The Haber-Bosch Reaction: An Early Chemical Impact On Sustainability. Chemical & Engineering News https://cen.acs.org/articles/86/i33/Haber-Bosch-Reaction-Early-Chemical.html.
3. Milton, R. D. et al. Bioelectrochemical Haber–Bosch Process: An Ammonia-Producing H2/N2 Fuel Cell. Angewandte Chemie International Edition 56, 2680–2683 (2017).
4. Renner, J. N., Greenlee, L. F., Ayres, K. E. & Herring, A. M. Electrochemical Synthesis of Ammonia: A Low Pressure, Low Temperature Approach. Interface magazine 24, 51–57 (2015).
5. Wang, L. et al. Greening Ammonia toward the Solar Ammonia Refinery. Joule 2, 1055–1074 (2018).
6. Piercing the haze. https://www.science.org/doi/10.1126/science.361.6407.1060?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed.
7. Ritter, S. Nabbing nitrogen from the air to make fertilizer on the farm. https://cen.acs.org/articles/95/i18/Nabbing-nitrogen-from-the-air-to-make-fertilizer-on-the-farm.html.
8. Global ammonia annual production capacity. Statista https://www.statista.com/statistics/1065865/ammonia-production-capacity-globally/.
9. Ghavam, S., Vahdati, M., Wilson, I. A. G. & Styring, P. Sustainable Ammonia Production Processes. Frontiers in Energy Research 9, (2021).
10. Boerner, L. Industrial ammonia production emits more CO2 than any other chemical-making reaction. Chemists want to change that. Chemical & Engineering News https://cen.acs.org/environment/green-chemistry/Industrial-ammonia-production-emits-CO2/97/i24.
11. Ammonia in agriculture: The engine of plant growth. thyssenkrupp https://www.thyssenkrupp.com/en/stories/sustainability-and-climate-protection/ammonia-in-agriculture:-the-engine-of-plant-growth.
12. Chen, H. et al. Fundamentals, Applications, and Future Directions of Bioelectrocatalysis. Chem. Rev. 120, 12903–12993 (2020).
13. Ma, R. et al. Mitigation potential of global ammonia emissions and related health impacts in the trade network. Nat Commun 12, 6308 (2021).
14. Liu, L. et al. Exploring global changes in agricultural ammonia emissions and their contribution to nitrogen deposition since 1980. Proceedings of the National Academy of Sciences 119, e2121998119 (2022).

15. New report: Oil and gas phase-out primer. International Institute for Sustainable Development https://www.iisd.org/articles/press-release/new-report-oil-gas-phase-out-primer.
16. The Paris Agreement | UNFCCC. https://unfccc.int/process-and-meetings/the-paris-agreement.
17. The Evolution of Smokeless Powder. U.S. Naval Institute https://www.usni.org/magazines/proceedings/1904/april/evolution-smokeless-powder (1904).
18. Zeldovich, L. A History of Human Waste as Fertilizer. JSTOR Daily https://daily.jstor.org/a-history-of-human-waste-as-fertilizer/ (2019).
19. Roush, G. A. Strategic Mineral Supplies 12. Nitrogen. The Military Engineer 29, 444–449 (1937).
20. Mann, C. C. 1491: New Revelations of the Americas before Columbus. (Knopf, New York, 2005).
21. Garvin, J., Buick, R., Anbar, A. D., Arnold, G. L. & Kaufman, A. J. Isotopic Evidence for an Aerobic Nitrogen Cycle in the Latest Archean. Science 323, 1045–1048 (2009).
22. War of the Pacific. Wikipedia (2024).
23. Smil, V. Enriching the Earth: Fritz Haber, Carl Bosch, and the Transformation of World Food Production. (The MIT Press, 2004).
24. Burén, S. & Rubio, L. M. State of the art in eukaryotic nitrogenase engineering. FEMS Microbiol Lett 365, fnx274 (2018).
25. Rees, D. C. & Howard, J. B. Nitrogenase: standing at the crossroads. Current Opinion in Chemical Biology 4, 559–566 (2000).
26. Einsle, O. & Rees, D. C. Structural Enzymology of Nitrogenase Enzymes. Chem Rev 120, 4969–5004 (2020).
27. Spatzal, T. et al. Evidence for Interstitial Carbon in Nitrogenase FeMo Cofactor. Science 334, 940–940 (2011).
28. Yang, K.-Y., Haynes, C. A., Spatzal, T., Rees, D. C. & Howard, J. B. Turnover-Dependent Inactivation of the Nitrogenase MoFe-Protein at High pH. Biochemistry 53, 333–343 (2014).
29. Kyriakou, V., Garagounis, I., Vourros, A., Vasileiou, E. & Stoukides, M. An Electrochemical Haber-Bosch Process. Joule 4, 142–158 (2020).
30. McEnaney, J. M. et al. Ammonia synthesis from N2 and H2O using a lithium cycling electrification strategy at atmospheric pressure. Energy Environ. Sci. 10, 1621–1630 (2017).
31. Chen, C., Liu, Y. & Yao, Y. Ammonia Synthesis via Electrochemical Nitrogen Reduction Reaction on Iron Molybdate under Ambient Conditions. European Journal of Inorganic Chemistry 2020, 3236–3241 (2020).
32. Kordali, V., Kyriacou, G. & Lambrou, C. Electrochemical synthesis of ammonia at atmospheric pressure and low temperature in a solid polymer electrolyte cell. Chem. Commun. 1673–1674 (2000) doi:10.1039/B004885M.

33. Cui, B. et al. Electrochemical synthesis of ammonia directly from N2 and water over iron-based catalysts supported on activated carbon. Green Chem. 19, 298–304 (2017).
34. Manjunatha, R., Karajić, A., Goldstein, V. & Schechter, A. Electrochemical Ammonia Generation Directly from Nitrogen and Air Using an Iron-Oxide/Titania-Based Catalyst at Ambient Conditions. ACS Appl. Mater. Interfaces 11, 7981–7989 (2019).
35. Roth, L. E., Nguyen, J. C. & Tezcan, F. A. ATP- and Iron−Protein-Independent Activation of Nitrogenase Catalysis by Light. J. Am. Chem. Soc. 132, 13672–13674 (2010).
36. Lee, Y. S., Yuan, M., Cai, R., Lim, K. & Minteer, S. D. Nitrogenase Bioelectrocatalysis: ATP-Independent Ammonia Production Using a Redox Polymer/MoFe Protein System. ACS Catal. 10, 6854–6861 (2020).
37. Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. Angewandte Chemie International Edition 57, 4143–4148 (2018).
38. Chen, H., Dong, F. & Minteer, S. D. The progress and outlook of bioelectrocatalysis for the production of chemicals, fuels and materials. Nat Catal 3, 225–244 (2020).
39. USDA ERS - Farming and Farm Income. https://www.ers.usda.gov/data-products/ag-and-food-statistics-charting-the-essentials/farming-and-farm-income/.
40. USDA. Highlights Farms and Farmland. nass.usda.gov https://www.nass.usda.gov/Publications/Highlights/2014/Highlights_Farms_and_Farmland.pdf (2014).
41. Davidson, D. Nitrogen Math: Simple Calculations Give You the Right Rates. DTN Progressive Farmer https://www.dtnpf.com/agriculture/web/ag/crops/article/2016/03/21/nitrogen-math-simple-calculations.
42. Climate. Cabot Chamber of Commerce https://www.cabotcc.org/climate/.
43. Agapakis, C. Scaling Bioprocesses. Ginkgo Bioworks https://www.ginkgobioworks.com/our-work/scaling-bioprocesses/.
44. Lee, Y. S. et al. Electroenzymatic Nitrogen Fixation Using a MoFe Protein System Immobilized in an Organic Redox Polymer. Angewandte Chemie International Edition 59, 16511–16516 (2020).
45. Schindelin, H., Kisker, C., Schlessman, J. L., Howard, J. B. & Rees, D. C. Structure of ADP·AIF4–-stabilized nitrogenase complex and its implications for signal transduction. Nature 387, 370–376 (1997).
46. Natural Gas Prices Only Account for 15% of Run-Up in Anhydrous Ammonia Prices, Shows New Texas A&M Study. AgWeb https://www.agweb.com/news/crops/corn/natural-gas-prices-only-account-15-run-anhydrous-ammonia-prices-shows-new-texas-am (2022).
47. Ibendahl, G. Fertilizer Price Outlook for 2024. (2024).
48. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. Nat Rev Genet 16, 379–394 (2015).

49. Miller, D. C., Athavale, S. V. & Arnold, F. H. Combining chemistry and protein engineering for new-to-nature biocatalysis. Nat Synth 1, 18–23 (2022).
50. Tetter, S. et al. Evolution of a virus-like architecture and packaging mechanism in a repurposed bacterial protein. Science 372, 1220–1224 (2021).
51. Longwell, C. K., Labanieh, L. & Cochran, J. R. High-throughput screening technologies for enzyme engineering. Current Opinion in Biotechnology 48, 196–202 (2017).
52. Schmidt-Dannert, C. & Arnold, F. H. Directed evolution of industrial enzymes. Trends in Biotechnology 17, 135–136 (1999).
53. Noar, J. D. & Bruno-Bárcena, J. M. Y. 2018. Azotobacter vinelandii: the source of 100 years of discoveries and many more to come. Microbiology 164, 421–436.
54. Nagpal, P., Jafri, S., Reddy, M. A. & Das, H. K. Multiple chromosomes of Azotobacter vinelandii. J Bacteriol 171, 3133–3138 (1989).
55. Brigle, K. E. et al. Site-directed mutagenesis of the nitrogenase MoFe protein of Azotobacter vinelandii. Proc. Natl. Acad. Sci. U.S.A. 84, 7066–7069 (1987).
56. Carruthers, B. M., Garcia, A. K., Rivier, A. & Kacar, B. Automated Laboratory Growth Assessment and Maintenance of Azotobacter vinelandii. Current Protocols 1, e57 (2021).
57. Zhang, L. et al. Directed Evolution of a Bacterial Laccase (CueO) for Enzymatic Biofuel Cells. Angewandte Chemie International Edition 58, 4562–4565 (2019).
58. Hickey, D. P. et al. Establishing a Thermodynamic Landscape for the Active Site of Mo-Dependent Nitrogenase. J. Am. Chem. Soc. 141, 17150–17157 (2019).
59. Dos Santos, P. C. Genomic Manipulations of the Diazotroph Azotobacter vinelandii. Methods Mol Biol 1876, 91–109 (2019).
60. Barnes, C. O. et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. Nature 588, 682–687 (2020).
61. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49, D480–D489 (2021).
62. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
63. Bhattacharya, S., Roche, R., Shuvo, M. H. & Bhattacharya, D. Recent Advances in Protein Homology Detection Propelled by Inter-Residue Interaction Map Threading. Frontiers in Molecular Biosciences 8, (2021).
64. Chatzou, M. et al. Multiple sequence alignment modeling: methods and applications. Briefings in Bioinformatics 17, 1009–1023 (2016).
65. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J. Ancestral sequence reconstruction for protein engineers. Current Opinion in Structural Biology 69, 131–141 (2021).

66. Furukawa, R., Toma, W., Yamazaki, K. & Akanuma, S. Ancestral sequence reconstruction produces thermally stable enzymes with mesophilic enzyme-like catalytic properties. Sci Rep 10, 15493 (2020).
67. Gumulya, Y. et al. Engineering highly functional thermostable proteins using ancestral sequence reconstruction. Nat Catal 1, 878–888 (2018).
68. Lin, M. T., Salihovic, H., Clark, F. K. & Hanson, M. R. Improving the efficiency of Rubisco by resurrecting its ancestors in the family Solanaceae. Science Advances 8, eabm6871 (2022).
69. Hie, B. L., Yang, K. K. & Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. cels 13, 274-285.e6 (2022).
70. Khan, R. T., Musil, M., Stourac, J., Damborsky, J. & Bednar, D. Fully Automated Ancestral Sequence Reconstruction using FireProtASR. Current Protocols 1, e30 (2021).
71. jackhmmer search | HMMER. https://www.ebi.ac.uk/Tools/hmmer/search/jackhmmer.
72. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7, 539 (2011).
73. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).
74. Garcia, A. K. et al. Nitrogenase resurrection and the evolution of a singular enzymatic mechanism. eLife 12, e85003 (2023).
75. Ruffolo, J. A. & Madani, A. Designing proteins with language models. Nat Biotechnol 42, 200–202 (2024).
76. Bepler, T. & Berger, B. Learning the Protein Language: Evolution, Structure and Function. Cell Syst 12, 654-669.e3 (2021).
77. Ziegler, C., Martin, J., Sinner, C. & Morcos, F. Latent generative landscapes as maps of functional diversity in protein sequence space. Nat Commun 14, 2222 (2023).
78. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023).
79. Schmirler, R., Heinzinger, M. & Rost, B. Fine-tuning protein language models boosts predictions across diverse tasks. 2023.12.13.571462 Preprint at https://doi.org/10.1101/2023.12.13.571462 (2023).
80. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 118, e2016239118 (2021).
81. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun 13, 4348 (2022).
82. Ray, P. P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems 3, 121–154 (2023).

83. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J Mol Biol 215, 403–410 (1990).
84. Scott, S. L. A Matter of Life(time) and Death. ACS Catal. 8, 8597–8599 (2018).
85. Hanson, A. D. et al. The number of catalytic cycles in an enzyme's lifetime and why it matters to metabolic engineering. Proceedings of the National Academy of Sciences 118, e2023348118 (2021).
86. Sterratt, D. C. Q10: the Effect of Temperature on Ion Channel Kinetics. in Encyclopedia of Computational Neuroscience (eds. Jaeger, D. & Jung, R.) 2551–2552 (Springer, New York, NY, 2015). doi:10.1007/978-1-4614-6675-8_236.
87. Modarres, H. P., Mofrad, M. R. & Sanati-Nezhad, A. Protein thermostability engineering. RSC Adv. 6, 115252–115270 (2016).
88. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. Proceedings of the National Academy of Sciences 103, 5869–5874 (2006).
89. Lavinder, J. J., Hari, S. B., Sullivan, B. J. & Magliery, T. J. High-Throughput Thermal Scanning: A General, Rapid Dye-Binding Thermal Shift Screen for Protein Engineering. J. Am. Chem. Soc. 131, 3794–3795 (2009).
90. Pham, D. N. & Burgess, B. K. Nitrogenase reactivity: Effects of pH on substrate reduction and carbon monoxide inhibition. Biochemistry 32, 13725–13731 (1993).
91. Warmack, R. A. et al. Structural consequences of turnover-induced homocitrate loss in nitrogenase. Nat Commun 14, 1091 (2023).

Chapter 4

DEVELOPMENT AND SCALE-UP OF A CATALYZED FORMATION OF *CIS*-TRIFLUOROMETHYL-SUBSTITUTED CYCLOPROPANES USING PROTOGLOBINS

L. Schaus, A. Das, A. M. Knight, G. Jimenez-Osés, K. N. Houk, M. Garcia-Borràs, F. H. Arnold, X. Huang, *Angew. Chem. Int. Ed.* 2023, 62, e202208936; *Angew. Chem.* **2023**, *135*, e202208936. doi: 10.1002/ange.202208936

Abstract

Trifluoromethyl-substituted cyclopropanes (CF_3-CPAs) constitute an important class of compounds for drug discovery. While several methods have been developed for synthesis of *trans*-CF_3-CPAs, stereoselective production of corresponding *cis*-diastereomers remains a formidable challenge. We report a biocatalyst for diastereo- and enantio-selective synthesis of *cis*-CF_3-CPAs with activity on a variety of alkenes. We found that an engineered protoglobin from *Aeropyrnum pernix* (*Ape*Pgb) can catalyze this unusual reaction at preparative scale with low-to-excellent yield (6–79%) and enantioselectivity (17–99% ee). Computational studies revealed that the steric environment in the active site of the protoglobin forced iron-carbenoid and substrates to adopt a pro-*cis* near-attack conformation. This work demonstrates the capability of enzyme catalysts to tackle challenging chemistry problems and provides a powerful means to expand the structural diversity of CF_3-CPAs for drug discovery.

Contributions

The project was designed and managed by Prof. Xiongyi Huang and Prof. Marc Garcia-Borràs. Starting variant discovery and directed evolution was performed by Dr. Anders Knight and Prof. Xiongyi Huang. Substrate scope, reaction scale-up, wet-lab mechanistic studies, transfers of mutations, and compound characterization was performed by Lucas Schaus with the help of Dr. Anuvab Das, in the lab of Prof. Frances Arnold. All computational studies were performed by Prof. Marc Garcia-Borràs and Dr. Gonzalo Jimenez-Osés in the lab of Prof. Ken Houk. Writing was done by Lucas Schaus with the help of Dr. Anuvab Das, Prof. Xiongyi Huang and Prof. Marc Garcia-Borràs.

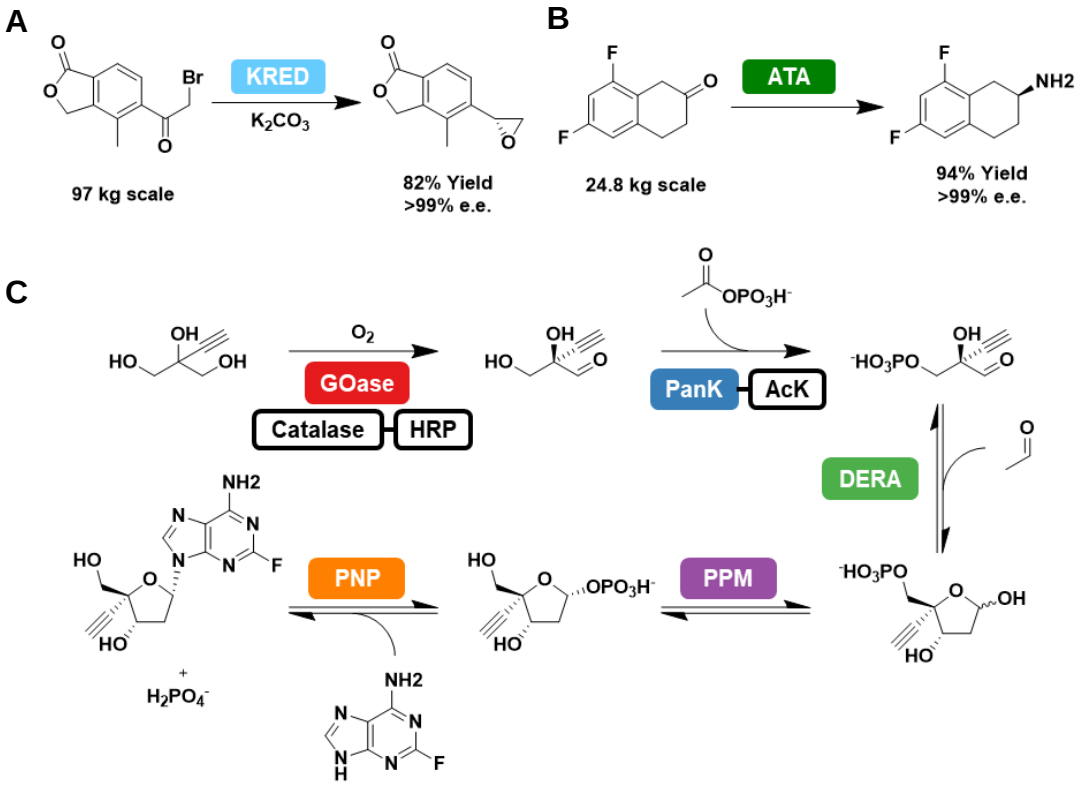
Biocatalysis in the Pharmaceutical Industry

Natural products and natural product derivatives represent 38% of all FDA-approved drugs.[1,2] While in nature, the molecular scaffolds that can help us fight infections or suppress coughing, are synthesized by cascades of enzymatic reactions, most of the drugs on the market are produced via non-biocatalytic methods. Medicinal chemists spend considerable time developing and testing multiple reaction steps to produce complex molecular scaffolds that enzymes can make in a only a few steps.[3,4]

On paper, biocatalytic strategies have many advantages over traditional chemical methods. Enzymes have excellent regio- and stereoselectivity allowing us to access functionalizations that are very challenging with conventional catalysts.[5–9] They can achieve diffusion-limited reaction rates and work under mild conditions which combined with their selectivity could allow the omission of protecting groups in the synthesis of a drug.[10,11] Enzyme promiscuity and their ability to be expressed from DNA can be exploited to optimize and modify the properties of a biocatalyst through directed evolution and other protein engineering techniques.[12] There are also significant environmental advantages to biocatalysis. The mild reaction conditions, mean that there is less energy expenditure. The production of the catalysts in an organism means they can be produced sustainably with lower carbon footprints.[10,11] Conventional transition metal catalysts can sometimes rival the stereoselectivity of enzymes, but the mining practices that provide these metals are often unsustainable and unethical.[13]

Despite these numerous advantages, adoption of biocatalytic processes pales in comparison to conventional catalysts. In a publication, Matthew Truppo outlines key factors that are responsible for the slow adoption of biocatalysts.[10] Access to biocatalysts is a challenge that can be seen as three related problems, the ability to obtain enzymes, the ability to use enzymes and range of reactions possible with accessible enzymes. For some categories of enzymes, these problems have been solved by companies such as Codexis, which sell screening kits for commonly used enzyme classes such as ketoreductases (KRED) and transaminases (ATA).[14,15] Access to biocatalysts is however far more limited when more bespoke reactions are required, especially for new-to-nature chemistry. Protein engineering lead times pose another problem. Even when a suitable enzyme for a reaction has been found, engineering through directed evolution is often too slow for the expected lead times in pharmaceutical industries.[10] Recent advances in computationally assisted protein engineering techniques, such as machine learning, have helped to shorten engineering lead times and may be a contributing factor to the increased adoption in the past decade.[16–20] The last key factor in the adoption of biocatalysis for the production of small molecule drugs is reaction scale-up. While many advances for biocatalytic reaction scale-up have been made in the past decade, it is still the case that processes that may look interesting in the laboratory can take a long time to get to a scale that is useful for industrial applications. One key problem is the necessity to use larger reactors in which the occurrence of nonideal mixing can produce

inconsistent reagent concentration, pH, or temperatures, which in turn can degrade the catalyst.[10,21]

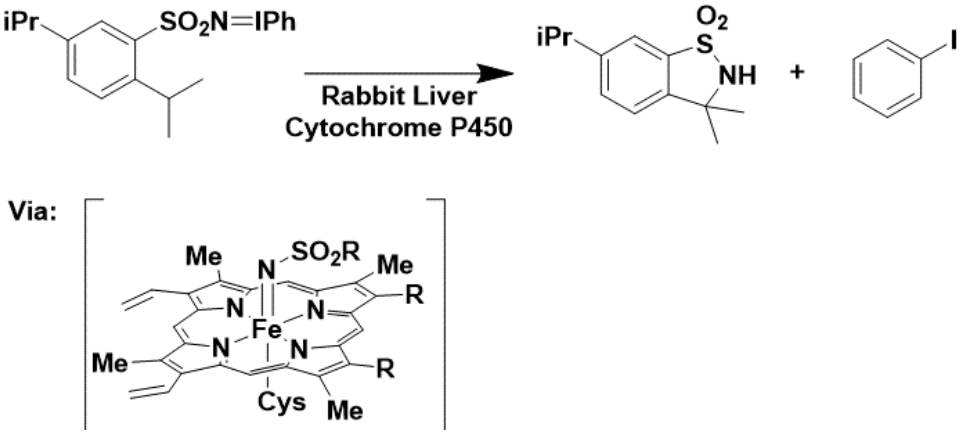


Scheme 4.1: Examples of biocatalytic reactions used in the synthesis of small molecule drugs. (**A**) Example of large-scale usage of a ketoreductases in the synthesis of an intermediate to produce a renal outer medullary potassium channel inhibitor candidate by Merck.[22] (**B**) Example of a large-scale usage of a transaminase in the synthesis of an intermediate of a gamma secretase inhibitor by Pfizer.[23] (**C**) Enzymatic cascade synthesis of Islatravir. GOase and PanK-AcK-fusion are immobilized via a polyhistidine tag. Conversion of sucrose and the phosphate product to glucose-1-phosphate not shown. This last step pulls the equilibrium towards the Islatravir product.[24] KRED: Ketoreductase; ATA: Transaminase; GOase: Galactose oxidase; HRP: Horseradish peroxidase; PanK: Pantothenate kinase; AcK: Acetate kinase; DERA: Deoxyribose 5-Phosphate Aldolase; PPM: Phosphopentomutase; PNP: Purine nucleoside phosphorylase.

Nevertheless, the past decade has seen considerable growth in the usage of biocatalytic processes to produce small molecule drugs. This can be largely contributed to advances in the first key point Matthew Truppo described about access to biocatalysts. Nowadays there are many off-the-shelf enzyme kits that are commercially available that help in rapidly finding a starting point for protein engineering campaigns. To the best of our knowledge, the most common enzymes that are used in kilogram-scale biocatalytic reactions are KREDs and ATAs, both of which catalyze reactions found in nature (Scheme 4.1A-B).[11,22,23] A notable recent example of the usage of enzymes for the production of a small molecule drug is the synthesis of Islatravir, an HIV reverse transcriptase translocation inhibitor (Scheme

4.1C).[24] The drug was produced via an enzymatic cascade, using nine enzymes in a six-step reaction, requiring no intermediate isolation. Even more impressive is the fact that no protective groups are used, and only a single stereoisomer of Islatravir is produced with 51% yield. Part of the selectivity comes from the excellent enantiomeric excess produced by the galactose oxidase (GOase) and the kinetic selectivity of the pantothenate kinase (PanK) towards the (R)-aldehyde intermediate. This example showcases the truly transformative potential that biocatalysis can have in the production of small molecule drugs.

The above enzymatic reactions showcase some of the advantages of using enzymes



Scheme 4.2: New-to-nature enzyme catalyzed reaction first discovered by Svastis *et al*. Rabbit liver cytochrome P450 catalyzes the insertion of a nitrene into a C-H bond on the same molecule, via an iron-nitrene intermediate that forms in the active site of the enzyme.

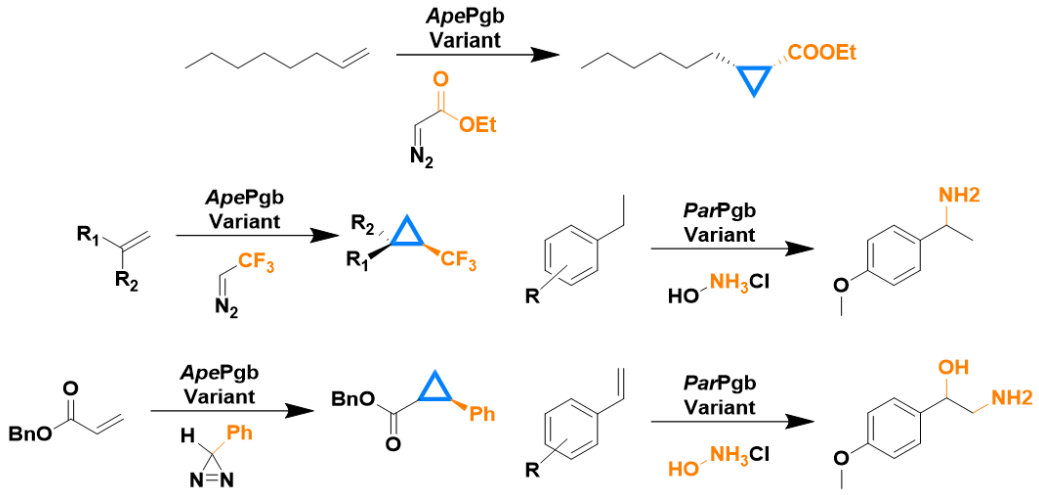
instead of conventional catalysts to achieve high selectivity and high yield. Despite these chemists require ever evolving tools to tackle new synthetic problems.[25] However, many commercially available enzymes catalyze reactions that are known in nature, which only provides a narrow reaction scope available to biocatalysis. In 1985 Svastits and his co-workers demonstrated that a rabbit liver cytochrome P450 could activate a nitrene precursor and catalyze a C-H amination reaction, showing that enzymes could be used for reactions that are new-to-nature (**Error! Reference source not found.**).[26] Followed by the practical application of John Maynard Smith's theoretical work on protein evolution by Frances Arnold in the 1990s, the gates were opened to explore a new reaction space catalyzed by engineered enzymes.[12,27] Since then, enzymes have been engineered to catalyze carbene addition across alkenes and alkynes,[7,28–32] nitrene C-H insertions,[5,33–36] Diels-Alder reactions,[37,38] and Kemp eliminations to name a few.[39,40] While much work has been done to engineer proteins to be capable of catalyzing new-to-nature reactions, in order to be truly useful in industry, protein engineers need to keep the key points of Matthew Truppo's perspective in mind. Ideally, the engineered proteins should be shelf-stable or even lyophilizable. The reactions should be as easy to setup as those using a conventional

small molecule catalyst. Lastly, protein engineers should demonstrate that the reactions work at a larger scale then the often-seen μmol-scale.[30,33,41]

Protoglobins

Much of the Arnold lab's pursuit to engineer enzymes for new-to-nature reactions focused on cytochromes P450 and cytochromes *c* as the parent enzyme.[25,42] The lab has shown that these enzymes can be evolved to catalyze a wide range of reactions. In his thesis work, Dr. Anders Knight wanted to explore the world of metalloenzymes beyond cytochromes.[29] Inspired by the concepts of evolvability described in a paper by Jesse Bloom,[43] Anders Knight was looking for iron-heme cofactor proteins from thermophiles. One of the enzymes that showed a very broad substrate scope for the cyclopropanation reaction of unactivated alkenes was a protoglobin from *Aeropyrnum pernix*.[29,30]

Protoglobins are iron-heme cofactor proteins found in Archaea with unknown biological function. The little that is known about them is limited to the protoglobin of *Methanosarcina acetivorans*. That protoglobin has been shown to bind the gases O_2, CO_2, and NO reversibly *in vitro*, as well as cyanide, azide and imidazole.[44,45] It is also observed to having a low O_2 dissociation constant. It has been hypothesized that protoglobins are the extant predecessors to globin-coupled sensors (GCS) and may be closely related to the globin of the last universal common ancestor (LUCA) due to its promiscuity in gas binding and the high oxygen sensitivity.[46,47] Regardless of their natural function, the two properties of high promiscuity and stability at high temperatures make them excellent candidates to explore and engineer non-natural functions.



Scheme 4.3: New-to-nature chemistry catalyzed by protoglobin variants discovered in the past five years. *Ape*Pgb: *Aeropyrnum pernix* protoglobin; *Par*Pgb: *Pyrobaculum arsenaticum* protoglobin.

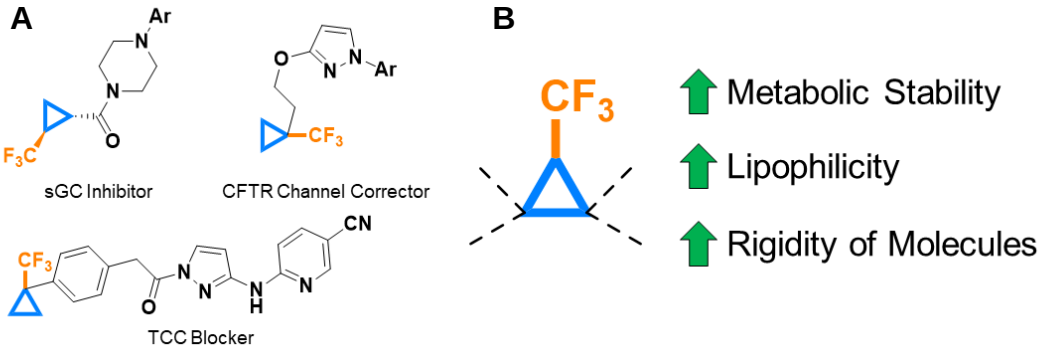
Since Anders Knight's first discovery of a new-to-nature reaction catalyzed by protoglobins,[30] these proteins have found a lot of success in protein engineering campaigns in the Arnold lab. In the past five years, protoglobins have been shown to catalyze carbene mediated cyclopropanations using diazo carbene precursors as well

as diazirine carbene precursors,[28,30,41] C-H primary amination and aminohydroxylation (Scheme 4.3).[33] In addition, as described in this work, protoglobins can be lyophilized and remain functional,[28] as opposed to cytochromes, and used at larger scale than has been shown with cytochromes.

Organofluorines

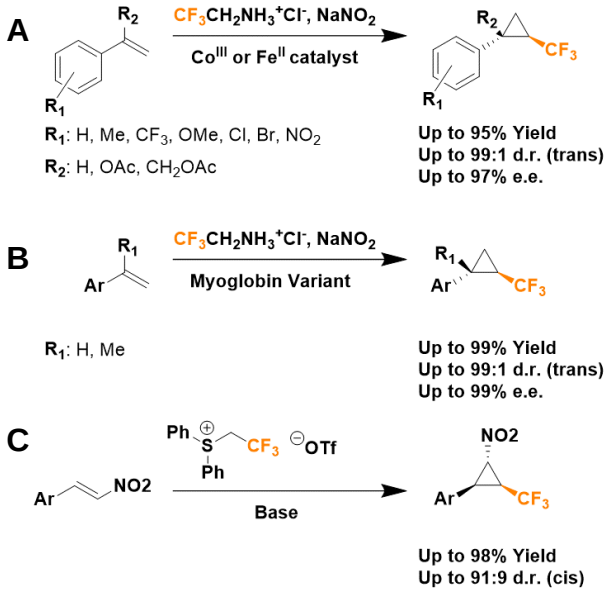
Fluorine-containing molecules (organofluorines) have become one of the most important classes of compounds in medicinal chemistry. Recent reports state that 20% of marketed drugs contain fluorine, a number that has only grown over the years versus conventional drug molecules.[48,49] This growing success rate of organofluorines suggests that these drug candidates can minimize the risk of unsuccessful drug trials.[48,50] Fluorine moieties can provide many desirable properties to a drug candidate. The C–F bond is one of the strongest bonds a carbon can form, significantly increasing metabolic stability. The electronegativity of fluorine leads to bond polarization which can shift the lipophilicity/hydrophobicity of a compound.[48,49] Fluorine can act as a weak hydrogen bond acceptor through its σ-hole,[51,52] which together with the large Van-der-Waals radius makes it a good carbonyl isostere.[49] Bioisosterism is an important concept in medicinal chemistry, which allows functional groups with similar shapes to be interchanged without inducing a large change in the group's biological behaviour.[53] As an example, trifluoromethyl-substituted cyclopropanes can act as a tert-butyl bioisostere to improve the bioavailability and metabolic stability of drug compounds, making them highly valuable for the pharmaceutical industry (Scheme 4.4).[54–56] Drug development is a challenging, time-consuming and expensive process, on the order of $1 billion per marketed drug. Thus, providing medicinal chemists with the tools to access new organofluorines can have a significant impact on the success of drug development.

Among various organofluorines for pharmaceutical development, trifluoromethyl-substituted cyclopropanes (CF_3-CPAs) have assumed a privileged position, as they combine the conformational rigidity of cyclopropanes and desirable medicinal properties of trifluoromethyl groups in one moiety.[55,57]



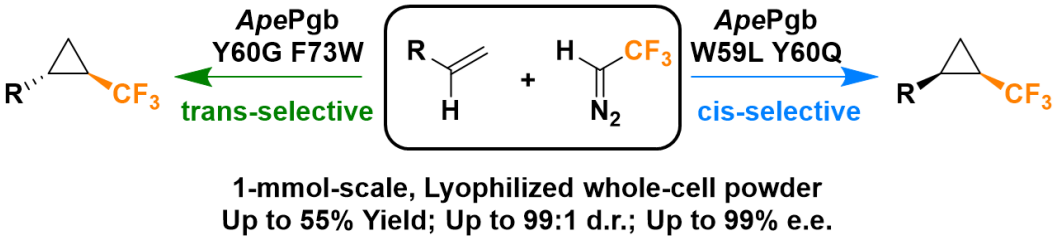
Scheme 4.4: Examples of CF_3-CPAs in the pharmaceutical industry. (**A**) Bioactive CF_3-CPA examples from left to right: Class of soluble guanylate cyclase (sGC) inhibitors by Bayer outlined in a recent patent (WO 2022/122916 Al); Bamocaftor, a cystic fibrosis transmembrane regulator (CFTR) channel corrector by Vertex Pharmaceuticals, currently in clinical trials (doi: 10.1056/NEJMoa1807119); Apinocaltamide, a T-type calcium channel (TCC) blocker by Idorsia Pharmaceuticals, currently in clinical trials (doi: 10.1111/epi.14732). (**B**) Set of property changes that are desired in drug candidates when including CF_3-CPAs in molecules.

Because of the prevalence of CF_3-CPAs in medicinal chemistry, there are continuous efforts in developing catalytic methods for enantioselective synthesis of CF_3-CPAs (Scheme 4.5).[58–60] Despite considerable progress, a few challenges remain. One is the limited substrate scope, as most methods developed so far have focused on cyclopropanation of styrenes and arene-substituted alkenes.[55,58,59,61–67] Furthermore, the majority of current catalytic methods only produced the *trans*-diastereomer of CF_3-CPAs, which is often thermodynamically and kinetically favored.[60,68–70] Accessing *cis*-CF_3-CPAs is considerably harder, most likely due to an additional steric challenge posed in the pro-*cis* transition state.[60] The only method reported to selectively synthesize the *cis*-CF_3-CPA used a Corey-Chaykovski reaction on highly electron-deficient *β*-nitrostyrenes;[63] it is neither catalytic nor enantioselective. Gaining access to *cis*-trifluoromethyl-substituted cyclopropanes will be valuable for full exploitation of CF_3-CPAs for drug development, as it is known that molecular topologies of *cis*- and *trans*-cyclopropanes are quite different and often lead to drastic differences in their biological properties.[71]



Scheme 4.5: Overview of strategies for the synthesis of CF_3-CPAs. (**A**) Transition-metal catalyzed strategies first reported by Le Maux *et al*.[72] in 2006 and expanded by Morandi *et al*.[58] in 2011. These first synthetic strategies were limited to aryl-substituted alkenes and showed strong preference for the trans-product across all substrates. (**B**) Enzyme-catalyzed strategy reported by Tinoco *et al*. in 2017.[59] While an overall improvement to the previous strategies in terms of yield and selectivity, the strategy is more limited in substrate scope and in scale (0.15-mmol reactions). (**C**) First report of a cis-selective synthesis of CF_3-CPAs by Hock *et al*. in 2017.[63] This strategy uses a Corey-Chaykovsky on highly electron-deficient β-nitrostyrenes. The reaction is neither catalytic nor stereoselective.

In this work, we present a method to synthesize *cis*-trifluoromethyl-substituted cyclopropanes using new laboratory-evolved variants of *Aeropyrnum pernix* and *Methanosarcina acetivorans* protoglobins (denoted as *Ape*Pgb and *Ma*Pgb respectively). Previous work showed that *Ape*Pgb could be engineered to catalyze the cyclopropanation of unactivated alkenes using ethyl diazoacetate, yielding corresponding *cis*-cyclopropanes.[30] Thus, we speculated that protoglobins can be evolved to catalyze the challenging synthesis of *cis*-CF_3-CPAs despite the fact that iron-porphyrin catalysts overwhelmingly produce the *trans* products.[58,72] The motivation to choose protoglobins to explore this reaction came from the fact that it is an iron-heme cofactor containing protein that is mainly found in thermophiles. Properties of proteins from thermophilic organisms make them more evolvable and, we hypothesized, more amenable to scale-ups. This approach enabled us to scale up the reaction to 1 mmol using lyophilized whole-cell powders, which can be used without cell-culture experience.



Scheme 4.6: In this chapter we will present a scalable biocatalytic synthetic strategy to form *trans*- and *cis*-CF_3-CPAs. While lower in overall yield to previously reported approaches, the substrate scope is much more expansive, going beyond aryl-substituted alkenes and includes unactivated-, electron-deficient, and electron-rich alkenes.

Engineering *Ape*Pgb for CF_3-CPA Reactions

We screened several of variants of protoglobin catalysts, previously engineered by Dr. Anders Knight, for initial activity for production *cis*-CF_3-CPAs. For screening conditions, we devised a method that would enable sufficiently high screening throughput during both initial activity search and consecutive enzyme engineering steps (**Error! Reference source not found.**A). A 96-deep-well plate screening assay was enabled by the preparation of solutions of trifluorodiazoethane via a protocol first reported by Gilman *et al.* (Figure 4.1B, see Materials and Methods for details).[73] This approach has been demonstrated to enable sufficiently high screening throughput during both initial activity search and consecutive enzyme engineering steps.[74] Through this initial screening, we found that wild-type *Ape*Pgb can catalyze the formation of *cis*-CF_3-CPA **1** with a total turnover number (TTN) of 110 (Figure 4.1**Error! Reference source not found.**C).

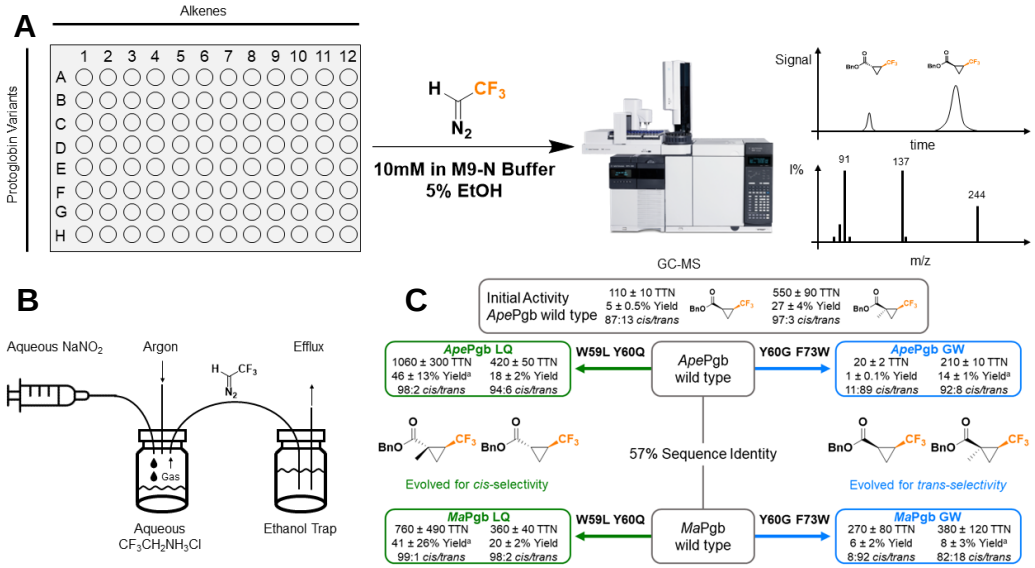


Figure 4.1: Variant screening strategy and results for the enzymatic formation of *cis*-CF_3-CPAs. (**A**) Initial screening for activity was conducted in 96-well plates with each row containing *E. coli* cells expressing a protoglobin variant at OD_{600}=30 in M9-N buffer. To this we added a carbene-precursor solution trapped in ethanol, as well as the respective alkene for each column, and ran the reaction for 16 hours. Results were analyzed on GCMS to confirm formation of the desired product at a favorable diastereomeric ratio. (**B**) Reaction setup for the formation of carbene-precursor trapped in ethanol as outlined in Gilman *et al.* (**C**) Results of reaction screening and directed evolution of *Ape*Pgb While evolving for the formation of trans-**1**, we discovered that the mutations to produce *trans*-CF_3-CPAs do not change diastereoselectivity for compound **2**, preferably forming the *cis* product. TTN: Total Turnover Number; [a] The major diastereomer is compound **2** as detailed in the compound characterization section.

We next carried out site-saturation mutagenesis (SSM) and screening on four sites including W59, Y60, F73, and F145. These four sites were chosen based on their proximity to the heme center. Previous studies also showed that they were important for regulating the stereoselectivity of carbene-transfer reactions catalyzed *Ape*Pgb.[30]

Through this engineering, we were able to improve the TTN of *Ape*Pgb for this abiological reaction to 420 (3.8-fold improvement) by introducing two mutations, W59L and Y60Q. This is the first example of a catalytic method that can selectively produce *cis*-CF_3-CPAs from alkenes. The *Ape*Pgb W59L Y60Q (*Ape*Pgb LQ) variant showed activity on a broad range of olefins including electron-rich and electron-deficient styrenes, unactivated alkenes, and heteroatom-substituted alkenes. The diastereo- and enantioselectivities were moderate to excellent for most tested substrates (Figure 4.2). The yields ranged from 6% to 55% in a 1-mmol-scale reaction using lyophilized powder of whole *Escherichia coli* cells expressing the *Ape*Pgb LQ variant. We also tested the reactions with clarified cell lysates. However, the yields dropped significantly presumably due to the instability of protein in cell lysates under gas bubbling conditions. Products **1**, **2,** and **3** were particularly interesting since neither the *cis* nor the *trans* forms of these compounds have been synthesized previously. And the successful production of **2** and **3** showed that this method could be used to synthesize quaternary chiral centers.

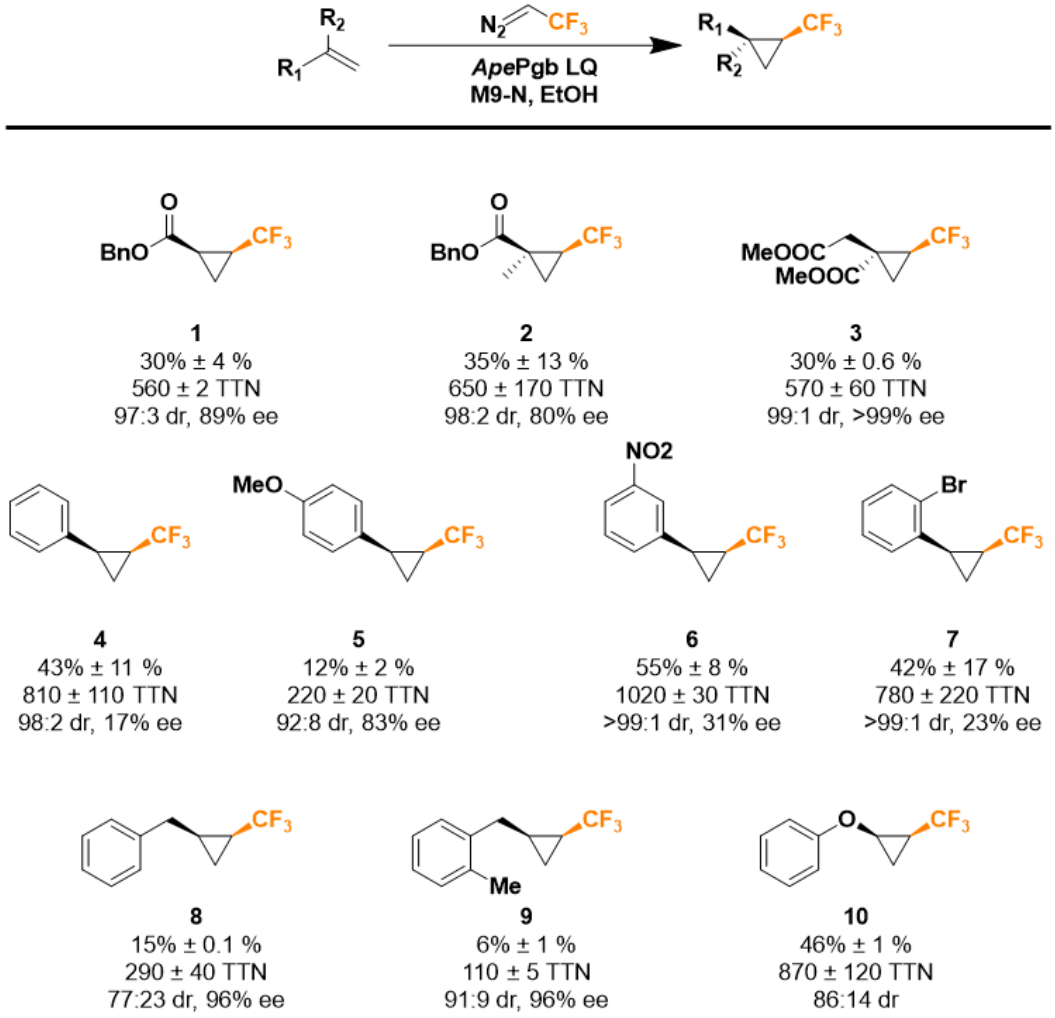
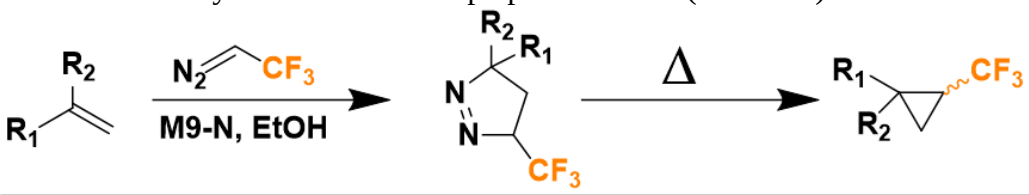


Figure 4.2: Substrate scope of the cyclopropanation reaction of trifluorodiazoethane with an alkene, catalyzed by *ApePgb* LQ. Yields, diastereomeric ratios (dr), enantiomeric excess (ee), and TTNs of the reactions are reported. Yields are reported as analytical yields measured via ^{19}F-NMR against 4-fluoroacetophenone of known concentration. Reactions were run at 1-mmol scale with lyophilized whole-cell powder at OD_{600} = 45 in M9-N buffer with 2.5% ethanol as co-solvent. Absolute configurations were not determined. We were not able to determine the ee of **10** due to difficulties in separation on chiral GC-FID

Scale-Up of Cis-Selective CF_3-CPA Formation

When we initially scaled the reaction from analytical to preparative scale (1 mmol), we observed an undesired 3+2 cycloaddition side reaction that was negligible in reactions at analytical scale. This cycloaddition is detrimental since the product is known to undergo a thermal contraction reaction to afford the corresponding CF_3-CPAs with no diastereo- and enantio-control (Table 4.1).[55,75] To investigate this, we performed a time-dependent analysis of the reaction outcome, which showed that *Ap*Pgb LQ catalyzed the cyclopropanation reaction rapidly (Figure 4.3). The desired *cis* product was formed within 20 minutes, and the yield was maximized at 55% at 4-µmol scale. The most interesting observation was the rapid accumulation of the 3+2 cycloaddition product in the initial five minutes of the reaction. We reasoned that this phenomenon indicated that the formation of cyclization products was favored when concentrations of alkenes and trifluorodiazoethane were high. Therefore, we resorted to *ex situ* generation of trifluorodiazoethane[59] and a slow addition of the alkene into the reaction mixture. This greatly improved the yield and diastereoselectivity of the reactions at preparative scale (Table 4.1).



Enzyme/Control	Substrate	Slow Addition	Yield (*cis* product)	d.r. (*cis*:*trans*)
*Ape*Pgb W59L, Y60Q (LQ)	Benzyl acrylate	No	6% ± 3%	88:12
M9-N	Benzyl acrylate	No	NR	ND
*Ape*Pgb W59L, Y60Q (LQ)	Benzyl methacrylate	No	56% ± 25%[a]	57:43
M9-N	Benzyl methacrylate	No	23% ± 0.01%[a]	50:50
*Ape*Pgb W59L, Y60Q (LQ)	Dimethyl itaconate	No	14% ± 1.5%[b]	92:8
M9-N	Dimethyl itaconate	No	3 ± 0.003%[b]	41:59
*Ape*Pgb W59L, Y60Q (LQ)	Benzyl acrylate	Yes	30% ± 4%	97:3
*Ape*Pgb W59L, Y60Q (LQ)	Benzyl methacrylate	Yes	35% ± 13%[a]	98:2
*Ape*Pgb W59L, Y60Q (LQ)	Dimethyl itaconate	Yes	30% ± 0.7%	99:1

Table 4.1 Proposed mechanism for loss of diastereoselectivity on preparative scale reactions. Reactions were carried out in preparative-scale format with lyophilized whole-cell powder. The slow-addition column indicates whether the alkene was added via slow addition, 5 minutes after generation of the trifluorodiazoethane gas. Yields for "slow addition: no" are reported as analytical yields calculated from a calibration curve on a GC-FID. Yields for "slow addition: yes" are reported as analytical yields from ^{19}F-NMR. For benzyl methacrylate, "*cis*" product refers to the diastereomer in which CF_3 and benzyl ester groups reside on the same side of the cyclopropane ring (compound **2**). For dimethyl itaconate, "*cis*" product refers to the diastereomer in which CF_3 and methyl ester groups reside on the opposite side of the cyclopropane ring (compound **3**). [a] The major diastereomer is compound **2** as detailed in the compound characterization section.

[b] The major diastereomer is compound **3** as detailed in the compound characterization section. NR: no reaction product detected. ND: not determined.

During the evolution of the *Ape*Pgb LQ variant, we learned that mutations at position F73 can dramatically influence the diastereoselectivity of *Ape*Pgb catalysts. We performed further rounds of SSM at residues W59, Y60, and F73, resulting in the discovery of an *Ape*Pgb Y60G F73W (GW) variant which selectively catalyzed the formation of the *trans* product for the benzyl acrylate model substrate (Figure 4.1C). To further expand the synthetic utility of this catalytic system, we tested whether the key LQ and GW mutations identified for *Ape*Pgb could be transferred to the protoglobin from *Methanosarcina acetivorans* (*Ma*Pgb, 57% sequence identity to *Ape*Pgb) and conserve catalytic properties. The choice of *Ma*Pgb, as opposed to other known protoglobins, was motivated by the fact that, at the time, an experimentally determined structure for this enzyme existed,[76,77] as opposed to *Ape*Pgb for which a structure was experimentally determined after completion of this project.[41] This allowed us to study the unique ability of protoglobins to produce *cis* conformations on cyclopropanes *in silico*. In all cases, *Ma*Pgb LQ and *Ma*Pgb GW achieved the diastereoselectivity observed with the *Ape*Pgb variants. One exception was benzyl methacrylate substrate, with which the *Ma*Pgb GW variant showed notably reduced preference for generating product **2** compared to its *Ape*Pgb counterpart (Figure 4.1). These results demonstrate the utility of protoglobins for stereoselective synthesis of CF_3-CPAs.

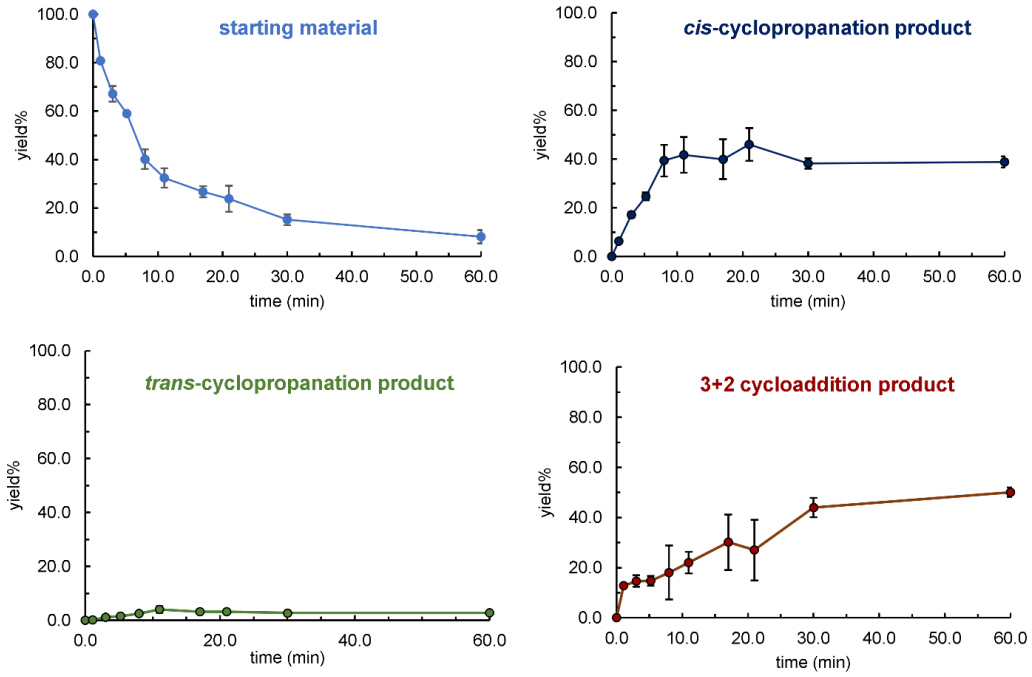


Figure 4.3: Time trace of the cyclopropanation reaction using *Ape*Pgb LQ. The reactions were setup on 4-µmol scale in 400 µL M9-N + *Ape*Pgb LQ (OD_{600} = 30). The reactions were extracted with

hexanes:ethyl acetate and measured on GC-FID. The depletion of the starting material stabilized in 30 min with most of the desired product synthesized after 20 min.

Investigating Divergent Selectivities

Next, we explored the origins of the divergent selectivity observed in these engineered protoglobins. Because of the high structural and sequence similarity between *Ma*Pgb (PDB: **2VEB**)[44,76,77] and *Ape*Pgb,[41] we based our *in-silico* studies on *Ma*Pgb variants since there were more structural data available for *Ma*Pgb. We first performed density functional theory (DFT) calculations on a truncated computational model to evaluate the intrinsic reactivity of histidine-ligated iron-heme for CF_3-cyclopropanation in a protein-free environment (see Materials and Methods: Molecular dynamics simulations). Our results revealed that a radical stepwise mechanism was likely, due to the presence of the strong electron-withdrawing CF_3 group on the iron-carbene intermediate and the electron-deficient character of the alkene substrates (Figure 4.4, see Appendix D Figures 1-2 for details), which was consistent with previous computational studies on truncated models of related heme-protein carbene systems.[70] Our transition-state analysis revealed that formation of the first C–C bond in the cyclopropane ring (TS1) was intrinsically favored for generating the *trans*-diastereomer over the *cis*-isomer for both benzyl acrylate ($\Delta\Delta G^{\ddagger}$ = 2.6 kcal·mol^{-1}) and benzyl methacrylate ($\Delta\Delta G^{\ddagger}$ = 2.0 kcal·mol^{-1}) substrates (see Appendix D Figure 3 for details). These calculations were in line with previous experimental studies which showed that iron-porphyrin catalysts overwhelmingly produced *trans* products for cyclopropanation.[64,72] These results were also consistent with previous computational studies in which *trans*-cyclopropanation was found to be the lowest-energy pathway and was preferred over the *cis*-cycloproanation pathway by about 1.7 kcal·mol^{-1}.[68,69] Our computational data also indicated that after generation of the radical intermediate after the first C−C bond formation, no stereo-scrambling is likely to occur prior to a fast second C–C bond formation step to generate the cyclopropane ring (Figure 4.4).[70] Overall, these results suggested that, for cyclopropanation mediated by a histidine-ligated heme-carbene complex in free solution, the formation of *trans*-CF_3-CPA products would be intrinsically favored. Therefore, the active-site environment of the proteins must play a crucial role in overcoming this intrinsic barrier and redirecting the reaction to selectively form *cis*-CF_3-CPAs.

To further investigate this, we first modelled the iron-carbene intermediate in the active sites of wild-type *Ma*Pgb and the *Ma*Pgb LQ and GW variants using molecular dynamics (MD) simulations (Figure 4.5, see Materials and Methods for details).[78–81] MD simulations showed that the iron-carbenoid explored two major conformations in the *Ma*Pgb active sites, whose geometric features were influenced directly by the mutations introduced. In wild-type *Ma*Pgb and *Ma*Pgb LQ, which are *cis*-selective, the iron-carbenoids mainly explore orientations with a ∠N-Fe-C-C(CF_3) dihedral angle around -25º and +50º, respectively, whereas the preferred conformer for the carbenoid in the *trans*-selective *Ma*Pgb GW variant is described by a ∠N-Fe-C-C(CF_3) angle of around -100º and -140º (Figure 4.5A-B). Under these iron-carbenoid conformations, assuming a similar binding pose for the substrate in the active site of different protoglobin variants (see Appendix D Figure 4 for details), wild-type

*Ma*Pgb and variant *Ma*Pgb LQ would preferentially lead to the *cis*-diastereomer and variant *Ma*Pgb GW would mainly afford the *trans*-diastereomer. These results suggest that the introduced mutations change the geometric constraints in the active site and switch iron-carbenoid orientation, which ultimately controlled the diastereoselectivity of the reaction (Figure 4.5C-D).

