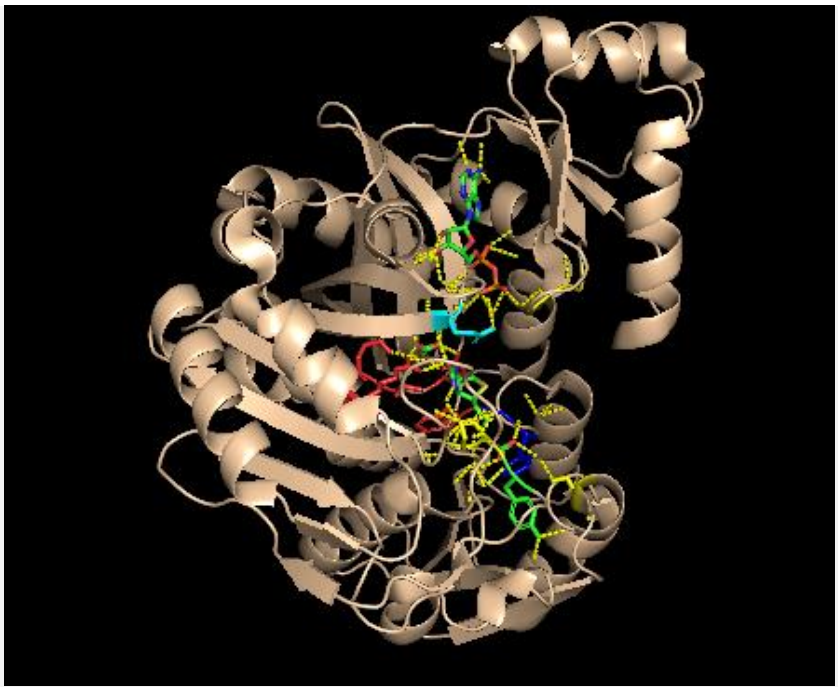
CHALMERS
UNIVERSITY OF TECHNOLOGY



<

Engineering of a Novel Substrate Specificity of Biotin Carboxylase by Machine Learning Assisted Directed Evolution

Written Report

Master's thesis in Biotechnology

LUIS M. LEAL GARZA

DEPARTMENT OF BIOLOGY AND BIOLOGICAL ENGINEERING
SYSTEMS BIOLOGY DIVISION

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021
www.chalmers.se

DEPARTMENT OF BIOLOGY AND BIOLOGICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021
www.chalmers.se

LUIS M. LEAL GARZA

Supervisors:
Martin Engqvist, Department of Biology and Biological Engineering
Examiner:
Martin Engqvist, Department of Biology and Biological Engineering

Department of Biology and Biological Engineering
Division of Systems Biology
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

CHALMERS
UNIVERSITY OF TECHNOLOGY

Abstract

The engineering of more efficient CO_2 fixation mechanisms is an important target that has been addressed by different approaches. Many of these approaches attempt to tackle the photorespiration process. In plants, photorespiration is a consequence of an oxygenation reaction catalysed by RuBisCO (main enzyme in charge of carbon fixation) instead of a carboxylation. This oxygenation occurs due to atmospheric O_2 competing with CO_2 for the active site of RuBisCO. In this project, the first step of a multi domain RuBisCO engineering approach is proposed. This first step consists in a substrate walk engineering on biotin carboxylase (accC) from *Escherichia coli*. Biotin-dependent carboxylases are attractive because they show high specificity for carboxylation and no interaction with atmospheric O_2. Biotin-dependent carboxylases have been the target of many different engineering endeavours, most of which focus in the carboxytransferase domain (accA), instead biotin carboxylase (accC). Additionally, many of these approaches focus on random directed evolution. This project uses a combination of Machine Learning-assisted and rational single-site mutagenesis directed evolution approaches to introduce a new-to-nature 2-imidazolidone carboxylase activity. The work presented here represents one of the first attempts to engineer the biotin carboxylase subunit (accC), one of the few to do machine learning assisted evolution, and the only known to the date to attempt a substrate walk in the biotin carboxylase subunit. The Machine Learning Assisted phase of the project managed to incorporate two simultaneous mutations (G163L and G164H) showing for the first time a 2-imidazolidone carboxylase activity 1.5-fold greater than the background. Results show that improvements in the methodology could have reduce the bias introduced into the machine learning models. After an additional round of site-saturation mutagenesis, one subsequent mutation (G83Y) increased the 2-imidazolidone carboxylase activity to be around 4-fold greater than the background. Unexpected ATPase activity was observed in the final mutant (G83Y, G163L, G164H, F279L) leading to an estimated futile ATP use of 4.4 molecules for each 1 molecule of carboxylated 2-imidazolidone. Future work on this enzyme should monitor both, spectrophotometrically and radiometrically to better control the mutant selection. Additionally, computational models could be used to pre-select the region of interest for mutagenesis.

Keywords: Machine Learning, Directed Evolution, Biotin, Biotin Carboxylase, Desthiobiotin, 2-imidazolidone

Acknowledgements

I would like to thank Martin Engqvist for his guidance during this project. Martin, your trust in my decisions for many of the parts of this project made me nervous several times, but I thank you for encouraging me always to take a step forward. Thank you for the time that you dedicated to supervising my project, I enjoyed our discussions and I think we were a very good team. Working with you was a great learning experience and a very fun time, I would love to collaborate once again with you in the future. Also, thank you for accepting me into your research group for this year, and for letting me participate in this project-idea that you have had for some time.

I would also like to thank Ela who shared with me some of the protocols that she refined for her own thesis. Thanks to all the people at the SysBio wetlab for always helping me find my way around and making my time at SysBio so enjoyable.

From all the people that I would like to thank there are some that have also a special place in my heart. To my parents, without which I would have never even dreamed of pursuing science. Thank you for deciding that making me love knowledge and pursue science would be the best thing for me.

And finally, to the love of my life, who also had to deal with all of my failed experiments and late nights writing this report, thank you Cecilia. You have believed in me since the beginning, and you were the one who pushed me to try to find a place overseas. Without your ambition and your trust in me, I would have never tried to leave my hometown and would have never learned all these marvellous things that I am learning here. This is the first of many academic works that I will dedicate to you.

Table of contents

1. Introduction

This master thesis focuses on using directed evolution to engineer a change in the substrate specificity of a biotin carboxylase (accC). This project is part of a larger RuBisCO engineering project at Engqvist Lab.

The key enzyme in charge of carbon fixation from CO_2 in photosynthesis is Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO). RuBisCo initiates the Calvin Cycle by catalyzing a series of reactions including the carboxylation of ribulose-1,5-bisphosphate (RuBP) to produce two 3-phosphoglycerate (3-PG) molecules. Unfortunately, this reaction is highly competed by an oxygenation reaction when O_2 is used as substrate instead of CO_2. The oxygenation reaction results in the production of one 3-PG and a 2-phosphoglycolate (2-PG) molecule. The production of 2-PG represents a waste of carbon and energy, which results in a decrease of carbon fixation efficiency around 50% [1].

According to molecular dynamics simulations, around 40% more CO_2 than O_2 molecules bind to RuBisCO when both gases are in a 1:1 molar ratio [2]. However, our O_2 rich atmosphere complicates carboxylation by favoring O_2: CO2 molar ratio to over 500:1.

For this reason, the understanding of the carboxylation activity of RuBisCO and its engineering have become important targets that could improve plant productivity.

1.1 RuBisCO Engineering

Different approaches have been attempted in the endeavor of improving RuBisCO's performance. Some first strategies included manipulating the amount of expression of RuBisCO and regulatory genes such as RuBisCO activase. However, more contemporary strategies rely on understanding and engineering the carboxylation process [3].

Enzymes with evolutionary relationship to RuBisCO, such as RuBisCO-like proteins (RLPs), have been studied to identify the residues in charge of carboxylation [4]. RLPs share many structural features with RuBisCO, as well as catalytic abilities, except for carboxylation. This has allowed scientists to determine that some residues close to the C-terminal region of RuBisCO are in charge of carboxylation [5]. However, in the attempts of engineering RuBisCO, rational engineering has underperformed in comparison to random strategies [6]. This is mainly because of our poor understanding of the relationships between structure and function of RuBisCO.

Different protein engineering and directed evolution strategies have also been evaluated to improve RuBisCO's catalytic performance. Random engineering by directed evolution has been able to find

the importance of non-obvious residues to CO_2/ O_2 specificity [6]. Also, random mutagenesis has been used to produce O_2 resisting variants, as well as variants with an increased affinity towards CO_2 [7], [8].

In plants, the lack of efficiency of RuBisCO is compensated by strategies such as CO_2 concentrating mechanisms and a highly abundant expression of the enzyme [1]. Similar to what happens naturally in plants, most of the directed evolution strategies have accidentally resulted in RuBisCO variants that have an increased expression or solubility, instead enzymes with increased activity [7]. This is often an issue, as many screening techniques do not account for protein expression levels. Therefore, one variant that was expressed more successfully could be selected over another one with the same specific activity, as the screening technique would show that the first one was more active. A different approach in directed evolution of RuBisCO's carboxylase activity linked carboxylation of RuBP to the growth of a RuBisCO-depending *Escherichia coli* (RDE). After screening over 15,000 mutants, this random mutagenesis approach showed an increase of specific activity of 85%, and 45% increase of specificity to CO_2 [9]. Nevertheless, this system can yield false positives in screening for activity-enhancing mutations, as it can still introduce solubility-enhancing mutations.

Instead of focusing on engineering RuBisCO variants that are more specific towards carboxylation, or variants that are insensitive to O_2, some strategies are trying to solve the energy and carbon waste that represents the oxygenation reaction. As mentioned previously, the 3-PG is the desired product from carboxylation and 2-PG is the unwanted by-product of oxygenation. When oxygenation occurs, an ATP-intensive pathway turns two 2-PG molecules back into 3-PG resulting also in a loss of carbon and nitrogen. New strategies have engineered new pathways such as the tartronyl-CoA pathway (TaCo pathway) as a ATP-consuming way to convert 2-PG into 3-PG without carbon and nitrogen waste [10].

This existing background, as well as the continuously increasing developments in gene editing and screening techniques, set ground to a feasible Directed Evolution engineering of the carboxylation reaction[11].

A different approach towards the creation of an engineered RuBisCO able to carboxylate RuBP with a higher CO_2 specificity is currently being explored by the Engqvist Lab. The approach involves using a carboxylation-deficient RuBisCO mutant. This carboxylation-deficient RuBisCO mutant was obtained by mutating Lysine 334. This mutant is able to use RuBP to form an enolate, which is an intermediate before the carboxylation/oxygenation, but unable to carboxylate it [3].

Afterwards, instead of relying in the promiscuity of RuBisCO for the carboxylation step, we propose the use of a biotin-dependent carboxylase to provide a carboxylated intermediate. The carboxylated intermediate produced by the biotin-dependent carboxylase would then act as a carboxylate-donor to the enolate intermediate produced by the RuBisCO mutant. Ideally, the substrate used by the biotin-dependent carboxylase should be small enough to diffuse into the active site of the RuBisCO mutant. In nature, the abundant biotin-dependent carboxylases employ a similar strategy by using biotin as intermediate for many carboxylation processes.

1.2 Biotin Carboxylases

Biotin-dependent carboxylases are a group of enzymes that carboxylate a broad range of substrates using biotin as a carboxylation intermediate. All of these enzymes contain three subunits (or domains of the same polypeptide, depending on the organism) that allow the use of a protein-bound biotin as an intermediate for carboxylation: Biotin carboxylase (BC), carboxyltransferase (CT), and biotin-carboxyl carrier protein (BCCP), as shown in Figure 1.

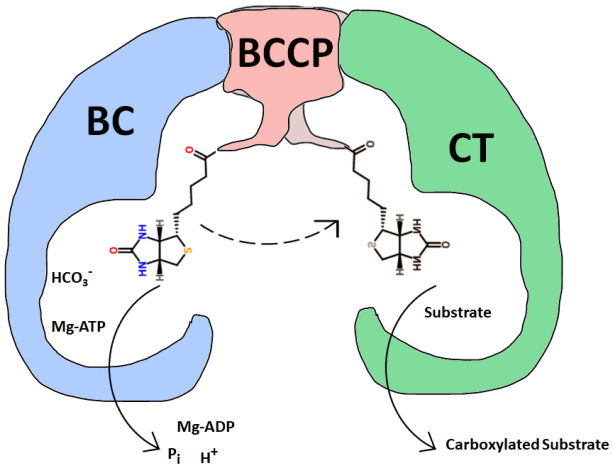


Figure 1. Structure of a typical biotin-dependent carboxylase. Biotin is covalently bound to the biotin-carboxyl-carrier protein subunit (BCCP, red/middle). BCCP is able to translocate its biotin into the active site of either the biotin carboxylase subunit (BC, blue/left) to get biotin carboxylated or the carboxyltransferase subunit (CT, green/right) to transfer its carboxylation to a final carboxyl acceptor.

The carboxylation process in biotin-dependent carboxylases starts at the BC subunit, where the MgATP-dependent carboxylation of biotin occurs, using a HCO_3^- ion as donor for the carboxylation. The carboxylated biotin, which is covalently bound to the BCCP subunit is translocated to the CT active site, where it donates the carboxyl group to a final substrate. There are multiple CT subunit variants of the biotin-dependent carboxylases, each able to catalyse the reaction into different final substrates such as acetyl-CoA, propionyl-CoA, pyruvate, and urea. However, the BC subunit has a high specificity only for BCCP bound biotin and has no reported interaction with atmospheric O_2[12].

1.2.1 Biotin Analogs

Typically, when engineering an enzyme from the biotin-dependent carboxylase family, the main target is the CT subunit. The CT subunit has been engineered via directed evolution in the past to increase catalytic activity over a known substrate[13]. As the CT subunit determines the whole enzyme's specificity, it is also an interesting target when engineering a change in substrate specificity (substrate walk). Recently, the CT subunit of a propionyl-CoA carboxylase has been engineered by a mixed strategy of site directed and random directed evolution to be able to catalyze the carboxylation of glycoyl-CoA instead [10].

