microPublication BIOLOGY

10/24/2025 - *Open Access*

Synthetic *Manihot esculenta* Rubisco activase proteins with increased thermotolerance identified via machine learning

Clayton Dilks[1§], Rhiannon LaVine[1], Claire Buchanan[1], Daniel Russo[1], Elizabete Carmo-Silva[2]

[1]Evozyne Inc, Chicago, Illinois, United States

[2]Environment Centre, Lancaster University, Lancaster, England, United Kingdom

[§]To whom correspondence should be addressed: claydilks@yahoo.com

Abstract

Adaptation to increasing environmental temperatures is essential to plant survival and human food production. Thermal tolerance is controlled by a complex network of factors in plants including but not limited to genetic variation and environmental context. Rubisco activase (Rca) is a key photosynthetic enzyme with low thermal tolerance. Here, we report a large machine learning-directed screen of >1,400 synthetic cassava Rca enzymes which identified mutations that convey increased thermal stability while minimizing introduced mutations. We demonstrate multiple synthetic proteins that maintain activity at 8°C higher than wildtype cassava Rca including a single mutation that retains most activity post heat-shock.

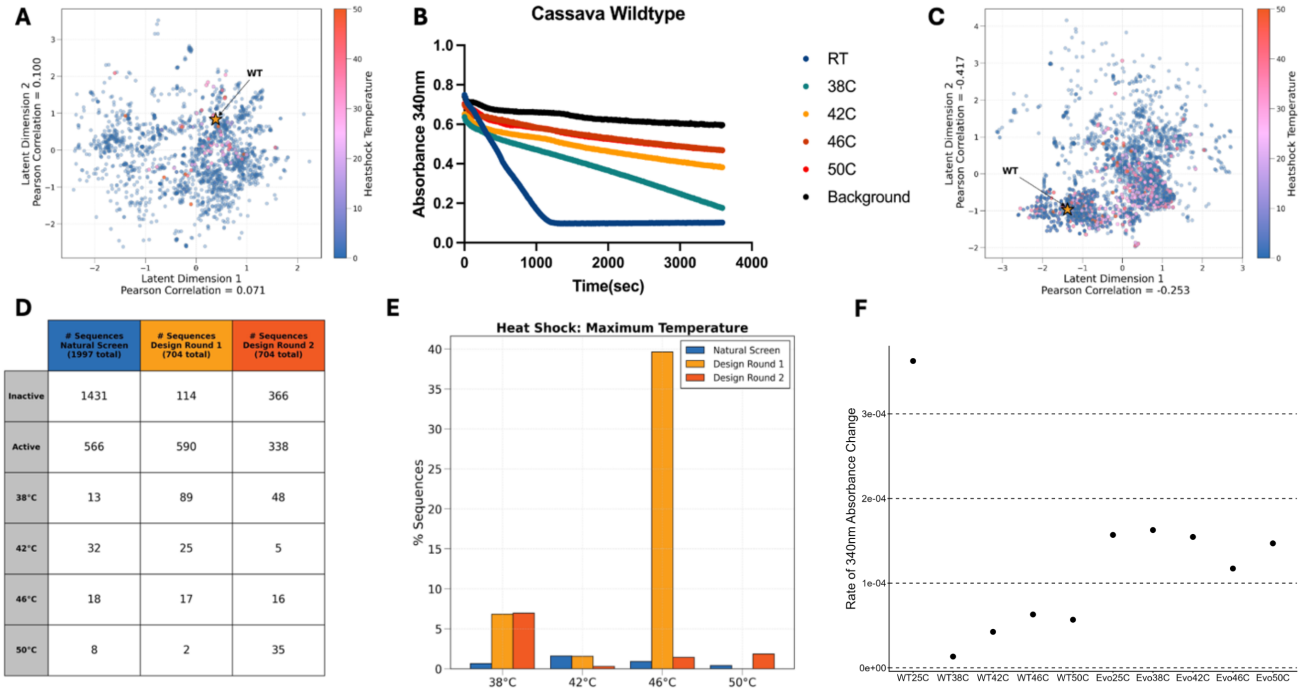

Figure 1. Thermotolerant synthetic Rubisco activase (Rca) proteins with minimal mutations identified via a large machine learning-directed screen.:

(A) Latent space showing two out of four latent dimensions with each Rca sequence color indicative of the highest thermal challenge under which it maintained ATPase activity. The wildtype cassava (WT) is indicated with a gold star. This latent space has not been trained on any data so low correlation is observed. (B) Wildtype Cassava Rca activity at the temperature indicated in the legend (the 46°C and 50°C traces are overlapping). The decrease in absorbance at 340 nm is proportional to the rate of ATP hydrolysis by Rca. (C) Latent space after training using Variational Autoencoder (VAE) deep generative models again showing two of the four latent dimensions each point representing a sequence and the color indicating the maximum temperature at which residual activity was observed. (D) The number of Rca sequences tested in each screening/design round that were active or inactive at room temperature, and that maintained activity at each heat shock temperature. (E) The percentage of sequences which maintained activity up to each heat shock temperature across each screening/design round. (F) The ATP hydrolysis activity of our top performing sequence compared to the wildtype cassava Rca sequence (Rca) with the heat shock temperature and sequence indicated on the x-axis.

Description

Food insecurity is a major problem for millions of people globally. Recent estimates suggest that over 800 million people do not have a secure supply of food (Qu et al., 2023). As climate change is predicted to decrease the production of many essential crops around the world, food insecurity is expected to grow. These crops include wheat (Qu et al., 2023), maize (Li et al., 2022), rice (Saud et al., 2022), cassava (Pipitpukdee et al., 2020), and many others. Cassava is an essential staple in the diet of millions of people in equatorial regions (Tize et al., 2021). Cassava grows optimally between 25-29°C and, although it can withstand temperatures up to 40°C, the yield is greatly reduced.

A promising avenue for crops is the engineering of Rubisco and its associated chaperones. Rubisco is one of the most abundant proteins in the world and is essential for conversion of atmospheric CO_2 into energy within plants, algae and cyanobacteria. A known limitation of Rubisco is the propensity for the complex to be inhibited by the incorrect binding of sugar phosphates (Mueller-Cajar, 2017). As temperatures rise, the concentration of inhibitory molecules also increases, which causes an overall decrease in photosynthetic rate (Waheeda et al., 2023). Rubisco activase (Rca), a chaperone of Rubisco, functions by hydrolyzing ATP and tying the energy released to the removal of the inhibitory molecules from Rubisco catalytic sites (Waheeda et al., 2023). Previous research has shown that the introduction of a more thermostable Rca increases the plant's ability to withstand higher temperatures (Kurek et al., 2007; Kumar et al., 2009). We aim to increase the thermal tolerance of the *Manihot esculenta* (cassava) Rca with three rounds of machine learning-directed engineering paired with a high-throughput assay which measures ATPase activity after a thermal challenge.

There are four Rca genes (XP_021625935.1, XP_021625936.1, XP_021624073.1, and XP_021628250.1) encoded in the cassava genome. XP_021625935.1 and XP_021625936.1 are splice variants that are identical besides a C-terminal extension involved in redox regulation. XP_021628250.1 encodes a more divergent and shorter protein than the other three Rca proteins. XP_021624073.1 is the gene we chose to focus our engineering efforts on because it does not have splice variants to consider, and it readily expressed in *E. coli* compared to the other cassava Rcas. All these genes were used as query sequences to retrieve orthologs to populate a multiple sequence alignment (MSA). The PSIBLAST query yielded 3,915 unique sequences (Schäffer et al., 2001; Altschul et al., 1997). 245,710 additional unique sequences were retrieved from the Mgnify database (Richardson et al., 2023) to supplement the dataset. Next, the dataset was refined to contain only sequences with greater than 20% identity to a reference sequence. This filter reduced the set to 2,573 sequences with a minimum length of 140 amino acids. Denmark Technical University's TargetP 2.0 model (Armenteros et al., 2019) was used to predict chloroplast transit peptides (cTP) in the sequences. 1,852 sequences were identified to contain a cTP which was removed prior to the preliminary alignment with FAMSA (Deorowicz et al., 2016). Further quality assurance and trimming steps were performed on the sequences (details in methods) before a final set of 1,997 sequences was chosen for initial temperature screening, referred to herein as the "natural screen" round.

