UC Davis
UC Davis Electronic Theses and Dissertations

Title

Applications of Protein Engineering in Synthetic Biology and Bioproduction

Permalink

https://escholarship.org/uc/item/8p08541h

Author

Minami, Shiaki

Publication Date

2023

Peer reviewed|Thesis/dissertation

Applications of Protein Engineering in Synthetic Biology and Bioproduction

By

SHIAKI MINAMI
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Priya Shah, Chair

Karen McDonald

Somen Nandi

Committee in Charge

2023

Copyright Acknowledgments

Chapter 2 contains material originally published in BMC Biotechnology:

Minami, S.A., Shah, P.S. Transient light-activated gene expression in Chinese hamster ovary cells. *BMC Biotechnol* **21**, 13 (2021). https://doi.org/10.1186/s12896-021-00670-1

Chapter 3 contains material originally submitted to Biotechnology Journal and bioRxiV:

Minami, S. A., & Shah, P. S. (2023). *Computational evaluation of light propagation in cylindrical bioreactors for optogenetic mammalian cell cultures* (p. 2023.02.01.526707). bioRxiv. https://doi.org/10.1101/2023.02.01.526707

Chapter 4 contains material originally published in Biotechnology Journal.

Minami, S. A., Jung, S., Huang, Y., Harris, B. S., Kenaston, M. W., Faller, R., Nandi, S., McDonald, K. A., & Shah, P. S. (2022). Production of novel SARS-CoV-2 Spike truncations in Chinese hamster ovary cells leads to high expression and binding to antibodies. *Biotechnology Journal*, *17*(9), 2100678. https://doi.org/10.1002/biot.202100678

Abstract

The use of engineered protein systems has become highly prevalent across a wide range of biological fields. In synthetic biology, engineered proteins and protein systems offer precise methods for regulation of cellular behavior, enabling high sensitivity, tunability, and temporal resolution. In bioproduction, proteins are engineered to have enhanced functional properties or other characteristics that facilitate production, purification, and storage. The work in this dissertation explores some of this vast landscape through several chapters that discuss the implementation of optogenetics in Chinese Hamster Ovary (CHO) cells and computationally guided design and production of higher performing severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) truncations.

In synthetic biology, many genetic, chemical, and environmental approaches have been developed to modulate cellular pathways to improve titers. However, these methods are often irreversible or have off-target effects. Development of synthetic biology techniques which are precise, tunable, and reversible will facilitate temporal regulation of target pathways to maximize titers and protein quality. In this study, we investigate the use of optogenetics in CHO cells. Chinese hamster ovary (CHO) cells are widely used for industrial production of biopharmaceuticals. The light-activated CRISPR-dCas9 effector (LACE) system was first transiently transfected to express eGFP in a light-inducible manner. Then, a stable system was tested using lentiviral transduction. Transient transfections resulted in increasing eGFP expression as a function of LED intensity, and fluorescence decreased once the LACE system was deactivated.

Optogenetic control of cellular pathways and gene circuits in mammalian cells is a new frontier in mammalian genetic engineering, and protein engineering has significantly increased the performance of these systems. As a low-cost, tunable, and reversible input, light is highly adept at spatiotemporal, orthogonal regulation of cellular behavior, advancing applications such as protein bioproduction and cultivated meat. However, light is absorbed and scattered as it travels through media and cells, and the applicability of optogenetics in larger mammalian bioreactors has not been determined. In this work, we computationally explore the size limit to which optogenetics can be applied in cylindrical bioreactors at relevant height-to-diameter ratios for mammalian cell culture. We model the propagation of light using the radiative transfer equation and consider changes in reactor volume, absorption coefficient, scattering coefficient, and scattering anisotropy. We observed sufficient light penetration for activation for bioreactor sizes of up to 80,000 L with maximal cell densities, with decreasing efficiency for larger bioreactors. For a 100,000 L bioreactor, we determined that lower cell densities of up to $1.5 \cdot 10^7$ cells/mL can be supported. We conclude that optogenetics can be applied to bioreactors at an industrial scale and may be a valuable tool for specific biomanufacturing applications.

Rather than using engineered proteins to regulate cellular behavior, protein engineering can also be performed to directly modify proteins of interest for enhanced bioproduction. During the SARS-CoV-2 pandemic, protein engineering has played a major role in Spike protein bioproduction. Spike is a key protein that mediates viral entry into cells and elicits antibody responses. Its importance in infection, diagnostics, and vaccinations has created a large demand for purified Spike for diagnostic, clinical and research applications. Spike is difficult to express, prompting modifications to the protein and expression platforms to improve yields. Alternatively, Spike receptor binding domain (RBD) is commonly expressed with higher titers, though it has lower sensitivity in serological assays. Engineered Spike proteins for higher stability have greatly increased expression levels, which is critical for rapid, cost-effective production. We first improve

transient Spike expression in Chinese hamster ovary (CHO) cells and demonstrate that Spike titers increase significantly over longer expression periods compared to RBD. Next, we developed 8 Spike truncations in pursuit of a truncation with both high expression and antibody binding. Truncations were designed such that the RBD sequence was conserved, and truncation points do not interrupt major secondary structures. Two truncations had higher expression than RBD, and one truncation had higher affinity to antibodies than did Spike. Binding of one truncation, T1, to ACE2-Fc was comparable to that of RBD, and N-linked glycosylation profiles resembled those of RBD and Spike very closely. These results suggest T1 is a promising Spike alternative for use in various applications.

Table of Contents

Acknowledgments

This work would not have been possible without the support of countless people throughout my Ph.D. journey. First, I would like to thank my advisor, Priya. This journey would not have been possible without your guidance and unwavering support for my scientific and professional aspirations. Thank you for allowing me to, at times, go on a tangent to explore other paths unpaved. I am grateful for your patience and the trust you have placed in me. My experience has been immeasurably brightened by your infectious cheerfulness and unwavering optimism.

I am deeply grateful to my closest collaborators and mentors Dr. Roland Faller, Dr. Karen McDonald, and Dr. Somen Nandi. You have truly been an inspiration to me, and it has been a privilege and opportunity to work with you. Reflecting on my Ph.D. trajectory, working with you was a pivotal process that has clearly shaped and developed me into the scientist I am today. Your mentorship and advice throughout the years has been invaluable.

I would like to express my gratitude to everyone in the Shah Lab for a positive, supportive, and laughter-filled lab environment. I am especially grateful to my mentees Kush Patel, Haley Heinemann, Nathaniel Burmas, Susannah Schaffer, and Shruthi Garimella for the privilege to work alongside you. Your dedication, hard work, and enthusiasm have been a constant source of inspiration to me, and this work would not have been possible without you. I would also like to thank Oanh Pham, Adam Fishburn, Ritika Gangaraju, Neil Adia, Matthew Kenaston, and Eshan Thilakaratne for your companionship both inside and outside the lab. Thank you all for the memories.

I extend my heartfelt appreciation to my close friends Nitin Sai Beesabathuni and Timothy Hui for their unwavering camaraderie throughout this journey together. You were an inspiration to continuously pursue excellence in science, and all the moments of laughter have made my time here all the more enjoyable. Thank you both for being such wonderful companions on this journey.

Finally, thank you to my family for supporting me throughout my life. Mom and Dad, thank you for all the sacrifices you have made for me. Had you not made the choice to stay in this country, my life would probably have unfolded very differently. Thank you for unconditionally supporting my passions. To my brother, Hataka, thank you for always believing in me. Your constant presence and support have been a source of comfort and encouragement.

Chapter 1: *Introduction*

Engineered protein systems are ubiquitous throughout many biological fields of study. The ability to tune functional properties or stability of proteins, add entirely new functions, or change the targets which the engineered proteins modulate has proven to be extremely powerful in applications such as advancing our knowledge of cell biology, developing synthetic biology systems, facilitating bioproduction, or understanding protein-protein interactions (PPI) during pathogenesis[1–4]. The applications of protein engineering are almost limitless.

The ability to engineer proteins relies heavily on the idea that the function of a protein is derived from its structure, which, in turn, is thought to be determined from its amino acid sequence[5–8]. Many computational tools have leveraged this idea to predict structures of proteins from their sequence alone[9–11]. Recent developments in protein structural prediction, such as AlphaFold2, AlphaFold-Multimer, and RoseTTAFold2NA, have shown major advancements in the ability to predict protein, protein complexes, or even protein-nucleic acid complex structures[12–14]. Alongside computational design of proteins, significant advances in experimental methods have also accelerated protein engineering. Techniques such as directed evolution have allowed rapid improvements in protein function, using methods such as surface display[15–18]. Other library methods such as deep mutational scanning have also accelerated our understanding of the roles of single amino acids in proteins[19]. In addition to high-throughput experimentation, the widespread use of cryogenic electron microscopy has led to the resolution of many more protein structures[20]. Not only can these structures be used directly for analysis, the expansion of protein databanks has also contributed greatly to the increased accuracy of structural prediction algorithms such as AlphaFold2.

Our understanding of the cell benefits greatly from the use of engineered proteins, which can provide specific outputs that provide valuable insights into cellular pathways. For instance, fluorescent proteins can be conjugated to other proteins to observe their expression and localization in the cell, which provides information on protein, protein complex, or organelle function and dynamics[21]. Additional information may also be gained by using biosensor systems, such as Förster resonance energy transfer systems that provide information on protein interactions[22]. These tools can be leveraged to advance our knowledge of cellular behavior, which is fundamental to all subsequent applications that rely on it, such as biotechnology, bioproduction, and medicine.

In this dissertation, two applications of protein engineering will be discussed. First, the use of an optogenetic tool comprised of a clustered regularly interspaced short palindromic repeats (CRISPR)-based optogenetics system is explored in mammalian cells. Second, the development of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Spike truncations for higher titers and antibody binding is discussed.

Synthetic Biology

In the context of synthetic biology, protein engineering enables the development of tools that specifically and precisely regulate gene circuits, cellular pathways, and protein-protein interactions. The advent of synthetic biology stems from major developments in molecular biology technologies, such as polymerase chain reaction, which allows amplification of genes and generation of mutants through site-directed mutagenesis[23,24]. Combined with the concept of

synthetic biological circuits introduced two decades ago[25,26], synthetic biology revolves around utilizing proteins, gene circuits, or cellular pathways that do not exist in nature.

CRISPR and optogenetics systems are both emerging synthetic biology tools that expand our ability to manipulate cells. CRISPR was originally used by prokaryotes as an antiviral system and was first used in synthetic biology as a gene editing tool. Though other gene editing methods existed previously such as applications of transcription activator-like effector nucleases and zinc finger nucleases, CRISPR offers higher efficiency and ease of use in many situations[27–30]. Optogenetics is another technique that adds to the synthetic biology toolbox. From light-sensitive cells that convert photon energy to chemical energy, their light-responsive proteins are expressed recombinantly in cells that are not naturally light-responsive[31,32]. This enables light induced regulation of cellular behavior without disrupting other pathways. Many different light-responsive proteins exist, such as phytochromes, cryptochromes, and light-oxygen-voltage-sensing domains[33]. When the light-sensitive proteins are illuminated with their activation wavelengths, the photon energy is absorbed, and the protein undergoes a conformational change. This leads to increased or decreased homo/hetero dimerization or oligomerization, and this basic property of proteins can be used to manipulate vast sets of gene expression or signaling pathways in many different ways[34–38].

Properties of optogenetic proteins, such as the kinetics of activation and deactivation, affinities to binding partners, and stability, can be altered through protein engineering. Controlling the kinetics of activation and deactivation increases the ease of use, as some systems may require faster on/off switching, while others may require stable on/off states, with minimal intervention. Optogenetic systems typically have activation and deactivation times from the order of

milliseconds to seconds, and these can be increased or decreased through truncations or mutations[39,40]. Binding affinities to binding partners can be regulated as well, such as through protein modifications that influence the binding interface[41]. Finally, stability of proteins can be increased by performing stabilizing mutations, or adding degrons, which promote protein degradation through proteasomal and autophagy pathways[42].

In the LACE system, the light-sensitive proteins cryptochrome 2 (CRY2) and CRY2-interacting bHLH 1(CIB) and the CRISPR Cas9 protein have been engineered to enhance transcription regulation[43]. Full-length CRY2 and its truncation have been fused to four tandem copies of Herpes Simplex Viral Protein 16 activation domain (VP64), which activates transcription from a minimal promoter[44]. For CIB, the N-terminal region of the protein (CIBN) was fused to catalytically deactivated Cas9 (dCas9), with the permutation dCas9-CIBN, CIBN-dCas9, or CIBN-dCas9-CIBN. The combination of full-length CRY2 fused to VP64 and CIBN-dCas9-CIBN led to the highest dynamic range, which is the ratio of expression level following activation to the expression level in the dark, inactivated state[43]. The mechanism of the LACE system for transcription activation will be described in detail in Chapter 2.

While the LACE system has been tested in HEK293T cells for mammalian cells, the applicability of the LACE system in other cell lines is unclear. The structure and function of proteins are dependent on the cell type and species, as protein folding is affected by the presence of chaperones, post-translational modifications, and other factors in the local environment[45–50]. The functionality of the system also depends on the transcription factors that are present in the host, which may influence the extent to which leaky expression in the dark may occur and extent to which activation occurs. The utility of the LACE system could be expanded to bioproduction if its

functionality is preserved in a model organism for bioproduction such as Chinese Hamster Ovary (CHO) cells, and if the light propagates sufficiently in industrial-scale systems[51–54].

Bioproduction

The second application of protein engineering covered in this dissertation is the design of protein truncations for bioproduction of SARS-CoV-2 Spike protein[55]. SARS-CoV-2 is an RNA virus that is decorated on its surface with the Spike membrane glycoprotein. The importance of Spike stems from its ability to cause viral infection of the cell. Spike is comprised of the S1 and S2 subunits, which are responsible for binding to the host receptor angiotensin converting enzyme 2 (ACE2) and for membrane fusion for entry into the cell, respectively. The COVID-19 pandemic has highlighted the critical need for rapid, large-scale, cost-effective production of proteins from the high demand of Spike. Spike production is essential for protein-based vaccines, diagnostic assays such as serology assays, and research efforts utilizing purified Spike to understand viral processes such as infection and immune system evasion[56–58].

Spike is a difficult protein to express, which may at least in part be caused by its large size and complex structure[59–61] Larger proteins have higher metabolic burden for expression, and both its size and complexity may lead to higher probability for incomplete folding or misfolding, which redirects proteins to degradation pathways. Its complex structure, including the formation of disulfide bonds, post-translational modifications through N-linked and O-linked glycosylation, and trimerization, may also increase the propensity for incorrect structures to form[62,63]. In particular, glycosylation is an important characteristic that has been shown to increase binding affinity of the receptor binding domain (RBD) of Spike to ACE2[64]. Glycans also modulate the RBD

5

conformational dynamics, balancing the availability of the RBD for interaction with ACE2 and potential recognition from the host immune system[65,66].

Given these characteristics of Spike, several engineering efforts have been undertaken to improve its expression levels in various hosts. First, Spike is a membrane protein, but high expression is favored by soluble, secreted proteins, and obviating the need to extract proteins from cell lysates facilitates purification processes. To create soluble Spike, the transmembrane domain has been removed, and secretion tags have been added onto the N-terminus of the protein. To retain the ability of the protein to trimerize without the transmembrane domain, a trimerization domain may be added to the C-terminus[67,68]. To produce the prefusion form of Spike, which is the form prior to cleavage at a furin site, the furin site may also be mutated to prevent cleavage. For purification of Spike, a C-terminal affinity tag can be utilized[69].

Next, mutations have been introduced in Spike to increase expression. Namely, the goal is to increase Spike stability such that protein degradation is reduced, and titers are subsequently increased. These mutations must be performed such that interactions with binding partners such as ACE2 and antibodies are not affected. A popular method to increase Spike stability is to introduce two proline substitutions for loop stabilization, which was previously effective for Middle East respiratory syndrome coronavirus and SARS-CoV-1[58,70,71]. Prolines have constrained dihedral angles compared to other amino acids, and its introduction can increase the stability of a protein fold by reducing the conformational entropy of the backbone[72,73].

Apart from mutations, the amino acid sequence can also be modified through truncations. Expressing truncations instead of a full-length protein have the potential to increase titers by

reducing the metabolic burden to create additional peptide chains, simplifying the folding and post-translational modification procedure, and eliminating regions of protein that may be contributing to its inherent difficulty for expression. For Spike, the RBD is an important region that mediates interactions with ACE2 and is targeted by many antibodies. Thus, the RBD alone has been expressed with a secretion tag and purification tag, which has led to titers that are an order of magnitude greater than that of Spike[67]. One drawback, however, is that the RBD has weaker binding to antibodies.

The RBD alone is less biologically relevant, as it does not exist in nature, but studies utilizing RBD are still incredibly useful for understanding interactions mechanisms, and purified RBD is useful as an antigen. Truncations have not been as extensively explored as stabilizing mutations. In Chapter 4, we discuss the design, expression, and characterization of 8 Spike truncations in pursuit of a truncation that has high expression like RBD and high binding to antibodies like full-length Spike.

References

1. Ebo, J. S., Guthertz, N., Radford, S. E., & Brockwell, D. J. (2020). Using protein engineering to understand and modulate aggregation. *Current Opinion in Structural Biology*, *60*, 157–166. https://doi.org/10.1016/j.sbi.2020.01.005

2. Foo, J. L., Ching, C. B., Chang, M. W., & Leong, S. S. J. (2012). The imminent role of protein engineering in synthetic biology. *Biotechnology Advances*, *30*(3), 541–549. https://doi.org/10.1016/j.biotechadv.2011.09.008

3. Bojar, D., & Fussenegger, M. (2020). The Role of Protein Engineering in Biomedical Applications of Mammalian Synthetic Biology. *Small*, *16*(27), 1903093. https://doi.org/10.1002/smll.201903093

4. Luke, J. (n.d.). *Protein Engineering Impending Implications in Synthetic Biology*.

5. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). Analyzing Protein Structure and Function. *Molecular Biology of the Cell. 4th Edition*. https://www.ncbi.nlm.nih.gov/books/NBK26820/

6. Spronk, C. A. E. M., Nabuurs, S. B., Krieger, E., Vriend, G., & Vuister, G. W. (2004). Validation of protein structures derived by NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *45*(3–4), 315–337. https://doi.org/10.1016/j.pnmrs.2004.08.003

7. Skolnick, J., Kolinski, A., Brooks, C. L., Godzik, A., & Rey, A. (1993). A method for predicting protein structure from sequence. *Current Biology*, *3*(7), 414–423. https://doi.org/10.1016/0960-9822(93)90348-R

8. Marks, D. S., Hopf, T. A., & Sander, C. (2012). Protein structure prediction from sequence variation. *Nature Biotechnology*, *30*(11), Article 11. https://doi.org/10.1038/nbt.2419

9. Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols*, *5*(4), Article 4. https://doi.org/10.1038/nprot.2010.5

10. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, *10*(6), Article 6. https://doi.org/10.1038/nprot.2015.053

11. Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M., Pieper, U., & Sali, A. (2006). Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *0 5*, Unit-5.6. https://doi.org/10.1002/0471250953.bi0506s15

12. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), Article 7873. https://doi.org/10.1038/s41586-021-03819-2

13. *Protein complex prediction with AlphaFold-Multimer | bioRxiv*. (n.d.). Retrieved March 29, 2023, from https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2

14. Baek, M., McHugh, R., Anishchenko, I., Baker, D., & DiMaio, F. (2022). *Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA* (p. 2022.09.09.507333). bioRxiv. https://doi.org/10.1101/2022.09.09.507333

15. Arnold, F. H. (1998). Design by Directed Evolution. *Accounts of Chemical Research*, *31*(3), 125–131. https://doi.org/10.1021/ar960017f

16. Schreuder, M. P., Brekelmans, S., van den Ende, H., & Klis, F. M. (1993). Targeting of a heterologous protein to the cell wall of Saccharomyces cerevisiae. *Yeast (Chichester, England)*, *9*(4), 399–409. https://doi.org/10.1002/yea.320090410

17. Gai, S. A., & Wittrup, K. D. (2007). Yeast surface display for protein engineering and characterization. *Current Opinion in Structural Biology*, *17*(4), 467–473. https://doi.org/10.1016/j.sbi.2007.08.012

18. Park, M. (2020). Surface Display Technology for Biosensor Applications: A Review. *Sensors (Basel, Switzerland)*, *20*(10), 2775. https://doi.org/10.3390/s20102775

19. Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: A new style of protein science. *Nature Methods*, *11*(8), 801–807. https://doi.org/10.1038/nmeth.3027

20. Bai, X., McMullan, G., & Scheres, S. H. W. (2015). How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*, *40*(1), 49–57. https://doi.org/10.1016/j.tibs.2014.10.005

21. Shaner, N. C., Steinbach, P. A., & Tsien, R. Y. (2005). A guide to choosing fluorescent proteins. *Nature Methods*, *2*(12), Article 12. https://doi.org/10.1038/nmeth819

22. Liu, L., He, F., Yu, Y., & Wang, Y. (2020). Application of FRET Biosensors in Mechanobiology and Mechanopharmacological Screening. *Frontiers in Bioengineering and Biotechnology*, *8*. https://www.frontiersin.org/articles/10.3389/fbioe.2020.595497

23. Schochetman, G., & Ou, C.-Y. (2023). *Polymerase Chain Reaction*.

24. Carter, P. (1986). Site-directed mutagenesis. *Biochemical Journal*, *237*(1), 1–7.

25. Gardner, T. S., Cantor, C. R., & Collins, J. J. (2000). Construction of a genetic toggle switch in Escherichia coli. *Nature*, *403*(6767), Article 6767. https://doi.org/10.1038/35002131

26. Kobayashi, H., Kærn, M., Araki, M., Chung, K., Gardner, T. S., Cantor, C. R., & Collins, J. J. (2004). Programmable cells: Interfacing natural and engineered gene networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(22), 8414–8419. https://doi.org/10.1073/pnas.0402940101

27. Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, *8*(11), Article 11. https://doi.org/10.1038/nprot.2013.143

28. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, *337*(6096), 816–821. https://doi.org/10.1126/science.1225829

29. Joung, J. K., & Sander, J. D. (2013). TALENs: A widely applicable technology for targeted genome editing. *Nature Reviews Molecular Cell Biology*, *14*(1), Article 1. https://doi.org/10.1038/nrm3486

30. Laity, J. H., Lee, B. M., & Wright, P. E. (2001). Zinc finger proteins: New insights into structural and functional diversity. *Current Opinion in Structural Biology*, *11*(1), 39–46. https://doi.org/10.1016/S0959-440X(00)00167-6

31. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., & Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience*, *8*(9), Article 9. https://doi.org/10.1038/nn1525

32. Deisseroth, K., Feng, G., Majewska, A. K., Miesenböck, G., Ting, A., & Schnitzer, M. J. (2006). Next-Generation Optical Technologies for Illuminating Genetically Targeted Brain Circuits. *The Journal of Neuroscience*, *26*(41), 10380–10386. https://doi.org/10.1523/JNEUROSCI.3863-06.2006

33. Christie, J. M., Gawthorne, J., Young, G., Fraser, N. J., & Roe, A. J. (2012). LOV to BLUF: Flavoprotein Contributions to the Optogenetic Toolkit. *Molecular Plant*, *5*(3), 533–544. https://doi.org/10.1093/mp/sss020

34. Sheets, M. B., Tague, N., & Dunlop, M. J. (2023). An optogenetic toolkit for light-inducible antibiotic resistance. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-36670-2

35. Zhang, X., Pang, G., Sun, T., Liu, X., Pan, H., Zhang, Y., Liu, J., Chang, J., Wang, H., & Liu, D. (2023). A red light-controlled probiotic bio-system for in-situ gut-brain axis regulation. *Biomaterials*, *294*, 122005. https://doi.org/10.1016/j.biomaterials.2023.122005

36. Andersen, T., Wörthmüller, D., Probst, D., Wang, I., Moreau, P., Fitzpatrick, V., Boudou, T., Schwarz, U. S., & Balland, M. (2023). Cell size and actin architecture determine force generation in optogenetically activated cells. *Biophysical Journal*, *122*(4), 684–696. https://doi.org/10.1016/j.bpj.2023.01.011

37. Crellin, H. A., & Buckley, C. E. (2023). Using Optogenetics to Investigate the Shared Mechanisms of Apical-Basal Polarity and Mitosis. *Cells Tissues Organs*, 1–19. https://doi.org/10.1159/000528796

38. Rojas, V., & Larrondo, L. F. (2023). Coupling Cell Communication and Optogenetics: Implementation of a Light-Inducible Intercellular System in Yeast. *ACS Synthetic Biology*, *12*(1), 71–82. https://doi.org/10.1021/acssynbio.2c00338

39. Motta-Mena, L. B., Reade, A., Mallory, M. J., Glantz, S., Weiner, O. D., Lynch, K. W., & Gardner, K. H. (2014). An optogenetic gene expression system with rapid activation and

deactivation kinetics. *Nature Chemical Biology*, *10*(3), Article 3. https://doi.org/10.1038/nchembio.1430

40. Packer, A. M., Roska, B., & Häusser, M. (2013). Targeting neurons and photons for optogenetics. *Nature Neuroscience*, *16*(7), Article 7. https://doi.org/10.1038/nn.3427

41. Berndt, A., Lee, S. Y., Wietek, J., Ramakrishnan, C., Steinberg, E. E., Rashid, A. J., Kim, H., Park, S., Santoro, A., Frankland, P. W., Iyer, S. M., Pak, S., Ährlund-Richter, S., Delp, S. L., Malenka, R. C., Josselyn, S. A., Carlén, M., Hegemann, P., & Deisseroth, K. (2016). Structural foundations of optogenetics: Determinants of channelrhodopsin ion selectivity. *Proceedings of the National Academy of Sciences*, *113*(4), 822–829. https://doi.org/10.1073/pnas.1523341113

42. Varshavsky, A. (2019). N-degron and C-degron pathways of protein degradation. *Proceedings of the National Academy of Sciences*, *116*(2), 358–366. https://doi.org/10.1073/pnas.1816596116

43. Polstein, L. R., & Gersbach, C. A. (2015). A light-inducible CRISPR/Cas9 system for control of endogenous gene activation. *Nature Chemical Biology*, *11*(3), 198–200. https://doi.org/10.1038/nchembio.1753

44. Beerli, R. R., Dreier, B., & Barbas, C. F. (2000). Positive and negative regulation of endogenous genes by designed transcription factors. *Proceedings of the National Academy of Sciences*, *97*(4), 1495–1500. https://doi.org/10.1073/pnas.040552697

45. Kang, T. S., & Kini, R. M. (2009). Structural determinants of protein folding. *Cellular and Molecular Life Sciences*, *66*(14), 2341–2361. https://doi.org/10.1007/s00018-009-0023-5

46. Chaudhuri, T. K., & Paul, S. (2006). Protein-misfolding diseases and chaperone-based therapeutic approaches. *The FEBS Journal*, *273*(7), 1331–1349. https://doi.org/10.1111/j.1742-4658.2006.05181.x

47. Welch, W. J., & Brown, C. R. (1996). Influence of molecular and chemical chaperones on protein folding. *Cell Stress & Chaperones*, *1*(2), 109–115.