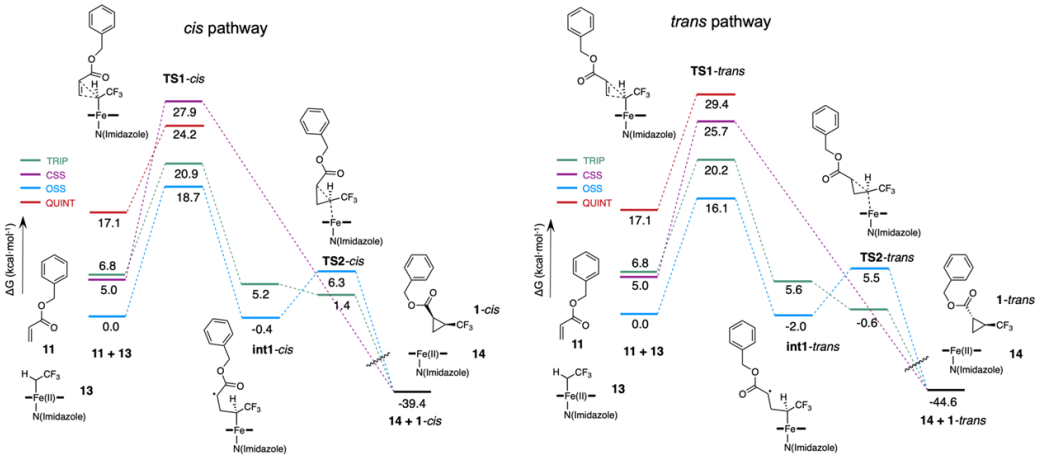


Figure 4.4: DFT-optimized intrinsic reaction profiles for the *trans* and *cis*- CF_3-cyclopropanation of benzyl acrylate for the two possible diastereomers (*cis* and *trans*). A computational truncated model was used. Calculations were performed at the uB3LYP-D3BJ / Def2-TZVP (PCM=DiethylEther) // uB3LYP / 6-31G(d)+SDD(Fe) (PCM=DiethylEther) level of theory. Three different electronic states (open-shell singlet, OSS; closed-shell singlet , CSS; and triplet) were considered. For TS1, the higher in energy quintet electronic state was also considered. Gibbs free energies are given in kcal·mol^{-1}.

Under this mechanistic scheme, however, there is one exception. *Ma*Pgb GW preferentially produces the *cis* diastereomer in the reaction with benzyl methacrylate (Figure 4.1C). To further explore the origins of opposite selectivities offered by *Ma*Pgb LQ and GW variants, we performed restrained MD simulations with both substrates bound in LQ and GW active sites. Starting from structures corresponding to the preferred carbenoid-bound major conformers in LQ and GW *Ma*Pgb variants, benzyl acrylate (**11**) and benzyl methacrylate (**12**) were docked into the active site. These structures were then used to start 500 ns MD trajectories in which the substrates were restrained at distances between 2.5–3.5 Å from the iron-carbenoid in order to analyze accessible near-attack conformations (NAC) that they could explore with respect to the iron-carbenoid and to avoid exploring unbinding events (see Materials and Methodsfor details).[82] To study the relative orientations of the alkene and iron-carbenoid, geometric parameters based on two dihedral angles were defined which allowed us to characterize the pro-*cis*/*trans* character of the NACs explored by the substrate and the carbene in each variant-substrate pair along the MD trajectories (Figure 4.6, relative orientation of the alkene is defined by the orange dihedral angle, and the relative orientation of the iron-carbenoid by the blue dihedral angle). These simulations revealed that when both substrates (benzyl acrylate **11**, and

benzyl methacrylate **12** in Figure 4.6) were bound in the *Ma*Pgb LQ active site, they mainly explored a major near-attack conformation with respect to the carbenoid, which corresponded to a *cis*-selective configuration (Figure 4.6A-B, dihedral values ca. +130º (alkene, in orange) and -90º (carbenoid, in blue)).

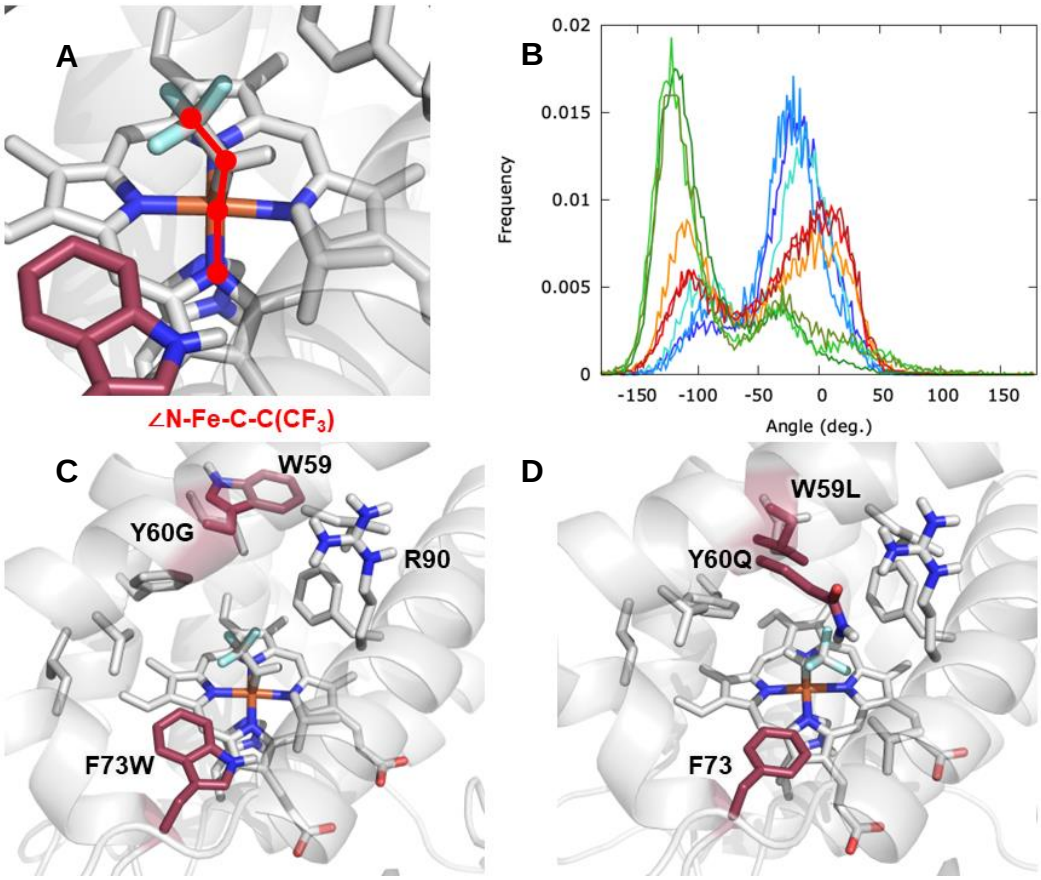


Figure 4.5: Computational modelling based on MD simulations to model the iron-carbenoid formed in the active site of wild-type *Ma*Pgb and *Ma*Pgb LQ and *Ma*Pgb GW variants. (**A**) Active site of *Ma*Pgb GW with the carbenoid bound to heme describing the ∠N-Fe-C-C(CF3) dihedral angle measured along the MD trajectories. This angle, highlighted in red, describes the relative orientation of the iron-carbene in the active site. (**B**) Histogram describing the dihedral angles explored by the carbenoid relative to heme along MD simulations. Three independent MD replicas of 500 ns each are conducted for each system, shown in three color shades: 3 red tones for wild-type *Ma*Pgb, 3 blue tones for *Ma*Pgb LQ, and 3 green tones for *Ma*Pgb GW (see SI for complete data). Representative snapshots of the major orientation explored in (**C**) *Ma*Pgb GW (∠N-Fe-C-C(CF3) = -117°); and in (**D**) *Ma*Pgb LQ (∠N-Fe-C-C(CF3) = +49°).

This was due to the preorganization of the iron-carbenoid intermediate in the active site and the steric requirements applied to the substrate when placed in the binding pocket in a catalytically competent pose. In contrast, *Ma*Pgb GW preferentially bound the benzyl acrylate substrate in a slightly different catalytic pose to that in the LQ variant, whereas the iron-carbenoid was rotated, as previously observed in carbene-bound simulations (Figure 4.6A-B, dihedral values ca. +130º (alkene, in

orange) and +90º (carbenoid, in blue)). This alternative near-attack conformation led to the preferential formation of the *trans*-diastereomer, in line with the experimentally observed selectivity switch. Consequently, we propose that different orientations explored by the iron-carbenoid in the active sites of the LQ and GW *Ma*Pgb variants are responsible for controlling the selectivity of these reactions and overcoming the intrinsic electronic preferences to yield almost exclusively the *cis*-diastereomer.

Finally, simulations with the benzyl methacrylate substrate bound in the *Ma*Pgb GW variant described a preferential near-attack conformation that led to the *cis*-cyclopropane product (Figure 4.6A-B, dihedral values ca. +130º (alkene, in orange) and -90º (carbenoid, in blue)). These simulations showed that due to the extra steric bulk of the methyl group at the alkene α-position and the *Ma*Pgb GW active-site environment, the iron-carbenoid was forced to rotate when the olefin approaches, as compared to the benzyl acrylate system. Consequently, the iron-carbenoid and benzyl methacrylate preferentially explored near-attack pro-*cis* conformations due to steric requirements in the GW variant active site. These results illustrate why *Ma*Pgb GW was not able to produce the *trans*-cyclopropane in the case of benzyl methacrylate (Figure 4.6A).

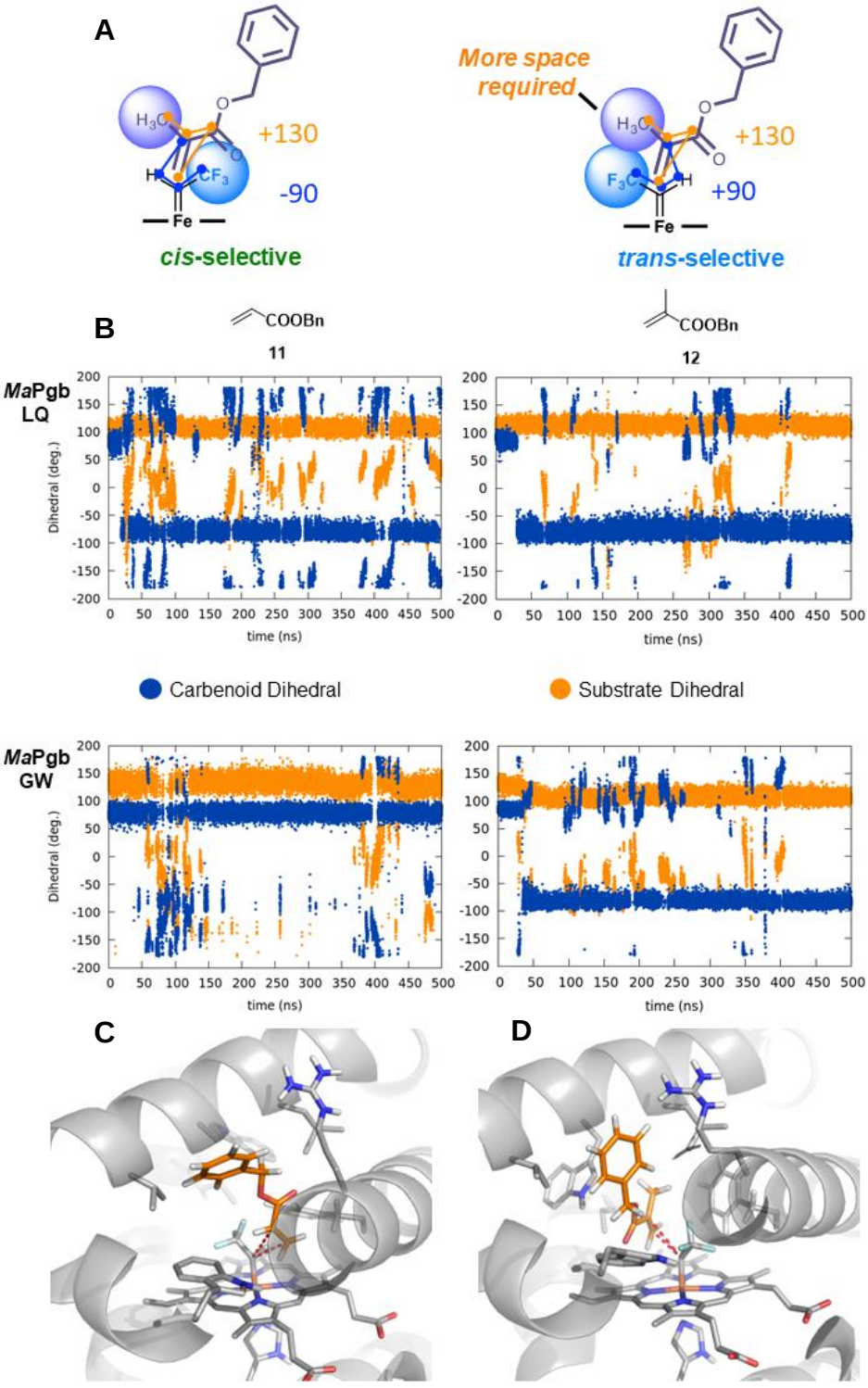


Figure 4.6: Computational modelling based on MD simulations to characterize the substrate bound in a catalytically competent pose relative to the iron-carbenoid in variants *Ma*Pgb LQ and GW. (**A**) Two different dihedral angles were defined to describe the relative orientation of the substituted alkene and the iron-carbenoid along substrate-bound MD simulations. These geometric parameters define which faces of the alkene and the carbenoid are exposed to each other. In orange:

∠C(carbene)-C(alkene)-C(=O)-C(CH_3) dihedral angle describes which face of the alkene is exposed to the carbenoid. In blue: ∠C(alkene)-H(carbene)-C(carbene)-C(CF_3) dihedral angle describes which face of the carbenoid is exposed to the alkene. Different combinations of dihedral angle values describe near attack conformations (NAC) that could produce *cis* or *trans* diastereomers. (**B**) Dihedral angles measured along 500 ns MD trajectories for *Ma*Pgb LQ/GW variants and benzyl acrylate/methacrylate substrates. Benzyl acrylate bound in *Ma*Pgb LQ mainly explores pro-*cis* NACs, while it mainly explores pro-*trans* NACs in the GW variant. Benzyl methacrylate mainly explores pro-*cis* NACs in both LQ and GW variants. (**C**) Representative snapshot taken from MD simulations of *Ma*Pgb GW with benzyl acrylate and (**D**) of *Ma*Pgb GW with benzyl methacrylate. For similar snapshots in *Ma*Pgb LQ, see Appendix D Figure 5.

Conclusion

Herein, we report a catalytic and enantioselective method for producing *cis*-trifluoromethyl-substituted cyclopropanes. Additionally, we have shown that the biocatalysts are active on a wide range of substrates, including non-aryl-substituted alkenes and unactivated alkenes. Their *cis*-selectivity is likely controlled by the active-site geometry, which preorganizes the iron-carbenoid and the substrate in a pro-*cis* near-attack conformation to overcome the intrinsic preference of the reaction. The reactions were performed using lyophilized whole-cell powders, which can be stored easily and used by those without any cell-culture experience. This new catalytic method to make *cis*-trifluoromethyl-substituted cyclopropanes provides a new, green route to their production and has the potential to become a valuable approach for producing new biologically active compounds.

Materials and Methods

Materials

Solvents and reagents were purchased from Sigma Aldrich, TCI, CombiBlocks, or Alfa Aesar and used without further purification. Nitrogen, argon, and carbon monoxide gas cylinders were ordered from Airgas. Phusion® High-Fidelity DNA polymerase was obtained from New England BioLabs (Ipswich, MA, USA). Oligonucleotides were synthesized by Integrated DNA Technologies (Coralville, Iowa). GC-FID data were collected on an Agilent 7820A GC system. GC-MS data were collected on a Shimadzu GCMS-QP2010 SE. NMR spectra were recorded on a Bruker Prodigy 400 MHz instrument or Varian 300 MHz instrument with CDCl3 as solvent. ^{1}H-NMR spectra were recorded at 400 MHz, ^{13}C-NMR spectra were recorded at 100 MHz and ^{19}F-NMR were recorded at 282 MHz. Chemical shifts were normalized to the chloroform solvent's proton impurity (^{1}H-NMR 7.26 ppm, ^{13}C-NMR 77.2 ppm) or to the fluorine in the 4-floroacetophenone internal standard (^{19}F NMR -113.15ppm). Protein concentrations were measured with a Tecan Spark® via carbon monoxide binding (Extinction coefficient $\varepsilon = 0.103\ \mu M^{-1}cm^{-1}$) as previously reported. Products were purified with a Biotage® Isolera One with either 10-g Biotage® SNAP Ultra columns or 25-g Biotage® Sfär Silica HC columns.

Molecular Cloning

Genes encoding *Methanosarcina acetivorans* and *Aeropyrnum pernix* protoglobins were ordered as codon-optimized gBlocks (Integrated DNA Technologies, Coralville, IA) and assembled into pET22b(+) with the pelB leader sequence removed. The gBlocks were amplified via polymerase chain reaction (PCR), and the PCR products were gel extracted and purified with the Zymoclean Gel DNA Recovery Kit (Zymo Research Corp, Irvine, CA). PCR was performed with the following protocol: 50 μL total volume, 50 ng template DNA, 0.5 μM primer, 0.2 mM dNTPs (Sigma), 1 μL Phusion), 2 μL DMSO (NEB), 2 μL formamide (Sigma). Annealing temperature was set at 60 °C for 15 s, while maintaining a ramp of 0.5 °C/s from the melting temperature (92 °C for 30 s) to the annealing temperature, followed by 15 s of elongation and 35 cycles were run to obtain product. The PCR product was subcloned into pET22b(+) via Gibson assembly.[2] Electrocompetent E. cloni EXPRESS BL21(DE3) cells (Lucigen, Middleton, WI) were transformed with the Gibson assembly products using a Gene Pulser Xcell (Bio-Rad, Hercules, CA). Aliquots of SOC medium (750 μL) were added, and the cells were incubated at 37 °C and 230 rpm for 45 minutes before being plated on LB ampicillin (LB_{amp}, 100 μg mL^{-1}) agar plates. Overnight cultures (5-mL LB_{amp} in culture tubes) were grown at 37 °C and 230 rpm for 12–18 hours. Overnight cultures were used to inoculate flask cultures, prepare glycerol stocks, and isolate plasmids. Plasmids were isolated with Qiagen Miniprep kits, and the genes were sequence verified (T7 promoter/terminator sequencing primers, Laragen, Inc.).

Site-Saturation Mutagenesis

Site-saturation mutagenesis was performed using the 22-codon method.[3] Briefly, oligonucleotides were ordered with NDT, VHG, and TGG codons in the coding strand at the amino acid position to be saturated. A reverse primer complementary to all three forward primers was also ordered. Two PCRs were performed for each library, the first containing a mixture of forward primers (12:9:1 NDT:VHG:TGG) and a pET22b(+) internal reverse primer and the second containing the complementary reverse primer and a pET22b(+) internal forward primer. The two PCR products were gel-purified with Zymoclean Gel DNA Recovery Kit (Zymo Research Corp, Irvine, CA) and ligated together via Gibson assembly. The Gibson assembly product was used to transform electrocompetent E. cloni EXPRESS BL21(DE3) cells (Lucigen, Middleton, WI). Aliquots of SOC medium (750 μL) were added, and the cells were incubated at 37 °C for 45 minutes before being plated on LB_{amp} (100 μg mL^{-1}) agar plates.

Protein Expression

For expression cultures, hyperbroth (HB, AthenaES) with 100 μg mL^{-1} ampicillin in unbaffled Erlenmeyer flasks was inoculated 1% (v/v) with stationary-phase overnight cultures and shaken in an Innova 428 shaker at 230 rpm and 37 °C. At an optical density of 600 (OD_{600}) = 0.8, the cultures were chilled on ice for 20 minutes. Protein expression was induced with 1 mM isopropyl β-d-1-thiogalactopyranoside (IPTG), and heme production was enhanced with supplementation of 1 mM 5-aminolevulinic acid (ALA). The cultures were shaken at 160 rpm and 22 °C overnight (18–24 hours). Cells were pelleted via centrifugation at 4000x*g* for 15 minutes at 4 °C. The supernatant was discarded, and the cells were resuspended in M9-N buffer.

96-Well Protein Expression

Single colonies from the LB_{amp} agar plates were picked using sterile toothpicks and grown in 300 μL LB_{amp} in 2-mL 96-well deep-well plates at 37 °C, 250 rpm, and 80% humidity overnight (12–18 hours). Multi-channel pipettes were used to transfer 20 μL of starter culture into deep-well plates containing 1 mL HB_{amp} per well. Glycerol stocks of these plates were prepared in parallel by adding starter culture (100 μL) and 50 % (v/v) sterile glycerol (100 μL) to a 96-well microplate, which was then stored at -80 °C. The deep-well expression culture plate was incubated at 37 °C, 250 rpm, and 80% humidity for 2.5 hours. The plate was then chilled on ice for 20 minutes. The cultures were induced with 1 mM IPTG and supplemented with 1 mM ALA to increase cellular heme production. The plate was incubated at 22 °C and 250 rpm overnight. The plate was centrifuged at 4000×*g* for 10 minutes at 4 °C.

Whole-Cell Lyophilization

Whole *Escherichia coli* (E. *coli*) cells containing *Ape*Pgb W59L Y60Q were expressed via the above protocol. The cell pellet was resuspended in M9-N buffer to OD_{600} = 45 in a 50-mL Falcon tube. The suspension was flash-frozen in liquid nitrogen and lyophilized for three days at 0.0018 mbar. Lyophilized whole-cell powder was stored at 4 °C. The lyophilized whole-cell powder can be reconstituted with dH_2O in a ratio of 36 mg lyophilized powder to 1 mL of water.

Synthesis of 2,2,2-Trifluorodiazoethane Trapped in Ethanol

To a 6-mL crimp vial, 4 mL EtOH were added and the vial subsequently sealed and cooled to -4 °C in an ice-salt bath (**Vial 1**). 2,2,2-Trifluoroethylamine HCl (800 mg, 6 mmol) was dissolved in 3 mL dH_2O and added to a crimp vial that was subsequently sealed (**Vial 2**). $NaNO_2$ (620 mg, 9 mmol) was dissolved in 2 mL dH_2O and taken up with a syringe. The needle of the syringe was introduced into **vial 2,** and $NaNO_2$ was added via slow addition over 4 hours. The developing gas of 2,2,2-Trifluorodiazoethane was bubbled through **vial 1** over the course of the reaction via canula transfer. A balloon was added to **vial 1** to regulate pressure. The resulting yellow solution was stored at -20 °C and used as such without further purification. ^{19}F-NMR: δ =-54 ppm

Analytical-Scale Enzymatic Reactions in 96-Well Plates

The reactions (400 μL) were carried out in 96-well plates in an anaerobic chamber (Coy). Whole-cell catalysts (380 μL, OD_{600} = 30 in M9-N minimal medium) were added to a 96-well plate. A solution of alkene reagent (10 μL, 400 mM in ethanol) was added, followed by a solution of 2,2,2-trifluorodiazoethane (10 μL, 400 mM in ethanol). The reaction plate was sealed using a pierceable foil cover (USA Scientific) and was left to shake on a plate shaker at 400 rpm for 12 h at room temperature. To quench the reaction, the plate was unsealed and a 1:1 mixture of pentane/diethyl ether (0.6 mL) was added, followed by 1,3,5-trimethoxybenzene (20 μL, 20 mM in toluene) as an internal standard. The plates were vortexed and centrifuged (4000x*g*, 5 min). The organic layer was analyzed by gas chromatography – mass spectrometry (GC-MS).

Time-Course Reactions

The reactions (400 μL) were carried out in 1.5-mL GC screw-cap vials in an anaerobic chamber at 24 °C and pH 7.4. Whole-cell catalysts (380 μL, OD_{600} = 30 in M9-N minimal medium) were added to a 1.5-mL GC screw-cap vial. A solution of alkene reagent (10 μL, 400 mM in ethanol) was added. As we added a solution of 2,2,2-trifluorodiazoethane (10 μL, 400 mM in ethanol), the timer was started. The vials were sealed, placed in a vial holder, and left to shake on a plate shaker at 400 rpm for the following times (minutes): 0, 1, 5, 10, 20, 30, 60, 90, 120, 180, 240, and

1440. After the time had elapsed, the reaction was quenched with the addition of a 1:1 mixture of hexane/ethyl acetate (0.6 mL) and vigorously shaken. The contents of each vial were transferred to a 1.7-mL microcentrifuge tube and vortexed, followed by centrifugation (14000x*g*, 5 min). The organic layer was analyzed by gas chromatography – flame induced detection (GC-FID). The reactions were performed in technical triplicated and biological duplicates for a total of 72 measurements.

Preparative-Scale Enzymatic Reactions

Lyophilized whole-cell powder (1.4 g) was reconstituted in 80 mL dH_2O via vortexing. The suspension was transferred to a 100-mL Erlenmeyer flask. To a 6-mL crimp vial, we added 270 mg of 2,2,2-trifluoroethylamine HCl dissolved in 1 mL dH_2O and subsequently sealed it. The crimp vial was placed on a stir plate and sparged with argon. A canula from the crimp vial was passed into the Erlenmeyer flask containing the enzyme, and this suspension was subsequently sparged with argon as well, making sure the gas stream did not make the mixture foam too much. $NaNO_2$ (207 mg) was dissolved in 0.8 mL dH_2O and taken up with a syringe. Sparging of the crimp vial solution was stopped and argon was blown on the surface of the solution in the vial. $NaNO_2$ was added to the crimp vial via slow addition over 2 hours. After 5 minutes, a 1 M solution of alkene in EtOH was added to the Erlenmeyer flask via slow addition over 20 minutes. After 2.5 hours, the reaction products were extracted with 1:1 pentane/diethyl ether (20 mL) via vortexing. The product was concentrated *in vacuo* (on ice) and purified via column chromatography (pentane/diethyl ether). The sample should always be kept cold to avoid evaporation of product.

Yield Determination of Preparative-Scale Enzymatic Reactions

Reactions were set up as described and extracted as above. The product was concentrated *in vacuo* (on ice) until around 1 to 3 mL of solution were left in the flask. The solution was then diluted in $CDCl_3$ to 10 mL total volume. For the substrates 'dimethyl itaconate', 3-nitrostyrene' and 'phenyl vinyl ether', we concentrated the crude *in vacuo* to below 1-mL volume and diluted the solution with $CDCl_3$ to 3 mL total volume. We transferred 900 µL of this solution to an NMR tube together with 100 µL of a 400 mM solution of 4-fluorobenzophenone as the standard. The NMR sample was measured at 282 MHz on a Varian 300 MHz instrument. The signal of the product was divided by three to account for the three fluorines of the cyclopropane versus one of the standards.
Yield was calculated via:

$$Yield = \frac{I(\mathrm{Cis})}{3 \cdot I(Standard)} \cdot Standard\ Concentration \cdot Volume(CDCl_3)$$

Where I is the integral from ^{19}F-NMR. Total Turnover Number (TTN) was calculated via:

$$TTN = \frac{Yield \cdot 1000}{Protein\ Concentration \cdot Reaction\ Volume}$$

Where the reaction was always from the same *Ape*Pgb LQ batch with a protein concentration of 5.3 μM and a reaction volume of 80 mL. Isolating the compounds typically reduces the yield about 10-fold, and the crude reaction extract should be always kept on ice to prevent evaporation. Ideally, isolation of products should be performed via distillation to avoid a large loss of product.

Synthesis of Authentic Standards: Method 1

As described previously:[4] $[Rh_2(OAc)_4]$ (21.5 mg, 0.0275 mmol) and NaOAc (18 mg, 0.22 mmol) were dissolved in degassed dH_2O (4 mL). Then 2,2,2-trifluoroethylamine HCl (300 mg, 2.2 mmol) and H_2SO_4 (6 μL, 0.11 mmol) were added. The solution was degassed for 1 minute by sparging with Ar(g). Alkene (1.1 mmol) was added next. $NaNO_2$ (180 mg, 2.6 mmol) was dissolved in 2.5 mL dH_2O and added over 10 hours, after which the products were extracted three times in DCM by vortexing in a 20-mL screw-cap vial. The organic phase was dried over $MgSO_4$, and solvents were removed *in vacuo*. The product was purified via column chromatography. Alternatively, after extraction and filtering of the reaction one can add 5% aqueous $KMnO_4$ (Warning: This is an exothermic reaction!), followed by washing of the organic phase with H_2O (three times). This consumes most of the unreacted alkene, making purification easier.

Synthesis of authentic standards: Method 2

As described previously:[5] A solution of $NaNO_2$ (345 mg, 5 mmol) in dH_2O (1 mL) and slowly added over 10 hours into a solution of 2,2,2-trifluoroethylamine HCl (677 mg, 5 mmol) in dH_2O (2 mL). Upon addition, the 2,2,2-trifluorodiazoethane formed was blown off with argon into a vial containing a stirring solution of alkene (1 mmol) and DCM (5 mL). Excess gas was trapped in a balloon. After the slow addition was completed, the solvent was removed *in vacuo* until only a clear oil was left. The oil was transferred into a sealed flask and heated to 70 °C for 1 hour in an oil bath. After cooling down the flask, the product was separated via column chromatography.

Synthesis of authentic standards: Method 3

As described previously:[4b] [Fe(TPP)Cl] (20 mg, 0.03 mmol), DMAP (12 mg, 0.1 mmol) and NaOAc (16 mg, 0.2 mmol) were dissolved in degassed dH_2O (4 mL). Then 2,2,2-trifluoroethylamine HCl (300 mg, 2.2 mmol) and H_2SO_4 (6 μL, 0.11 mmol) were added. Then the alkene (1.1 mmol) was added to the reaction mixture. $NaNO_2$ (180 mg, 2.6 mmol) was dissolved in 2.5 mL dH_2O and added over 10 hours. The product was extracted three times in DCM (10 mL), and the aqueous phase was separated from the organic phase via separatory funnel. The organic phase was dried with Na_2SO4, and solvents were removed *in vacuo*. The product was purified via column chromatography (pentane/diethyl ether).

Density functional theory calculations

Density Functional Theory (DFT) calculations were carried out using Gaussian09.[6] A truncated computational model was used, which includes the Fe-porphyrin pyrrole core, an imidazole coordinated to the Fe center to mimic the histidine axial ligand, the CF3-carbene bound to the Fe, and the corresponding benzyl acrylate substrate. Geometry optimizations and frequency calculations were performed using (U)B3LYP[7] functional with the SDD basis set for iron and 6-31G(d) on all other atoms. Transition states had one negative force constant corresponding to the desired reaction coordinate. Enthalpies and entropies were calculated for 1 atm and 298.15 K. Single point energy calculations were performed using the dispersion-corrected functional (U)B3LYP-D3(BJ)[8] with the Def2TZVP basis set on all atoms and within the CPCM polarizable conductor model (diethyl ether, $\varepsilon = 4$)[9] to have an estimation of the dielectric permittivity in the enzyme active site. The use of a dielectric constant ε=4 has been proven to be a good and general model to account for electronic polarization and small backbone fluctuations in enzyme active sites.*[10]*
The DFT-based approaches employed in this study are very similar to the ones previously used by us and other groups for the study of heme-iron carbene transfer reaction mechanisms.[3,28–32] Independent benchmark studies by Prof. Shaik[28] and Prof. Liu's[31] groups demonstrated that this method performs very well in the computational modelling of these carbene transfer reactions.
The modeling of the open-shell electronic state was done by using a Gaussian09 "stable = opt" calculation[11] to generate a singlet open-shell orbital guess from the triplet optimized geometry, followed by a full optimization of the system starting from this guess.

Molecular dynamics simulations

Molecular Dynamics simulations were performed using the GPU code (*pmemd*)[13] of the AMBER 16 package.[14] The *Ma*Pgb X-ray structure available from the PDB Bank (2VEB.pdb) was used as starting point in its monomeric form. Initial structures for double mutants LQ and GW were built using PyMOL[15] mutagenesis tool.
Parameters for the iron-carbenoid and substrates were generated within the *antechamber* and MCPB.py[16] modules in AMBER16 package using the general AMBER force field (*gaff*),[17] with partial charges set to fit the electrostatic potential generated at the B3LYP/6-31G(d) level by the RESP model.[18] The charges were calculated according to the Merz–Singh–Kollman scheme[19] using the Gaussian 09 package.[6] Protonation states of protein residues were predicted using H++ server. Each protein was immersed in a pre-equilibrated truncated cuboid box with a 10-Å buffer of TIP3P[20] water molecules using the *leap* module, resulting in the addition of around 9,600 solvent molecules. The systems were neutralized by addition of explicit counter ions (Na^+). All subsequent calculations were done using the widely tested Stony Brook modification of the Amber14 force field (*ff14sb*).[21] A two-stage geometry optimization approach was performed. The first stage minimizes the positions of solvent molecules and ions imposing positional restraints on the solute

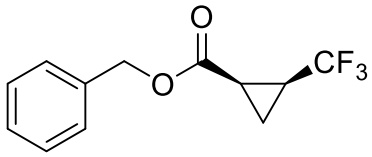
by a harmonic potential with a force constant of 500 kcal·mol^{-1}·Å^{-2} and the second stage minimizes all the atoms in the simulation cell except those involved in the harmonic distance restraint. The systems were gently heated using six 50 ps steps, incrementing the temperature by 50 K for each step (0–300 K) under constant-volume and periodic-boundary conditions. Water molecules were treated with the SHAKE algorithm such that the angle between the hydrogen atoms was kept fixed. Long-range electrostatic effects were modelled using the particle-mesh-Ewald method.[22] An 8-Å cutoff was applied to Lennard–Jones and electrostatic interactions. Harmonic restraints of 10 kcal·mol^{-1} were applied to the solute, and the Langevin equilibration scheme was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each system was then equilibrated for 2 ns with a 2-fs time step at a constant pressure. Once the systems were equilibrated in the NPT ensemble, production trajectories were then run under the NVT ensemble and periodic-boundary conditions for an additional 500 ns (0.5 μs). Trajectories were processed and analyzed using the cpptraj[23] module from Ambertools utilities.

Substrate-bound constrained MD simulations included a restrained distance between the center of mass of the substrate C–C double bond and the central C atom of the iron-carbenoid (3.0 – 3.2 Å) that was defined by adding a harmonic potential with k = 100 mol^{-1} Å^{-2} to this coordinate during the respective equilibrations and production runs.

Compound Characterization Data

Note: For compound characterization spectra see Appendix E.

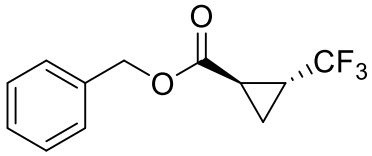
Cis-Benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate (cis-1):



Synthesized via method "Preparative Scale Enzymatic Reactions". After removal of solvent in the crude reaction mixture, the product was purified via column chromatography on a 10-g silica column using a pentane/diethyl ether gradient: 3 column volumes (CV) with pure pentane, 15 CV gradient up to 5% diethyl ether in pentane, 10 CV with 5% diethyl ether in pentane, 5 CV gradient up to 10% diethyl ether in pentane, 5 CV with 10% diethyl ether in pentane, 5 CV gradient up to 25% diethyl ether in pentane, 5 CV with 25% diethyl ether in pentane.

^{1}H-NMR (400 MHz, $CDCl_3$): δ 7.37 (m, 5H), 5.16 (dd, J = 12.2, 6.7 Hz, 2H), 2.06 (dddd, J = 9.4, 8.4, 6.8, 1.0 Hz, 1H), 1.93 (tp, J = 9.3, 7.5 Hz, 1H), 1.70 (q, J = 9.7 Hz, 3H), 1.27 (m, 1H).
^{13}C-NMR (100 MHz, $CDCl_3$): δ 168.6, 135.6, 128.7, 128.6, 128.5, 123.9 (q, J = 272.7 Hz), 67.4, 22.1 (q, J = 38.8 Hz), 18.8 (q, J = 1.8 Hz), 8.7 (q, J = 2.6 Hz).
^{19}F-NMR (282 MHz, $CDCl_3$): δ -61.15 (d, J = 7.4 Hz);
MS (FAB) m/z $[M]^{\bullet+}$ calcd for $C_{12}H_{11}F_3O_2$: 244.07057, found 244.06947.

Trans-Benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate (trans-1):

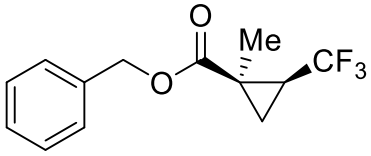


Synthesized via method "Synthesis of authentic standards: Method 3". After removal of solvent in the crude reaction mixture, the product was purified via column chromatography on a 10-g silica column using a pentane/diethyl ether gradient: 3 column volumes (CV) with pure pentane, 15 CV gradient up to 5% diethyl ether in pentane, 10 CV with 5% diethyl ether in pentane, 5 CV gradient up to 10% diethyl ether in pentane, 5 CV with 10% diethyl ether in pentane, 5 CV gradient up to 25% diethyl ether in pentane, 5 CV with 25% diethyl ether in pentane.

^{1}H-NMR (400 MHz, $CDCl_3$): δ 7.37 (m, 5H, Ar), 5.17 (m, 2H, $PhCH_2$), 2.18 (m, 1H, COCH), 2.09 (ddd, J = 9.5, 5.5, 4.3 Hz, 1H, CF_3CH), 1.38 (m, 2H, CH_2-cyclopropane).

^{13}C-NMR (100 MHz, $CDCl_3$): δ 171.3, 135.4, 128.8, 128.7, 128.5, 123.3 (q, J = 271.6 Hz), 67.3, 22.1, 17.0 (q, J = 2.6 Hz), 10.66 (q, J = 2.6 Hz).
^{19}F-NMR (282 MHz, $CDCl_3$): δ -67.14.

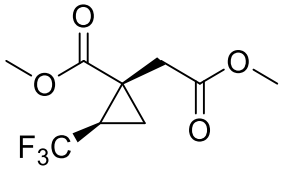
Benzyl 1-methyl-2-(trifluoromethyl)cyclopropane-1-carboxylate (2):



Synthesized via method "Preparative Scale Enzymatic Reactions". After removal of solvent in the crude reaction mixture, the product was purified via column chromatography on a 10-g silica column using a pentane/diethyl ether gradient: 3 column volumes (CV) with pure pentane, 15 CV gradient up to 5% diethyl ether in pentane, 5 CV with 5% diethyl ether in pentane, 5 CV gradient up to 25% diethyl ether in pentane, 5 CV with 25% diethyl ether in pentane.

^{1}H-NMR (400 MHz, $CDCl_3$): δ 7.36 (m, 5H, Ar), 5.14 (dd, J = 12.3, 11.0 Hz, 2H, $PhCH_2$), 1.82 (t, J = 6.2 Hz, 1H, CF_3CH), 1.68 (m, 1H, CH_2-cyclopropane), 1.42 (s, 3H, Me), 1.05 (ddq, J = 8.5, 5.6, 1.3 Hz, 1H, CH_2-cyclopropane).
^{13}C-NMR (100 MHz, $CDCl_3$): δ 170.5, 135.6, 128.6, 128.4, 126.5, 125.1 (q, J = 272.5 Hz), 67.4, 28.8 (q, J = 38.0 Hz). 25.5 (q, J = 1.9 Hz), 15.9 (q, J = 2.5 Hz).
^{19}F-NMR (282 MHz, $CDCl_3$): δ -61.4 (d, J = 7.7 Hz).
MS (FAB) m/z $[M]^{\bullet+}$ calcd for $C_{13}H_{13}F_3O_2$: 258.08622, found 258.08484.

Methyl 1-(2-methoxy-2-oxoethyl)-2-(trifluoromethyl)cyclopropane-1-carboxylate (3):

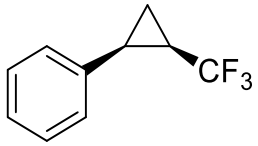


Synthesized via method "Preparative Scale Enzymatic Reactions". After removal of solvent in the crude reaction mixture, the product was purified via column chromatography on a 10-g silica column using a pentane/diethyl ether gradient: 3 column volumes (CV) with pure pentane, 10 CV gradient up to 5% diethyl ether, 10 CV with 5% diethyl ether in pentane, 10 CV gradient up to 25% diethyl ether in pentane, 15 CV with 25% diethyl ether in pentane.
^{1}H-NMR (400 MHz, $CDCl_3$): δ 3.69 (s, 3H, COOMe), 3.69 (s, 3H, COOMe), 2.80 (dd, J = 60, 18, 2H, CH_2COOMe), 2.34 (m, 1H, CF_3CH), 1.71 (ddt, J = 9.6, 5.5, 1.2, 1H, CH_2-cyclopropane), 1.27 (m, 1H, CH_2-cyclopropane).

^{13}C-NMR (100 MHz, $CDCl_3$): δ 172.3, 171.6, 125.1 (q, J = 272.4 Hz), 52.9, 52.0, 32.9 (d, J = 2.0 Hz), 26.0 (q, J = 37.5 Hz), 24.6 (d, J = 1.7 Hz), 16.9 (q, J = 2.7 Hz).
^{19}F-NMR (282 MHz, $CDCl_3$): δ -60.9 (d, J = 7.8 Hz).

MS (FAB) m/z $[M]^{\bullet+}$ calcd for $C_9H_{11}F_3O_4$: 240. 06039, found 240. 05937.

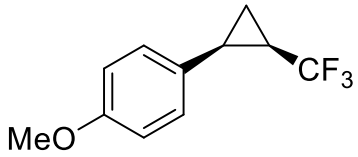
(2-(Trifluoromethyl)cyclopropyl)benzene (4):



Synthesized via method "Preparative Scale Enzymatic Reactions". Reactions were extracted in pentane, and the organic phase was washed with 1% aqueous $KMnO_4$. After removal of solvent in the crude reaction mixture, the product was purified via column chromatography on a 10-g silica column using pentane as eluent. Compound is volatile, and pentane is present in the NMR spectrum.

^{1}H-NMR (400 MHz, $CDCl_3$): δ 7.31 (m, 4H, Ar), 7.26 (m, 1H, Ar), 2.49 (tdd, J = 8.8, 7.5, 1.2 Hz, 1H, PhCH), 1.88 (m, 1H, CF_3CH), 1.44 (dd, J = 7.3, 5.9, 1H, CH_2-cyclopropane), 1.34 (m, 1H, CH_2-cyclopropane).
^{13}C-NMR (100 MHz, $CDCl_3$): δ 135.4, 129.5, 128.0, 127.0, 126.4 (q, J = 272.0 Hz), 20.6 (q, J = 1.8 Hz), 20.38 (q, J = 35.5 Hz), 6.5 (q, J = 2.5 Hz).
^{19}F-NMR (282 MHz, $CDCl_3$): δ -61.2 (d, J = 7.5 Hz).
NMR values are in accordance with literature. [4a]

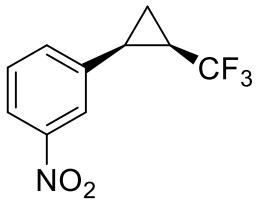
1-Methoxy-4-(2-(trifluoromethyl)cyclopropyl)benzene (5):



Synthesized via method "Preparative Scale Enzymatic Reactions". Reactions were extracted in pentane, and the organic phase was washed with 1% aqueous $KMnO_4$. After removal of solvent in the crude reaction mixture, the product was purified via column chromatography on a 10-g silica column using pentane as eluent. Compound is volatile, and pentane is present in the NMR spectrum.

^{1}H-NMR (400 MHz, $CDCl_3$): δ 7.22 (m, 2H, Ar), 6.83 (m, 2H, Ar), 3.79 (s, 3H, OMe), 2.42 (m, 1H, ArCH), 1.81 (tqd, J = 8.8, 7.7, 5.7, 1H, CF_3CH), 1.36 (m, 1H, CH_2-cyclopropane), 1.26 (m, 1H, CH_2-cyclopropane).
^{13}C-NMR (100 MHz, $CDCl_3$): δ 158.6, 130.5, 127.4, 126.43 (q, J = 272.1 Hz), 113.7, 55.3, 20.23 (q, J = 35.1 Hz), 19.86 (q, J = 1.8 Hz), 6.74 (q, J = 2.5 Hz).
^{19}F-NMR (282 MHz, $CDCl_3$): δ -60.49 (d, J = 7.4 Hz).
MS (FAB) m/z $[M]^{\bullet+}$ calcd for $C_{11}H_{11}F_3O$: 216.07565, found 216.07384.

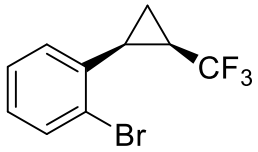
1-Nitro-3-(2-(trifluoromethyl)cyclopropyl)benzene (6):



Synthesized via method "Preparative Scale Enzymatic Reactions". Reactions were extracted in pentane, and the organic phase was washed with 1% aqueous $KMnO_4$. After removal of solvent in the crude reaction mixture, the product was purified via column chromatography on a 10-g silica column using pentane as eluent. Compound is volatile, and pentane is present in the NMR spectrum.

^{1}H-NMR (400 MHz, $CDCl_3$): δ 8.17 (m, 1H, Ar), 8.11 (m, 1H, Ar), 7.65 (d, J = 7.7 Hz, 1H, Ar), 7.47 (t, J = 7.9 Hz, 1H, Ar), 2.55 (tdt, J = 9.1, 7.1, 1.0 Hz, 1H, ArCH), 1.99 (m, 1H, CF_3CH), 1.51 (dt, J = 7.2, 6.0, Hz, 1H, CH_2-cyclopropane), 1.40 (tdq, J = 8.7, 6.0, 1.3 Hz, 1H, CH_2-cyclopropane).
^{13}C-NMR (100 MHz, $CDCl_3$): δ 148.3, 137.7, 135.9, 129.2, 125.9 (q, J = 272.2 Hz), 124.4, 122.3, 20.75 (q, J = 35.7 Hz), 20.10(q, 1.8 Hz), 6.9 (q, J = 2.7 Hz).
^{19}F-NMR (282 MHz, $CDCl_3$): δ -61.31 (d, J = 7.5 Hz).
MS (FAB) m/z $[M]^{\bullet+}$ calcd for $C_{10}H_8F_3NO_2$: 231.05016, found 231.04902.

1-Bromo-2-(2-(trifluoromethyl)cyclopropyl)benzene (7):



Synthesized via method "Preparative Scale Enzymatic Reactions". Reactions were extracted in pentane and the organic phase was washed with 1% aqueous $KMnO_4$. After removal of solvent in the crude reaction mixture, the product was purified via column chromatography on a 10-g silica column using pentane as eluent. Compound is volatile, and pentane is present in the NMR spectrum.

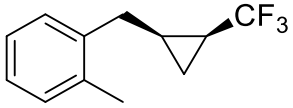
^{1}H-NMR (400 MHz, $CDCl_3$): 7.59 (dd, J = 7.9, 1.3, 1H, Ar), 7.23 (m, 2H, Ar), 7.13 (tdd, J = 7.2, 1.9, 0.6 Hz, 1H, Ar), 2.48 (q, J = 8.3, 1H, ArCH), 2.04 (m, 1H, CF_3CH), 1.52 (m, 1H, CH_2-cyclopropane), 1.36 (m, 1H, CH_2-cyclopropane).
^{13}C-NMR (400 MHz, $CDCl_3$): δ 135.09, 132.65, 129.81 (d, J = 1.6 Hz), 128.69, 127.12, 126.07 (q, J = 272.5 Hz), 27.29 (q, J = 899.1 Hz), 20.94 (q, J = 35.5 Hz), 6.91 (q, J = 2.6 Hz).
^{19}F-NMR (300 MHz, $CDCl_3$): -61.70 (d, J = 7.5 Hz).
MS (FAB) m/z $[M]^{\bullet+}$ calcd for $C_{10}H_8BrF_3$: 263.97560, found 263.97443.

((2-(Trifluoromethyl)cyclopropyl)methyl)benzene (8):

Synthesized via method "Preparative Scale Enzymatic Reactions". Reactions were extracted in pentane, and the organic phase was washed with 1% aqueous $KMnO_4$. After removal of solvent in the crude reaction mixture, the product was purified via column chromatography on a 10-g silica column using pentane as eluent. Compound is volatile, and pentane is present in the NMR spectrum. As reported before,[25] the trans- and cis-isomer cannot be separated via column chromatography. At large enough scale, distillation can be performed as previously suggested [4a].

^{1}H-NMR (400 MHz, $CDCl_3$): δ 7.28 (m, Ar), 2.98 (dd, J = 15.0, 5.5 Hz, *cis*, $PhCH_2$), 2.73 (dd, J = 14.8, 6.5 Hz, *trans*, $PhCH_2$), 2.64 (dt, J = 14.9, 9.7 Hz, $PhCH_2$), 1.87 (dq, J = 14.7, 7.4 Hz, *cis*, BnCH-cyclopropane), 1.66–1.44 (m, CF_3CH-cyclopropane), 1.12-1.06 (m, *cis*, CH_2-cyclopropane), 1.06-0.99 (m, *trans*, CH_2-cyclopropane), 0.97 (dd, J = 6.6, 0.8 Hz, *cis*, CH_2-cyclopropane), 0.71 (dd, J = 9.5, 5.2 Hz, *trans*, CH_2-cyclopropane).
^{13}C-NMR (100 MHz, $CDCl_3$): δ 141.11, 139.78, 128.64, 128.61, 128.46, 128.41, 126.57, 126.37, 37.95 (*trans*, $PhCH_2$), 33.44 (q, J = 1.7 Hz, *cis*, $PhCH_2$), 19.69 (*trans*, CF_3CH), 18.17 (*cis*, CF_3CH), 17.57 (*cis*, BnCH), 16.24 (q, J = 2.3 Hz, *trans*, BnCH), 8.98 (q, J = 2.5 Hz, *cis*, CH_2-cyclopropane), 8.61 (*trans*, CH_2-cyclopropane).
^{19}F-NMR (282 MHz, $CDCl_3$): δ -59.54 (dd, J = 8.6 Hz, *cis*), -66.29 (d, J = 6.7 Hz, *trans*).
MS (FAB) m/z $[M]^{\bullet+}$ calcd for $C_{11}H_{11}F_3$: 200.08074, found 200.07894.

1-Methyl-2-((2-(trifluoromethyl)cyclopropyl)methyl)benzene (9):

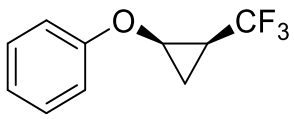


Synthesized via method "Preparative Scale Enzymatic Reactions". Reactions were extracted in pentane, and the organic phase was washed with 1% aqueous $KMnO_4$. After removal of solvent in the crude reaction mixture, the product was purified via column chromatography on a 10-g silica column using pentane as eluent. Compound is volatile, and pentane is present in the NMR spectrum.

^{1}H-NMR (400 MHz, $CDCl_3$): δ 7.30 (dd, J = 7.7, 2.0 Hz, 1H, Ar), 7.18 (m, 3H, Ar), 3.03 (dd, J = 15.4, 4.9 Hz, 1H, $ArCH_2$), 2.60 (dd, J = 15.2, 9.5 Hz, 1H, $ArCH_2$), 2.35 (s, 3H, Me), 1.65 (qd, J = 8.6, 5.7 Hz, 1H, CF_3CH), 1.41 (m, 1H, ArCH), 1.09 (tdd, J = 8.7, 5.2, 1.4 Hz, 1H, CH_2-cyclopropane), 1.00 (d, 6.6 Hz, 1H, CH_2-cyclopropane).
^{13}C-NMR (100 MHz, $CDCl_3$): δ 139.32, 136.09, 130.30, 128.44, 127.24 (q, J = 271.9 Hz), 126.45, 126.21, 31.77, 30.55, 18.39 (q, J = 35.9 Hz), 16.34 (q, J = 1.5 Hz), 9.18 (q, J = 2.6 Hz).

^{19}F-NMR (282 MHz, $CDCl_3$): δ -59.6 (d, J = 8.4 Hz).
MS (FAB) m/z $[M]^{\bullet+}$ calcd for $C_{12}H_{12}F_3$: 214.09639, found 214.09549.

(2-(Trifluoromethyl)cyclopropoxy)benzene (10):



Synthesized via method "Preparative Scale Enzymatic Reactions". After removal of solvent in the crude reaction mixture, the product was purified via column chromatography on a 10-g silica column using pentane. Compound is volatile, and pentane is present in the NMR spectrum.

^{1}H-NMR (400 MHz, $CDCl_3$): δ 7.31 (m, 2H, Ar), 7.03 (m,3 H, Ar), 3.96 (dtq, *J* = 6.4, 4.3, 2.1 Hz, 1H, OCH), 1.79 (dpd, J = 10.0, 7.2, 6.4 Hz, 1H, CF_3CH), 1.36 (td, J = 7.0, 4.3 Hz, 2H, CH_2-cyclopropane), 1.29 (m, 1H, CH_2-cyclopropane).
^{13}C-NMR (100 MHz, $CDCl_3$): δ 158.1, 129.5, 125.50 (q, J = 271.9 Hz), 121.9, 115.16, 52.3 (q, J = 2.1 Hz), 19.7 (q, J = 36.4 Hz), 9.43 (q, J = 2.6 Hz), 1.18.
^{19}F-NMR (300 MHz, $CDCl_3$): δ -60.6 (dt, J = 7.2, 1.7 Hz).
MS (FAB) m/z $[M]^{\bullet+}$ calcd for $C_{10}H_9F_3O$: 202.06000, found 202.05855.

References

1. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J. Nat. Prod. 83, 770–803 (2020).
2. Patridge, E., Gareiss, P., Kinch, M. S. & Hoyer, D. An analysis of FDA-approved drugs: natural products and their derivatives. Drug Discovery Today 21, 204–207 (2016).
3. Batth, R., Nicolle, C., Cuciurean, I. S. & Simonsen, H. T. Biosynthesis and Industrial Production of Androsteroids. Plants (Basel) 9, 1144 (2020).
4. Helfrich, E. J. N., Lin, G.-M., Voigt, C. A. & Clardy, J. Bacterial terpene biosynthesis: challenges and opportunities for pathway engineering. Beilstein J Org Chem 15, 2889–2906 (2019).
5. Athavale, S. V. et al. Enzymatic Nitrogen Insertion into Unactivated C–H Bonds. J. Am. Chem. Soc. 144, 19097–19105 (2022).
6. Lewis, J. C., Coelho, P. S. & Arnold, F. H. Enzymatic Functionalization of Carbon-Hydrogen Bonds. Chem Soc Rev 40, 2003–2021 (2011).
7. Hartwig, J. F. & Larsen, M. A. Undirected, Homogeneous C–H Bond Functionalization: Challenges and Opportunities. ACS Cent. Sci. 2, 281–292 (2016).
8. Zhang, J. et al. Chemodivergent C(sp3)–H and C(sp2)–H cyanomethylation using engineered carbene transferases. Nat Catal 6, 152–160 (2023).
9. Zhang, Y. et al. Enantioselective oxidation of unactivated C–H bonds in cyclic amines by iterative docking-guided mutagenesis of P450BM3 (CYP102A1). Nat. Synth 1, 936–945 (2022).
10. Truppo, M. D. Biocatalysis in the Pharmaceutical Industry: The Need for Speed. ACS Med. Chem. Lett. 8, 476–480 (2017).
11. France, S. P., Lewis, R. D. & Martinez, C. A. The Evolving Nature of Biocatalysis in Pharmaceutical Research and Development. JACS Au 3, 715–735 (2023).
12. Arnold, F. H. Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). Angewandte Chemie International Edition 58, 14420–14426 (2019).
13. Lèbre, É. et al. The social and environmental complexities of extracting energy transition metals. Nat Commun 11, 4823 (2020).
14. Micklitsch, C., Duan, D. & Borra-Garske, M. KREDs: Toward Green, Cost-Effective, and Efficient Chiral Alcohol Generation. in Industrial Enzyme Applications 351–383 (John Wiley & Sons, Ltd, 2019). doi:10.1002/9783527813780.ch5_1.
15. Truppo, M. D., Rozzell, J. D. & Turner, N. J. Efficient Production of Enantiomerically Pure Chiral Amines at Concentrations of 50 g/L Using Transaminases. Org. Process Res. Dev. 14, 234–237 (2010).

16. Wittmann, B. J., Yue, Y. & Arnold, F. H. Informed training set design enables efficient machine learning-assisted directed protein evolution. cels 12, 1026-1045.e7 (2021).
17. Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. Advances in machine learning for directed evolution. Curr Opin Struct Biol 69, 11–18 (2021).
18. Yang, J. et al. DeCOIL: Optimization of Degenerate Codon Libraries for Machine Learning-Assisted Protein Engineering. ACS Synth. Biol. 12, 2444–2454 (2023).
19. Freschlin, C. R., Fahlberg, S. A. & Romero, P. A. Machine learning to navigate fitness landscapes for protein engineering. Current Opinion in Biotechnology 75, 102713 (2022).
20. Tachibana, R., Zhang, K., Zou, Z., Burgener, S. & Ward, T. R. A Customized Bayesian Algorithm to Optimize Enzyme-Catalyzed Reactions. ACS Sustainable Chem. Eng. 11, 12336–12344 (2023).
21. Woodley, J. M. Integrating protein engineering into biocatalytic process scale-up. TRECHEM 4, 371–373 (2022).
22. Ruck, R. T. et al. Bio- and Chemocatalysis for the Synthesis of Late Stage SAR-Enabling Intermediates for ROMK Inhibitors and MK-7145 for the Treatment of Hypertension and Heart Failure. Org. Process Res. Dev. 25, 405–410 (2021).
23. Burns, M. et al. A Chemoenzymatic Route to Chiral Intermediates Used in the Multikilogram Synthesis of a Gamma Secretase Inhibitor. Org. Process Res. Dev. 21, 871–877 (2017).
24. Huffman, M. A. et al. Design of an in vitro biocatalytic cascade for the manufacture of islatravir. Science 366, 1255–1259 (2019).
25. Miller, D. C., Athavale, S. V. & Arnold, F. H. Combining chemistry and protein engineering for new-to-nature biocatalysis. Nat Synth 1, 18–23 (2022).
26. Svastits, E. W., Dawson, J. H., Breslow, R. & Gellman, S. H. Functionalized nitrogen atom transfer catalyzed by cytochrome P-450. J. Am. Chem. Soc. 107, 6427–6428 (1985).
27. Maynard Smith, J. Natural Selection and the Concept of a Protein Space. Nature 225, 563–564 (1970).
28. Schaus, L. et al. Protoglobin-Catalyzed Formation of cis-Trifluoromethyl-Substituted Cyclopropanes by Carbene Transfer. Angewandte Chemie International Edition 62, e202208936 (2023).
29. Knight, A. M. Expanding the Scope of Metalloprotein Families and Substrate Classes in New-to-Nature Reactions.
30. Knight, A. M. et al. Diverse Engineered Heme Proteins Enable Stereodivergent Cyclopropanation of Unactivated Alkenes. ACS Cent. Sci. 4, 372–377 (2018).