The BC subunit however, has not been engineered for a change in substrate specificity, unlike the CT subunit. Engineering the BC subunit to use either free biotin or smaller biotin analogs would allow its use in fusion protein designs, as it will not require the formation of the typical three subunit complex. In this context, it is important to understand the activity of BC towards biotin and biotin analogs.

Previous studies have assessed the activity of BC towards BCCP-bound biotin, free biotin, and biotin bound to smaller peptides than BCCP (peptide-biotin). These have shown that the reactivity for free biotin has 8000-fold less catalytic efficiency (V_{max}/K_m) than BCCP bound biotin [14]. Using peptide-biotin, which includes just 5 residues from BCCP, it was found just a 20% increase in catalytic efficiency in comparison to free biotin [15].

The specificity of BC has been measured by testing its activity towards biotin analogs of different sizes. While BC seems to keep some activity with some larger biotin analogs like biocytin (15% in comparison to biotin), its carboxylating activity has been reported to be almost completely absent for smaller analogs such as desthiobiotin (0.2% in comparison to biotin) and 2-imidazolidone (0% in comparison to biotin)[16]. Thus, a deeper understanding of BC catalysis is needed if the engineering goal is to increase the activity to either free biotin or any of its smaller analogs of interest for this project (Figure 2).

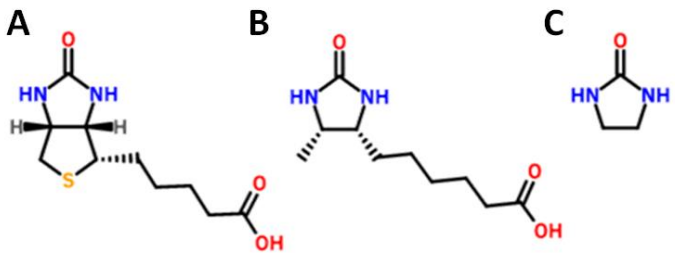


Figure 2. Structure of biotin (A), dethiobiotin (B), and 2-imidazolidone (C).

1.2.2 Biotin Carboxylase Structure and Engineering

One of the biotin-dependent carboxylases that has been studied the most and for which there are many crystal structures available is acetyl-CoA carboxylase from *E. coli*. This enzyme has the three typical subunits of a biotin-dependent carboxylase BC, BCCP, and CT encoded by accC, accB, and accA respectively. In vivo these subunits form a quaternary structure that contains two copies of each polypeptide. Crystal structures for isolated BC confirm that it forms a homodimer [17]. Structures of co-purified BC and BCCP also show that the BC homodimer allocates two BCCP subunits [18]. These crystal structures, together with variants generated by point mutations have helped to understand more about which residues are more relevant for the activity of the BC subunit, as shown in **Figure 3**. Identifying these residues is important to perform engineering that will not hinder the enzyme's activity.

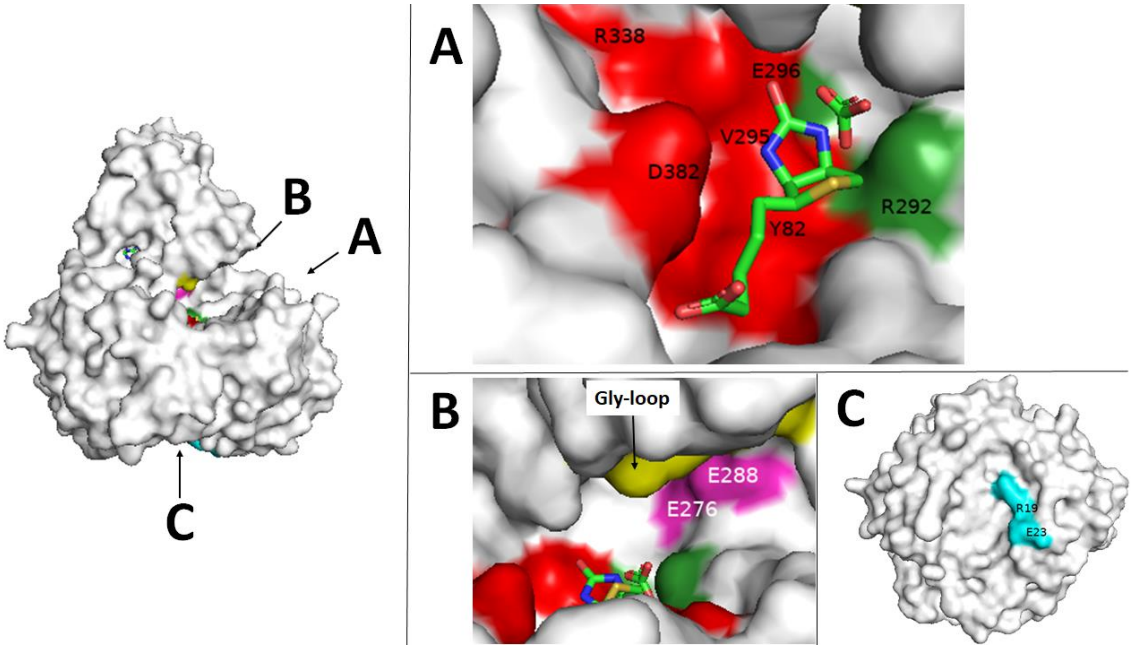


Figure 3. Crystal structure and important residues in biotin carboxylase (BC) subunit of acetyl-CoA carboxylase (accC) from *E.coli*. R19 and E23 in cyan, R292 and E296 in green, E276 and E288 in magenta, Y82, V295, D382, and R388 in red, and the Gly-rich loop (residues 162 to 166) in yellow.

Some residues provide a binding site for the substrates. R292 and E296 hold the HCO_3^- ion, E276 and E288 form a complex with Mg-ATP, and Y82 and V295 stabilize both, biotin and the HCO_3^- ion. Carboxylate atoms from the tail part of biotin also interact with D382. Residues like R338 have several functions, as it shows interaction with biotin and HCO_3^-, and catalytic activity together with E296. Residues such as R19 and E23 play a structural role, as it has been found that mutating them results in the formation of monomers instead of homodimers with just a 3-fold loss of activity [17]. Some other residue structures like the glycine-rich loop (residues 162-166) do not have a clear function but seem to rearrange when the substrates fit in place [19].

1.2.3 Activity Assays in Biotin Carboxylases

The reaction catalysed by the BC subunit of every biotin-dependent carboxylase (as shown in Figure 1) has different substrates and products which can be used to monitor the reaction. The formation of the main product of the reaction, carboxybiotin, is most commonly monitored using a radiometric assay. To perform this assay, radioactively labeled $NaH^{14}CO_3$ is used so a [^{14}C]-carboxybiotin can be measured when terminating the reaction [16], [20]–[22]. Another radiometric assay can be used to measure the consumption of ATP indirectly by using radioactively labeled [γ-^{32}P]ATP and measuring the release of $^{32}P_i$ after precipitation [22], [23]. However, both assays are endpoint and kinetic data is not easily obtainable from them.

To obtain kinetic data from a BC, the most commonly used method is a spectrophotometric method that monitors the formation of ADP by using a coupled reaction system with pyruvate kinase (PK) and lactate dehydrogenase (LDH) as seen in Figure 4. This coupled reaction system measures the formation of ADP indirectly by monitoring the oxidation of NADH at 340nm [15], [20], [21], [23]. The advantages of this method is that is easily adaptable to high-throughput settings, which are frequently used in directed evolution experiments [17].

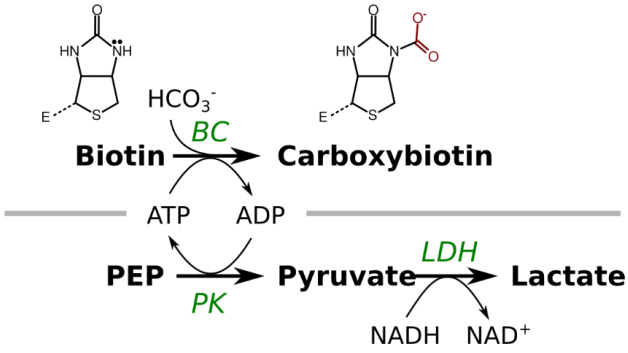


Figure 4. Coupled enzymatic reactions for spectrophotometric monitoring of NADH oxidation. The main reaction (Top) produces carboxybiotin and ADP. ADP enables the conversion of phospho(enol)pyruvate (PEP) by PK. Finally, lactate is produced from pyruvate by LDH, using a NADH molecule in the process. The decrease in NADH is monitored a 340nm.

1.3 Directed Evolution

Directed evolution of enzymes is a process which consists in generating a library of variants of an enzyme and later high throughput screening it for a desired property [24]. Directed evolution of enzymes can be used to improve catalytic activity for an existing substrate or to engineer for different substrate specificities [25]. Directed evolution works as an iterative process that uses the best variants from each round as parents for the next round. Typically, even a minimal activity for an enzyme towards a substrate is considered enough to start a directed evolution project [26]. Directed Evolution can be done in many different ways depending on the method to generate the variant library and the high throughput screening technique used.

1.3.1 Mutagenesis in Directed Evolution

There are three main methods to generate the variant library in a Directed Evolution experiment: error-prone PCR, site-saturation mutagenesis, and DNA shuffling. While error-prone PCR relies in complete randomness, DNA shuffling relies in recombination processes, both methods emulating the natural evolution of enzymes. However, directed evolution via site-saturation mutagenesis (SSM) implies a deeper previous knowledge of the enzyme, thus considered semi-rational evolution [27].

SSM is a technique that can be done via PCR in which a previously selected specific site (amino acid residue) is substituted by all the possible 20 amino acids in one single experiment. To be able to do this, PCR must be performed as usual, except for the primer design. Primers used for SSM should be complementary to the template DNA but must include a degenerate codon for the selected residue [28]. In its simplest form, the degenerate codon used for SSM primer design can be NNN, which codes for the 64 possible codons. However, the disproportionally redundant nature of the standard genetic code means not only that stop codons can be introduced and some amino acids will be overrepresented, but also that 64 redundant variants will be screened, instead of just 20. A solution for this situation is to use a different degenerate codon, such as NNK. The use of NNK shrinks the sample space to just 32 different codons, which simplify the screening efforts. Nevertheless, this strategy still disproportionally codes for the 20 amino acids and allows for a stop codon to be introduced. Strategies to reduce overrepresentation of some amino acids and avoid stop codons completely exist, but often require the use of more than one degenerate codon. A strategy called “22-c trick” uses two different degenerate codons and an additional non-degenerate codon (NDT, VHG and TGG) to achieve a scenario where the overrepresentation is minimal, and no stop codons are introduced [29]. While some of these strategies require more than one set of degenerate primers to be designed, the reduction in screening efforts justifies their use.

The size of the sample space of the variants created by SSM is not only influenced by the degenerate codon strategy used to create the library, but also by the number of sites mutated at once. Cassette mutagenesis, which is SSM applied over several neighboring residues, often results in large sample spaces. To illustrate this, a single residue SSM requires the screening of 192 colonies to make sure all possible variants are screened with a 95% confidence using a NNN degenerate codon. In contrast, a 4-residue cassette mutagenesis requires $5x10^7$ colonies to be screened to achieve the same feat under the same conditions [29].

1.3.2 Machine Learning in Directed Evolution

The use of computational tools to aid directed evolution is nothing new. Semi-rational engineering can be helped by the use of structure homology modelling, and important residues can be identified by substrate docking [27]. However, a recent trend has started to use the computational power of

Machine Learning (ML) to reduce the need of screening the complete sample space of a directed evolution experiment.

In small directed evolution experiments where of 3 or 4 amino acids are targeted each round, a machine learning assisted approach has shown positive results in evolving towards a stereo-selective variant of an enzyme that typically yields a racemic product [30]. By testing on a diverse set of variants and obtaining their sequence-function information, machine learning can be used to generate models and *in silico* screen the complete sample space. The use and training of a diverse set of models such as random forests, linear, and kernel modelling methods is suggested, with as many as 15 models being trained to use the information of the top performing models. It is important not to generalize and evaluate with multiple models for every different protein when using this approach, as every protein has different properties [30]. More complex methods can be used to engineer larger amounts of amino acids. Using deep learning to model local and global sequence landscapes even allow for the screening of as few as 24 variants to enable *in silico* screen as much as ten million sequences [31].

1.4 Objective

This project proposes the use of several of these strategies to assist directed evolution. The final goal of the directed evolution was to engineer an BC subunit that either shows greater activity for free biotin or, ideally, a BC subunit able to use a smaller biotin analog (Figure 2). Divided in two different phases, this project consists in a machine learning assisted phase (MLA) and active site engineering phase (ASE). An outline of the methodology of this project is shown in Figure 5.

The MLA phase consisted of a machine learning assisted directed evolution of a 4-residue region from the Gly-rich loop of the BC subunit. This target region is made up by the residues G162, G163, G164, and G165. The main objective from the MLA phase was to find the best set of variants that will bulk up the entrance to the active site without damaging the ability of ATP to bind, as it has previously been shown that G165 can be mutated to hamper this interaction [19]. The MLA phase was performed in two mutation rounds.

The first round's objective was to create a broad library of mutants (Training library). The strategy followed to generate this library of mutants uses four consecutive degenerate NNK codons as a cassette for site saturation mutagenesis and is explained with more detail in the next section. From this library, two plates of 96-well plates (176 variants) are sequenced and screened for activity towards free biotin to create function-sequence data. Afterwards, the function-sequence information from the first round was used to train the ML models and the models are used to *in silico* screen for the rest of the 20^4 variants. The results from the *in silico* screening were then used to start the second round of mutations, keeping only the amino acids that will most likely retain function.