48. Walsh, G. (2010). Post-translational modifications of protein biopharmaceuticals. *Drug Discovery Today*, *15*(17), 773–780. https://doi.org/10.1016/j.drudis.2010.06.009

49. Sarkar, A., & Wintrode, P. L. (2011). Effects of glycosylation on the stability and flexibility of a metastable protein: The human serpin α1-antitrypsin. *International Journal of Mass Spectrometry*, *302*(1–3), 69–75. https://doi.org/10.1016/j.ijms.2010.08.003

50. Shental-Bechor, D., & Levy, Y. (2008). Effect of glycosylation on protein folding: A close look at thermodynamic stabilization. *Proceedings of the National Academy of Sciences*, *105*(24), 8256–8261. https://doi.org/10.1073/pnas.0801340105

51. Pachauri, N., Singh, V., & Rani, A. (2017). Two degree of freedom PID based inferential control of continuous bioreactor for ethanol production. *ISA Transactions*, *68*, 235–250. https://doi.org/10.1016/j.isatra.2017.03.014

52. Stephenson, M., & Grayson, W. (2018). Recent advances in bioreactors for cell-based therapies. *F1000Research*, *7*, F1000 Faculty Rev-517. https://doi.org/10.12688/f1000research.12533.1

53. Asllanaj, F., Contassot-Vivier, S., Hohmann, A., & Kienle, A. (2019). Light propagation in biological tissue. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *224*, 78–90. https://doi.org/10.1016/j.jqsrt.2018.11.001

54. Stolfa, G., Smonskey, M. T., Boniface, R., Hachmann, A.-B., Gulde, P., Joshi, A. D., Pierce, A. P., Jacobia, S. J., & Campbell, A. (2018). CHO-Omics Review: The Impact of Current and Emerging Technologies on Chinese Hamster Ovary Based Bioproduction. *Biotechnology Journal*, *13*(3), 1700227. https://doi.org/10.1002/biot.201700227

55. Hu, B., Guo, H., Zhou, P., & Shi, Z.-L. (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, *19*(3), Article 3. https://doi.org/10.1038/s41579-020-00459-7

56. Esposito, D., Mehalko, J., Drew, M., Snead, K., Wall, V., Taylor, T., Frank, P., Denson, J.-P., Hong, M., Gulten, G., Sadtler, K., Messing, S., & Gillette, W. (2020). Optimizing high-yield production of SARS-CoV-2 soluble spike trimers for serology assays. *Protein Expression and Purification*, *174*, 105686. https://doi.org/10.1016/j.pep.2020.105686

57. Cibelli, N., Arias, G., Figur, M., Khayat, S. S., Leach, K., Loukinov, I., Shadrick, W., Chuenchor, W., Tsybovsky, Y., Gulla, K., & Gowetski, D. B. (2022). Advances in purification of SARS-CoV-2 spike ectodomain protein using high-throughput screening and non-affinity methods. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-07485-w

58. Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C.-L., Abiona, O., Graham, B. S., & McLellan, J. S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science (New York, N.y.)*, *367*(6483), 1260–1263. https://doi.org/10.1126/science.abb2507

59. Writer, G. S. (2015, January 30). *Difficult-to-Express Proteins, Made Easy*. GEN - Genetic Engineering and Biotechnology News. https://www.genengnews.com/magazine-issues/february-1-2015-vol-35-no-3/difficult-to-express-proteins-made-easy/

60. Poothong, J., Jang, I., & Kaufman, R. J. (2021). Defects in protein folding and/or quality control cause protein aggregation in the ER. *Progress in Molecular and Subcellular Biology*, *59*, 115–143. https://doi.org/10.1007/978-3-030-67696-4_6

61. Díaz-Villanueva, J. F., Díaz-Molina, R., & García-González, V. (2015). Protein Folding and Mechanisms of Proteostasis. *International Journal of Molecular Sciences*, *16*(8), 17193–17230. https://doi.org/10.3390/ijms160817193

62. Dorai, H., & Ganguly, S. (2014). Mammalian cell-produced therapeutic proteins: Heterogeneity derived from protein degradation. *Current Opinion in Biotechnology*, *30*, 198–204. https://doi.org/10.1016/j.copbio.2014.07.007

63. Coghlan, J., Benet, A., Kumaran, P., Ford, M., Veale, L., Skilton, St. J., Saveliev, S., & Schwendeman, A. A. (2022). Streamlining the Characterization of Disulfide Bond Shuffling and Protein Degradation in IgG1 Biopharmaceuticals Under Native and Stressed Conditions. *Frontiers in Bioengineering and Biotechnology*, *10*. https://www.frontiersin.org/articles/10.3389/fbioe.2022.862456

64. Huang, Y., Harris, B. S., Minami, S. A., Jung, S., Shah, P. S., Nandi, S., McDonald, K. A., & Faller, R. (2022). SARS-CoV-2 spike binding to ACE2 is stronger and longer ranged due to glycan interaction. *Biophysical Journal*, *121*(1), 79–90. https://doi.org/10.1016/j.bpj.2021.12.002

65. Gong, Y., Qin, S., Dai, L., & Tian, Z. (2021). The glycosylation in SARS-CoV-2 and its receptor ACE2. *Signal Transduction and Targeted Therapy*, *6*(1), Article 1. https://doi.org/10.1038/s41392-021-00809-8

66. Sztain, T., Ahn, S.-H., Bogetti, A. T., Casalino, L., Goldsmith, J. A., Seitz, E., McCool, R. S., Kearns, F. L., Acosta-Reyes, F., Maji, S., Mashayekhi, G., McCammon, J. A., Ourmazd, A., Frank, J., McLellan, J. S., Chong, L. T., & Amaro, R. E. (2021). A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nature Chemistry*, *13*(10), Article 10. https://doi.org/10.1038/s41557-021-00758-3

67. Amanat, F., Stadlbauer, D., Strohmeier, S., Nguyen, T. H. O., Chromikova, V., McMahon, M., Jiang, K., Arunkumar, G. A., Jurczyszak, D., Polanco, J., Bermudez-Gonzalez, M., Kleiner, G., Aydillo, T., Miorin, L., Fierer, D. S., Lugo, L. A., Kojic, E. M., Stoever, J., Liu, S. T. H., … Krammer, F. (2020). A serological assay to detect SARS-CoV-2 seroconversion in humans. *Nature Medicine*, *26*(7), Article 7. https://doi.org/10.1038/s41591-020-0913-5

68. Dalton, A. C., & Barton, W. A. (2014). Over-expression of secreted proteins from mammalian cell lines. *Protein Science : A Publication of the Protein Society*, *23*(5), 517–525. https://doi.org/10.1002/pro.2439

69. Kimple, M. E., Brill, A. L., & Pasker, R. L. (2013). Overview of Affinity Tags for Protein Purification. *Current Protocols in Protein Science*, *73*(1), 9.9.1-9.9.23. https://doi.org/10.1002/0471140864.ps0909s73

70. Pallesen, J., Wang, N., Corbett, K. S., Wrapp, D., Kirchdoerfer, R. N., Turner, H. L., Cottrell, C. A., Becker, M. M., Wang, L., Shi, W., Kong, W.-P., Andres, E. L., Kettenbach, A. N., Denison, M. R., Chappell, J. D., Graham, B. S., Ward, A. B., & McLellan, J. S. (2017). Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proceedings of the*

National Academy of Sciences of the United States of America, *114*(35), E7348–E7357. https://doi.org/10.1073/pnas.1707304114

71. Kirchdoerfer, R. N., Wang, N., Pallesen, J., Wrapp, D., Turner, H. L., Cottrell, C. A., Corbett, K. S., Graham, B. S., McLellan, J. S., & Ward, A. B. (2018). Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Scientific Reports*, *8*, 15701. https://doi.org/10.1038/s41598-018-34171-7

72. Choi, E. J., & Mayo, S. L. (2006). Generation and analysis of proline mutants in protein G. *Protein Engineering, Design & Selection: PEDS*, *19*(6), 285–289. https://doi.org/10.1093/protein/gzl007

73. Prajapati, R. S., Das, M., Sreeramulu, S., Sirajuddin, M., Srinivasan, S., Krishnamurthy, V., Ranjani, R., Ramakrishnan, C., & Varadarajan, R. (2007). Thermodynamic effects of proline introduction on protein stability. *Proteins: Structure, Function, and Bioinformatics*, *66*(2), 480–491. https://doi.org/10.1002/prot.21215

Chapter 2: *Transient light-activated gene expression in Chinese hamster ovary cells*

Introduction

Chinese hamster ovary (CHO) cells are a major industrial workhorse for mammalian glycoprotein production. Improving protein yields through regulation of gene expression is a major focus of CHO cell engineering[1]. Reversible and scalable control of gene expression would represent a major step forward in CHO cell engineering, yet mature methods typically rely on constitutive regulation or on chemical induction that cannot be easily reversed unless the media is replaced[2]. For instance, sodium butyrate is a highly utilized chemical in industrial CHO cell cultures to increase specific productivity[3]. However, the increase in specific productivity is often accompanied by a decrease in cell viability[4]. Removing sodium butyrate towards the end of the culture duration may improve viability while retaining the benefits of increased specific productivity[5], but media replacement is economically disadvantageous in large bioreactors. A system in which stimuli are easily removed would allow greater control over the entire culture period, which is crucial given the dynamic nature of production cultures.

Recently developed systems, such as the light-activated CRISPR-dCas9 effector (LACE), enable regulation of gene expression with a tunable and easily reversed mechanism[6]. Briefly, the LACE system relies on light-activated dimerization of CRY2 and CIBN. CRY2 is a light-sensitive protein which undergoes a conformational change upon excitation with 450nm light. Here, CRY2 is fused to VP64 as CRY2-VP64. VP64 is a viral transactivation domain that activates expression from a minimal CMV promoter. CIBN binds to the activated form of CRY2. In this system, CIBN is fused to dCas9 as CIBN-dCas9-CibN and integrated into a CRISPR dCas-9 system. The guide RNA (gRNA) localizes the CIBN-dCas9-CIBN complex to a minimal CMV promoter. Upon light-

activated dimerization of CRY2 and CIBN, VP64 is recruited to the minimal CMV promoter, activating the expression of a target gene. Once the light is removed, CRY2 and CIBN dissociate, and gene expression is turned off. This is a modular system in which light can be used to regulate the expression of any transgene of interest (Fig 1).
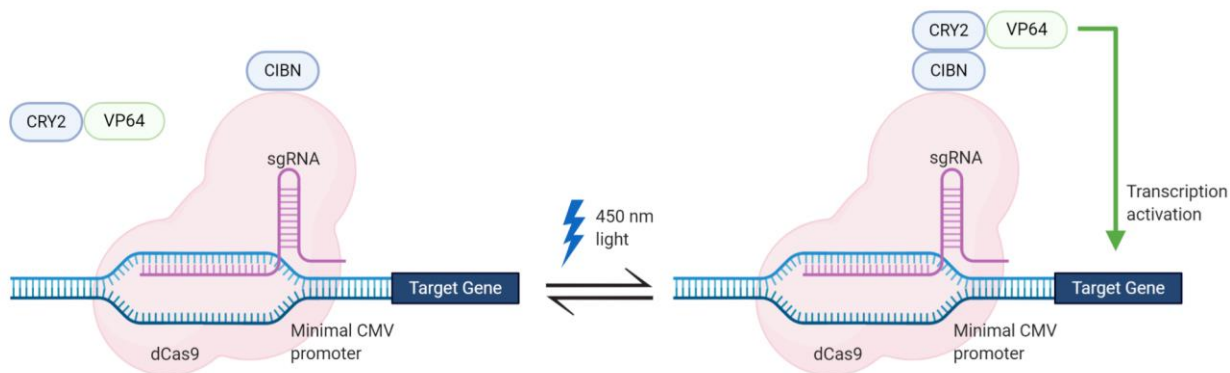
Figure 1. Schematic of LACE system. The LACE system relies on light-activated dimerization of CRY2 and CIBN (blue). CIBN is fused to dCas9 and targeted to the minimal CMV promoter using a synthetic gRNA sequence. CRY2 is fused to VP64 so that it activates transcription following light activation and dimerization with CIBN.

Here, we use this previously described LACE system to control gene expression in CHO cells. We show that transient light-activated expression of enhanced green fluorescent protein (eGFP) is possible through transient transfection of a four-plasmid system, and this system is tunable. However, only a small percentage of cells received adequate amounts of all four plasmids, and the transient nature of the system limits its utility for CHO cell engineering applications. We therefore tested stable expression relying on lentiviral transduction. This system did not result in

light-activated expression of eGFP, suggesting additional modifications are required for stable light-activated gene expression.

Results

LACE activity in CHO cells using transient transfection

We first tested the LACE system in CHO-DG44 (DG44) cells using transient transfection and eGFP as a reporter for expression. The percent of eGFP-positive cells varies with gate position (Appendix A Fig A-1A). At high gate thresholds, the percent of eGFP-positive cells is higher for the activated LACE system and clearly separated from the dark state. This separation decreases as we lower the gating threshold, similar to previous studies using the LACE system in HEK 293T cells [6]. Interestingly, at much lower gate thresholds, the activated LACE system has slightly lower percent eGFP-positive cells. This may be because cells that receive the plasmid encoding eGFP but not all four plasmids required for light activation are capable of low-level eGFP expression that is photobleached by the LEDs used for LACE activation. Given that gating on percent eGFP-positive cells using a high threshold could result in analyzing a very small number of cells in the dark state (<10 cells), we considered mean fluorescence intensity (MFI) as an alternative measure. LACE cells consistently had a higher MFI for cells in the light, regardless of the gate threshold (Appendix A Fig A-1B). Gating for subsequent experiments was performed such that >99.9% of untransfected cells were excluded. This ensures that the most LACE cells are captured for analysis.

Since the minimal CMV promoter can result in leaky transcription, we determined the background level of eGFP expression by transfecting only the plasmid encoding eGFP (8x gRNA eGFP). Background MFI of eGFP signal was slightly lower than that of the full four-plasmid LACE system without light activation. Following light activation for 24 hours, eGFP expression was measured

by flow cytometry. We observed an approximately 4-fold increase in MFI for the population of cells that were gated positive for eGFP expression (Fig 2A).
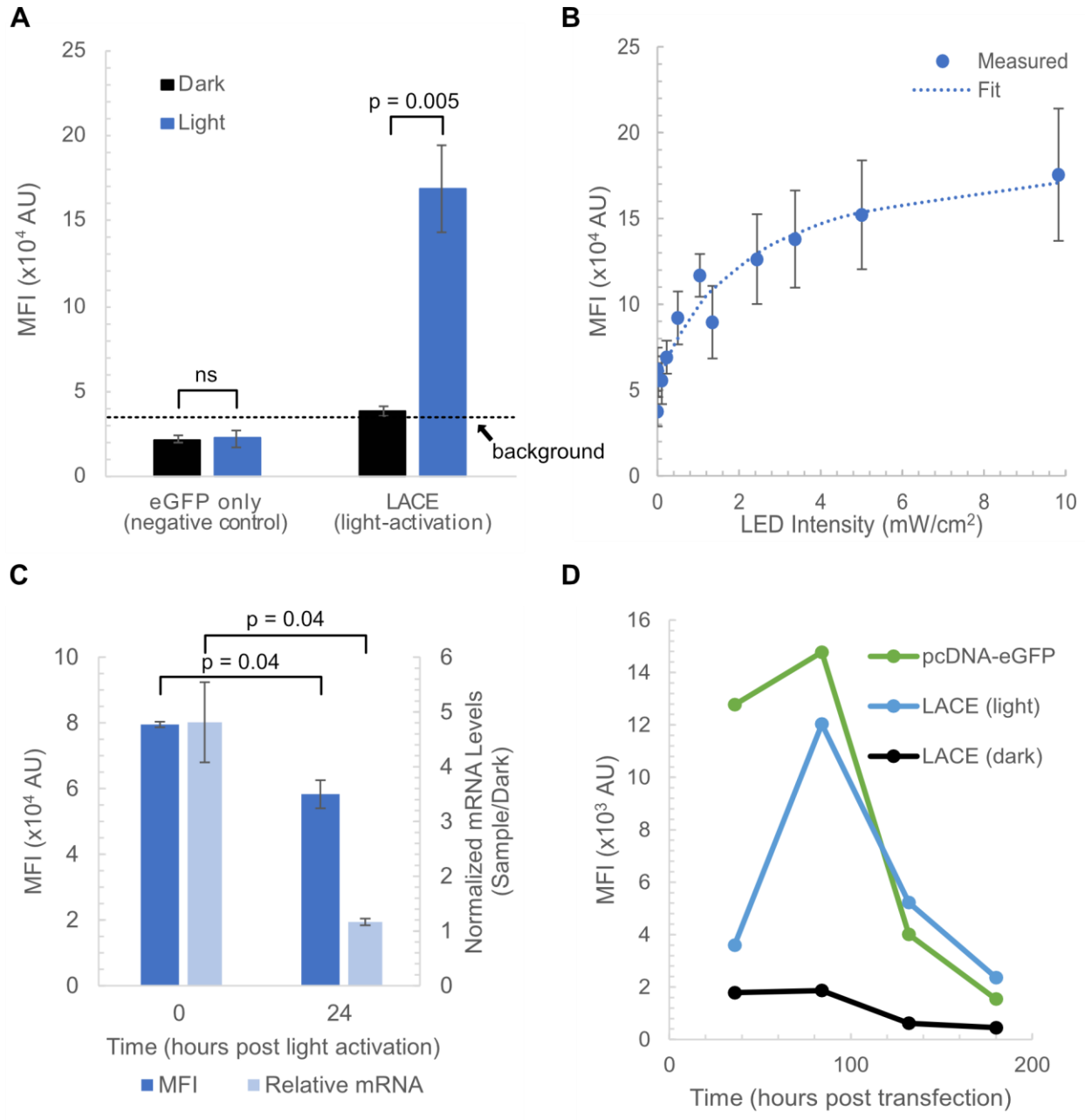

Figure 2. Transient light-active gene expression in CHO cells. The LACE system was used to express eGFP in CHO cells in a light-activated manner. (**A**) DG44 cells were transfected with

LACE plasmids. Mean fluorescence intensity (MFI) of eGFP signal was measured by flow cytometry with (blue) and without (black) 24 hours of light activation with 465 nm light using 1 s pulses at 0.067 Hz pulse frequency and 9.8 mW/cm^2 intensity. eGFP only (negative control) represents the background fluorescence of cells (dashed line). (**B**) Tunability of eGFP expression was measured by altering LED intensity. MFI of eGFP signal from DG44 cells was measured by flow cytometry following 24 hours of light activation at the indicated intensity. A fit of the data was performed assuming Michaelis-Menten kinetics. (**C**) Reversibility of eGFP expression was measured by monitoring eGFP MFI from DG44 and mRNA levels from CHO-K1 cells after turning off light activation. MFI and mRNA were measured by flow cytometry and qRT-PCR, respectively, at 0 and 24 hours after light activation was terminated. (**D**) CHO-K1 cells were transfected with the indicated plasmids. Cells transfected with LACE plasmids were either exposed to light or kept in the dark. MFI of eGFP signal was measured at the indicated times. Data from (**A**) and (**B**) represent the mean values of five and three independent experiments, respectively. Data from (**C**) represent the mean values of three technical replicates in an experiment. Error bars represent standard error of the mean. Background fluorescence is indicated with a dashed line and the upper limit of the 95% confidence interval for negative control samples. P values were calculated with a paired two-tailed T-test.

Once we confirmed that the LACE system functions in CHO cells, we sought to test tunability, reversibility, and duration of light-activation. We tested whether eGFP expression could be tuned in CHO cells using light intensity. We varied light intensity while keeping pulse width and frequency constant. We observed saturating behavior in which more than 80% of maximum MFI of eGFP signal was achieved with only 50% of the maximum light intensity possible (Fig 2B). We also showed that the system is reversible in CHO cells. MFI of eGFP fluorescence and relative eGFP mRNA levels significantly decrease when light activation is terminated (Fig 2C and Appendix A Fig A-2). Using an exponential decay fit of flow cytometry data, the protein half-life of

eGFP intensity was calculated to be ~14.9 hours. This is within the range of commonly reported values in the literature[7,8]. Together, our results show that the LACE system is tunable and reversible in CHO cells. We next tested how long light activation could be maintained in CHO cells using transient transfection. CHO-K1 cells were transfected with the LACE plasmids and subjected to light activation for the indicated times post-transfection or kept in the dark. MFI of eGFP signal was measured by flow cytometry. The decay in MFI of eGFP signal for the LACE system following light activation (light) mirrored the decay by constitutively active eGFP (pCDNA-eGFP). We observed a peak MFI of eGFP signal at 84 hours post-transfection, which was a 6-fold increase compared to control cells kept in the dark. While the signal was still moderately high and well above control cells kept in the dark at 132 hours post-transfection, at 180 hours post-transfection the signal was indistinguishable from control cells kept in the dark (Fig 2D). Another flow cytometer was used to collect the data in Fig 2D, resulting in eGFP intensities which are an order of magnitude lower compared to other data in Figure 2. As the units of fluorescence intensity are arbitrary, flow cytometer acquisition parameters were adjusted to ensure that fluorescence intensity values were within the linear range of detection for both instruments. Together, our results show that the LACE system is tunable and reversible in CHO cells, with light-activated expression viable for up to 5 days post-transfection.

LACE activity in CHO cells using stable lentiviral transduction

Many applications in CHO cells require the generation of stable cell lines and cultures that last more than one week[9,10]. Transient systems are becoming more popular but given the decay of the LACE signal following transfection (Fig 2D), LACE activity would only be accessible in the first five days post-transfection. Moreover, transient transfection of the LACE plasmids would result in only a small percentage of the cells receiving all four plasmids required for LACE activity. For example, even with a transfection efficiency as high as 90%, only 66% (0.9^4) of cells would receive four plasmids of an equimolar mixture. In these cases, stable expression of LACE components

would be desirable. We therefore cloned the LACE components into a lentiviral expression system using Gibson Assembly (Appendix A Fig A-3). Each component was introduced to DG44 by serial transduction. We then split cells into two samples with one sample cultured in the dark and the other sample activated with light for 24 hours. While we did not observe significant light-activation, a small population of cells (~0.1%) had increased fluorescence following light activation (Fig 3A). To test whether this result was cell-line dependent, we tested transient transfection and serial transfection in CHO-K1 cells. Very similar results were obtained for CHO-K1, with robust transient light-activated expression of eGFP (data not shown), but low signal-to-noise for stably integrated cells (Fig 3B). We hypothesized that the population of light-responsive cells was very small because only a small percentage of them received and appropriately expressed all four plasmids using serial transduction. We therefore sorted single CHO-K1 cells in the top 0.1% of fluorescence intensity following light activation to further characterize their response to light. Clones were allowed to recover and expand for more than two weeks before analysis. Surprisingly, light activation of the sorted clones resulted in the same MFI of eGFP signal as the clones without light activation. This almost perfect 1:1 relationship between dark and light for the stably-transduced clones stands in contrast to the approximately 4-fold change in MFI for the transient system and indicates that the sorted clones are not light-responsive (Fig 3C).
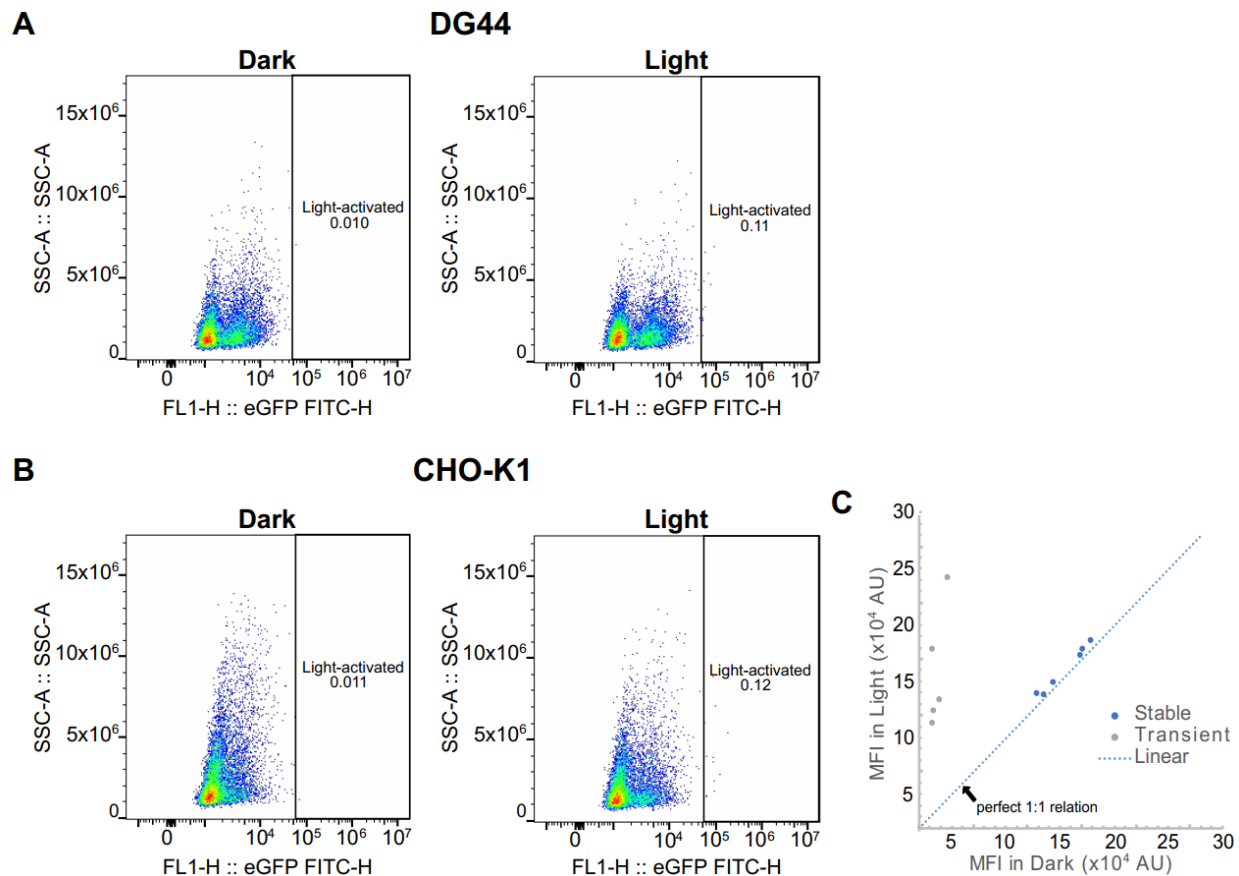
Figure 3. Lentiviral transduction does not result in stable light-activated gene expression. (**A**) DG44 and (**B**) CHO-K1 cells serially transduced with lentiviral vectors encoding the LACE system were incubated in the dark or with light activation for 24 hours. Fluorescence intensity of cells was measured by flow cytometry. (**C**) Cells in the top 0.1% of fluorescence from (**B**) were single-cell sorted and expanded. The surviving clonal populations were incubated in the dark or with light activation for 24 hours. Fluorescence intensity of cells was measured by flow cytometry. Transient transfection data from Fig 2A is included as a reference for robust light activation.