Variational Autoencoders (VAEs) are deep generative models composed of an encoder and a decoder. The encoder compresses complex protein sequence data into a low-dimensional latent space, which the decoder then uses to reconstruct the input sequences. During training, VAEs use variational inference to optimize a loss function that balances reconstruction accuracy and regularizes the latent space with a prior distribution (Kingma and Welling, 2013, 2019; Zhao et al., 2019). An additional semi-supervised prediction layer can be included to predict experimentally measured targets, which enables the model to be iteratively re-trained with new data from synthesized sequences (Castro et al., 2018; Gómez-Bombarelli et al., 2018; Frassek et al., 2021; Lian et al., 2022). This approach allows the latent space to capture key patterns of amino acid mutations that characterize the sequence set of interest and provide an interpretable, low-dimensional embedding to facilitate design of novel sequences with desirable functionality (Sevgen et al., 2023). An unsupervised VAE was trained on the MSA, resulting in no strong latent space organization with respect to thermal tolerance of the Rca sequences (Figure 1A). Gaussian sampling was performed to generate synthetic Rca variants within 20 hamming distance of the primary wildtype (XP_021624073.1) to further characterize the sequence of interest and provide training data for design round 1.

The respective DNA sequences were obtained from IDT, cloned, expressed recombinantly, and the proteins were purified in high throughput (see methods). After purification, we used an ATPase activity assay that couples the production of ADP to the oxidation of NADH to NAD+ which can be monitored through the decrease in absorbance at 340 nm (Barta et al., 2011). At 5 µM the wildtype Rca sequence demonstrated robust activity at room temperature (RT) but lost activity when subject to a heat shock at 42°C or higher temperatures (Figure 1B (the 46°C and 50°C traces are overlapping)). We found 566 sequences from the natural screen that showed ATPase activity at room temperature (Figure 1C/D). We also performed a series of temperature challenges to identify the thermal tolerance of each Rca sequence tested. For these tests, we placed 20 µL of purified protein into a 384-well plate and placed the plate into a pre-equilibrated thermocycler (38°C, 42°C, 46°C, and 50°C) for one hour. After temperature challenges, the samples were centrifuged (2,000 rcf for 2 min) and then 10 µL of sample was moved into a plate pre-loaded with 40 µL of assay buffer and absorbance was measured after ~16 hours in assay buffer. This permissive screen was performed to give sequences with low activity the opportunity to complete the reaction if any active protein remained post heat shock. We found that 92 sequences remained active after any of these temperatures, 18 of which were active above 46°C. Sequential design rounds (Design Round 1 and Design Round 2) proceeded with training a semi-supervised VAE using ATPase activity data. The resultant latent spaces showed

more thermotolerance-related organization (Figure 1C), allowing for data-informed generation of synthetic Rca sequences with the desired thermal tolerant phenotype.

After the model and sequence generation were completed, these sequences were tested for activity as described above with one additional screening step. These sequences were concentration-normalized prior to thermal challenges. We first measured the concentration with the high-throughput microfluidic system (Revvity LabChip). After the concentration was measured, the samples were diluted in elution buffer to 5 µM of protein in 6 µL total with the ECHO liquid handling system. Sequences with a concentration <5 µM were run at the highest possible concentration. This set of sequences showed an 83.8% active rate (590/704 tested) and showed activity even at the highest tested temperatures. We identified 19 sequences with activity after a 46°C temperature challenge, two of which were active after 50°C. These data were included in the semi-supervised VAE and a second design round was generated.

The sequences generated in design round two were screened in the same manner as design round one. We found 35 additional sequences that had activity after a 50°C heat shock, some of which retained close to 100% activity after the challenge at the highest temperature (Figure 1D/E/F). A sequence of particular interest is designated as evozyne_rca-1 (EVO in Figure 1F), this sequence has only a single mutation from the wildtype sequence with a proline to glycine swap at position 250 after transit peptide removal. A notable observation with this sequence is that increasing thermal stability has led to a loss in overall activity, a trade off which has been previously documented (Figure 1F) (Degen et al., 2020; Vanella et al., 2024). The remaining sequences from design rounds 1 and 2 as well as the natural round are included in Data table 1, along with if each sequence showed activity and the highest temperature at which the protein retained activity.