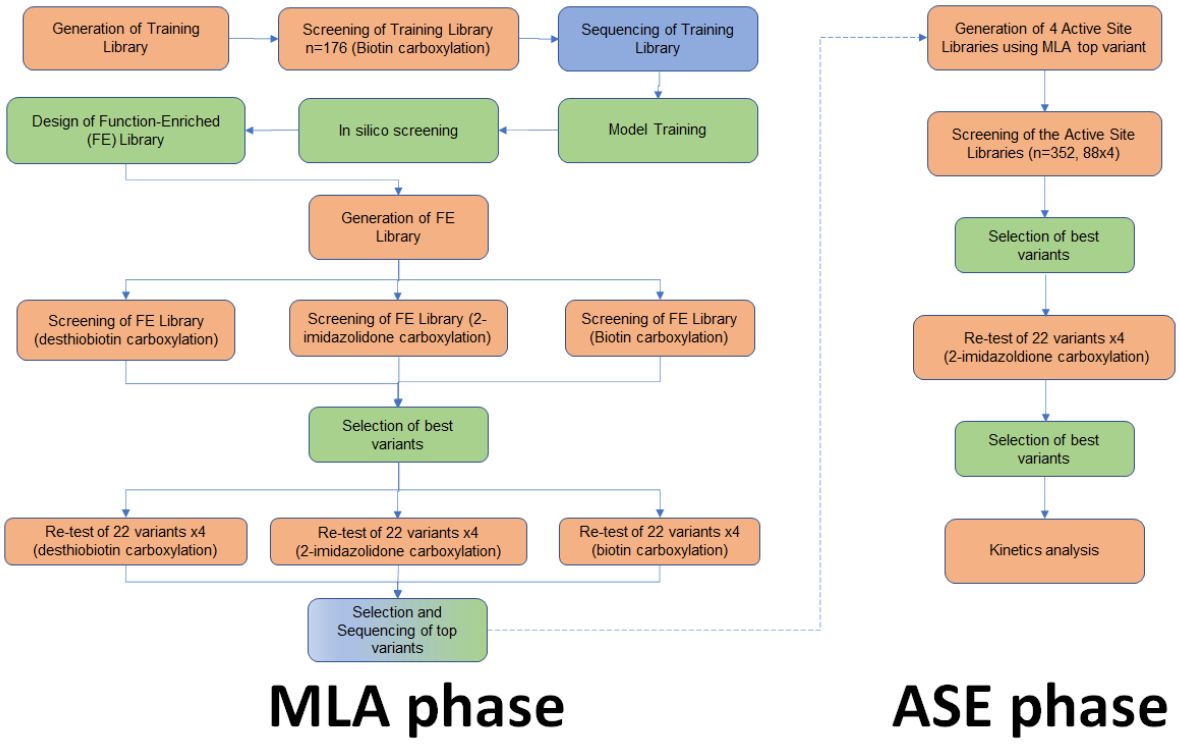


Figure 5. Complete outline of the project. Wet-lab, dry-lab, and external work is highlighted in red, green, and blue, respectively.

The second round used the output information from the machine learning models and created a second function-enriched library. This new library was screened for activity towards biotin, desthiobiotin and 2-imidazolidone, with the expectation that now some activity towards the smaller analogs would be detected. These substrates are structurally similar but smaller than biotin, and 2-imidazolidone has been used to model biotin reactions previously [32]. The mutant with greatest activity for free biotin and/or ideally to the smaller biotin analogs (desthiobiotin and 2-imidazolidone), from this MLA phase served as the starting point of the ASE phase.

The ASE phase's objective was to create libraries of 4 non-consecutive residues located in the active site of the enzyme. The expectation is that, because these residues surround the biotin tail, an enhanced activity for the 2-imidazolidone would be obtained. The 4 residues of interest for this round were selected by avoiding the known important and catalytic residues and using structural analysis to explore the best candidates that had a higher interaction with the biotin tail. The target residues were S56, G83, F84, D382 and its interaction with biotin as a substrate in the active site is shown in Figure 6.

The strategy for ASE library generation, described in more detail in the next section, uses the "22-c trick" strategy. This strategy implies using 3 degenerate codons to code for all amino acids with minimal redundancy. After the generation of the 4 libraries, 88 variants from each library were

screened for its activity on 2-imidazolidone. The best variants were sequenced and assayed to obtain kinetic information.

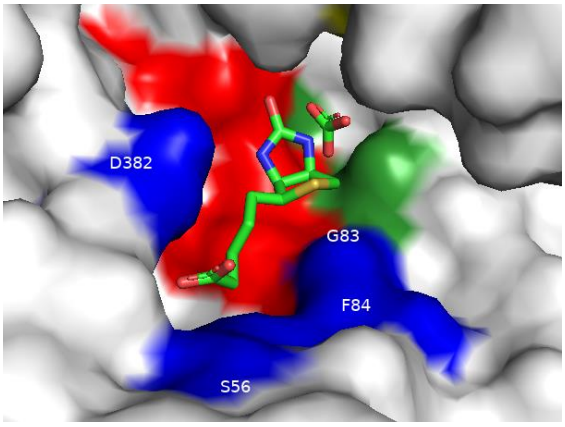


Figure 6. Active site of BC subunit. The target residues surrounding the biotin tail for directed evolution are shown in blue.

The results of this project provide the first known directed evolution engineering of the biotin carboxylase subunit. The promiscuity for other substrates has been measured previously, and site-directed mutagenesis has been used to explore the structure-function relationships in the BC subunit. However, to the best of our knowledge, there has not been any previous work attempting to change BC subunit substrate specificity outside of this study.

If a substrate walk is successfully engineered, the resulting variant can undergo further engineering rounds to increase its catalytic activity and eventually be used for the original purpose of working as an intermediate of carboxylation in engineered enzymes.

2. Materials and Methods

As described in Figure 5, the different phases of the project share several experimental procedures. In this section, the different experimental procedures are described as they were carried out. The initial parent sequence on which directed evolution was performed was an already mutated variant of accC provided by the Engqvist Lab. This variant contains already a F279L mutation, which has shown an increased activity towards free biotin in comparison to accC.

2.1 Generation of Training Library

The strategy chosen for the simultaneous 4-residue SSM was to design a mutation cassette with 4 consecutive NNK degenerate codons. This cassette was contained within an oligonucleotide with a 20 bases region for assembly and an 18 bases region to serve as primer for the PCR as seen in Figure 7. All PCR reactions in this work were performed with the Phusion High-Fidelity DNA Polymerase (Thermo Scientific) using the recommended running conditions unless stated otherwise. The PCR reactions were all performed using the recommended HF buffer 10X, 200µM of each dNTP, 0.5µM of each primer, 5% DMSO and 20ng of template DNA. The primers used for this PCR are under the rows labeled as “MLA Training Library” shown in Table 1, together with the rest of the primers used throughout this project. The specific annealing temperature for this PCR was 57°C and 30s for extension were used.

This PCR reaction results in two fragments of the accC gene, each containing a minimal ~20bp part of the plasmid backbone. The plasmid backbone was obtained by an enzymatic digestion of the original plasmid (as shown in Figure 7) with NdeI. The resulting PCR fragments and the pETME31 backbone obtained from NdeI digestion were assembled together via Gibson Assembly (isothermal assembly) for 1h at 50°C [33], this was possible because the fragments and plasmid have overlapping regions. The assembled library was then transformed into OneShot *E. coli* BL21 (DE3) chemically competent cells (Invitrogen) using the recommended transformation protocol and plated into LB media plates containing 100µg/mL carbenicillin as antibiotic.

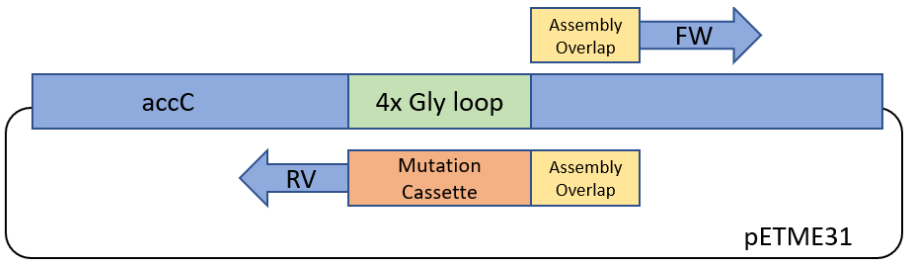


Figure 7. Primer and Mutation Cassette design for library generation. The difference between Training Library and Function-Enhanced Library lies in the degenerate codons used for the Mutation Cassette.

Table 1. Primers used during the project. Primers used throughout the three rounds of mutagenesis. The primers used for the ASE round were further divided by the residue that they targeted.

Mutation Round	ID	5' to 3' DNA sequence
MLA Training Library	Gib_FW2	GGTCGCGGTATGCGCGTAGTGCGCGGCGACGCTGAACTGG
	Gib_mut_RV	ACTACGCGCATACCGCGACCMNNMNNMNNMNNGGAGGCTTTGATAATCACCGG
MLA Function Enriched	Gib_FW2	GGTCGCGGTATGCGCGTAGTGCGCGGCGACGCTGAACTGG
	Gib_0000_RV	ACTACGCGCATACCGCGACCTVCGKKTSCGTDGGAGGCTTTGATAATCACCGG
	Gib_1000_RV	ACTACGCGCATACCGCGACCTVCGKKTSCACCGGAGGCTTTGATAATCACCGG
	Gib_0100_RV	ACTACGCGCATACCGCGACCTVCGKKGWRGTDGGAGGCTTTGATAATCACCGG
	Gib_1100_RV	ACTACGCGCATACCGCGACCTVCGKKGWRACCGGAGGCTTTGATAATCACCGG
	Gib_0010_RV	ACTACGCGCATACCGCGACCTVCTMCTSCGTDGGAGGCTTTGATAATCACCGG
	Gib_1010_RV	ACTACGCGCATACCGCGACCTVCTMCTSCACCGGAGGCTTTGATAATCACCGG
	Gib_0110_RV	ACTACGCGCATACCGCGACCTVCTMCGWRGTDGGAGGCTTTGATAATCACCGG
	Gib_1110_RV	ACTACGCGCATACCGCGACCTVCTMCGWRACCGGAGGCTTTGATAATCACCGG
	Gib_0001_RV	ACTACGCGCATACCGCGACCGWKGKKTSCGTDGGAGGCTTTGATAATCACCGG
	Gib_1001_RV	ACTACGCGCATACCGCGACCGWKGKKTSCACCGGAGGCTTTGATAATCACCGG
	Gib_0101_RV	ACTACGCGCATACCGCGACCGWKGKKGWRGTDGGAGGCTTTGATAATCACCGG
	Gib_1101_RV	ACTACGCGCATACCGCGACCGWKGKKGWRACCGGAGGCTTTGATAATCACCGG
	Gib_0011_RV	ACTACGCGCATACCGCGACCGWKTMCTSCGTDGGAGGCTTTGATAATCACCGG
	Gib_1011_RV	ACTACGCGCATACCGCGACCGWKTMCTSCACCGGAGGCTTTGATAATCACCGG
	Gib_0111_RV	ACTACGCGCATACCGCGACCGWKTMCGWRGTDGGAGGCTTTGATAATCACCGG
	Gib_1111_RV	ACTACGCGCATACCGCGACCGWKTMCGWRACCGGAGGCTTTGATAATCACCGG
ASE Library 83	GIB83_RV	GTAACCCGGATGGATTGCTACTGCGCCGGTGATTTCAG
	GIB_G83.1_FW	TAGCAATCCATCCGGGTTACNDTTTCCTCTCCGAGAACGCC
	GIB_G83.2_FW	TAGCAATCCATCCGGGTTACVHGTTCCTCTCCGAGAACGCC
	GIB_G83.3_FW	TAGCAATCCATCCGGGTTACTGGTTCCTCTCCGAGAACGCC
ASE Library 84	GIB83_RV	GTAACCCGGATGGATTGCTACTGCGCCGGTGATTTCAG
	GIB_F84.1_FW	TAGCAATCCATCCGGGTTACGGCNDTCTCTCCGAGAACGCCAAC
	GIB_F84.2_FW	TAGCAATCCATCCGGGTTACGGCVHGCTCTCCGAGAACGCCAAC
	GIB_F84.3_FW	TAGCAATCCATCCGGGTTACGGCTGGCTCTCCGAGAACGCCAAC
ASE Library 56	GIB56_RV	CGGAGCAGGGCCAATACAGACCGTTTCATCTGCCAGTAATACG
	GIB_S56.1_FW	TCTGTATTGGCCCTGCTCCGNDTGTAAAAAGTTATCTGAACATCCCGGCAATC
	GIB_S56.2_FW	TCTGTATTGGCCCTGCTCCGVHGGTAAAAAGTTATCTGAACATCCCGGCAATC
	GIB_S56.3_FW	TCTGTATTGGCCCTGCTCCGTGGGTAAAAAGTTATCTGAACATCCCGGCAATC
ASE Library 382	GIB382_RV	ATAGTACGGCGGTACGGTATAGCCCGCGTAGATATGAGAC
	GIB_D382.1_FW	ATACCGTACCGCCGTACTATNDTTCAATGATCGGTAAGCTGATTTGC
	GIB_D382.2_FW	ATACCGTACCGCCGTACTATVHGTCAATGATCGGTAAGCTGATTTGC
	GIB_D382.3_FW	ATACCGTACCGCCGTACTATTGGTCAATGATCGGTAAGCTGATTTGC

2.2 Biotin Carboxylase Expression and Recovery for library screening

The culture, protein expression, and protein recovery protocols used for this project were adapted from Rembeza's work [34]. A day after the library was transformed into *E. coli* BL21 (DE3) (as mentioned at the end of section 2.1), individual colonies were picked into wells from a 96-deep well plate (2mL/well) containing 500uL of LB media with 100 µg/mL carbenicillin and cultivated over night at 37°C 700rpm in an Eppendorf ThermoMixer C shaker (3mm orbit diameter).