Given the absence of light-activation in the lentiviral system, we sought to determine if each of the lentiviral constructs retained activity. We therefore tested the lentiviral plasmids in a transient transfection setting for responsiveness to light activation. Each of the lentiviral constructs except

24

for the eGFP-encoding construct was tested in combination with the transient transfection-based LACE plasmids. When only one lentiviral construct was included, we observed between 2- and 6-fold increases in MFI of eGFP signal following light activation. The lentiviral construct encoding the gRNA resulted in a light-activated MFI similar to that of the signal of cells left in the dark for other combinations, suggesting that this construct may not be functional. However, in experiments in which two lentiviral constructs were used in tandem with two transient transfection-based LACE plasmids, we still observed robust 4-fold light activation. Interestingly, MFI of eGFP signal for cells kept in the dark increased as we added more lentiviral constructs to the transient transfection. This increase in background fluorescence was not always accompanied with proportional increases in signal following light activation, suggesting that diminishing signal-to-noise ratio for the lentiviral system limits its utility (**Fig 4A**). We were not able to test the lentiviral construct encoding eGFP in the transient transfection system because of its intrinsic fluorescence observed during transient transfection scenarios such as lentiviral packaging (**Fig 4B**).
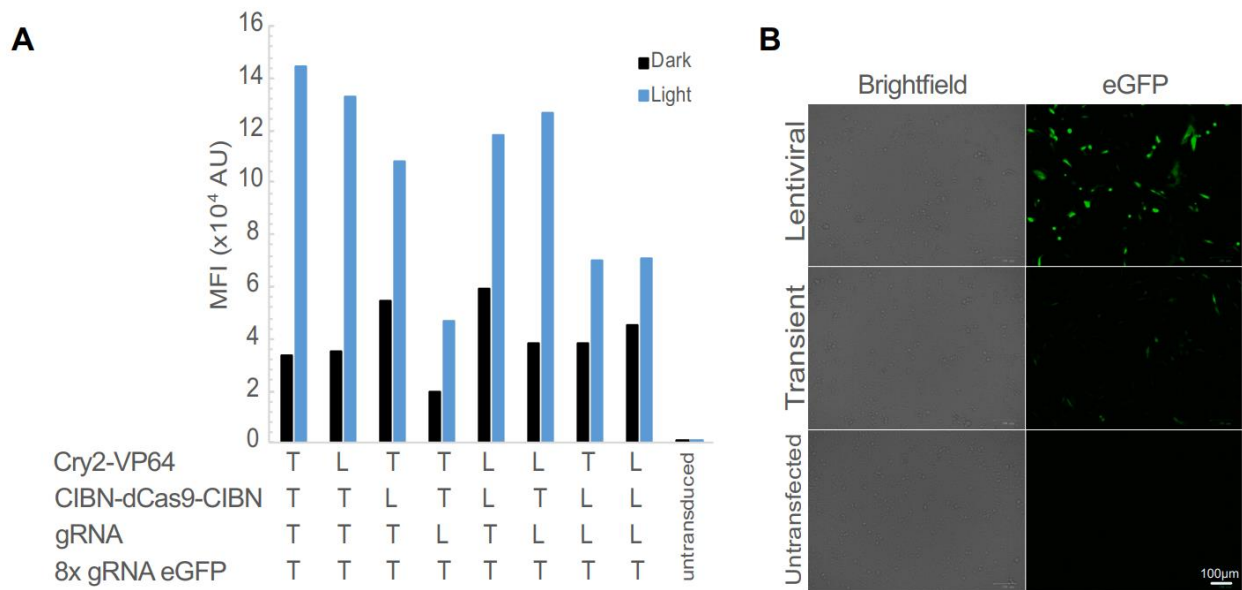
Figure 4. Lentiviral constructs function in a transient setting. (A) The indicated combinations of transient (T) and lentiviral (L) LACE plasmids were transiently transfected into CHO cells and incubated in the dark or with light activation. MFI of eGFP signal was measured by flow cytometry after 24 hours of light activation. The 8xgRNA eGFP lentiviral construct was not tested because of its intrinsic eGFP expression in a transient setting (see **(B)**). **(B)** eGFP expression of the 8x gRNA eGFP lentiviral construct. CHO cells were transfected with the lentiviral or transient 8x gRNA eGFP construct. eGFP and bright field images were taken 24 hours post-transfection.

Discussion

Here we test the LACE system for light-activated gene expression in CHO cells. We find that transient transfection of LACE plasmids results in tunable and reversible light-activated gene expression that can last approximately five days (**Fig 2**). The deactivation of the LACE system is dependent on the half-life of the protein being expressed because the time scale for LACE component interaction is much smaller than the time scale for protein degradation. Additionally, eGFP mRNA degradation is much faster compared to protein degradation (**Fig 2**). Therefore, in this experiment, protein half-life dictates the off-time of the LACE system and if desired, the rate of reversion of the system into the dark state can be increased by using destabilized proteins. High transfection efficiencies are necessary (>95%) because this system relies on co-transfection of four separate plasmids. Moreover, this transient system would not allow for applications involving stable cell lines, or experiments spanning more than one week. We therefore tested the LACE system in lentiviral constructs. We found that the lentiviral LACE system had low signal-to-noise that limited its application (**Figs 3 and 4**). We allowed clones to recover for at least two weeks to allow them to relax back into the dark state, but it is still possible that the previous light activation required to sort individual clones created an epigenetic memory that increases the background fluorescence of these clones, even in a prolonged dark state[11]. Identifying cells that

include all four components of the LACE system using antibiotic selection markers and fewer constructs could be used to circumvent this potential technical barrier.

We hypothesize that the lentiviral LACE system exhibits low signal-to-noise because of increased intrinsic background eGFP expression from the minimal CMV promoter. Our observations of high eGFP expression in lentiviral LACE clones compared to transiently-transfected LACE cells in the dark state (Fig 3C) and high eGFP expression from the 8x gRNA eGFP lentiviral construct compared to the transient plasmid in a transfection setting (Fig 4B) support this hypothesis. Although the LACE expression constructs were expected to behave similarly between the transient and stable settings since the original promoters were also cloned into the lentiviral vectors (Appendix A Fig A-3), it is possible that other regulatory elements of the lentiviral vectors that were also integrated into the cells caused an increase in expression levels in the dark. In this case, a lentiviral system may not be a viable option as a stable LACE system. While less efficient than lentiviral transduction, transient transfection and stable cell line generation via antibiotic selection is another approach. This approach is also more widely accepted than lentiviral transduction for industrial therapeutic protein production. Alternatively, incorporating mammalian origins of replication into all LACE plasmids could allow for episomal plasmid maintenance following transient transfection[12].

Conclusion

In summary, transient light-activated gene expression in CHO cells is tunable and reversible using the LACE system. While stable expression of LACE components would expand their utility, lentiviral transduction of LACE components resulted in increased background and a low signal-to-noise ratio. Other approaches to stable expression of LACE components, such as antibiotic selection of transfected LACE plasmids, could overcome this technical challenge.

Methods

Plasmids

pcDNA3.1-CRY2FL-VP64, pcDNA3.1-CibN-dCas9-CibN, pGL3-Basic-8x-gRNA-eGFP, and gRNA-eGFP-Reporter were gifted from Charles Gersbach (Addgene plasmid # 60554, 60553, 60718, and 60719, respectively). These plasmids were used directly for transient transfection. Lentiviral vectors were constructed from PCR amplification of the LACE components. CRY2FL-VP64 was amplified with primers 1 and 2, CibN-dCas9-CibN with primers 3 and 4, and gRNA-eGFP-Reporter with primers 5 and 6. A synthesized eGFP-IRES-mTagBFP2 fragment was PCR-amplified with primers 7 and 8 (**Table 1**). mTagBFP2 was introduced so that LACE system activation could be monitored once eGFP was replaced with another gene of interest. PCR-amplified fragments were cloned using Gibson Assembly into pFUGW vectors, gifted from Isei Tanida[13] (Addgene plasmid # 61460), which were digested with PacI (New England Biolabs) and EcoRI-HF (New England Biolabs). Diagrams of the transient and lentiviral LACE plasmids are shown in Appendix A Fig A-3.

Table 1. Primer sequences for generation of lentiviral vectors

Primer 1	5'-cagagatccagtttggttaatcgctgcttcgcgatgtacggg-3'
Primer 2	5'-gattatcgataagcttgatatcgtcaagaagcgtagtccggaacgtcgtacg-3'
Primer 3	5'-cagagatccagtttggttaatcgcgatgtacgggccagatatacgc-3'
Primer 4	5'-gattatcgataagcttgatatcgcagtcgaggctgatcagcggtttaaacttaagt-3'
Primer 5	5'-cagagatccagtttggttaatggcgattaagttgggtaacgccaggg-3'
Primer 6	5'-gctccatgtttttctaggtcgaacaaaagctggagctccaccgcg-3'
Primer 7	5'-cagagatccagtttggttaattaagcgccgggctcgagaaac-3'
Primer 8	5'-gattatcgataagcttgatatcgcattaagcttgtgccccagtttgctag-3'

Cells

CHO-K1 cells (ATCC-No. CCL-61) were maintained in IMDM (Gibco) supplemented with 10% FBS (Gibco) and 1% penicillin/streptomycin (Lonza). CHO-DG44 cells[14] were maintained in IMDM supplemented with 10% FBS, 1% penicillin/streptomycin, 13.6 mg/L hypoxanthine (Alfa Aesar), and 3.9 mg/L thymidine (Sigma). HEK293T cells (ATCC-No. CRL-11268) were maintained in DMEM (Gibco) supplemented with 10% FBS and 1% penicillin/streptomycin. Cells were cultured in a humidified incubator at 37 °C and 5% CO_2. Cell density and viability were measured using trypan blue and an automated cell counter (TC20, Bio-Rad).

Transient transfection

At 24 hours before transfection, cells were seeded in 12-well plates at a density 0.12 x 10^6 cells per well. To introduce the LACE system, cells were transfected with 206 ng pcDNA3.1-CRY2FL-VP64, 294 ng pcDNA3.1-CibN-dCas9-CibN, 100 ng pGL3-Basic-8x-gRNA-eGFP, and 400 ng gRNA-eGFP-Reporter. Transfections were performed using PolyJet DNA *In Vitro* Transfection Reagent (SignaGen Laboratories) with a PolyJet/DNA ratio of 2.

Lentiviral packaging and transduction

Lentivirus was packaged in HEK293T cells using sodium phosphate co-transfection of 3.5 ug transgene, 1.8 µg gag/pol-, 1.25 µg vsv-g-, and 0.5 µg rsv-rev-expressing plasmids[15]. Medium exchange was performed 6-12 post-transfection. Medium containing lentivirus was harvested 48-60 hours post-transfection via centrifugation at 1000 rpm for 10 minutes and filtration through 0.45 µm syringe filters. Packaged lentivirus was stored at -80 °C until further use.

For transduction, cells were plated on 12-well plates at a density 0.12 x 10⁶ cells per well and transduced serially with the LACE system components. Cell culture medium was replaced by 250 µL medium containing CRY2FL-VP64 lentivirus and 250 µL fresh medium. After 1 hour, 500 µL fresh medium was added. The following day, cells were passaged back to a density of 0.12 x 10⁶ cells per well to prepare for the next transduction. CIBN-dCas9-CIBN, gRNA-eGFP-Reporter, and 8x-gRNA-eGFP-IRES-BFP were then individually transduced in this manner.

Light activation

24 hours post-transfection, cells were activated via 8mm LEDs (465 nm, L.C. LED). LEDs were mounted on a breadboard, and intensities and pulse frequencies were regulated using an Arduino Uno microcontroller. For all experiments, cells were illuminated with 1-second pulses of light, with a pulse frequency of 0.067 Hz. Unless otherwise specified, LED intensities were 9.8 mW/cm^2. LED intensities were measured using an optical power meter (1931-C, Newport). Samples in the dark were wrapped in aluminum foil to ensure isolation from light.

Flow cytometry

Cells were trypsinized (Gibco) and resuspended in culture medium. Following centrifugation at 300 rcf for 5 minutes at 4 °C, supernatants were discarded, and cell pellets were resuspended in cold DPBS containing 1% FBS. A 488 nm laser was used for all flow cytometry analysis.

BD FACSCanto II was used to analyze LACE system activity over 180 hours. CytoFLEX S Flow Cytometer was used for all other analytical cytometry experiments. Gating for eGFP-positive cells were performed such that >99.9% of untransfected cells were excluded.

Astrios MoFlo EQ was used to sort single cells. Following the activation of serially transduced cells, cells were activated with 465 nm light for 24 hours. Cells with eGFP fluorescence intensities in the top 0.1% of the population were sorted into 96-well plates containing cell culture medium. Surviving colonies were scaled up to 12-well cultures before light-activation and flow cytometry analysis.

qRT-PCR

RNA was extracted from LACE cells activated for 24 hours and from LACE cells that were activated, then deactivated for 24 hours, using Quick-RNA Miniprep Kit (Zymo Research). 400ng of RNA were used to produce cDNA using iScript cDNA Synthesis Kit (Bio-Rad). qPCR reactions were set up using iTaq Universal SYBR Green Supermix (Bio-Rad) and samples were run on a Roche LightCycler 480. Relative quantifications were determined using the Delta-Delta Ct Method[16]. As the control gene, Actin was amplified using primers 9 and 10, and eGFP was amplified using primers 11 and 12 (**Table 2**).

Table 2. Primer sequences for qRT-PCR

Primer 9	5'-gtcgtaccactggcattgtg-3'
Primer 10	5'-agggcaacatagcacagctt-3'
Primer 11	5'-gacaaccactacctgagcac-3'
Primer 12	5'-caggaccatgtgatcgcg-3'

Fitting eGFP MFI vs LED intensity and determining the half-life of eGFP

Production of eGFP was modeled assuming a saturation production rate model with first-order degradation and a constant term for low expression that is observed in the dark:

$$\frac{d[G]}{dt} = \frac{k_1 L}{k_2 + L} - k_3[G] + k_4 \qquad (1)$$

$[G]$ is the mean fluorescence intensity of eGFP, L is the intensity of LED light, k_1 and k_2 are constants describing the saturation, k_3 is the degradation rate constant, and k_4 is the constant for constitutive expression. To fit the parameter scan for eGFP MFI vs. LED intensity, the steady-state solution was used:

$$[G(L)] = \frac{\frac{k_1}{k_3} L}{k_2 + L} + \frac{k_4}{k_3} \qquad (2)$$

The equation was fit using fitnlm in MATLAB, yielding $k_1 = 14.4 \frac{F}{t}$, $k_2 = 2.13 \frac{mW}{cm^2}$, $k_3 = 1.01\, t^{-1}$, and $k_4 = 5.35 \frac{F}{t}$. F is fluorescence intensity of eGFP in arbitrary units.

To calculate protein half-life, the time-dependent equation was solved after setting $L = 0$ to apply the effect of turning the LEDs off:

$$[G(t)] = \frac{k_4}{k_3} + \left([G_0] - \frac{k_4}{k_3}\right) e^{-k_3 t} \qquad (3)$$

$[G_0]$ is the fluorescence intensity of eGFP at the time the LEDs are turned off. Fitting this expression, it was found that $k_1 = 15.0000 \frac{F cm^2}{mW}$, $k_2 = 1.0000\, F$, $k_3 = 1.2581\, t^{-1}$, and $k_4 =$

$6.2724 \frac{F}{t}$. The degradation rate constant, k_3, is similar between both fits performed. The half-life was then calculated with first order degradation kinetics:

$$\frac{\ln 2}{k_3} = t_{\frac{1}{2}} \tag{3}$$

Using k_3 values from the LED intensity and reversibility experiments, half-life of eGFP was determined to be an average of 14.9 hours.

References

1. Majors, B. S., Betenbaugh, M. J., Pederson, N. E., & Chiang, G. G. (2008). Enhancement of transient gene expression and culture viability using Chinese hamster ovary cells overexpressing Bcl-x L. *Biotechnology and Bioengineering*, *101*(3), 567–578. https://doi.org/10.1002/bit.21917

2. Mellahi, K., Cambay, F., Brochu, D., Gilbert, M., Perrier, M., Ansorge, S., Durocher, Y., & Henry, O. (2019). Process development for an inducible rituximab-expressing Chinese hamster ovary cell line. *Biotechnology Progress*, *35*(1), e2742. https://doi.org/10.1002/btpr.2742

3. Sung, Y. H., Song, Y. J., Lim, S. W., Chung, J. Y., & Lee, G. M. (2004). Effect of sodium butyrate on the production, heterogeneity and biological activity of human thrombopoietin by recombinant Chinese hamster ovary cells. *Journal of Biotechnology*, *112*(3), 323–335. https://doi.org/10.1016/j.jbiotec.2004.05.003

4. Jiang, Z., & Sharfstein, S. T. (2008). Sodium butyrate stimulates monoclonal antibody over-expression in CHO cells by improving gene accessibility. *Biotechnology and Bioengineering*, *100*(1), 189–194. https://doi.org/10.1002/bit.21726

5. Wright, J. A. (1973). Morphology and growth rate changes in Chinese hamster cells cultured in presence of sodium butyrate. *Experimental Cell Research*, *78*(2), 456–460. https://doi.org/10.1016/0014-4827(73)90091-8

6. Polstein, L. R., & Gersbach, C. A. (2015). A light-inducible CRISPR-Cas9 system for control of endogenous gene activation. *Nature Chemical Biology*, *11*(3), 198–200. https://doi.org/10.1038/nchembio.1753

7. Danhier, P., Krishnamachary, B., Bharti, S., Kakkad, S., Mironchik, Y., & Bhujwalla, Z. M. (2015). Combining Optical Reporter Proteins with Different Half-lives to Detect Temporal Evolution of Hypoxia and Reoxygenation in Tumors. *Neoplasia*, *17*(12), 871–881. https://doi.org/10.1016/j.neo.2015.11.007

8. Corish, P., & Tyler-Smith, C. (1999). Attenuation of green fluorescent protein half-life in mammalian cells. *Protein Engineering*, *12*(12), 1035–1040. https://doi.org/10.1093/protein/12.12.1035

9. Clincke, M.-F., Mölleryd, C., Zhang, Y., Lindskog, E., Walsh, K., & Chotteau, V. (2011). Study of a recombinant CHO cell line producing a monoclonal antibody by ATF or TFF external filter perfusion in a WAVE Bioreactor™. *BMC Proceedings*, *5*(Suppl 8), P105. https://doi.org/10.1186/1753-6561-5-s8-p105

10. Wang, I. K., Hsieh, S. Y., Chang, K. M., Wang, Y. C., Chu, A., Shaw, S. Y., Ou, J. J., & Ho, L. (2006). A novel control scheme for inducing angiostatin-human IgG fusion protein production using recombinant CHO cells in a oscillating bioreactor. *Journal of Biotechnology*, *121*(3), 418–428. https://doi.org/10.1016/j.jbiotec.2005.07.025

11. D'Urso, A., & Brickner, J. H. (2017). Epigenetic transcriptional memory. In *Current Genetics* (Vol. 63, Issue 3, pp. 435–439). Springer Verlag. https://doi.org/10.1007/s00294-016-0661-8

12. Codamo, J., Munro, T. P., Hughes, B. S., Song, M., & Gray, P. P. (2011). Enhanced CHO cell-based transient gene expression with the Epi-CHO expression system. *Molecular Biotechnology*, *48*(2), 109–115. https://doi.org/10.1007/s12033-010-9351-9

13. Tanida, I., Ueno, T., & Uchiyama, Y. (2014). A super-ecliptic, pHluorin-mKate2, tandem fluorescent protein-tagged human LC3 for the monitoring of mammalian autophagy. *PloS One*, *9*(10), e110600. https://doi.org/10.1371/journal.pone.0110600

14. Chromikova, V., Zaragoza, M. A., & Krammer, F. (2017). Generation of a serum free CHO DG44 cell line stably producing a broadly protective anti-influenza virus monoclonal antibody. *PLoS ONE*, *12*(9). https://doi.org/10.1371/journal.pone.0183315

15. Yu, J. H., & Schaffer, D. V. (2006). Selection of Novel Vesicular Stomatitis Virus Glycoprotein Variants from a Peptide Insertion Library for Enhanced Purification of Retroviral and Lentiviral Vectors. *Journal of Virology*, *80*(7), 3285–3292. https://doi.org/10.1128/jvi.80.7.3285-3292.2006

16. Livak, K. J., & Schmittgen, T. D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2 C T Method. *METHODS*, *25*, 402–408. https://doi.org/10.1006/meth.2001.1262

Chapter 3: *Computational evaluation of light propagation in cylindrical bioreactors for optogenetic mammalian cell cultures*

Introduction

Electromagnetic radiation in the visible range regulates cellular behavior in organisms that utilize light, such as plants, algae, and bacteria, as well as eyes in animals. The responsiveness to light is caused by opsins, which absorb photon energy and undergo conformational changes[1]. In optogenetics and other light-inducible synthetic biology systems, opsins that homo/hetero-dimerize, oligomerize, or undergo conformational changes upon light activation are used to control gene expression or protein activity. Optogenetics has enabled advancements in a variety of fields[2,3], such as regulating and understanding cell motility in cell biology[4], regulating specific pathways such as autophagy in cellular engineering[5], or restoring vision in a blind patient in a biomedical application[6]. The spatiotemporal control and ease of use that optogenetics offers are clear advantages that have been demonstrated at a small scale in mammalian cell culture.

While the applications of optogenetics continue to expand, industrial applications will rely on bioreactors capable of transmitting light, or photobioreactors. Yet, as light travels through media and cells, its intensity decreases through absorption and scattering. Thus, understanding the design constraints of such photobioreactors is critical to establishing the feasibility of optogenetic approaches in industrial applications. In other bioproduction systems such as algae and cyanobacteria, cellular growth in photobioreactors has been studied, taking into consideration the attenuation of light in the bioreactor through absorption and scattering[7–10]. On small scales, optogenetics in model organisms such as *Saccharomyces cerevisiae* and *Escherichia coli* have also been performed on milliliter and liter scales to increase production[11–14]. However, propagation of light in mammalian bioreactors has not yet been evaluated. Given that reactor

geometry heavily differs between microbial and mammalian bioreactors, it is critical to explore light transmission specific to typical mammalian bioreactor designs. Microbial photobioreactors often employ flat panel or tubular geometries to overcome limitations in light propagation, but mammalian cells often require cylindrical bioreactors with a limited range of tolerated aspect ratios for optimal cell growth or bioproduction[15–17]. If optogenetics is compatible with such reactor design, this technology may be useful for new biomanufacturing areas in which reactor design is still ongoing, including cultivated meat. This field is also constrained by considerably lower cost targets. Light-inducible systems could be far more economical than chemically-induced systems, which have added costs of the chemicals and potentially higher media costs if media exchange is required to remove the chemicals [16,18,19].

Here, we use computer simulations to explore light propagation in bioreactors with parameters tuned to mammalian cultures and bioreactor geometry. We find numerical solutions to the radiative transfer equation (RTE) and assess relevant ranges of absorption coefficients, scattering coefficients, scattering anisotropy, and reactor height-to-diameter ratios for mammalian cells. We determine that optogenetic systems are applicable to industrial scale bioreactors with high efficiencies up to 80,000 L.

Results

Paramter scans on reactor size and optical parameters

To model light propagation through a bioreactor, we simulated a cylinder filled with medium, with light emitted inwards from the walls of the cylinder (Fig. 1A). As light travels through the medium, which includes the cells, its intensity is decreased through absorption and scattering (Fig. 1B).
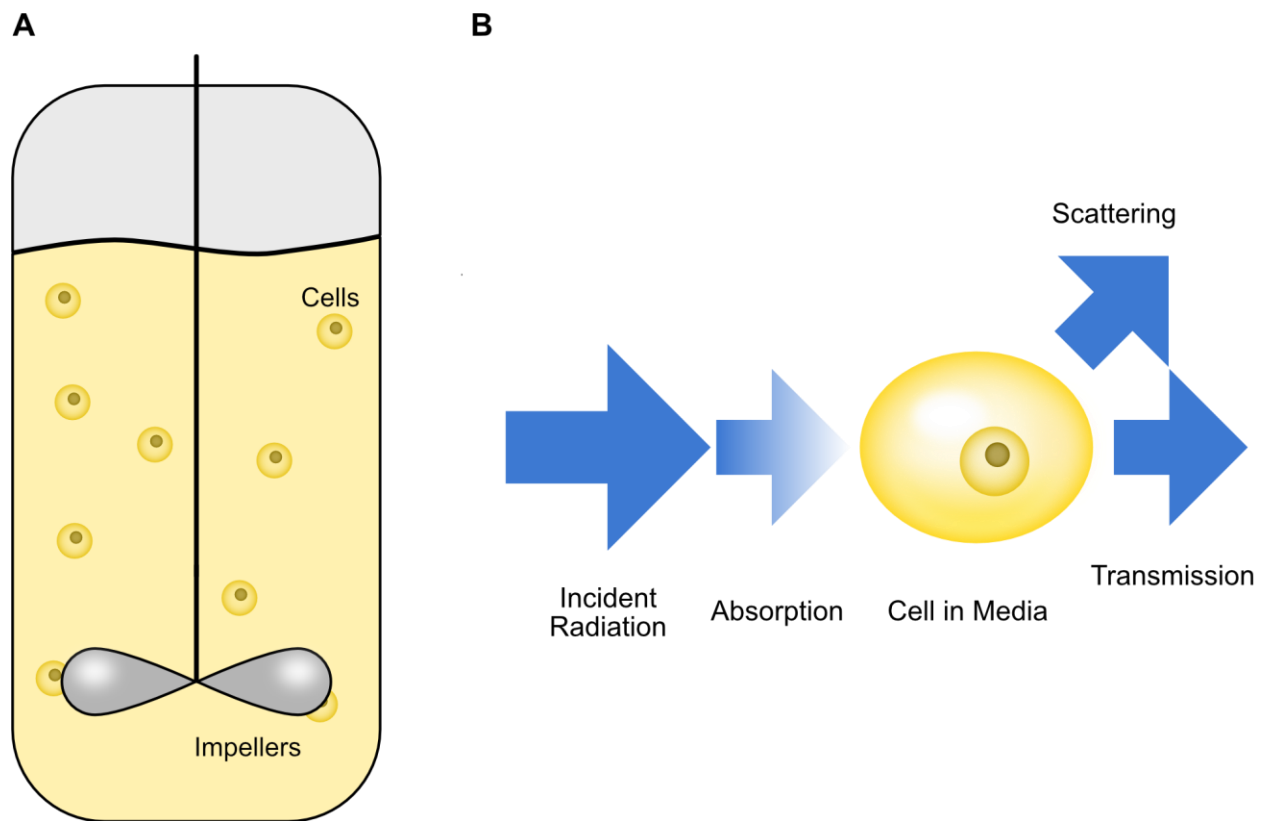
Figure 1. Light is attenuated through absorption and scattering as it travels through a bioreactor. (A) Cells dispersed in a well-mixed cylindrical bioreactor. (B) Radiation is absorbed, scattered, or transmitted while traveling through media containing cells.