One difficulty of engineering enzymes with increased thermostability is that we currently do not know the level of Rca activity needed *in planta*. If only a small amount of activity is enough to cause a large improvement in crop stability, this increase may be enough to improve crop yield. If activity equivalent to the wildtype Rca enzyme is required, additional engineering on the thermally stable sequences should be performed to maintain the stability gains while reintroducing the level of wildtype activity. Overall, we believe the engineered sequences presented here represent an excellent starting point for further engineering of cassava to maintain high yields in the face of climate change.

Methods

Reagents

All chemicals including adenosine 5’ triphosphate (ATP), magnesium chloride (MgCl2), phosphoenyol-pyruvate (PEP), potassium chloride (KCL), polyethylene glycol (PEG), pyruvate-kinase (PK), lactate dehydrogense (LDH) lysozyme, DNAse, β-nicotinamide adenine dinucleotide reduced disodium salt (NADH), and imidazole were purchased from Sigma Aldrich. Other reagents were purchased from GoldBio or Thermo Fisher. Cloning and transformation reagents are purchased mainly from New England BioLabs and Teknova.

Cloning

A proprietary codon optimization algorithm was used to optimize the nucleic acid codon usage of the sequences for expression in *E. coli* BL21 (DE3) cells. After optimization, 5’ and 3’ adapters were added to each sequence which included cloning sites for introduction into a modified pDV4 plasmid using golden-gate cloning with BsaI. This cloning design put sequences in frame with a 6x HIS tag for purification and a HRV 3C protease site for tag removal. The sequences were then cloned into the modified pDV4 plasmid and transformed into BL21 (DE3) *E. coli*. The chaperone plasmid pGro7, which encodes the groEL and groES chaperones under an arabinose inducible plasmid, was used to assist in protein folding.

Protein Expression

Protein expression was initiated by inoculating the transformed cells in autoinduction media with selective antibiotics (6.7 g/L $NaHPO_4$, 3.4 g/L KH_2PO_4, 20 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl, 5 g/L dextrose, 20 g/L lactose, 6% w/v of 100% glycerol, 0.5 mg/mL arabinose, 100µg/mL carbenicillin, and 6.25µg/mL chloramphenicol) (Grabski & Drott, 2005). Cultures were grown at 37°C with shaking at 800RPM in 96-deep well plates until the optical density (OD600) of the culture reached ~0.8. The temperature was then dropped to 20°C with continuous shaking overnight for ~16 hours. After expression the plates were centrifuged at 4000 rcf for 10 minutes at 4°C, decanted, and then frozen at –80°C. Plates were thawed on ice and the cells were lysed using a SoluLyse-based system (0.10 µL/mL Lysonase, 5 mM DTT, 1 mM PMSF, 5 mM $MgCl_2$, 300 mM NaCl, 2 mM ATP) with manual pellet resuspension. The samples were clarified by centrifugation (60 minutes at 4°C, 5500 rcf).

Ni-NTA Protein Purification

Cell pellets were gently re-suspended in 120 µL of lysis buffer (50mM sodium phosphate, 90mM NaCl, 0.15mg/mL DNAse, 0.8mg/mL Lysozyme, 10mM MgCl2, 1mM PMSF, 2mM ATP, pH 7.8), transferred to 96-well PCR plates and held on ice for 10 minutes. PCR plates were then centrifuged at 5500 rcf, 4°C, for 60 minutes to pellet cellular debris in

the sample. The clarified lysate was transferred to a deep well plate pre-loaded with 25uL of charged Ni-NTA magnetic beads (Genscript L00295) using the VIAFLOW liquid handler. Plates were then placed in a 4°C incubator, 800RPM for 60 minutes to bind the protein to the charged Ni-NTA beads. After this binding step, the plate is placed on a magnetic plate to allow removal of clarified lysate from the beads. To each well, 500 µL of wash buffer (1X PBS pH 7.8, 200mM NaCl, 50mM Imidazole, 2mM ATP, 3mM MgCl2) was added to wash the beads with an 800RPM, 4°C, 10 minutes incubation. The wash buffer is aspirated out by placing the deep well plate on the magna-rack. This wash step was repeated twice. The elution step consisted of 45 µL of elution buffer (1X PBS pH 7.8, 200mM NaCl, 50mM Imidazole, 2mM ATP, 3mM MgCl2) with an incubation time of 10 minutes, 4°C. Once the elution step is completed, the plate is placed again on the magna-rack and the eluted protein solutions transferred into a fresh 96-well PCR plate.