The following day, to induce expression, 30µL of the culture from each well was used to inoculate a new 96-deep well plate with wells containing 1mL of Terrific Broth Auto Induction Media (AIM) with 100µg/mL carbenicillin and cultivated at 30°C 700rpm for 9h. Cells were then harvested by centrifugation at 3000g for 10min at 10°C and frozen at -80°C overnight. The next morning, frozen cells were thawed 15 min in water bath at room temperature before adding 200µL of Lysis Buffer (50mM HEPES pH 7.4, 5% Glycerol, 300mM NaCl, 0.5mM TCEP) with 0.5ng/µL Lysozyme, 10U/mL DNAse I, and 2mM MgCl, and incubated 30min at room temperature. After the incubation, Triton-X-

100 was added at 0.125% and the 96-deep well plate was block-vortexed to be frozen again for 30min at -80°C. After thawing for 15min using a water bath at room temperature, the plate was shaken at 600rpm using Eppendorf ThermoMixer C, and the lysate is recovered by centrifuging at 3000g at 10°C for 10min and recovering the supernatant.

2.3 Biotin Carboxylase Expression and Purification for variant analysis

Protein expression and purification for variant analysis was performed in a low throughput setting. A colony of each variant was used to inoculate in 5mL of LB media with 100µg/mL carbenicillin in either 15mL or 50mL tubes and cultivated over night at 37°C 180rpm.

The following day, to induce expression, 500µL of the culture from the previous day was used to inoculate a new baffled 250mL shake flask containing 50mL of Terrific Broth Auto Induction Media (AIM) with 100µg/mL carbenicillin and cultivated first at 37°C 200rpm for 2.5h and then at 30°C 200rpm for 4.5h. Cells are then harvested by centrifugation at 3000g for 10min at 10°C and frozen at -80°C overnight.

For protein extraction the next morning, frozen cells were thawed 15min in water bath at room temperature before adding 5mL of Base Buffer (50mM HEPES pH 8, 300mM NaCl, 5% glycerol, 0.5mM TCEP, 20mM imidazole) with 0.5mg/ml lysozyme, 10U/mL DNAseI, 2mM MgCl2, and left 30min incubating at room temperature. The resuspended pellets were then sonicated using an amplitude of 40% for 1min of active sonication (10s on, 20s off). To finish with the lysis process, the lysate was centrifuged at 4°C 4500g for 30min.

Previous to protein purification, 1mL of 50% Talon resin was washed twice with water and twice with Base Buffer. Also, a PD10 desalting column was equilibrated by washing once with 25mL water and 25mL Storage Buffer (20mM HEPES pH 8, 50mM NaCl, 0.5mM EDTA, 5% glycerol, 0.5mM TCEP).

For protein purification, the soluble lysate resulting from the centrifugation (around 5mL) and the Talon resin (1mL) were mixed for 30min in a 4°C cold room. Afterwards, the lysate-resin mixture was poured into a glass column. To remove non-specific binding proteins from the resin, the glass column was washed with 5mL of Base Buffer once and with 5mL of Wash Buffer (Storage Buffer + 40mM imidazole). Finally, protein was eluted with 2.5mL Elution Buffer (Storage buffer + 250mM imidazole).

For desalting, the flowthrough with the eluted protein was poured directly into a PD10 desalting column using a labmate extender. The PD10 desalting columns retain protein for the first 2-3mL of volume poured and elute protein in the next 3-4 mL added, so 3.5mL of Storage Buffer was added to elute and collect the final flowthrough with purified and desalted protein. The purified protein was stored as aliquots of 400µL at -80°C.

To assess expression and solubility, 1uL of total lysate (before centrifugation), 1uL of soluble lysate (after centrifugation) and 10uL of purified protein was run in a SDS-PAGE gel using a Bio-Rad Mini-PROTEAN Precast 4-20% gradient Tris-Glycine gel and running it at 180V for 50 min. The protein ladder used in all protein gels was Spectra Multicolor Broad Protein Ladder. Protein was quantified with the Thermo Scientific™ NanoDrop 2000, by using a built-in program to measure absorbance at 280nm.

2.4 Carboxylation Assay for library screening

Carboxylation of biotin is an MgATP dependent process. As such, a coupled assay with pyruvate kinase (PK) and lactate dehydrogenase (LDH) can be used to monitor ADP production by measuring the oxidation of NADH spectrophotometrically at 340nm as shown in Figure 4 in section 1.2.3. The assay used in this project is adapted from the work of Blanchard [35], and Guchhait [20]. Using Greiner UV-Star 96 well half area transparent plates, the assay was carried out in a total volume of 100µL containing 8mM MgCl2, 100mM HEPES-KOH at pH 8, 8mM $NaHCO_3$ as HCO_3^- donor, 2mM EDTA, 10% ethanol, 1mM phosphoenolpyruvate (PEP), 0.5mM NADH, 1mM ATP, 10U/mL LDH, 6U/mL PK, and 50mM of substrate, were biotin, desthiobiotin or 2-imidazolidone, depending on the experiment. The addition of 10% ethanol is recommended throughout literature as it is said to activate 5- to 6- fold the carboxylation reaction for biotin when the BC subunit is expressed by itself [20]. The use of ethanol was preserved only for the MLA phase, as removing it did not change the activity later when the substrate was changed from biotin to 2-imidazolidone. Therefore, all the carboxylation assays carried out in the ASE phase do not contain ethanol in the reaction mixture.

The reaction was started by the addition of 10uL of diluted (either 1:5 for desthiobiotin and 2-imidazolidone or 1:50 for biotin) soluble cell lysate. The reaction was monitored for 20 minutes with 1 reading per minute at room temperature at 340nm using the SPECTROstar Nano (BMG LABTECH).

2.5 Machine Learning models training

The sequence information was encoded in three different ways: identity, property short (size, side chain type), and property (size, side chain type, tertiary structure preference, functional properties, secondary structure preference) as done by Heil *et al.* [36]. The function information (from the assay) was encoded as “Active” or “Inactive” according to a threshold defined as one standard deviation away from the mean of a “minimally-active” enzyme group, as shown in Figure 8.

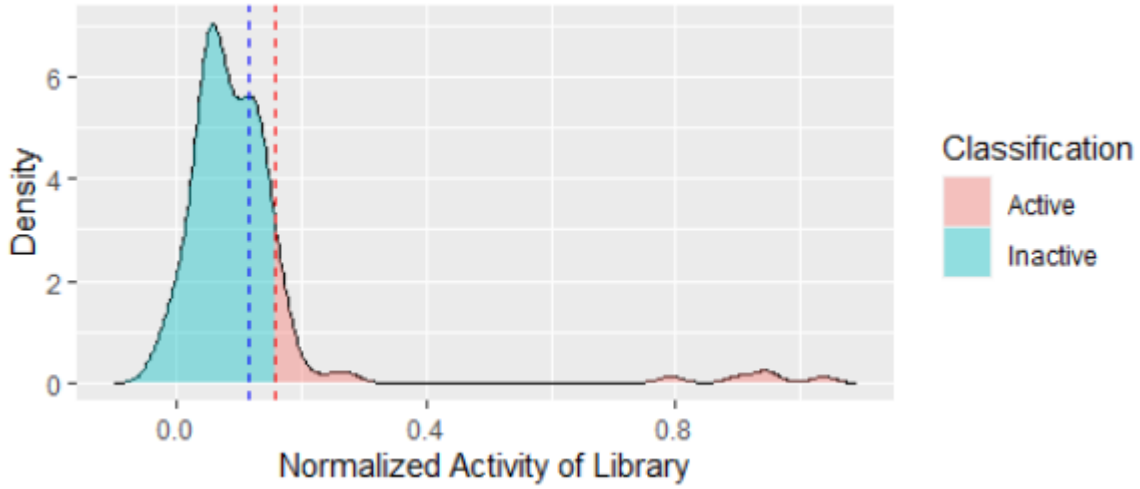


Figure 8. Distribution of normalized activity of training library. In the right-most area, two peaks can be observed, around 0.09 and around 0.12 which we believe represent the background lysate activity and "minimally-active" enzyme activity, respectively. The blue dotted line represents the mean of the "minimally-active" enzyme distribution, while the red dotted line lies one standard deviation from it.

The Scikit-learn Python package was used to train a panel of linear and non-linear models including Lasso and Ridge logistic regressions, Support Vector Classifier (SVC) using linear and radial basis function (rbf) kernel, and random forests. Each of the 5 models was trained with the sequence-function information encoded by each of the 3 sequence encoding types, resulting in 15 different models. The Python libraries and versions used throughout the project shown in Table 2.

Table 2. Software and library versions used for this project.

Programming Language	Library name	Library Version
R 3.6.1	dplyr	0.8.5
	reshape2	1.4.4
	ggplot2	3.3.2
	gridExtra	2.3
	ggseqlogo	0.1
	drc	3.0-1
Python 3.8.5	pandas	1.1.3
	matplotlib	3.3.2
	numpy	1.19.2
	os	-
	biopython	1.78
	sklearn	0.23.2
	itertools	8.6.0
	datetime	-
	joblib	0.17.0

2.6 Generation of Function-Enriched Library

As the results from property and property short encoding type models were very similar, property encoding results were discarded, leaving us with 10 models (5 types of models of 2 different

encoding types). Each of these models generated using the training library was used to *in silico* screen for the complete theoretical library of variants (160,000 variants). R was then used to analyze the results from the models and create sequence logos with the most frequent amino acids from the sequences classified as “Active” by each model. This led to the creation of multiple function-enriched libraries. As each model learns using different criteria, each one might model a different part of the sequence fitness landscape, coming up with different but equally relevant solutions [30]. For this reason, R was used to create a consensus sequence logo of the results from each model to generate a single function-enriched library. The Python libraries and versions used throughout the project shown in Table 2.

After the function-enriched library was designed, a degenerate codon strategy was used to design the primers for making the library. The optimal degenerate codons needed for the specific target amino acids were found using the Python-based ANT tool, which is able to generate degenerate codons for a set of input amino acids minimizing off-target amino acids [37]. Finally, the mutation cassettes were designed similar as the training library cassettes (Figure 7), varying only in the degenerate codons used. Two different degenerate codons were used for each of the target residue positions of the mutagenesis cassette to achieve minimal redundancy and minimal off-target amino acid coding. The combination of these two different degenerate codons for each of the 4 positions resulted in 16 reverse primers that were mixed equimolarly before its use. The primers used for this second round of mutations can be found in Table 1 in the rows labeled as MLA Function Enriched Library. The PCR conditions, the plasmid preparation, library construction, and transformation were performed in the same way as described in section 2.1.

2.7 Generation of Active Site Libraries

The strategy chosen for the generation of the Active Site libraries was to perform SSM for each amino acid position separately. The mutations were be performed on the most active variant from the Function-Enriched Library for 2-imidazolidone. The degenerate codons used for the SSM used are the ones suggested by the “22-c trick” strategy [29]. This strategy is meant to reduce amino acid redundancy in the resulting libraries. The primers for this strategy were designed similar to the first round. The forward primer consists of 20 bases region for assembly followed by a degenerate codon, and an 18 bases region to serve as primer for the PCR. The reverse primer in 5’ to 3’ direction consists of 20 bases region for assembly, complementary to the ones in the forward primer, followed by an 18 bases region to serve as primer for the PCR similar to Figure 7. The primers used for this PCR are under the rows labeled as “ASE Library” shown in Table 1, together with the rest of the primers used throughout this project.

The PCR reactions performed for the generation of the ASE libraries used the circular plasmid as DNA template. For the PCR conditions used in the generation of these libraries, the denaturing

phase of each cycle was changed to 10s per cycle, the specific annealing temperature was changed to 62°C, and only 25 cycles were run. After the PCR was finished 0.5uL DpnI was added and incubated overnight at 37°C to eliminate template DNA and prepare the amplified region to reassemble with itself using isothermal assembly, similar to the process described in section 2.1.

2.8 Carboxylation Assay for variant analysis

When the best performing variant was found from the ASE library screening, a more detailed series of experiments were carried out to find the specific behavior of the variant towards its substrates. This set of experiments was also performed in order to obtain kinetic information. The carboxylation assay for these experiments used the same conditions of mentioned previously on section 2.4 with the exception of the substrates. The amounts of ATP, $NaHCO_3$, and biotin or 2-imidazolidone were varied in some experiments. It is important to highlight that the biotin or 2-imidazolidone were dissolved in HEPES KOH pH 8 buffer, so the concentration of HEPES KOH pH 8 in the reaction mix had to be kept constant even when fewer or no substrate was used. All these experiments were done without the 10% ethanol used for activation.

After testing different purified protein dilutions that could be used for the experiment it was determined that the carboxylation assays should be initiated by the addition of 10µL of undiluted purified protein (~2-5µg). The experiments were monitored with the same conditions than as described in section 2.4.

3. Results and Discussion

3.1 Training Library

The purpose of this first round of mutations is not yet to find an improved strain, but rather to obtain sequence-function data from as many diverse variants for our target region library (Training Library) as possible. The target region to be mutated was the Gly-rich loop (Gly162-165), which is in the entrance to the active site of the BC subunit as shown in Figure 3. The resulting sequence-function data from this experiment was later used to train the ML models to learn which mutant amino acid residues in the target region would yield a functional BC enzyme.

As it has been previously reported that Machine Learning-assisted directed evolution requires a low number of screened variants of a library [30], we decided to test 176 colonies resulting from the transformation of the Training Library (created as described in section 2.1) (2 96-deep well plates – 6 parent enzyme repetitions and 2 negative controls per plate).

The assay used to screen these experiments introduces variability in different stages. First, the growth in the 96-well plate might not be homogeneous due to differences in aeration. Second, the expression might not be homogeneous. One of the reasons expression might not be homogeneous, other than different solubility for each variant, is that induction of expression is linked to growth when using Auto Induction Media, as the inductor will not work until the glucose is depleted. To account the variability introduced by the screening method, a previous experiment was performed to calculate a coefficient of variance or relative standard deviation (CV). To perform this,94 wells out of a 96-deep well plate was inoculated with colonies from the same parent enzyme and were assayed for biotin carboxylase activity. The CV was calculated to be between 8 and 15.7% after repeated measurements. For this reason, lines at a distance of 2*CV are shown in the plots as reference.