First, we investigated the effects of the absorption coefficient, κ, scattering coefficient, σ_s, and scattering anisotropy, g, on the intensity profile across a 100,000 L reactor, which is the order of magnitude size for the largest commercial bioreactors (Fig. 2). Incident radiation of 7.958 W/m^2·sr was used for all simulations. We note that changing g also affects σ_s when calculating from eq. 4, but for the purpose of observing the sensitivity of the intensity profile on g, σ_s was held constant while scanning g. While scanning the values of one parameter, the others were kept at more transparent values such that the changes in intensity distribution from the parameter being

scanned are more apparent. The range in values used for absorption and scattering coefficients were chosen such that they spanned values attained at minimum and maximum cell concentrations. Scattering anisotropy is known to be very high for mammalian cell suspensions, but lower values that tissues may have were also explored[20–22].
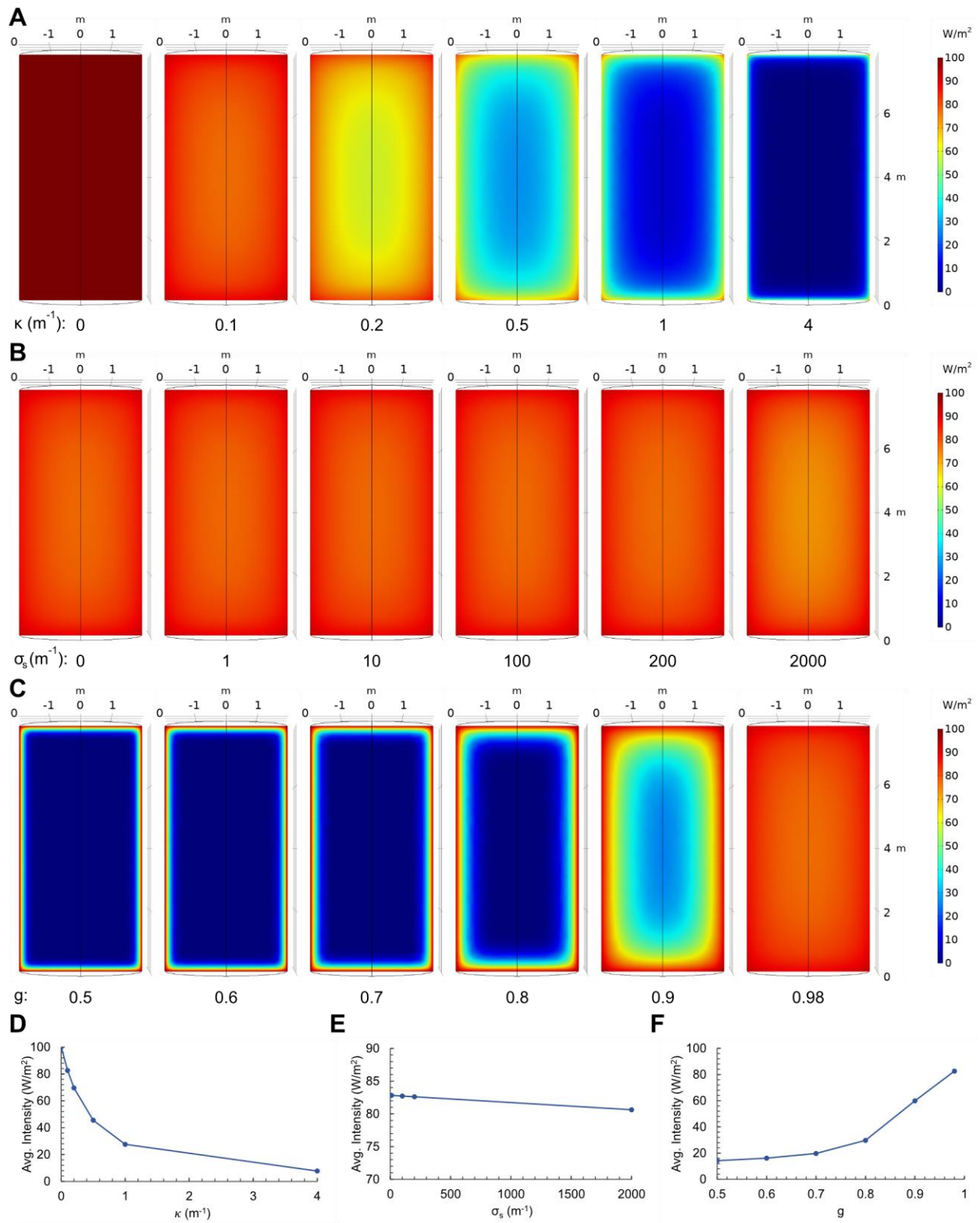
Figure 2. Light propagation over various κ, σ_s, and g. (A) Parameter scan over κ with constant σ_s and g. (B) Parameter scan over σ_s with constant κ and g. (C) Parameter scan over g with constant κ and σ_s. (D) Average intensity as a function of κ with constant σ_s and g. (E) Average intensity as a function of σ_s with constant κ and g. (F) Average intensity as a function of g with constant κ and σ_s. When parameters were held constant, they were set to $\kappa = 0.1 \text{ m}^{-1}$, $\sigma_s = 200 \ m^{-1}$, and $g = 0.98$.

The parameter scan through relevant κ values indicated that absorption had a strong effect on light attenuation (Fig. 2A and 2D). As κ increased, the intensity distribution shifted from mostly maximal throughout the reactor to mostly 0. In contrast, increases in σ_s did not noticeably impact light propagation until the maximum value of 2000 m^{-1} was used (Fig. 2B and 2E). Although scattering coefficients were high, the high scattering anisotropy may alleviate decreases in intensity along the radial direction by scattering the radiation strongly in the forward direction. Indeed, scanning lower scattering anisotropy values caused lower intensities in the center of the reactor by redirecting the radiation away from the radial path (Fig. 2C and 2F). This indicates that σ_s was not negligible, even at the intermediate value of 200 m^{-1} that was used in the anisotropy sweep, as changing anisotropy would not impact the intensity distribution if σ_s was too low. Rather, strong scattering occurred but simply in the forward direction, and decreasing the scattering anisotropy had a very strong effect on the light intensity distribution. The change in intensity distribution from changing the scattering anisotropy is only attributed to scattering, and not absorption.

Mammalian cell bioreactors are utilized at a wide range of sizes and aspect ratios[15–17]. The surface area to volume ratio of a cylinder decreases as $\frac{1+2n}{nr}$, where n is the height to radius aspect

ratio and r is the radius. For this system, in which the light is emitted inwards from the reactor walls, the average intensity of light in the reactor is expected to decrease as the reactor size increases. We first observed the effect of reactor size and height-to-diameter (HD) ratio on intensity, using the maximum cell density $2 \cdot 10^{13}$ cells/m^3 and absorption and scattering coefficients at that cell density (Fig. 3).
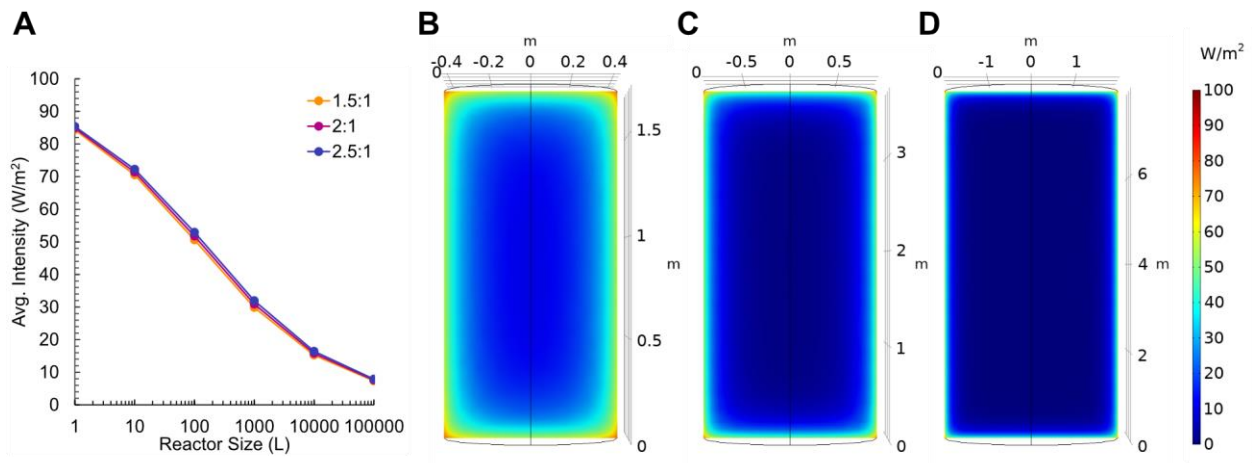

Figure 3. Light propagation vs HD ratios and volume. (A) Average intensity in the reactor as a function of reactor volume for three different HD ratios. (B-D) Slices of the 3D radial intensity distributions for (B) 1,000, (C) 10,000, and (D) 100,000 L reactors.

As reactor volume increased, the volume average intensity decreased. The HD ratio also affected the average intensity, where larger HD ratios, which yield a higher surface area to volume ratio, had higher average intensity, as expected. Overall, however, the aspect ratios in the range of interest have little effect on the average intensity in the relevant range for mammalian cell culture, where in a well-mixed reactor, cells sample all regions in the reactor equally.

Threshold analysis for sufficient light activation

As a metric for sufficient light-activation, we used the light intensity at which 90% of maximal expression was obtained in previous work, which corresponds to 50 W/m^2[23]. For the 1,000 L and larger bioreactors, the average intensity was below the threshold of 50 W/m^2, which indicates that there may be insufficient activation. For these cases, we inspected slices of the 3D radial intensity distributions for the bioreactor with 2:1 HD ratio (Fig. 3B-3D). For the larger reactors, most reactor volumes had near-zero intensity, but positions near the boundaries still had high intensity. Under this constraint, we now consider that in many optogenetic applications[13,23,24], cells are not required to be constantly illuminated, but are sufficiently activated following short pulses of light followed by extended periods in the dark. Short millisecond to second pulses of light followed by seconds or minutes in the dark are sufficient for activation because the timescale for the optogenetic proteins to deactivate through conformational changes and dissociation are longer, with half-lives on the order of seconds to minutes[25]. Therefore, reactors that have a lower average intensity than the required intensity may still be able to activate cells if regions near the boundary meeting the threshold intensity are sufficiently large. As an order-of-magnitude estimate, we considered a case in which 1% of time spent under illumination is sufficient for activation[25]. In the context of a bioreactor with constant illumination, if a cell occupies a region of the reactor with the threshold intensity for 1% of the time, the cell is adequately activated. Then, assuming a well-mixed system, the volume fraction of the reactor achieving threshold intensity is equal to the time-averaged probability that a cell will experience that intensity. Thus, for the larger bioreactors, reactor volume fractions achieving threshold intensity were calculated (Fig. 4A). The 2:1 HD ratio is considered as a representative case, and the other aspect ratios are expected to yield similar results as the intensity distributions were similar (Fig. 3A).
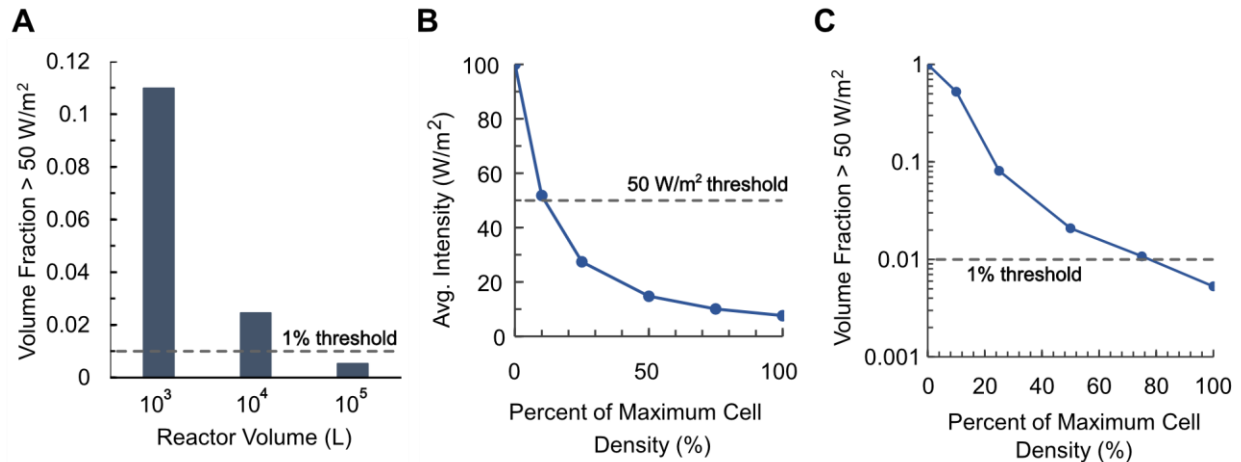
Figure 4. Volume fractions achieving threshold intensity for 2:1 HD ratio reactors. (A) Volume fractions achieving threshold intensity at maximum cell density. (B) Average intensity as a function of cell density. (C) Volume fractions achieving threshold intensity as a function of cell density. Maximum cell density was taken to be $2 \cdot 10^7$ cells/mL. Dashed lines indicate threshold volume fraction and threshold average intensity for a well-mixed system.

The 1,000 L and 10,000 L reactors now met the threshold for adequate activation, as more than 1% of the reactor volumes had intensities > 50 W/m². The 100,000 L bioreactor did not meet the volume fraction threshold. Thus, for the largest of industrial bioreactors with the highest cell densities, there was a decrease in the dynamic range for light activation. Linearly interpolating as a first approximation, the largest reactor size that met the volume fraction threshold for light activation with maximum cell culture density occurred at approximately 80,000 L for the parameters tested.

Finally, we explored the effects of increasing cell density. As cell density increases, absorption and scattering increase proportionally, as given in eq. 3 and eq. 5. Considering the range of cell density from 0 cells/mL – $2 \cdot 10^7$ cells/mL, the effects of cell growth on the intensity profile in the 100,000 L bioreactor were explored through simultaneous increases in κ and σ_s. As cell density

was increased, κ and σ_s were recalculated from α_a and α_s and the average intensities inside the reactor were calculated (Fig. 4B). From the parameter scan, we determined that the highest density of cells that can be sufficiently activated in a 100,000 L bioreactor is approximately 75% of maximum density, or $1.5 \cdot 10^7$ cells/mL (Fig. 4C).

Discussion

Our data suggest that light-inducible systems are viable in large-scale mammalian suspension culture bioreactors. The average intensity drops quickly as reactor size or cell density increases, but only short pulses of light are required for robust light-activation. This allows the high intensities near the boundaries of the reactor to sufficiently activate cells, even when the average intensity within the overall volume is low.

In general, absorption dramatically limits light propagation as the absorption coefficient or cell density increases. However, scattering does not have as large an effect, even though scattering coefficients of mammalian cells are very high. The lower effect of scattering on the intensity distribution is due to the scattering anisotropy also being high. High scattering anisotropy causes most of the radiation to be forward scattered, mitigating the attenuation of radiation in the radial direction of the reactor. The effects of scattering became more evident when the scattering anisotropy was decreased, which rapidly decreased the intensity throughout the volume (Fig. 2C). Future directions include experimental corroboration of these simulations, which can be achieved from sophisticated experimental setups that can discern light attenuation from absorption and scattering. Examples include utilizing an integrating sphere and measuring thin samples such that multiple scattering events do not occur, [26–29]

Sparging was not considered in this work, but the presence of bubbles will increase scattering. Scattering from bubbles will require careful and more intensive computation for several reasons. First, the scattering depends on the bubble distribution throughout the reactor, which may not be as uniformly dispersed as the cells[30]. Bubble breakage and coalescence may contribute to changing bubble size distributions as bubbles travel upwards along the reactor height[31]. Modelling bubbles as distinct phases may be required, given the larger bubble sizes and potentially changing size distributions as the bubbles travel upwards through the reactor[32]. The bubble size distribution also affects scattering, where scattering is affected by the width of the size distribution and cannot be calculated solely from the mean bubble size[20,33,34]. The sparging rate will also affect the size and total bubble volume in the reactor. The intricacy in reactor design and variability that scattering may have across different bioreactor designs suggests a need to focus on specific bioreactor designs for computing scattering phenomena, where for specific reactor designs, scaling analysis no longer straightforward as many parameters such as aspect ratio are not constant across reactor size. Perhaps a specific line of reactors, such as Ambr bioreactors, could be used as a case study for investigating light propagation with a defined set of reactor geometries and operational constraints[35,36].

The maximum reactor size that can support optogenetics based on the geometry and parameters used in this work is approximately 80,000 L, but other elements that may be present in a real reactor may be utilized to further enable optogenetics. For instance, bioreactors are often equipped with impellers, and LEDs may be mounted on the impeller shaft, such that illumination occurs from the center of the reactor, as well as from the reactor walls. Baffles may also be added to the geometry, which increases the surface area available for LED integration without increasing the volume. Finally, LED intensities can be increased, though care must be taken to avoid phototoxicity to the cells[37]. Phototoxicity studies on the intensities and pulse frequencies at which optogenetics becomes cytotoxic for model organisms would contribute valuable information for

optogenetics and determine the extent to which optogenetics could be scaled up for cultures allowing less light penetration, such as suspension cells with lower scattering anisotropy or tissues on scaffolds for cultivated meat applications, or for other systems that similarly inhibit light propagation.

Simulations in this work explore maximal mammalian bioreactor sizes, which are several orders of magnitudes larger than photobioreactors for algae and cyanobacteria [7,9,10]. Larger reactors can be supported by optogenetics because sufficient activation only requires illumination comprising a small fraction of overall time, such as 1% (Fig. 4). We speculate that lower illumination fractions are sufficient for optogenetics because less photon energy is required to saturate single gene circuits or signaling pathways, whereas light-dependent cell growth may require more energy in algal and cyanobacterial systems.

The extent to which light propagates through mammalian suspension cultures heavily depends on the wavelength of the incident radiation. The simulations performed in this work use parameter values assuming blue light at approximately 450 nm. Blue light optogenetics systems are the most widely used, but systems utilizing other wavelengths are also gaining popularity. In particular, red light optogenetics is promising because mammalian cells have lower absorption and scattering for red light, which would increase the penetration of red light through larger systems [20,26]. Absorption and scattering generally decrease as the wavelength increase to the infrared region, and the results in this work may be interpreted as conclusions from conservative calculations. As the use of higher wavelength optogenetics systems increases, further scaleup will become feasible in mammalian suspension cultures or enable deeper penetration of light into tissues, such as in cultivated meat applications.

Conclusion

We demonstrate that light propagates sufficiently in large bioreactors of mammalian suspension cultures. While we did observe the limits of optogenetic activation in the largest scales of industrial bioreactors, modifications to the lighting conditions and bioreactor geometries may extend the limits of optogenetics into volumes of higher orders of magnitudes. Future efforts to model more realistic mammalian bioreactor setups, including sparging, impellers, and baffles, will improve the accuracy of results. Ultimately, such simulations can be used to inform new photobioreactor design in growing industries requiring innovation to lower production costs of complex products.

Methods

Simulating radiation transfer in absorbing and scattering media

Computations were performed using COMSOL Multiphysics 5.5. Light propagation was modeled using Radiation in Absorbing Scattering Medium (RASM) physics in 3D. To model the propagation of radiation through absorbing, scattering medium, the RTE can be used, which takes the form:

$$\Omega \cdot \nabla I(\Omega) = -(\kappa + \sigma_s)I(\Omega) + \frac{\sigma_s}{4\pi} \int_0^{4\pi} I(\Omega')\phi(\Omega',\Omega)d\Omega' \tag{1}$$

where Ω is the solid angle, I is intensity, κ is the absorption coefficient, σ_s is the scattering coefficient, and ϕ is the scattering phase function.

Optical properties are highly wavelength-dependent. In this work we focus on values pertinent to blue light radiation, which has wavelengths of approximately 450 nm - 500 nm and is commonly used in mammalian light-inducible systems[23,24,38,39].

48

The angular space in eq. 1 was discretized using the discrete ordinates method:

$$\frac{\sigma_s}{4\pi}\int_0^{4\pi} I(\Omega')\phi(\Omega',\Omega)d\Omega' = \frac{\sigma_s}{4\pi}\sum_{j=1}^{N}\omega_j I_j \phi(s_j, s_i).$$ (2)

In eq. 2, the S6 level-symmetric even quadrature set and a linear discretization of the radiative intensity were used. Incident radiation was set to 7.958 W/m^2·sr, which yields 100 W/m^2, an order of magnitude of light intensities often used in optogenetics[23,24,38,40]. The radiation was emitted inwards from all boundaries of the domain. Bubbles were not considered, and internal components such as impellers and shafts were also omitted from the simulations.

Calculation of absorption and scattering coefficients

The values of absorption and scattering coefficients used in this work are based on available data on mammalian cells in the literature[26]. Detailed calculations of parameters are available in Appendix B. Given the high optical transparency of PBS, the mass absorption coefficient α_a of mammalian cells is estimated from the absorption coefficient κ and the cell number density ρ_c:

$$\alpha_a = \frac{\kappa}{\rho_c}.$$ (3)

$\alpha_a = 2 \cdot 10^{-13}\frac{m^2}{cell}$ is used in this work, and this value of α_a is used again in eq. 3 to calculate κ for various cell densities. Given culture medium also has a low absorption coefficient, calculations for the absorption coefficient as a function of cell density were also performed assuming negligible contributions from media[41]. An underlying assumption made for optically transparent media is the media is free of serum and phenol red, which is typically the case in industrial bioreactors.

For scattering, we use a reduced scattering coefficient of 2 m^{-1} and a scattering anisotropy g of 0.98[20]. The scattering coefficient was calculated as:

$$\sigma_s = \frac{\sigma_s'}{1-g}$$

(4)

From eq. 4, the mass scattering coefficient α_s is subsequently calculated:

$$\alpha_s = \frac{\sigma_s}{\rho_c}.$$

(5)

$\alpha_s = 10^{-10} \frac{m^2}{cell}$ is used in this work. Similarly to α_a, this α_s value is reused in eq. 5 to calculate σ_s for various cell densities. Scattering coefficients were also assumed to be unaffected by culture media[41].

Calculation of the scattering phase function

The scattering phase function was calculated using the Henyey-Greenstein function[20,21]:

$$\Phi = \frac{1}{K} \frac{(1-g^2)}{1+g^2-2g\mu_0^{\left(\frac{3}{2}\right)}}$$

(6)

where g is the anisotropy factor, μ_0 is the cosine of the scattering angle, and K is defined as the following to normalize the phase function:

$$K = \frac{1}{4\pi} \int_0^{4\pi} \frac{(1 - g^2)}{1 + g^2 - 2g\mu_0^{\left(\frac{3}{2}\right)}} \, d\Omega.$$ (7)

Eq. 6 and eq. 7 were substituted into eq. 1 to calculate radiative transfer.