31. Coelho, P. S., Brustad, E. M., Kannan, A. & Arnold, F. H. Olefin Cyclopropanation via Carbene Transfer Catalyzed by Engineered Cytochrome P450 Enzymes. Science 339, 307–310 (2013).
32. Key, H. M. et al. Beyond Iron: Iridium-Containing P450 Enzymes for Selective Cyclopropanations of Structurally Diverse Alkenes. ACS Cent. Sci. 3, 302–308 (2017).
33. Gao, S. et al. Enzymatic Nitrogen Incorporation Using Hydroxylamine. J. Am. Chem. Soc. 145, 20196–20201 (2023).
34. Cho, I. et al. Enantioselective Aminohydroxylation of Styrenyl Olefins Catalyzed by an Engineered Hemoprotein. Angew Chem Int Ed Engl 58, 3138–3142 (2019).
35. Jia, Z.-J., Gao, S. & Arnold, F. H. Enzymatic Primary Amination of Benzylic and Allylic C(sp3)-H Bonds. J Am Chem Soc 142, 10279–10283 (2020).
36. Hyster, T. K., Farwell, C. C., Buller, A. R., McIntosh, J. A. & Arnold, F. H. Enzyme-controlled nitrogen-atom transfer enables regiodivergent C-H amination. J Am Chem Soc 136, 15505–15508 (2014).
37. Basler, S. et al. Efficient Lewis acid catalysis of an abiological reaction in a de novo protein scaffold. Nat Chem 13, 231–235 (2021).
38. Preiswerk, N. et al. Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. Proceedings of the National Academy of Sciences 111, 8013–8018 (2014).
39. Blomberg, R. et al. Precision is essential for efficient catalysis in an evolved Kemp eliminase. Nature 503, 418–421 (2013).
40. Bunzel, H. A. et al. Emergence of a Negative Activation Heat Capacity during Evolution of a Designed Enzyme. J. Am. Chem. Soc. 141, 11745–11748 (2019).
41. Porter, N. J., Danelius, E., Gonen, T. & Arnold, F. H. Biocatalytic Carbene Transfer Using Diazirines. J. Am. Chem. Soc. 144, 8892–8896 (2022).
42. Athavale, S. V., Chen, K. & Arnold, F. H. Engineering Enzymes for New-to-Nature Carbene Chemistry. in Transition Metal-Catalyzed Carbene Transformations 95–138 (John Wiley & Sons, Ltd, 2022). doi:10.1002/9783527829170.ch4.
43. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. Proceedings of the National Academy of Sciences 103, 5869–5874 (2006).
44. Pesce, A., Bolognesi, M. & Nardini, M. Protoglobin: structure and ligand-binding properties. Adv Microb Physiol 63, 79–96 (2013).
45. Nardini, M. et al. Archaeal protoglobin structure indicates new ligand diffusion paths and modulation of haem-reactivity. EMBO Rep 9, 157–163 (2008).
46. Freitas, T. A. K., Saito, J. A., Hou, S. & Alam, M. Globin-coupled sensors, protoglobins, and the last universal common ancestor. Journal of Inorganic Biochemistry 99, 23–33 (2005).

47. Walker, J. A., Rivera, S. & Weinert, E. E. Mechanism and Role of Globin Coupled Sensor Signaling. Adv Microb Physiol 71, 133–169 (2017).
48. Inoue, M., Sumii, Y. & Shibata, N. Contribution of Organofluorine Compounds to Pharmaceuticals. ACS Omega 5, 10633–10640 (2020).
49. Müller, K., Faeh, C. & Diederich, F. Fluorine in pharmaceuticals: looking beyond intuition. Science 317, 1881–1886 (2007).
50. Purser, S., Moore, P. R., Swallow, S. & Gouverneur, V. Fluorine in medicinal chemistry. Chem. Soc. Rev. 37, 320–330 (2008).
51. Murray, J. S., Lane, P., Clark, T. & Politzer, P. Sigma-hole bonding: molecules containing group VI atoms. J Mol Model 13, 1033–1038 (2007).
52. Elakkat, V. et al. The first two examples of halogen bonding with a sigma hole-donating fluorine in the Csp3–F···Osp3 interaction from polyfluorinated trans-dihalo-palladium(II) di-substituted pyridine complexes. Chem. Commun. 55, 14259–14262 (2019).
53. Patani, G. A. & LaVoie, E. J. Bioisosterism: A Rational Approach in Drug Design. Chem Rev 96, 3147–3176 (1996).
54. Westphal, M. V., Wolfstädter, B. T., Plancher, J.-M., Gatfield, J. & Carreira, E. M. Evaluation of tert-butyl isosteres: case studies of physicochemical and pharmacokinetic properties, efficacies, and activities. ChemMedChem 10, 461–469 (2015).
55. Bos, M., Poisson, T., Pannecoucke, X., Charette, A. B. & Jubault, P. Recent Progress Toward the Synthesis of Trifluoromethyl- and Difluoromethyl-Substituted Cyclopropanes. Chemistry 23, 4950–4961 (2017).
56. Barnes-Seeman, D. et al. Metabolically Stable tert-Butyl Replacement. ACS Med. Chem. Lett. 4, 514–516 (2013).
57. Talele, T. T. The 'Cyclopropyl Fragment' is a Versatile Player that Frequently Appears in Preclinical/Clinical Drug Molecules. J Med Chem 59, 8712–8756 (2016).
58. Morandi, B., Mariampillai, B. & Carreira, E. M. Enantioselective cobalt-catalyzed preparation of trifluoromethyl-substituted cyclopropanes. Angew Chem Int Ed Engl 50, 1101–1104 (2011).
59. Tinoco, A., Steck, V., Tyagi, V. & Fasan, R. Highly Diastereo- and Enantioselective Synthesis of Trifluoromethyl-Substituted Cyclopropanes via Myoglobin-Catalyzed Transfer of Trifluoromethylcarbene. J Am Chem Soc 139, 5293–5296 (2017).
60. Costantini, M. & Mendoza, A. Modular Enantioselective Synthesis of cis-Cyclopropanes through Self-Sensitized Stereoselective Photodecarboxylation with Benzothiazolines. ACS Catal 11, 13312–13319 (2021).
61. Duncton, M. A. J. & Singh, R. Synthesis of trans-2-(trifluoromethyl)cyclopropanes via Suzuki reactions with an N-methyliminodiacetic acid boronate. Org Lett 15, 4284–4287 (2013).

62. Mykhailiuk, P., Afonin, S., Ulrich, A. & Komarov, I. A Convenient Route to Trifluoromethyl-Substituted Cyclopropane Derivatives. Synthesis 2008, 1757–1760 (2008).
63. Hock, K. J. et al. Corey-Chaykovsky Reactions of Nitro Styrenes Enable cis-Configured Trifluoromethyl Cyclopropanes. J Org Chem 82, 8220–8227 (2017).
64. Morandi, B. & Carreira, E. M. Iron-catalyzed cyclopropanation with trifluoroethylamine hydrochloride and olefins in aqueous media: in situ generation of trifluoromethyl diazomethane. Angew Chem Int Ed Engl 49, 938–941 (2010).
65. Fuchibe, K., Oki, R., Hatta, H. & Ichikawa, J. Single C-F Bond Activation of the CF3 Group with a Lewis Acid: CF3 -Cyclopropanes as Versatile 4,4-Difluorohomoallylating Agents. Chemistry 24, 17932–17935 (2018).
66. Risse, J., Fernández-Zúmel, M. A., Cudré, Y. & Severin, K. Synthesis of trifluoromethyl-substituted cyclopropanes via sequential Kharasch-dehalogenation reactions. Org Lett 14, 3060–3063 (2012).
67. Duan, Y., Lin, J.-H., Xiao, J.-C. & Gu, Y.-C. A Trifluoromethylcarbene Source. Org Lett 18, 2471–2474 (2016).
68. Wei, Y., Tinoco, A., Steck, V., Fasan, R. & Zhang, Y. Cyclopropanations via Heme Carbenes: Basic Mechanism and Effects of Carbene Substituent, Protein Axial Ligand, and Porphyrin Substitution. J Am Chem Soc 140, 1649–1662 (2018).
69. Tinoco, A. et al. Origin of high stereocontrol in olefin cyclopropanation catalyzed by an engineered carbene transferase. ACS Catal 9, 1514–1524 (2019).
70. Carminati, D. M., Decaens, J., Couve-Bonnaire, S., Jubault, P. & Fasan, R. Biocatalytic Strategy for the Highly Stereoselective Synthesis of CHF2 -Containing Trisubstituted Cyclopropanes. Angew Chem Int Ed Engl 60, 7072–7076 (2021).
71. Kazuta, Y., Matsuda, A. & Shuto, S. Development of versatile cis- and trans-dicarbon-substituted chiral cyclopropane units: synthesis of (1S,2R)- and (1R,2R)-2-aminomethyl-1-(1H-imidazol-4-yl)cyclopropanes and their enantiomers as conformationally restricted analogues of histamine. J Org Chem 67, 1669–1677 (2002).
72. Maux, Ap., Juillard, S. & Simonneaux, G. Asymmetric Synthesis of Trifluoromethylphenyl Cyclopropanes Catalyzed by Chiral Metalloporphyrins. New York (2006).
73. Gilman, H. & Jones, R. G. 2,2,2-Trifluoroethylamine and 2,2,2-Trifluorodiazoethane. J. Am. Chem. Soc. 65, 1458–1460 (1943).
74. Zhang, J., Huang, X., Zhang, R. K. & Arnold, F. H. Enantiodivergent α-Amino C-H Fluoroalkylation Catalyzed by Engineered Cytochrome P450s. J Am Chem Soc 141, 9798–9802 (2019).
75. Artamonov, O. S., Mykhailiuk, P. K., Voievoda, N. M., Volochnyuk, D. M. & Komarov, I. V. Simple and Efficient Procedure for a Multigram

Synthesis of Both trans- and cis-1-Amino-2-(trifluoromethyl)cyclopropane-1-carboxylic Acid. Synthesis 2010, 443–446 (2010).
76. Pesce, A. et al. Structure and haem-distal site plasticity in Methanosarcina acetivorans protoglobin. PLoS One 8, e66144 (2013).
77. Ciaccio, C. et al. Functional and structural roles of the N-terminal extension in Methanosarcina acetivorans protoglobin. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics 1834, 1813–1823 (2013).
78. Garcia-Borràs, M. et al. Origin and Control of Chemoselectivity in Cytochrome C-Catalyzed Carbene Transfer into Si–H and N–H Bonds. (2021) doi:10.26434/chemrxiv.14102363.v1.
79. Vargas, D. A. et al. Biocatalytic strategy for the construction of sp3-rich polycyclic compounds from directed evolution and computational modelling. Nat. Chem. 1–10 (2024) doi:10.1038/s41557-023-01435-3.
80. Liu, Z. et al. Dual-function enzyme catalysis for enantioselective carbon-nitrogen bond formation. Nat Chem 13, 1166–1172 (2021).
81. Calvó-Tusell, C., Liu, Z., Chen, K., Arnold, F. H. & Garcia-Borràs, M. Reversing the Enantioselectivity of Enzymatic Carbene N−H Insertion Through Mechanism-Guided Protein Engineering**. Angewandte Chemie International Edition 62, e202303879 (2023).
82. Garcia-Borràs, M. et al. Origin and Control of Chemoselectivity in Cytochrome c Catalyzed Carbene Transfer into Si-H and N-H bonds. J Am Chem Soc 143, 7114–7123 (2021).
83. Athavale, S. V. et al. Biocatalytic, Intermolecular C−H Bond Functionalization for the Synthesis of Enantioenriched Amides. Angewandte Chemie International Edition 60, 24864–24869 (2021).
84. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr. ,. J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman & Fox, D. J. Gaussian 09, Revision A.02. (2016).
85. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. Phys. Rev. A 38, 3098–3100 (1988).
86. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. The Journal of Chemical Physics 98, 5648–5652 (1993).

87. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys. Rev. B 37, 785–789 (1988).
88. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. The Journal of Chemical Physics 132, 154104 (2010).
89. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. Journal of Computational Chemistry 32, 1456–1465 (2011).
90. Cossi, M., Rega, N., Scalmani, G. & Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. J Comput Chem 24, 669–681 (2003).
91. Barone, V. & Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. J. Phys. Chem. A 102, 1995–2001 (1998).
92. Schutz, C. N. & Warshel, A. What are the dielectric "constants" of proteins and how to validate electrostatic models? Proteins: Structure, Function, and Bioinformatics 44, 400–417 (2001).
93. Li, L., Li, C., Zhang, Z. & Alexov, E. On the Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. J. Chem. Theory Comput. 9, 2126–2136 (2013).
94. Chen, K., Zhang, S.-Q., Brandenberg, O. F., Hong, X. & Arnold, F. H. Alternate heme ligation steers activity and selectivity in engineered cytochrome P450-catalyzed carbene-transfer reactions. J. Am. Chem. Soc. 140, 16402–16407 (2018).
95. Sharon, D. A., Mallick, D., Wang, B. & Shaik, S. Computation sheds insight into iron porphyrin carbenes' electronic structure, formation, and N–H insertion reactivity. J. Am. Chem. Soc. 138, 9597–9610 (2016).
96. Lewis, R. D. et al. Catalytic iron-carbene intermediate revealed in a cytochrome c carbene transferase. Proc. Natl. Acad. Sci. USA 115, 7308–7313 (2018).
97. Huang, X. et al. A biocatalytic platform for synthesis of chiral α-trifluoromethylated organoborons. ACS Cent. Sci. 5, 270–276 (2019).
98. Yang, Y., Cho, I., Qi, X., Liu, P. & Arnold, F. H. An enzymatic platform for the asymmetric amination of primary, secondary and tertiary C(sp3)–H bonds. Nat. Chem. 11, 987–993 (2019).
99. Bauernschmitt, R. & Ahlrichs, R. Stability analysis for solutions of the closed shell Kohn–Sham equation. The Journal of Chemical Physics 104, 9047–9052 (1996).
100. Seeger, R. & Pople, J. A. Self-consistent molecular orbital methods. XVIII. Constraints and stability in Hartree–Fock theory. The Journal of Chemical Physics 66, 3045–3050 (1977).

101. Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput. 9, 3878–3888 (2013).
102. D.A. Case, T. E. Cheatham, III, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K. M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, C. L. Simmerling, W. M. Botello-Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, L. Xiao, D. M. York, P. A. Kollman. Amber 2016 Reference Manual. (2017).
103. The Pymol Molecular Graphics System. Schrodinger LLC (2017).
104. Li, P. & Merz, K. M. Jr. MCPB.py: A Python Based Metal Center Parameter Builder. J. Chem. Inf. Model. 56, 599–604 (2016).
105. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. Journal of Computational Chemistry 25, 1157–1174 (2004).
106. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J. Phys. Chem. 97, 10269–10280 (1993).
107. Besler, B. H., Merz Jr., K. M. & Kollman, P. A. Atomic charges derived from semiempirical methods. Journal of Computational Chemistry 11, 431–439 (1990).
108. Singh, U. C. & Kollman, P. A. An approach to computing electrostatic charges for molecules. Journal of Computational Chemistry 5, 129–145 (1984).
109. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics 79, 926–935 (1983).
110. Maier, J. A. et al. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. J Chem Theory Comput 11, 3696–3713 (2015).
111. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. The Journal of Chemical Physics 98, 10089–10092 (1993).
112. Roe, D. R. & Cheatham, T. E. I. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J. Chem. Theory Comput. 9, 3084–3095 (2013).

Appendix A: Explanation of Common (ML) Software Engineering Terms

Abstract Class:
A type of parent class from which objects cannot be instantiated from. Abstract classes are used to define a set of classes that should all have the same behavior. A key concept of an abstract class is the abstract method, which a class inheriting from the abstract class has to implement its own version of for the program to compile.

Attribute:
The set of variables that define an object.

Child Class:
A class that inherited from another class. In other words, a class that defines its own methods but that also has access to the methods of the parent class.

Class:
A class in object-oriented programming is a template that defines methods and attributes of an object that is an instance of it.

Dependency Injection:
A type of coding pattern in which the required input of an object is defined by an abstract class. This ensures that whatever the exact nature of the input is does not matter as long as it has implemented the methods of the abstract class.

Dictionary:
A dictionary is a data structure that stores data in key/value pairs. The value in a dictionary can be accessed by providing the dictionary with a key. An important property of dictionaries is that each key appears at most once in the data structure.

Factory Pattern:
A type of coding pattern in which relays information to be processed to the correct function. The factory pattern achieves this by using a dictionary in which the key is used to define what function should be executed and the value of the dictionary is the function to be executed. The factory pattern is useful in not having to write long if/else statements.

Method:
A function of a class or object. Just like a function, a method takes inputs and manipulates them. However, a method also has access to the attributes of the object it is executed from.

MLM:
A masked language model (MLM) is an unsupervised machine learning model in which a token in a sequence is masked and the task for the model is to guess the correct token in the sequence.

NLP:
A class of machine learning models that learns the structure of written and spoken language.

Object:
An object is a data structure that stores specific variables (attributes) and methods on how to manipulate them. An object is always in instance of a class which defines it.

Private Attribute:
An attribute that can only be accessed by the object itself and no external objects.

PLM:
A class of machine learning models that learns the language of protein sequences, their properties and structures.

Appendix B: Other Contributions

In addition to the work presented in this thesis, I also contributed to the following paper:
S. Petrovic, D. Samanta, T. Perriches, C. J. Bley, K. Thierbach, B. Brown, S. Nie, G. W. Mobbs, T. A. Stevens, X. Liu, G. P. Tomaleri, L. Schaus, A. Hoelz, Architecture of the linker-scaffold in the nuclear pore, *Science* 2022, 376, 6598. doi:10.1126/science.abm9798

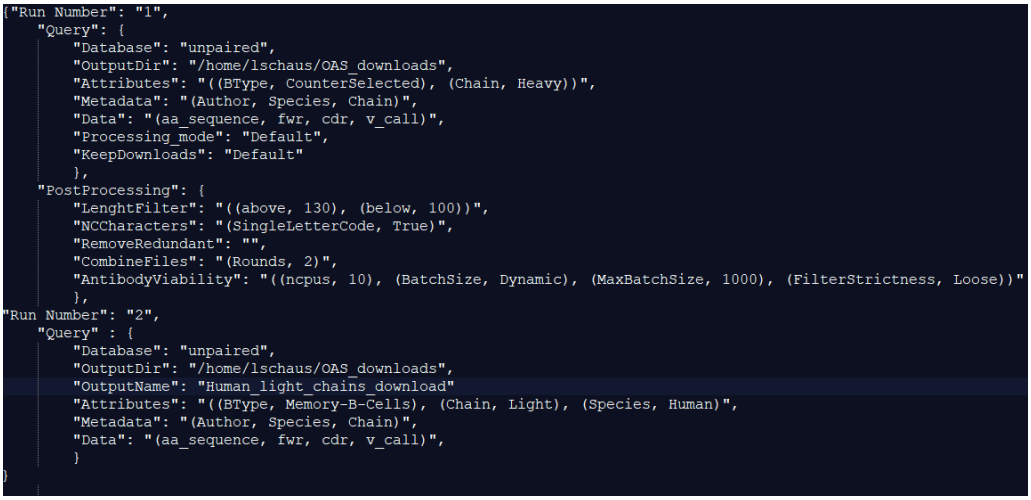
Together with S. Petrovic, I solved determined the crystal growth conditions and solved the structure of H. sapiens $Nup93^{SOL}$ in complex with $Nup53^{R2}$.

I also contributed to the development of projects and helped in writing grants and fellowships for the following programs, all of which have been funded:

- Resnick Impact Grant
- AFR PhD Fellowship
- NSF CBE Grant
- DOE Data Science to Advance Chemical and Material Sciences Grant
- Amgen Foundation Grant
- Nvidia GPU Grant Program

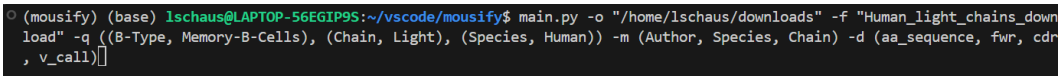
Appendix C: Chapter 2 Supplementary Figures

OAS C/S



```
{"Run Number": "1",
    "Query": {
        "Database": "unpaired",
        "OutputDir": "/home/lschaus/OAS_downloads",
        "Attributes": "((BType, CounterSelected), (Chain, Heavy))",
        "Metadata": "(Author, Species, Chain)",
        "Data": "(aa_sequence, fwr, cdr, v_call)",
        "Processing_mode": "Default",
        "KeepDownloads": "Default"
        },
    "PostProcessing": {
        "LenghtFilter": "((above, 130), (below, 100))",
        "NCCharacters": "(SingleLetterCode, True)",
        "RemoveRedundant": "",
        "CombineFiles": "(Rounds, 2)",
        "AntibodyViability": "((ncpus, 10), (BatchSize, Dynamic), (MaxBatchSize, 1000), (FilterStrictness, Loose))"
        },
"Run Number": "2",
    "Query" : {
        "Database": "unpaired",
        "OutputDir": "/home/lschaus/OAS_downloads",
        "OutputName": "Human_light_chains_download"
        "Attributes": "((BType, Memory-B-Cells), (Chain, Light), (Species, Human)",
        "Metadata": "(Author, Species, Chain)",
        "Data": "(aa_sequence, fwr, cdr, v_call)",
        }
}
```

Figure SC.1: Example of a query file for OAS C/S. The program will parse this file and start two downloading runs. The first run will download all available heavy chains from B-Cells that have been counter-selected (i.e. Memory B-Cells, Plasma B-Cells, Germinal Center B-Cells etc.). The data kept from the OAS files will be information on the author of the study, the species of origin, the chain type (i.e. IGHG, IGHE etc.), the amino acid sequence, the framework and CDR sequences, and the V-gene type. After processing the files, the listed post-processing modules will be passing over the data in the order listed, with the options listed.



```
(mousify) (base) lschaus@LAPTOP-56EGIP9S:~/vscode/mousify$ main.py -o "/home/lschaus/downloads" -f "Human_light_chains_down
load" -q ((B-Type, Memory-B-Cells), (Chain, Light), (Species, Human)) -m (Author, Species, Chain) -d (aa_sequence, fwr, cdr
, v_call)
```

Figure SC.2: Example of the command line interface (CLI) of OAS C/S. This CLI query results in the same download as “Run Number: 2” from Figure SC.1. In this case OAS C/S has not been added to PATH and therefore main.py has to be run in the folder it is stored in.

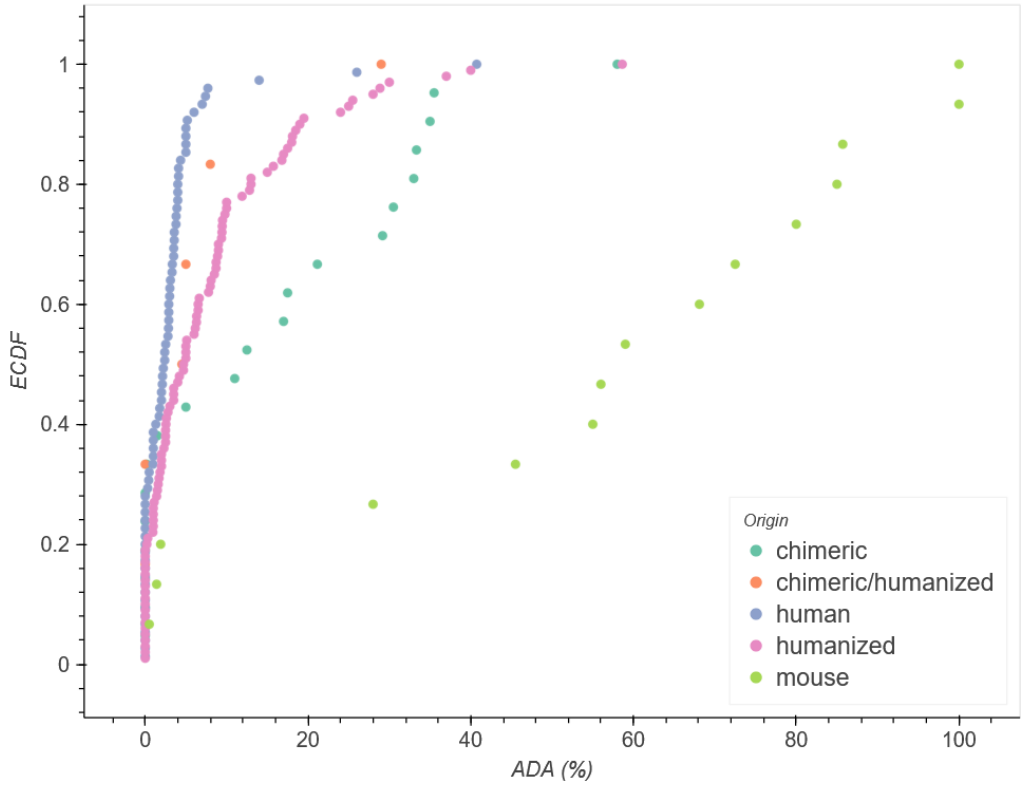


Figure SC.3: Empirical cumulative density function (ECDF) of the percentage of patients that produce anti-drug antibodies (ADA) when treated with a monoclonal antibody, separated by the origin of the antibody.

Antibody	VL Grafted Gene	VH Grafted Gene
Certolizumab	IGKV1-16	IGHV7-4
Omalizumab	IGKV1-3	IGHV4-38
Palivizumab	IGKV1-39	IGHV2-5
M8a-3	IGKV4-1	IGVH1-3

Table SC2.1: Framework genes that monoclonal antibodies' CDRs have been grafted onto. CDR grafting was performed via the BioPhi web application.

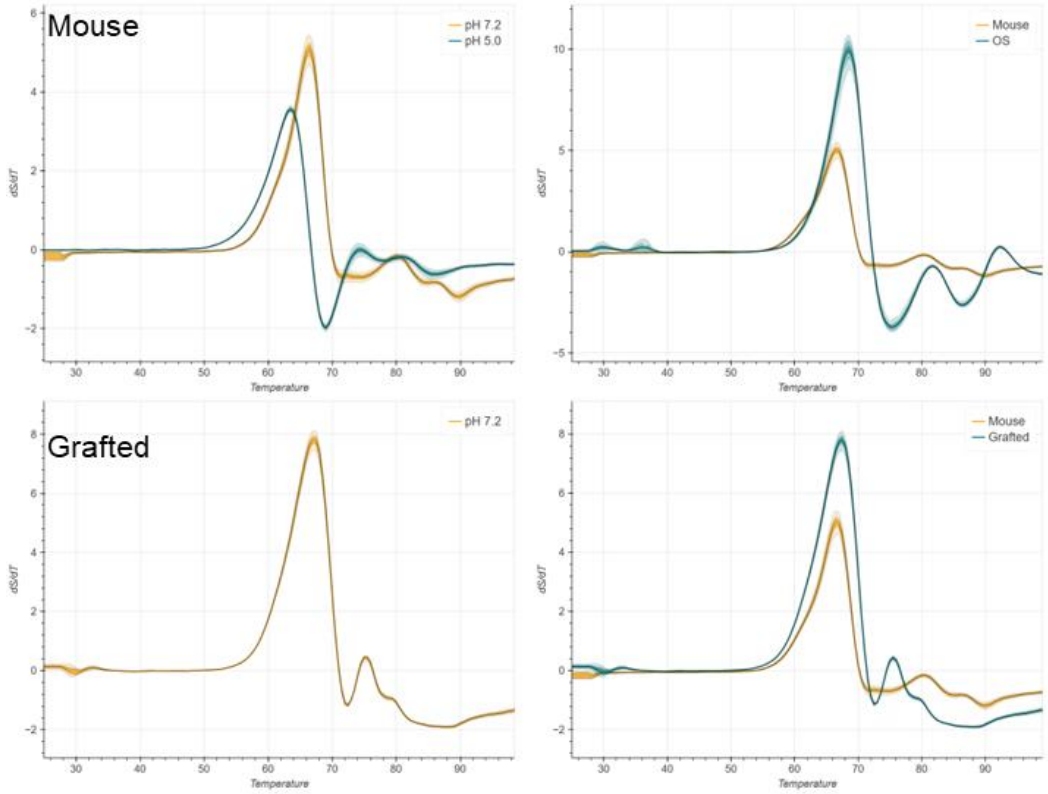


Figure SC.4: Plots used for peak assignments for M8a-3 “Mouse” and “Grafted” thermofluor assays.

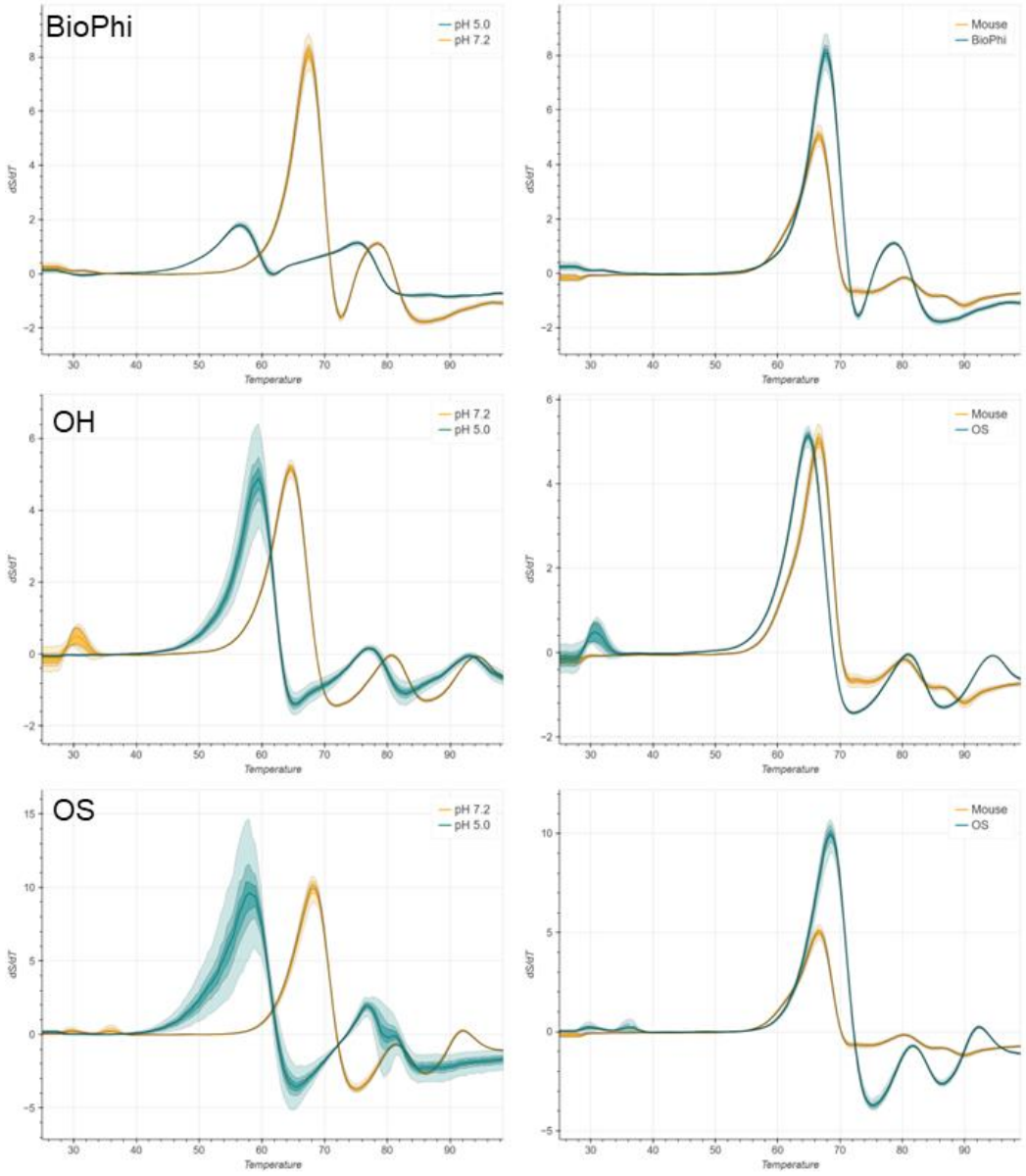


Figure SC.5: Plots used for peak assignments for M8a-3 “BioPhi”, “Mousify: OASis/Hu-mAb”, and “Mousify: OASis/Sapiens” thermofluor assays.

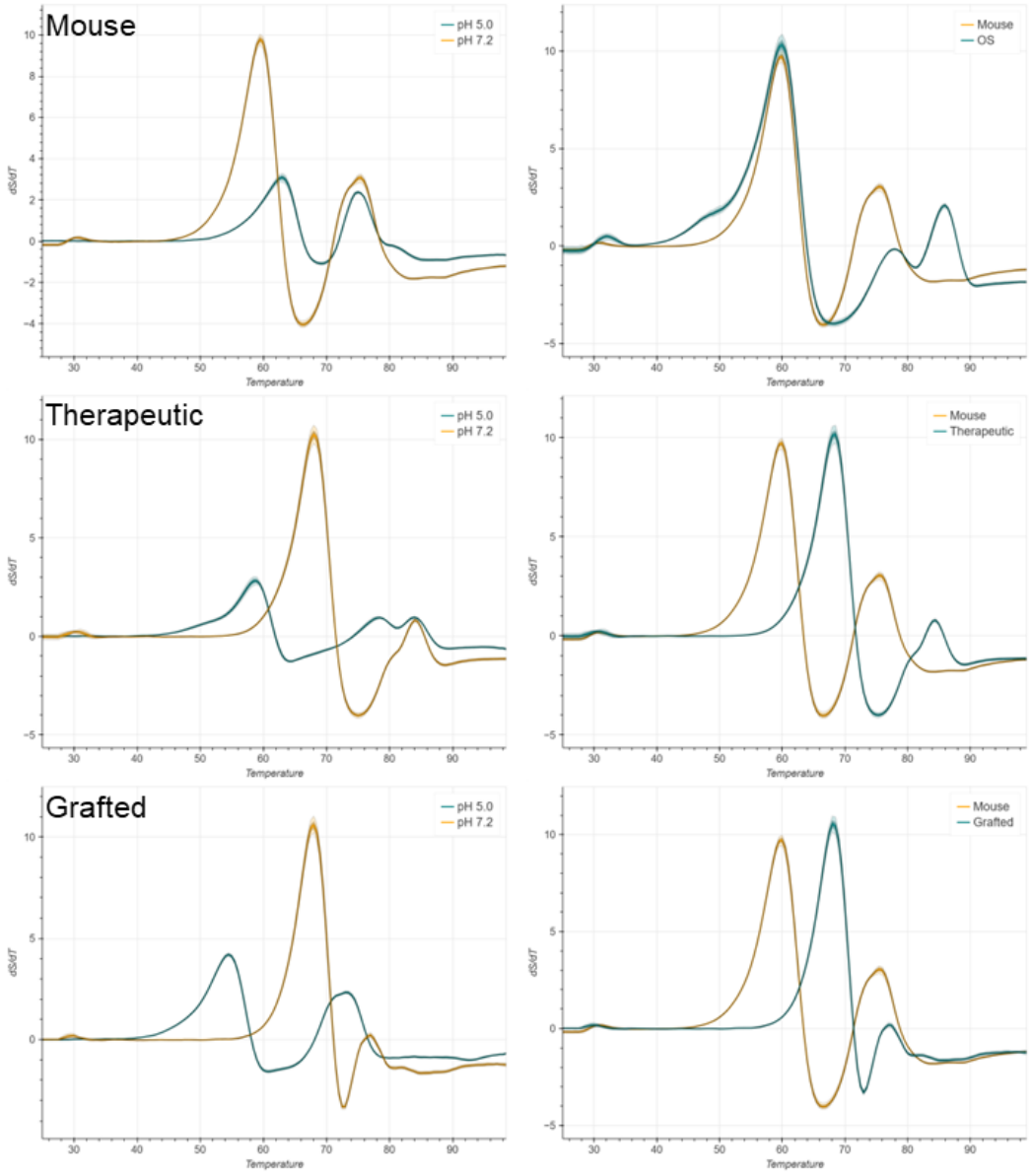


Figure SC.6: Plots used for peak assignments for Certolizumab "Mouse", "Therapeutic", and "Grafted" thermofluor assays.

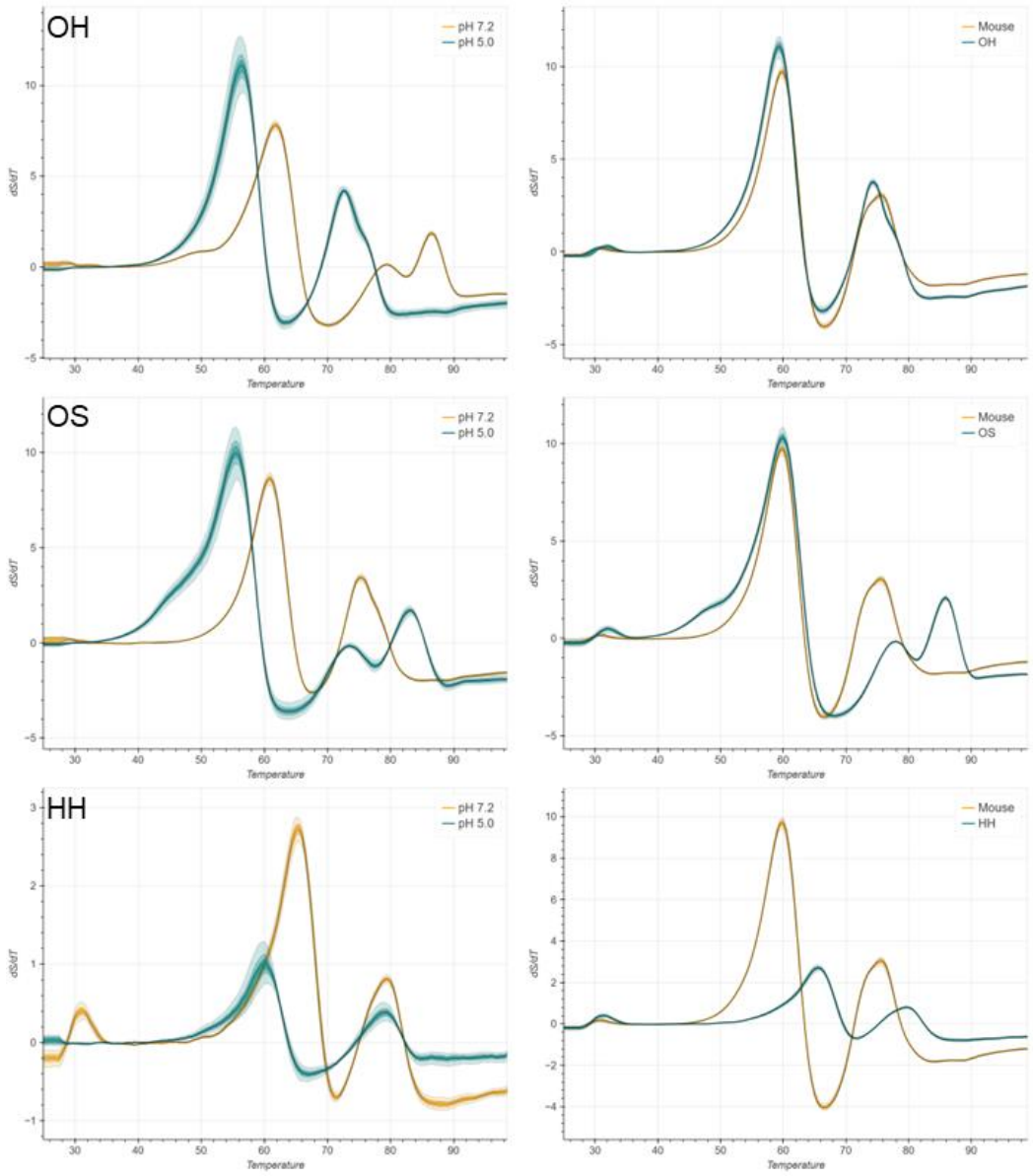


Figure SC.7: Plots used for peak assignments for Certolizumab “Mousify: OASis/Hu-mAb”, “Mousify: OASis/Sapiens” and “Mousify: Hu-mAb/Hu-mAb” thermofluor assays.

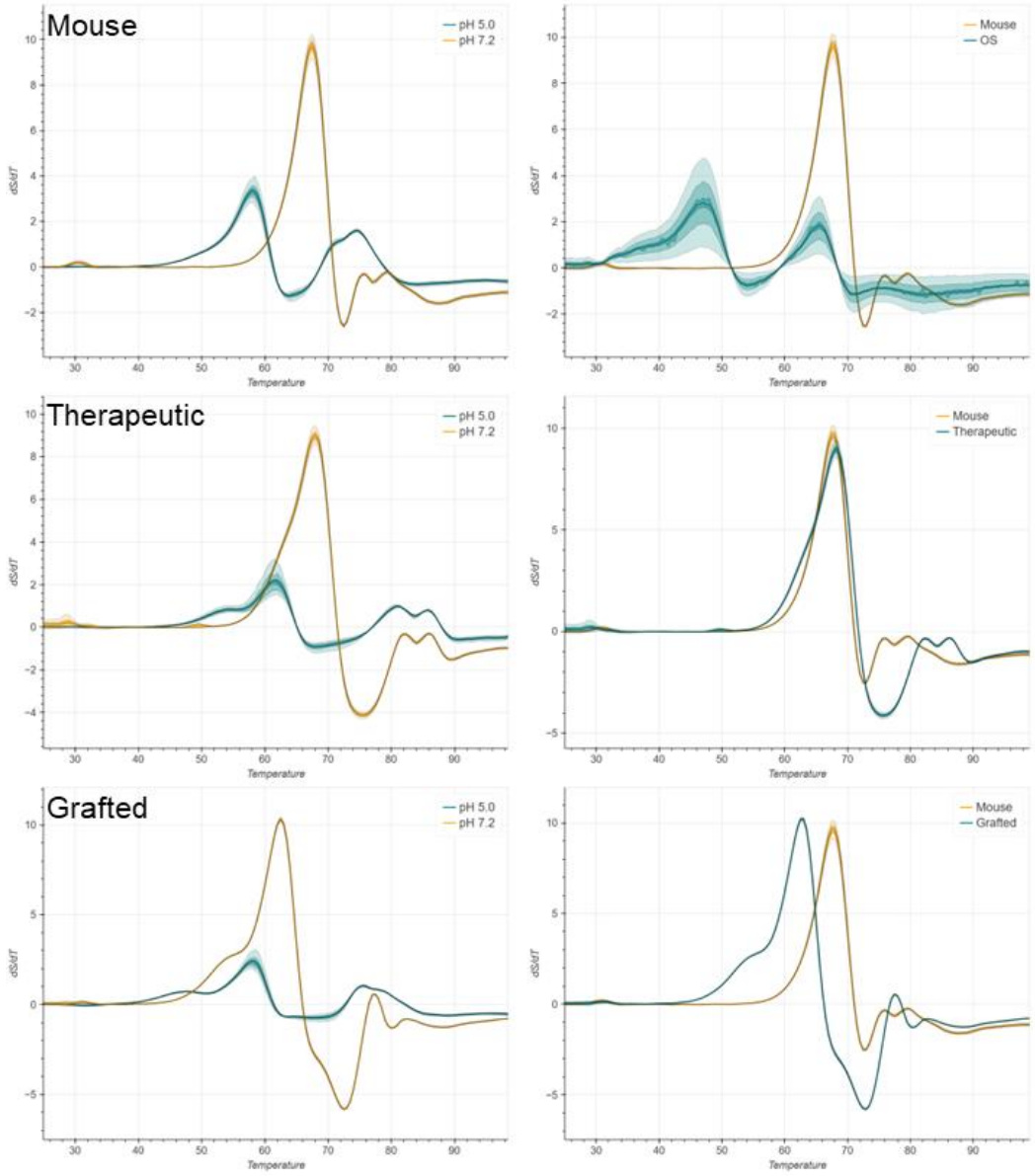


Figure SC.8: Plots used for peak assignments for Omalizumab "Mouse", "Therapeutic" and "Grafted" thermofluor assays.

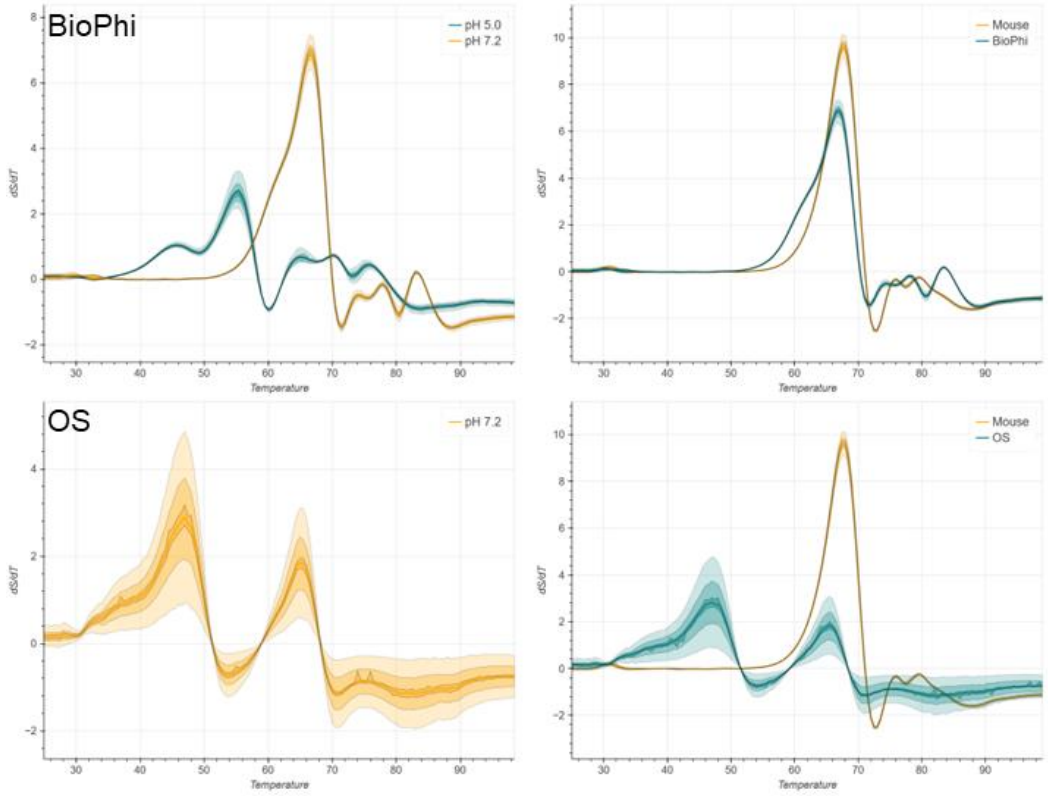


Figure SC.9: Plots used for peak assignments for Omalizumab “BioPhi” and “Mousify: OASis/Sapiens” thermofluor assays.

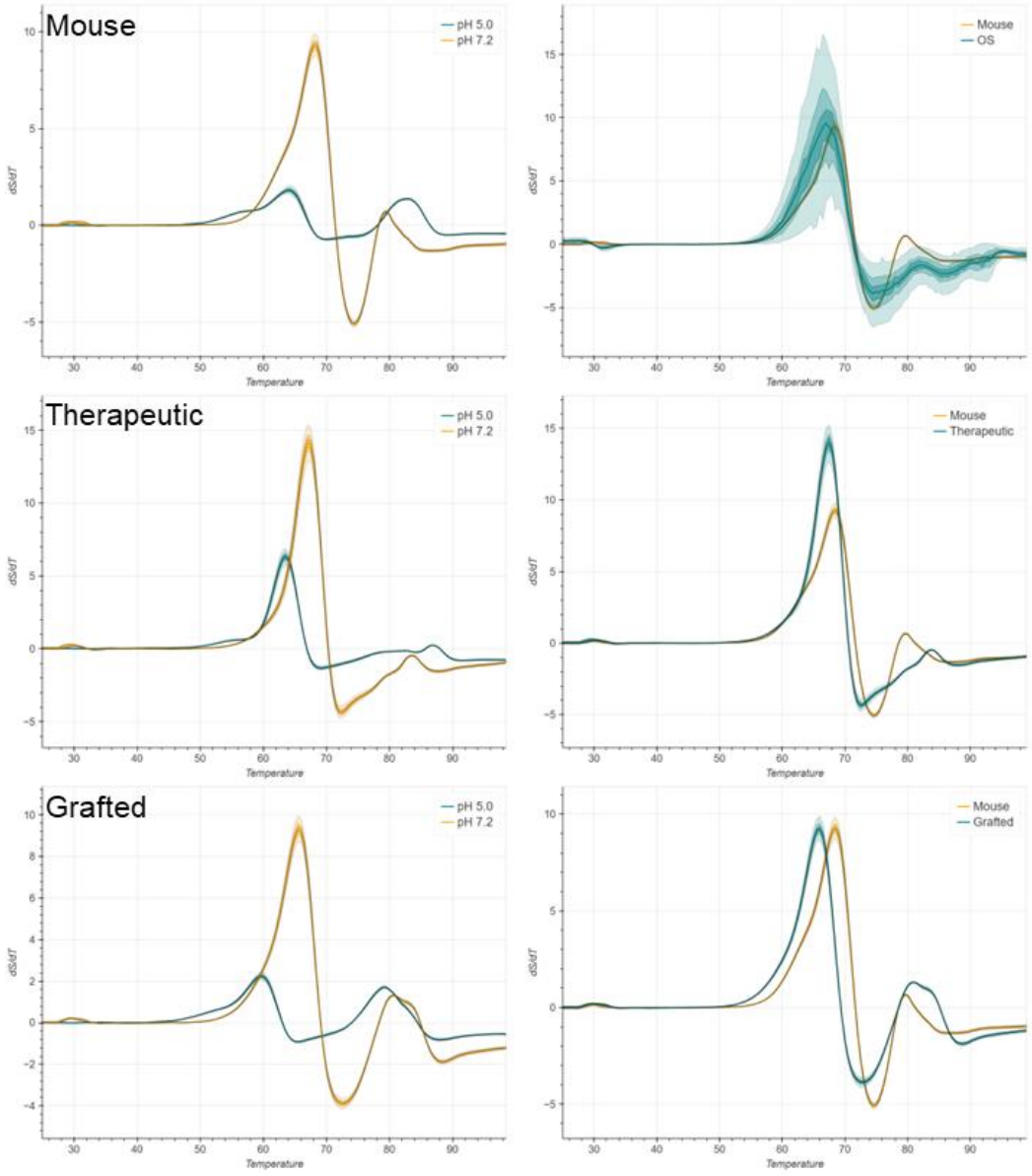


Figure SC.10: Plots used for peak assignments for Palivizumab “Mouse”, “Therapeutic”, and “Grafted” thermofluor assays.

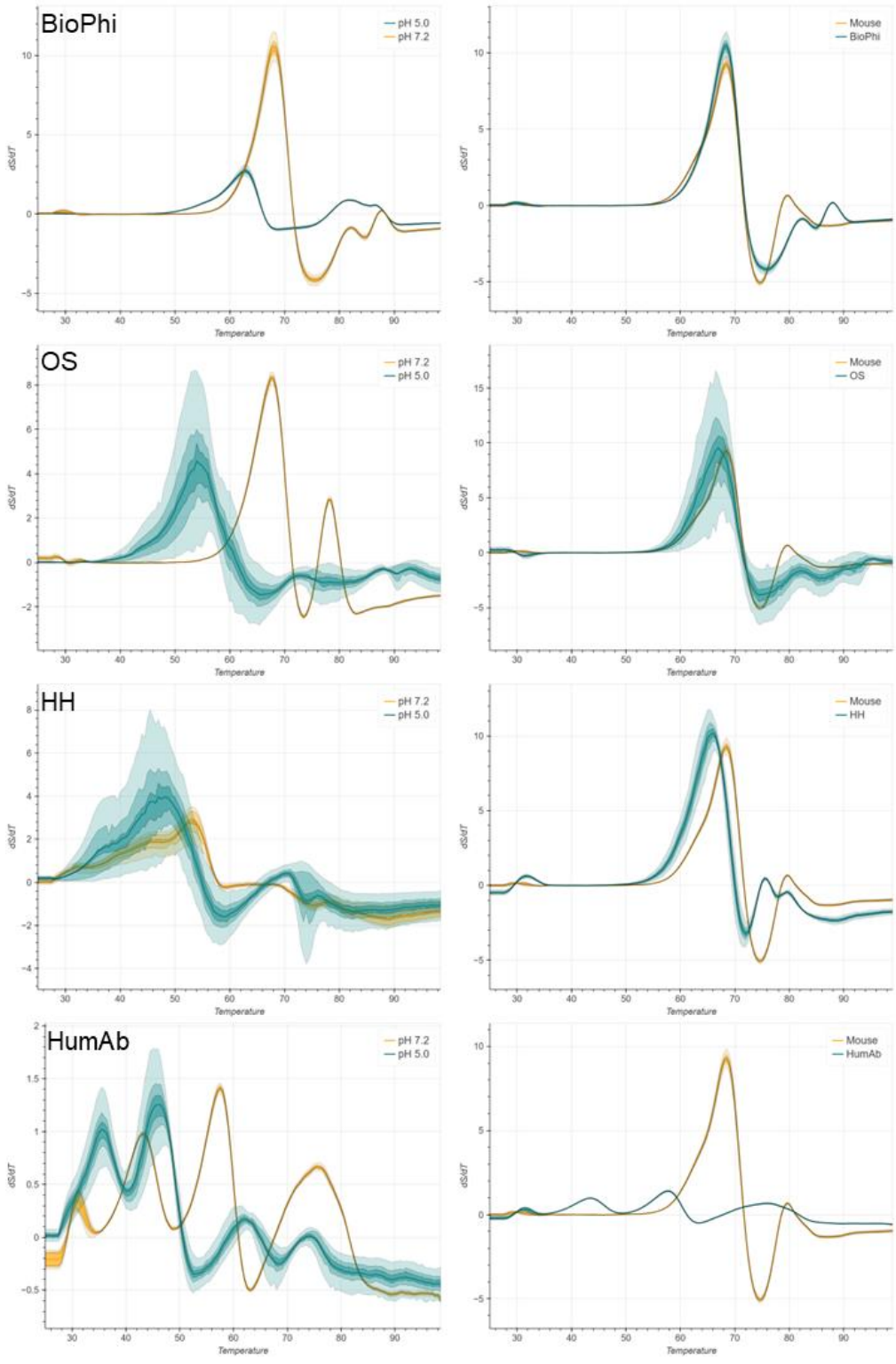


Figure SC.11: Plots used for peak assignments for Palivizumab "BioPhi", "Mousify: Oasis/Sapiens", "Mousify: Hu-mAb/Hu-mAb", and "Hu-mAb" thermofluor assays.

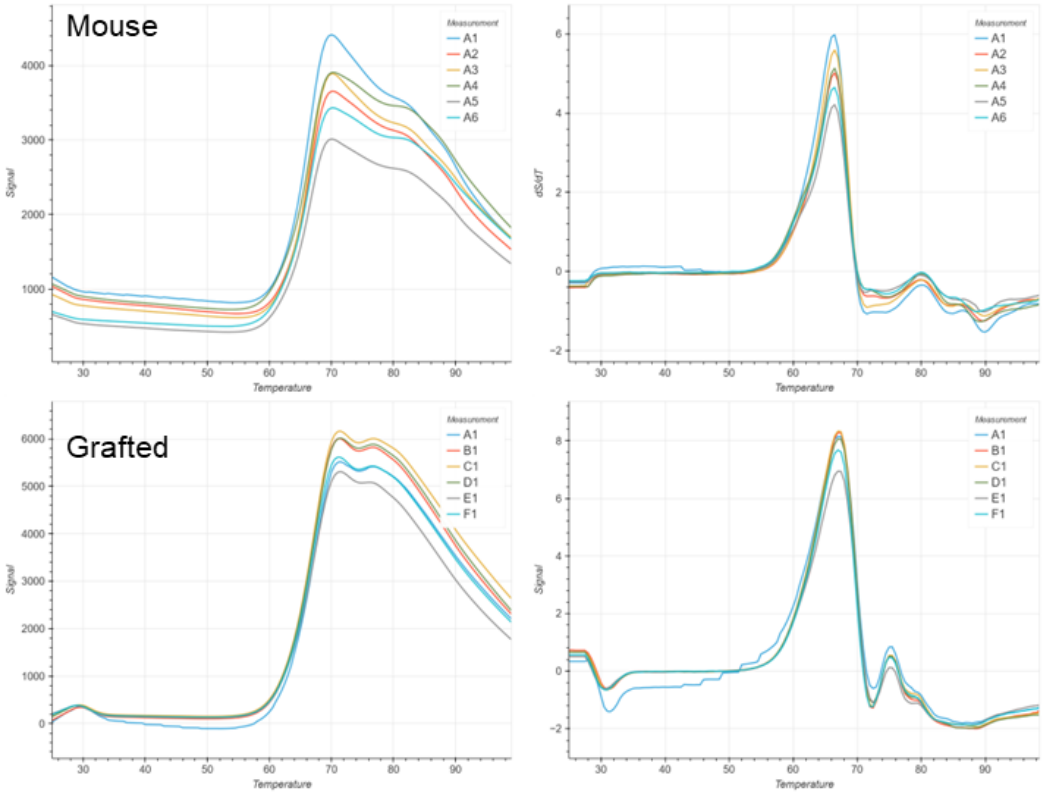


Figure SC.12: Raw data of thermofluor assay data and the derivative of the data of M8a-3 “Mouse” and “Grafted” antibodies.

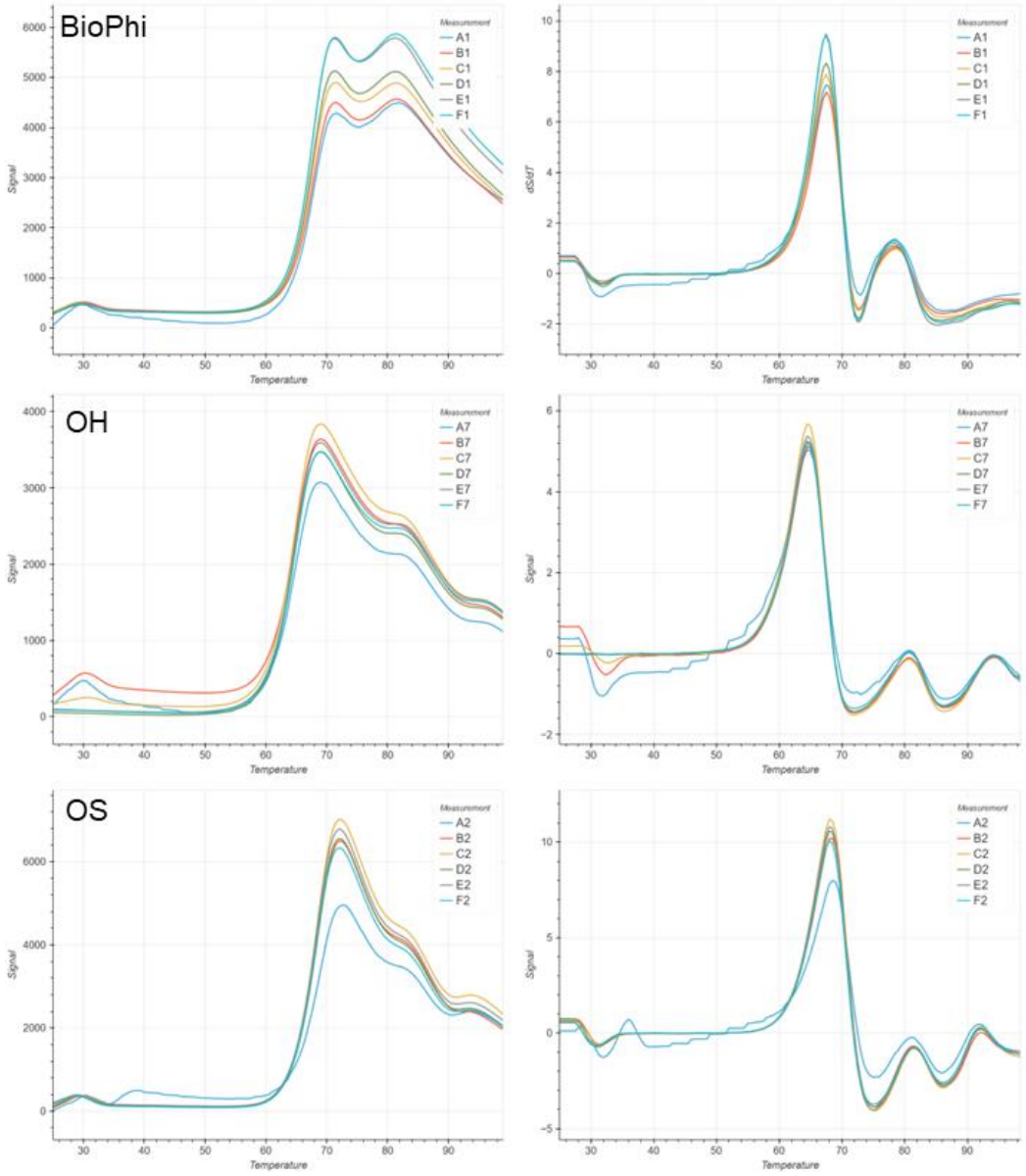


Figure SC.13: Raw data of thermofluor assay data and the derivative of the data of M8a-3 “BioPhi”, “Mousify: OASis/Hu-mAb” and “Mousify: OASis/Sapiens” antibodies.