ATPase activity assay

For testing Rca activity using the ATPase assay, a reaction mixture of 45 µL containing 100mM Tris-KOH, 10mM $MgCl_2$, 20mM KCL, 5mM DTT, 2mM PEP, 5% w/v PEG, 300uM NADH, 5U/mL PK, 5.75U/mL LDH, and 2mM ATP is loaded into a 384-well optical plate using the VIAFLOW liquid handler (Barta et al., 2011). 5 µL of each eluted protein was added to the respective wells on the optical plate and mixed briefly with the liquid handler. The plate was quickly spun on a tabletop centrifuge for 30 seconds and read on a plate reader at 340 nm. The reaction couples ATP to ADP conversion to NADH to NAD+ oxidation using pyruvate kinase and lactate dehydrogenase. The NADH oxidation rate is continuously monitored by measuring the absorbance at 340 nm.

Acknowledgements: We would like to acknowledge the Gates' foundation for funding this research effort. Additionally, we would like to thank Dr. Adam Zmyslowski for his help with the Rubisco purification protocol. We would also like to thank Dr. Tom Speltz who came up with the idea for the collaboration with the Gates' foundation.

Extended Data

Description: Dataset containing sequence IDs, sequences, information on each sequence, and activity data used throughout manuscript. . Resource Type: Dataset. File: 20251004_Data_table.csv. DOI: 10.22002/kpejh-m5258

References

Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402. DOI: 10.1093/nar/25.17.3389

Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, Nielsen H. 2019. Detecting sequence signals in targeting peptides using deep learning. Life Science Alliance 2: e201900429. DOI: 10.26508/lsa.201900429

Barta C, Carmo-Silva AE, Salvucci ME. 2010. Rubisco Activase Activity Assays. Methods in Molecular Biology,Photosynthesis Research Protocols : 375-382. DOI: 10.1007/978-1-60761-925-3_29

Castro E, Godavarthi A, Rubinfien J, Givechian K, Bhaskar D, Krishnaswamy S. 2022. Transformer-based protein generation with regularized latent space optimization. Nature Machine Intelligence 4: 840-851. DOI: 10.1038/s42256-022-00532-1

Degen GE, Worrall D, Carmo-Silva E. 2020. An isoleucine residue acts as a thermal and regulatory switch in wheat Rubisco activase. The Plant Journal 103: 742-751. DOI: 10.1111/tpj.14766

Deorowicz S, Debudaj-Grabysz A, Gudyś A. 2016. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. Scientific Reports 6: 10.1038/srep33964. DOI: 10.1038/srep33964

Frassek M, Arjun A, Bolhuis PG. 2021. An extended autoencoder model for reaction coordinate discovery in rare event molecular dynamics datasets. The Journal of Chemical Physics 155: 10.1063/5.0058639. DOI: 10.1063/5.0058639

Gomez Bombarelli R, Wei JN, Duvenaud D, Hernandez Lobato JM, Sanchez Lengeling B, Sheberla D, et al., Aspuru Guzik A. 2018. undefined. ACS Central Science. 4: 268. DOI: 10.1021/acscentsci.7b00572

Kingma DP, Welling M. 2013. Auto-Encoding Variational Bayes. arXiv e-prints: arXiv:1312.6114. DOI: 10.48550/arXiv.1312.6114

Kingma DP, Welling M. 2019. An Introduction to Variational Autoencoders. DOI: 10.1561/2200000056

Kumar A, Li C, Portis AR. 2009. Arabidopsis thaliana expressing a thermostable chimeric Rubisco activase exhibits enhanced growth and higher rates of photosynthesis at moderately high temperatures. Photosynthesis Research 100: 143-153. DOI: 10.1007/s11120-009-9438-y

Kurek I, Chang TK, Bertain SM, Madrigal A, Liu L, Lassner MW, Zhu G. 2007. Enhanced Thermostability of *Arabidopsis* Rubisco Activase Improves Photosynthesis and Growth Rates under Moderate Heat Stress. The Plant Cell 19:

3230-3241. DOI: 10.1105/tpc.107.054171

Li K, Pan J, Xiong W, Xie W, Ali T. 2022. The impact of 1.5 °C and 2.0 °C global warming on global maize production and trade. Scientific Reports 12: 10.1038/s41598-022-22228-7. DOI: 10.1038/s41598-022-22228-7

Lian X, Praljak N, Subramanian SK, Wasinger S, Ranganathan R, Ferguson AL. 2022. Deep learning-enabled design of synthetic orthologs of a signaling protein. : 10.1101/2022.12.21.521443. DOI: 10.1101/2022.12.21.521443

Mueller-Cajar O. 2017. The Diverse AAA+ Machines that Repair Inhibited Rubisco Active Sites. Frontiers in Molecular Biosciences 4: 10.3389/fmolb.2017.00031. DOI: 10.3389/fmolb.2017.00031

Richardson L, Allen B, Baldi G, Beracochea M, Bileschi ML, Burdett T, et al., Finn. 2022. MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Research 51: D753-D759. DOI: 10.1093/nar/gkac1080

Pipitpukdee S, Attavanich W, Bejranonda S. 2020. Impact of Climate Change on Land Use, Yield and Production of Cassava in Thailand. Agriculture 10: 402. DOI: 10.3390/agriculture10090402

Qu Y, Mueller-Cajar O, Yamori W. 2022. Improving plant heat tolerance through modification of Rubisco activase in C3 plants to secure crop yield and food security in a future warming world. Journal of Experimental Botany 74: 591-599. DOI: 10.1093/jxb/erac340

Saud S, Wang D, Fahad S, Alharby HF, Bamagoos AA, Mjrashi A, et al., Hassan. 2022. Comprehensive Impacts of Climate Change on Rice Production and Adaptive Strategies in China. Frontiers in Microbiology 13: 10.3389/fmicb.2022.926059. DOI: 10.3389/fmicb.2022.926059

Schaffer AA. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Research 29: 2994-3005. DOI: 10.1093/nar/29.14.2994

Sevgen E, Moller J, Lange A, Parker J, Quigley S, Mayer J, et al., Ferguson. 2023. ProT-VAE: Protein Transformer Variational AutoEncoder for Functional Protein Design. : 10.1101/2023.01.23.525232. DOI: 10.1101/2023.01.23.525232

Tize I, Fotso AK, Nukenine EN, Masso C, Ngome FA, Suh C, et al., Hanna. 2021. New cassava germplasm for food and nutritional security in Central Africa. Scientific Reports 11: 10.1038/s41598-021-86958-w. DOI: 10.1038/s41598-021-86958-w

Vanella R, Küng C, Schoepfer AA, Doffini V, Ren J, Nash MA. 2024. Understanding activity-stability tradeoffs in biocatalysts by enzyme proximity sequencing. Nature Communications 15: 10.1038/s41467-024-45630-3. DOI: 10.1038/s41467-024-45630-3

Waheeda K, Kitchel H, Wang Q, Chiu PL. 2023. Molecular mechanism of Rubisco activase: Dynamic assembly and Rubisco remodeling. Frontiers in Molecular Biosciences 10: 10.3389/fmolb.2023.1125922. DOI: 10.3389/fmolb.2023.1125922

Zhao S, Song J, Ermon S. 2019. InfoVAE: Balancing Learning and Inference in Variational Autoencoders. Proceedings of the AAAI Conference on Artificial Intelligence 33: 5885-5892. DOI: 10.1609/aaai.v33i01.33015885

Funding: This work was funded by the Bill and Melinda Gates Foundation.

Author Contributions: Clayton Dilks: conceptualization, data curation, project administration, supervision, writing - original draft. Rhiannon LaVine: data curation, formal analysis, software, visualization. Claire Buchanan: methodology, investigation. Daniel Russo: investigation, validation. Elizabete Carmo-Silva: visualization, writing - review editing.

Reviewed By: Anonymous

History: Received July 29, 2025 **Revision Received** October 4, 2025 **Accepted** October 22, 2025 **Published Online** October 24, 2025 **Indexed** November 7, 2025

Citation: Dilks C, LaVine R, Buchanan C, Russo D, Carmo-Silva E. 2025. Synthetic *Manihot esculenta* Rubisco activase proteins with increased thermotolerance identified via machine learning. microPublication Biology. 10.17912/micropub.biology.001773