References

1. Fenno, L., Yizhar, O., & Deisseroth, K. (2011). The Development and Application of Optogenetics. *Annual Review of Neuroscience*, *34*, 389–412. https://doi.org/10.1146/annurev-neuro-061010-113817

2. Deisseroth, K. (2015). Optogenetics: 10 years of microbial opsins in neuroscience. *Nature Neuroscience*, *18*(9), Article 9. https://doi.org/10.1038/nn.4091

3. Goglia, A. G., & Toettcher, J. E. (2019). A bright future: Optogenetics to dissect the spatiotemporal control of cell behavior. *Current Opinion in Chemical Biology*, *48*, 106–113. https://doi.org/10.1016/j.cbpa.2018.11.010

4. Bell, G. R. R., Rincón, E., Akdoğan, E., & Collins, S. R. (2021). Optogenetic control of receptors reveals distinct roles for actin- and Cdc42-dependent negative signals in chemotactic signal processing. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-26371-z

5. Pan, H., Wang, H., Yu, J., Huang, X., Hao, Y., Zhang, C., Ji, W., Yang, M., Gong, X., Wu, X., & Chang, J. (2019). Near-infrared light remotely up-regulate autophagy with spatiotemporal precision via upconversion optogenetic nanosystem. *Biomaterials*, *199*, 22–31. https://doi.org/10.1016/j.biomaterials.2019.01.042

6. Sahel, J.-A., Boulanger-Scemama, E., Pagot, C., Arleo, A., Galluppi, F., Martel, J. N., Esposti, S. D., Delaux, A., de Saint Aubert, J.-B., de Montleau, C., Gutman, E., Audo, I., Duebel, J., Picaud, S., Dalkara, D., Blouin, L., Taiel, M., & Roska, B. (2021). Partial recovery of visual function in a blind patient after optogenetic therapy. *Nature Medicine*, *27*(7), Article 7. https://doi.org/10.1038/s41591-021-01351-4

7. Wheaton, Z. C., & Krishnamoorthy, G. (2012). Modeling radiative transfer in photobioreactors for algal growth. *Computers and Electronics in Agriculture*, *87*, 64–73. https://doi.org/10.1016/j.compag.2012.05.002

8. Gao, X., Kong, B., & Vigil, R. D. (2018). Simulation of algal photobioreactors: Recent developments and challenges. *Biotechnology Letters*, *40*(9), 1311–1327. https://doi.org/10.1007/s10529-018-2595-3

9. Choi, S. Y., Sim, S. J., Ko, S. C., Son, J., Lee, J. S., Lee, H. J., Chang, W. S., & Woo, H. M. (2020). Scalable Cultivation of Engineered Cyanobacteria for Squalene Production from Industrial Flue Gas in a Closed Photobioreactor. *Journal of Agricultural and Food Chemistry*, *68*(37), 10050–10055. https://doi.org/10.1021/acs.jafc.0c03133

10. Melnicki, M. R., Pinchuk, G. E., Hill, E. A., Kucek, L. A., Stolyar, S. M., Fredrickson, J. K., Konopka, A. E., & Beliaev, A. S. (2013). Feedback-controlled LED photobioreactor for photophysiological studies of cyanobacteria. *Bioresource Technology*, *134*, 127–133. https://doi.org/10.1016/j.biortech.2013.01.079

11. Pouzet, S., Banderas, A., Le Bec, M., Lautier, T., Truan, G., & Hersen, P. (2020). The Promise of Optogenetics for Bioproduction: Dynamic Control Strategies and Scale-Up Instruments. *Bioengineering*, *7*(4), 151. https://doi.org/10.3390/bioengineering7040151

12. Bertaux, F., Sosa-Carrillo, S., Gross, V., Fraisse, A., Aditya, C., Furstenheim, M., & Batt, G. (2022). Enhancing bioreactor arrays for automated measurements and reactive control with ReacSight. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-31033-9

13. Lalwani, M. A., Ip, S. S., Carrasco-López, C., Day, C., Zhao, E. M., Kawabe, H., & Avalos, J. L. (2021). Optogenetic control of the lac operon for bacterial chemical and protein production. *Nature Chemical Biology*, *17*(1), Article 1. https://doi.org/10.1038/s41589-020-0639-1

14. Zhao, E. M., Zhang, Y., Mehl, J., Park, H., Lalwani, M. A., Toettcher, J. E., & Avalos, J. L. (2018). Optogenetic regulation of engineered cellular metabolism for microbial chemical production. *Nature*, *555*(7698), 683–687. https://doi.org/10.1038/nature26141

15. Ahleboot, Z., Khorshidtalab, M., Motahari, P., Mahboudi, R., Arjmand, R., Mokarizadeh, A., & Maleknia, S. (2021). Designing a Strategy for pH Control to Improve CHO Cell Productivity in

Bioreactor. *Avicenna Journal of Medical Biotechnology*, *13*(3), 123–130. https://doi.org/10.18502/ajmb.v13i3.6365

16. Shafiei Kaleybar, L., Khoshfetrat, A. B., & Nozad Charoudeh, H. (2020). Modeling and performance prediction of a conceptual bioprocess for mass production of suspended stem cells. *Food and Bioproducts Processing*, *122*, 254–268. https://doi.org/10.1016/j.fbp.2020.04.012

17. Sandadi, S., Pedersen, H., Bowers, J. S., & Rendeiro, D. (2011). A comprehensive comparison of mixing, mass transfer, Chinese hamster ovary cell growth, and antibody production using Rushton turbine and marine impellers. *Bioprocess and Biosystems Engineering*, *34*(7), 819–832. https://doi.org/10.1007/s00449-011-0532-0

18. Li, F., Vijayasankaran, N., Shen, A. (Yijuan), Kiss, R., & Amanullah, A. (2010). Cell culture processes for monoclonal antibody production. *MAbs*, *2*(5), 466–477. https://doi.org/10.4161/mabs.2.5.12720

19. Allan, S. J., De Bank, P. A., & Ellis, M. J. (2019). Bioprocess Design Considerations for Cultured Meat Production With a Focus on the Expansion Bioreactor. *Frontiers in Sustainable Food Systems*, *3*. https://doi.org/10.3389/fsufs.2019.00044

20. Mourant, J. R., Freyer, J. P., Hielscher, A. H., Eick, A. A., Shen, D., & Johnson, T. M. (1998). Mechanisms of light scattering from biological cells relevant to noninvasive optical-tissue diagnostics. *Applied Optics*, *37*(16), 3586–3593. https://doi.org/10.1364/AO.37.003586

21. Xu, M., Wu, T. T., & Qu, J. Y. (2008). Unified Mie and fractal scattering by cells and experimental study on application in optical characterization of cellular and subcellular structures. *Journal of Biomedical Optics*, *13*(2), 024015. https://doi.org/10.1117/1.2907790

22. Jacques, S. L., Roussel, S., & Samatham, R. V. (2016). Polarized light imaging specifies the anisotropy of light scattering in the superficial layer of a tissue. *Journal of Biomedical Optics*, *21*(7), 071115. https://doi.org/10.1117/1.JBO.21.7.071115

23. Minami, S. A., & Shah, P. S. (2021). Transient light-activated gene expression in Chinese hamster ovary cells. *BMC Biotechnology*, *21*(1), 13. https://doi.org/10.1186/s12896-021-00670-1

24. Polstein, L. R., & Gersbach, C. A. (2015). A light-inducible CRISPR/Cas9 system for control of endogenous gene activation. *Nature Chemical Biology*, *11*(3), 198–200. https://doi.org/10.1038/nchembio.1753

25. Taslimi, A., Zoltowski, B., Miranda, J. G., Pathak, G., Hughes, R. M., & Tucker, C. L. (2016). Optimized second generation CRY2/CIB dimerizers and photoactivatable Cre recombinase. *Nature Chemical Biology*, *12*(6), 425–430. https://doi.org/10.1038/nchembio.2063

26. Cone, M. T., Mason, J. D., Figueroa, E., Hokr, B. H., Bixler, J. N., Castellanos, C. C., Noojin, G. D., Wigle, J. C., Rockwell, B. A., Yakovlev, V. V., & Fry, E. S. (2015). Measuring the absorption coefficient of biological materials using integrating cavity ring-down spectroscopy. *Optica*, *2*(2), 162–168. https://doi.org/10.1364/OPTICA.2.000162

27. Röttgers, R., & Gehnke, S. (2012). Measurement of light absorption by aquatic particles: Improvement of the quantitative filter technique by use of an integrating sphere approach. *Applied Optics*, *51*(9), 1336–1351. https://doi.org/10.1364/AO.51.001336

28. Jacques, S. L. (2013). Optical properties of biological tissues: A review. *Physics in Medicine and Biology*, *58*(11), R37–R61. https://doi.org/10.1088/0031-9155/58/11/R37

29. Hohmann, M., Lengenfelder, B., Muhr, D., Späth, M., Hauptkorn, M., Klämpfl, F., & Schmidt, M. (2021). Direct measurement of the scattering coefficient. *Biomedical Optics Express*, *12*(1), 320–335. https://doi.org/10.1364/BOE.410248

30. Huang, Y., Zhao, S., Ding, Y., Liao, Q., Huang, Y., & Zhu, X. (2017). Optimizing the gas distributor based on CO2 bubble dynamic behaviors to improve microalgal biomass production in an air-lift photo-bioreactor. *Bioresource Technology*, *233*, 84–91. https://doi.org/10.1016/j.biortech.2017.02.071

31. Krýsa, P., & Šoóš, M. (2022). Modelling of bubble breakage and coalescence in stirred and sparged bioreactor using the Euler-Lagrange approach. *International Journal of Heat and Mass Transfer*, *199*, 123466. https://doi.org/10.1016/j.ijheatmasstransfer.2022.123466

32. Gray, D. R., Chen, S., Howarth, W., Inlow, D., & Maiorella, B. L. (1996). CO(2) in large-scale and high-density CHO cell perfusion culture. *Cytotechnology*, *22*(1–3), 65–78. https://doi.org/10.1007/BF00353925

33. Davis, G. E. (1955). Scattering of Light by an Air Bubble in Water. *JOSA*, *45*(7), 572–581. https://doi.org/10.1364/JOSA.45.000572

34. Baker, E. T., & Lavelle, J. W. (1984). The effect of particle size on the light attenuation coefficient of natural suspensions. *Journal of Geophysical Research: Oceans*, *89*(C5), 8197–8203. https://doi.org/10.1029/JC089iC05p08197

35. Delouvroy, F., Le Reverend, G., Fessler, B., Mathy, G., Harmsen, M., Kochanowski, N., & Malphettes, L. (2013). Evaluation of the advanced micro-scale bioreactor (ambr^{TM}) as a highthroughput tool for cell culture process development. *BMC Proceedings*, *7*(6), P73. https://doi.org/10.1186/1753-6561-7-S6-P73

36. Nienow, A. W., Rielly, C. D., Brosnan, K., Bargh, N., Lee, K., Coopman, K., & Hewitt, C. J. (2013). The physical characterisation of a microscale parallel bioreactor platform with an industrial CHO cell line expressing an IgG4. *Biochemical Engineering Journal*, *76*, 25–36. https://doi.org/10.1016/j.bej.2013.04.011

37. Icha, J., Weber, M., Waters, J. C., & Norden, C. (2017). Phototoxicity in live fluorescence microscopy, and how to avoid it. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *39*(8). https://doi.org/10.1002/bies.201700003

38. Guntas, G., Hallett, R. A., Zimmerman, S. P., Williams, T., Yumerefendi, H., Bear, J. E., & Kuhlman, B. (2015). Engineering an improved light-induced dimer (iLID) for controlling the localization and activity of signaling proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(1), 112–117. https://doi.org/10.1073/pnas.1417910112

39. Hartzell, E. J., Terr, J., & Chen, W. (2021). Engineering a Blue Light Inducible SpyTag System (BLISS). *Journal of the American Chemical Society*, *143*(23), 8572–8577. https://doi.org/10.1021/jacs.1c03198

40. Uhelski, M. L., Bruce, D. J., Séguéla, P., Wilcox, G. L., & Simone, D. A. (2017). In vivo optogenetic activation of Nav1.8+ cutaneous nociceptors and their responses to natural stimuli. *Journal of Neurophysiology*, *117*(6), 2218–2223. https://doi.org/10.1152/jn.00083.2017

41. Silva, D. F. T., Mesquita-Ferrari, R. A., Fernandes, K. P. S., Raele, M. P., Wetter, N. U., & Deana, A. M. (2012). Effective Transmission of Light for Media Culture, Plates and Tubes. *Photochemistry and Photobiology*, *88*(5), 1211–1216. https://doi.org/10.1111/j.1751-1097.2012.01166.x

Chapter 4: Production of novel SARS-CoV-2 Spike truncations in Chinese hamster ovary cells leads to high expression and binding to antibodies

Introduction

The emergence of coronavirus infectious disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has resulted in over 764 million infections and 6.9 million deaths globally since November 2019. Major aspects of containing this global pandemic are surveillance (large-scale and rapid asymptomatic testing) and herd immunity (immunity achieved in a large portion of the population with protective antibodies resulting from vaccination or natural infection). Many of these containment efforts require generating large amounts of viral glycoproteins. Consequently, the COVID-19 pandemic has highlighted the critical need for rapid, scalable, and cost-effective production of recombinant glycoproteins for use as antigens in diagnostic kits, research reagents, and even the active pharmaceutical ingredient in protein-based vaccines.

For SARS-CoV-2, diagnosis and vaccination strategies involve scalable production of the Spike glycoprotein. Spike is the structural protein responsible for protecting the viral genome and for entry into cells. Spike contains the S1 and S2 domains, which mediate host receptor binding and membrane fusion, respectively [1]. The receptor binding domain (RBD) of Spike lies within the S1 domain (Fig. 1A). Spike is a major antigen and the primary target for antibody binding. Consequently, immunoassays to assess immunity of individuals or a community require a SARS-CoV-2 antigen, most commonly the Spike protein. Protein-based SARS-CoV-2 vaccines also rely on delivering Spike protein with adjuvant for immunization [2].

A major limitation to scaling these approaches is generating Spike protein at high titers in a cost-effective manner. Several forms of full-length Spike have been produced in mammalian cell lines, including modifications to increase stability and expression, but titers remain at a low range of approximately 5-30 mg/L, with one report of 150 mg/L [3–5]. A possible alternative is to express the Spike RBD, which can have expression levels of an order of magnitude higher than those of Spike, but is less sensitive than Spike in serological assays [3]. This suggests that RBD may not have the same activity as Spike for such applications. Mutational scanning has been performed on RBD, which resulted in higher expression and stability [6,7]. Rational structure-based approaches have also been used to improve stability of full-length Spike [4]. Nonetheless, identifying sequence-independent methods to increase expression is essential for comparisons with existing variants.

In this work, we transiently express Spike and RBD in Chinese hamster ovary (CHO) cells. To find high expressing and antibody binding forms of Spike, we also design and express 8 truncations of Spike, which include the RBD and additional residues. Two of these truncations express at high levels. Using simulation and experimentation, we find that one of the high-producing truncations also has more structural similarity to full-length Spike than the other and has higher binding to anti-Spike antibodies. The N-linked glycosylation profiles present on T1 are similar to those of full-length Spike and RBD. Taken together, these truncations may provide an additional avenue for lower cost production of COVID-19 biologics with improved expression and antibody binding.

Results

Expression and purification of Spike and RBD

We first compared the expression of Spike and RBD in ExpiCHO-S cells transfected in 96-well format. Spike and RBD were expressed with N-terminal secretion signals and C-terminal 6x His tags for downstream purification (Fig. 1A). We also replaced the previously tested secretion signal with an alternative secretion signal to determine whether it affects expression and secretion of Spike and RBD (Spike-1, Spike-2, RBD-1, and RBD-2, Fig. 1B).
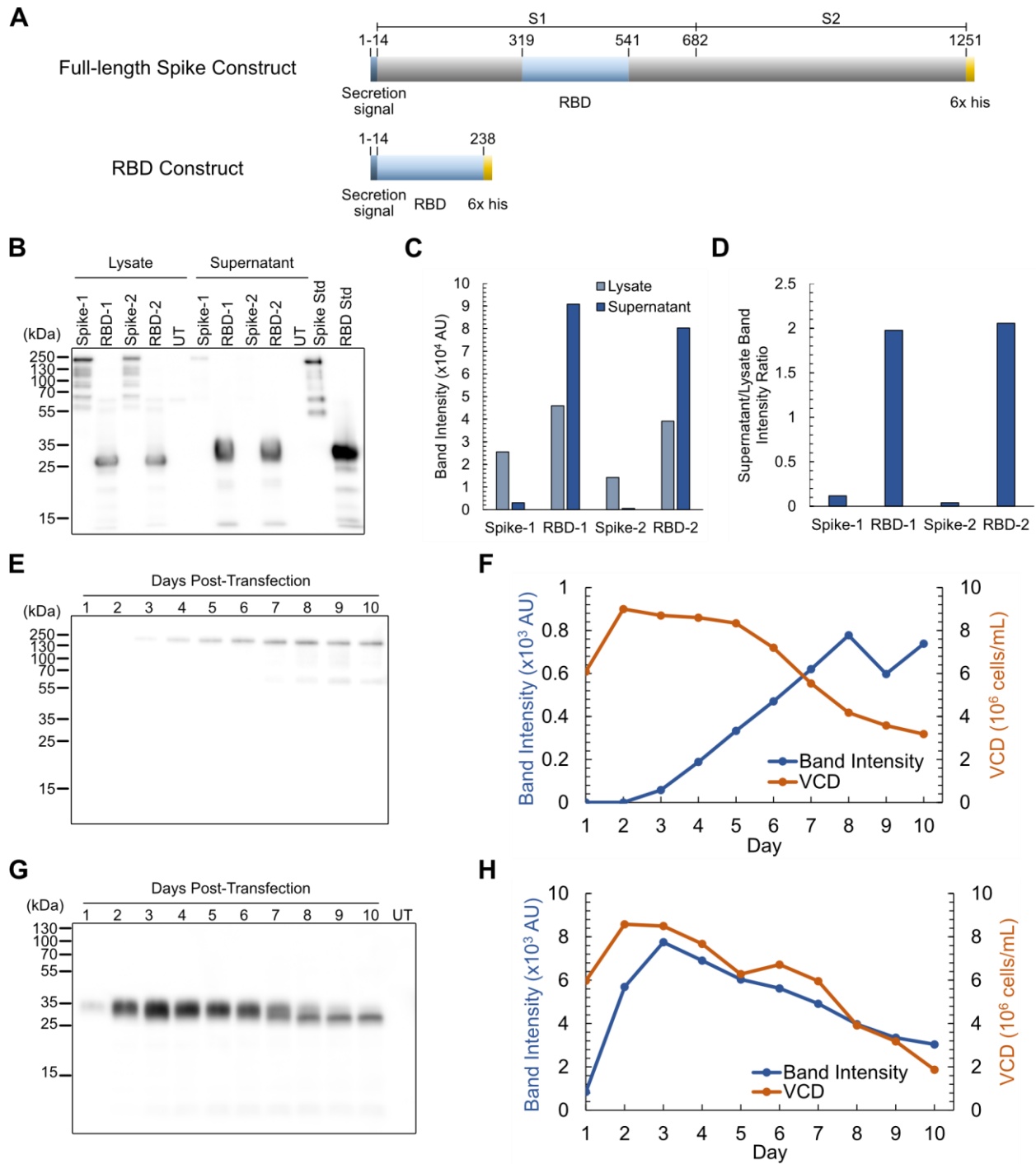
Figure 1. Transient Spike and RBD production in CHO cells. (**A**) Diagram of full-length Spike (1257 aa) and RBD (244 aa) constructs. Residues are labeled starting from the beginning of the secretion signal. (**B**) Western blot and (**C**) densitometry comparing two secretion signals for Spike and RBD. (**D**) Ratio of band intensities of supernatants and lysates. (**E**) Western blot and (**F**)

densitometry on Spike expression time course. (**G**) Western blot and (**H**) densitometry on RBD expression time course. Abbreviations: aa (amino acids); untransfected (UT); standard (std); viable cell density (VCD).

Relative amounts of Spike and RBD in the supernatant and cell lysate were determined by western blot 5 days post-transfection (Fig. 1B and 1C). Spike had significantly lower expression than RBD and was particularly less abundant in the supernatant. Comparison of ratios of supernatants over lysates also showed that Spike is significantly retained in the cells compared to RBD (Fig. 1D). Both the expression and supernatant/lysate ratio remained the same for Spike and RBD with either signal sequence, indicating that low expression and high retention of Spike in the cells may be due to the protein sequence itself, and not a consequence of the tag used. All following experiments were performed with Spike-1 and RBD-1, hereafter referred to as Spike and RBD, respectively.

Next, cultures were scaled up to 25 mL and a time course experiment was performed to determine the optimal harvest time for maximum titers. Cells were transfected with plasmids encoding Spike and RBD, and a sample of the supernatant was collected every 24 hours over 10 days. Western blots were performed on the supernatants and band intensities were plotted over time (Fig. 1E-H). Spike concentration in the supernatant increased steadily until 7 days post-transfection, after which time it remained stable (Fig. 1E and 1F). In comparison, RBD concentration in the supernatant peaked at day 3, then decreased (Fig. 1G and 1H). From these results, Spike and RBD harvests were determined to be 7 and 3 days post-transfection, respectively.

To produce large quantities of Spike and RBD for purification and downstream analysis, 150 mL of supernatants were prepared from pooled 25 mL cultures. Crude titers of Spike and RBD were measured using sandwich ELISAs on filtered crude, yielding 14 mg/L and 54 mg/L, respectively

(Fig. A-1). The crude supernatants were purified through FPLC (Fig. 2 and A-2). For purification of Spike, SDS-PAGE revealed bands in elution fractions E3, E4, and E5 (Fig. 2A), which were confirmed by western blot to include immunoreactive bands consistent with full-length Spike (Fig. 2B). For RBD, SDS-PAGE (Fig. 2C) and western blot (Fig. 2D) showed bands in elution fractions E3, E4, and E7. FPLC-purified samples were dialyzed using PBS and concentrated through centrifugal filter units. E3 fractions of both proteins were used for subsequent experiments.
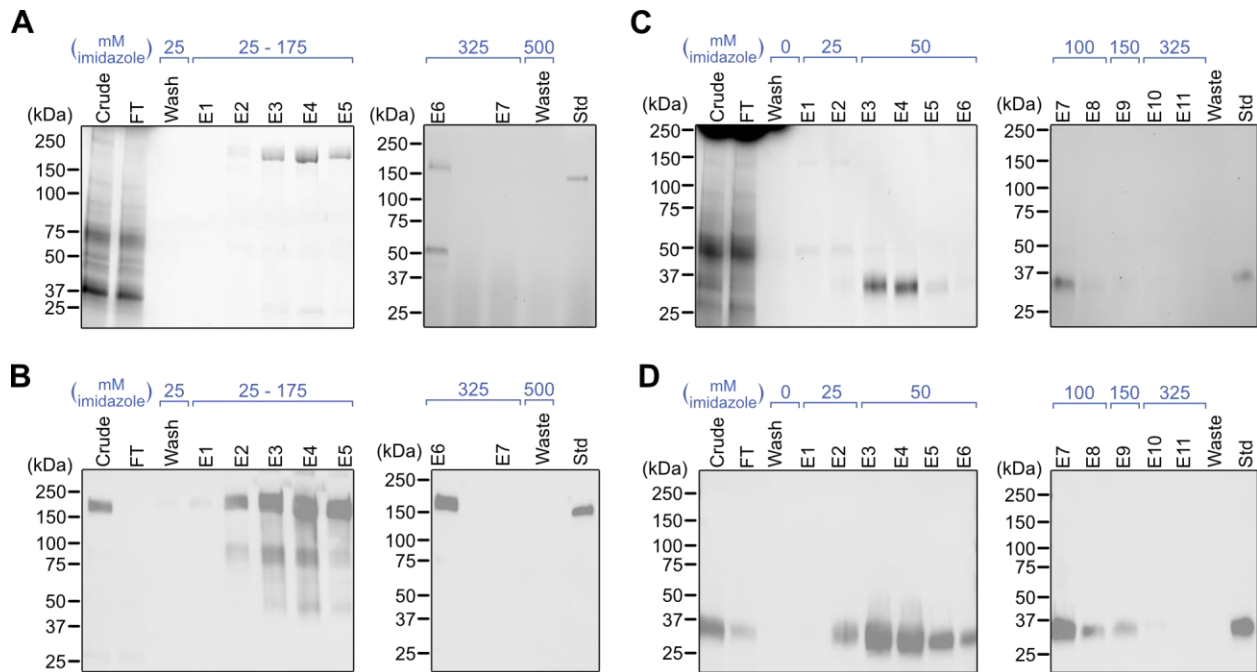

Figure 2. Purification of Spike and RBD. (**A**) SDS-PAGE and (**B**) western blot on Spike fractions. (**C**) SDS-PAGE and (**D**) western blot on RBD fractions. FT samples were collected during sample loading onto the column. FT and wash samples were pooled from multiple fractions at equal volumes. Abbreviations: flow-through (FT); elution (E); standard (std).

Novel truncations to improve protein titers

Given the difficulty in expressing full-length Spike but its higher sensitivity in serological assays

[3] , we sought to determine whether other truncations that are larger than the RBD could be highly expressed and maintain high sensitivity. We developed eight truncations, T1-T8, by adding increments of approximately 50 amino acids on the N- or C-terminal ends of RBD (Fig. 3A). We also expressed the full S1 subunit of Spike. Secretion signals and 6x His tags were added to N- and C-termini, respectively.
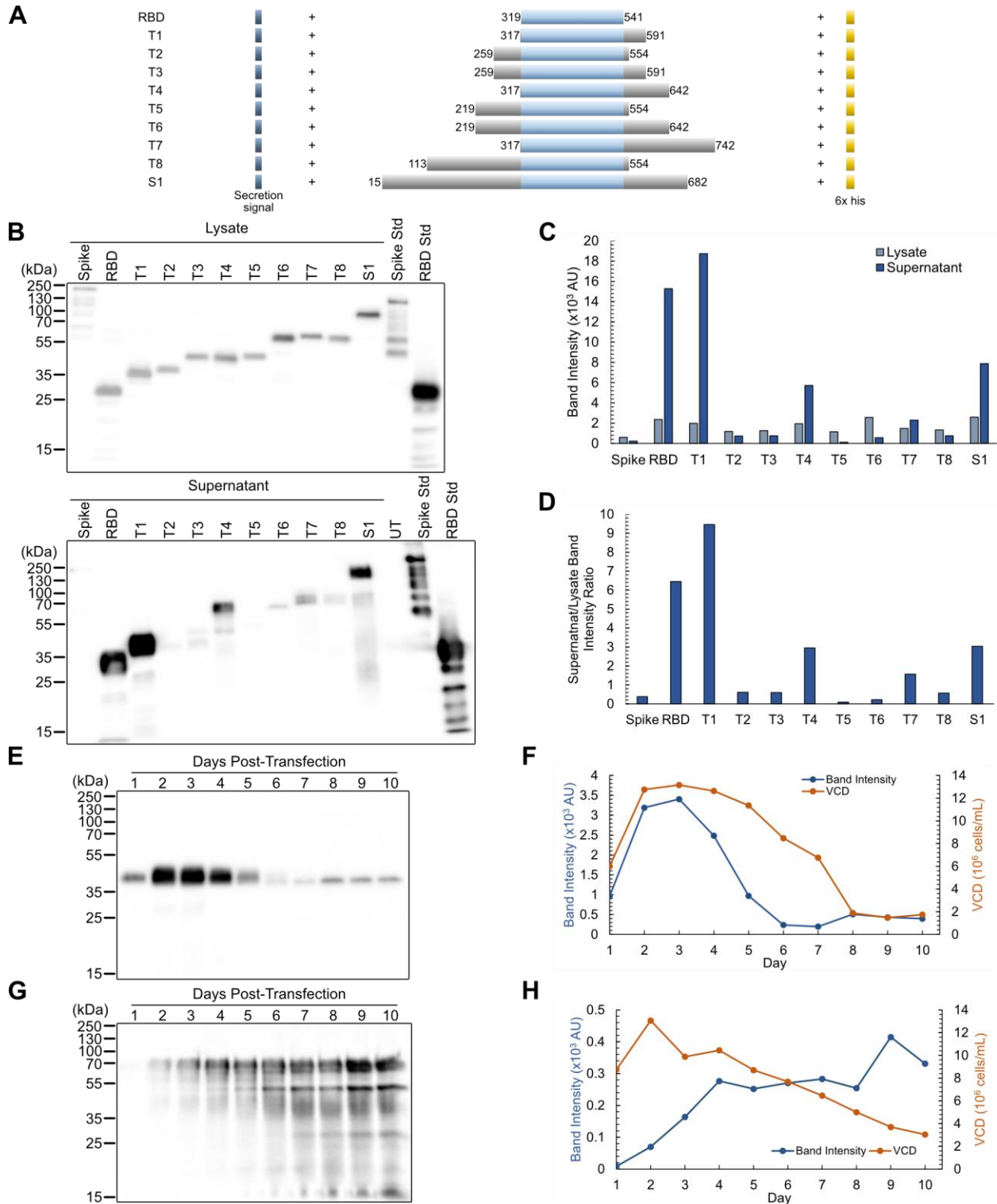
Figure 3. Expression of Spike truncations. (**A**) Construct diagram of Spike truncations. Residue numbers are relative to the position within full-length Spike, including the secretion tag. (**B**) Western blot and (**C**) densitometry on truncations for lysates and supernatants. (**D**) Ratio of band

intensities of supernatants over lysates. (**E**) Western blot and (**F**) densitometry of T1 expression over time. (**G**) Western blot and (**H**) densitometry of T4 expression over time. Abbreviations: untransfected (UT); standard (std); viable cell density (VCD).