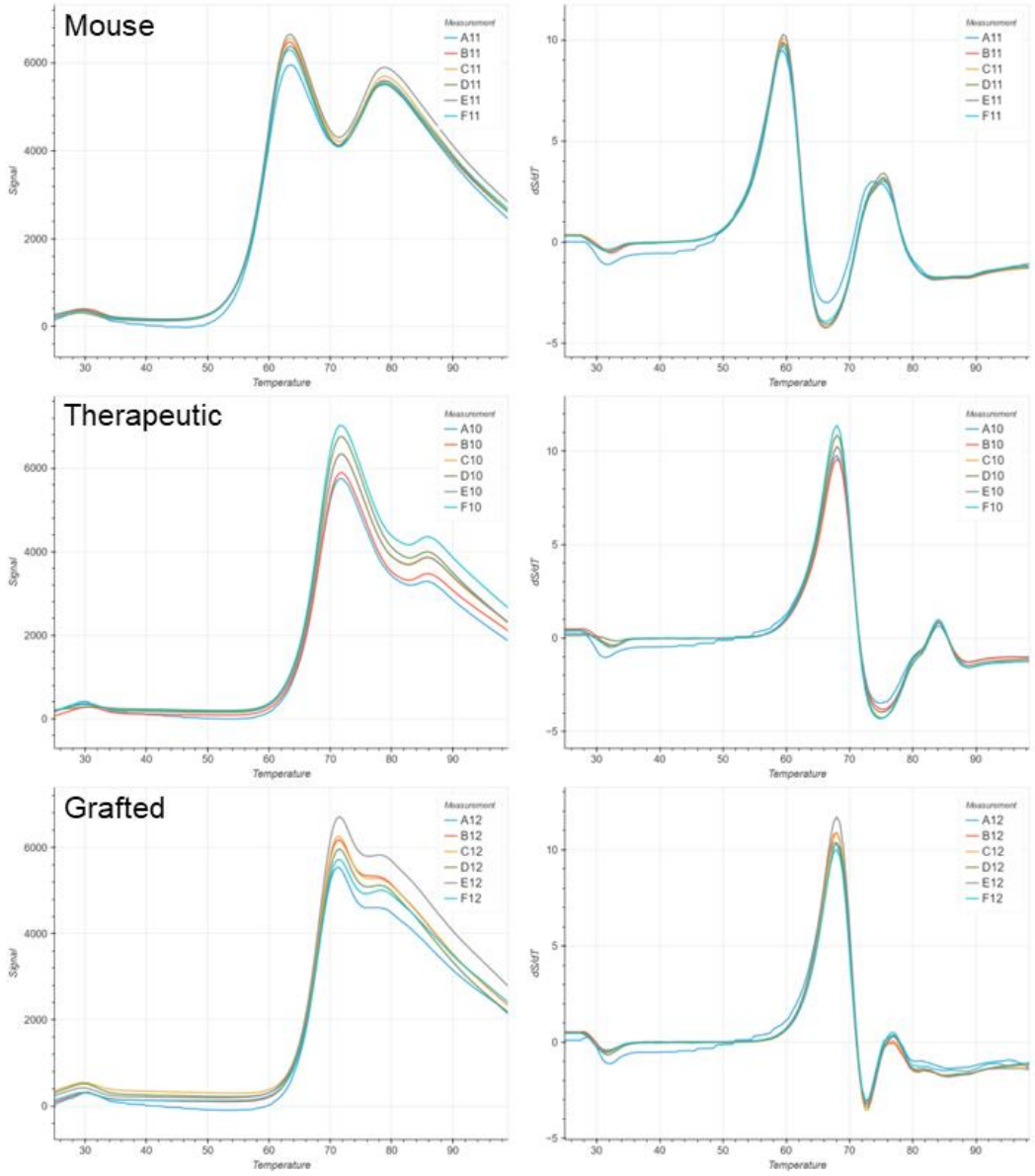


Figure SC.14: Raw data of thermofluor assay data and the derivative of the data of Certolizumab "Mouse", "Therapeutic", and "Grafted" antibodies.

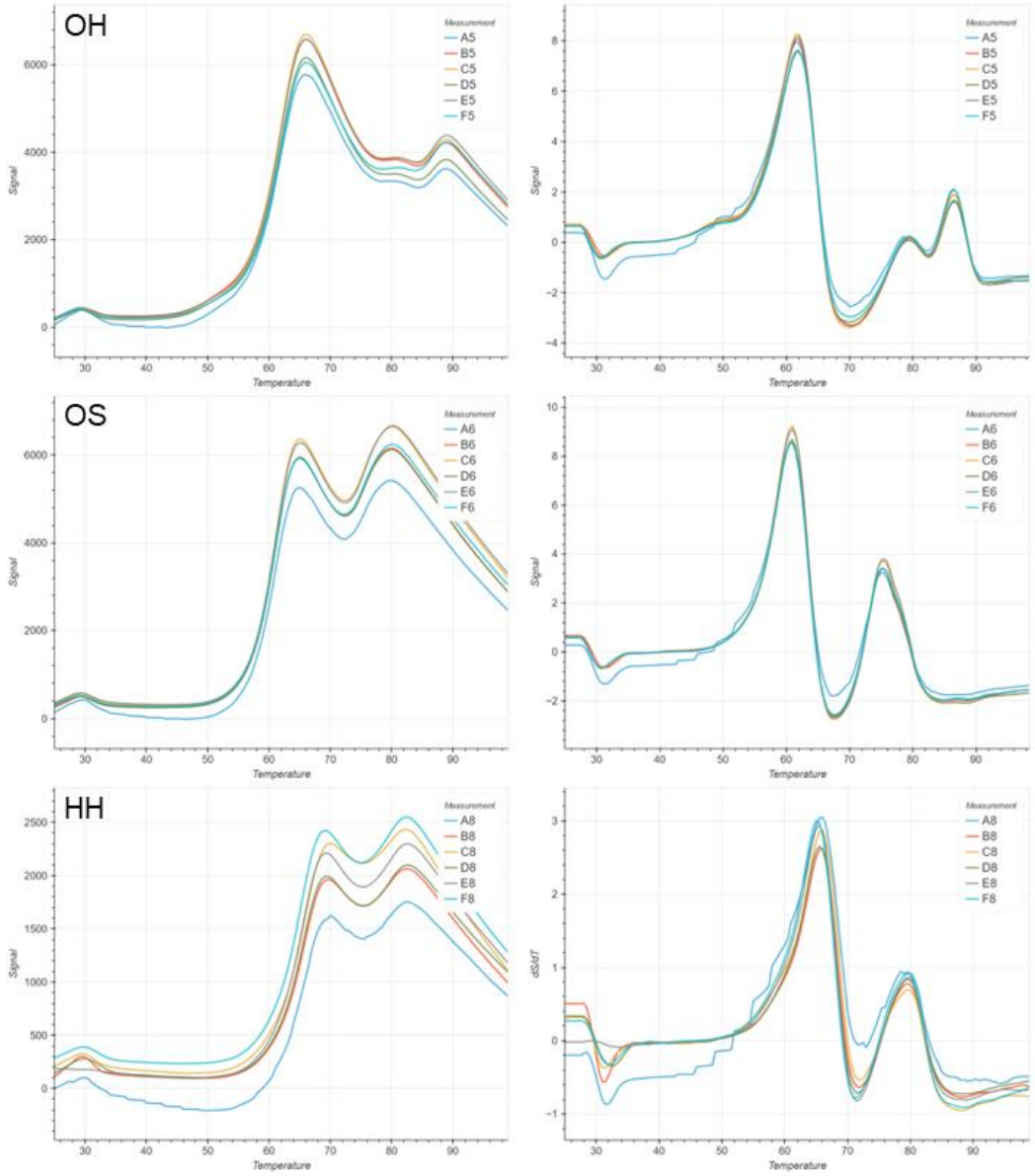


Figure SC.15: Raw data of thermofluor assay data and the derivative of the data of Certolizumab "Mousify: OASis/Hu-mAb", "Mousify: OASis/Sapiens", and "Mousify: Hu-mAb/Hu-mAb" antibodies.

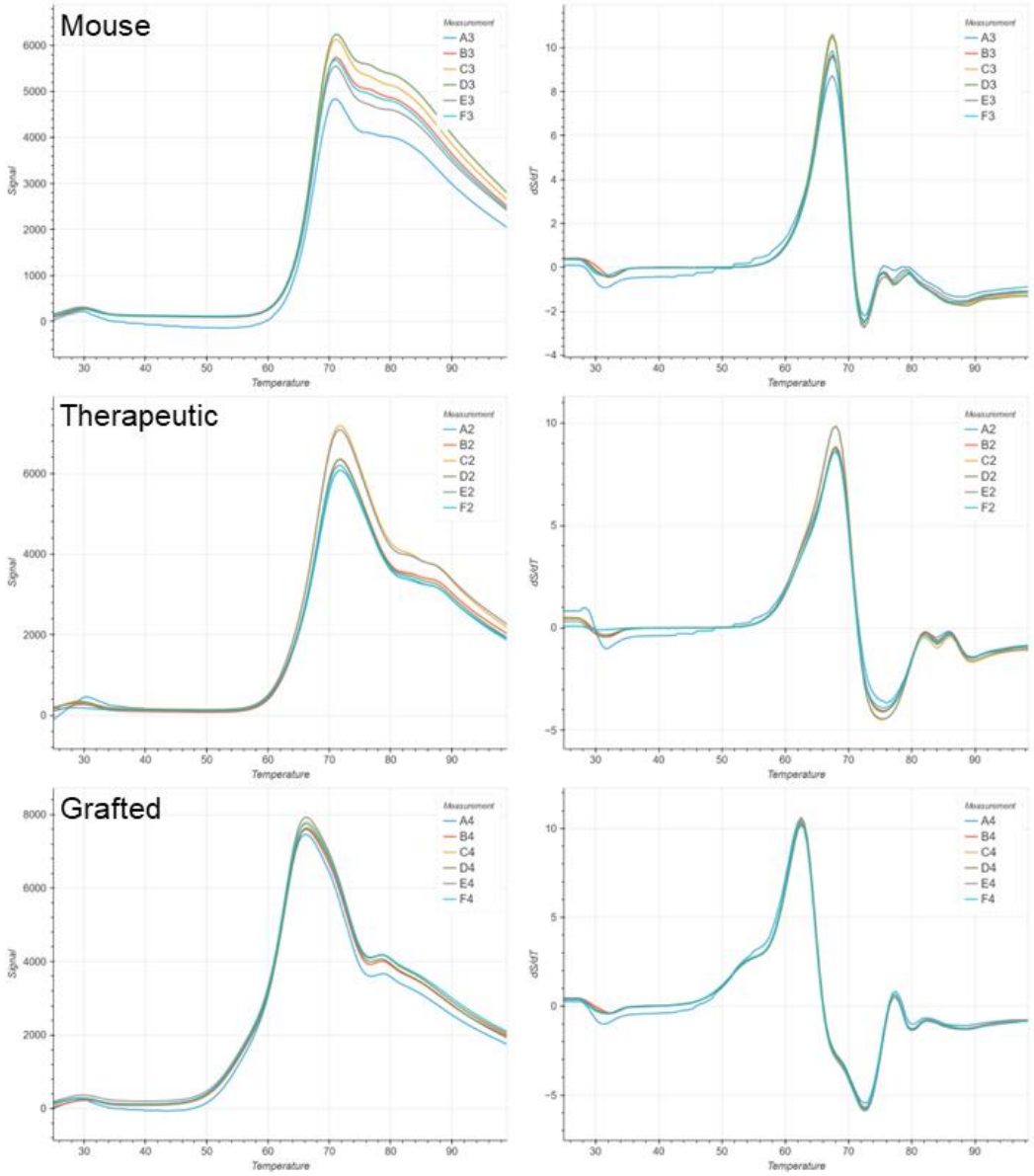


Figure SC.16: Raw data of thermofluor assay data and the derivative of the data of Omalizumab "Mouse", "Therapeutic", and "Grafted" antibodies.

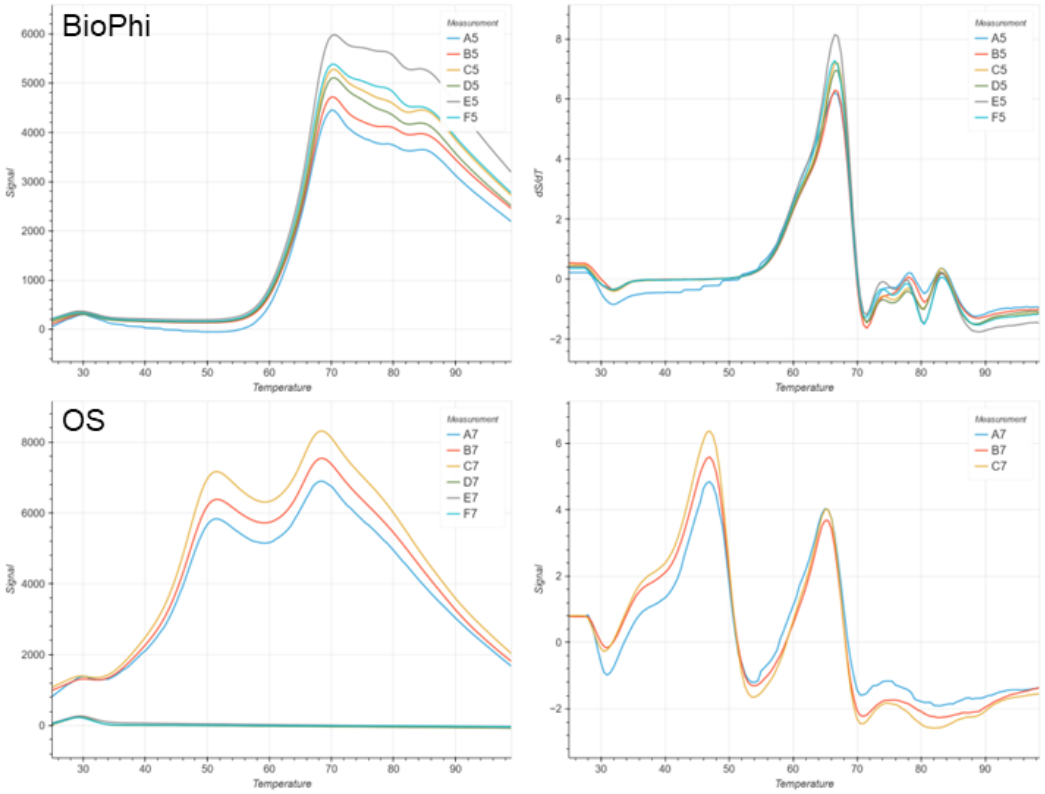


Figure SC.17: Raw data of thermofluor assay data and the derivative of the data of Omalizumab "BioPhi", and "Mousify: OASis/Sapiens" antibodies.

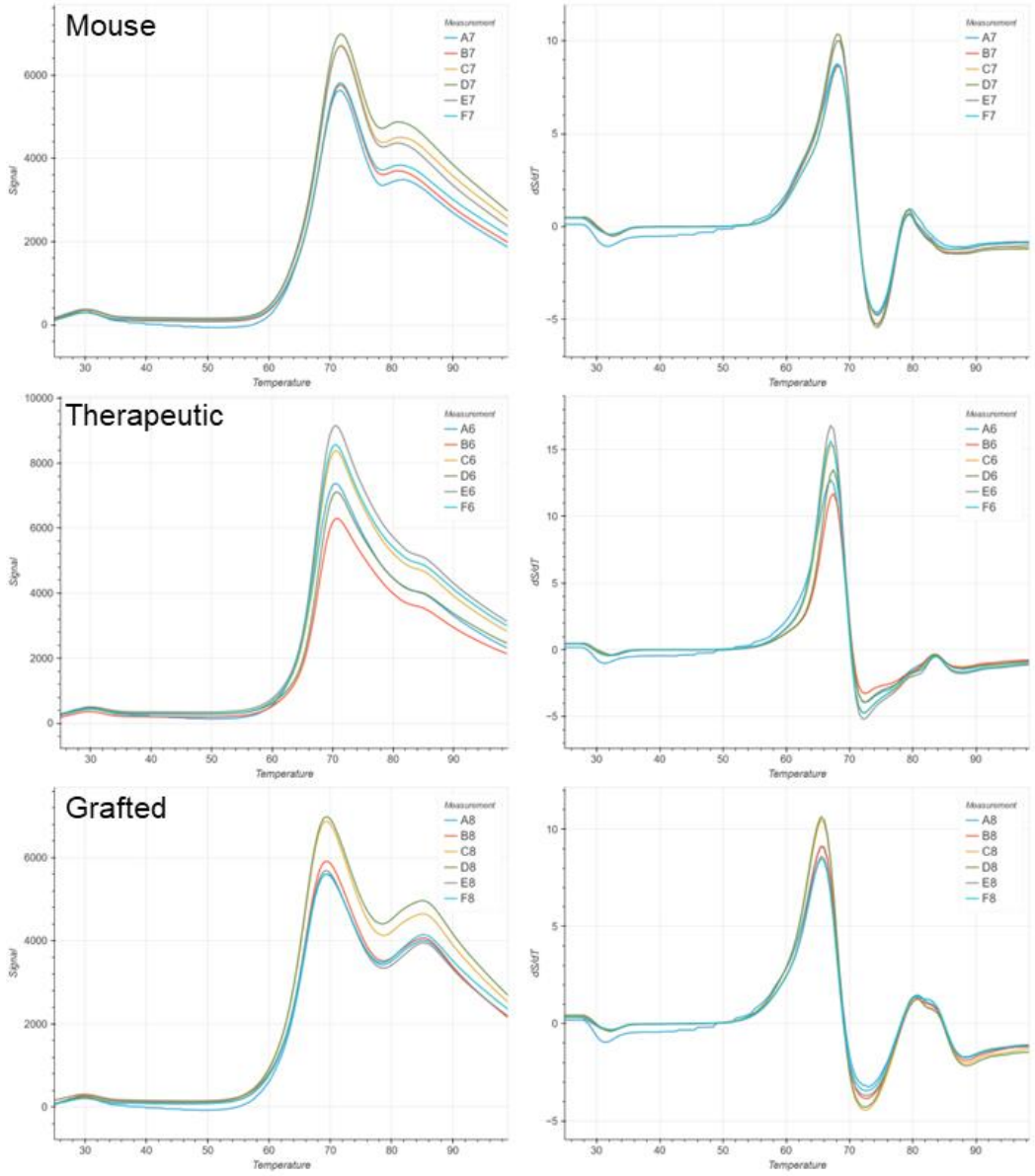


Figure SC.18: Raw data of thermofluor assay data and the derivative of the data of Palivizumab "Mouse", "Therapeutic", and "Grafted" antibodies.

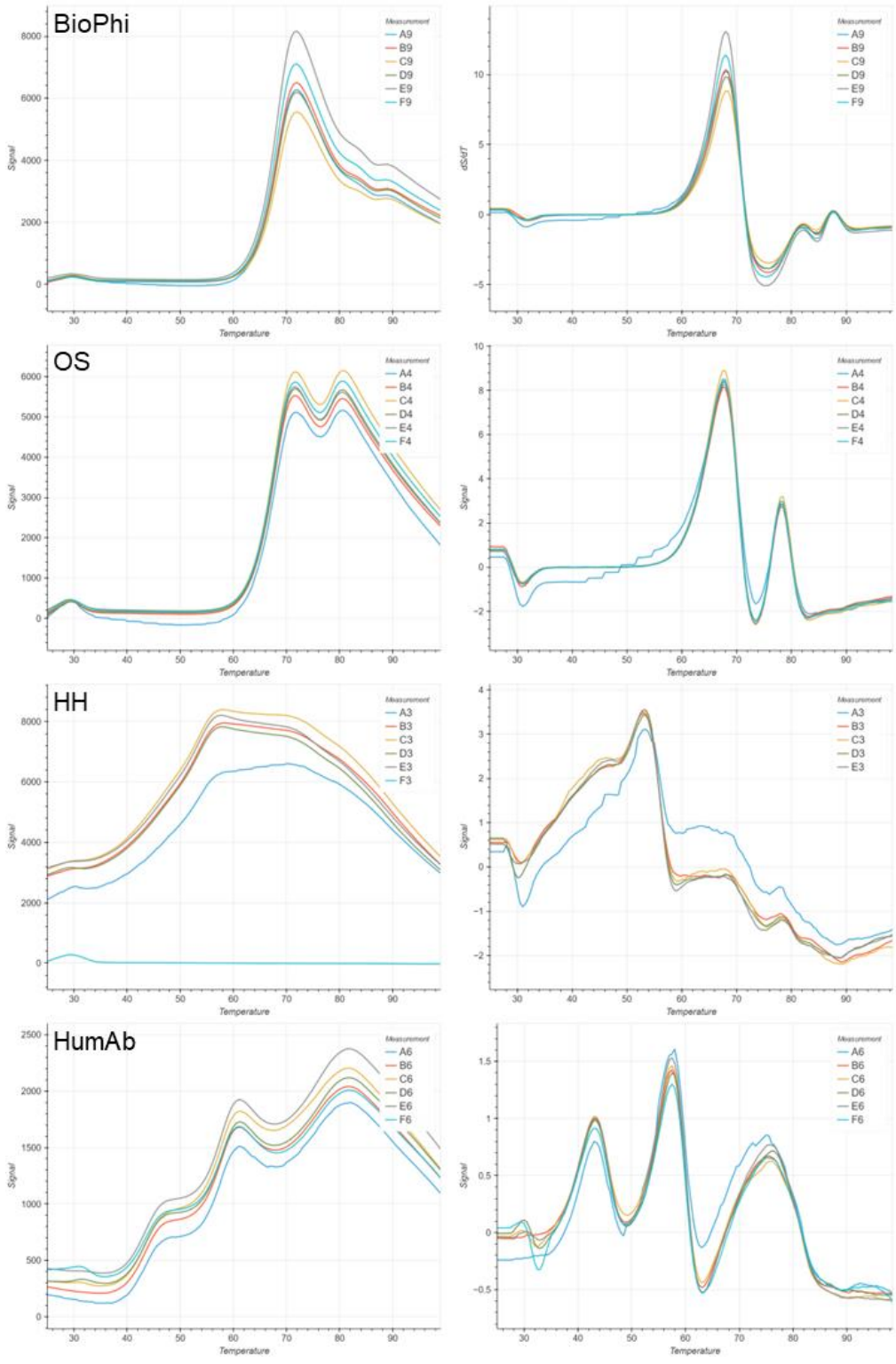
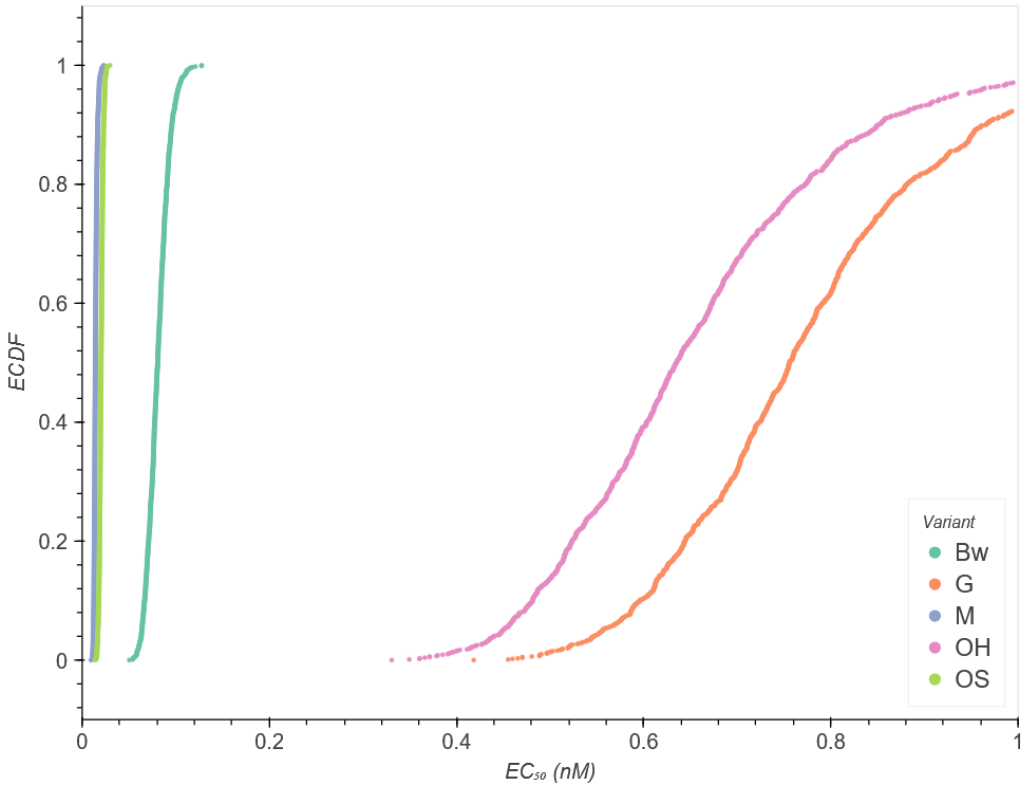
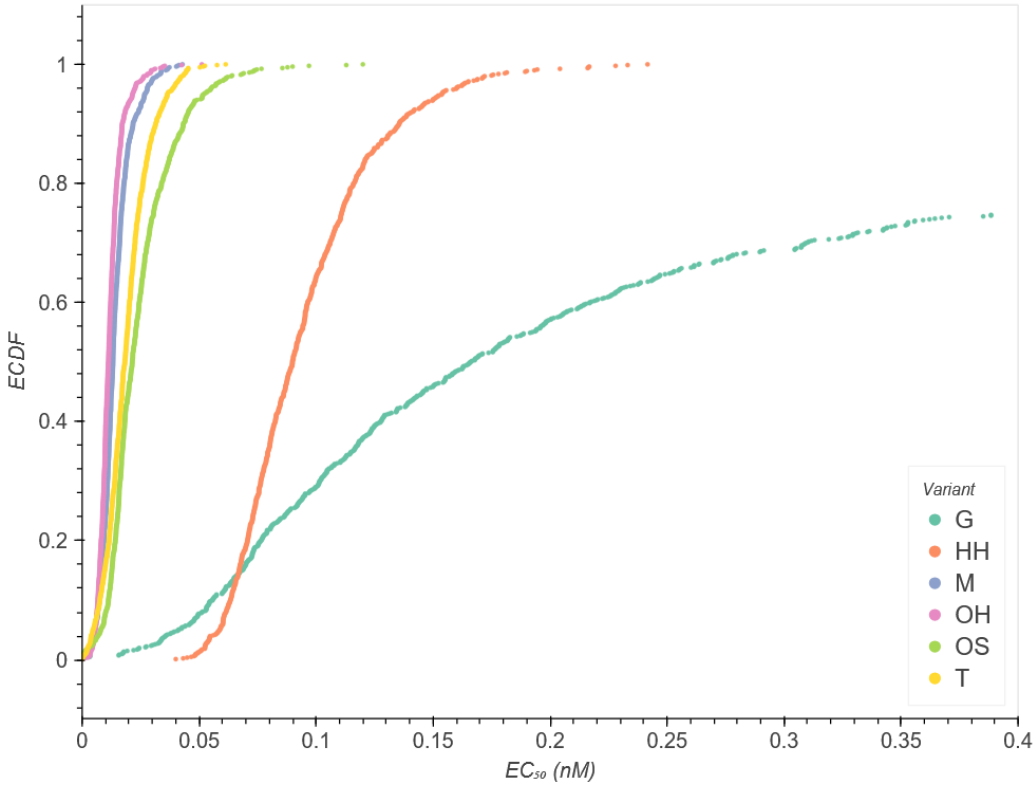


Figure SC.19: Raw data of thermofluor assay data and the derivative of the data of Palivizumab "BioPhi", "Mousify: OASis/Sapiens", "Mousify: Hu-mAb/Hu-mAb" and "Hu-mAb" antibodies.


ECDF
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
EC₅₀ (nM)
Variant
Bw
G
M
OH
OS


ECDF
1
0.8
0.6
0.4
0.2
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
EC₅₀ (nM)
Variant
G
HH
M
OH
OS
T

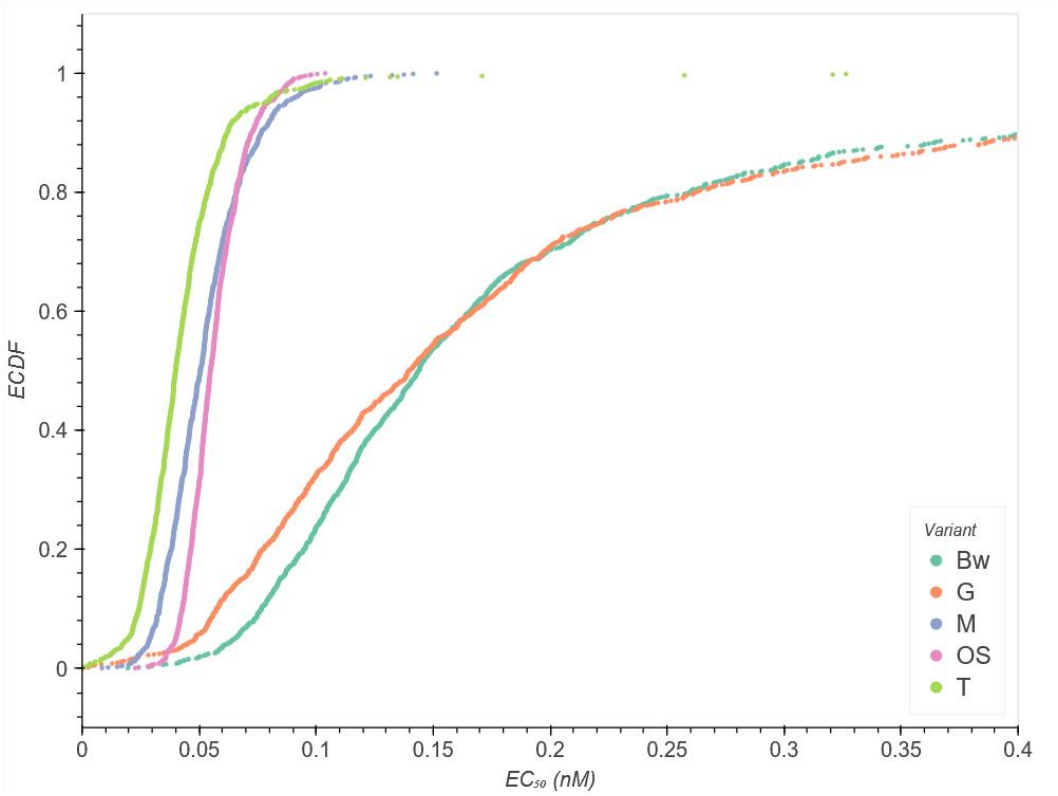


Figure SC.20: ECDF of EC_{50} values calculated from ELISA. (Top) M8a-3 EC_{50} values. (Middle) Certolizumab EC_{50} values. (Bottom) Omalizumab EC_{50} values.

Mousify Library Generation

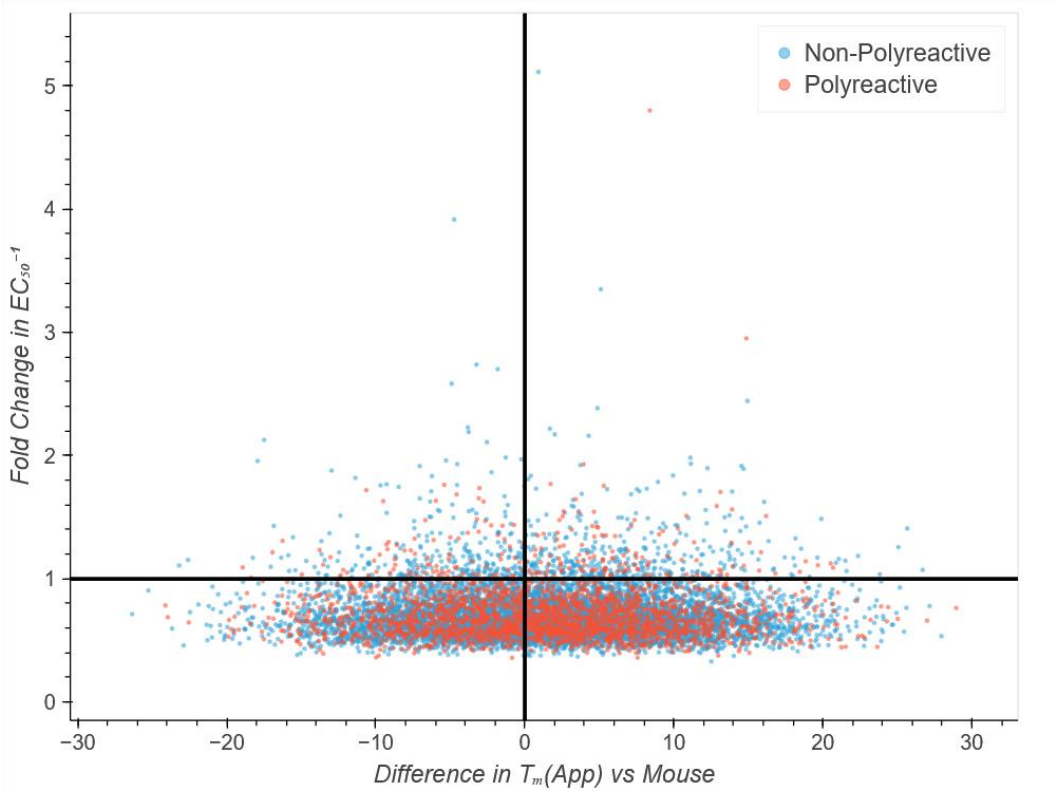


Figure SC.21: Sampled data from the Mousify OS model. Only 9.6% of antibodies would pass all experiments (At least no change in T_M, not polyreactive, expresses and maximum EC_{50} increase of 20%). Each point would represent an antibody with the properties listed. Samples were considered iid. to generate this figure. It is worth noting that the samples are likely not iid. in reality. For example, a very unstable antibody will likely also have very poor EC_{50}, or a very low EC_{50} antibody also has a higher chance of being polyreactive than an antibody that does not bind to the target antigen at all.

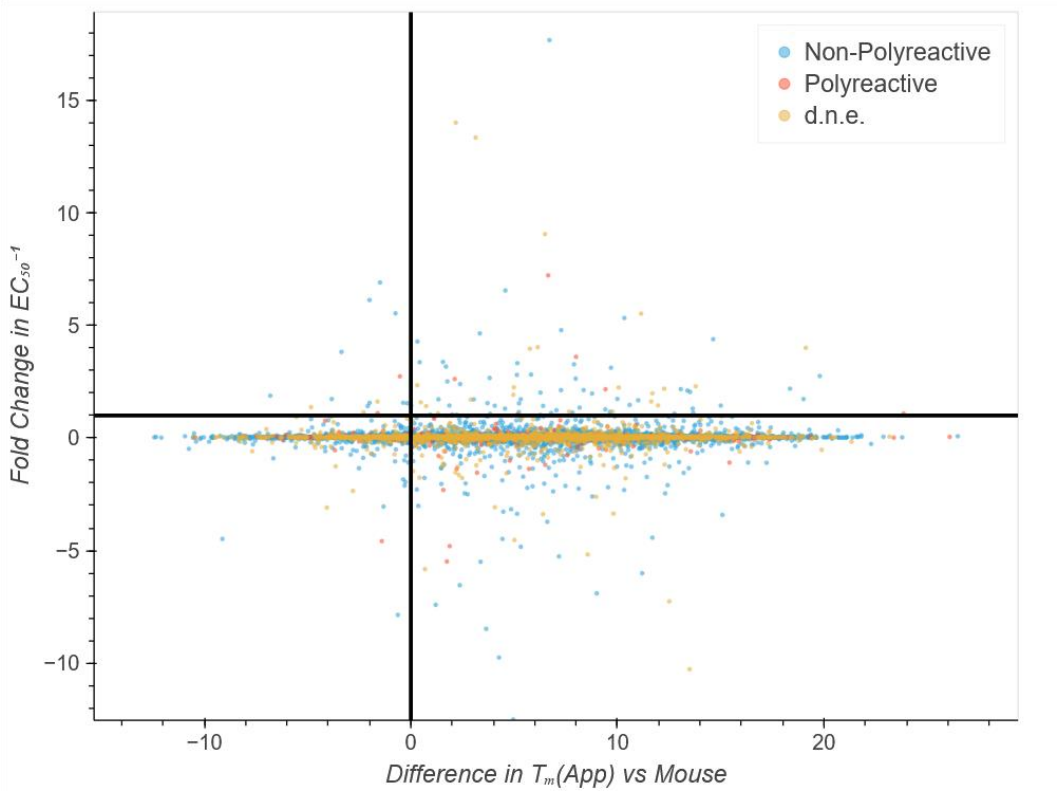


Figure SC.22: Sampled data from the Mousify OH model. Only 0.7% of antibodies would pass all experiments (At least no change in T_M, not polyreactive, expresses and maximum EC_{50} increase of 20%). Each point would represent an antibody with the properties listed. Samples were considered iid. to generate this figure. It is worth noting that the samples are likely not iid. in reality. For example, a very unstable antibody will likely also have very poor EC_{50}, or a very low EC_{50} antibody also has a higher chance of being polyreactive than an antibody that does not bind to the target antigen at all. D.n.e.: Does not express.

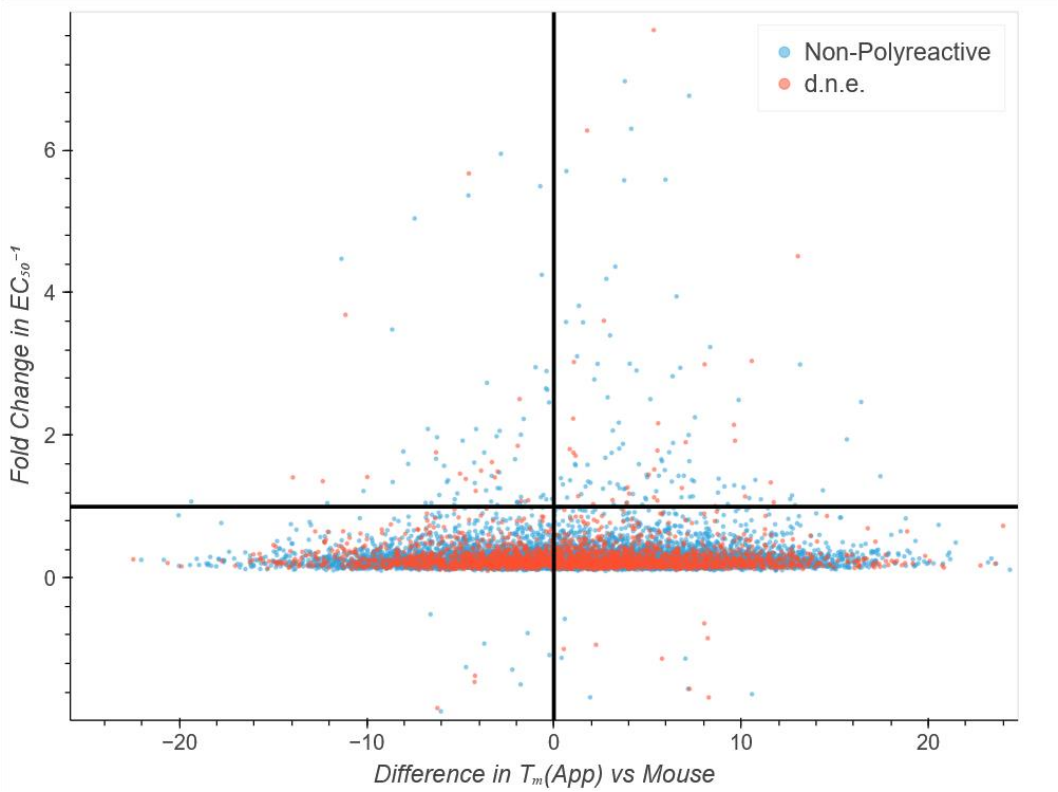


Figure SC.23: Sampled data from the Mousify OH model. Only 1% of antibodies would pass all experiments (At least no change in T_M, not polyreactive, expresses and maximum EC_{50} increase of 20%). Each point would represent an antibody with the properties listed. Samples were considered iid. to generate this figure. It is worth noting that the samples are likely not iid. in reality. For example, a very unstable antibody will likely also have very poor EC_{50}, or a very low EC_{50} antibody also has a higher chance of being polyreactive than an antibody that does not bind to the target antigen at all. D.n.e.: Does not express.

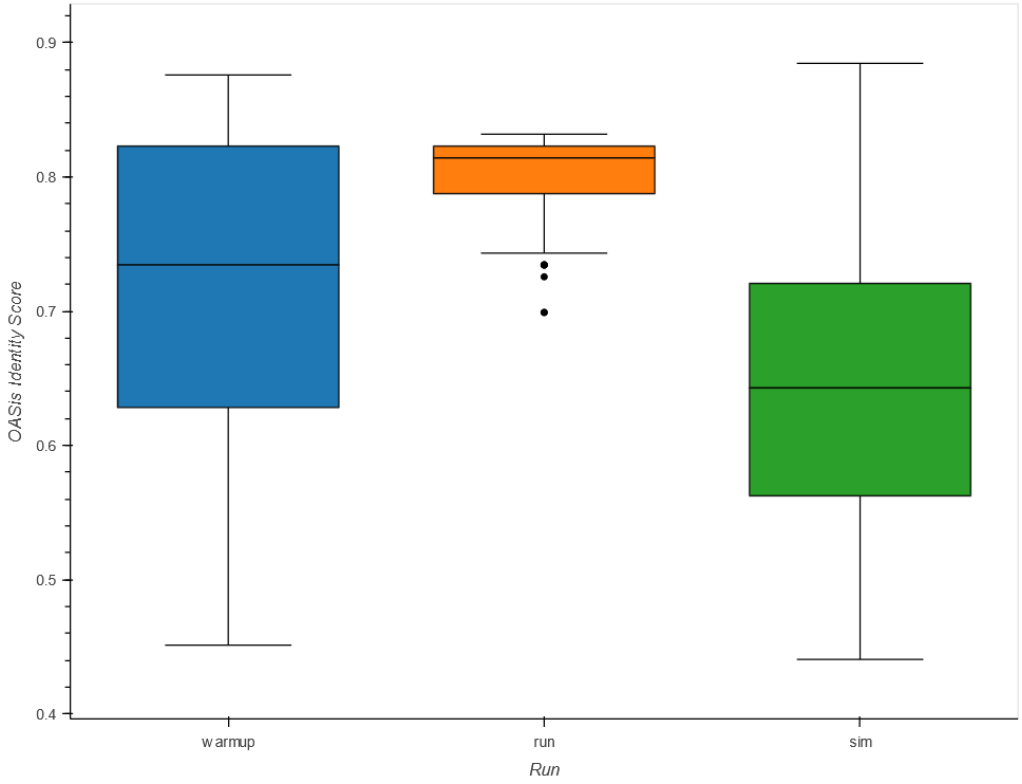


Figure SC.24: Boxplot of OASis Identity Scores of the M8a-3 library run. Mousify Library Parameters: Discriminator: OASis (percentile minimum: 0.15); Map: Sapiens; Mutator: Simulated Annealing (Temperature schedule of 0.005 (warmup), 0.0025 (Run, first 10,000), 0.000001 (Run, up to 200,032). Warmup: First 10,000 sequences; Run: Second phase of MCMC, 200,032 sequences; Sim: OASis Identity score distribution of therapeutically available mAbs.

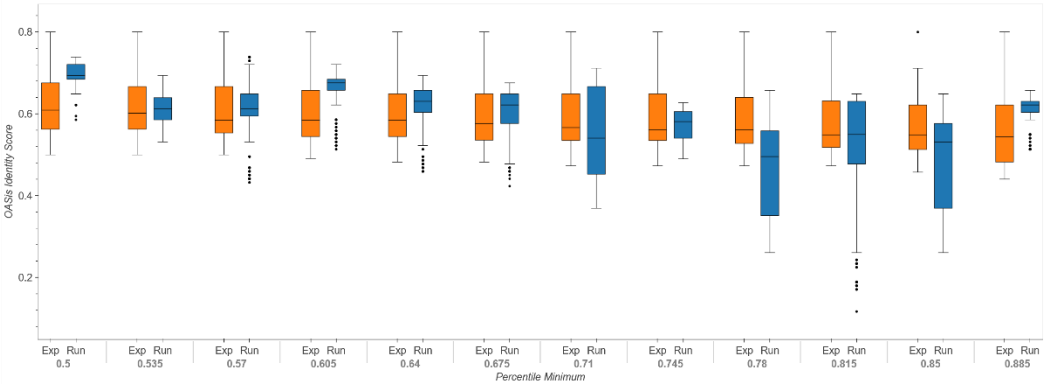


Figure SC.25: Boxplot of OASis scores vs Percentile Minimum values in an SAMC Mousify run. Percentile minimum was cut-off at ≥ 0.5 for clarity. This figure is complementary to Figure 2.11.

AbVAE

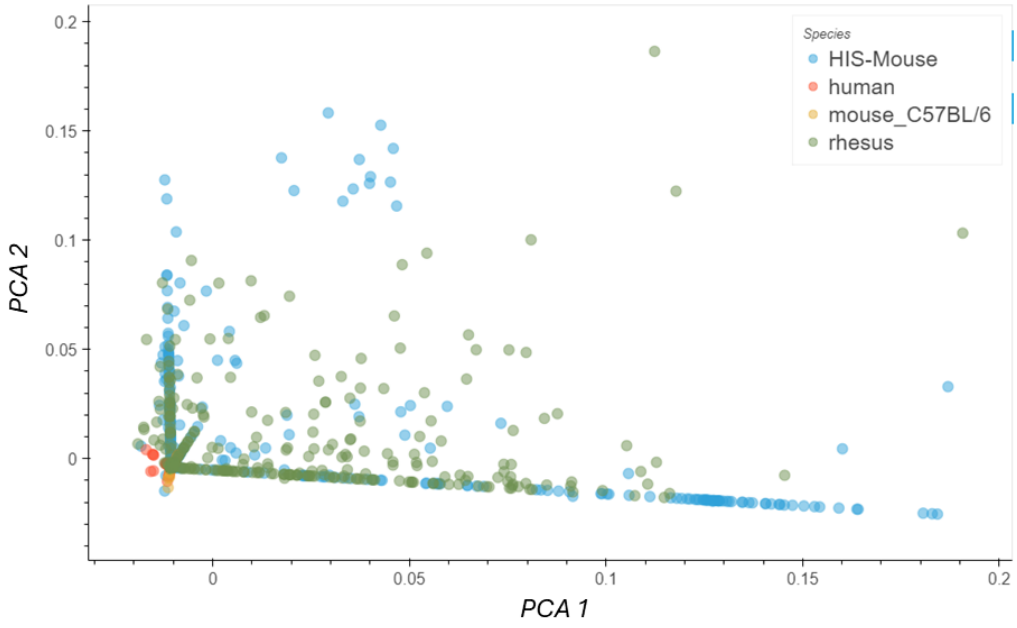


Figure SC.26: AbVAE trained without PID control algorithm to limit the contribution of KL divergence to the total loss. The model cannot learn any structure of the latent space as well as cannot reduce the reconstruction loss due to KL vanishing.

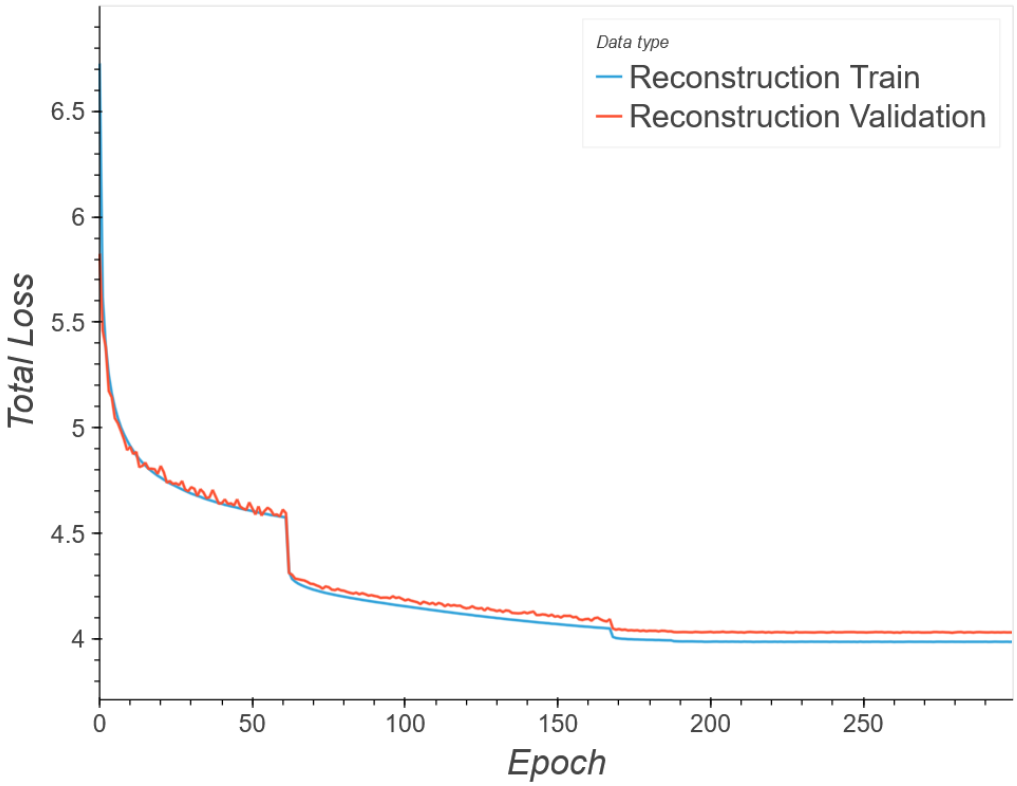


Figure SC.27: Reconstruction loss of AbVAE.

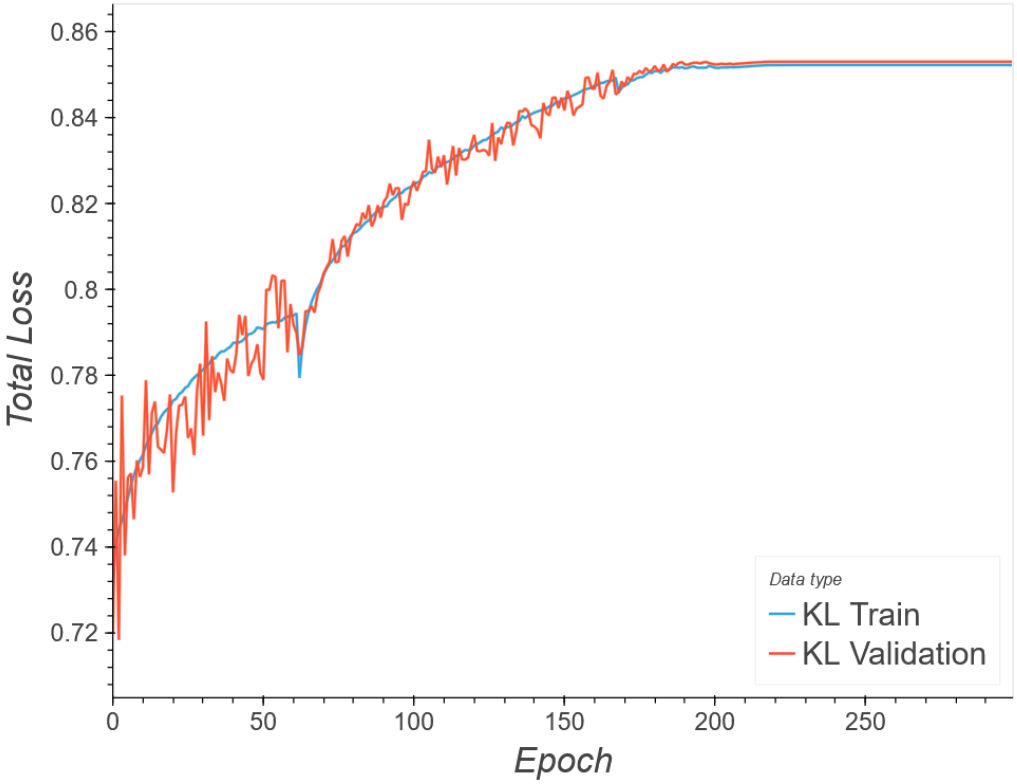


Figure SC.28: KL divergence loss of AbVAE.

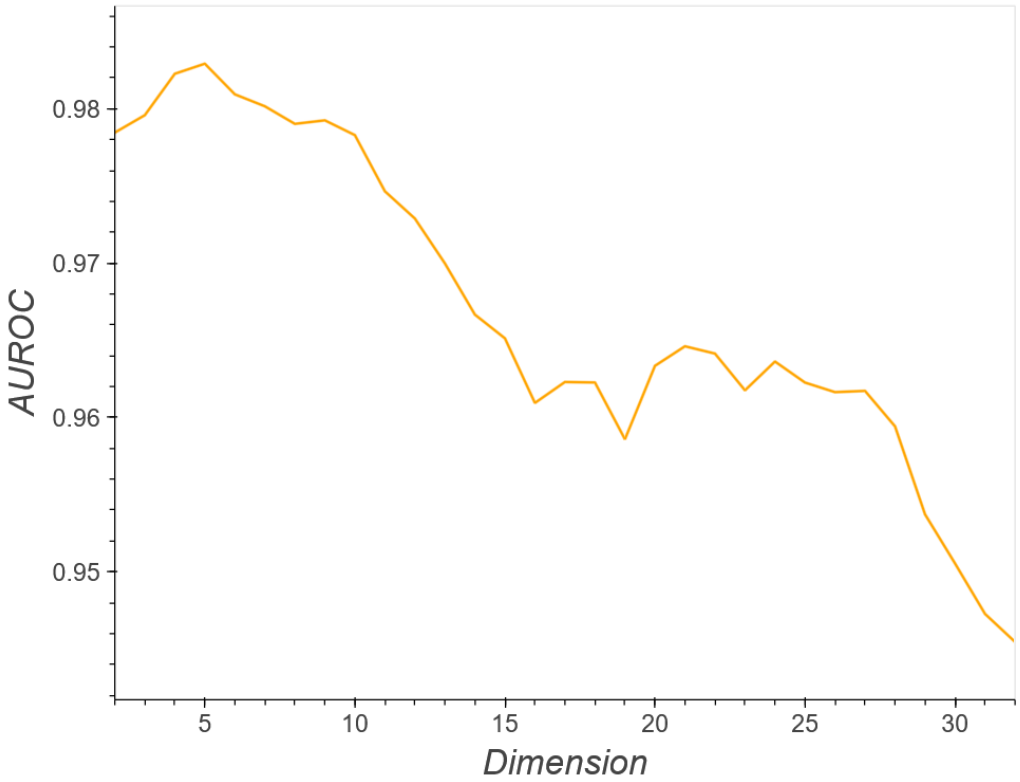


Figure SC.29: AbVAE classification performance of human vs non-human calculated as the AUROC score plotted against the number of PCA dimensions used.

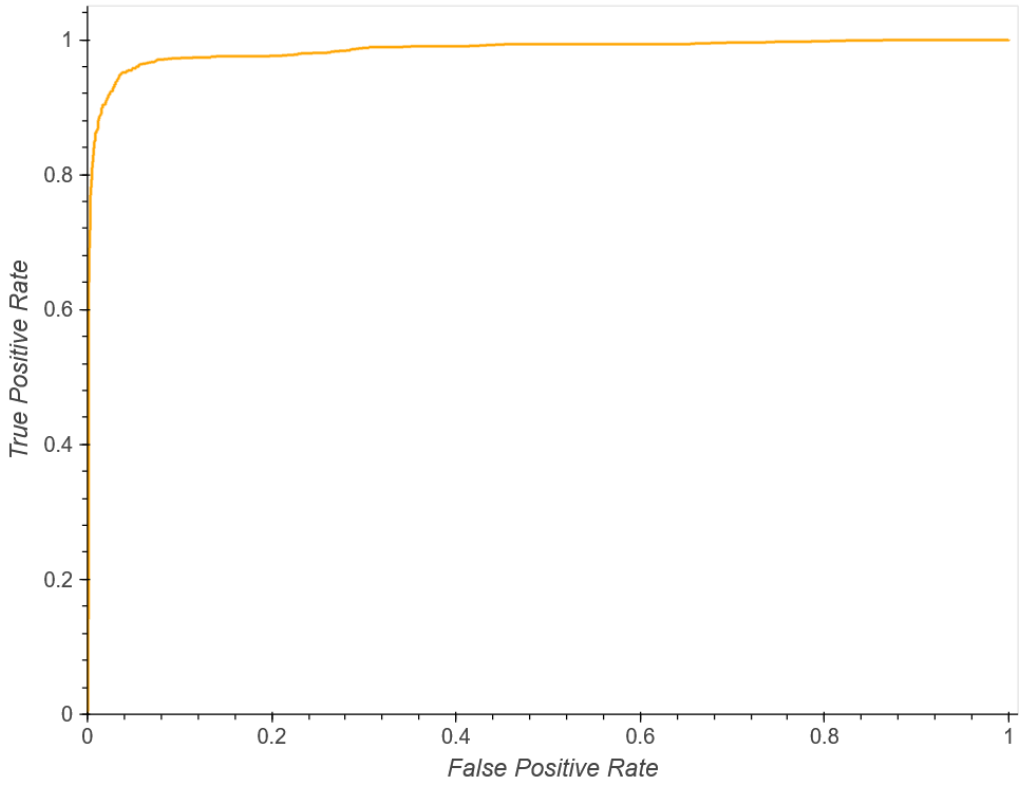


Figure SC.30: AbVAE AUROC plot at 4 PCA dimensions. AUROC score at this dimension is 0.9823.

	AbVAE	Hu-mAb	OASis
AUROC	**0.981**	0.977	0.966

Table SC.1: Comparison of AUROC scores of three different models compared in this thesis. AbVAE at 4 latent dimensions outperforms the other models in terms of classification task.