The activity of the biotin carboxylase library was calculated using the rate of oxidation of NADH and normalized to the activity of the parent enzyme within the plate. These results are displayed in a negative scale from 0 (no NADH oxidation) to 1 (parent enzyme activity). The results, seen in Figure 9, show a range of activities with most variants having a dramatic decrease in function. This implies that the glycine residues at the glycine-rich loop are relevant for the BC activity. After sequencing all the variants, some turned out to be identical to the parent sequence and some variants were duplicated, which means only 110 unique variants were actually tested.

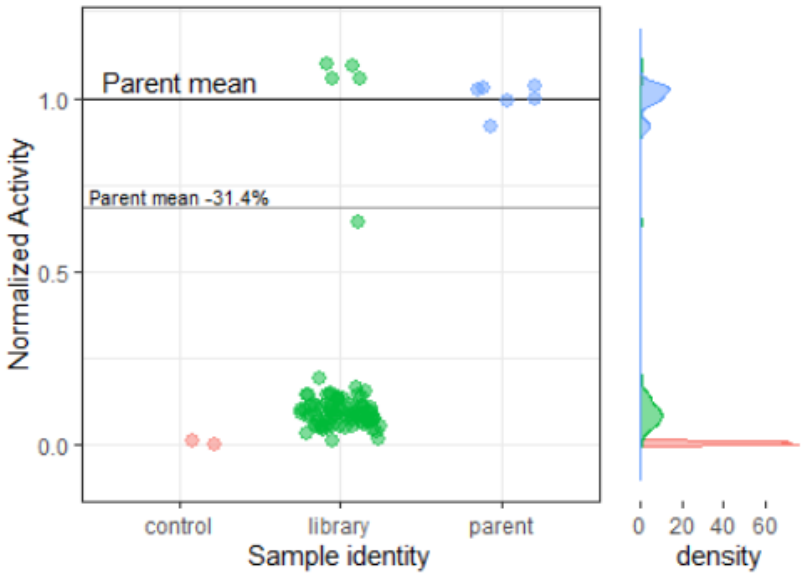


Figure 9. Normalized biotin carboxylase activity (left) and density distribution (right) of the Training Library. Control, Library, and Parent samples are indicated with red, green, and blue respectively.

3.2 Machine Learning models

The purpose of these models is to learn using the sequence-function data from the Training Library. After learning using different modelling algorithms, the model's purpose was to predict whether a specific mutation would yield an active or inactive enzyme. This would be later used to create a function-enriched library.

Ideally, the model should be trained with data having a uniform distribution of the 20 amino acids per position, so that each model would learn better from the effect of each amino acid. However, as only 110 variants were tested, and due to the NNK strategy being biased towards several amino acids, the models were trained with a biased data set. The amino acid distribution for each amino acid position within the mutation cassette (4 amino acids long) is shown in Figure 10. An even greater and more concerning bias is the proportion of active and inactive variants fed into the model. As it can be seen in Figure 8 in section 2.5, the "Inactive" variants were in a vastly greater amount (98 Inactive vs 12 Active) after removing duplicated variants. These results introduce several layers of bias, now in terms of amino acids and in terms of activity, to our models. To compensate for the bias introduced by the imbalanced classes (98 Inactive vs 12 Active), the parameter "class_weight" for all the used models was set to "balanced". This parameter allows for the model to assign each class a weight that is inversely proportional to their frequency. Therefore, even if the "Active" class is underrepresented, its upweight compared to the "Inactive" class reduces the bias introduced into the model.

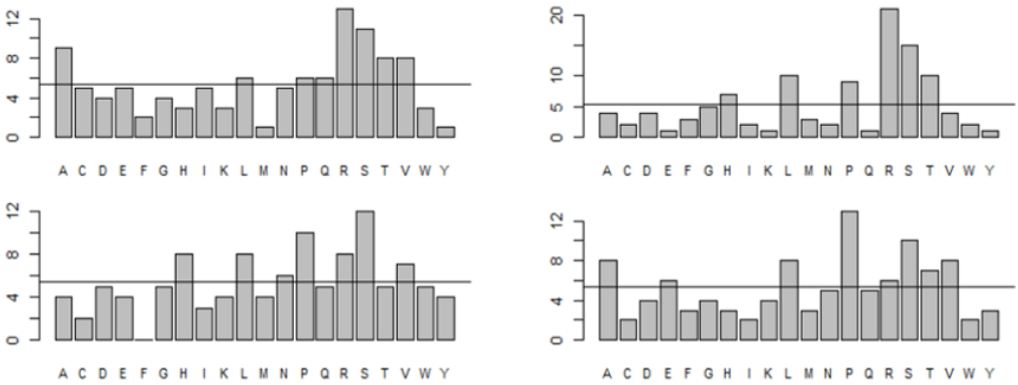


Figure 10. Amino acid frequency distribution of the sequence information fed into the machine learning models. The horizontal black line represents uniform distribution. This distribution has already omitted sequence replicates. The amino acid positions within the region of interest (Gly162-165-loop) are the First (162 top-left), Second (163 top-right), Third (164 bottom-left), and Fourth (165 bottom-right).

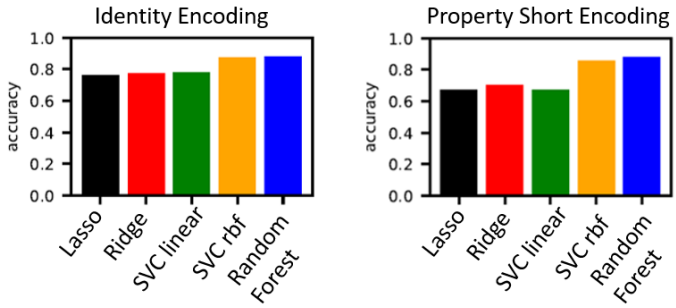


Figure 11. Accuracy metric for the 10 different models used to generate the consensus Function-Enriched Library. The 10 models come from 2 different sequence encoding types used to train 5 different machine learning models each.

Models were trained using a leave-one-out approach and the accuracy of classification on the training data was measured as fitness indicator for each model as seen in Figure 11. As described before, given that all models have similar metrics and that each model learns from a different part of the fitness landscape, all the models were taken into consideration. A consensus of the amino acids from the sequences classified as “Active” by the *in silico* screening, for each amino acid position within the Gly162-165 loop, was made as shown in Figure 12.

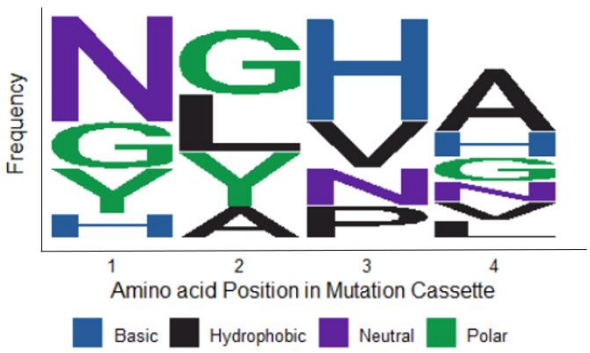


Figure 12. Consensus sequence logo of the Function-Enriched Library.

3.3 Function-Enriched Library

The purpose of this library is to find a variant with an improved carboxylase activity over either free biotin, desthiobiotin or 2-imidazolidone. As the function-enriched library was designed to retain the known catalytic residues but introduce amino acid variability in the entrance of the active site, we expect more promiscuity from this library.

The function-enriched library was designed according to the consensus from the models as was shown in Figure 12. Following a 16-primer strategy for the mutation cassettes (2 degenerate codons per amino acid position), we were able to minimize the presence of off-target residues, as shown in Table 3. For this library, we also decided to test 176 variants using a similar 96-well plate distribution. As this library was not going to be used for model training, we decided not to sequence all the variants, but just focus on the ones that exhibited higher activity for any of the biotin analogs. The activity of these assays was also normalized using the activity of the parent enzyme within the plate and kept negative to show oxidation of NADH.

Table 3. Function Enriched Library Mutation Cassette Design. This table shows the degenerate codon strategy used to code for each of the target amino acids for each of the 4 positions within the Mutation Cassette. Off-target amino acids were minimized using the ANT tool [37].

Function Enriched Library Mutation Cassette Design			
Position 1	Target Amino Acids	Asn-Tyr-His-Gly	
	Amino Acids per codon	Asn-Tyr-His	Gly
	Degenerate codon(s) used	HAC	GGT
	off-target amino acids	-	-
Position 2	Target Amino Acids	Gly-Ala-Tyr-Leu	
	Amino Acids per codon	Gly-Ala	Tyr-Leu
	Degenerate codon(s) used	GSA	YWC
	off-target amino acids	-	Phe-His
Position 3	Target Amino Acids	His-Asn-Pro-Gly-Val	
	Amino Acids per codon	His-Asn-Pro	Gly-Val
	Degenerate codon(s) used	MMC	GKA
	off-target amino acids	Thr	-
Position 4	Target Amino Acids	Gly-Val-Ala-Leu-Asn-His	
	Amino Acids per codon	Gly-Val-Ala	Leu-Asn-His
	Degenerate codon(s) used	GBA	MWC
	off-target amino acids	-	Ile

Even though biotin carboxylase has reported activity for BCCP-bound and free biotin, there are no previous reports of any activity for desthiobiotin or 2-imidazolidone. For this reason, our intention of creating a function enriched library of BCs, is to find a variant capable of carboxylation (as key residues for carboxylation are not been modified), but with an active site more suited for the docking of smaller biotin analogs. The activity of the function-enriched library for free biotin, desthiobiotin, and 2-imidazolidone is shown in Figure 13.

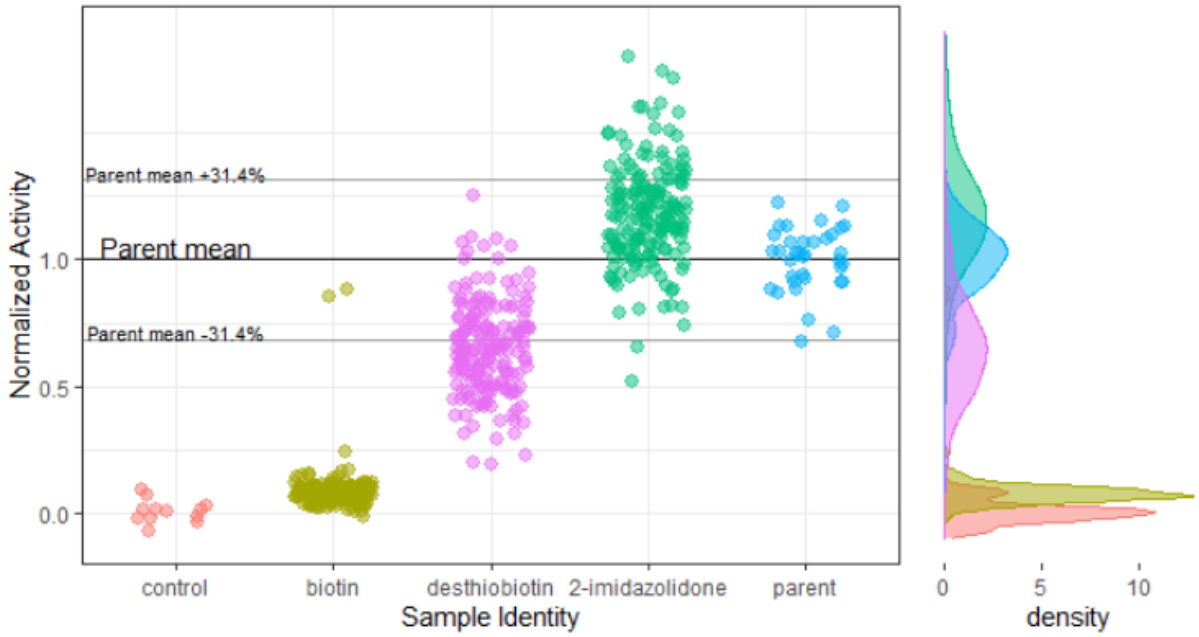


Figure 13. Function-Enhanced library activity towards biotin analogs normalized using parent enzyme activity. Biotin carboxylase activity (yellow), desthiobiotin carboxylase activity (purple), and 2-imidazolidone carboxylase activity (green). Control and Parent samples are indicated with red and blue respectively.

The variants with greatest activity for each substrate (biotin, desthiobiotin, and 2-imidazolidone) were chosen to be re-tested, as a single colony assay for each cannot be considered for statistical significance. A total of 22 variants were chosen for re-test with 4 biological repetitions each. The re-screening was performed in the same 96-well format with 6 repetitions of the parent enzyme and 2 negative control wells. This re-test experiment screened the 22 variants and the parent enzyme for their activity on the three substrates to confirm their activity. The results of the re-test experiment can be seen in Figure 14 for each substrate. The top 2 variants from the re-test of each substrate were sequenced and the results can be seen in Table 4.

Two out of the three variants that displayed a higher activity than the parent showed two point mutations in positions between the first and the third position of the Mutation Cassette (Gly162-164). This means that via single step SSM, we would have to run at least three rounds of mutations to find these variants. As single mutations are not guaranteed to work when combined, it is possible that these variants would have not been found via single step SSM strategy. While the Function-Enriched

Library did not yield any variant with higher activity towards free-biotin, 3 variants with improved function for smaller analogs were found. These three variants were internally called 2A4 (G163L, G164H, F279L), 1E11 (G163L, F279L), and 2E10 (G162Y, G163L, F279L) and the G163L mutation was found to be introduced in this mutation round in all of these improved variants. It is important to remember that carboxylase activity for these smaller biotin analogs has previously been reported to be non or close to non-existing; 0.2% in comparison to biotin for desthiobiotin and 0% in comparison to biotin for 2-imidazolidone [16]. To the best of our knowledge, this represents the first directed evolution on a BC subunit that has targeted and obtained some results on a change of substrate specificity.