We first screened the truncations for expression levels. Cells were transfected in 96-well format, and lysates and supernatants were analyzed by western blot 5 days post-transfection (Fig. 3B). T1 and T4 had the highest expression, as well as the highest supernatant/lysate signals (Fig. 3C and 3D). In particular, T1 had even higher expression and relative secretion than RBD did. Given that S1 has previously been studied [8], T1 and T4 were selected for scaleup and purification. To determine optimal harvest dates for T1 and T4, and expression time course was performed as previously described for Spike and RBD. For T1, western blot and densitometry (Fig. 3E and 3F) showed that expression peaked 3 days post-transfection with a single band. For T4, expression peaked after 4 days post-transfection, with degradation bands also increasing after this point (Fig 3G and 3H). Therefore, T1 and T4 harvest dates were determined to be 3 and 4 days post-transfection, respectively.

For purification, transfections with T1 and T4 were performed in 150 mL cultures and samples were purified through FPLC (Fig. 4 and S3). SDS-PAGE revealed two bands for T1, one of which was determined to be T1 through western blot (Fig. 4A and 4B). To remove the impurity, T1 was dialyzed and repurified. For T4, three bands were detected: one band at 70 kDa, one band at 50 kDa, and one band at slightly below 50 kDa (Fig. 4C). Western blot detected bands in both the 70 kDa and 50 kDa regions (Fig. 4D), indicating two forms of T4. The impurity was removed through dialysis and repurification. Pure T1 and T4 were obtained in this manner and prepared for further characterization through dialysis and spin column concentration. The bands detected for T1 and T4 through western blot were cut out of an SDS-PAGE gel and submitted for proteomic analysis. Coverage for T1 and T4 were high (Fig. S4), and the two bands of T4 on the SDS-PAGE gel

suggest different post-translational modifications. Quantifying crude titers, we found that T1 and T4 expressed at 130 mg/L and 73 mg/L, respectively, which are higher than our RBD titers (Fig. S5). For further characterization and binding assays, concentrations of purified samples were determined using Sandwich ELISAs and SDS-PAGE densitometry (Fig. S6).
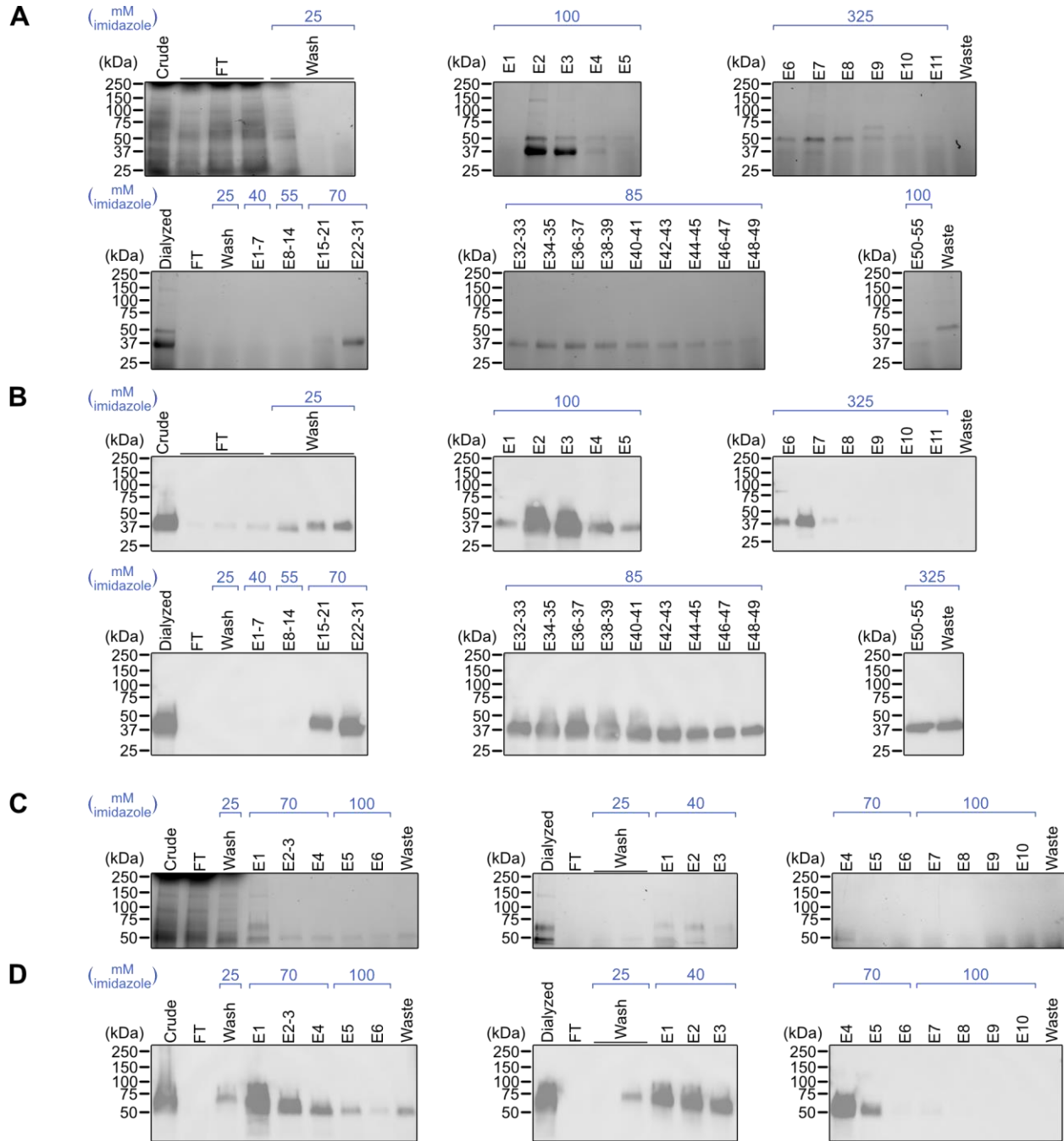
Figure 4. Purification of T1 and T4. (**A**) SDS-PAGE and (**B**) western blot on T1 fractions from crude purification and repurification. (**C**) SDS-PAGE and (**D**) western blot on T4 fractions from crude purification and repurification. FT and wash samples were pooled from multiple fractions at equal volumes. Abbreviations: flow-through (FT); elution (E).

Binding sensitivities against antibodies

The activities of the CHO-expressed proteins were evaluated via indirect ELISAs with antibodies raised against full-length Spike. First, the monoclonal antibody CR3022 was tested, which binds to the receptor binding domain of Spike[9]. Serial dilutions of CR3022 were incubated with Spike, RBD, T1, and T4 (Fig. 5A). Binding sensitivities were compared by taking the areas under the curves (Fig. 5B). To compare with another recombinant source of Spike, Sf9 insect cell-expressed Spike was also used in the assays.
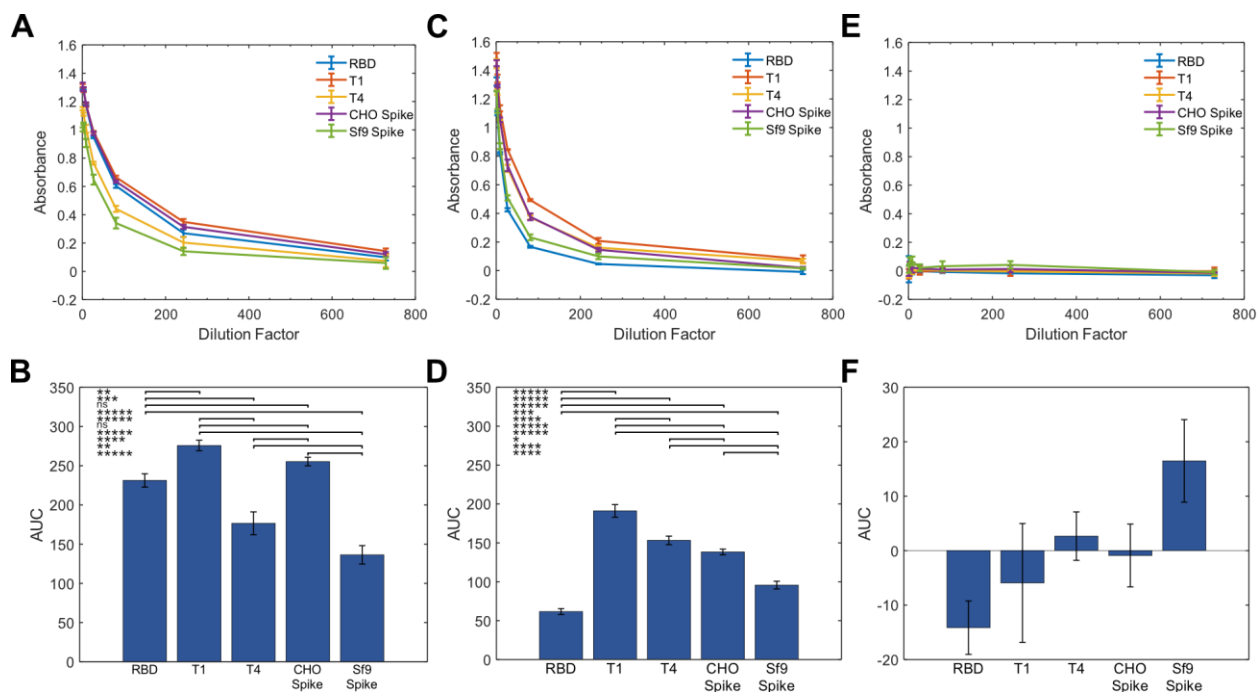
Figure 5. Binding assays of Spike truncations with anti-Spike antibodies. (**A**) Absorbance as a function of dilution factor of CR3022. (**B**) AUC calculated from (A). (**C**) Absorbance as a function of dilution factor of an anti-Spike polyclonal antibody (PAb). (**D**) AUC calculated from (C). (**E**) Absorbance against dilution factor of a rabbit normal IgG antibody. (**F**) AUC calculated from (E). Error bars represent ± SD of technical triplicates. p-values were calculated using a one-way ANOVA followed by Tukey's Test. * indicates $p < 5 \times 10^{-2}$, ** $p < 5 \times 10^{-3}$, *** $p < 5 \times 10^{-4}$, **** $p < 5 \times 10^{-5}$, and ***** $p < 5 \times 10^{-6}$. Abbreviations: area under curve (AUC); ns (not significant).

CHO-expressed Spike had higher binding to CR3022 than Sf9-expressed Spike. This may be due to differences in folding or glycosylation between the CHO- and insect-expressed proteins. Among the CHO-expressed proteins, T1 had higher binding to CR3022 than RBD did and is comparable to the performance of Spike. T4 had lower signal but still outperformed Sf9-expressed Spike. Next, serial dilutions of a polyclonal antibody raised against full-length Spike were tested (Fig. 5C and 5D). Given that PAbs may recognize multiple binding epitopes in a protein, larger forms of Spike were expected to have higher performance. Strikingly, however, T1 and T4 had very high signal across dilutions of the antibody, and T1 outperformed full-length CHO-expressed Spike. The increased sensitivities were not due to non-specific binding of T1 and T4 to rabbit antibodies, since a control rabbit IgG did not produce significant signal (Fig. 5E and 5F).

Structural characterization of truncations

To determine whether structural similarities are maintained between the truncations and the relevant regions of Spike, structures of T1 and T4 were predicted using MD. Snapshots of simulated structures of RBD, T1, and T4 at 0 ns and 100 ns simulation times are shown (Fig. 6A-D). The RBD portion of all structures remained stable during this time. T1 and T4 showed similarly stable secondary structures in the additional residues at the bottom of the structure. The more flexible turn features curled in and stabilized over the course of the trajectory. To quantify this

behavior, the root mean squared deviation (RMSD) of the whole structures and RBD subdomains for each truncation were evaluated (Fig. 6E and 6F). These RMSD plots show deviation relative to initial structures and provide further evidence that the RBD subdomains are stable or reach a stable structure early in the simulation trajectory. The T1 and T4 RMSD plots show more conformational change, likely due to the flexibility of the turn features observed in the snapshots. An interesting observation is that the RBD with a 6x His tag appeared to be more stable compared to the RBD without the tag. Based on this RMSD data, T1 appeared to stabilize the RBD in line with this 6x His tag, while the T4 structure aligned more closely to the RBD without a 6x His tag.
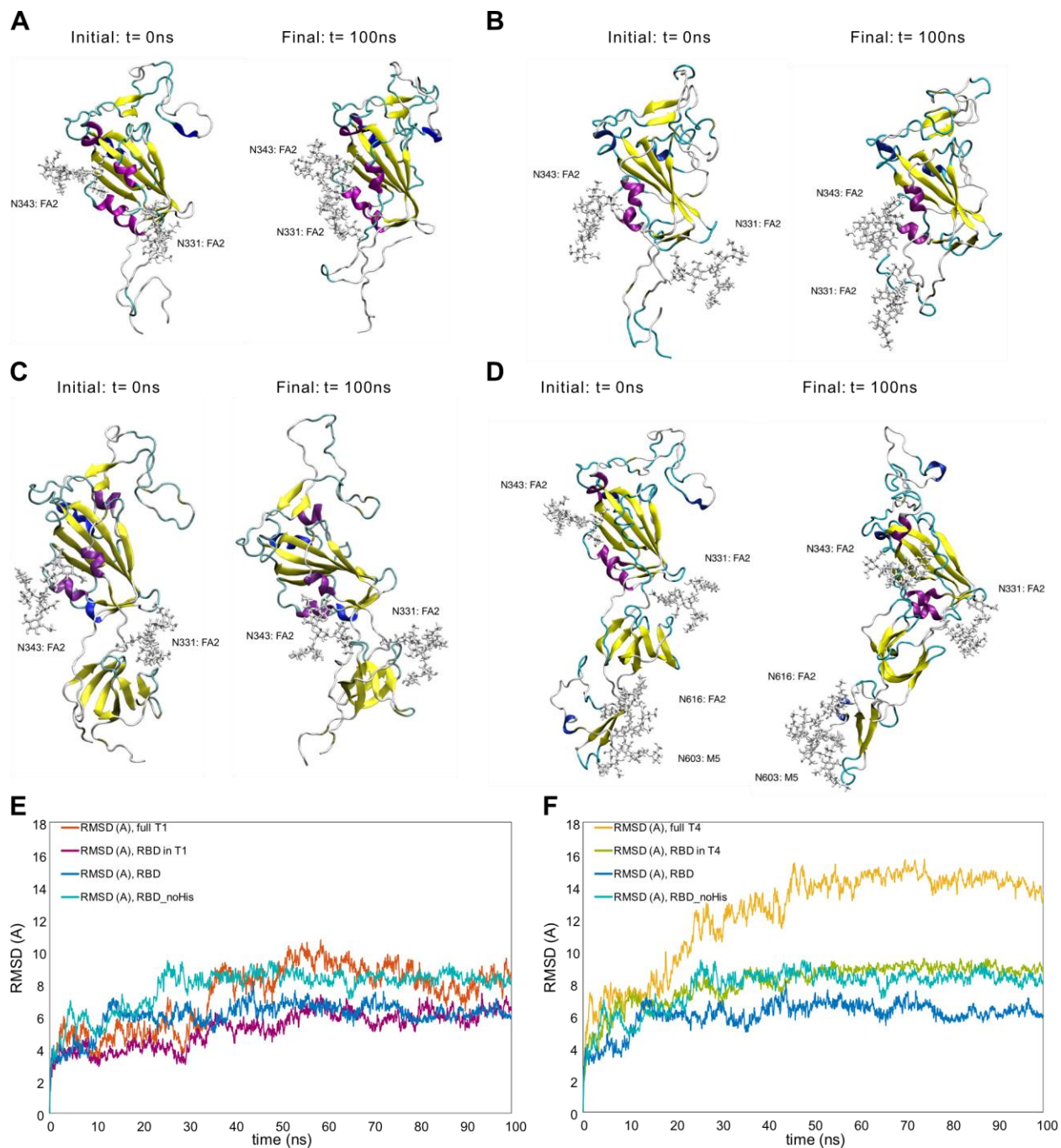
Figure 6. MD structural stability snapshots and analysis. MD snapshots are visualized for (**A**) RBD, (**B**) RBD without the 6x His tag (RBD_noHis), (**C**) T1, and (**D**) T4 at 0 ns and 100 ns. Backbone RMSD profiles of (**E**) full T1 and T1 RBD subdomain and (**F**) full T4 and T4 RBD subdomains are compared against RBD with and without His tag referenced to initial configurations.

71

We hypothesize that some truncations did not express well because of structural differences. To explore this idea, we compared T1 and T3, which only differ by ~50 amino acids at the N-terminus but had vastly different expression (Fig. 3B and Fig. 3C). MD was used to determine whether structural differences may have caused the discrepancy in expression. RMSD analysis showed that T3 had much higher RMSD compared to T1 (Fig. S7). Visualization of T3 revealed that the difference in RMSD was due to the additional FA3 glycan binding to its own RBD, which could contribute to low expression. Removal of the FA3 glycan from T3 resulted in a secondary structure that matched more closely to T1 and a more stable RBD within T3 (Figure S7). It is possible that other truncations also had incorrect folding.

Experimentally, secondary structure compositions of CHO-expressed Spike, RBD, T1, and T4 were obtained using circular dichroism (CD). $\Delta\varepsilon$ values were obtained, which is a measure of the difference in absorbance of left- and right-circularly polarized light. Using the BeStSel server, $\Delta\varepsilon$ as a function of wavelength was analyzed to predict the secondary structure compositions. The distributions of observed secondary structures were similar for most proteins (Fig. 7A and S8). CHO Spike and Sf9 Spike had very similar compositions, suggesting high structural similarity. RBD and T1 also had similar compositions. T4 was slightly dissimilar, with low beta sheet content compared to other proteins. CD-analyzed proteins were also compared to a structure of Spike determined through cryo-electron microscopy (PDB 6VXX, Walls et al., 2020). 6VXX had similar alpha helix and beta sheet content as CHO and Sf9 Spike but had much higher turn content and lower "other" content, which includes coils, bends, irregular loops, β-bridges, 3_{10} helices, and π-helices.
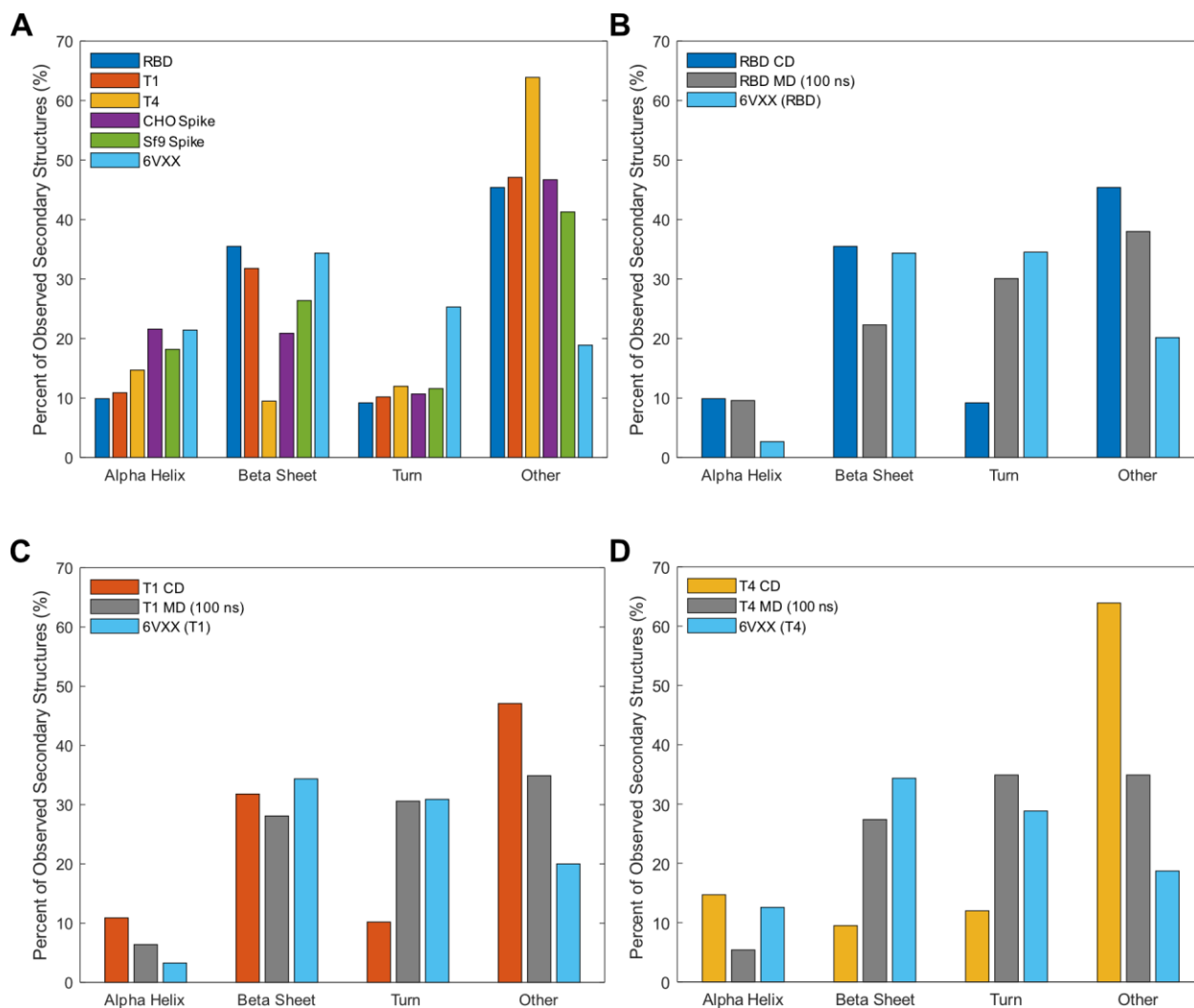
Figure 7. Structural composition of Spike and truncations produced in CHO cells. (**A**) Overall secondary structure compositions of CHO-expressed proteins and 6VXX. (**B**) Comparison of RBD MD to CD data and the RBD region of 6VXX. (**C**) Comparison of T1 MD to CD data and the T1 region of 6VXX. (**D**) Comparison of T4 MD to CD data and the T4 region of 6VXX. MD structural data represent proteins including the 6x His tags with the final structural compositions at 100 ns.

CD-derived structural information was also compared with MD secondary structures for RBD, T1, and T4 determined using DSSP [11] (Fig. 7B-D). Truncated 6VXX structures containing the relevant residues were also included, which represent structural composition had truncating Spike not resulted in any structural changes. For all three proteins, 6VXX and MD structures had high

73

similarity, with CD-derived structures having lower turn content. Overall, MD and CD results suggest that T1 and T4 retain accurate RBD structure, and consistent with their high sensitivities in the ELISAs against anti-Spike antibodies.

Finally, we analyzed the N-linked glycosylation profiles of T1 and T4 using glycoproteomics. Glycosylation is an important protein attribute that influences Spike structure, binding, and immunogenicity[12–14]. The N-glycosylation profiles of full-length Spike have been studied by several groups[15,16], with markedly different glycosylation profiles resolved on the various glycosylation sites present on Spike. This suggests that specific glycans may be needed on different glycosylation sites for the functionality of Spike.

Glycoproteomic profiles of Spike truncations

Glycoproteomics analysis was performed on CHO Spike, RBD, T1, and T4 to compare glycan distributions among the various forms of Spike. As a first inspection of the glycosylation profiles, glycans on each site were clustered according to the following types: undecorated, high mannose, fucosylated, sialyalated, and sialofucosylated (Fig. 1). Abundances were normalized with respect to the specific glycosylation site.
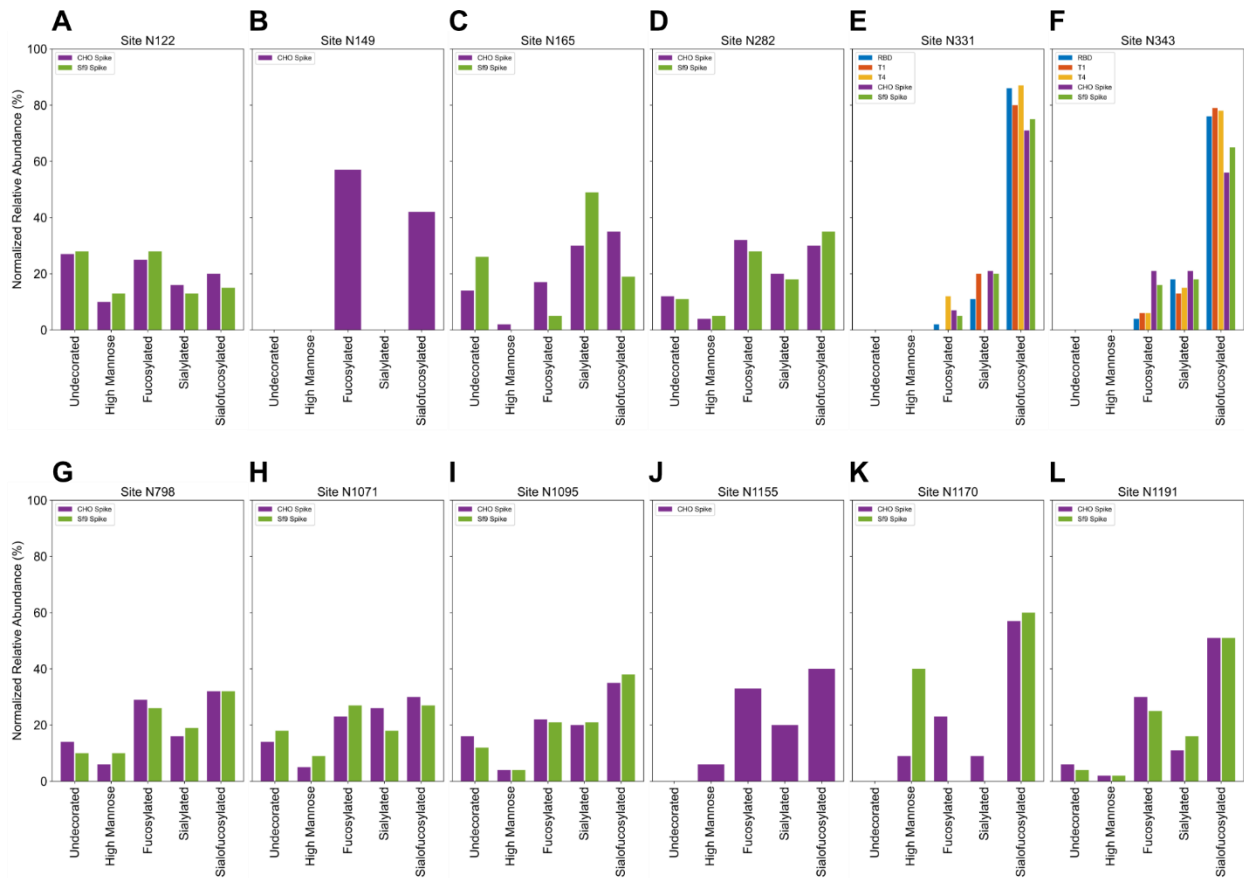
Figure 1. Bar clusters of glycan types on site (**A**) N122, (**B**) N149, (**C**) N165, (**D**) N282, (**E**) N331, (**F**) N343, (**G**) N798, (**H**) N1071, (**I**) N1095, (**J**) N1155, (**K**) 1170, (**L**) N1191. Legends indicate the protein, and the absence of a protein in any of the charts indicates that the protein does not contain the glycosylation site or glycosylation was not detected at that site.