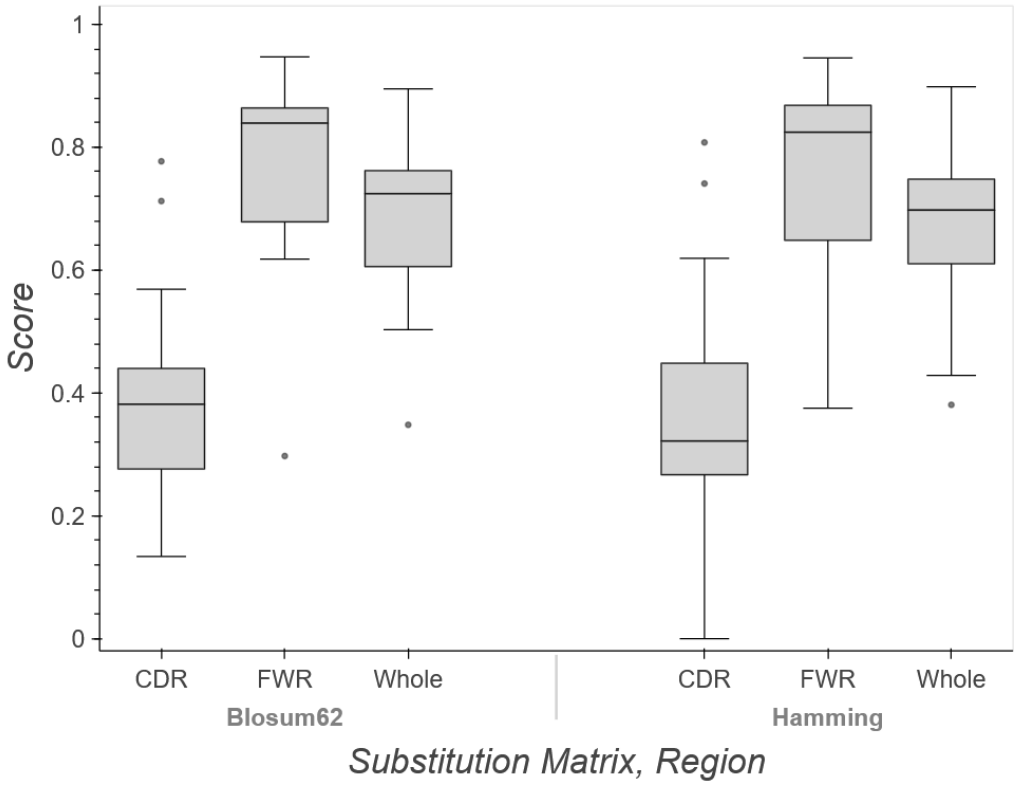


Figure SC.31: Reconstruction performance of AbVAE-ESM. This model performs worse than AbVAE and AbVAE-ByteNet in terms of reconstruction performance.

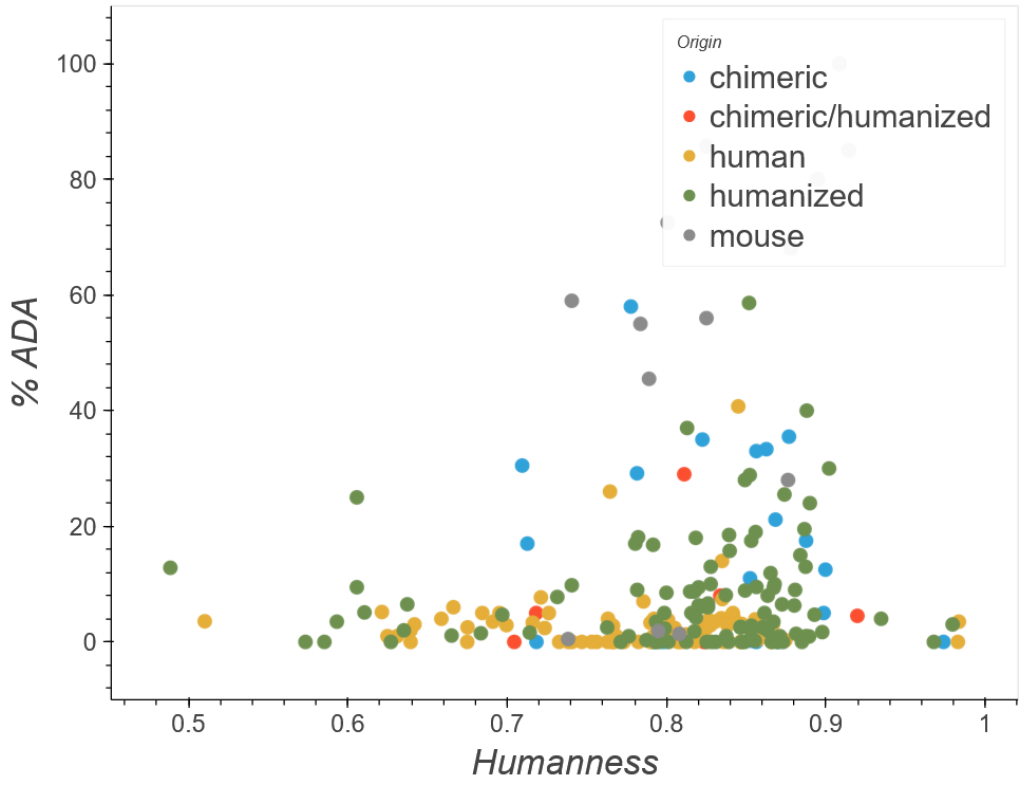


Figure SC.32: AbVAE-ESM humanness score correlation to %ADA. Pearsons correlation coefficient was 0.16 for this model which is of the inverse sign of what one would expect the correlation to be.

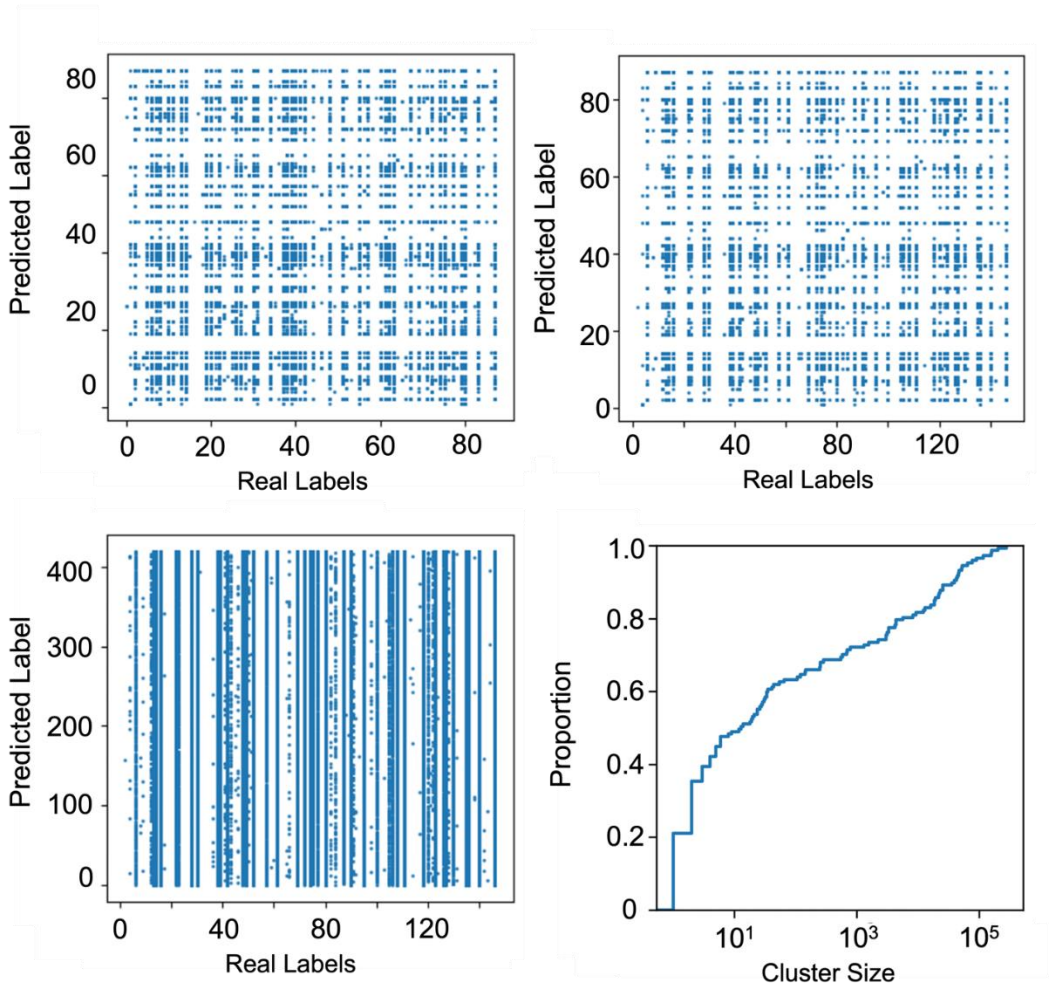


Figure SC.33: Thermosurf classifier performance. The simple classifiers used in this ensemble model (Support Vector Classifier, XGBoost, Random Forest Classifier) cannot find a good solution for a complex classification task. This is most likely due to the distribution of sequence cluster sizes as show in the lower right panel. 20% of clusters only have a single sequence as a member and nearly 50% of the clusters have 10 or less members.

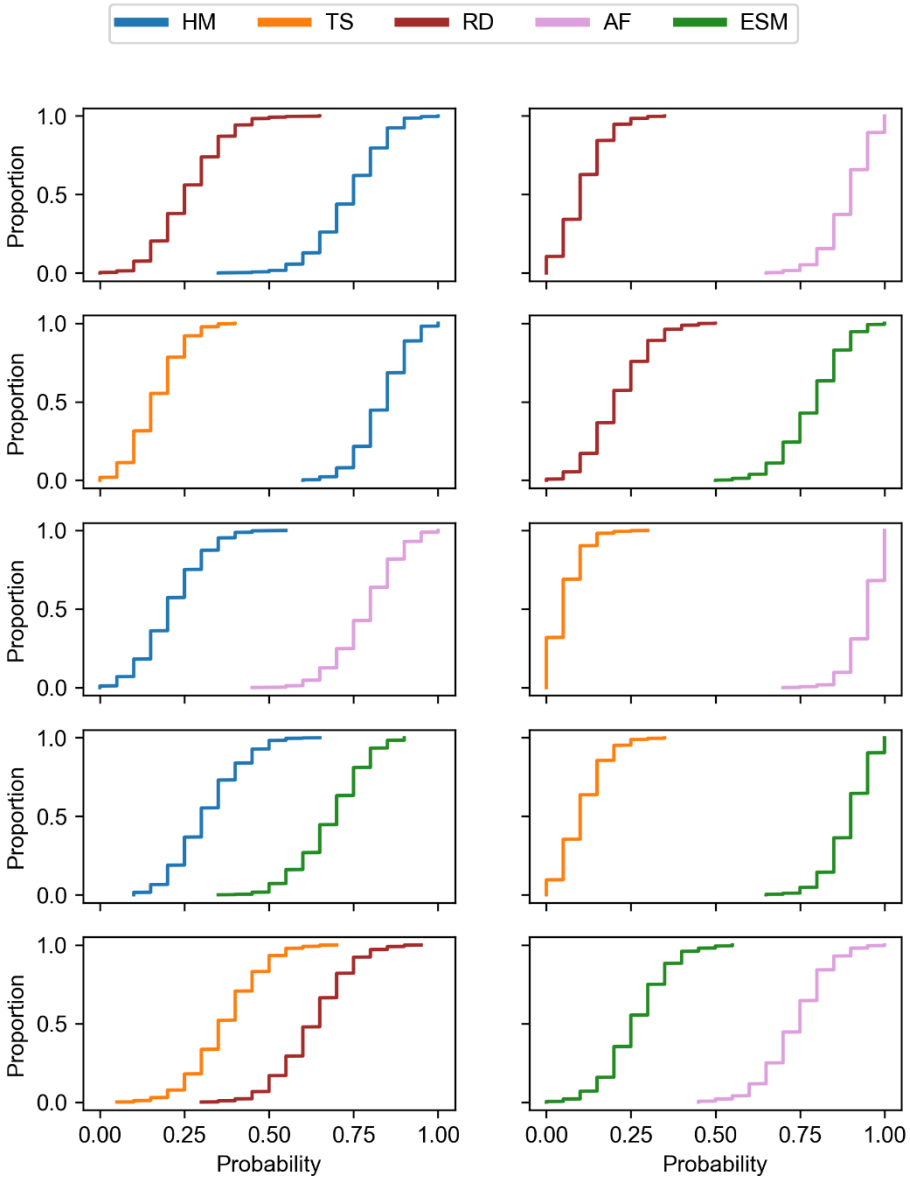


Figure SC.34: Empirical Cumulative Distribution Function (ECDF) plots of the probability that one structure prediction model performs better than the other. This was performed via the one-to-one comparison of each model's predictions on a given query sequence, we then counted the number of occurrences of one performing better than the other. We performed statistical bootstrapping to generate an empirical distribution of each model's performance. These results were used to make Figure 2.21C.

Appendix D: Chapter 4 Supplementary Figures

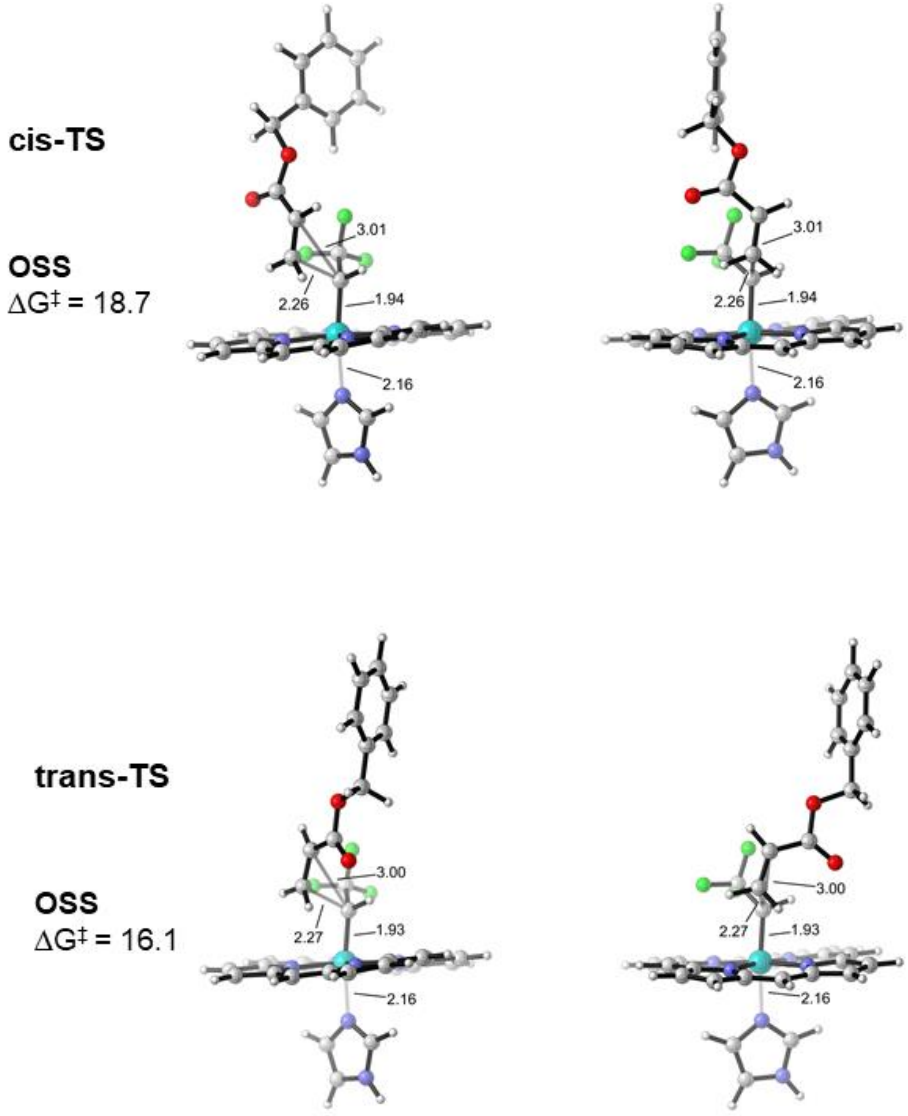


Figure SD.1: DFT-optimized lowest in energy transition state (TS1) structures for the cyclopropanation reaction involving benzyl acrylate for the two possible diastereomers (*cis* and *trans*). A computational truncated model was used. Calculations were performed at the uB3LYP-D3BJ / Def2-TZVP (PCM=DiethylEther) // uB3LYP / 6-31G(d)+SDD(Fe) (PCM=DiethylEther) level of theory. Distances are shown in Å, and energies are reported in kcal·mol^{-1}.

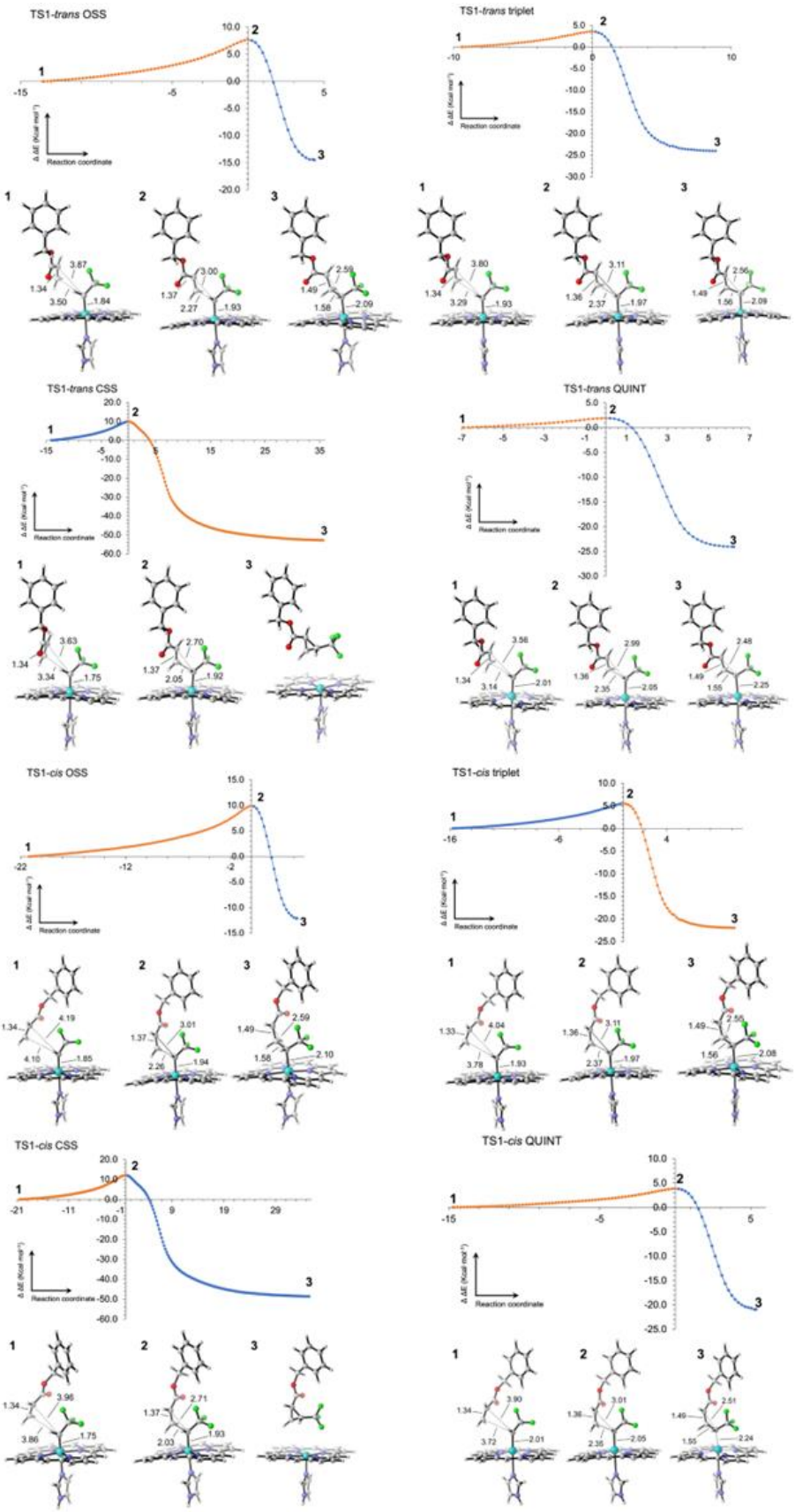

TS1-trans OSS
TS1-trans triplet
TS1-trans CSS
TS1-trans QUINT
TS1-cis OSS
TS1-cis triplet
TS1-cis CSS
TS1-cis QUINT
Reaction coordinate

Figure SD.2: Intrinsic reaction coordinate (IRC) calculations for the DFT-optimized transition state TS1 (*trans* and *cis*) structures for the cyclopropanation reaction involving benzyl acrylate for the two possible diastereomers (*cis* and *trans*) at the uB3LYP / 6-31G(d)+SDD(Fe) (PCM=DiethylEther) level of theory. Four different electronic states (open-shell singlet, OSS; closed-shell singlet , CSS; quintet, QUINT; and triplet) have been considered for TS1. Distances are shown in Å, and electronic energies are reported in kcal·mol^{-1}. IRC calculations describe a stepwise mechanism for the open-shell singlet (OSS), quintet and triplet electronic states, where a covalent intermediate is formed from the corresponding TS1-*trans/cis*. On the other hand, IRC calculations describe a concerted mechanism for the closed-shell singlet (CSS) electronic state for both *trans*- and *cis*- CF_3-cyclopropanations.

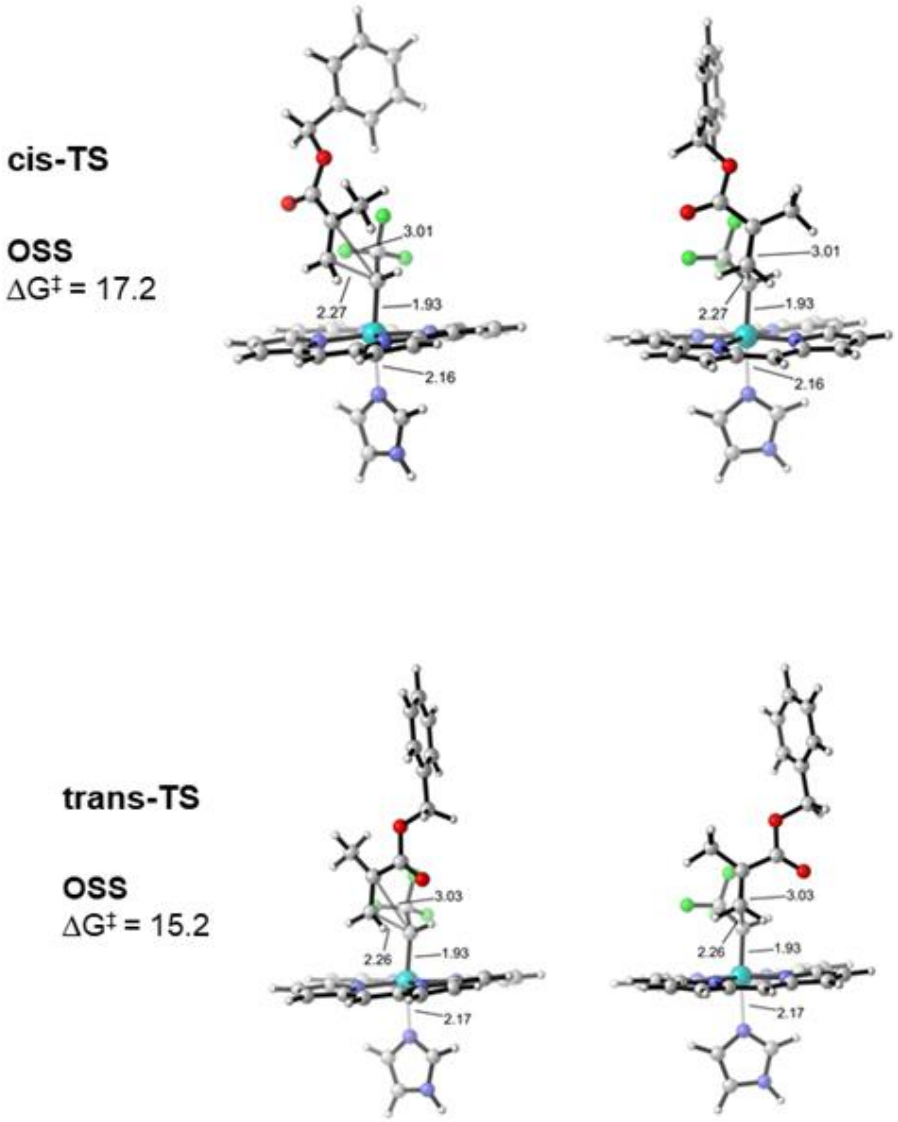


Figure SD.3: DFT-optimized lowest in energy transition state (TS1) structures for the cyclopropanation reaction involving benzyl acrylate for the two possible diastereomers (*cis* and *trans*). A computational truncated model has been used. Calculations were performed at the

uB3LYP-D3BJ / Def2-TZVP (PCM=DiethylEther) // uB3LYP / 6-31G(d)+SDD(Fe) (PCM=DiethylEther) level of theory. Distances are shown in Å, and energies are reported in kcal·mol^{-1}.

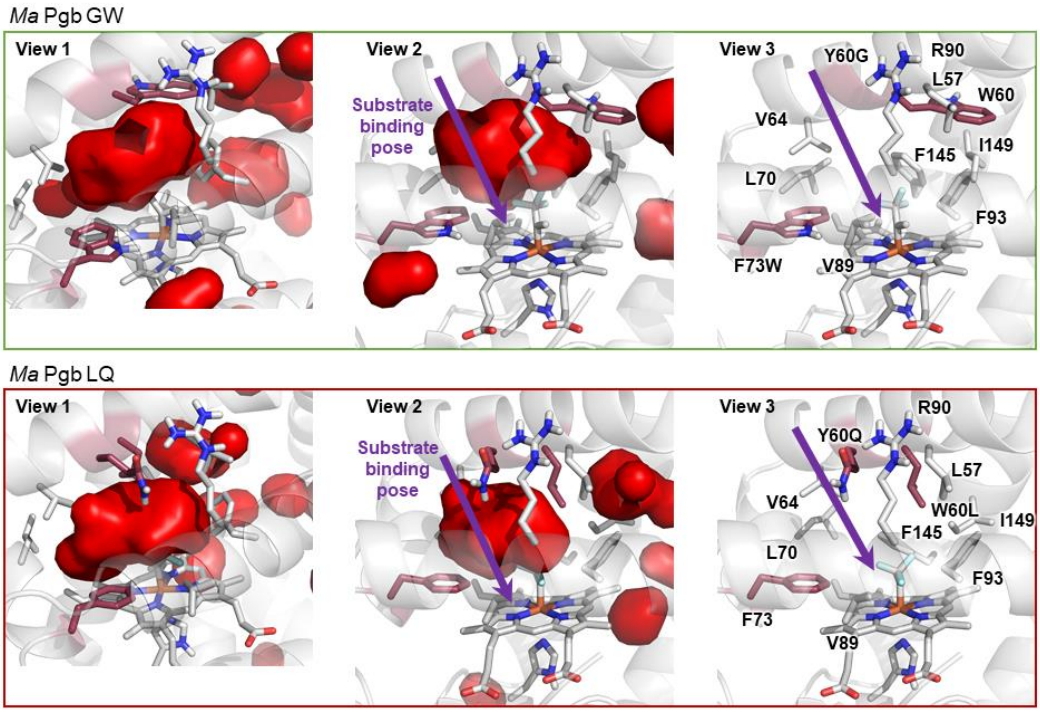


Figure SD.4: Representative snapshots obtained from MD simulations (replica 1, at t = 60 ns) showing the accessible active site volume in two protoglobins variants with an iron-carbenoid bound (View 1 and 2, red blobs), where substrate is expected to bind and approach the carbenoid species in a catalytically competent pose (View 2 and 3, purple arrow). View 2 and 3 are equivalent and correspond to a 90° rotation from View 1.

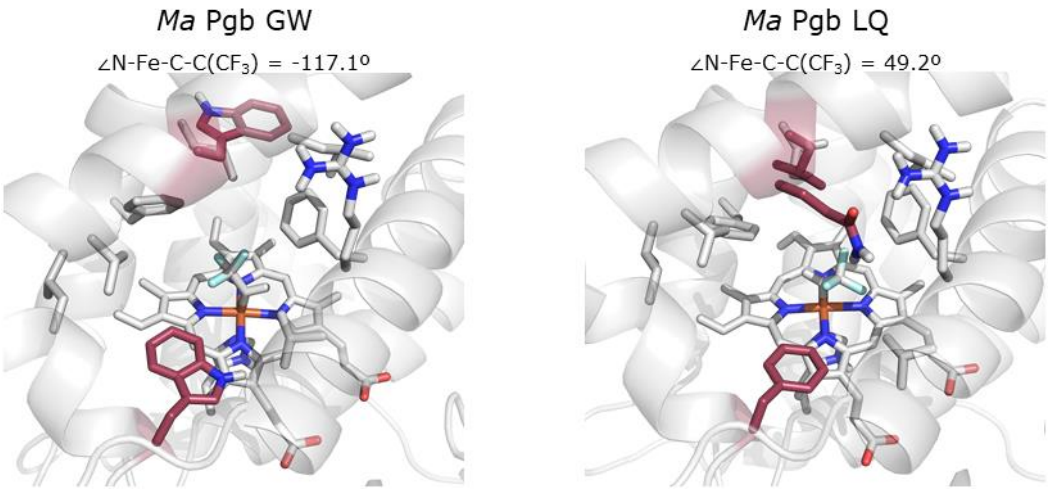
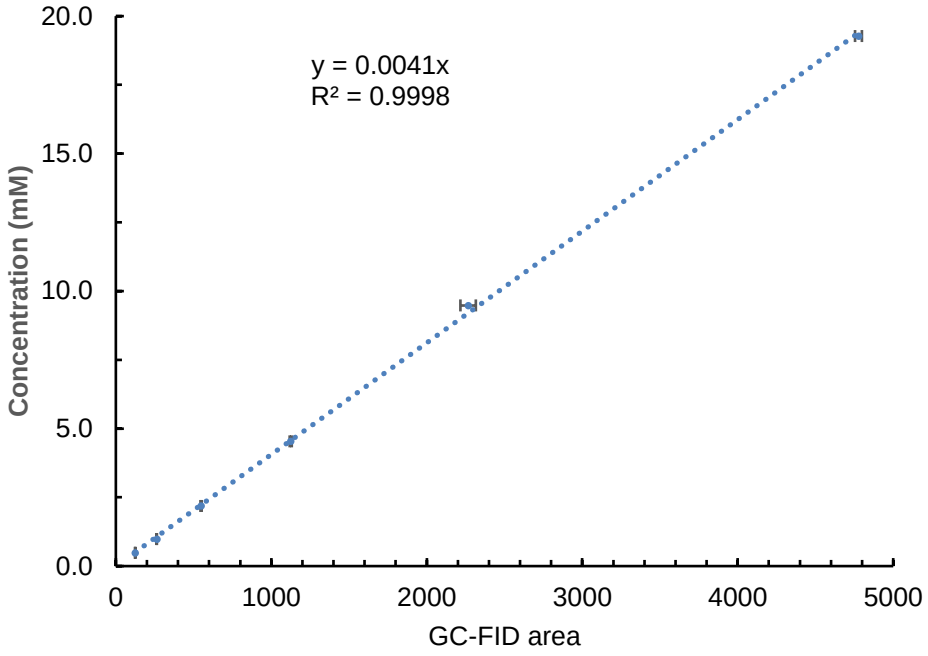


Figure SD.5: Snapshots describing the most visited and representative conformers of the iron-carbenoid as observed from MD simulations in of *Ma*Pgb GW (left) and in *Ma*Pgb LW (right). The mutated amino acids from wild type are shown in maroon color.

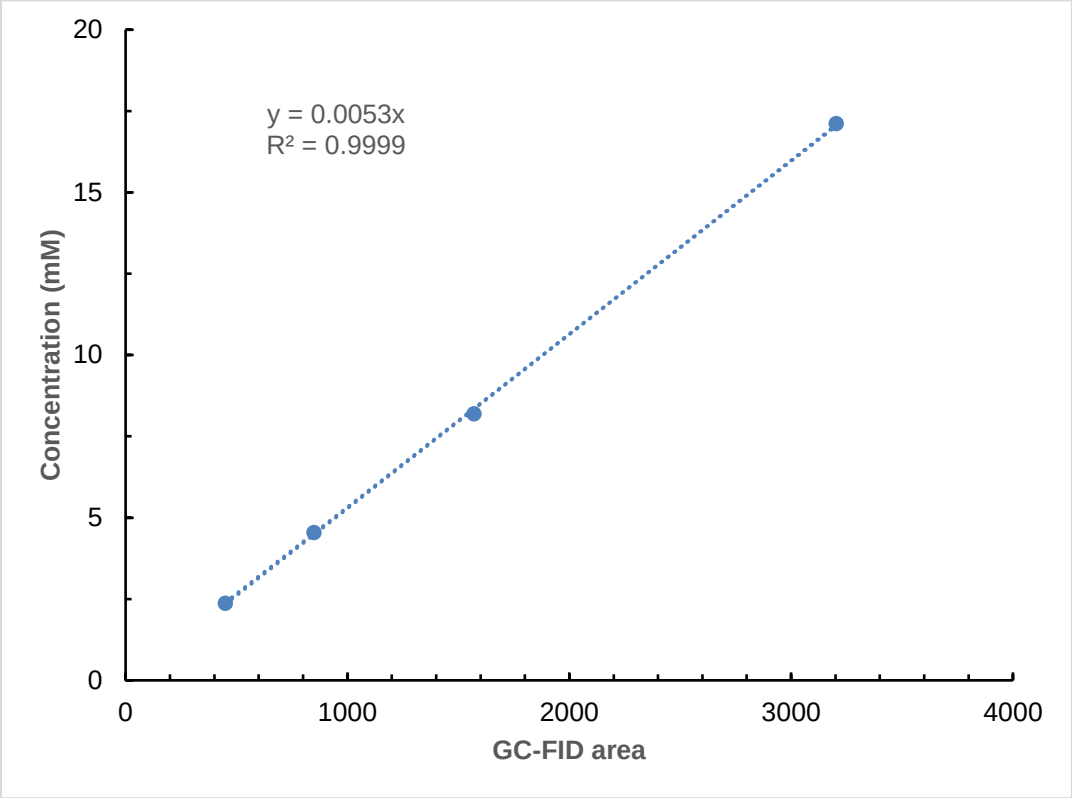
Calibration Curves

To determine the standard calibration curves, stock solutions of chemically synthesized organoborane products were prepared at various concentrations (0.4–20 mM in 4:6 hexanes/EtOAc). All data points represent the average of triplicate runs. The standard curves plot product concentration in mM (y-axis) against the ratio of product area to internal standard area on GC-FID (x-axis).

Calibration curve of benzyl 1-methyl-2-(trifluoromethyl)cyclopropane-1-carboxylate


y = 0.0041x
R² = 0.9998
Concentration (mM)
GC-FID area
20.0
15.0
10.0
5.0
0.0
0
1000
2000
3000
4000
5000

Calibration curve of benzyl 1-methyl-2-(trifluoromethyl)cyclopropane-1-carboxylate

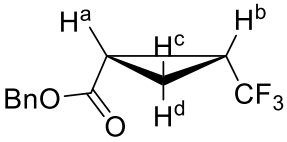


Appendix E: Chapter 4 Compound Characterization Data

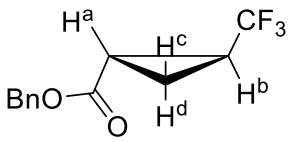
Explanation of NMR-NOESY Interpretation

We added this section to the compound characterization to demonstrate that the *cis* diastereomer is the major product yielded by *Ape*Pgb LQ. First, we use the example of *trans*-benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate and *cis*-benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate to demonstrate how the ^{1}H-NMR shifts differ in both diastereomers. Second, we show the difference in NOESY spectra for the two diastereomers to demonstrate that the *cis* and *trans* labels have been correctly assigned. Lastly, we show the NOESY spectra for all other reported *cis* diastereomers. NOESY spectra were measured on a Bruker Prodigy 400 MHz instrument at 400 MHz and a mixing time of 1 s.

cis conformation

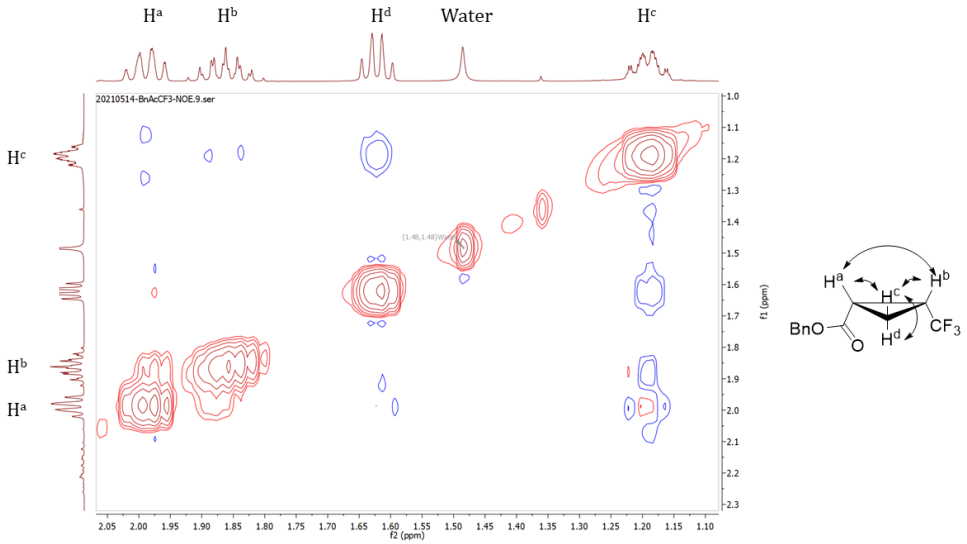


trans conformation

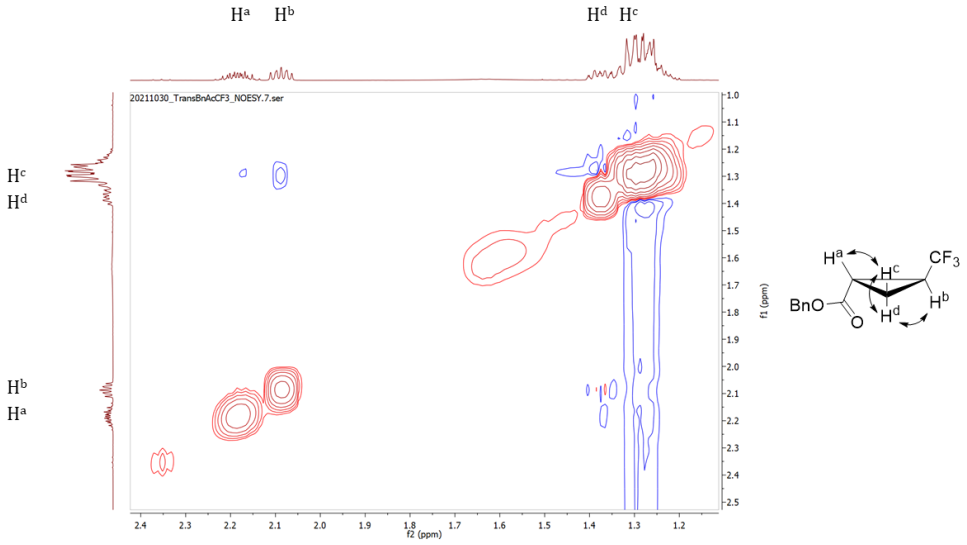


δ (ppm)	H^a	H^b	H^c	H^d
cis	2.21	1.88	1.27	1.62
trans	2.21	1.70	1.23	1.41

NOESY spectrum of *cis*-benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate shows strong NOE signal of H^a with H^b, indicating that both protons are on the same side of the cyclopropane plane. Additionally, a weaker signal is observed of H^c with H^a and H^b.

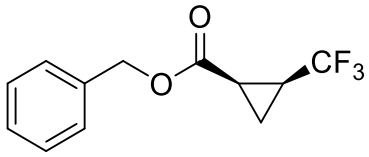


NOESY spectrum of *trans*-benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate shows no NOE signal between H^a and H^b, indicating that these two protons are on opposite sides of the cyclopropane plane. In the *trans* diastereomers, H^d is also shifted upfield compared to the *cis* diastereomer, providing another indicator for the diastereomer.

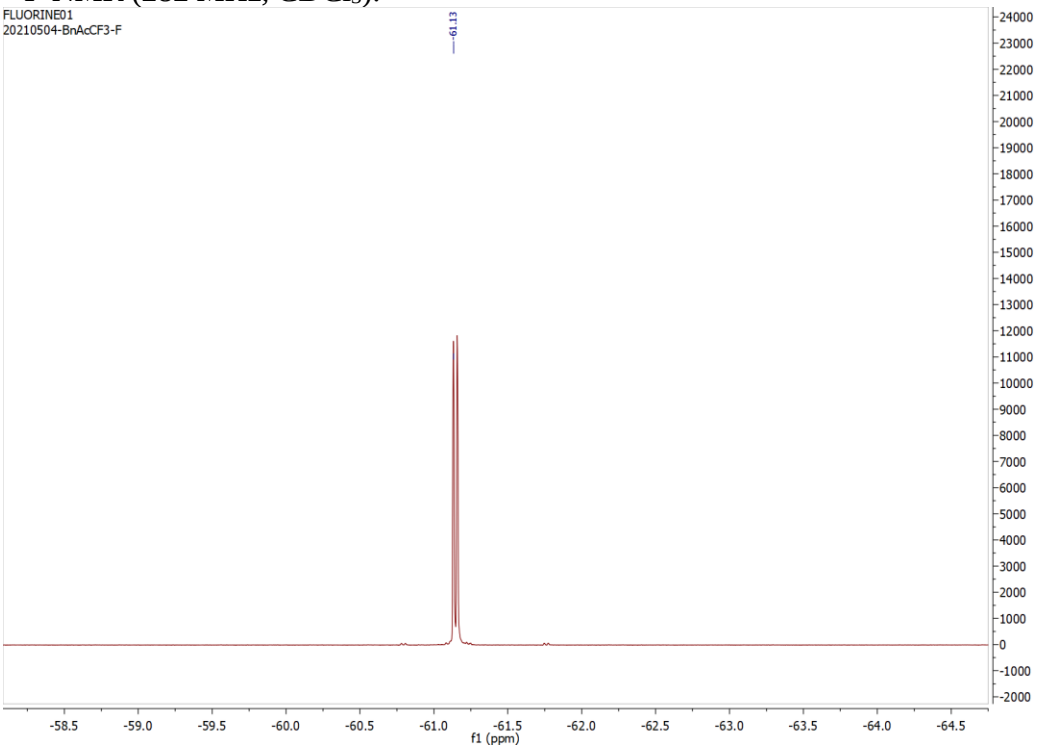


NMR Spectra of new compounds

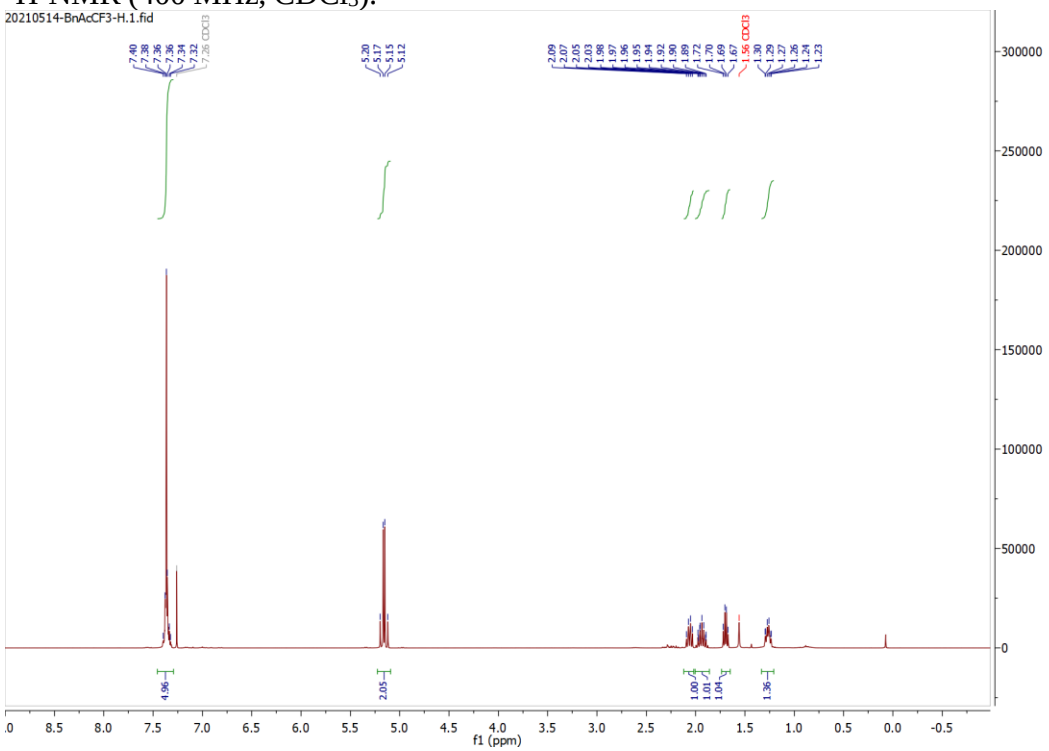
cis-benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate (Cis-1):



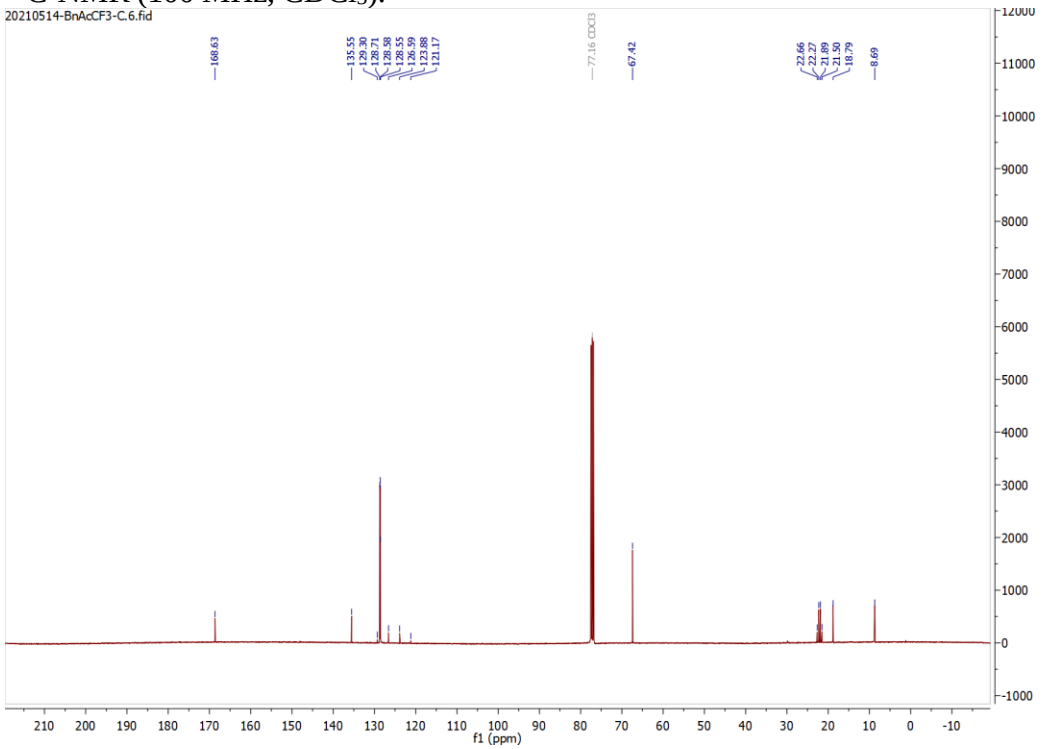
^{19}F-NMR (282 MHz, $CDCl_3$):



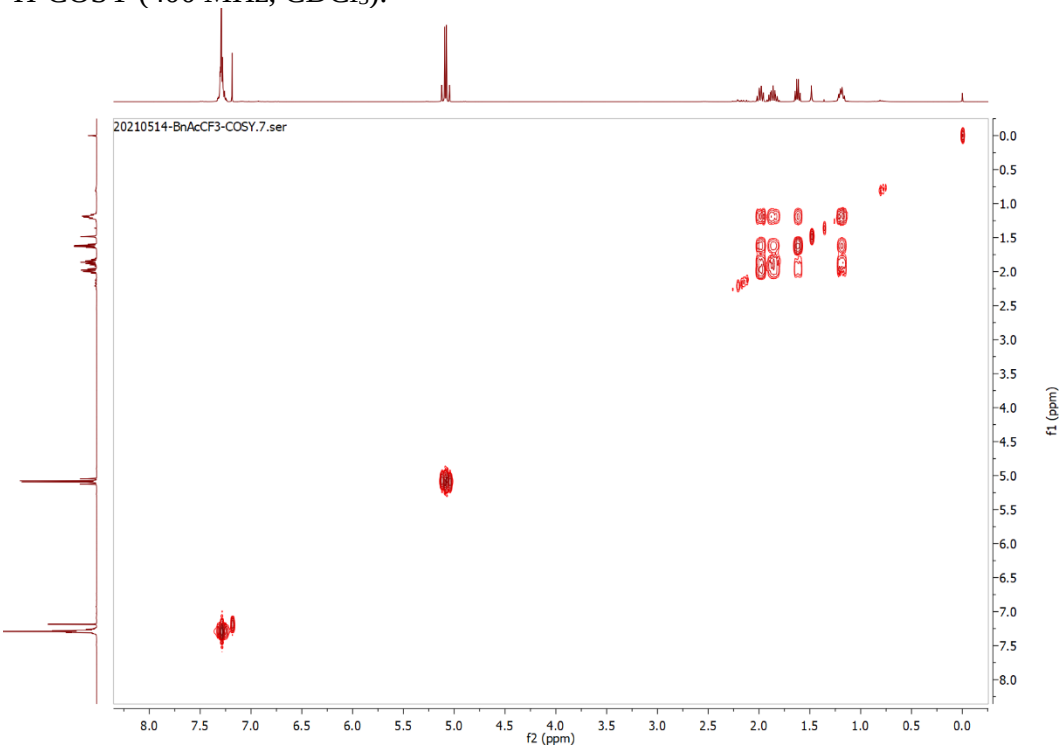
^{1}H-NMR (400 MHz, $CDCl_3$):



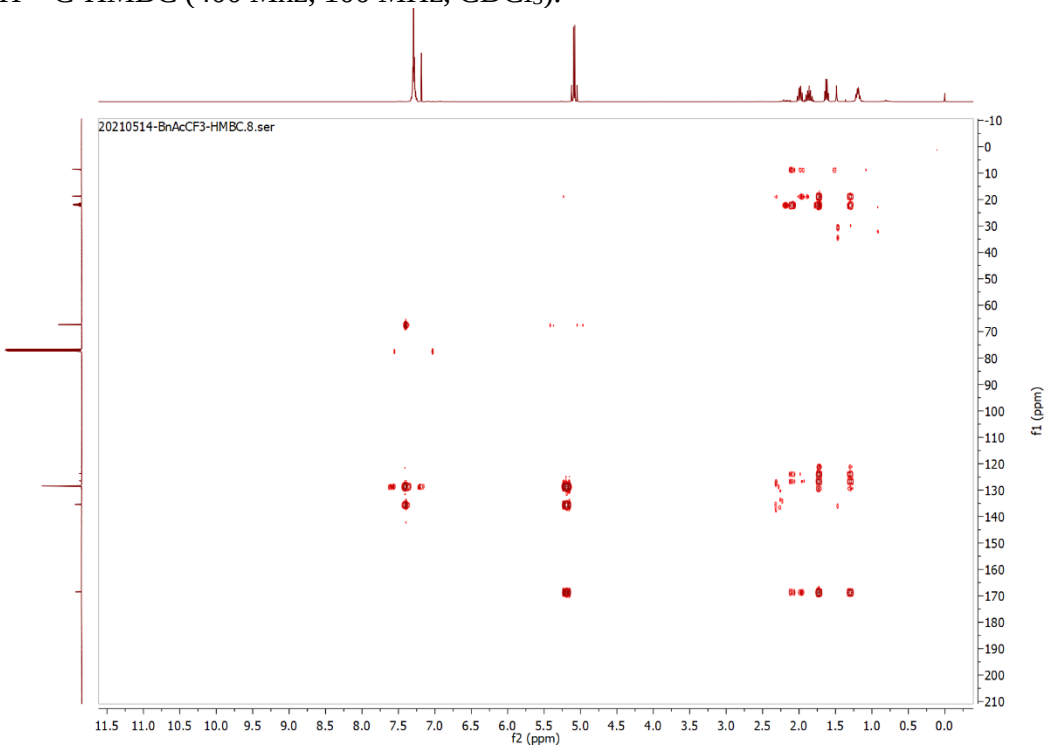
^{13}C-NMR (100 MHz, $CDCl_3$):



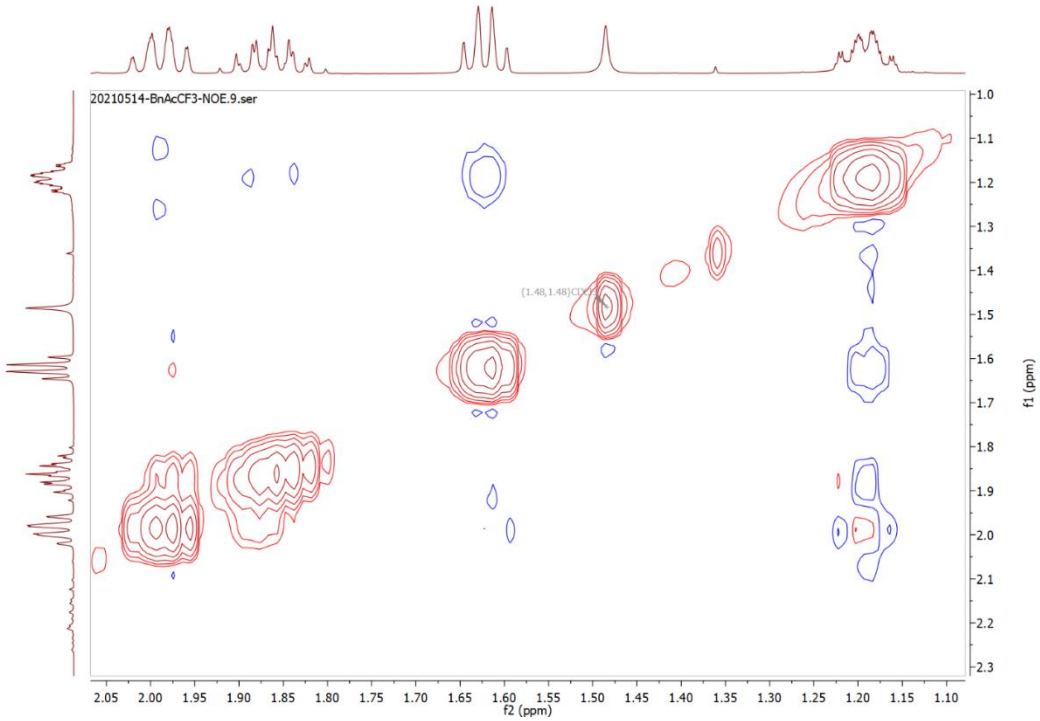
^{1}H-COSY (400 MHz, $CDCl_3$):



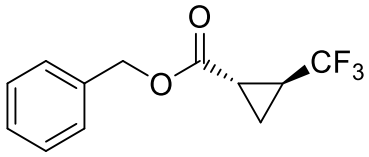
^{1}H-^{13}C-HMBC (400 Mhz, 100 MHz, $CDCl_3$):



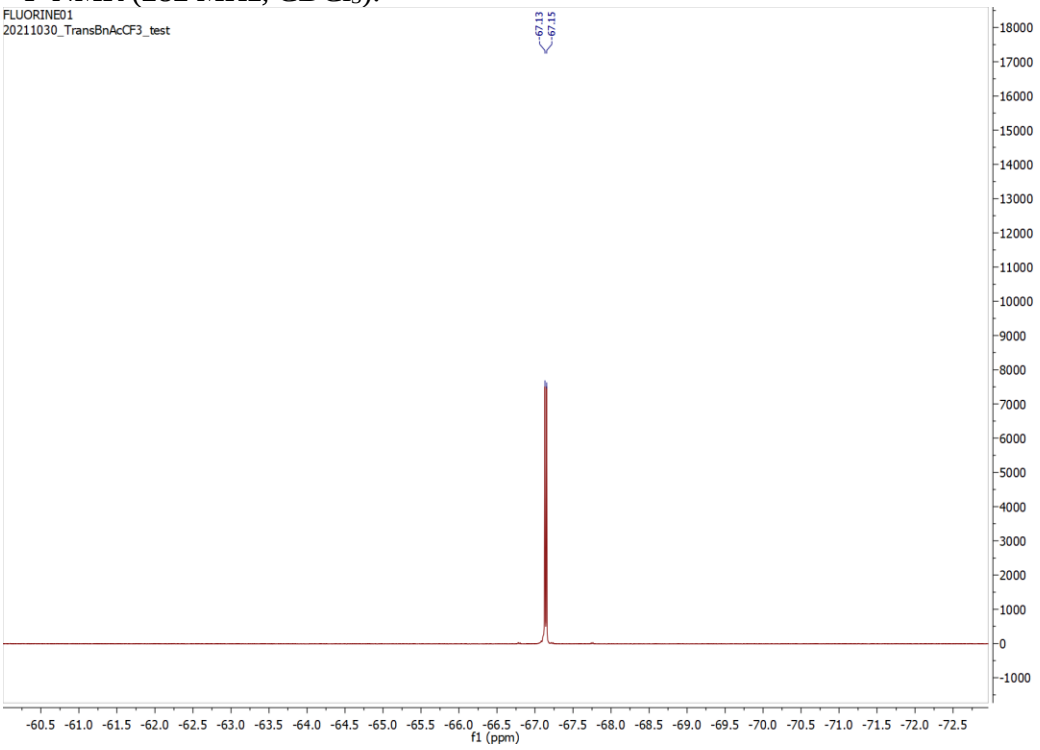
NOESY (400 MHz, $CDCl_3$):



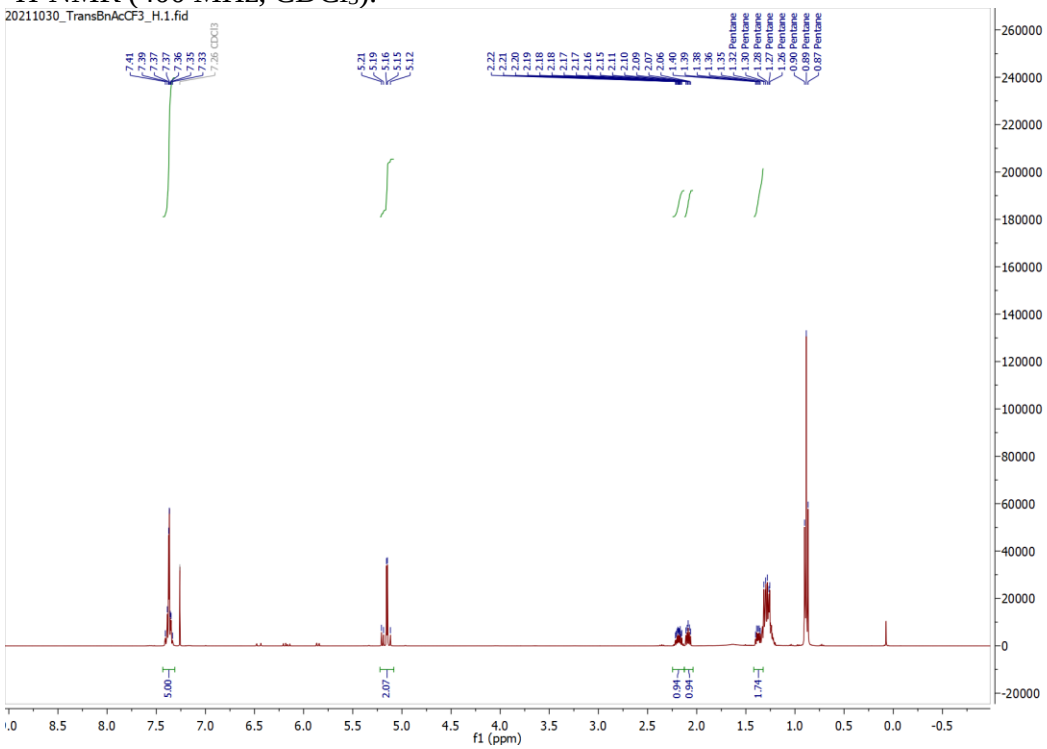
Trans-benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate (Trans-1):