These results also suggest that the Glycine-rich loop region (Gly162-165) of the BC is highly conserved not only because BCCP-biotin is the original substrate but because some other unidentified function. Although, this must not be interpreted as conclusive, as our models are highly biased, and the tested library and its results are a consequence of that bias. In literature, Gly165 and Gly166 have been reported as increasing K_m towards MgATP when mutated [19]. It is not easy to conclude whether this region is a good or a bad candidate for directed evolution for increasing affinity to free biotin due to our reduced screening. However, it seems to be a good target for substrate specificity.

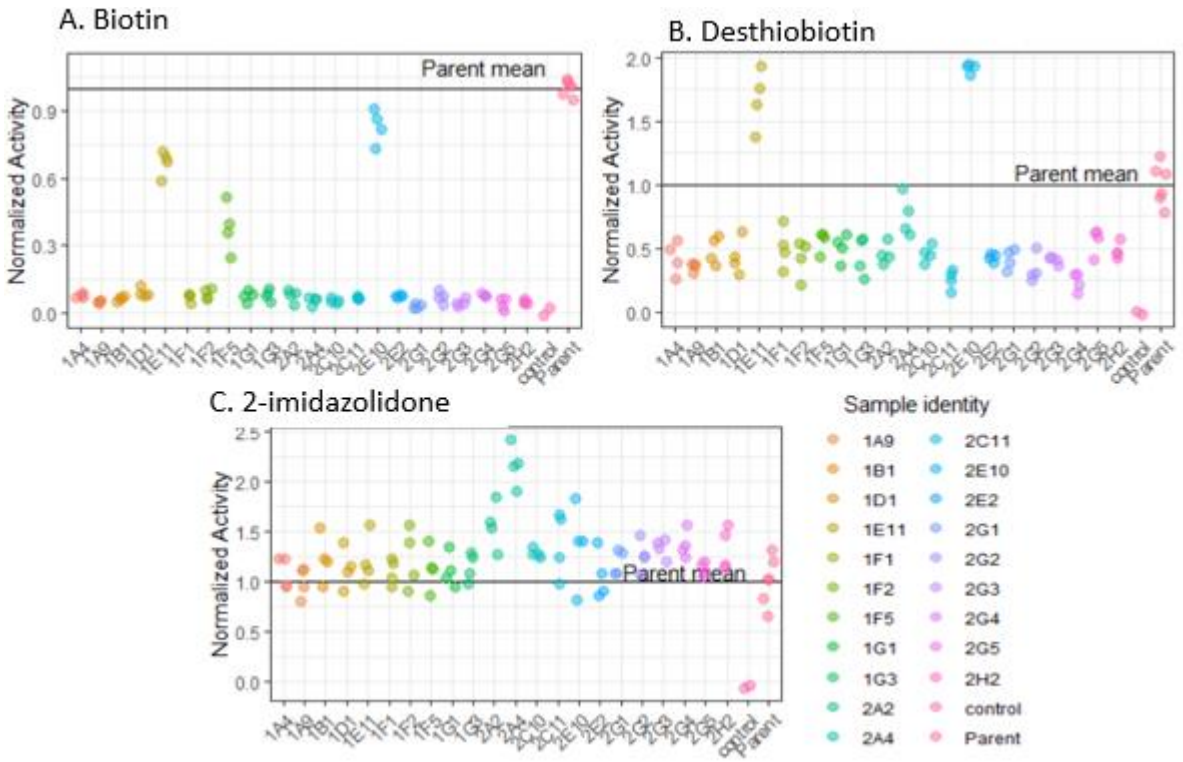


Figure 14. Normalized biotin carboxylase activity (top-left), desthiobiotin carboxylase activity (top-right), and 2-imidazolidone carboxylase activity (bottom) of the 22 selected variants. Each variant was screened with 4 biological repetitions, except from the parent enzyme and the negative control which had 6 and 2 biological repetitions, respectively.

Table 4. Sequence-Function information from the top variants for all substrate carboxylation assays. The activity of each variant is normalized so that it is displayed as a percentage of the parent activity. No variant showed improved free-biotin carboxylase activity. Variant 1E11 and 2E10 excelled in desthiobiotin carboxylase activity, while variant 2A4 was the best in 2-imidazolidone activity. Parent enzyme sequence-function information showed in the last row for reference. Other than the mutations shown in the Sequence column, all variants, including Parent enzyme, have a F279L mutation.

Internal ID	Sequence (DNA)	Sequence (amino acid)	Biotin Carboxylase Activity	Desthiobiotin Carboxylase Activity	2-imidazoldione Carboxylase Activity
2A2	GGTGCAGTACTC	GAVL	7%	40%	160%
2A4	GGTCTCCACGGT	GLHG	5%	70%	**220%**
1E11	GGTCTCGGAGGA	GLGG	**70%**	**170%**	120%
2E10	TACCTCGGAGGA	YLGG	**80%**	**190%**	140%
1F5	GGTGCSGGAGGA	GAGG	40%	60%	110%
Parent	GGAGGAGGAGGA	GGGG	100%	100%	100%

3.5 Active Site Engineering Libraries

For this next round of engineering, variant 2A4 (G163L, G164H, F279L) (Table 4) was chosen to be the new parent enzyme, as it showed the greatest activity towards 2-imidazolidone. The purpose of this mutation round is to see if any of the four libraries produced by individual SSM of the four selected residues (S56, G83, F84, and D382) would yield an increase of carboxylase activity of variant 2A4 (new parent enzyme) towards 2-imidazolidone.

The four Active Site Engineering Libraries (S56, G83, F84, and D382) were tested for 2-imidazolidone carboxylase activity. For each library, the 2-imidazolidone carboxylase assay was performed on a 96-well plate inoculated with 6 of the new parent enzyme (2A4 variant) colonies, 2 empty wells (as negative control) and 88 colonies resulting from the transformation of each library. The results from these assays were normalized per plate to the activity of the new parent enzyme (2A4) and represent the ADP production by measuring oxidation of NADH. The results of the four libraries are shown pooled together in Figure 15.

Similar to what was done with the function enriched library, 22 variants were selected from these results to undergo a re-test. In this re-test, each colony will be re-plated in a LB-carbenicillin plate and used to inoculate 4 wells in a 96-well plate. From the rest of the wells of the 96 well plate, 6 were inoculated with the new parent (2A4 variant) enzyme, and 2 were left blank. This re-test is done to add statistical significance to the test and have greater certainty of the activity of the best variants.

The re-test assay results were normalized using the parent's (variant 2A4) activity and are shown below in Figure 16.

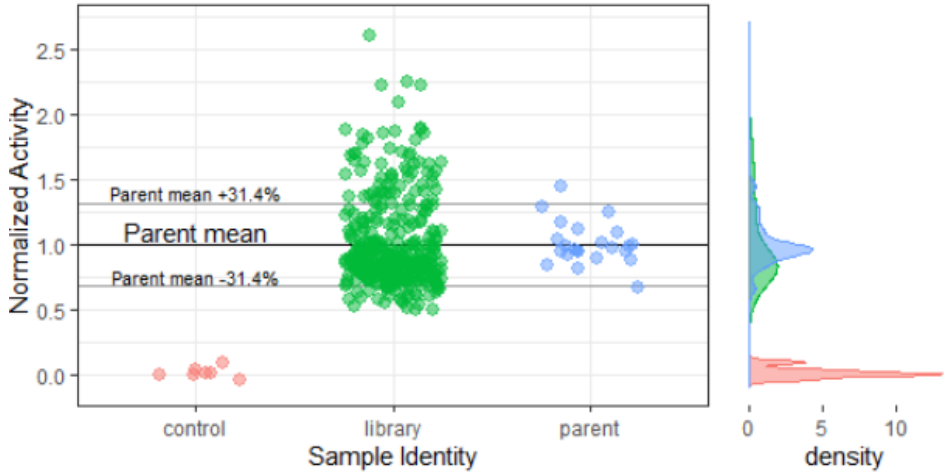


Figure 15. Normalized 2-imidazolidone carboxylase activity dot plot (left) and density distribution (right) of the Active Site Engineering Library. Control, Library, and Parent samples are indicated with red, green, and blue, respectively.

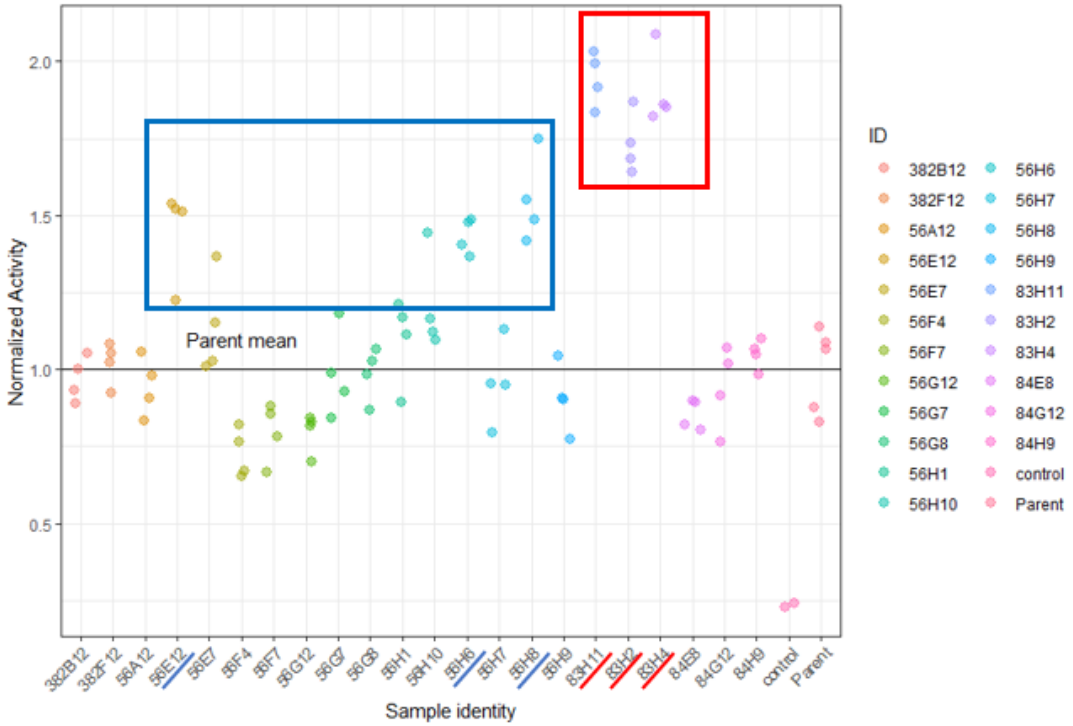


Figure 16. Normalized 2-imidazolidone carboxylase activity dot plot of the variants chosen for re-test from the Active Site Engineering Libraries. Two groups of variants which showed higher activity than the parent enzyme were identified. These groups belong each to a different ASE library and are identified in blue (from library 56) and red (from library 83).

The re-test results show that most of the variants previously selected from library 56 had overestimated activity values in the screening. While this could happen because of the number of repetitions being just one on the screening, most likely there were unknown variations in the growth conditions which resulted in an increased activity, even after normalization.

From the results of the re-test there were two groups of variants specifically showed very similar activity. Each of these groups was formed by three variants from the same library, and each of these groups comes from a different library. The reduction in amino acid redundancy due to the “22-c trick” makes possible that these six variants actually came from just two different mutations. This was shown to be true after sequencing, where the variants 56E12, 56H6, and 56H8 (Figure 16, blue rectangle) were all the same S56E mutation, and variants 83H11, 83H2, and 83H4 (Figure 16, red rectangle) were all the same G83Y mutation. It is important to remember that these variants all contain the mutations of the parent enzyme used (G163L, G164H, F279L).

The overall mutant with greatest activity was found to be G83Y (G163L, G164H, F279L), and is estimated to have a 4- fold activity on 2-imidazolidone in comparison to the original parent enzyme (F279L) of this project.

3.6 Best Variants Analysis

After the ASE phase of engineering was finished, variant G83Y (including also G163L, G164H, F279L mutations) and the wild type accC (WT) were assayed to determine the differences in activity towards the different substrates caused by the introduced mutations.

An initial test revealed that the purified protein had to be used undiluted, straight from the purification protocol, in order for activity to be easily detectable for both, WT and G83Y purified proteins and both, 2-imidazolidone and biotin substrates.

To try to obtain kinetic data from our G83Y variant and the WT, they were assayed as described in section 2.8. Biotin concentration was varied while the other substrates (as HCO_3^- and ATP) were kept constant. The results are shown in Figure 17.

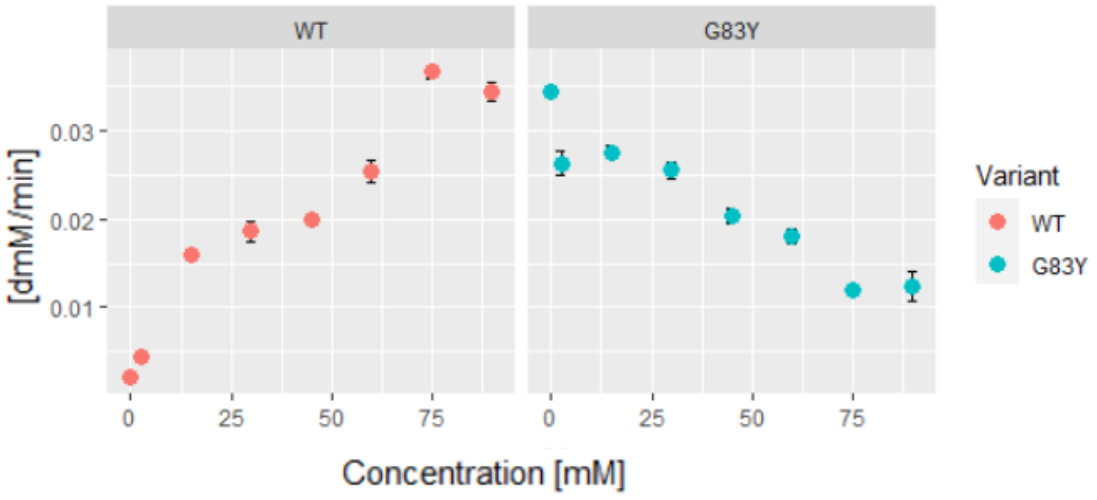


Figure 17. Kinetic results from test with varied biotin concentrations. Activity is reported in d(mM of produced ADP)/min. Wild Type results are on the left (red), while G83Y results are on the right (blue). Error bars represent standard error per timepoint.