Initial analysis of the glycans reveals that distribution of glycan types are similar across Spike proteins and suggests that the glycosylation distributions may be highly similar for all proteins where glycans were detected. Particularly for Spike produced in CHO and Sf9 systems, which have many glycosylation sites present for comparison, the glycan types are remarkably similar. Different glycosylation sites also appear to have diverse but nonrandom distributions of glycan types, based on the similarities of the two Spike proteins. For instance, a large fraction of glycans at N122 are undecorated and high mannose, which is consistent between CHO and Sf9 Spike

(Fig. 1A). Residue 122 is not within the residue ranges of T1 and T4, and therefore no glycans are present for the truncations. In contrast, glycans at site N331 and N343 are heavily sialofucosylated, which is also the same for the Spike truncations (Fig. 1E and 1F). N331 and N343 are of particular interest as a point of comparison, as these sites are also present on both truncations. A couple of sites have slightly different glycan type distributions, such as N1170. At N1170, CHO has more complex glycans compared to Sf9 Spike, with less high mannose glycans and more fucosylated and sialylated glycans (Fig. 1K). Some glycosylation sites were not detected to be glycosylated with the methodology used in this work.

As the glycan types are similar for all Spike proteins, we expected that the full glycosylation profiles with individual glycans would also be similar. However, the deeper inspection of the glycosylation distribution reveals that there are differences in the specific glycans that occupy each protein, which are shown in clustered heatmaps. Sites N331 and N343 are explored in detail, which are contained in all proteins studied in this work (Fig. 2A and 2B). All other glycosylation sites for Spike are shown in Appendix C (Fig. A-9).
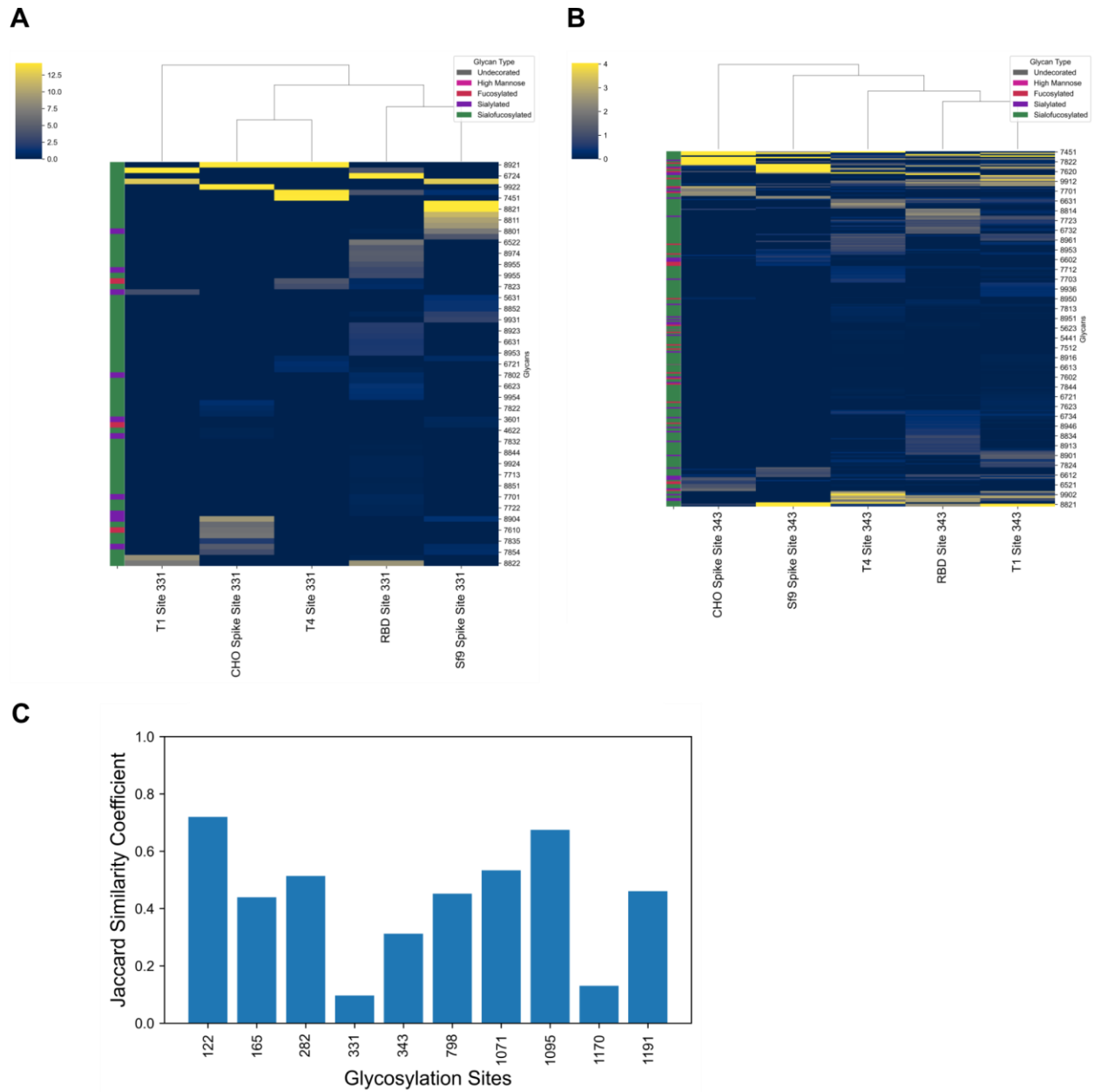
Figure 2. Comparisons on N331 and N343 glycosylation. (**A**) Heatmaps of N331 glycosylation. (**B**) Heatmaps of N343 glycosylation. (**C**) Jaccard similarity coefficient comparing CHO and Sf9 Spike.

The clustered heatmaps for sites N331 and N343 clearly indicate that the clusters of glycans on each protein are largely absent in the other proteins, which is unexpected since the glycan types are similar across proteins. When comparing all glycosylation sites on the full-length Spike proteins, similarity varies considerably depending on the glycosylation site, as shown by Jaccard similarity coefficients on each site (Fig. 2C). Overall, however, the types of glycans are similar among the proteins and suggest that the proteins have similar glycosylation profiles.

Binding to human ACE2-Fc

Binding kinetics of Spike truncations to human ACE2-Fc were measured using biolayer interferometry, which is a label-free assay that measures the rate of protein binding and unbinding through a shift in interference pattern of white light (Fig. 4).
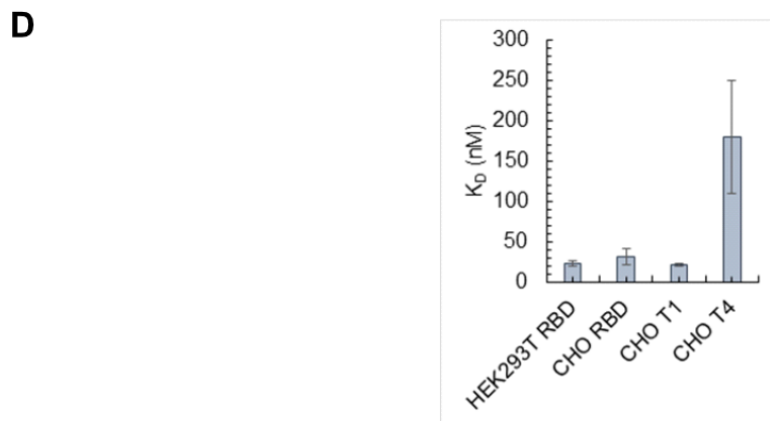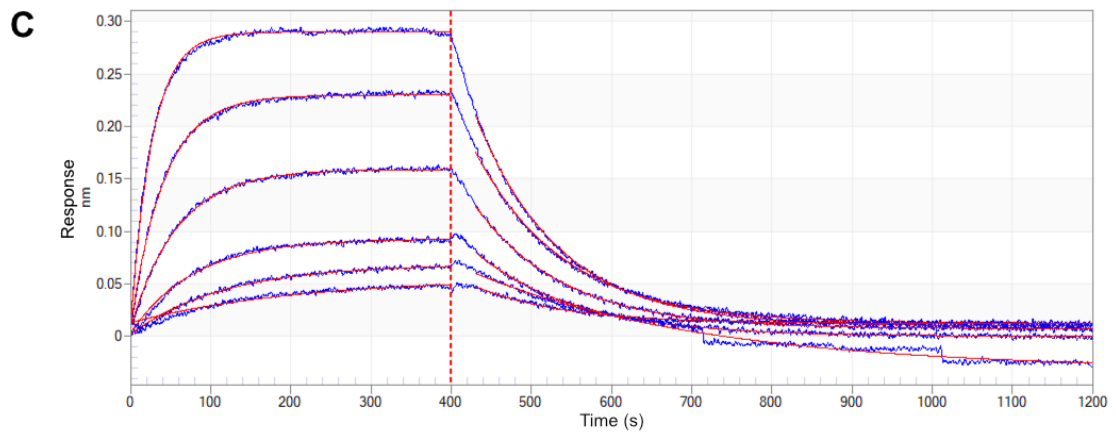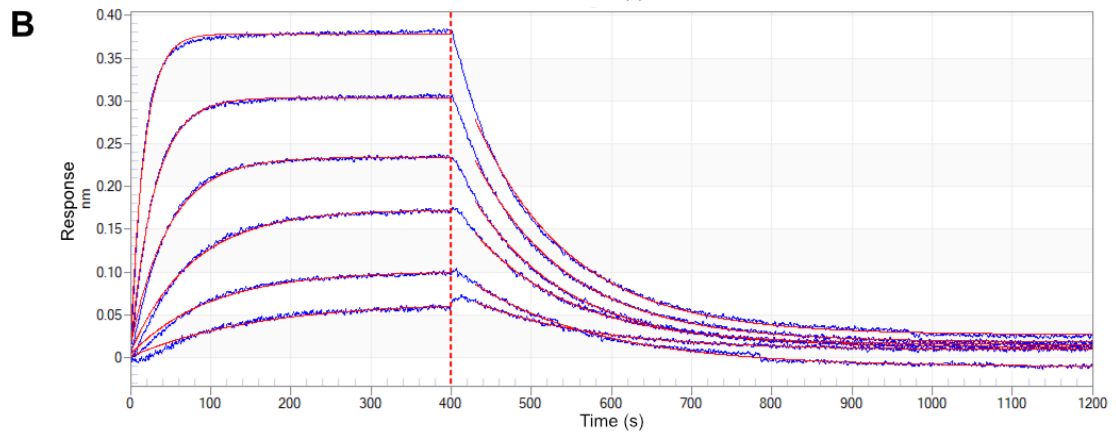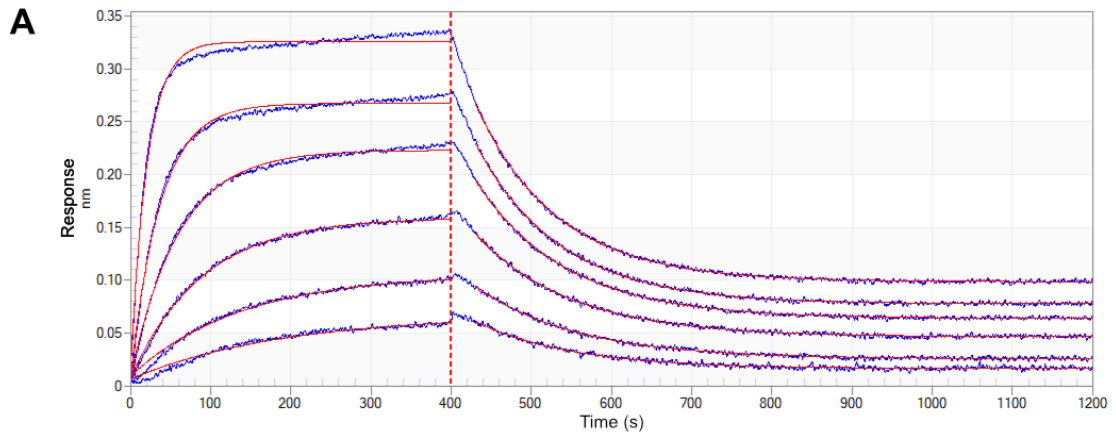
Figure 4. Biolayer interferometry on RBD, T1, and T4. Processed sensorgrams plotting the response over time on (**A**) RBD, (**B**) T1, and (**C**) T4. (**D**) Steady-state analysis was performed on the data from 390-395 s. Error bars represent standard error of the mean.

RBD and T1 both have similar binding affinities to ACE2-Fc with K_D values of approximately 20-30 nM. T4 has a lower binding affinity, with a K_D value of 180 nM. K_D values for full-length Spike could not be obtained because of lack of dissociation of Spike from ACE2-Fc (data not shown), but it is expected that Spike has higher binding affinity to ACE2-Fc than does RBD, and therefore higher binding affinity than the truncations as well.

Discussion

Production of Spike fragments is important for its use in diagnostics, protein subunit vaccines, and research. In addition, high affinities of the Spike fragments are critical in these applications. Several approaches have been used to increase Spike yields, including stabilizing mutations [4], comparison of different cell lines [5], and optimization of production conditions, such as temperature shifts [17]. Here, we expressed full-length Spike and RBD transiently in CHO cells to determine the intracellular and extracellular production kinetics. In addition, we developed 8 truncations in pursuit of a truncation which exhibits both high expression and binding to antibodies.

The regions of Spike that cause lower expression and higher sensitivity compared to RBD are not known, but the initial screen of the truncations showed that T1 is highly expressed and secreted compared to other truncations, with T4 following at much lower titers (Fig. 3C and 3D). This suggests that residues downstream of T1 may be contributing to decreased titers. Comparing T1

to T2 and T3, residues upstream of T1 also appear to decrease titers. The additional residues in T2, T3, and T4 contain predicted glycosylation sites, which may introduce avenues for protein retention such as incomplete glycosylation. This is supported by the lysate proteins running at their expected molecular weights and the supernatant proteins much higher, though protein size did not appear to correlate with relative retention in the cell (Fig. 3B). In contrast, T1 only contains the same glycosylation sites as RBD and was found in the crude at much higher titers compared to other truncations. Interestingly, the MD simulations of T3 suggest that lower stability may result from unexpected intramolecular glycan-protein interactions for fully glycosylated truncations.

In the ELISA sensitivity assays for CR3022 and the PAb, CHO-expressed Spike has higher AUCs for both antibodies compared to Sf9-expressed Spike (Fig. 5). The discrepancy may be due to different glycosylation profiles, which would be consistent with the idea that CHO-expressed proteins tend to have more human-like glycosylation patterns [18]. We also found that CHO Spike produces higher signal than RBD when probed with the PAb, consistent with results from serological assays (Amanat, 2020). This was also expected because polyclonal antibodies target multiple epitopes, and full-length Spike may contain more binding epitopes than RBD. Surprisingly, T1 and T4 have higher sensitivities to the PAb, outperforming full-length Spike. One possibility is that T1 and T4 contain an additional epitope, not present on RBD, that has high affinity but is sterically hindered when additional residues are present. This may also be the reason for the higher performance of T1 over T4. Visualization of binding through crystallography and analysis of binding kinetics and thermodynamics through methods such as steered MD may elucidate the reason for their high affinities.

Different glycosylation sites on Spike were determined to have distinct glycosylation profiles, and the distinct glycans on different glycosylation sites may be important for the structure and function of Spike, as CHO and Sf9 Spike both had similar glycosylation profiles across the analyzed sites.

Not all glycosylation sites were detected to have glycans, but a combination of proteases may allow detection of glycans on the remaining sites[15]. Spike and its truncations were determined to have similar glycan types, but there were differences in the diversity of glycosylation distributions and the specific glycan types that were most abundant for each protein. For instance, on site N343, full-length Spike had a much tighter glycan distribution compared to the truncations. N343 glycosylation mediates RBD conformational changes from the down to up state, and the dynamics may be influenced by the specific glycan that occupies the site. Glycoengineering followed by structural analysis or molecular dynamics studies may help discern whether specific glycans are required on Spike for optimal conformational dynamics.

Biolayer interferometry on T1 and T4 shows that T1 has a similar binding affinity as RBD, but T4 has a lower binding affinity than T1 and RBD. Given that the abundances of the types of glycans between the truncations are highly similar (Fig. 2A and 2B), different glycosylation may not be the reason for the higher binding affinity of full-length Spike to ACE2-Fc. However, the specific glycans on the truncations are different, and further investigation by methods such as steered molecular dynamics is needed to determine whether specific glycans enhance binding to ACE2-Fc.

Conclusion

We expressed SARS-CoV-2 Spike and RBD in CHO cells and optimized harvest dates. Additionally, we expressed 8 new truncations and found that T1 and T4 have high expression and secretion, where T1 has even higher expression than RBD. T1 and T4 also have higher binding sensitivity to a PAb compared to Spike. Overall, T1 had the highest performance in all expression and binding experiments conducted in this work. Its high expression and sensitivity suggest T1 may be a promising Spike alternative in research and clinical applications. Further work is needed

to understand why T1 has higher affinity to antibodies and whether the higher affinity translates to assays with convalescent sera.

Methods

Plasmids

pCAGGS-Spike and pCAGGS-RBD were gifted from Florian Krammer [3]. Spike and RBD both contain an N-terminal signal peptide for secretion and a hexahistidine (6x His) tag for purification. Spike-1 and RBD-1 contain the signal sequence MFVFLVLLPLVSSQ. Spike-2 and RBD-2 contain the signal sequence MEFGLSWLFLVAILKGVQC. Spike has two stabilizing mutations (K983P and V984P), and its polybasic furin site has been removed (RRAR to R). Truncations T1-T8 were synthesized (GenScript, Piscataway, NJ) with overhangs for insertion into pCAGGS vectors (Table S1). Truncations were inserted into pCAGGS vectors via Gibson Assembly of pCAGGS-RBD digested with XbaI and XhoI. Spike truncations were designed by adding increments of approximately 50 amino acids to the N- and/or C-termini of RBD. Each truncation includes an N-terminal signal peptide and a C-terminal 6x His tag. Possible structural and binding motifs for the truncations were determined with PredictProtein [19]. Starting and ending residues were selected to avoid interrupting major secondary structures present in Spike [20–22].

Cell culture and transfection

ExpiCHO-S cells (Thermo Fisher Scientific, Waltham, MA) were maintained in a 125 mL vented shake flask with 30 mL of culture in ExpiCHO Expression Medium (Thermo Fisher Scientific). Cells were cultured in a humidified incubator at 37°C and 8% CO_2, on a 19 mm shaking diameter orbital shaker at 120 rpm (Ohaus, Parsippany, NJ).

For transfection in 125 mL shake flasks, cultures were transfected using the Expifectamine CHO Transfection Kit (Thermo Fisher Scientific), following manufacturer instructions for the Standard Protocol. For time course experiments, 0.5 mL of culture was harvested each day. Viable cell densities were measured using trypan blue and a TC20 automated cell counter (Bio-Rad, Hercules, California). Samples were harvested by centrifuging at 300 rcf for 5 min and collecting the supernatant. For samples to undergo purification, entire cultures were centrifuged at 4,000 rcf for 20 minutes at 4°C and filtered through 0.22 µm filters. For transfection in 2.0 mL 96-well deep well blocks (Genesee Scientific, El Cajon, CA), 0.8 mL of cells at 6×10^6 cells/mL were plated on the day of transfection. Cells were cultured on a 3 mm shaking diameter orbital shaker at 900 rpm (Benchmark Scientific, Sayreville, NJ) and transfected according to manufacturer instructions. Samples were harvested 5 days post-transfection by centrifuging the cultures at 300 rcf for 5 minutes and collecting the supernatant.

Protein purification and concentration

Filtered samples were column purified using an AKTA Pure fast protein liquid chromatography (FPLC) system with a 5 mL prepacked Ni Sepharose HP column (Cytiva, Marlborough, MA), using imidazole to elute the proteins [23]. Samples were loaded onto the column at a flow rate of 5 mL/min, the resin was washed for 10 column volumes (CV), and proteins were eluted using imidazole. Detailed procedures are available in Appendix C. Purified proteins were dialyzed against phosphate-buffered saline (PBS) using dialysis cassettes at 4°C (Thermo Fisher Scientific). Spike was dialyzed with a 20 kDa molecular weight cutoff (MWCO) membrane. RBD, T1, and T4 were dialyzed with 10 kDa MWCO membranes. Dialyzed samples were concentrated using centrifugal filter units (Millipore Sigma, Burlington, MA) at 4,000 rcf for 20 minutes at 4°C. Spike was concentrated using centrifugal filter units with a MWCO of 30 kDa. RBD, T1, and T4 were concentrated with 10 kDa MWCO centrifugal filter units.

SDS-PAGE and western blot

Samples from time course experiments and the truncation screening were prepared for sodium dodecyl-sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) by adding 12 µL of NuPAGE LDS Sample Buffer (Thermo Fisher Scientific) and 3 µL of tris(2-carboxyethyl)phosphine (Thermo Fisher Scientific) to 30 µL of sample. Mixtures were heated at 95°C for 10 minutes and 10 µL of samples were loaded into gels cast in-house, with a 12% acrylamide resolving layer and 4% acrylamide stacking layer. Samples were run through the gel for 15 minutes at 115 V, then 50 minutes at 150 V. Proteins were transferred onto polyvinylidene difluoride membranes in a wet sandwich and membranes were blocked using 5% non-fat milk. Membranes were stained overnight at 4°C with a 1:1000 diluted mouse anti-his primary antibody (MCA1396, RRID:AB_322084, Bio-Rad) and then for 1 hour at room temperature with a 1:4000 diluted rabbit anti-mouse HRP secondary antibody (SouthernBiotech Cat# 6170-05, RRID:AB_2796243, Birmingham, AL). Membranes were developed using Pierce ECL Western Blotting Substrate (Thermo Fisher Scientific) and imaged using an Amersham Imager 600 (Cytiva).

Purified samples were analyzed by SDS-PAGE with a method previously described [24]. Images of the gels were taken using a ChemiDoc Imaging system (Bio-Rad), and proteins were transferred onto nitrocellulose membranes using Trans-Blot Turbo Packs (Bio-Rad) and Trans-Blot Turbo System (Bio-Rad). Membranes were blocked overnight in 1% casein, stained with 1:1000 diluted mouse anti-his primary antibody and stained with 1:4000 diluted rabbit anti-mouse secondary antibody. The chemiluminescent reactions were performed using Clarity ECL Substrate (Bio-Rad).

Concentrations for purified proteins were estimated using a combination of ELISA, Bradford Assay, and scanning densitometry on SDS-PAGE gels. Spike and RBD concentrations were first calculated using sandwich ELISA. Purified T1 and T4 concentrations were determined using

Bradford Assay since ELISA standard curves could not be generated for these novel protein truncations. Next, 1 µg of proteins, as determined by the two methods, were loaded into each lane of a 4%-20% gradient stain-free gel (Bio-Rad). Dilutions of RBD standard from 1.5 µg to 0.5 µg were also loaded into the gel. Samples were run at 200 V for 36 minutes and imaged using a ChemiDoc imaging system (Bio-Rad). A standard curve was generated via densitometry through ImageJ, and primary band intensities for the samples were interpolated to quantify concentrations.

Enzyme-linked immunosorbent assay (ELISA)

Sandwich ELISAs were performed to quantify purified Spike and RBD and crude supernatants. 1:1000 mouse anti-his capture antibody in PBS was coated onto Immulon 2 HB 96-well plates (Thermo Fisher Scientific) at 4°C overnight. Plates were blocked with 200 µL/well PBS with 3% BSA for 30 minutes. Plates were loaded with serial dilutions of purified protein samples or crude supernatants. Plates were incubated with 1:1000 rabbit anti-RBD primary antibody (Sino Biological Cat# 40592-R001, RRID:AB_2857936, Wayne, PA), then 1:6000 or 1:4000 goat anti-rabbit, HRP secondary antibody (SouthernBiotech Cat# 4030-05, RRID:AB_2687483) in PBS with 1% BSA for purified or crude proteins, respectively. Plates were developed with 1-step Turbo TMB-ELISA Substrate Solution (Thermo Fisher Scientific) and 2N HCl. Absorbance at 450 nm was measured using a Spectramax 250 spectrophotometer (Molecular Devices, San Jose, CA). Plates were washed 3 times with 200 µL/well PBS with 0.05% Tween20 (PBS-T) between each step and incubations were performed using volumes of 100 µL/well for 1 hour at room temperature unless specified otherwise. Standard curves for quantifying Spike and RBD were generated using serial dilutions of Sf9 insect Spike (NR-52308, BEI Resources, Manassas, VA) and HEK293F human RBD (NR-52366, BEI Resources), respectively.

Indirect ELISAs were performed to assess the sensitivities of CHO-expressed proteins to a human anti-Spike monoclonal antibody CR3022 (NR-52392, BEI Resources, RRID:AB_2848080) and a

rabbit anti-Spike polyclonal antibody (PAb, eEnzyme, SCV2-S-100, RRID:AB_2893135, Gaithersburg, MD). For CR3022, antigens were first coated onto plates at 4°C overnight. After blocking, serial dilutions of CR3022 in PBS with 1% BSA were loaded from 100 ng/well. Plates were loaded with 100 µL/well goat anti-hIgG, HRP secondary antibody at 1:4000 in PBS containing 1% BSA. For the PAb, 3-fold serial dilutions starting at 400 ng/well of rabbit anti-Spike primary antibody were used (PAb, SCV2-S-100, eEnzyme), and a 1:4000 goat anti-rabbit IgG, HRP secondary antibody was used instead.