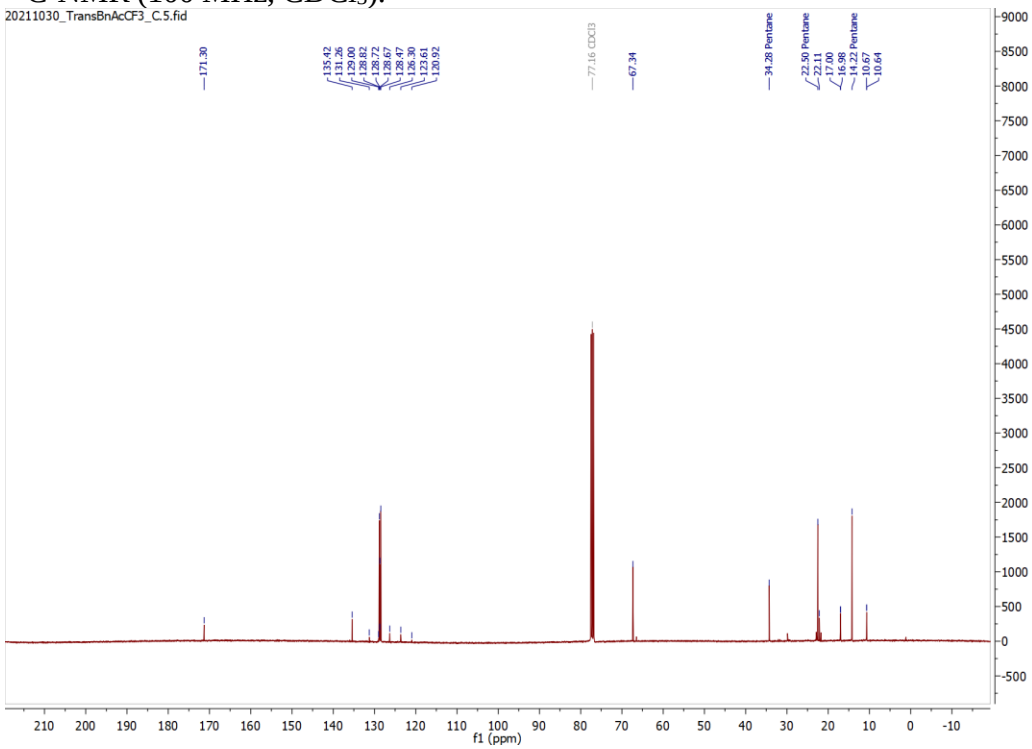
^{19}F-NMR (282 MHz, $CDCl_3$):



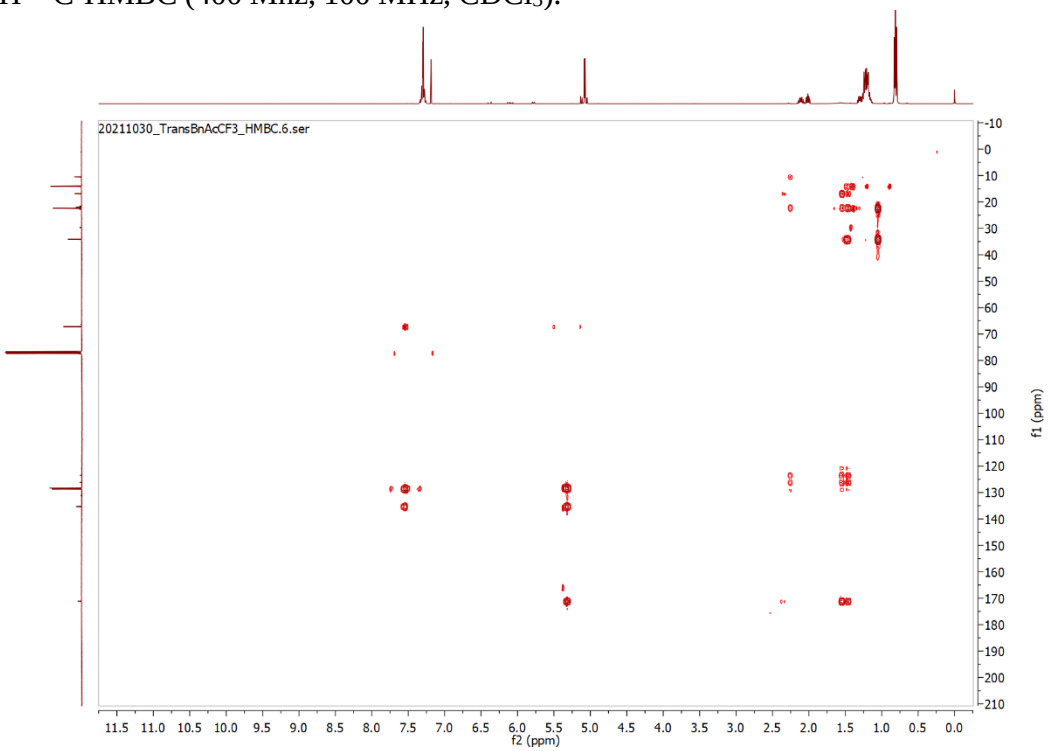
^{1}H-NMR (400 MHz, $CDCl_3$):



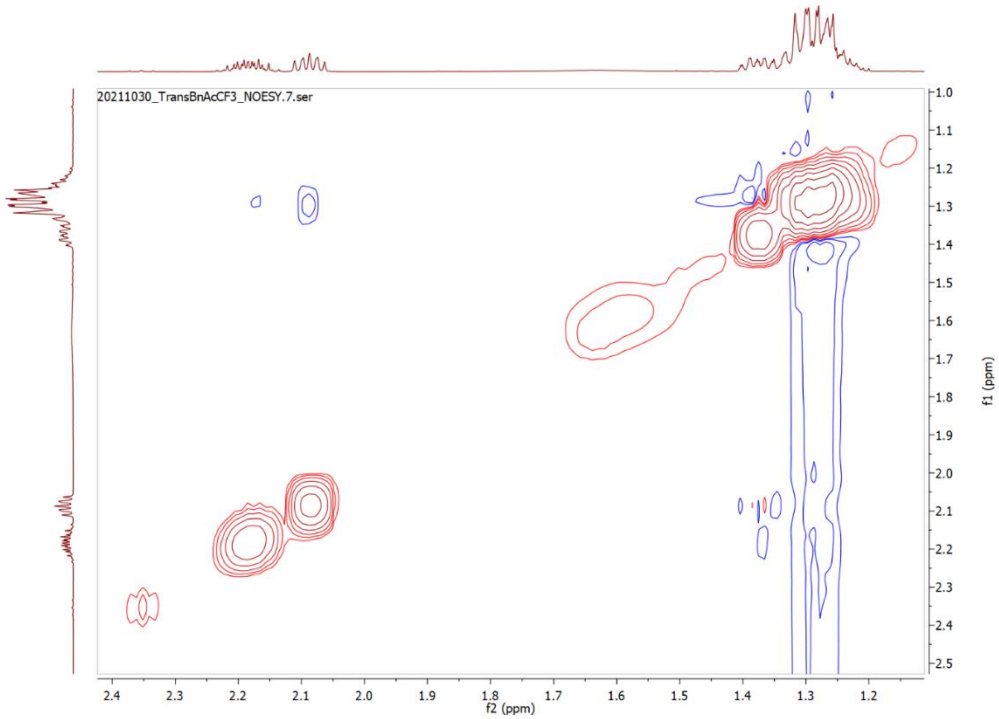
^{13}C-NMR (100 MHz, $CDCl_3$):



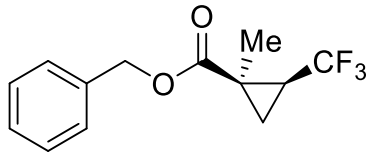
^{1}H-^{13}C-HMBC (400 Mhz, 100 MHz, $CDCl_3$):



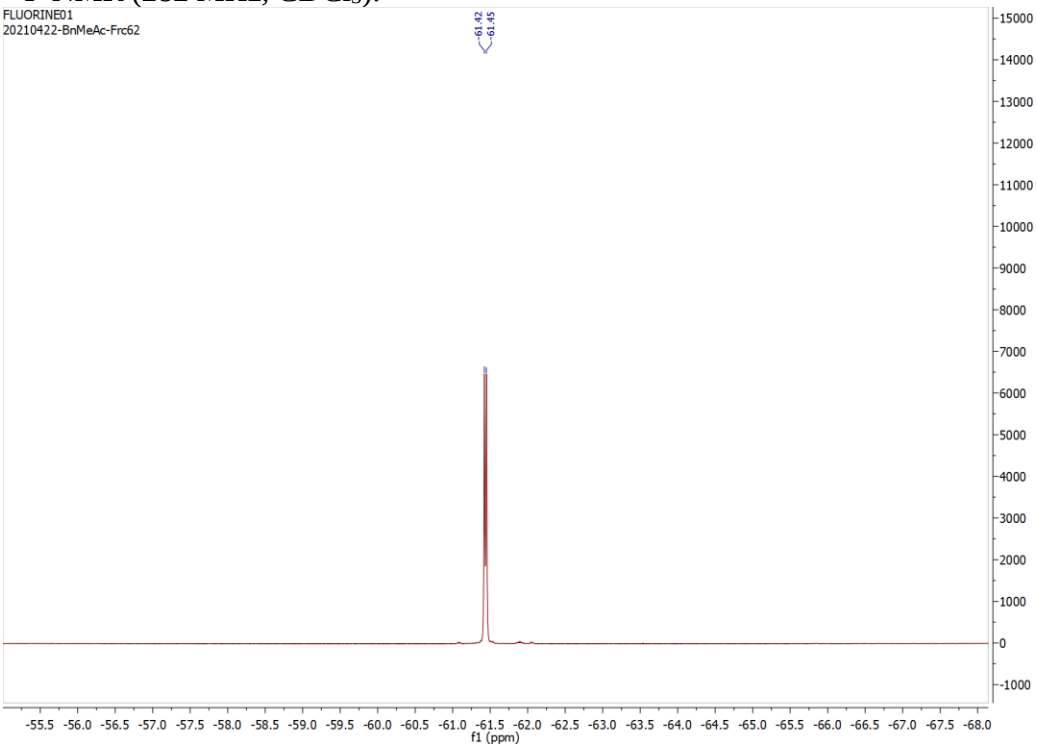
NOESY (400 MHz, $CDCl_3$):



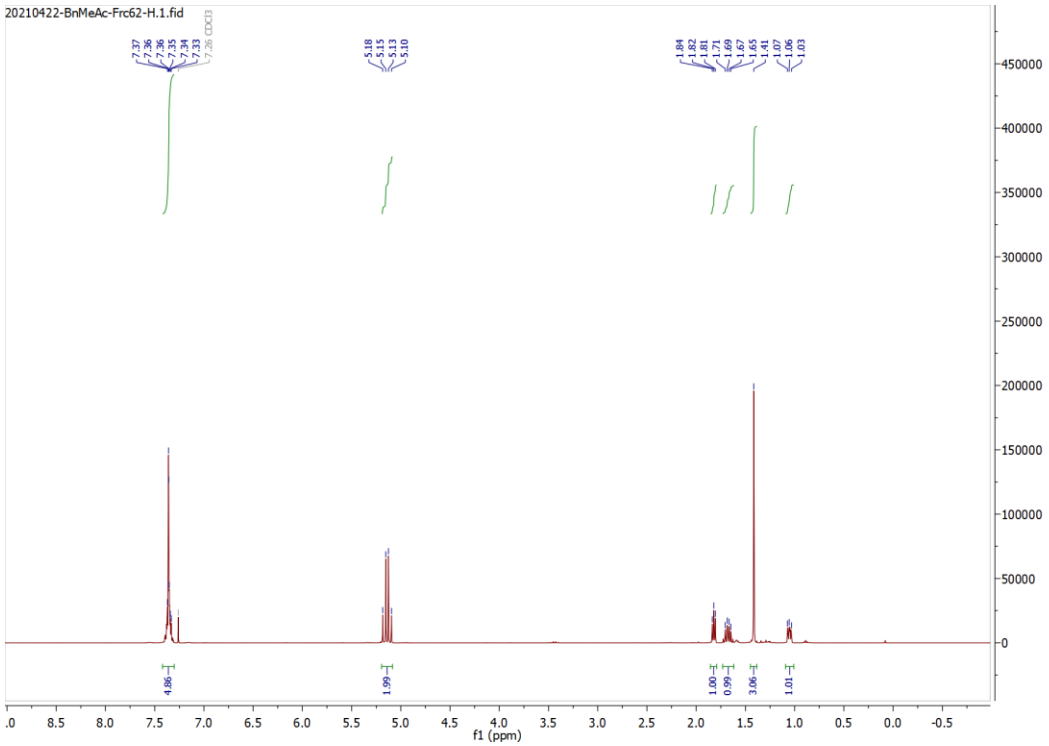
Benzyl 1-methyl-2-(trifluoromethyl)cyclopropane-1-carboxylate (2):



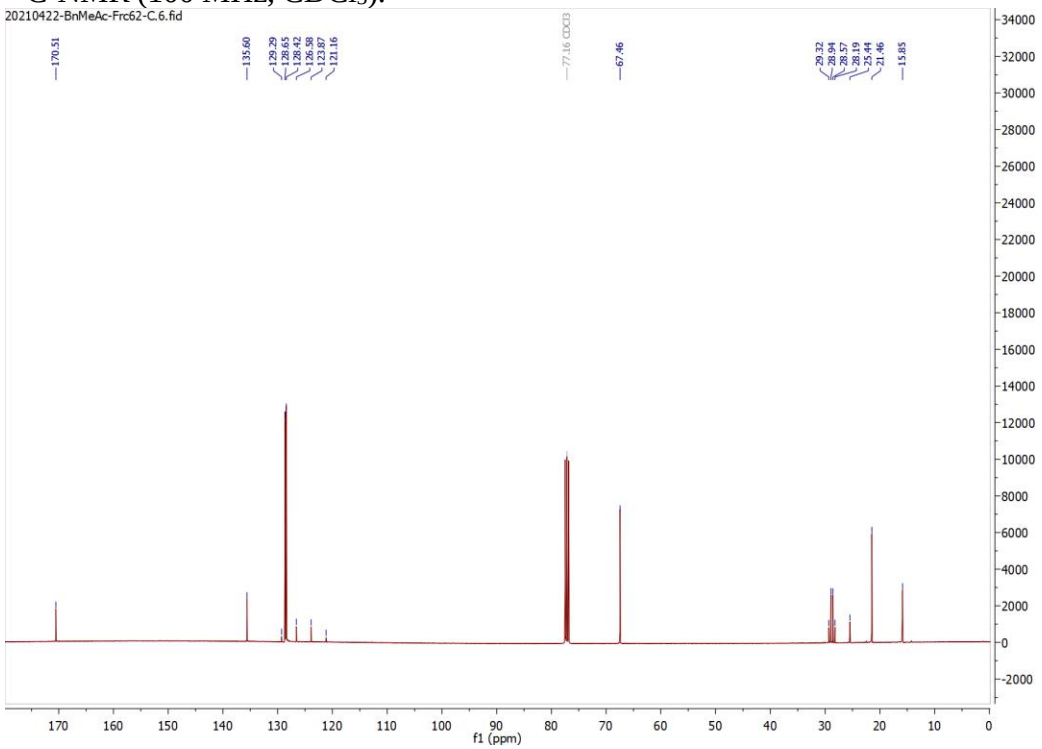
^{19}F-NMR (282 MHz, $CDCl_3$):



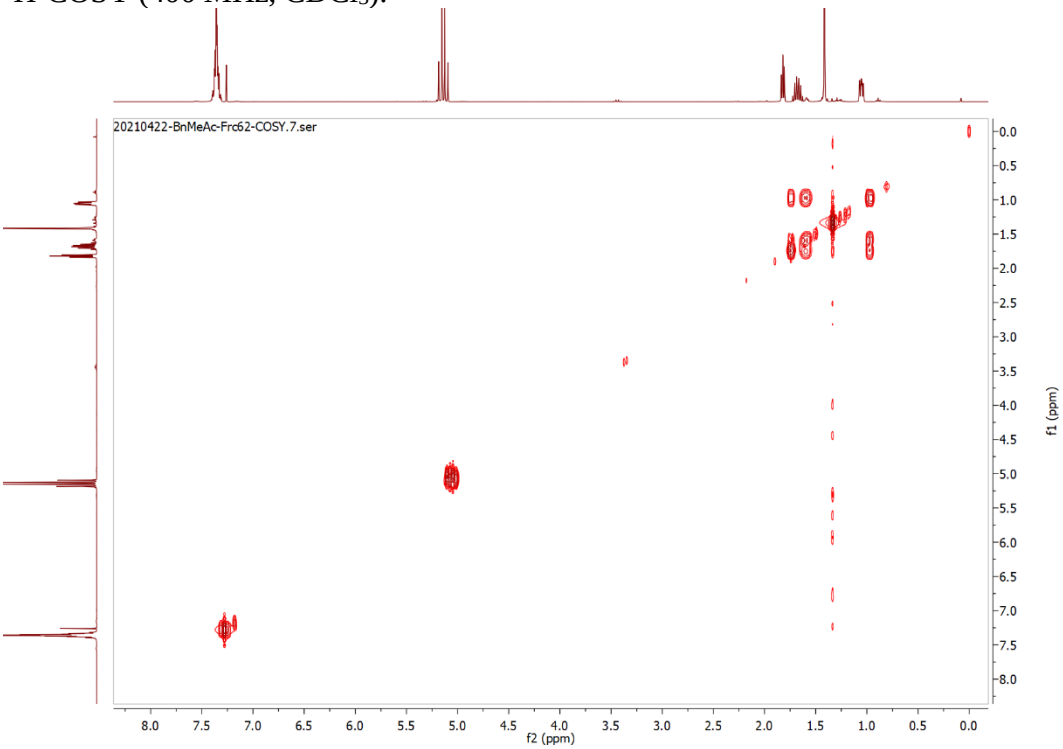
^{1}H-NMR (400 MHz, $CDCl_3$):



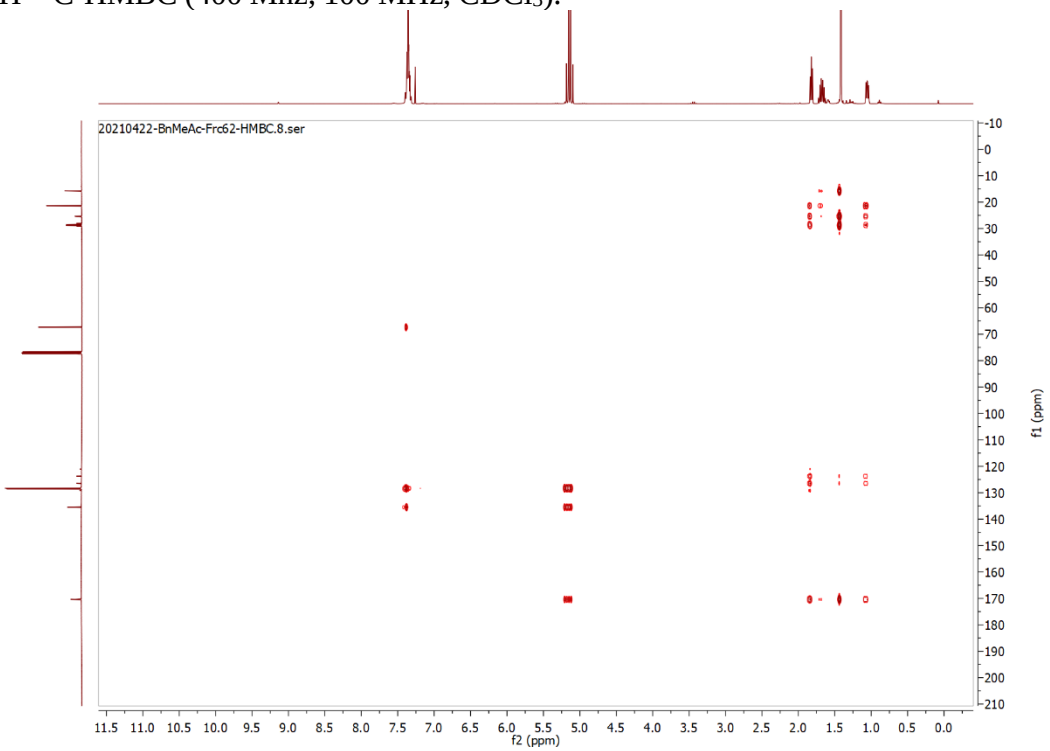
^{13}C-NMR (100 MHz, $CDCl_3$):



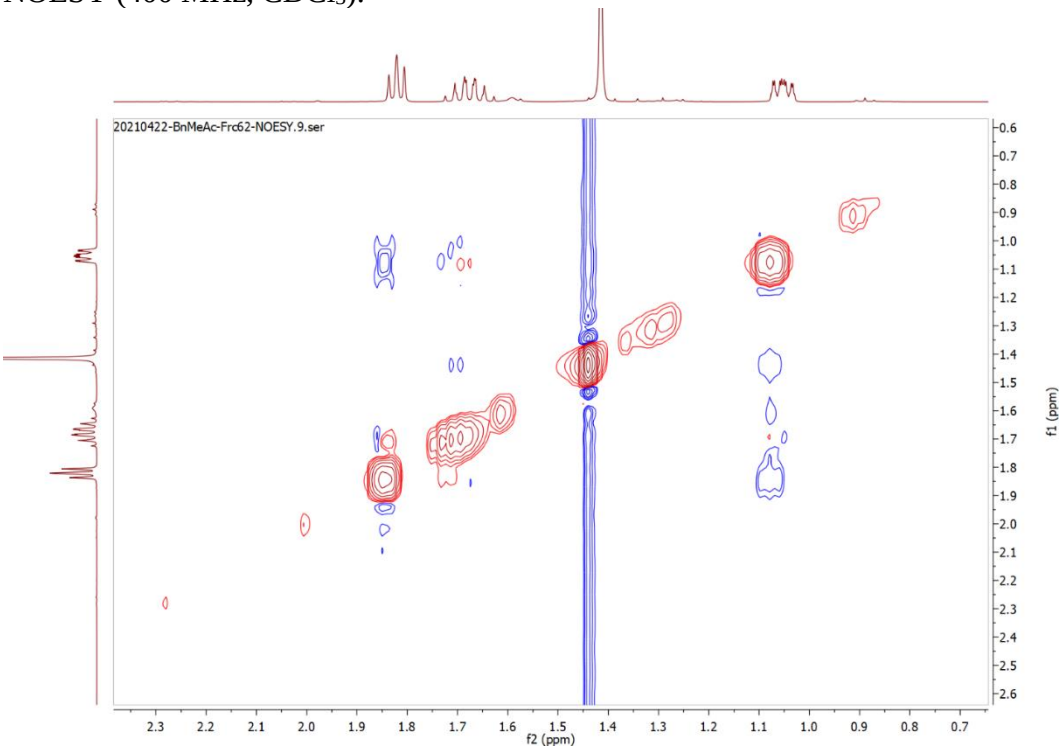
^{1}H-COSY (400 MHz, $CDCl_3$):



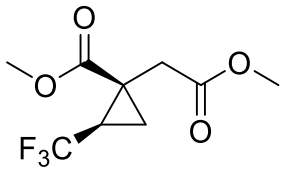
^{1}H-^{13}C-HMBC (400 Mhz, 100 MHz, $CDCl_3$):



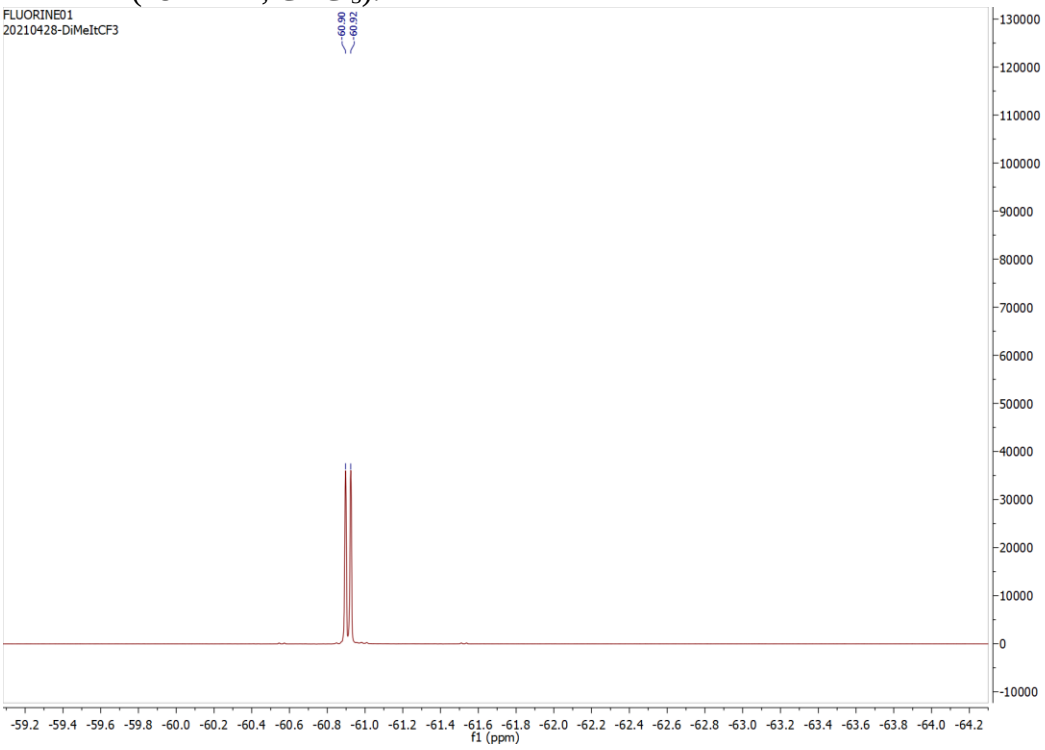
NOESY (400 MHz, $CDCl_3$):



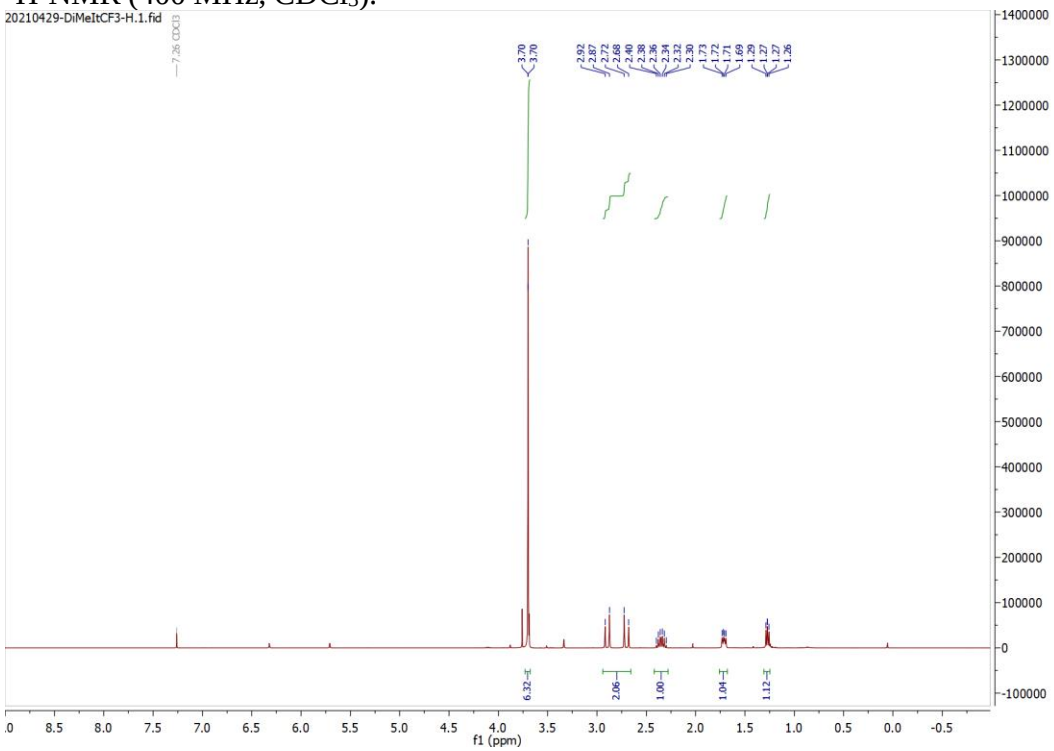
Methyl 1-(2-methoxy-2-oxoethyl)-2-(trifluoromethyl)cyclopropane-1-carboxylate (3):



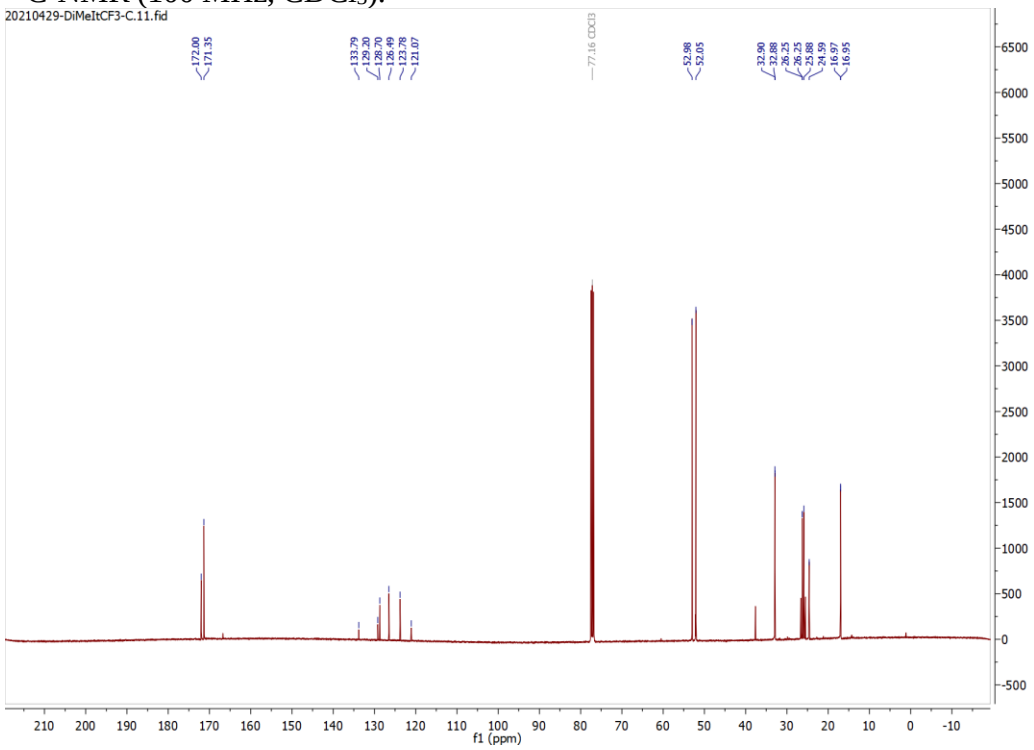
^{19}F-NMR (282 MHz, $CDCl_3$):



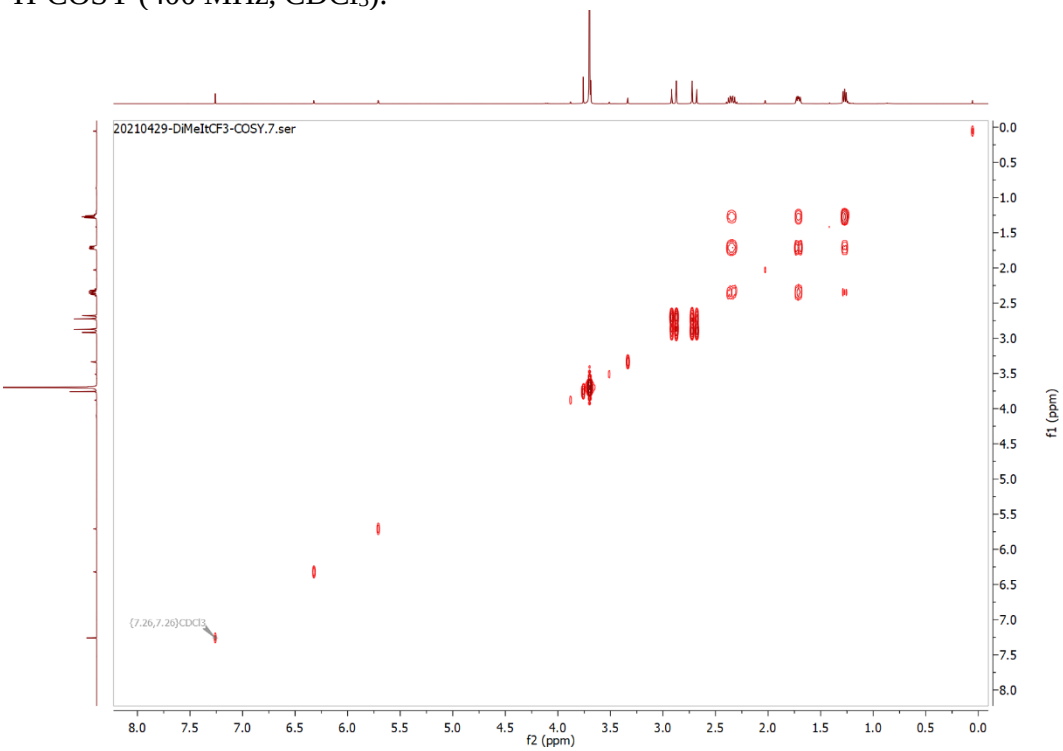
^{1}H-NMR (400 MHz, $CDCl_3$):



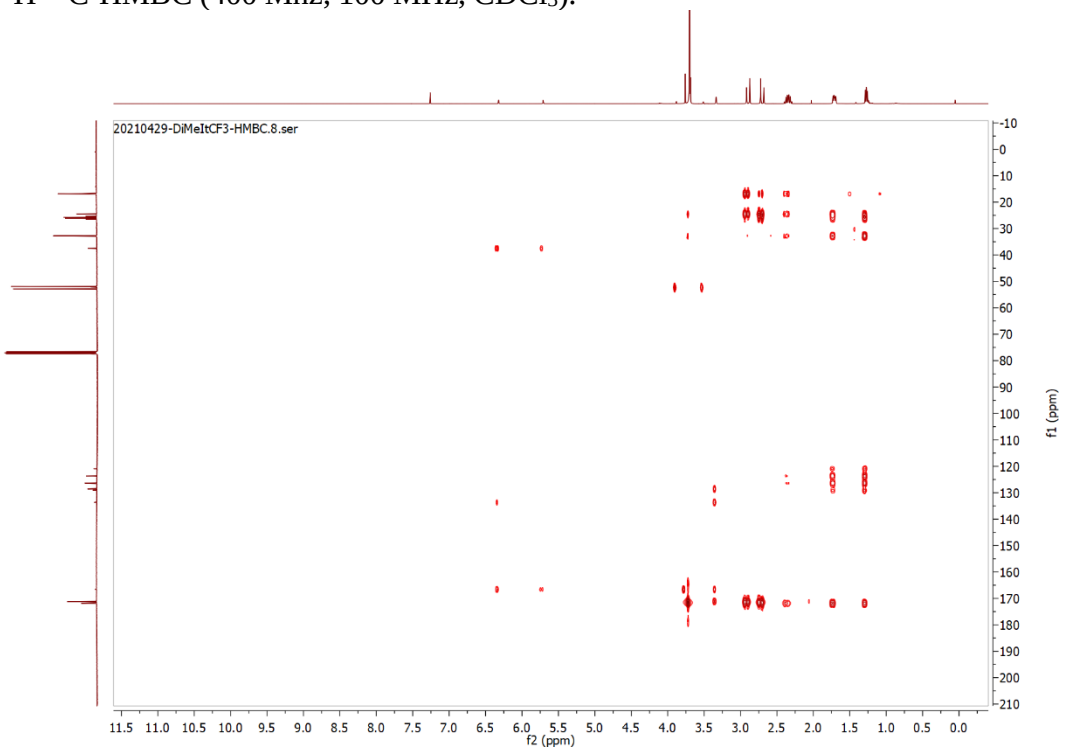
^{13}C-NMR (100 MHz, $CDCl_3$):



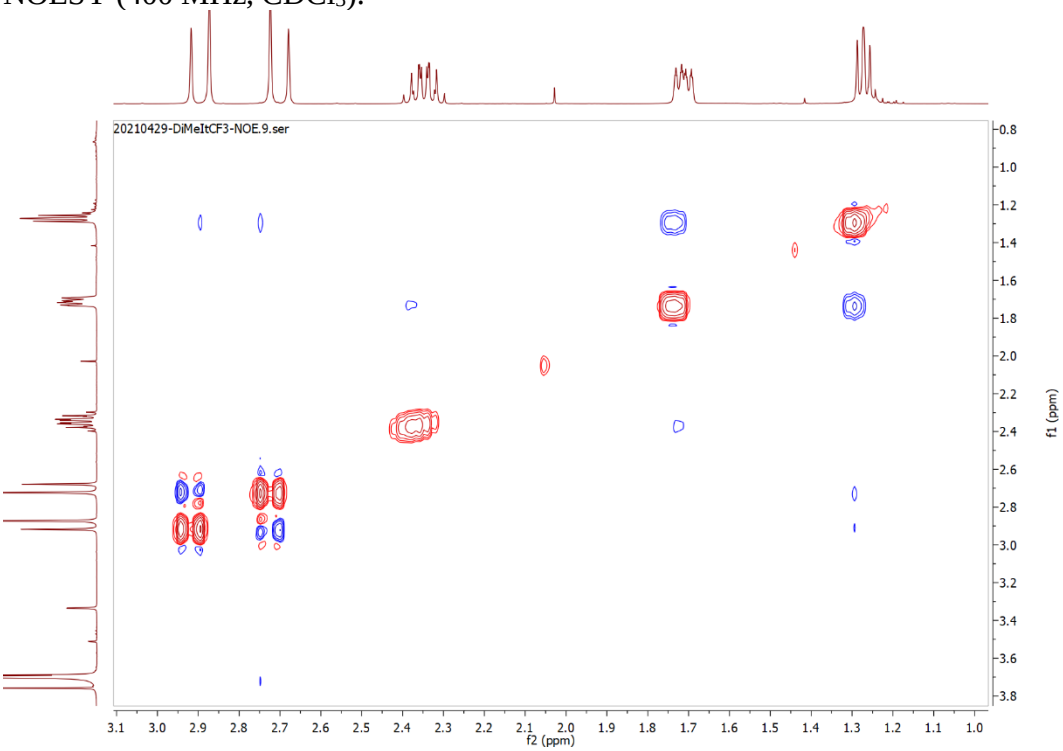
^{1}H-COSY (400 MHz, $CDCl_3$):



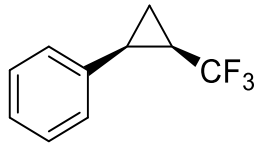
^{1}H-^{13}C-HMBC (400 Mhz, 100 MHz, $CDCl_3$):



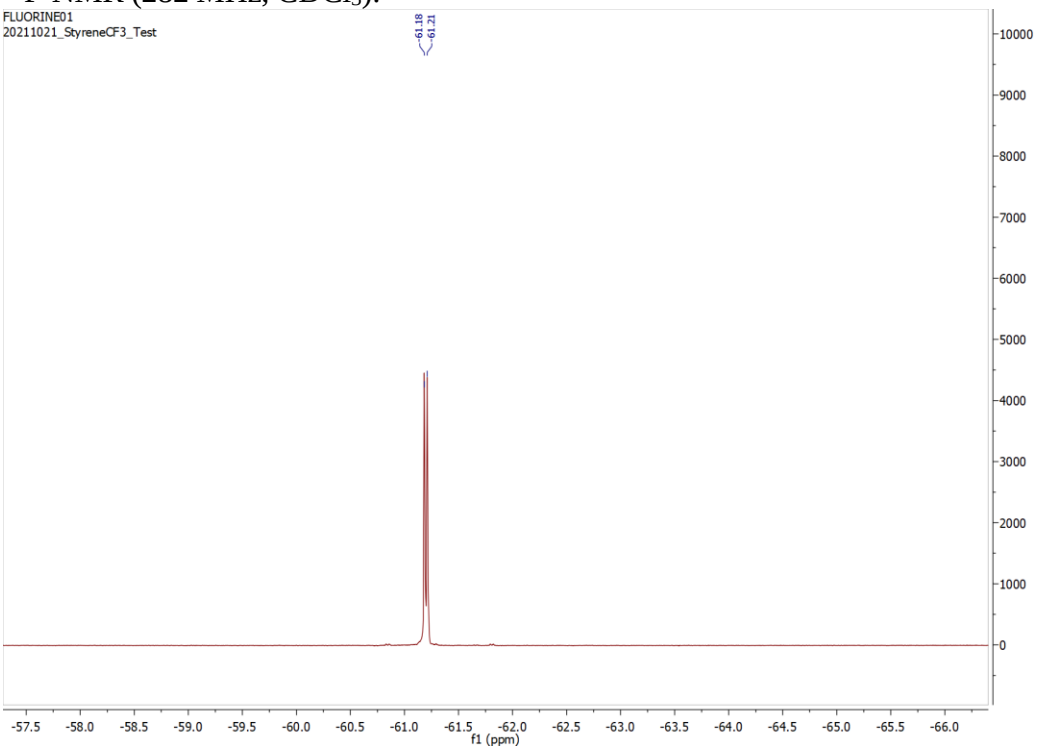
NOESY (400 MHz, $CDCl_3$):



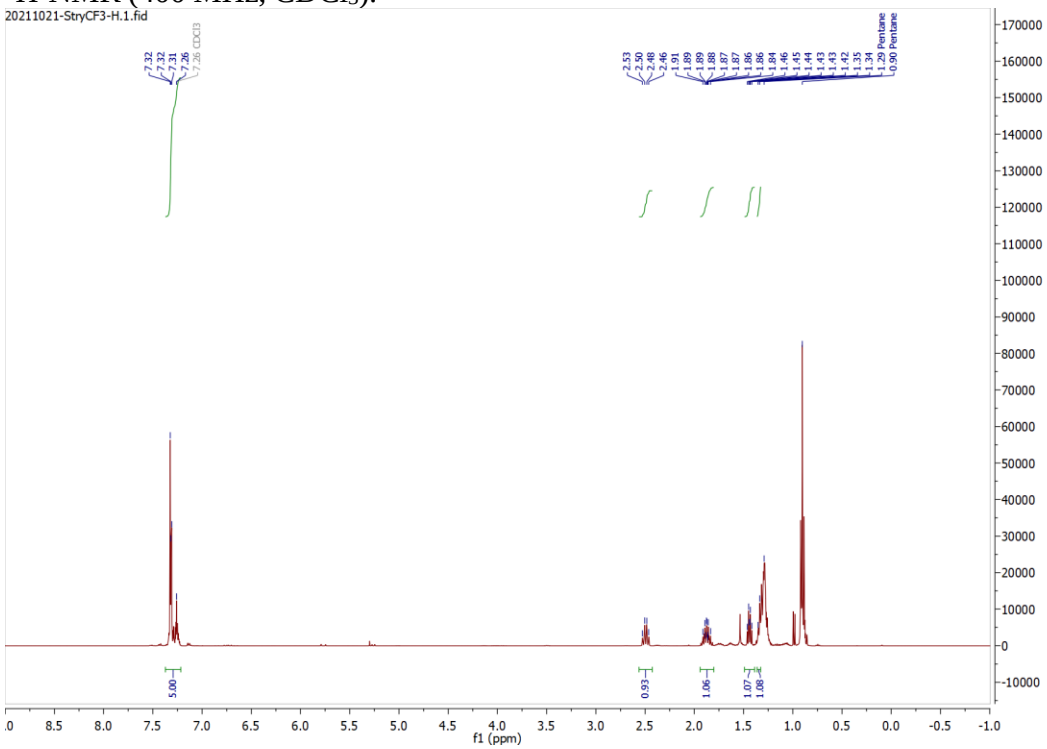
(2-(Trifluoromethyl)cyclopropyl)benzene (4):



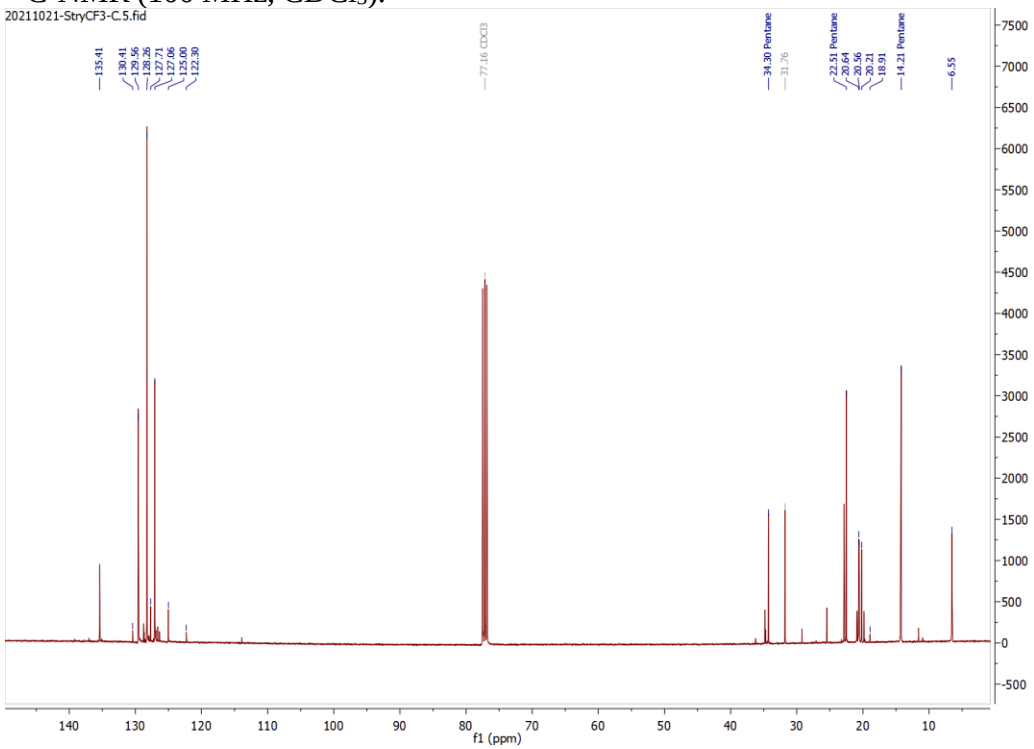
^{19}F-NMR (282 MHz, $CDCl_3$):



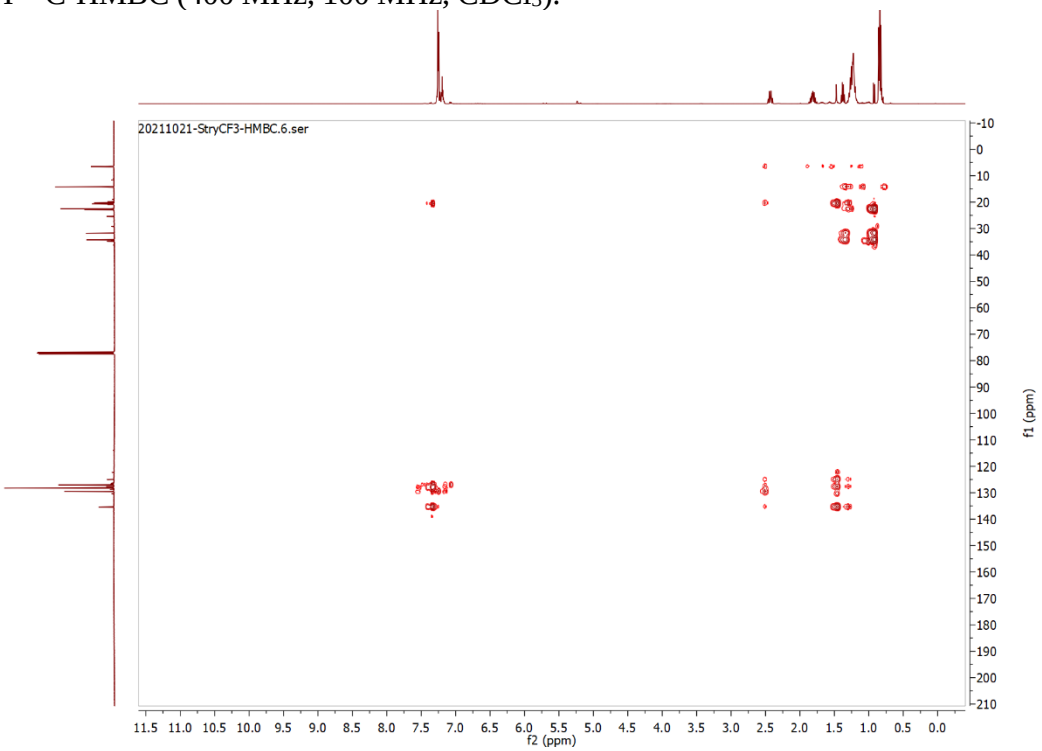
^{1}H-NMR (400 MHz, $CDCl_3$):



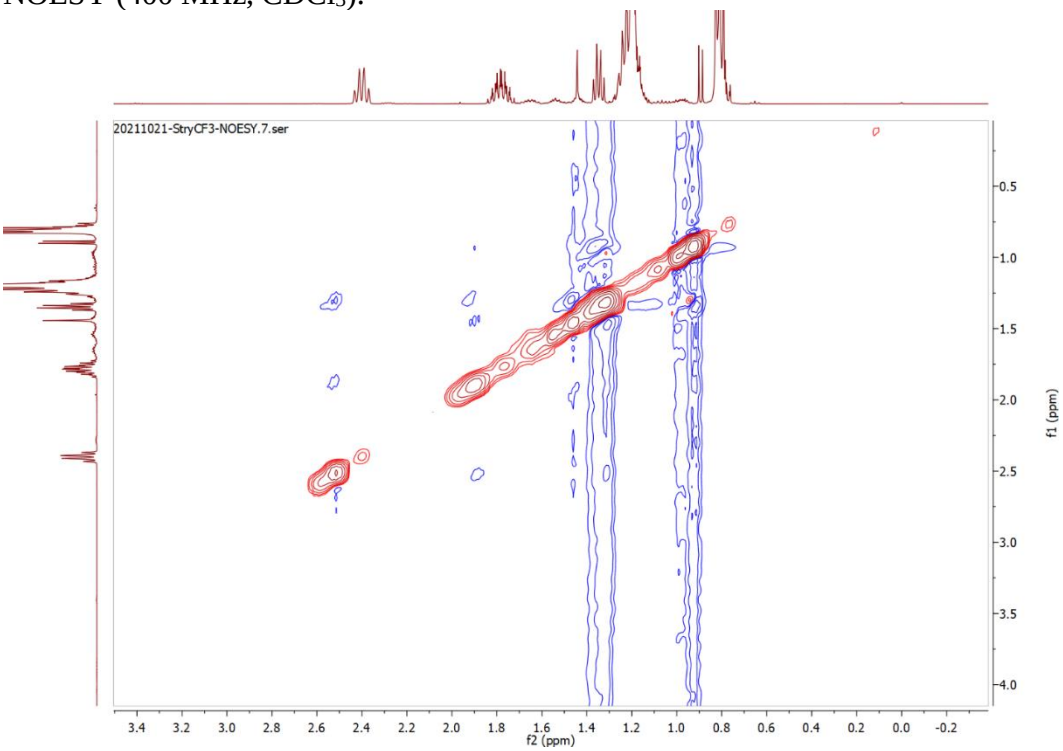
^{13}C-NMR (100 MHz, $CDCl_3$):



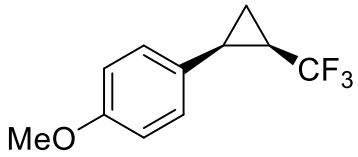
^{1}H-^{13}C-HMBC (400 MHz, 100 MHz, $CDCl_3$):



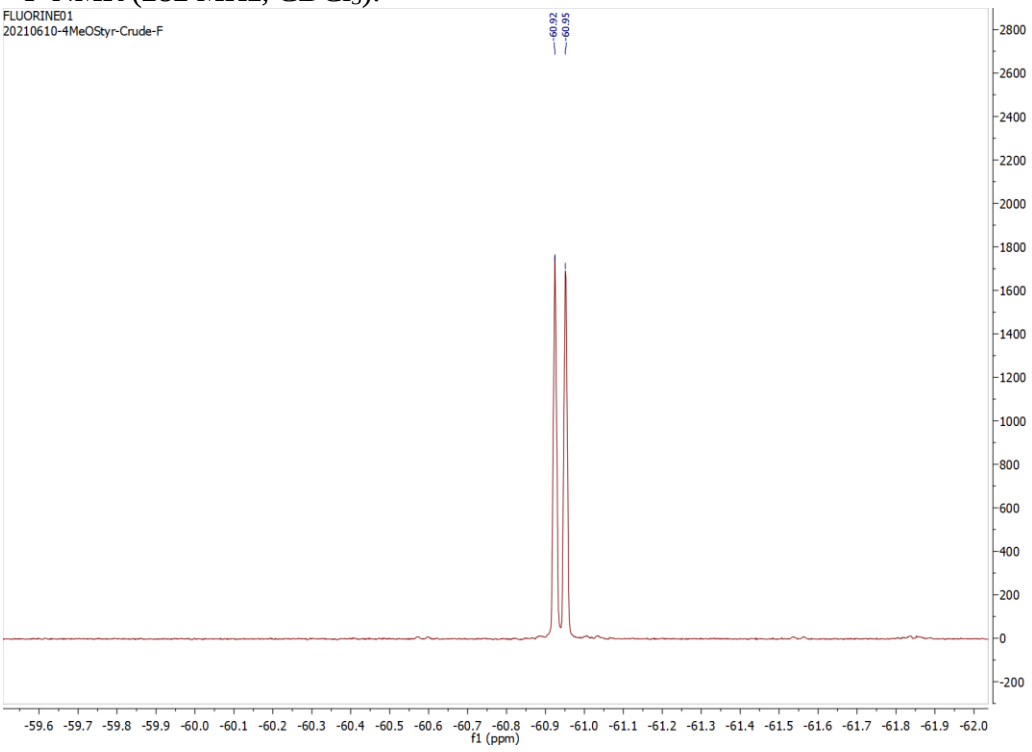
NOESY (400 MHz, $CDCl_3$):



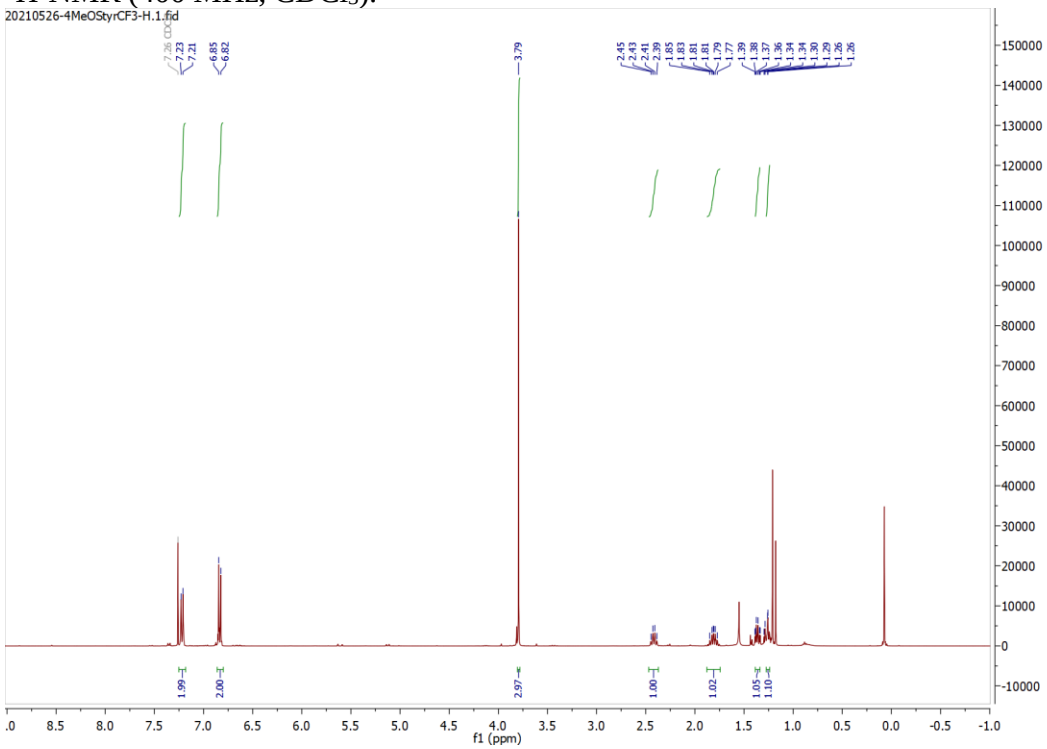
1-Methoxy-4-(2-(trifluoromethyl)cyclopropyl)benzene (5):



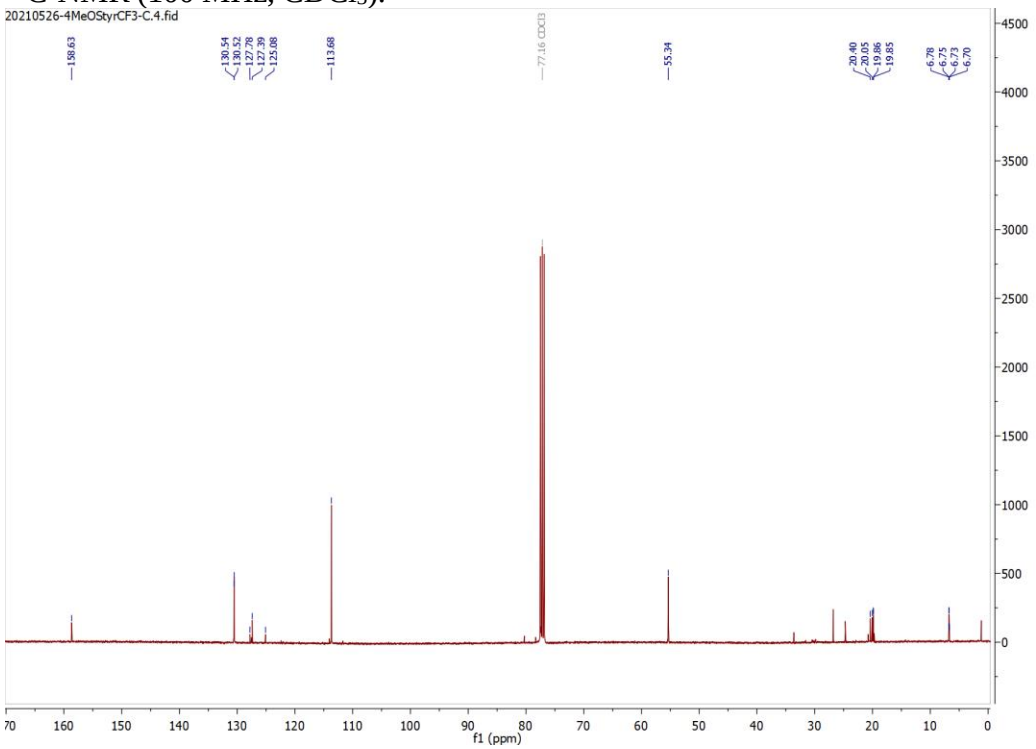
^{19}F-NMR (282 MHz, $CDCl_3$):