The results for WT kinetics are consistent with literature [14], [19], as fitting the data under the Michaelis-Menten equation (Figure 18) yielded a K_m of 61.64+-17.9mM and a V_{max} of 0.058+-0.008min^{-1}.

Surprisingly, Michaelis-Menten constants were not able to be obtained for G83Y variant's activity with biotin. The reason for this, is that the kinetic data observed for that variant does not follow a typical Michaelis-Menten pattern where, as the initial concentration of the substrate is increased, the velocity increases as well until reaching saturation. Instead, variant G83Y follows an inverse pattern. According to our results, the G83Y mutant has a basal ATP-hydrolysis activity of around 0.03mM/min. This ATP-hydrolysis activity seems to be inhibited by biotin almost until the point of complete inhibition. Assays with a higher biotin concentration were not achieved as biotin was insoluble in concentrations greater than 90mM in the conditions used in our assay. Therefore, it is not possible for us to say whether biotin inhibits G83Y's ATPase activity completely, or if there is a saturation for this inhibition of the ATPase activity. As the enzyme used for this assay was purified rather than a simple lysate, we discard the behavior as an action of background enzymes or background ATP. While we did not expect G83Y to have this behavior, we did expect it not to be able to carboxylate biotin, as is parent enzyme (variant 2A4) retained 5% of biotin carboxylase activity from the original enzyme.

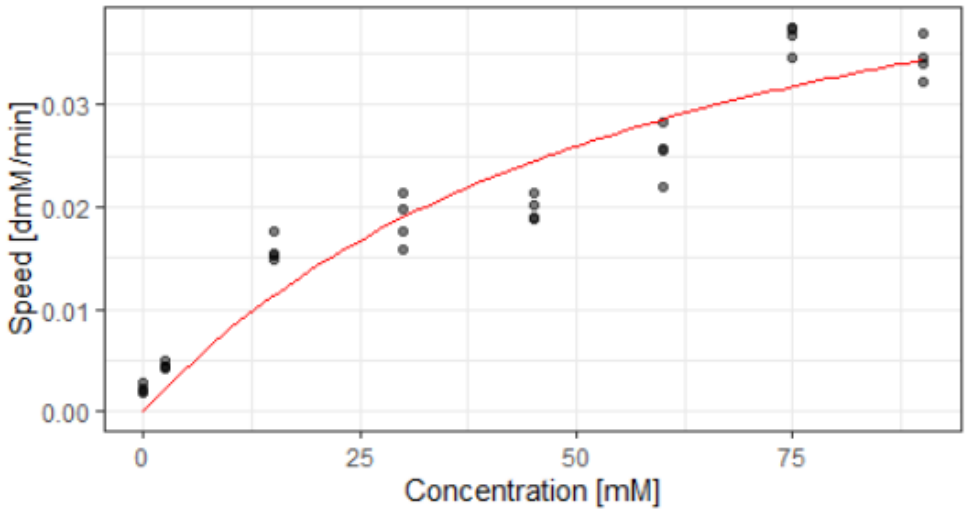


Figure 18. Michaelis-Menten fitting to the wild-type kinetic data with biotin. K_m 61.64+-17.9mM V_{max} 0.058+- 0.008 min^{-1}.

Similar experiments were also performed using 2-imidazolidone as substrate. As 2-imidazolidone is more soluble than biotin, higher concentrations were achieved. The results are shown in Figure 19. Kinetic results from WT are consistent with literature (0% activity)[16]. As expected, kinetic results from G83Y show the same basal ATPase activity of 0.03mM/min. However, the ADP production activity increases upon addition of 2-imidazolidone, which suggests either an activation of the ATPase activity or the coexistence of ATPase and 2-imidazolidone carboxylase activity. Even if a Michaelis-

Menten equation fitting is not possible for the present data, due to the unexpected behavior. The combined results of G83Y with biotin and 2-imidazolidone suggest a strong futile ATP-hydrolysis activity.

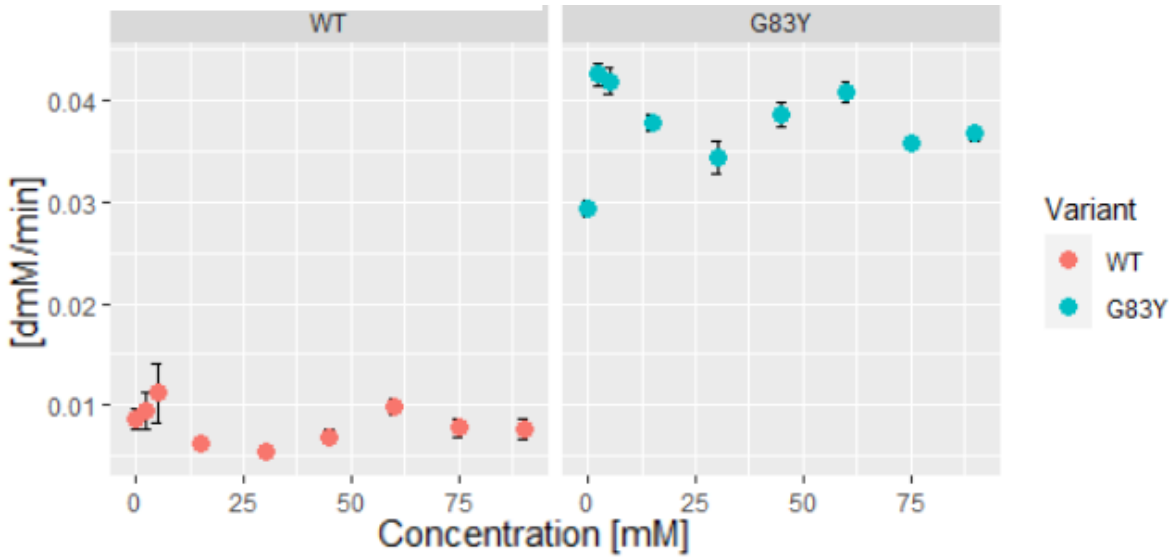


Figure 19. Kinetic results from test with varied 2-imidazolidone concentrations. Activity is reported in d(mM of produced ADP)/min. Wild Type results are on the left (red), while G83Y results are on the right (blue). Error bars represent standard error per timepoint.

Futile ATP-hydrolysis activity (ATPase activity) is a commonly observed in biotin-dependent carboxylases [22]. Mutations made in *E.coli* BC subunit have been used as models to show that, in some human diseases, these mutations provoke a misalignment of the substrates and show ATPase activity. In these human disease simulations, WT BC and mutant M169K show a 1:1 stoichiometry of ADP and carboxybiotin formed. However, some of the mutants tested (R338S and R338Q) showed a between 3-4:1 ratio in ADP to carboxybiotin production[21]. This allows for the possibility of carboxylation and futile ATPase activity to coexist.

The futile ATPase activity has also been reported in other engineering endeavors of biotin-dependent carboxylases, such as propionyl-carboxylase. In that project, mutations that achieved a change in substrate specificity (from propionyl-CoA to glycoyl-CoA), also resulted in a ADP to carboxybiotin production ratio of 100:1, which was later improved to around 6:1 [10]. This phenomenon has also been reported in pyruvate-carboxylase, which is also a member of the biotin-dependent carboxylases family [22]. An explanation to such futile ATPase activity, is that the reaction occurring at BC is actually not one, but two coupled reactions. In this model, ATP reacts with HCO_3^- ion to produce a carboxyphosphate intermediate [23]. This carboxyphosphate later reacts with biotin (or biotin analog) releasing P_i [21], [22]. In the absence of biotin, neither of these reactions seems to happen for most enzymes, but for some mutants the first reaction occurs and carboxyphosphate later degrades.

All of these studies that show an ADP to carboxybiotin ratio were able to do so by monitoring the reaction spectrophotometrically but also determining the amount of fixed [^{14}C] carbon radiometrically. However, due to out experimental set up for high throughput screening, radiometric monitoring was not feasible during the screening phases, introducing bias in our selection of the best variants. By using the data available to us from these experiments, we can calculate a pseudo ADP to 2-imidazolidone ratio. This can be calculated by observing the ATPase activity (0.0293mM/min) and the average ATPase activity when 2-imidazolidone is in the reaction (0.038mM/min). Under the assumption that the difference in activities is just due to 2-imidazolidone carboxylation, an estimated ratio would be around 4.4:1 (4.4 molecules of ADP produced for each 1 molecule of 2-imidazolidone carboxylated). Whether the assumptions are correct, or 2-imidazolidone is "enhancing" the ATPase activity, or 2-imidazolidone is being carboxylated while futile ATPase activity is diminished, we do not have enough evidence to conclude.

3.7 Futile ATPase Mutation Analysis

A final experiment was performed on purified versions of the most relevant variants used during this project. The objective of this assay was to identify which mutation introduced the ATPase activity. For this assay four different treatments were used in which the full reaction was used first and then the biotin analog, HCO_3^- , and ATP were alternatively removed from the reaction.

This assay was performed using 4 replicates of each of the following variants: WT accC, the original parent enzyme provided by Engqvist lab (labeled as 14D8, with mutation F279L), 2A4 (best variant from MLA phase, with mutations G163L, G164H, F279L) and the two best variants from the ASE phase G83Y (G83Y, G163L, G164H, F279L) and S56E (S56E, G163L, G164H, F279L). The results of this test are shown in Figure 20.

It is important to note that for these experiments, 2µg of protein was used for WT and G83Y, however this was not possible for the rest of the variants, as our assay was not able to detect any protein after expression and purification of 2A4, 14D8, or S56E. For the variants with unknown protein concentration, 10µL of the purified fraction were used instead.

A protein gel was used to visually assess for protein in the purified fraction, and while strong bands were identified for WT and G83Y, very faint bands of 14D8 and S56E were detected, but no band was visible for 2A4. This could be a result of the optimization of the growth conditions used for the protein purification, which was performed specifically for WT and G83Y.

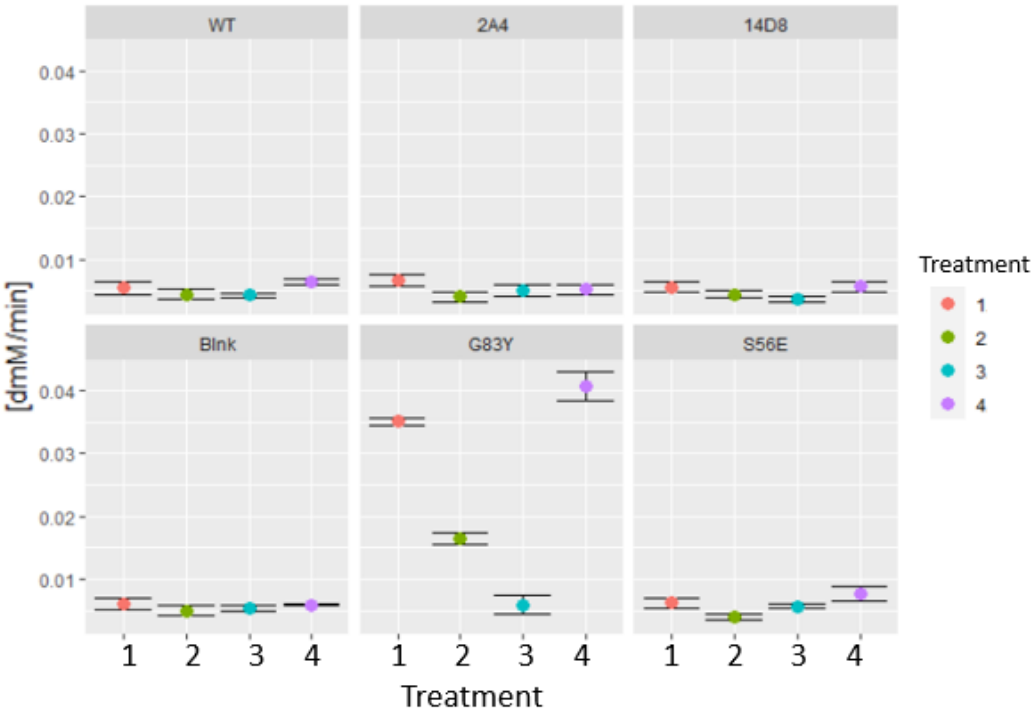


Figure 20. Kinetic results from test with 4 different treatments. Activity is reported in d(mM of produced ADP)/min. Treatments 1: 2-imidazolidone substrate. Treatment 2: no HCO3-. Treatment 3: no ATP. Treatment 4: Full reaction with all substrates. Error bars represent standard error for each group.