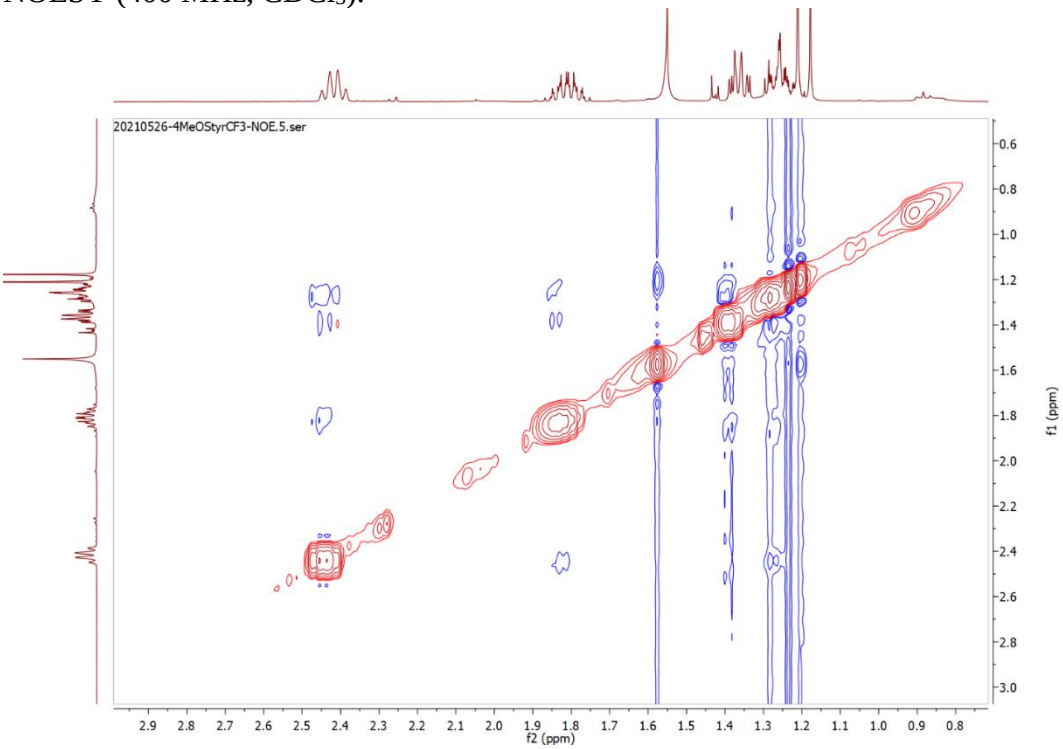
^{1}H-NMR (400 MHz, $CDCl_3$):



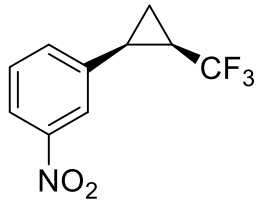
^{13}C-NMR (100 MHz, $CDCl_3$):



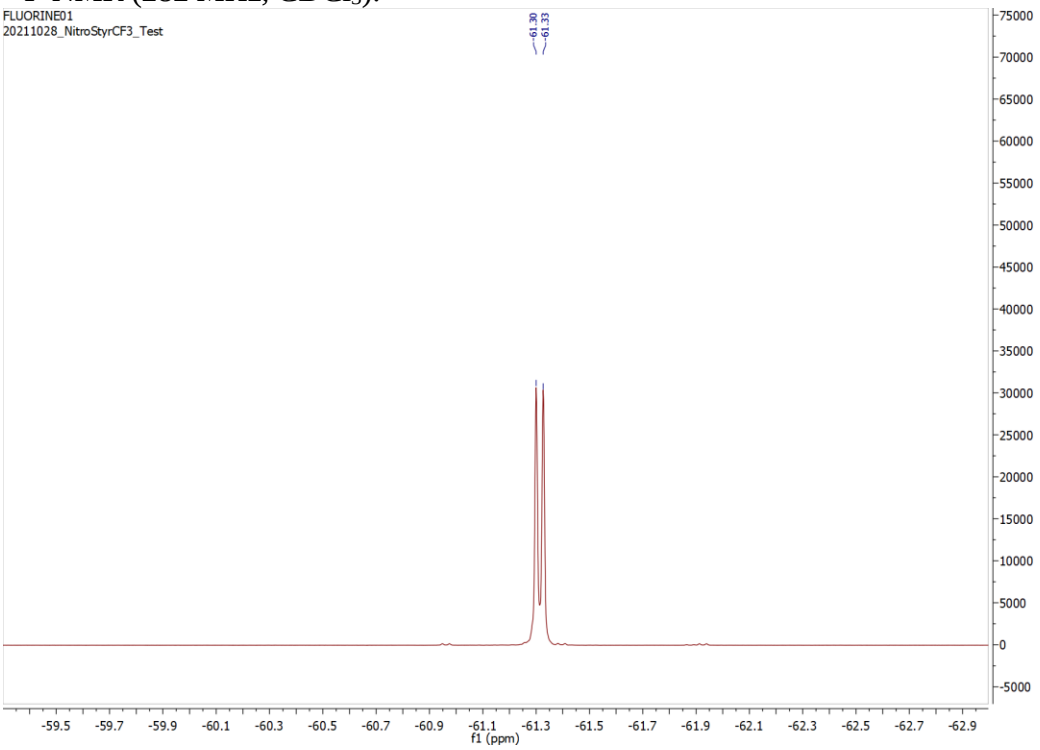
NOESY (400 MHz, $CDCl_3$):



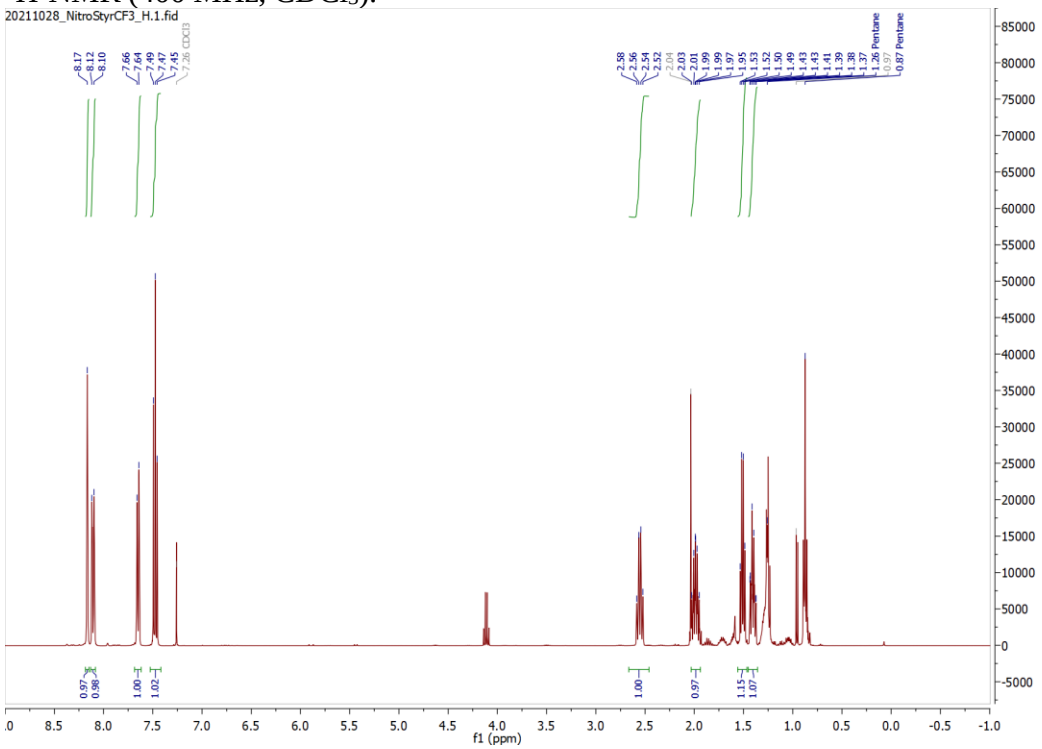
1-Nitro-3-(2-(trifluoromethyl)cyclopropyl)benzene (6):



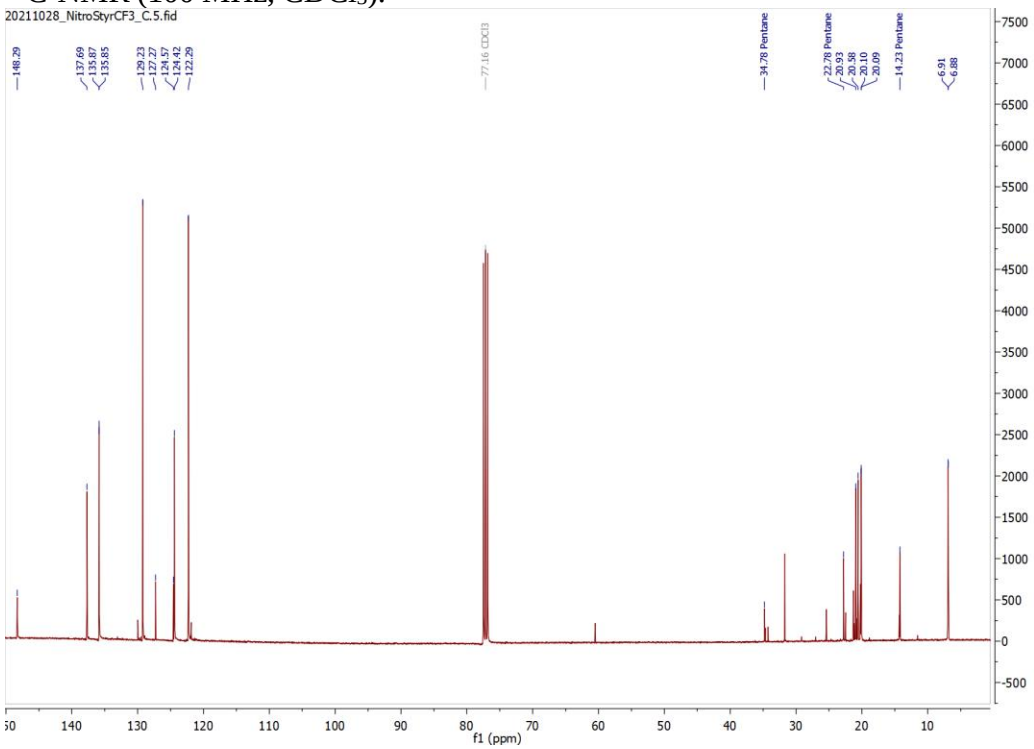
^{19}F-NMR (282 MHz, $CDCl_3$):



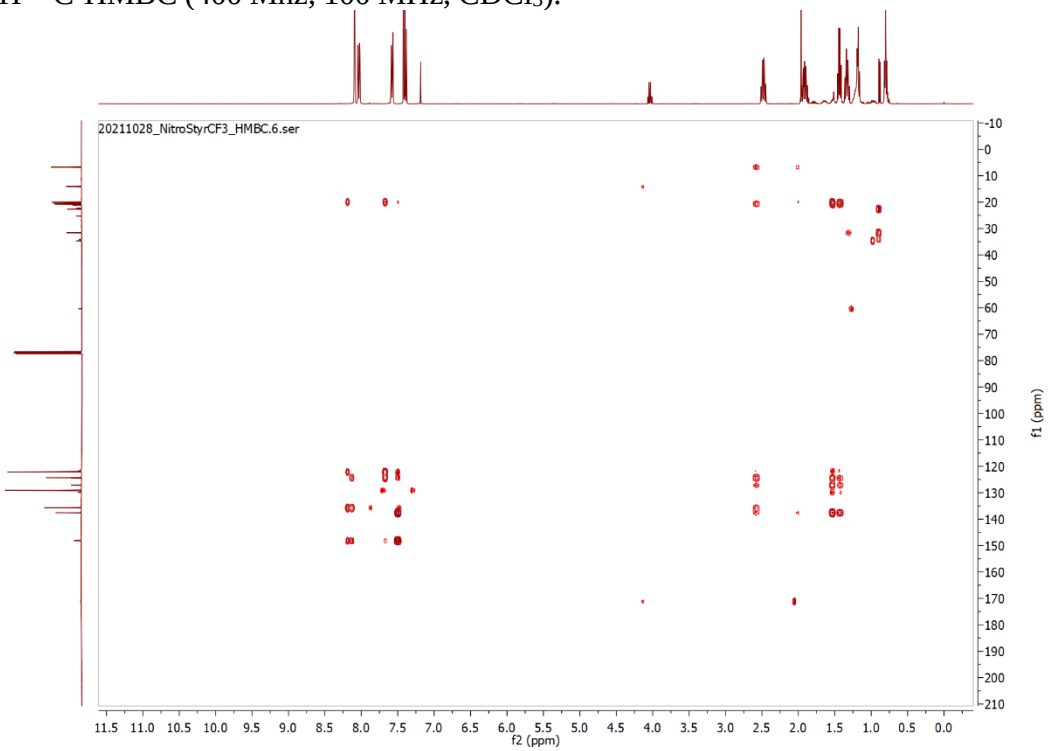
^{1}H-NMR (400 MHz, $CDCl_3$):



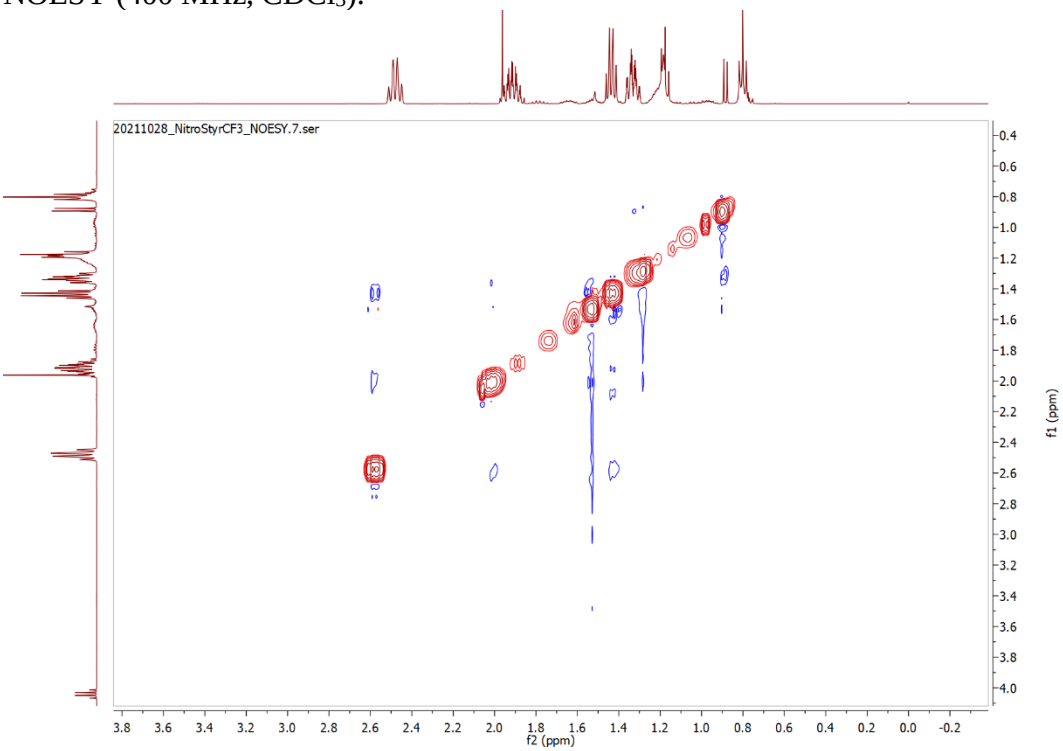
^{13}C-NMR (100 MHz, $CDCl_3$):



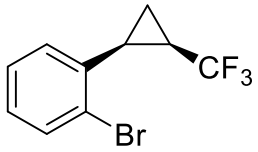
^{1}H-^{13}C-HMBC (400 Mhz, 100 MHz, $CDCl_3$):



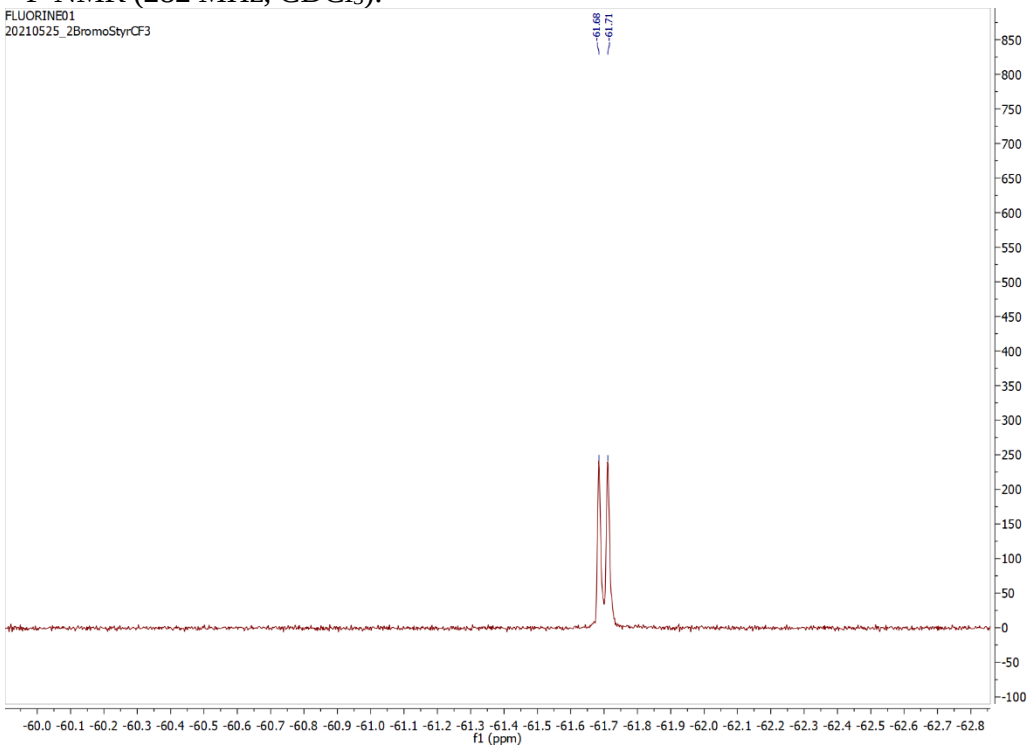
NOESY (400 MHz, $CDCl_3$):



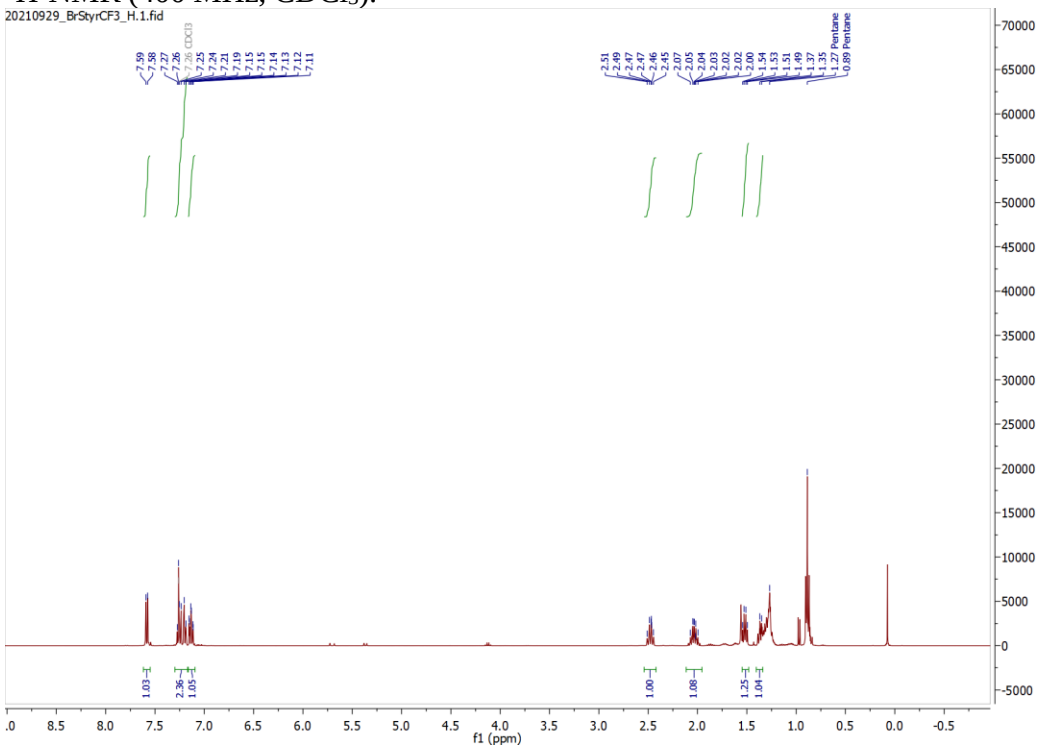
1-Bromo-2-(2-(trifluoromethyl)cyclopropyl)benzene (7):



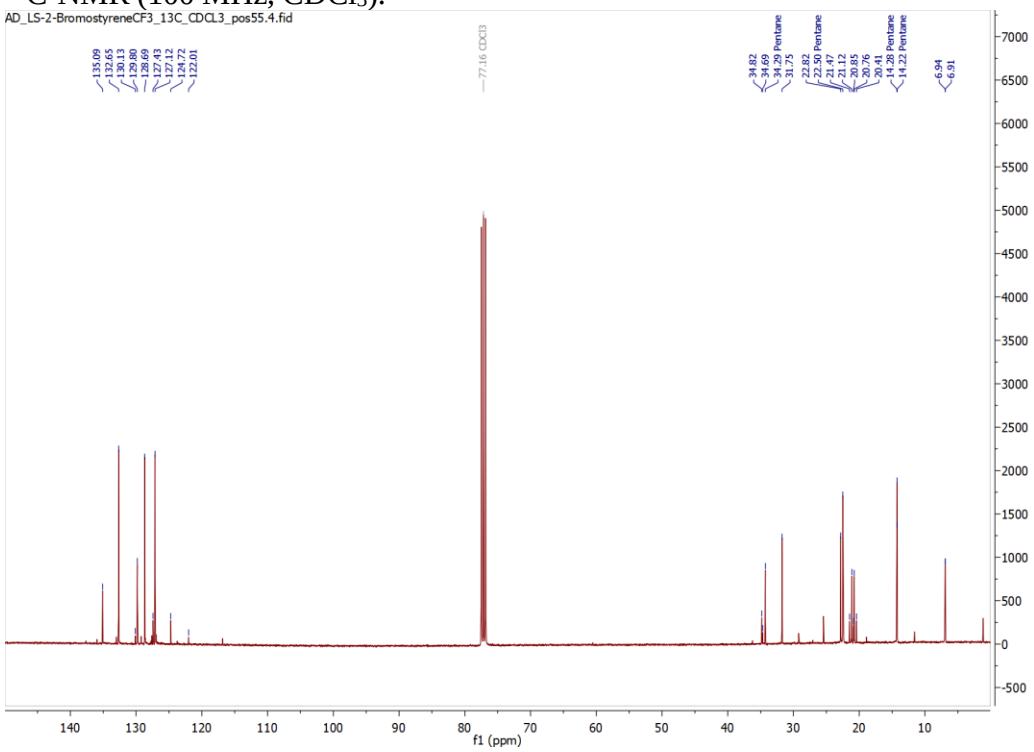
^{19}F-NMR (282 MHz, $CDCl_3$):



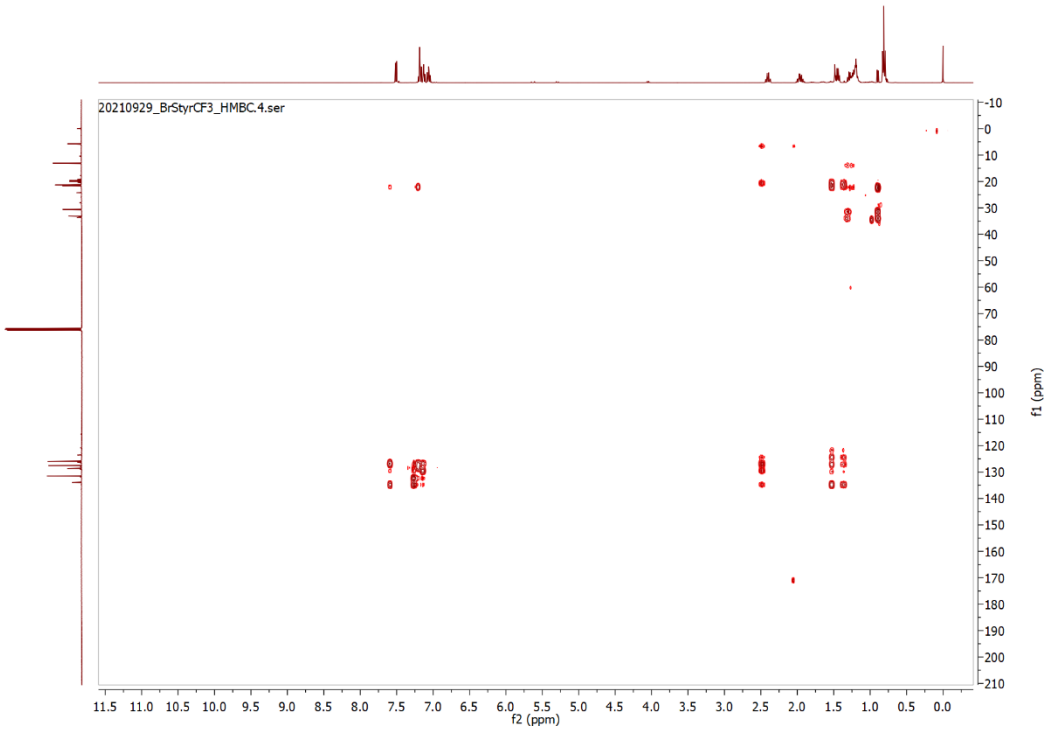
^{1}H-NMR (400 MHz, $CDCl_3$):



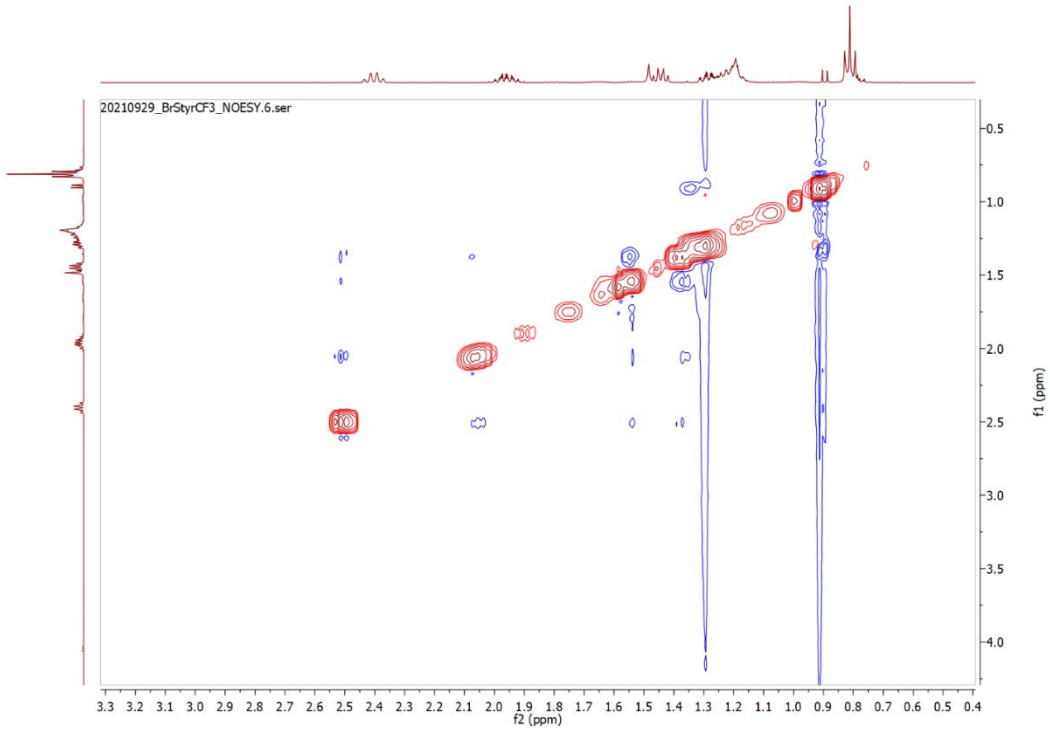
^{13}C-NMR (100 MHz, $CDCl_3$):



^{1}H-^{13}C-HMBC (400 Mhz, 100 MHz, $CDCl_3$):

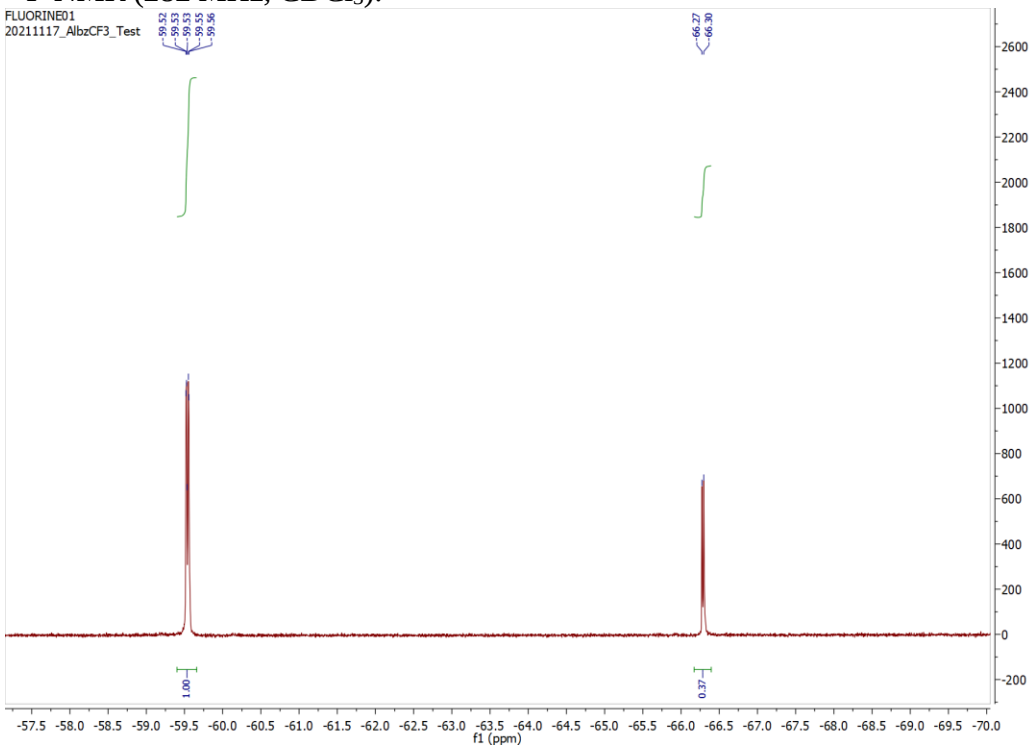


NOESY (400 MHz, $CDCl_3$):

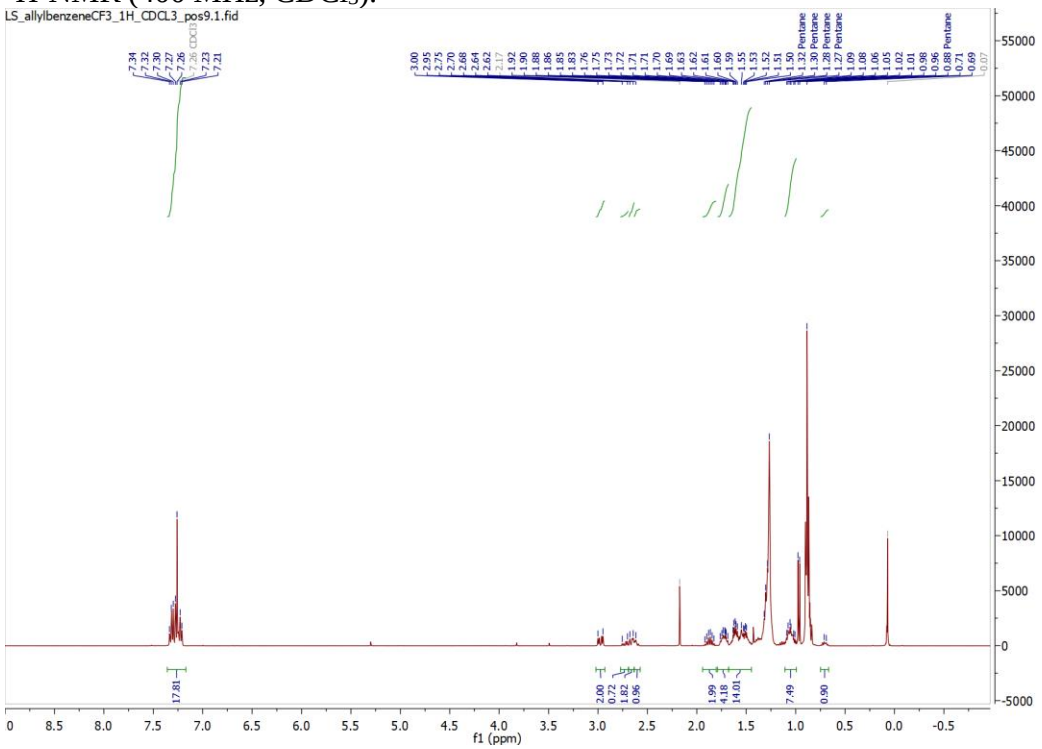


((2-(Trifluoromethyl)cyclopropyl)methyl)benzene (8):

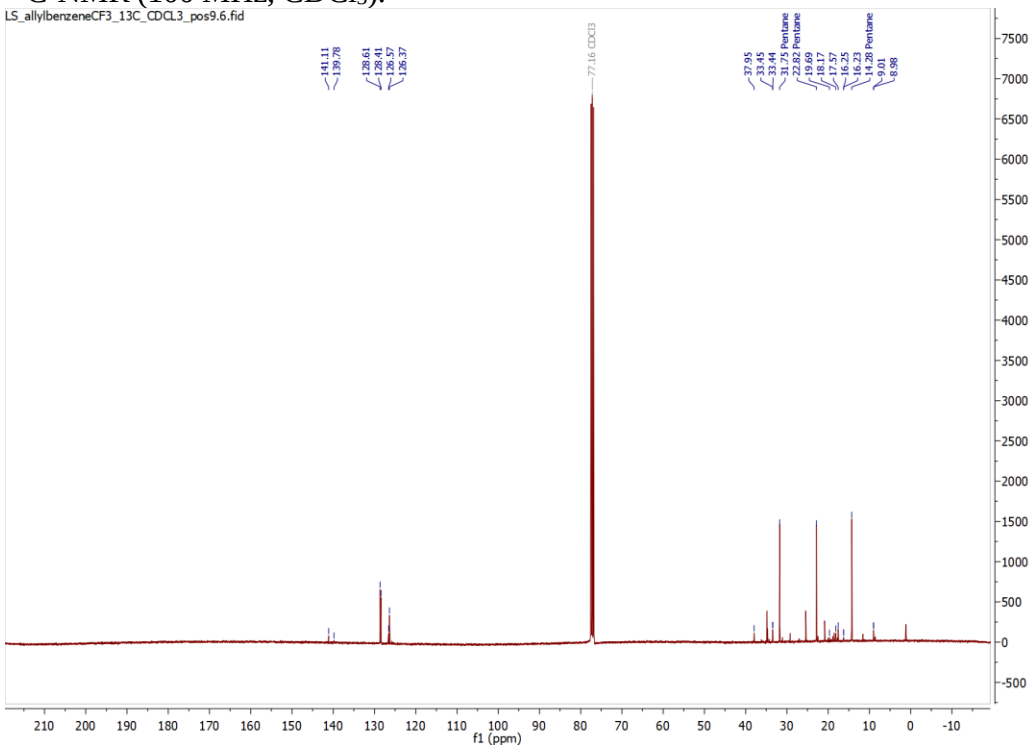
^{19}F-NMR (282 MHz, $CDCl_3$):



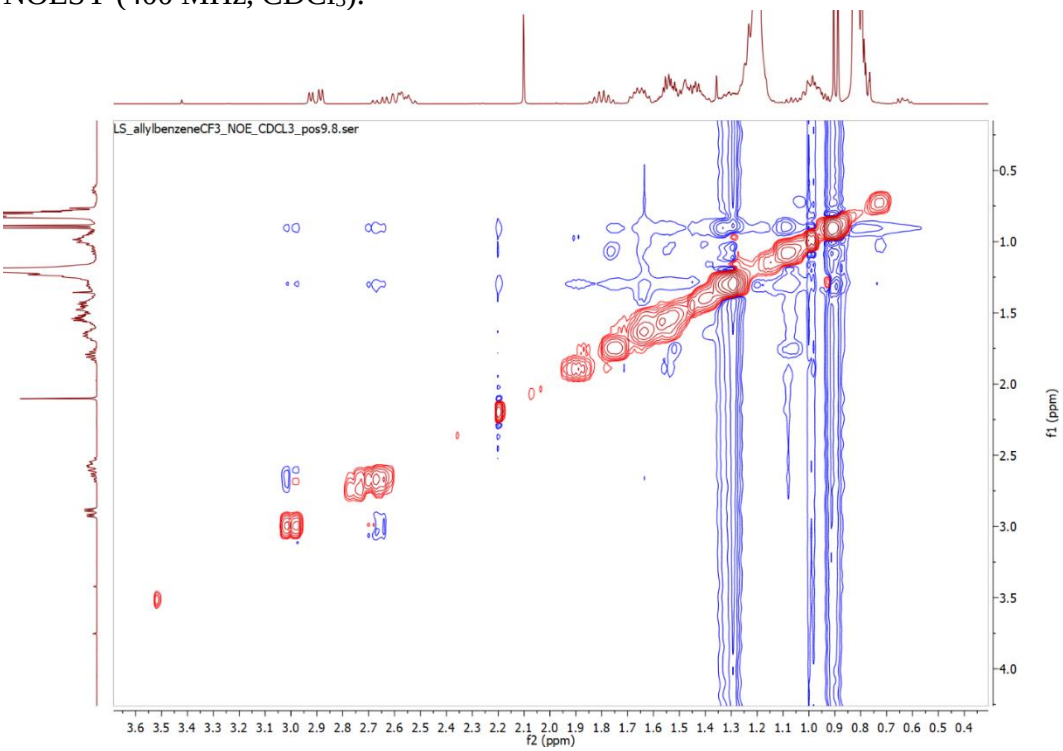
^{1}H-NMR (400 MHz, $CDCl_3$):



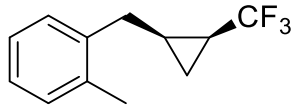
^{13}C-NMR (100 MHz, $CDCl_3$):



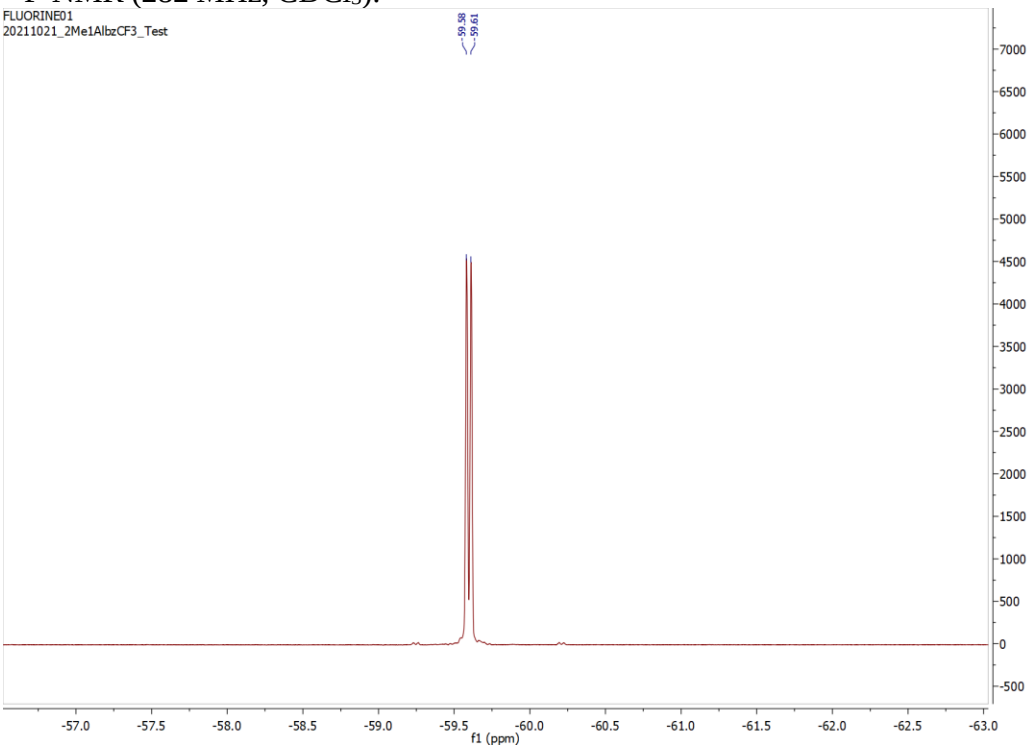
NOESY (400 MHz, $CDCl_3$):



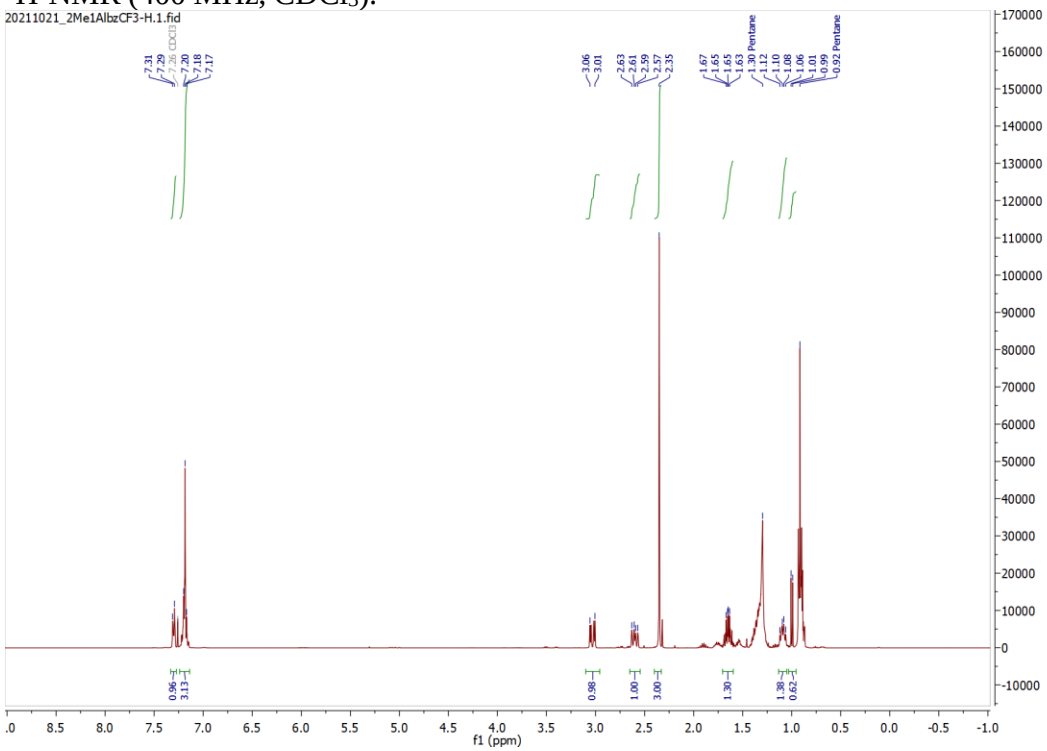
1-Methyl-2-((2-(trifluoromethyl)cyclopropyl)methyl)benzene (9):



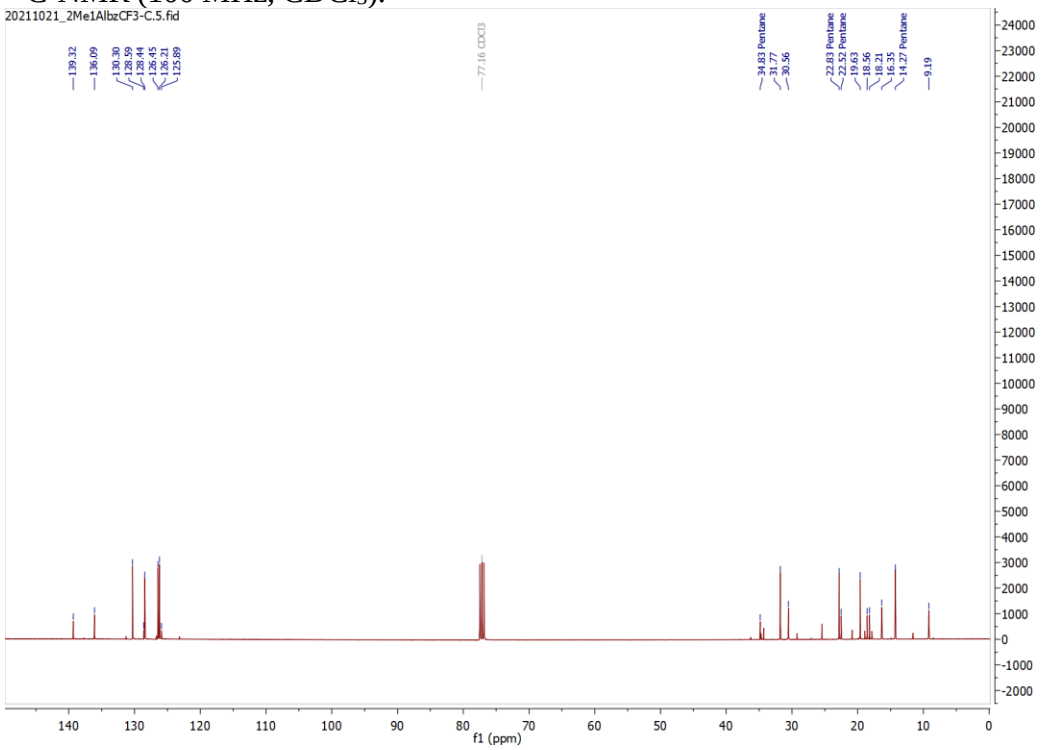
^{19}F-NMR (282 MHz, $CDCl_3$):



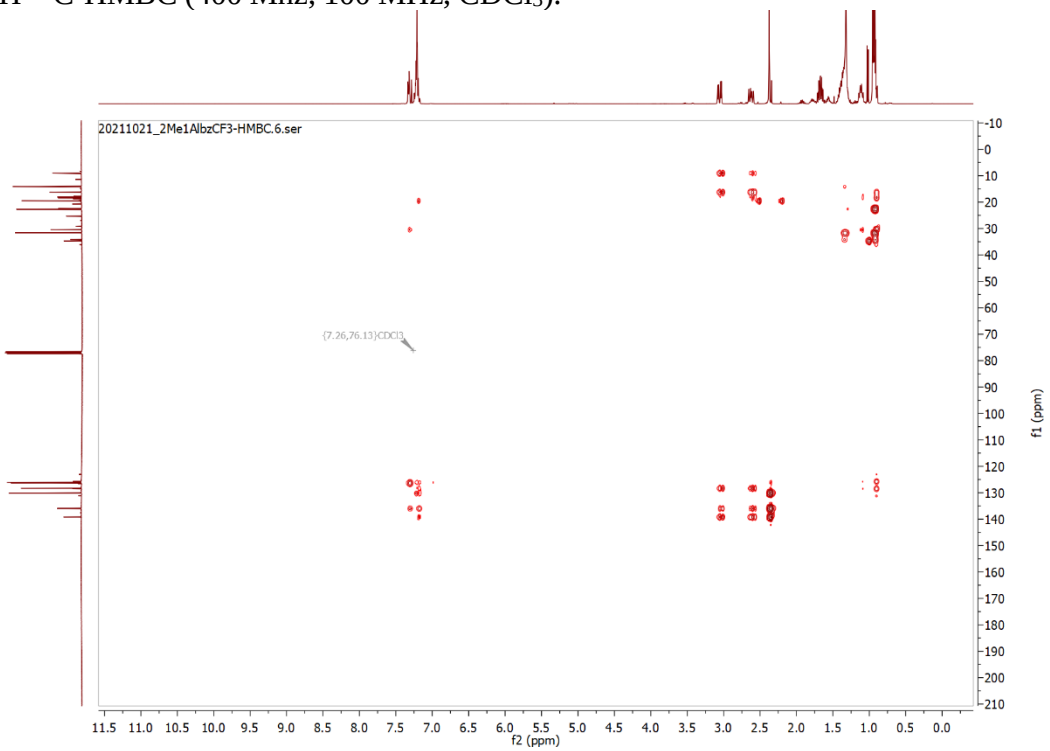
^{1}H-NMR (400 MHz, $CDCl_3$):



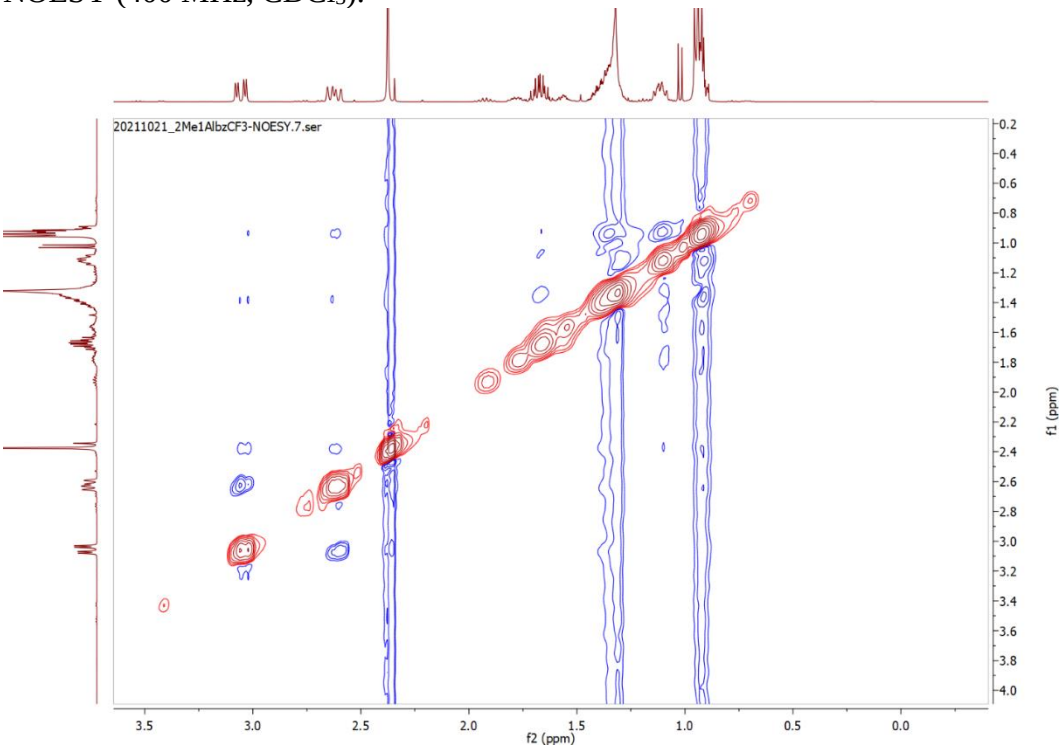
^{13}C-NMR (100 MHz, $CDCl_3$):



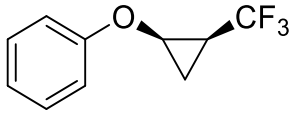
^{1}H-^{13}C-HMBC (400 Mhz, 100 MHz, $CDCl_3$):



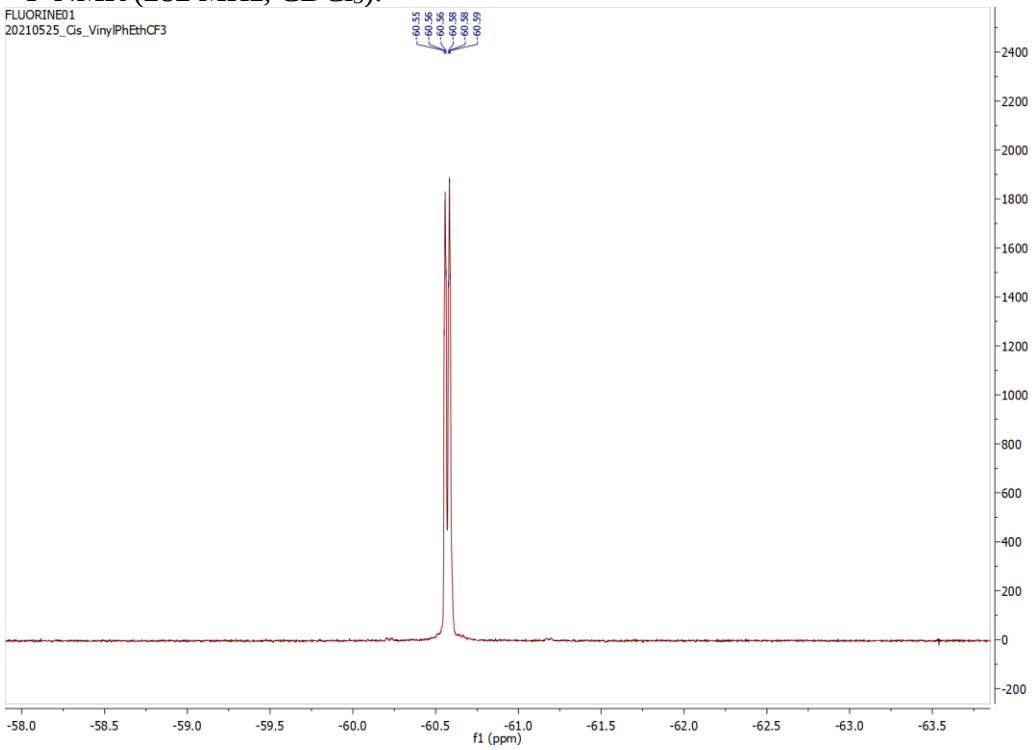
NOESY (400 MHz, $CDCl_3$):



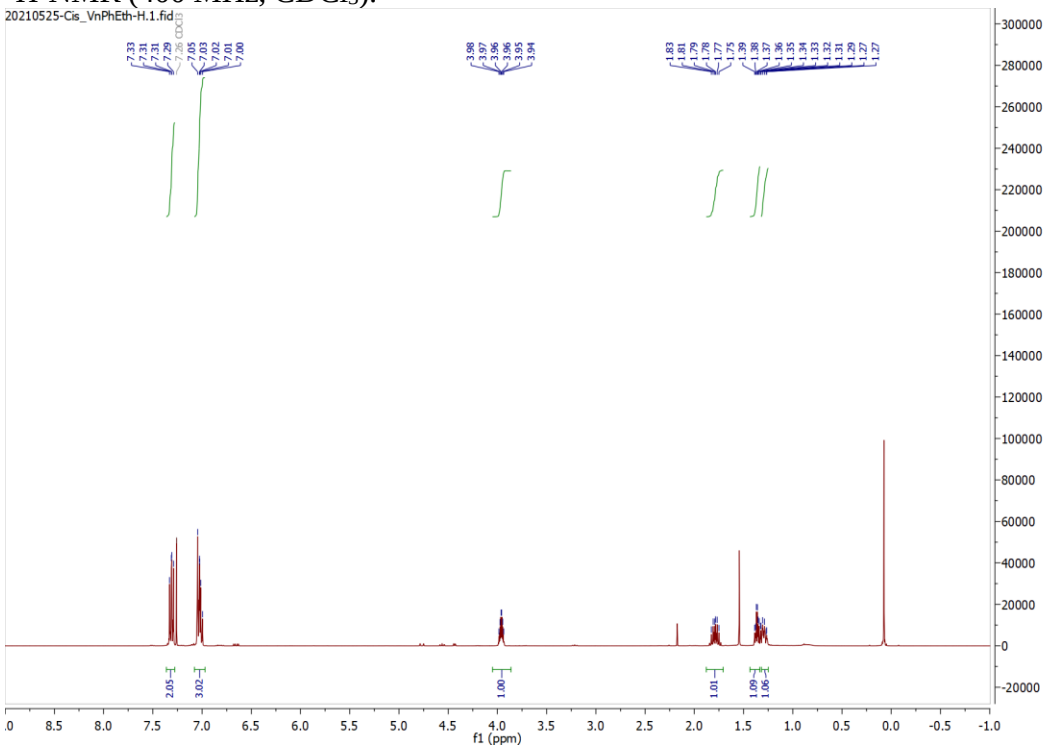
(2-(Trifluoromethyl)cyclopropoxy)benzene (10):



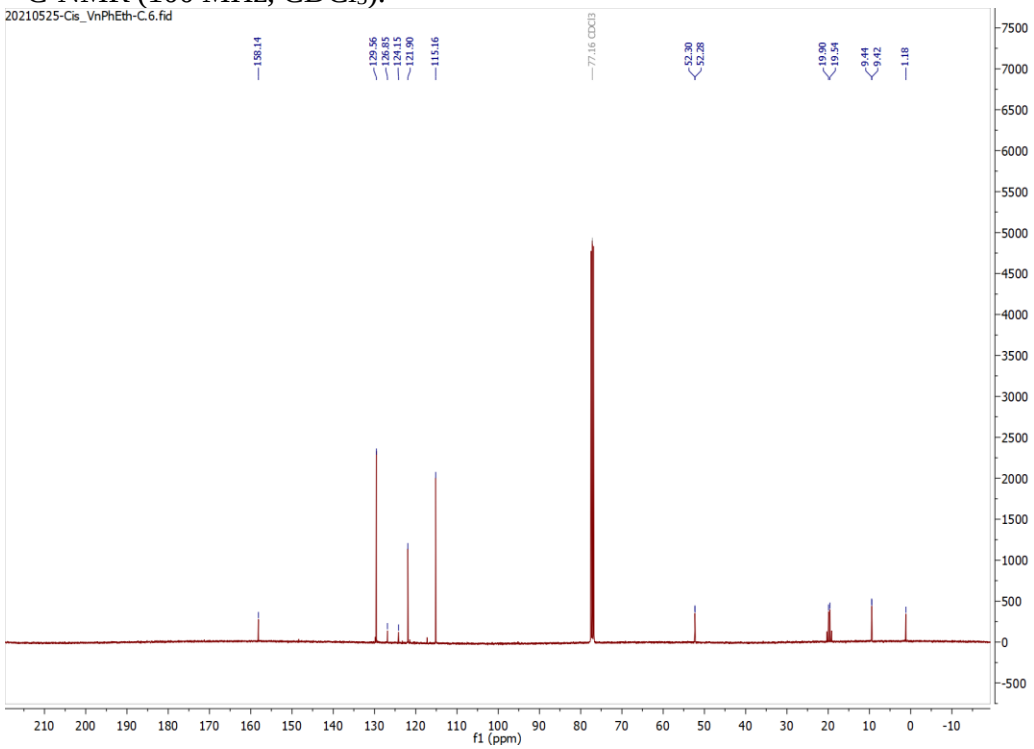
^{19}F-NMR (282 MHz, $CDCl_3$):



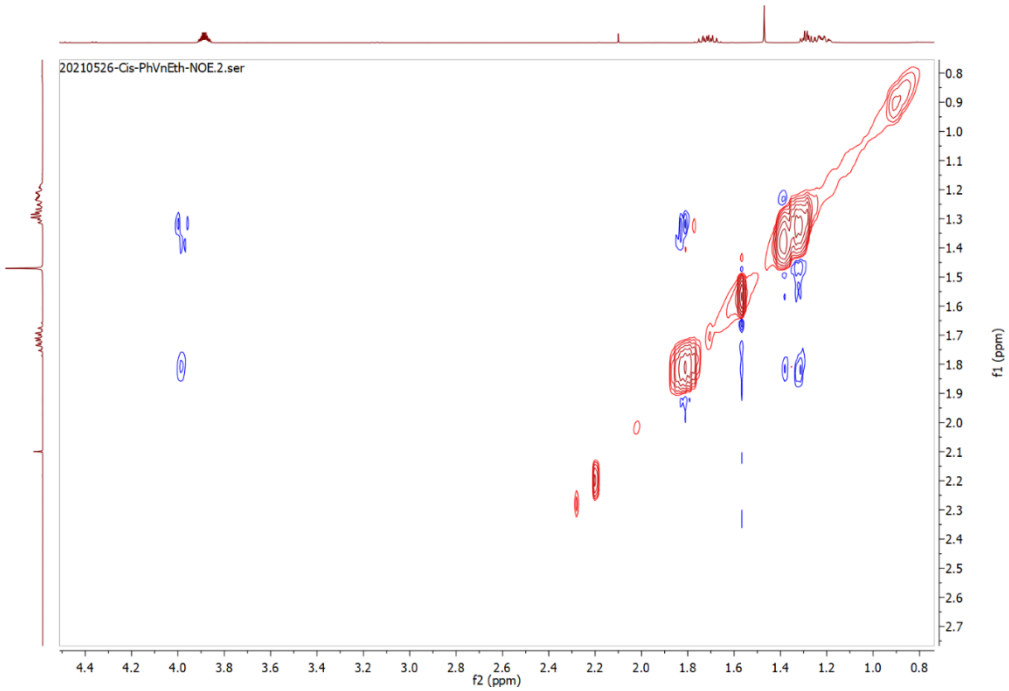
^{1}H-NMR (400 MHz, $CDCl_3$):



^{13}C-NMR (100 MHz, $CDCl_3$):

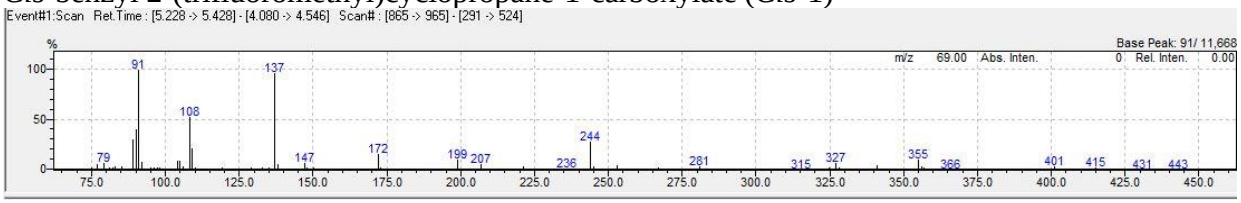


NOESY (400 MHz, $CDCl_3$):

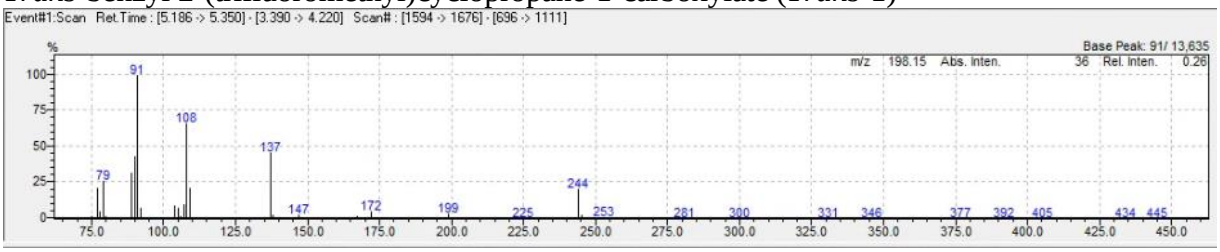


Mass Spectra from GC-MS

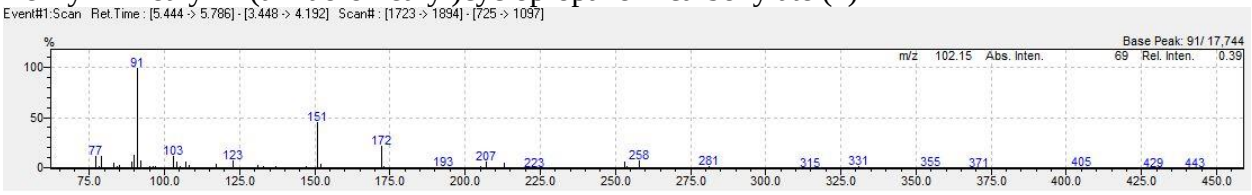
Cis-benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate (*Cis*-1)



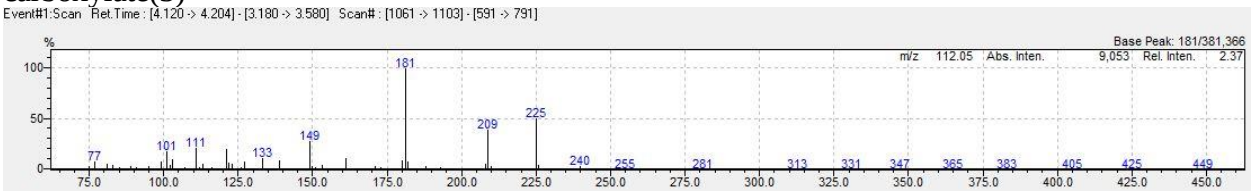
Trans-benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate (*Trans*-1)



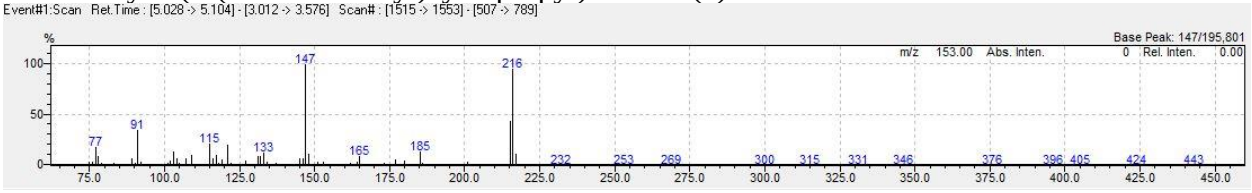
Benzyl 1-methyl-2-(trifluoromethyl)cyclopropane-1-carboxylate (2)



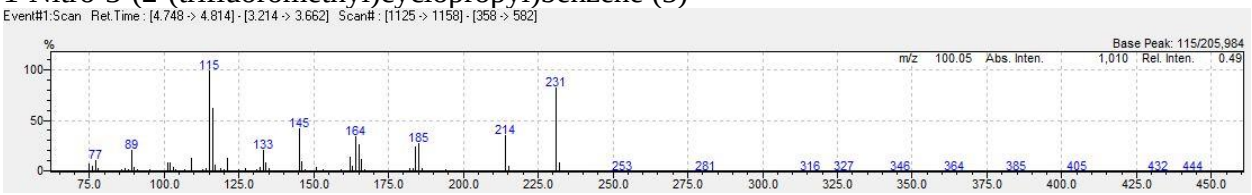
Methyl 1-(2-methoxy-2-oxoethyl)-2-(trifluoromethyl)cyclopropane-1-carboxylate(3)



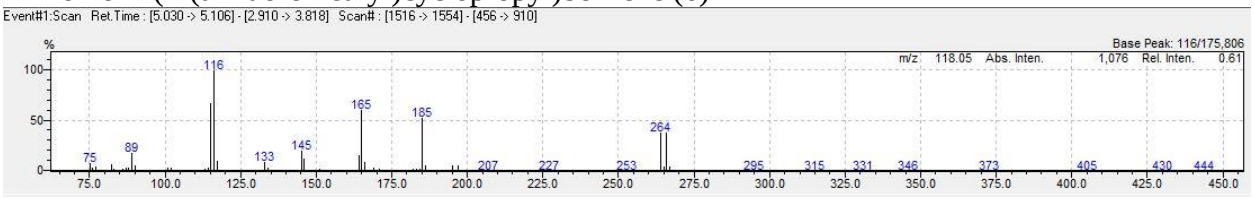
1-Methoxy-4-(2-(trifluoromethyl)cyclopropyl)benzene(4)



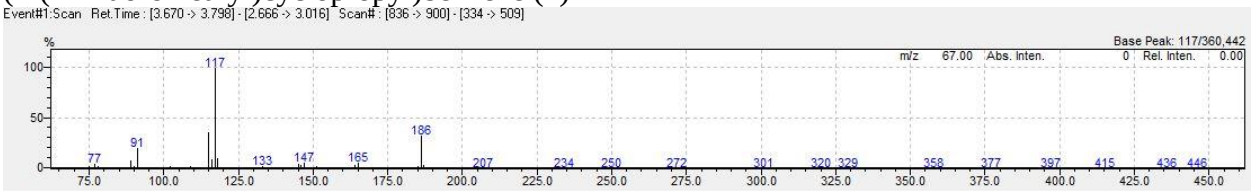
1-Nitro-3-(2-(trifluoromethyl)cyclopropyl)benzene (5)



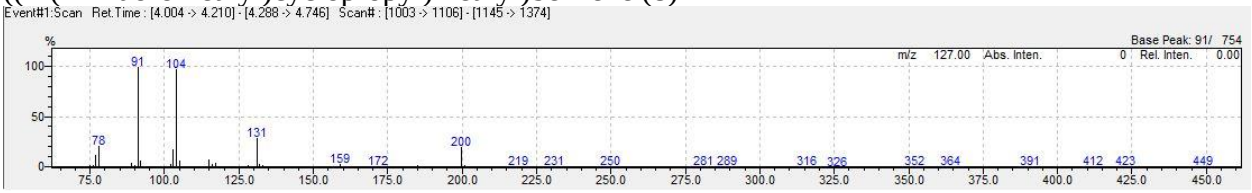
1-Bromo-2-(2-(trifluoromethyl)cyclopropyl)benzene (6)



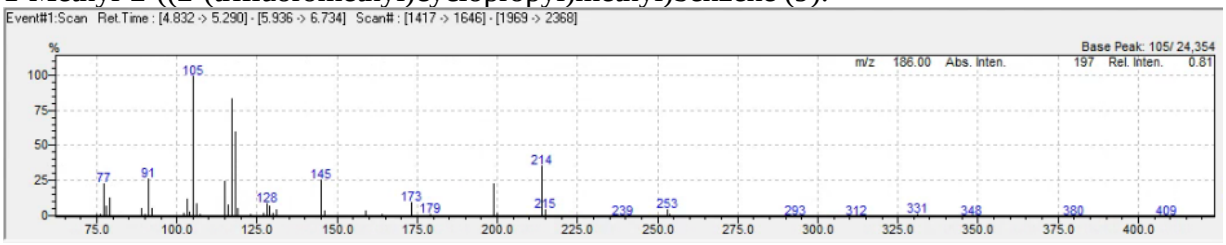
(2-(Trifluoromethyl)cyclopropyl)benzene (7)



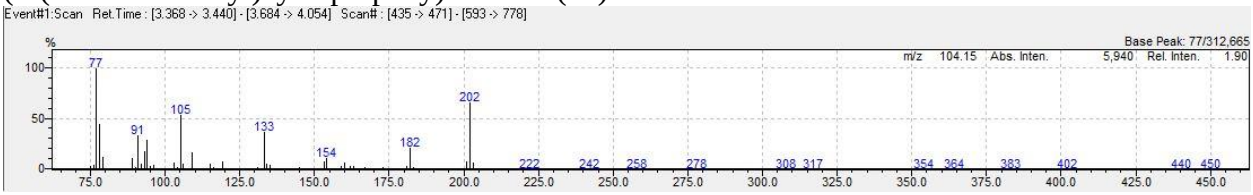
((2-(Trifluoromethyl)cyclopropyl)methyl)benzene (8)



1-Methyl-2-((2-(trifluoromethyl)cyclopropyl)methyl)benzene (9):

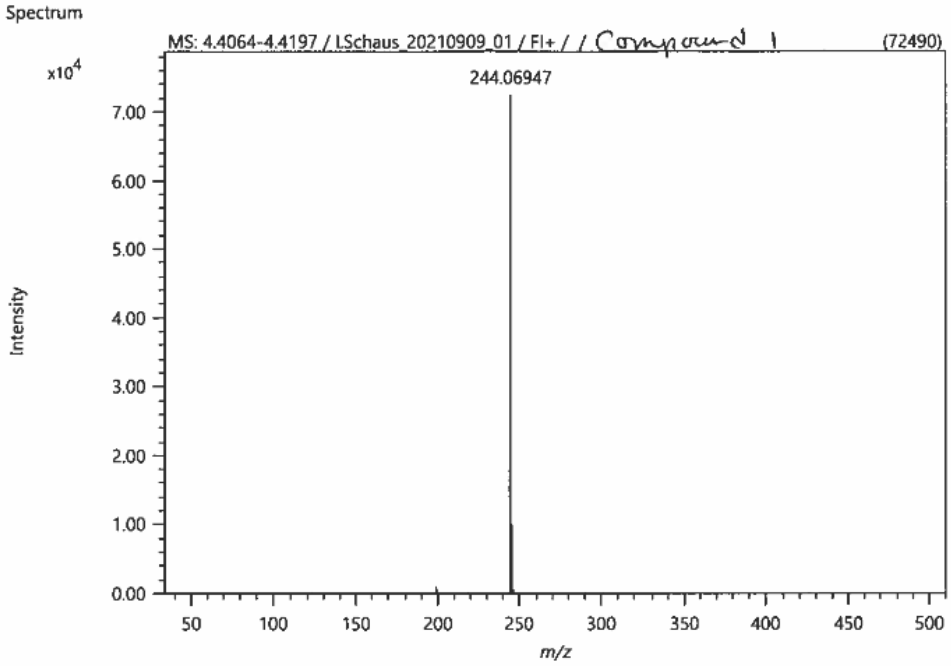


(2-(Trifluoromethyl)cyclopropoxy)benzene (10)



High Resolution Mass Spectra

Benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate (1)



Elemental Composition

Parameters	
Tolerance:	±10.00 mDa
Electron:	Odd/Even
Charge:	+1
DBE:	-1.5 - 20.0

Elements Set 1:

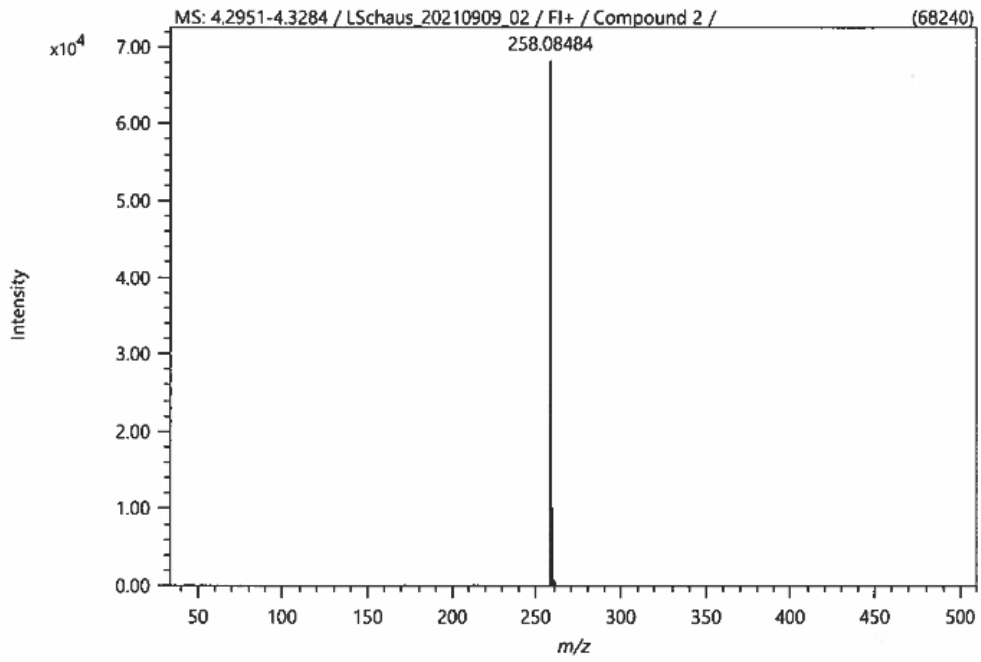
Symbol	C	H	O	F
Min	0	0	0	3
Max	50	100	3	3

Results

Mass	Intensity	Formula	Calculated Mass	Mass Difference [mDa]	Mass Difference [ppm]	DBE
244.06947	72489.85	C12 H11 O2 F3	244.07057	-1.09	-4.48	6.0

Benzyl 1-methyl-2-(trifluoromethyl)cyclopropane-1-carboxylate (2)

Spectrum



Elemental Composition

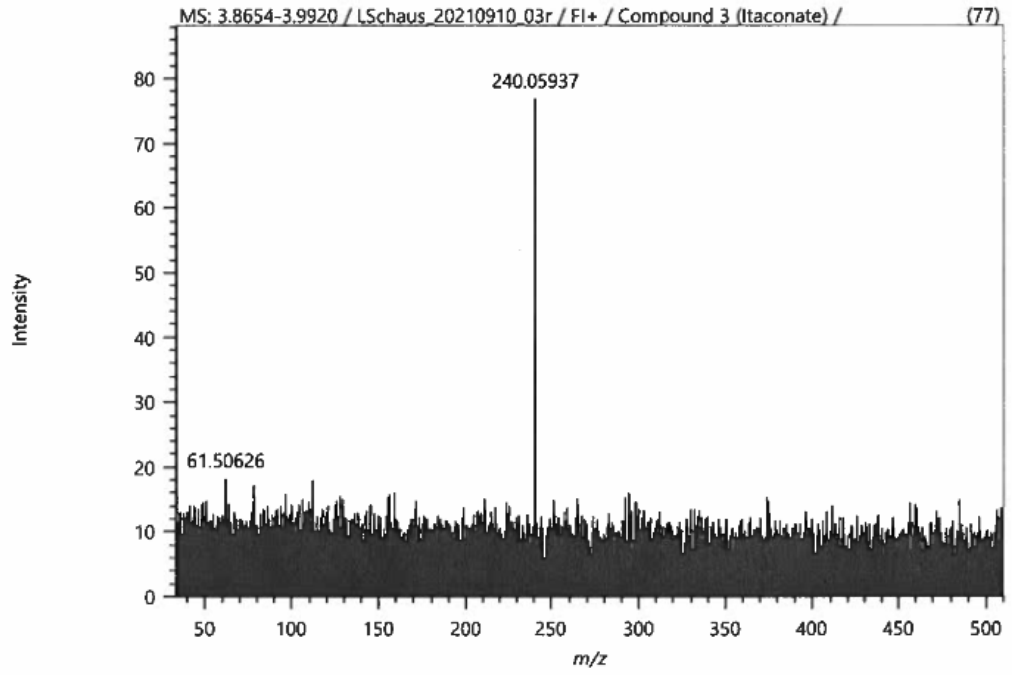
Parameters		Elements Set 1:				
Tolerance:	±5.00 mDa	Symbol	C	H	O	F
Electron:	Odd/Even	Min	0	0	0	3
Charge:	+1	Max	50	100	3	3
DBE:	-1.5 - 20.0					

Results

Mass	Intensity	Formula	Calculated Mass	Mass Difference [mDa]	Mass Difference [ppm]	DBE
258.08484	68239.77	C13 H13 O2 F3	258.08622	-1.38	-5.34	6.0

Methyl 1-(2-methoxy-2-oxoethyl)-2-(trifluoromethyl)cyclopropane-1-carboxylate (3)

Spectrum



Elemental Composition

Parameters	
Tolerance:	±5.00 mDa
Electron:	Odd/Even
Charge:	+1
DBE:	-1.5 - 20.0

Elements Set 1:

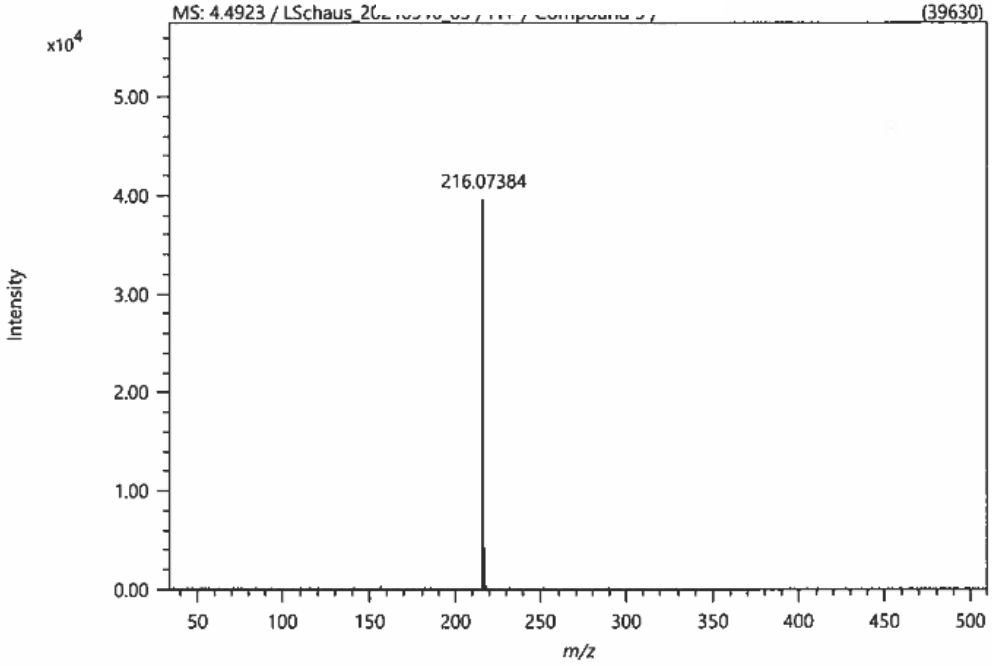
Symbol	C	H	O	F
Min	0	0	0	3
Max	50	100	5	3

Results

Mass	Intensity	Formula	Calculated Mass	Mass Difference [mDa]	Mass Difference [ppm]	DBE
240.05937	76.97	C9 H11 O4 F3	240.06039	-1.03	-4.29	3.0

1-Methoxy-4-(2-(trifluoromethyl)cyclopropyl)benzene (5)

Spectrum



Elemental Composition

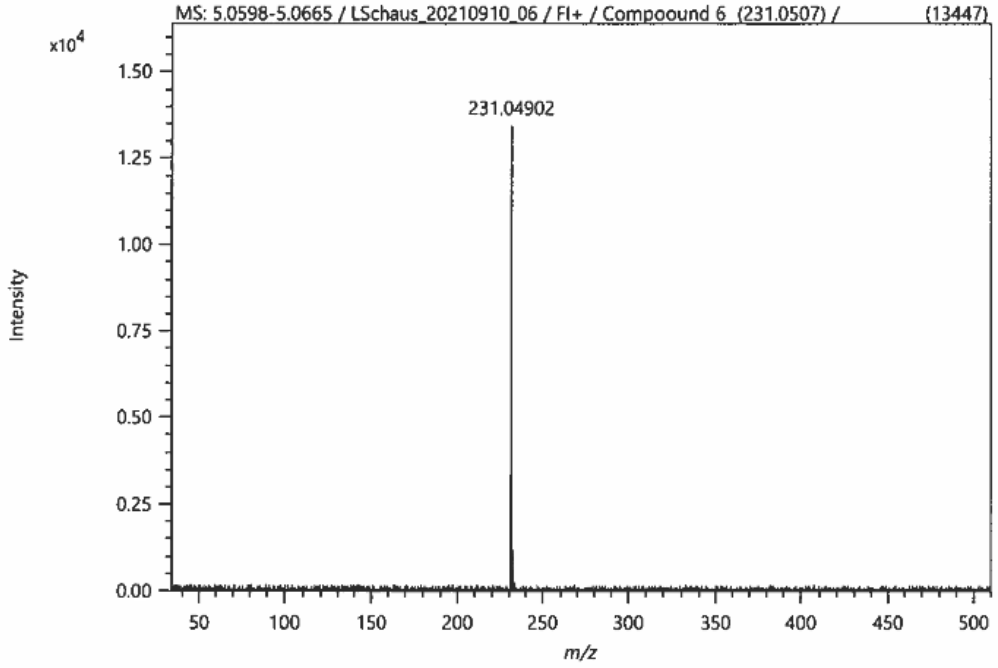
Parameters		Elements Set 1:				
Tolerance:	±5.00 mDa	Symbol	C	H	O	F
Electron:	Odd/Even	Min	0	0	0	3
Charge:	+1	Max	50	100	3	3
DBE:	-1.5 - 20.0					

Results

Mass	Intensity	Formula	Calculated Mass	Mass Difference [mDa]	Mass Difference [ppm]	DBE
216.07384	39630.00	C11 H11 O F3	216.07565	-1.81	-8.38	5.0

1-Nitro-3-(2-(trifluoromethyl)cyclopropyl)benzene (6)

Spectrum



Elemental Composition

Parameters	
Tolerance:	±5.00 mDa
Electron:	Odd/Even
Charge:	+1
DBE:	-1.5 - 20.0

Elements Set 1:

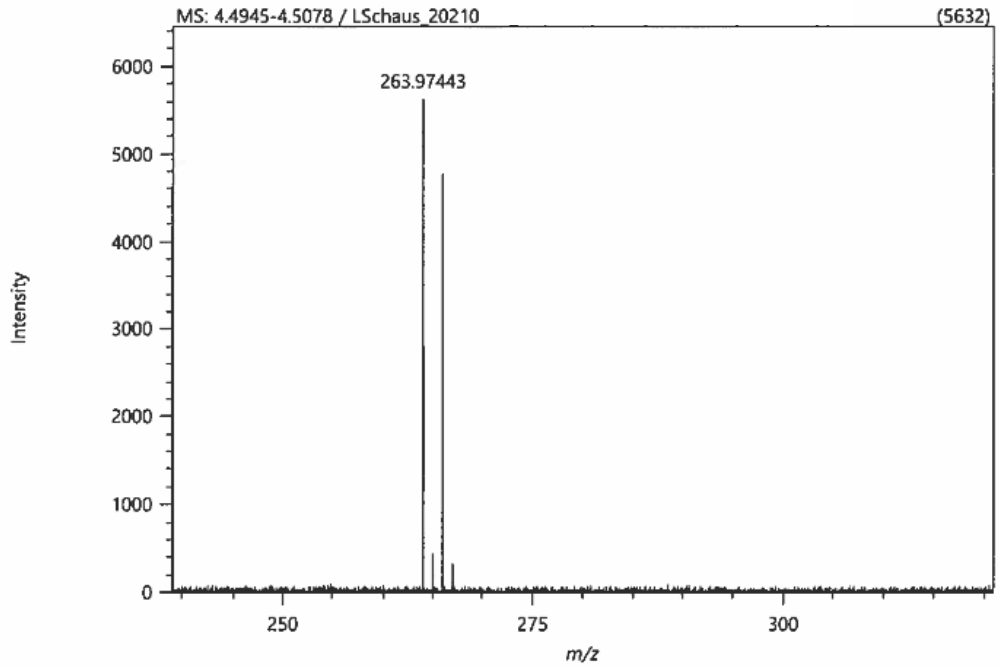
Symbol	C	H	O	F	N
Min	0	0	0	3	0
Max	50	100	3	3	2

Results

Mass	Intensity	Formula	Calculated Mass	Mass Difference [mDa]	Mass Difference [ppm]	DBE
231.04902	13446.57	C10 H8 N O2 F3	231.05016	-1.14	-4.94	6.0

1-Bromo-2-(2-(trifluoromethyl)cyclopropyl)benzene (7)

Spectrum



Elemental Composition

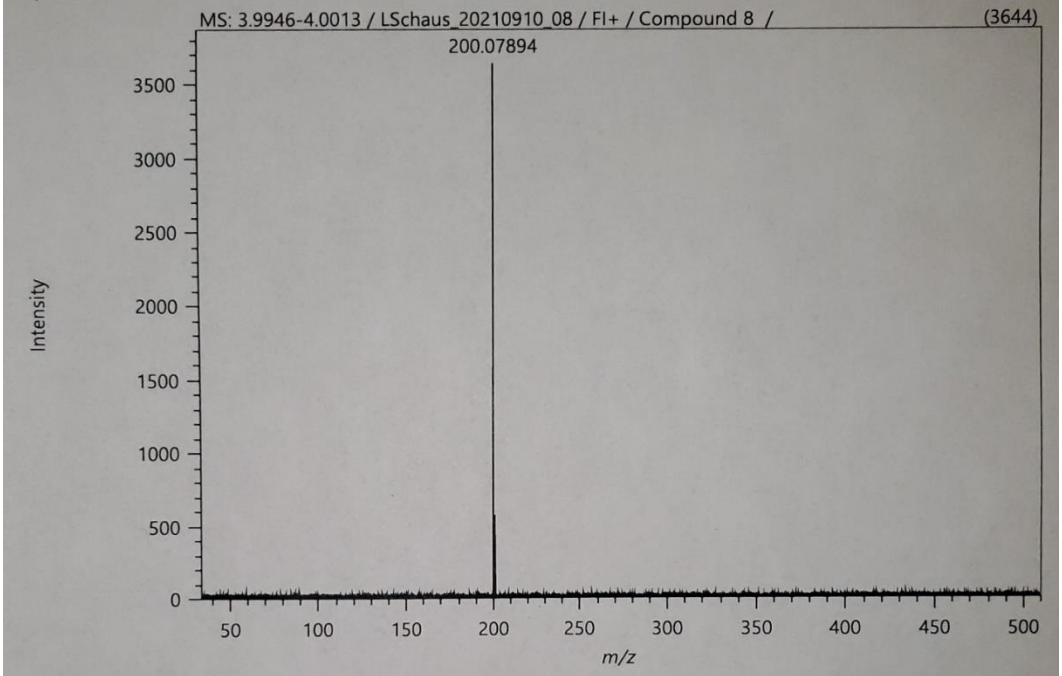
Parameters		Elements Set 1:					
Tolerance:	±5.00 mDa	Symbol	C	H	O	F	Br
Electron:	Odd/Even	Min	0	0	0	3	0
Charge:	+1	Max	50	100	3	3	1
DBE:	-1.5 - 20.0						

Results

Mass	Intensity	Formula	Calculated Mass	Mass Difference [mDa]	Mass Difference [ppm]	DBE
263.97443	5631.82	C10 H8 F3 Br	263.97560	-1.17	-4.43	5.0

(2-(Trifluoromethyl)cyclopropyl)benzene (8)

Spectrum



Elemental Composition

Parameters	
Tolerance:	±5.00 mDa
Electron:	Odd/Even
Charge:	+1
DBE:	-1.5 - 20.0

Elements Set 1:

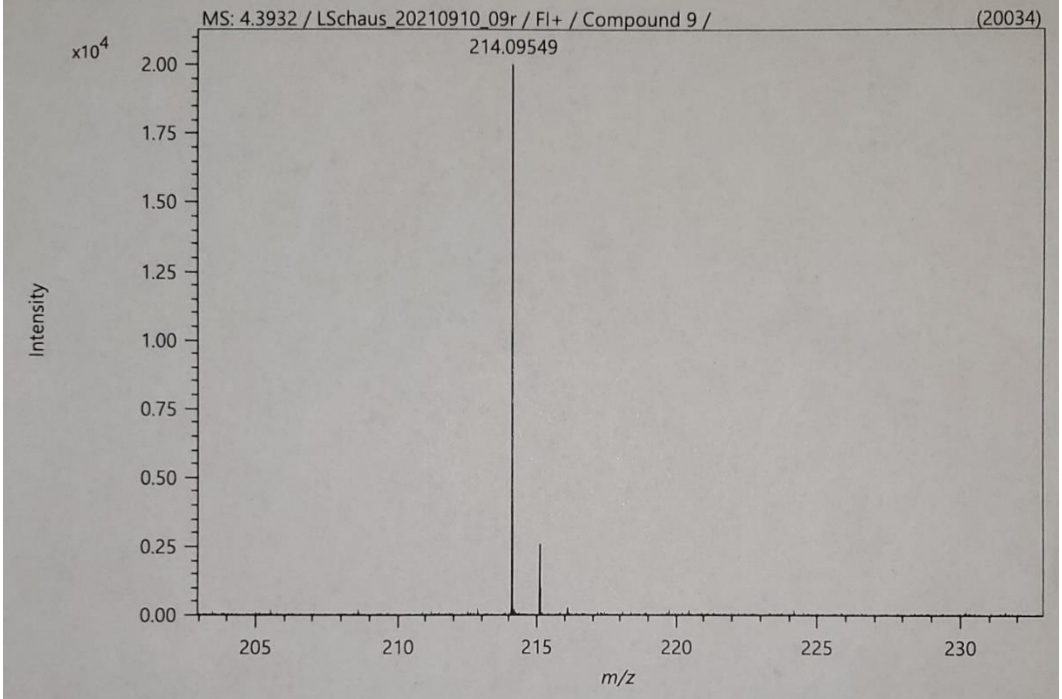
Symbol	C	H	O	F
Min	0	0	0	3
Max	50	100	5	3

Results

Mass	Intensity	Formula	Calculated Mass	Mass Difference [mDa]	Mass Difference [ppm]	DBE
200.07894	3644.12	C11 H11 F3	200.08074	-1.80	-9.00	5.0

((2-(Trifluoromethyl)cyclopropyl)methyl)benzene (9)

Spectrum



Elemental Composition

Parameters	
Tolerance:	±5.00 mDa
Electron:	Odd/Even
Charge:	+1
DBE:	-1.5 - 20.0

Elements Set 1:

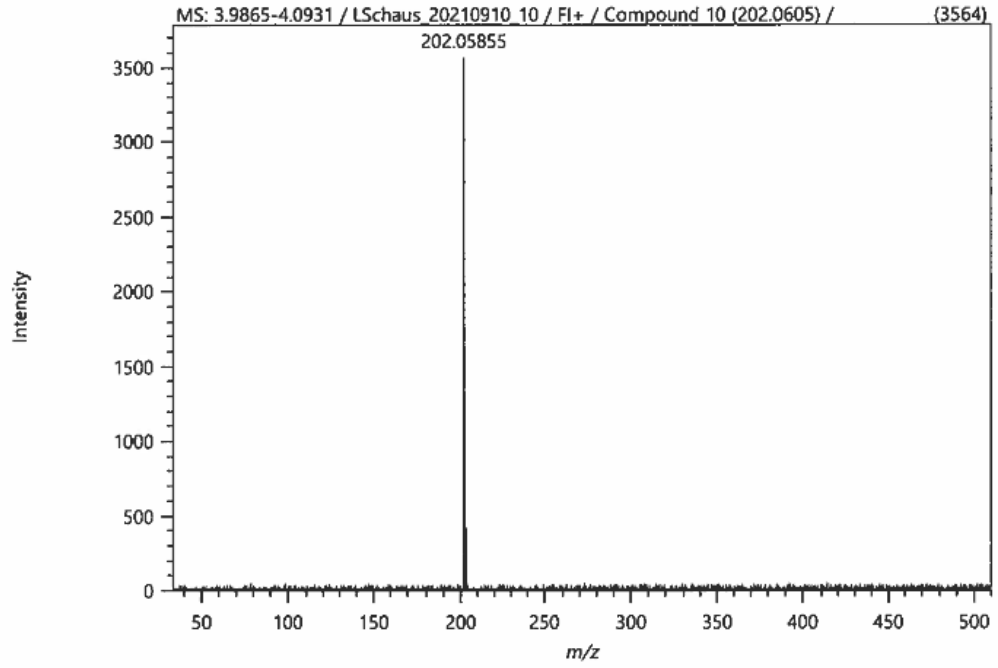
Symbol	C	H	O	F
Min	0	0	0	3
Max	50	100	5	3

Results

Mass	Intensity	Formula	Calculated Mass	Mass Difference [mDa]	Mass Difference [ppm]	DBE
214.09549	20034.00	C12 H13 F3	214.09639	-0.90	-4.18	5.0

(2-(Trifluoromethyl)cyclopropoxy)benzene (10)

Spectrum



Elemental Composition

Parameters	
Tolerance:	±5.00 mDa
Electron:	Odd/Even
Charge:	+1
DBE:	-1.5 - 20.0

Elements Set 1:

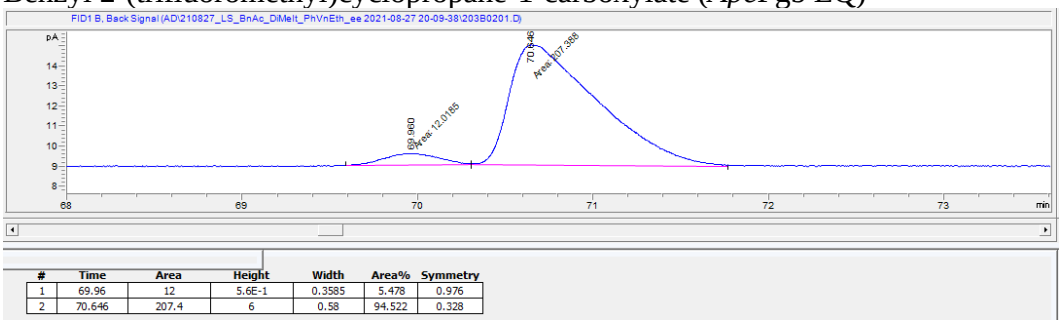
Symbol	C	H	O	F
Min	0	0	0	3
Max	50	100	3	3

Results

Mass	Intensity	Formula	Calculated Mass	Mass Difference [mDa]	Mass Difference [ppm]	DBE
202.05855	3564.11	C10 H9 O F3	202.06000	-1.45	-7.17	5.0

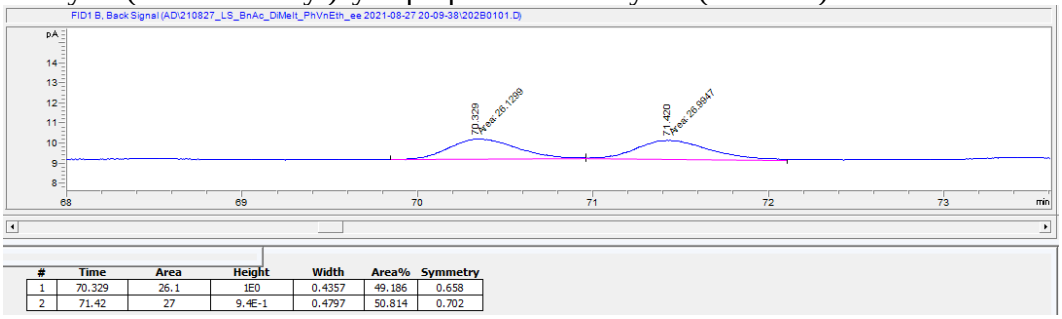
Chiral GC-FID Traces

Benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate (*Ape*Pgb LQ)



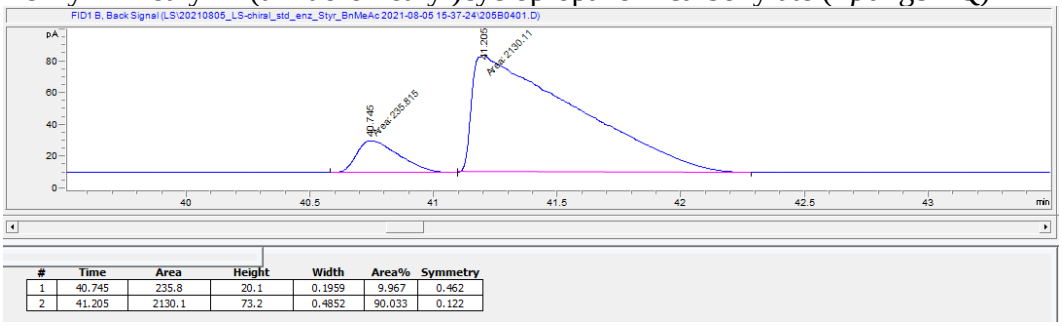
#	Time	Area	Height	Width	Area%	Symmetry
1	69.96	12	5.6E-1	0.3585	5.478	0.976
2	70.646	207.4	6	0.58	94.522	0.328

Benzyl 2-(trifluoromethyl)cyclopropane-1-carboxylate (Racemic)



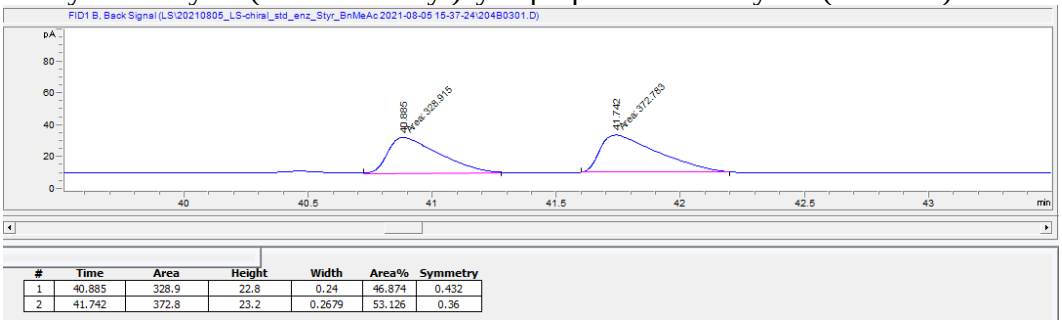
#	Time	Area	Height	Width	Area%	Symmetry
1	70.329	26.1	1E0	0.4357	49.186	0.658
2	71.42	27	9.4E-1	0.4797	50.814	0.702

Benzyl 1-methyl-2-(trifluoromethyl)cyclopropane-1-carboxylate (*Ape*Pgb LQ)



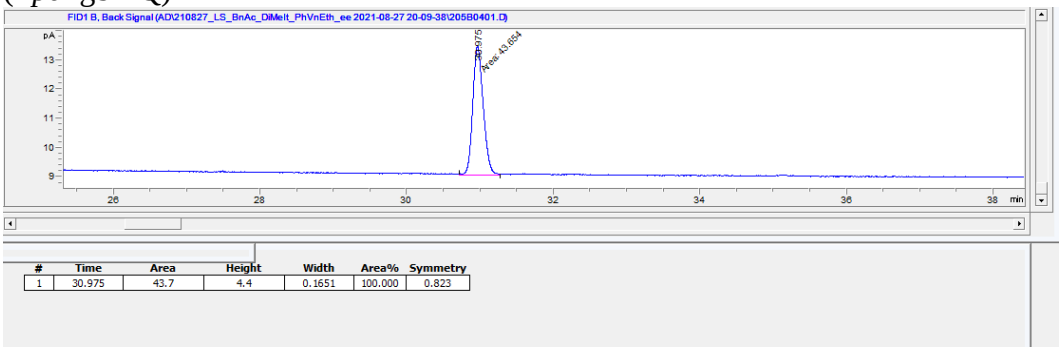
#	Time	Area	Height	Width	Area%	Symmetry
1	40.745	235.8	20.1	0.1959	9.967	0.462
2	41.205	2130.1	73.2	0.4852	90.033	0.122

Benzyl 1-methyl-2-(trifluoromethyl)cyclopropane-1-carboxylate (Racemic)



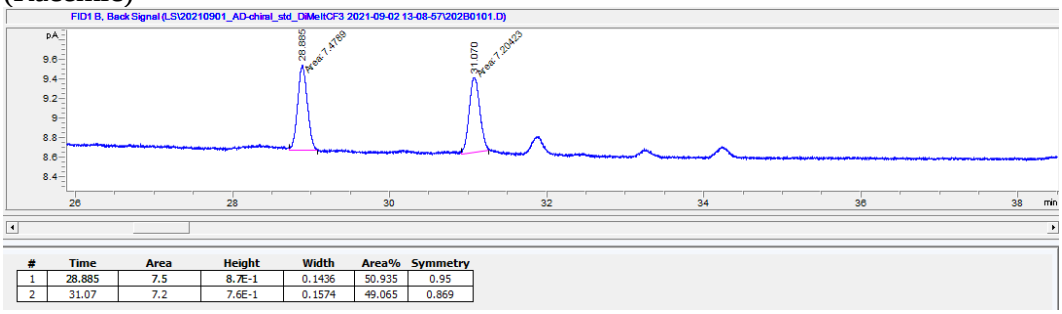
#	Time	Area	Height	Width	Area%	Symmetry
1	40.885	328.9	22.8	0.24	46.874	0.432
2	41.742	372.8	23.2	0.2679	53.126	0.36

Methyl 1-(2-methoxy-2-oxoethyl)-2-(trifluoromethyl)cyclopropane-1-carboxylate (*Ape*Pgb LQ)



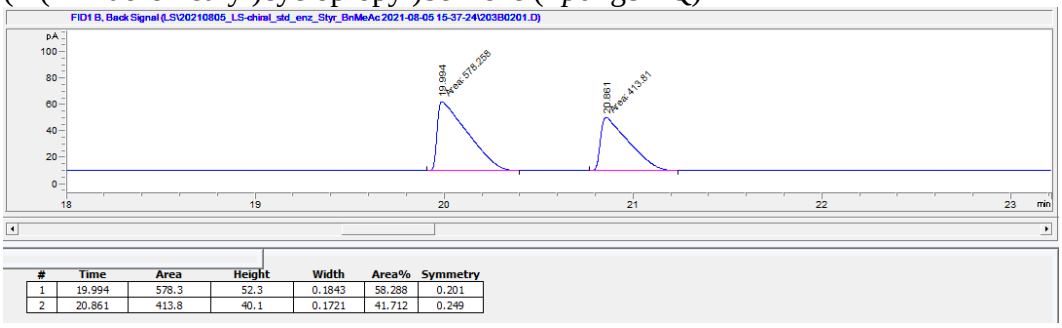
#	Time	Area	Height	Width	Area%	Symmetry
1	30.975	43.7	4.4	0.1651	100.000	0.823

Methyl 1-(2-methoxy-2-oxoethyl)-2-(trifluoromethyl)cyclopropane-1-carboxylate (Racemic)



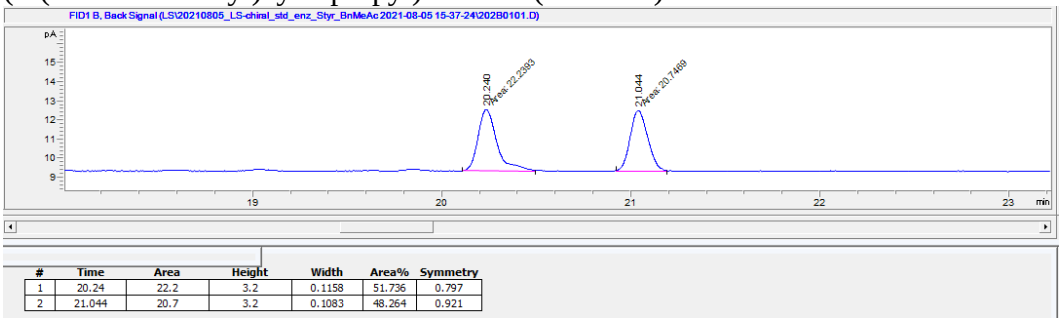
#	Time	Area	Height	Width	Area%	Symmetry
1	28.885	7.5	8.7E-1	0.1436	50.935	0.95
2	31.07	7.2	7.6E-1	0.1574	49.065	0.869

(2-(Trifluoromethyl)cyclopropyl)benzene (*Ape*Pgb LQ)



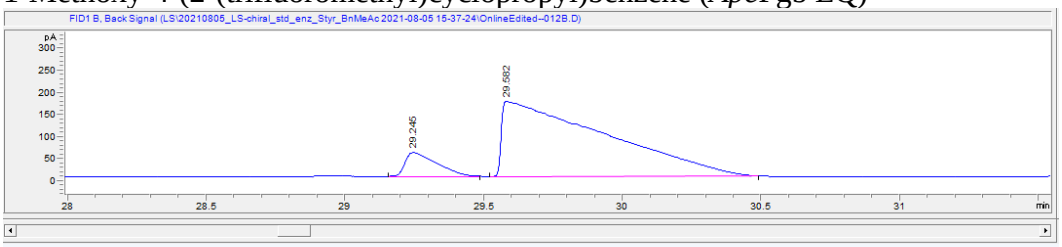
#	Time	Area	Height	Width	Area%	Symmetry
1	19.994	578.3	52.3	0.1843	58.288	0.201
2	20.861	413.8	40.1	0.1721	41.712	0.249

(2-(Trifluoromethyl)cyclopropyl)benzene (Racemic)



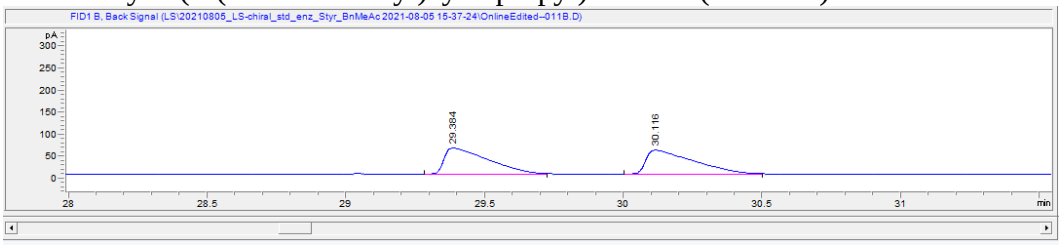
#	Time	Area	Height	Width	Area%	Symmetry
1	20.24	22.2	3.2	0.1158	51.736	0.797
2	21.044	20.7	3.2	0.1083	48.264	0.921

1-Methoxy-4-(2-(trifluoromethyl)cyclopropyl)benzene (*Ape*Pgb LQ)



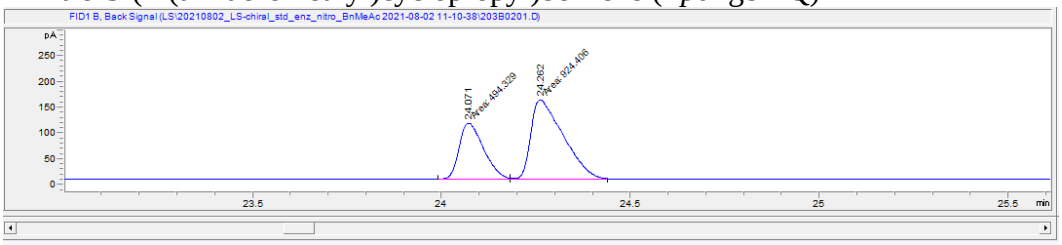
#	Time	Area	Height	Width	Area%	Symmetry
1	29.245	420	53.6	0.1054	8.695	0.323
2	29.582	4410	169.2	0.3131	91.305	4.7E-2

1-Methoxy-4-(2-(trifluoromethyl)cyclopropyl)benzene (Racemic)



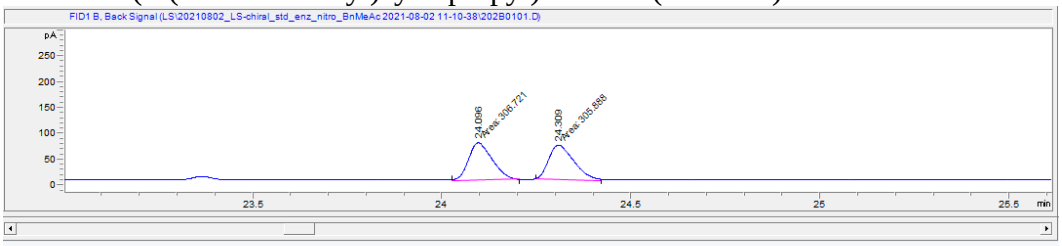
#	Time	Area	Height	Width	Area%	Symmetry
1	29.384	635.9	58.5	0.1489	49.999	0.232
2	30.116	635.9	53.8	0.162	50.001	0.238

1-Nitro-3-(2-(trifluoromethyl)cyclopropyl)benzene (*Ape*Pgb LQ)



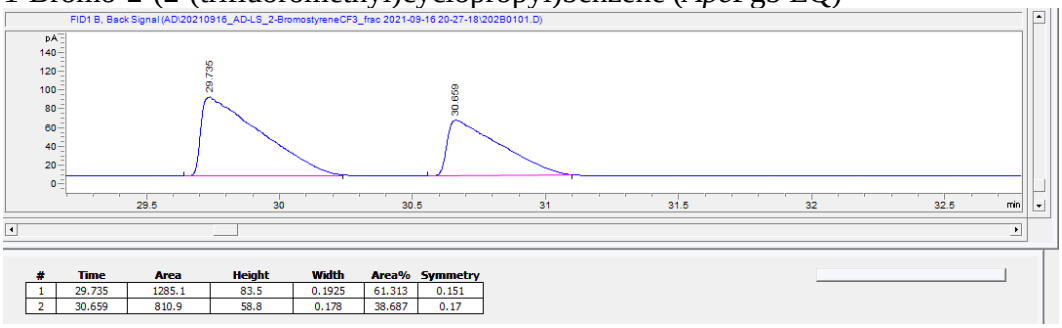
#	Time	Area	Height	Width	Area%	Symmetry
1	24.071	494.3	109.3	0.0754	34.843	0.575
2	24.262	924.4	153.7	0.1003	65.157	0.374

1-Nitro-3-(2-(trifluoromethyl)cyclopropyl)benzene (Racemic)



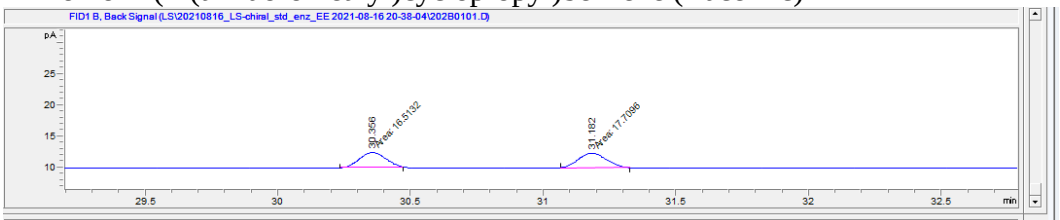
#	Time	Area	Height	Width	Area%	Symmetry
1	24.096	306.7	71.7	0.0713	50.068	0.639
2	24.309	305.9	66.5	0.0766	49.932	0.657

1-Bromo-2-(2-(trifluoromethyl)cyclopropyl)benzene (*Ape*Pgb LQ)



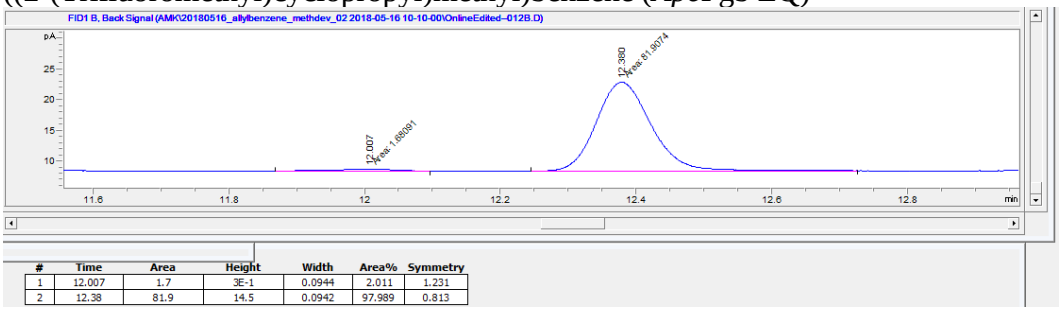
#	Time	Area	Height	Width	Area%	Symmetry
1	29.735	1285.1	83.5	0.1925	61.313	0.151
2	30.659	810.9	58.8	0.178	38.687	0.17

1-Bromo-2-(2-(trifluoromethyl)cyclopropyl)benzene (Racemic)



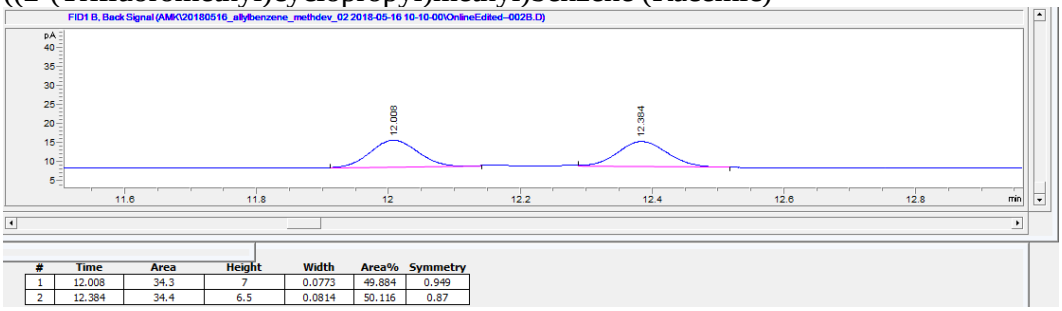
#	Time	Area	Height	Width	Area%	Symmetry
1	24.708	25.6	4.6	0.0925	20.328	0.713
2	26.864	33.2	5.1	0.0999	26.295	0.904
3	27.387	33.1	5	0.1	26.251	0.883
4	30.356	16.5	2.4	0.1155	13.088	1.014
5	31.182	17.7	2.4	0.1237	14.037	0.959

((2-(Trifluoromethyl)cyclopropyl)methyl)benzene (*Ape*Pgb LQ)



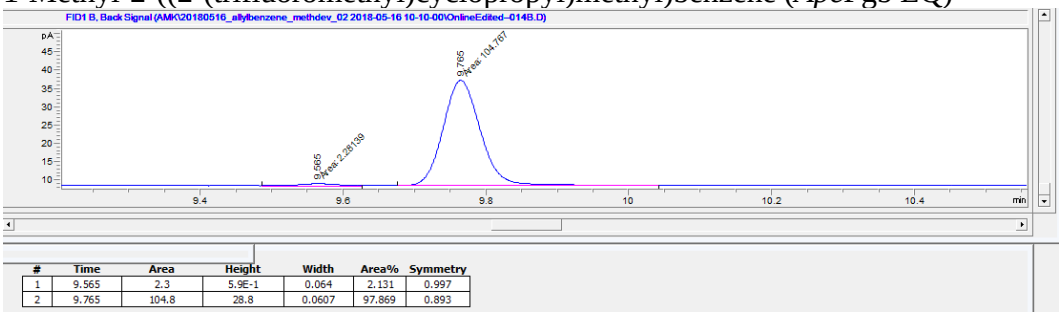
#	Time	Area	Height	Width	Area%	Symmetry
1	12.007	1.7	3E-1	0.0944	2.011	1.231
2	12.38	81.9	14.5	0.0942	97.989	0.813

((2-(Trifluoromethyl)cyclopropyl)methyl)benzene (Racemic)



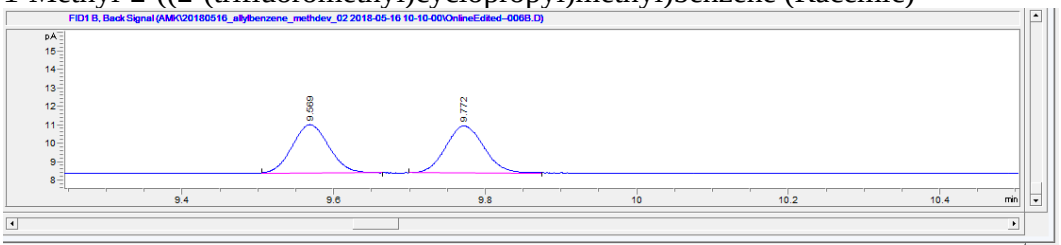
#	Time	Area	Height	Width	Area%	Symmetry
1	12.008	34.3	7	0.0773	49.884	0.949
2	12.384	34.4	6.5	0.0814	50.116	0.87

1-Methyl-2-((2-(trifluoromethyl)cyclopropyl)methyl)benzene (*Ape*Pgb LQ)



#	Time	Area	Height	Width	Area%	Symmetry
1	9.565	2.3	5.9E-1	0.064	2.131	0.997
2	9.765	104.8	28.8	0.0607	97.869	0.893

1-Methyl-2-((2-(trifluoromethyl)cyclopropyl)methyl)benzene (Racemic)



#	Time	Area	Height	Width	Area%	Symmetry
1	9.569	8.8	2.6	0.0532	49.817	0.916
2	9.772	8.9	2.5	0.0557	50.183	0.882