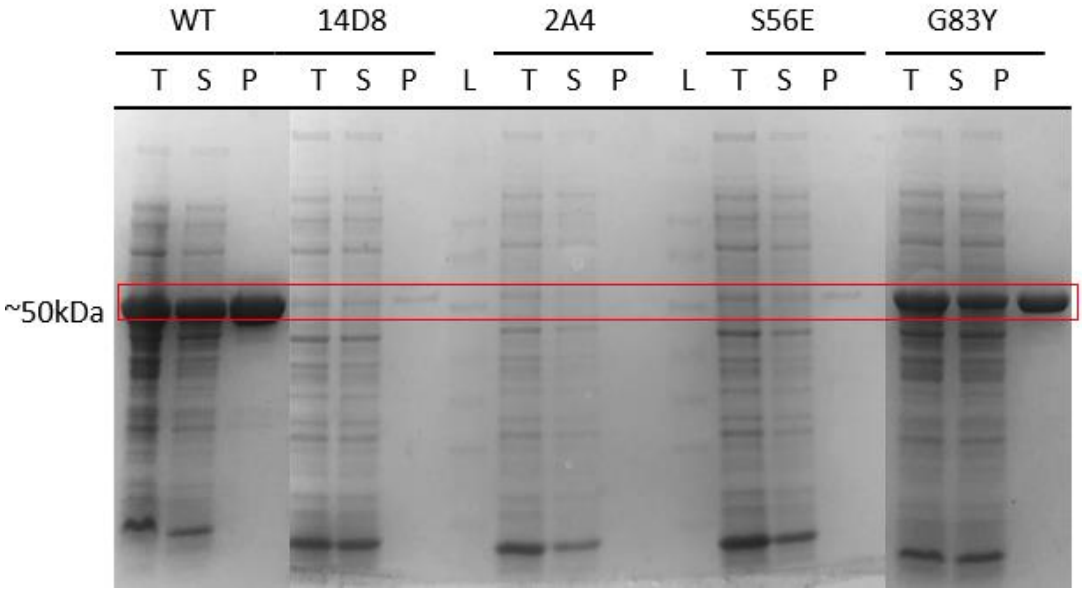


Figure 21. Protein Gel for the variants used on the Futile ATPase Mutation Assay. For each variant, 1µL of Total Lysate (T), 1uL of Soluble Lysate (S) and 10µL of Purified protein were run. The ladder used was Spectra multicolor broad protein ladder.

The results from the protein gel suggest that the expression conditions were sub-optimal. Due to time constraints, we were unable to repeat this experiment. For this reason, our ability to make conclusions regarding which mutation introduced the ATPase activity are limited.

The results also show the basal ATPase activity for G83Y and the previously observed increase in activity after the addition of 2-imidazolidone. This suggests that the carboxylation of 2-imidazolidone could be happening. Also, activity drops significantly when removing non-basal HCO_3^-, showing that the ATPase activity is HCO_3^- dependent. ATPase activity being HCO_3^- dependent further supports the idea of the carboxylation reaction occurring in two phases, as mentioned previously. The residual activity in treatment 2 for G83Y could be explained by the naturally occurring HCO_3^- in the HEPES-KOH pH 8 buffer at room temperature conditions.

With the experiments performed in this project, we were able to engineer a BC to change its substrate specificity. This is the first time a biotin-dependent carboxylase is engineered via the BC subunit for substrate specificity, to the best of our knowledge. The experimental results suggest that the desired activity is present, even when an undesired ATPase activity complicates to its detection. These results are supported by an ANOVA performed on the activity of the G83Y variant as dependent variable and the 4 experimental treatments as independent variable. The ANOVA indicates that the treatments are a variable that affects the activity with a significant p value of 3.466e-09. Additionally, pair-wise t-tests were performed between the different treatments to assure that there were significant differences between all of them, as shown in Table 5. This was especially needed to confirm that the ATPase activity was different with and without 2-imidazolidone, which is confirmed with p value of 0.0208.

Table 5. Pair-wise t-tests between the experimental treatments for activity of variant G83Y. Treatments 1: 2-imidazolidone substrate. Treatment 2: no HCO3-. Treatment 3: no ATP. Treatment 4: Full reaction with all substrates. Error bars represent standard error for each group.

Pair-wise t-tests			
Treatments	**1**	**2**	**3**
2	1.30E-06	-	-
3	9.10E-09	2.70E-04	-
4	2.08E-02	7.70E-08	1.20E-09

4. Final remarks and future work

To the date, and to the best of our knowledge, this is the first time 2-imidazolidone and desthiobiotin carboxylase enzymes has been reported. This work provides several point mutations that can be useful in the future engineering of other biotin-dependent carboxylase variants. The Machine Learning-assisted directed evolution approach was effective to produce these variants but failed to produce a variant with higher affinity to free biotin. While this could be due to several reasons, we suggest computational analysis to predict a better region (instead of Gly-loop) for the use of this strategy. Additionally, to improve the MLA approach, the use of strategies such as “22-c trick” or NDT degenerate codons to reduce redundancy would be useful to increase the screening coverage.

The Active Site Engineering approach resulted in a significant increase in activity. However, due to time and experimental limitations, we were not able to conclude whether this increment in activity was completely the desired activity. For this reason, we recommend any future work on this to do a parallel monitoring of the reaction. By doing so, one can make sure that the directed evolution screening is actually selecting what the project wants, and not what the experiment allows one to see.

The use of several directed evolution strategies allowed us to perform the initial steps toward a substrate walk in just 3 rounds of mutation (from which 1 was for model training), by introducing three mutations. The variants engineered in this project hold the potential to be further improved towards less ATPase activity and increased 2-imidazolidone carboxylase activity. This represents a valuable step in the creation of key-enzymes in the strategy of engineering a new RuBisCO enzyme with greater carbon-fixating activity.

5. Bibliography

[1] I. Andersson, “Catalysis and regulation in Rubisco,” *J. Exp. Bot.*, vol. 59, no. 7, pp. 1555–1568, 2008.

[2] M. Van Lun, J. S. Hub, D. Van Der Spoel, and I. Andersson, “CO2 and O2 distribution in rubisco suggests the small subunit functions as a CO2 reservoir,” *J. Am. Chem. Soc.*, vol. 136, no. 8, pp. 3165–3171, 2014.

[3] M. A. J. Parry, P. J. Andralojc, R. A. C. Mitchell, P. J. Madgwick, and A. J. Keys, “Manipulation of Rubisco: The amount, activity, function and regulation,” *J. Exp. Bot.*, vol. 54, no. 386, pp. 1321–1333, 2003.

[4] H. Ashida *et al.*, “RuBisCO-like proteins as the enolase enzyme in the methionine salvage pathway: Functional and evolutionary relationships between RuBisCO-like proteins and photosynthetic RuBisCO,” *J. Exp. Bot.*, vol. 59, no. 7, pp. 1543–1554, 2008.

[5] Y. Saito *et al.*, “Structural and functional similarities between a ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO)-like protein from Bacillus subtilis and photosynthetic RuBisCO,” *J. Biol. Chem.*, vol. 284, no. 19, pp. 13256–13264, 2009.

[6] O. Mueller-Cajar, M. Morell, and S. M. Whitney, “Directed evolution of Rubisco in Escherichia coli reveals a specificity-determining hydrogen bond in the form II enzyme,” *Biochemistry*, vol. 46, no. 49, pp. 14067–14074, 2007.

[7] Y. Zhou and S. Whitney, “Directed evolution of an improved Rubisco; in vitro analyses to decipher fact from fiction,” *Int. J. Mol. Sci.*, vol. 20, no. 20, 2019.

[8] F. R. T. S. Satagopan, S. S. Scott, T. G. Smith, “A Rubisco mutant that confers growth under a normally ‘inhibitory’ oxygen concentration,” *Biochemistry*, vol. 23, no. 1, pp. 1–7, 2009.

[9] Z. Cai, G. Liu, J. Zhang, and Y. Li, “Development of an activity-directed selection system enabled significant improvement of the carboxylation efficiency of Rubisco,” *Protein Cell*, vol. 5, no. 7, pp. 552–562, 2014.

[10] M. Scheffen *et al.*, “A new-to-nature carboxylation module to improve natural and synthetic CO2 fixation,” *Nat. Catal.*, vol. 4, no. 2, pp. 105–115, 2021.

[11] M. K. M. Engqvist and K. S. Rabe, “Applications of protein engineering and directed evolution in plant research,” *Plant Physiol.*, vol. 179, no. 3, pp. 907–917, 2019.

[12] L. Tong, “Structure and function of biotin-dependent carboxylases,” *Cell Mol Life Sci*, vol. 23, no. 1, pp. 1–7, 2013.

[13] X. Liu *et al.*, “Characterization and directed evolution of propionyl-CoA carboxylase and its application in succinate biosynthetic pathway with two CO2 fixation reactions,” *Metab. Eng.*, vol. 62, no. April, pp. 42–50, 2020.

[14] C. Z. Blanchard, A. Chapman-Smith, J. C. Wallace, and G. L. Waldrop, “The biotin domain peptide from the biotin carboxyl carrier protein of Escherichia coli acetyl-CoA carboxylase causes a marked increase in the catalytic efficiency of biotin carboxylase and carboxyltransferase relative to free biotin,” *J. Biol. Chem.*, vol. 274, no. 45, pp. 31767–31769, 1999.

[15] D. T. H. Kondo, Shingo U, F. Moriuchi, J. Sunamoto, S. Ogushi, “Synthesis and Enzymatic Carboxylation of a Biotin-containing Peptide Representing the Coenzyme Binding Site of E. coli Acetyl-CoA Carboxylase.” The Chemical Society of Japan, pp. 1176–1180, 1983.

[16] P. Dimroth, R. B. Guchhait, E. Stoll, and M. D. Lane, “Enzymatic carboxylation of biotin: molecular and catalytic properties of a component enzyme of acetyl CoA carboxylase.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 67, no. 3, pp. 1353–1360, 1970.

[17] Y. Shen, C. Y. Chou, G. G. Chang, and L. Tong, “Is Dimerization Required for the Catalytic Activity of Bacterial Biotin Carboxylase?,” *Mol. Cell*, vol. 22, no. 6, pp. 807–818, 2006.

[18] T. C. Broussard *et al.*, “The three-dimensional structure of the biotin carboxylase-biotin carboxyl carrier protein complex of E. coli acetyl-CoA carboxylase,” *Structure*, vol. 21, no. 4, pp. 650–657, 2013.

[19] C. Y. Chou, L. P. C. Yu, and L. Tong, “Crystal structure of biotin carboxylase in complex with substrates and implications for its catalytic mechanism,” *J. Biol. Chem.*, vol. 284, no. 17, pp. 11690–11697, 2009.

[20] M. D. L. R. B. Guchhait, S. E. Polakis, P. Dimroth, E. Stoll, J. Moss, “Acetyl Coenzyme A Carboxylase System of Escherichia coli,” *J. Biol. Chem.*, vol. 249, no. 20, pp. 6633–6645, 1974.
[21] V. Sloane and G. L. Waldrop, “Kinetic characterization of mutations found in propionic acidemia and methylcrotonylglycinuria: Evidence for cooperativity in biotin carboxylase,” *J. Biol. Chem.*, vol. 279, no. 16, pp. 15772–15778, 2004.
[22] P. V. Attwood and B. D. L. A. Graneri, “Bicarbonate-dependent ATP cleavage catalysed by pyruvate carboxylase in the absence of pyruvate,” *Biochem. J.*, vol. 287, no. 3, pp. 1011–1017, 1992.
[23] I. Climent and V. Rubio, “ATPase activity of biotin carboxylase provides evidence for initial activation of HCO3- by ATP in the carboxylation of biotin,” *Arch. Biochem. Biophys.*, vol. 251, no. 2, pp. 465–470, 1986.
[24] H. Z. R. E. Cobb, R. Chao, “Directed Evolution: Past, Present, and Future Ryan,” *AIChE J.*, vol. 59, no. 4, pp. 215–228, 2013.
[25] C. A. Tracewell and F. H. Arnold, “Directed enzyme evolution: climbing fitness peaks one amino acid at a time,” *Curr. Opin. Chem. Biol.*, vol. 13, no. 1, pp. 3–9, 2009.
[26] M. Janasch and E. P. Hudson, “CO2 fixation gets a second chance,” *Nat. Catal.*, vol. 4, no. 2, pp. 94–95, 2021.
[27] P. A. Dalby, “Strategy and success for the directed evolution of enzymes,” *Curr. Opin. Struct. Biol.*, vol. 21, no. 4, pp. 473–480, 2011.
[28] N. E. Labrou, “Random Mutagenesis Methods for In Vitro Directed Enzyme Evolution,” *Curr. Protein Pept. Sci.*, vol. 999, no. 999, pp. 1–12, 2009.
[29] S. Kille *et al.*, “Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis,” *ACS Synth. Biol.*, vol. 2, no. 2, pp. 83–92, 2013.
[30] Z. Wu, S. B. Jennifer Kan, R. D. Lewis, B. J. Wittmann, and F. H. Arnold, “Machine learning-assisted directed protein evolution with combinatorial libraries,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 18, pp. 8852–8858, 2019.
[31] S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, and G. M. Church, “Low-N protein engineering with data-efficient deep learning,” *bioRxiv*, 2020.
[32] M. Caplow and M. Yager, “Studies on the Mechanism of Biotin Catalysis. II,” *J. Am. Chem. Soc.*, vol. 89, no. 17, pp. 4513–4521, 1967.
[33] D. G. Gibson, L. Young, R. Y. Chuang, J. C. Venter, C. A. Hutchison, and H. O. Smith, “Enzymatic assembly of DNA molecules up to several hundred kilobases,” *Nat. Methods*, vol. 6, no. 5, pp. 343–345, 2009.
[34] E. Rembeza, “High-throughput protein production,” *Adv. Technol. Biosci. Chalmers Univ. student Pap.*, 2020.
[35] C. Z. Blanchard, Y. M. Lee, P. A. Frantom, and G. L. Waldrop, “Mutations at four active site residues of biotin carboxylase abolish substrate-induced synergism by biotin,” *Biochemistry*, vol. 38, no. 11, pp. 3393–3400, 1999.
[36] B. Heil, J. Ludwig, H. Lichtenberg-Fraté, and T. Lengauer, “Computational recognition of potassium channel sequences,” *Bioinformatics*, vol. 22, no. 13, pp. 1562–1568, 2006.
[37] M. K. M. Engqvist and J. Nielsen, “ANT: Software for Generating and Evaluating Degenerate Codons for Natural and Expanded Genetic Codes,” *ACS Synth. Biol.*, vol. 4, no. 8, pp. 935–938, 2015.