Bradford assay

Bradford assays were performed to quantify the concentration of total soluble protein (TSP) by using a protein assay dye reagent (Bio-Rad). For each BSA standard, sample, and diluted sample, 10 µL/well of sample and 190 µL/well of Bradford dye were loaded into 96-well plates. After incubating for 10 minutes at room temperature, the absorbances of standards and samples were measured at 450 nm and 590 nm [25], using a Spectramax M4 spectrophotometer (Molecular Devices). Standard curves for quantifying samples were generated by using serial dilutions of BSA from 0-0.5 mg/mL with 0.05 mg/mL steps.

Liquid chromatography-tandem mass spectrometry (LC-MS/MS)

Sequences of purified T1 and T4 were obtained via LC-MS/MS. 10 µg of T1 and 20 µg of T4 were run on a 4%-20% gradient SDS-PAGE gel. Bands were extracted and submitted to the Genome Center Proteomics Core at the University of California, Davis. Briefly, proteins were digested with tryspin and analyzed on a Dionex UltiMate 3000 RSLC system (Thermo Fisher Scientific) using a PepSep (PepSep, Denmark) ReproSil 8 cm 150 µm I.D. C18 column with 1.5 µm particle size (120 Å pores). Mass spectra analysis is described in Appendix C. Searches were conducted

against the known sequences of T1 and T4, and alignments were performed using Multiple Alignment using Fast Fourier Transform [26].

Circular dichroism (CD)

Concentrated samples were prepared for CD analysis by diluting 150 µg of protein in 50% PBS and 50% CD buffer (25mM of phosphate and 40mM of NaF). Single spectrum data were obtained using a Jasco J-715 CD spectrometer (Jasco, Easton, MD). Data were analyzed using BeStSel [27]. Spectra of buffer were subtracted before analysis. To obtain secondary structure data for the PDB Spike structure, the PDB file 6VXX was analyzed using the STRIDE server [28].

Simulations

Starting configurations for molecular dynamics (MD) simulations were obtained by trimming the full Spike protein structure obtained from the protein data bank (6VXX). Structures were reduced to a single monomer and cut at the amino acid sequences corresponding to RBD, T1, T3, and T4. 6x His tags were added using modeller [29], which modifies amino acid sequences of proteins. The new His-tagged structures were prepared and had glycans attached using Glycam [30]. The N-glycosylation sites of RBD and the RBD portion of all truncations had the glycoform FA2 attached. T1 contained no additional N-glycosylation sites, T3 contained an additional FA3 glycoform, and T4 contained an additional M5 glycoform. Amber ff14SB and Glycam06 forcefields [31,32] were used and generated using acpype.py following the method shown previously [33,34]. Simulations were conducted using the Gromacs 2019.1 suite with similar energy minimization procedure as in previous simulations [35–37] including ones involving glycosylated RBD [12,33]. Simulation runs after equilibration were carried out for 100 ns.

Glycoproteomic Analysis With LC-MS/MS

Details of protein digestion for the glycoproteomics analysis have been described previously[38,39]. Briefly, buffer exchange was performed using a molecular weight (3K) cutoff spin column (Merck Millipore, MA) to remove salts and dilute recombinant proteins with 50 uL of 50 mM ammonium bicarbonate solution, then the proteins were reduced with 2 µl of 550 mM dithiothreitol and alkylated with 4 µl of 450 mM iodoacetamide. The samples were incubated with trypsin at 37°C for 18 hours. The resulting peptides and glycopeptides were dried using a miVac (SP Scientific, PA, United States) prior to mass spectrometry analysis.

The peptide and glycopeptide samples were reconstituted with nanopure water and directly characterized using UltiMate™ WPS-3000RS nanoLC 980 system coupled to the Nanospray Flex ion source of an Orbitrap Fusion Lumos Tribrid Mass Spectrometer system (Thermo Fisher Scientific, MA, United States). The analytes were separated on an Acclaim™ PepMap™ 100 C18 LC Column (3 µm, 0.075 mm × 150 mm, ThermoFisher Scientific). A binary gradient was applied using 0.1% (v/v) formic acid in (A) water and (B) 80% acetonitrile: 0–5 min, 4–4% (B); 5–133 min, 4–32% (B); 133–152 min, 32%–48% (B); 152–155 min, 48–100% (B); 155–170 min, 100–100% (B); 170–171 min, 100–4% (B); 171–180 min, 4–4% (B). The instrument was run in data-dependent mode with 1.8 kV spray voltage, 275°C ion transfer capillary temperature, and the acquisition was performed with the full MS scanned from 700 to 2000 m/z in positive ionization mode. Stepped higher-energy C-trap dissociation (HCD) at 30 ± 10% was applied to obtain tandem MS/MS spectra with m/z values starting from 120.

Glycopeptide fragmentation spectra were annotated using Byonic software (Protein Metrics, CA, United States) against the protein sequences. Common modifications, including cysteine carbamidomethyl, methionine oxidation, asparagine deamidation and glutamine deamidation

were assigned. A published in-house N-glycan librarywas utilized for the glycopeptide identification, and relative abundance values were calculated based on the glycopeptide precursor peak areas [38].

The full glycosylation profiles are available in Appendix C (Figure S1). Glycans were clustered into undecorated, high mannose, fucosylated, sialylated, and sialofucosylated types according to the following rules: undecorated glycans contain less than 5 mannose, no fucose, and no sialic acid. High mannose glycans contain at least 5 mannose, no fucose, and no sialic acid. Fucosylated glycans contain at least 1 fucose. Sialylated glycans contain at least 1 sialic acid. Sialofucosylated glycans contain at least 1 fucose and 1 sialic acid.

Heatmaps and dendrograms were made using the seaborn library in Python 3.8, using the average method for clustering.

Biolayer Interferometry

Purified and concentrated proteins were prepared for biolayer interferometry similarly to as previously described[12]. Anti h-IgG-Fc biosensors were used to immobilize ACE2-Fc, and interactions with the analytes HEK293T RBD, CHO RBD, T1, and T4 were performed using 2x serial dilutions of analytes from 250 nM to 7.81 nM. Steady-state analysis was performed using the response from 390 to 395 s.

References

1. Huang, Y., Yang, C., Xu, X., Xu, W., & Liu, S. (2020). Structural and functional properties of SARS-CoV-2 spike protein: Potential antivirus drug development for COVID-19. *Acta Pharmacologica Sinica*, *41*(9), 1141–1149. https://doi.org/10.1038/s41401-020-0485-4

2. Heath, P. T., Galiza, E. P., Baxter, D. N., Boffito, M., Browne, D., Burns, F., Chadwick, D. R., Clark, R., Cosgrove, C., Galloway, J., Goodman, A. L., Heer, A., Higham, A., Iyengar, S., Jamal, A., Jeanes, C., Kalra, P. A., Kyriakidou, C., McAuley, D. F., … Toback, S. (2021). Safety and Efficacy of NVX-CoV2373 Covid-19 Vaccine. *New England Journal of Medicine*, *0*(0), null. https://doi.org/10.1056/NEJMoa2107659

3. Amanat, F., Stadlbauer, D., Strohmeier, S., Nguyen, T. H. O., Chromikova, V., McMahon, M., Jiang, K., Arunkumar, G. A., Jurczyszak, D., Polanco, J., Bermudez-Gonzalez, M., Kleiner, G., Aydillo, T., Miorin, L., Fierer, D. S., Lugo, L. A., Kojic, E. M., Stoever, J., Liu, S. T. H., … Krammer, F. (2020). A serological assay to detect SARS-CoV-2 seroconversion in humans. *Nature Medicine*, *26*(7), 1033–1036. https://doi.org/10.1038/s41591-020-0913-5

4. Hsieh, C.-L., Goldsmith, J. A., Schaub, J. M., DiVenere, A. M., Kuo, H.-C., Javanmardi, K., Le, K. C., Wrapp, D., Lee, A. G., Liu, Y., Chou, C.-W., Byrne, P. O., Hjorth, C. K., Johnson, N. V., Ludes-Meyers, J., Nguyen, A. W., Park, J., Wang, N., Amengor, D., … McLellan, J. S. (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science (New York, N.y.)*, eabd0826. https://doi.org/10.1126/science.abd0826

5. Stuible, M., Gervais, C., Lord-Dufour, S., Perret, S., L'Abbé, D., Schrag, J., St-Laurent, G., & Durocher, Y. (2021). Rapid, high-yield production of full-length SARS-CoV-2 spike ectodomain by transient gene expression in CHO cells. *Journal of Biotechnology*, *326*, 21–27. https://doi.org/10.1016/j.jbiotec.2020.12.005

6. Smaoui, M. R., & Yahyaoui, H. (2021). Unraveling the stability landscape of mutations in the SARS-CoV-2 receptor-binding domain. *Scientific Reports*, *11*(1), 9166. https://doi.org/10.1038/s41598-021-88696-5

7. Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D., & Bloom, J. D. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on

Folding and ACE2 Binding. *Cell*, *182*(5), 1295-1310.e20. https://doi.org/10.1016/j.cell.2020.08.012

8. Ren, W., Sun, H., Gao, G. F., Chen, J., Sun, S., Zhao, R., Gao, G., Hu, Y., Zhao, G., Chen, Y., Jin, X., Fang, F., Chen, J., Wang, Q., Gong, S., Gao, W., Sun, Y., Su, J., He, A., … Sun, L. (2020). Recombinant SARS-CoV-2 spike S1-Fc fusion protein induced high levels of neutralizing responses in nonhuman primates. *Vaccine*, *38*(35), 5653–5658. https://doi.org/10.1016/j.vaccine.2020.06.066

9. Yuan, M., Wu, N. C., Zhu, X., Lee, C.-C. D., So, R. T. Y., Lv, H., Mok, C. K. P., & Wilson, I. A. (2020). A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science*, *368*(6491), 630–633. https://doi.org/10.1126/science.abb7269

10. Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*, *181*(2), 281-292.e6. https://doi.org/10.1016/j.cell.2020.02.058

11. Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/bip.360221211

12. Huang, Y., Harris, B. S., Minami, S. A., Jung, S., Shah, P., Nandi, S., McDonald, K., & Faller, R. (2021). *SARS-Cov-2 Spike binding to ACE2 is stronger and longer ranged due to glycan interaction* (p. 2021.07.15.452507). https://doi.org/10.1101/2021.07.15.452507

13. Shental-Bechor, D., & Levy, Y. (2008). Effect of glycosylation on protein folding: A close look at thermodynamic stabilization. *Proceedings of the National Academy of Sciences*, *105*(24), 8256–8261. https://doi.org/10.1073/pnas.0801340105

14. Casalino, L., Gaieb, Z., Goldsmith, J. A., Hjorth, C. K., Dommer, A. C., Harbison, A. M., Fogarty, C. A., Barros, E. P., Taylor, B. C., McLellan, J. S., Fadda, E., & Amaro, R. E. (2020). Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Central Science*, *6*(10), 1722–1734. https://doi.org/10.1021/acscentsci.0c01056

15. Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S., & Crispin, M. (2020). Site-specific glycan analysis of the SARS-CoV-2 spike. *Science*, *369*(6501), 330–333. https://doi.org/10.1126/science.abb9983

16. Gong, Y., Qin, S., Dai, L., & Tian, Z. (2021). The glycosylation in SARS-CoV-2 and its receptor ACE2. *Signal Transduction and Targeted Therapy*, *6*(1), Article 1. https://doi.org/10.1038/s41392-021-00809-8

17. Johari, Y. B., Jaffé, S. R. P., Scarrott, J. M., Johnson, A. O., Mozzanino, T., Pohle, T. H., Maisuria, S., Bhayat-Cammack, A., Lambiase, G., Brown, A. J., Tee, K. L., Jackson, P. J., Wong, T. S., Dickman, M. J., Sargur, R. B., & James, D. C. (2021). Production of trimeric SARS-CoV-2 spike protein by CHO cells for serological COVID-19 testing. *Biotechnology and Bioengineering*, *118*(2), 1013–1021. https://doi.org/10.1002/bit.27615

18. Esko, J. D., & Stanley, P. (2015). Glycosylation Mutants of Cultured Mammalian Cells. In A. Varki, R. D. Cummings, J. D. Esko, P. Stanley, G. W. Hart, M. Aebi, A. G. Darvill, T. Kinoshita, N. H. Packer, J. H. Prestegard, R. L. Schnaar, & P. H. Seeberger (Eds.), *Essentials of Glycobiology* (3rd ed.). Cold Spring Harbor Laboratory Press. http://www.ncbi.nlm.nih.gov/books/NBK453088/

19. Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., Hönigschmid, P., Schafferhans, A., Roos, M., Bernhofer, M., Richter, L., Ashkenazy, H., Punta, M., Schlessinger, A., Bromberg, Y., Schneider, R., Vriend, G., Sander, C., Ben-Tal, N., & Rost, B. (2014). PredictProtein—An open resource for online prediction of protein structural and functional features. *Nucleic Acids Research*, *42*(W1), W337–W343. https://doi.org/10.1093/nar/gku366

20. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015.053

21. Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C., & Ferrin, T. E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics*, *7*(1), 339. https://doi.org/10.1186/1471-2105-7-339

22. Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., & Ferrin, T. E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science*, *30*(1), 70–82. https://doi.org/10.1002/pro.3943

23. Esposito, D., Mehalko, J., Drew, M., Snead, K., Wall, V., Taylor, T., Frank, P., Denson, J.-P., Hong, M., Gulten, G., Sadtler, K., Messing, S., & Gillette, W. (2020). Optimizing high-yield production of SARS-CoV-2 soluble spike trimers for serology assays. *Protein Expression and Purification*, *174*, 105686. https://doi.org/10.1016/j.pep.2020.105686

24. Xiong, Y., Karuppanan, K., Bernardi, A., Li, Q., Kommineni, V., Dandekar, A. M., Lebrilla, C. B., Faller, R., McDonald, K. A., & Nandi, S. (2019). Effects of N-Glycosylation on the Structure, Function, and Stability of a Plant-Made Fc-Fusion Anthrax Decoy Protein. *Frontiers in Plant Science*, *10*, 768. https://doi.org/10.3389/fpls.2019.00768

25. Ernst, O., & Zor, T. (2010). Linearization of the Bradford Protein Assay. *Journal of Visualized Experiments : JoVE*, *38*, 1918. https://doi.org/10.3791/1918

26. Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166. https://doi.org/10.1093/bib/bbx108

27. Micsonai, A., Wien, F., Kernya, L., Lee, Y.-H., Goto, Y., Réfrégiers, M., & Kardos, J. (2015). Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proceedings of the National Academy of Sciences*, *112*(24), E3095–E3103. https://doi.org/10.1073/pnas.1500851112

28. Frishman, D., & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, *23*(4), 566–579. https://doi.org/10.1002/prot.340230412

29. Webb, B., & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, *54*(1), 5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3

30. Woods, R. (2005). *GLYCAM*. Complex Carbohydrate Research Center, University of Georgia. http://glycam.org

31. Kirschner, K. N., Yongye, A. B., Tschampel, S. M., González-Outeiriño, J., Daniels, C. R., Foley, B. L., & Woods, R. J. (2008). GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *Journal of Computational Chemistry*, *29*(4), 622–655. https://doi.org/10.1002/jcc.20820

32. Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., & Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, *11*(8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255

33. Bernardi, A., Kirschner, K. N., & Faller, R. (2017). Structural analysis of human glycoprotein butyrylcholinesterase using atomistic molecular dynamics: The importance of glycosylation site ASN241. *PLOS ONE*, *12*(11), e0187994. https://doi.org/10.1371/journal.pone.0187994

34. Bernardi, A., Faller, R., Reith, D., & Kirschner, K. N. (2019). ACPYPE update for nonuniform 1–4 scale factors: Conversion of the GLYCAM06 force field from AMBER to GROMACS. *SoftwareX*, *10*, 100241. https://doi.org/10.1016/j.softx.2019.100241

35. Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

36. Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B., & Lindahl, E. (2013). GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, *29*(7), 845–854. https://doi.org/10.1093/bioinformatics/btt055

37. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. C. (2005). GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*, *26*(16), 1701–1718. https://doi.org/10.1002/jcc.20291

38. Li, Q., Xie, Y., Wong, M., Barboza, M., & Lebrilla, C. B. (2020). Comprehensive structural glycomic characterization of the glycocalyxes of cells and tissues. *Nature Protocols*, *15*(8), Article 8. https://doi.org/10.1038/s41596-020-0350-4

39. Kasper, D. M., Hintzen, J., Wu, Y., Ghersi, J. J., Mandl, H. K., Salinas, K. E., Armero, W., He, Z., Sheng, Y., Xie, Y., Heindel, D. W., Joo Park, E., Sessa, W. C., Mahal, L. K., Lebrilla, C., Hirschi, K. K., & Nicoli, S. (2020). The N-glycome regulates the endothelial-to-hematopoietic transition. *Science (New York, N.Y.)*, *370*(6521), 1186–1191. https://doi.org/10.1126/science.aaz2121

Chapter 5: *Conclusions and Future Work*

Chapters 2 through 4 describe applications of protein engineering for light-inducible CRISPR-based transcription regulation and an engineering effort of SARS-CoV-2 Spike for higher expression and antibody binding. In the context of these two works, we reflect upon their contributions to protein engineering and consider implications for the future.

Optogenetics

Optogenetics is a relatively new field that is still growing, and the LACE system allows a novel mechanism for gene expression regulation. Demonstrating the effectiveness of the system in CHO cells also expands the utility of optogenetics, as CHO cells are a model system for industrial biotechnology. The LACE system is also orthogonal to other methods such as chemical additions to media and changes in environmental conditions, indicating that these approaches are complementary and may be used in conjunction for complex systems. One limitation discussed in Chapter 2 is the LACE system is sensitive to the stoichiometry of its components that are expressed in the cell, which is controlled by regulating the ratios of plasmid transfection during a simultaneous 4-plasmid transfection procedure. However, the need to transfect four separate plasmids causes high variability that decreases the population of cells that express optimal proportions of the components, or even receives all components of the LACE system. The upper limit on efficiency of LACE activation is dictated by the transfection efficiency raised to the fourth power. To improve upon the LACE system, we are developing a 2-plasmid expression system for utilizing the LACE system (2pLACE), which improves the efficiency and consistency.

Our previous efforts to create a stable expression system for the four plasmid LACE system (4pLACE) using lentiviral vectors were hampered due to leaky expression from lentiviral vectors

in the dark state. We therefore pursued a stable plasmid strategy by including the selection markers Hygromycin and Bleocin into each plasmid. As the plasmid ratios of CRY2 and eGFP are now fixed to be equal, as are the ratios of the ratios of CIBN and gRNA, the transfection ratios of each plasmid must be reoptimized. Expected results of the 2pLACE system are that the fraction of cells that are positive for the 2pLACE system will be higher, as only two plasmids need to be transfected. The response of cells to light are expected to be more uniform and have a tighter distribution, since the expression of the LACE components will be more uniform. Similarly, with more cells having optimal ratios of plasmids, the dynamic range of 2pLACE is expected to be higher than that of 4pLACE. However, a slight penalty to the dynamic range is possible, since the stoichiometry of the 2pLACE system is not as flexible as that of the 4pLACE system.

The light sensitive proteins, CRY2 and CIBN, are also under constant development, where mutations, truncations, and other amino acid sequence modifications are made in pursuit of higher stability and different kinetics. The proteins used in this system have activation and deactivation times on the order of seconds and minutes, respectively, but faster or slower kinetics may be desired for various applications[1]. The ability to design and produce these engineered proteins further extends the capabilities of synthetic biology. Ultimately, *de novo* design of proteins for specific functions may allow complete control over cellular pathways[2].

Propagation of light

Optogenetics offers new approaches by which production can be modulated. One novel application could be lab grown organs or cultivated meat. In this application, light would enable regulation of differentiation with high spatiotemporal resolution by regulating the expression of differentiation factors or related genes, and the spatial control over cellular behavior is a unique advantage that may allow properties such as marbling in cultivated meat. To take advantage of

the spatial control that optogenetics offers, adherent cells on a surface or scaffold may be utilized. A limitation of the current work is that the propagation of light was evaluated assuming suspension cells, where tissues would have different optical properties that decrease light penetration[3]. Further work on the propagation of light through biological tissues is necessary. One potential method to overcome the limitation of light propagation through thicker tissues could be to combine optogenetics with 3D bioprinting, through which thinner layers of cells can be printed and light-activated in an alternating manner to activate cells in ultimately thicker tissues.

The use of optogenetics may circumvent complex media development that is often used to regulate growth and differentiation of cells. Future work involves developing cell lines that are equipped with gene expression systems for light-activated expression of differentiation factors, rather than eGFP.

Another avenue of exploration is how the intensity of light and length of exposure is related to phototoxicity, which is the limiting factor for large-scale optogenetics. One major barrier to scaling up optogenetics is cell stress and death, which can occur with higher intensity light[4]. Many parameters such as intensity, wavelength, and length of exposure to radiation influence cytotoxicity, but it is unclear how these parameters relate to phototoxicity. A plausible but wishful solution may be that the net energy absorbed by the media and the cell, which accounts for additional phototoxicity when fluorescent proteins are present, is well correlated with cell death. Perhaps a dimensionless variable may be developed to relate the two. A solution to overcoming phototoxicity challenges include using red light optogenetics, which has lower energy and causes less phototoxicity[5].

Spike Bioproduction

In the bioproduction of SARS-CoV-2 Spike protein, engineered full-length Spike proteins have exhibited orders of magnitudes of higher expression than wild-type Spike, including the 2P mutation and the more recent HexaPro Spike with additional proline substitutions[6]. The design of Spike truncations has enabled further titer increases, with T1 having higher expression than that of the engineered 2P Spike and the RBD.

T1 also exhibited higher binding to an anti-Spike polyclonal antibody than did full-length Spike. This behavior was unexpected, as polyclonal antibodies often target multiple binding epitopes on a protein, and therefore the full-length protein would be expected to have more binding epitopes and have higher avidity to the antibody. We hypothesize that T1 has an additional binding epitope that is sterically hindered when more residues are added. This may explain the decrease in avidity once residues are added for T4 and full-length Spike. Discovering new binding epitopes would be important for designing engineered proteins for antigen tests or protein-based vaccines, as well as for evaluating whether mutations in Spike may result in new strains of SARS-CoV-2 that have stronger binding to ACE2 or weaker binding to anitbodies.

Protein engineering extends beyond the manipulation of amino acids, and the consideration of other parameters such as post-translational modifications are critical. In SARS-CoV-2 Spike engineering efforts, many groups have focused on glycosylation, as Spike is a heavily glycosylated protein with 22 known N-linked glycosylation sites[7]. Spike glycosylation is critical for evading recognition by the host immune response through glycan shielding[8]. Glycosylation also influences the conformational dynamics of Spike, in which the RBD conformation is modulated by glycan interactions with the protein backbone[9]. The conformational dynamics of RBD affects the binding of Spike to ACE2, and it also affects the immune response, as the

exposed RBD has higher immunogenicity[10,11]. The RBD dynamics illustrate a complex interplay between a tradeoff in ability to bind ACE2 and evasion of the host immune system. O-linked glycosylation has also been found on Spike, and while O-linked glycosylation is less understood than N-linked glycosylation, O-linked glycosylation may also play a major role in Spike structure and binding dynamics[12,13].

The truncation and glycoproteomics analysis of Spike demonstrates the ability to produce higher performing proteins for expression and binding, but it is still unclear whether T1 is the optimal truncation. Molecular dynamics simulations of additional truncations may elucidate whether other truncations may be more stable and bind well to antibodies, but this would be a large body of work that is currently impractical. Advances in high-throughput experimental screening, machine learning methods, and computational power are all potential ways to overcome this hurdle. For instance, the higher throughput of structure prediction possible by deep learning methods such as AlphaFold2 and ESMFold may allow computational screening of truncations near T1.

Concluding Remarks

Protein engineering provides a rich toolset for investigating many biological fields of research, and macroscopic experimental approaches dovetail with molecular dynamics and deep learning computational approaches, leading to molecular-level understanding of macroscopic behavior. The results from our synthetic biology and Spike engineering works address some limitations, but also remind us that the journey to understanding protein folding, stability, and binding is far from complete.

Deep learning algorithms such as AlphaFold2, OpenFold, ESMFold, and RoseTTAFold are major milestones in protein structure prediction efforts and have been critical in accelerating protein

engineering efforts. In particular, RoseTTAFold2NA represents the next step forward with its ability to predict protein-nucleic acid complexes. Beyond increasing the accuracy of protein monomer and multimer structure predictions further, potential future directions include enhancing predictions of protein complexes with other types of molecules, such as nucleic acids, small molecules, and lipids.

However, several challenges remain to be solved. Predicting different conformational states of proteins is one difficulty, in which vanilla AlphaFold2 tends to have predict more conformational heterogeneity. Modifications to the multiple sequence alignment (MSA) may allow AlphaFold2 to explore greater conformational diversity, but relies on experimental data to validate the set of predicted structures[14,15]. Custom templates may also be used to guide structural prediction to distinct conformations, though a decrease in maximum MSA may be needed for the templates to strongly influence the predicted structure[16]. Another challenge is that proteins are dynamic structures, which is not represented in static predictions. Although the predicted local distance difference test (pLDDT) score can give insight into the local flexibility of a protein structure, it is different from a B-factor in that pLDDT scores do not explicitly measure structural flexibility, but rather, predict model error[17]. Decoupling of model error and protein flexibility would greatly increase the utility to which AlphaFold2 and other deep learning algorithms can predict conformational diversity.

Paired with advances in structural prediction algorithms, significant advances have also been continuously occurring in computational hardware. Particularly, graphical processing units (GPUs) have become extremely powerful for parallelization, and tensor cores are heavily utilized for deep learning applications. The higher throughput of modern algorithms is a result that is co-derived from software development and compatible hardware. Other non-deep learning software such as the GROMACS molecular dynamics package, are also incorporating more GPU

computation by enabling Compute Unified Device Architecture (CUDA) graphs. These advances allow longer simulations that can capture large conformational changes in proteins over time or can simulate larger systems.

Beyond the prediction of protein structuers, applications of engineered proteins for modulating biological systems is the next step. Increasing our ability to finely tune and design proteins would greatly facilitate efforts to make engineered systems more stable, precise, and efficacious.

References

1. Taslimi, A., Zoltowski, B., Miranda, J. G., Pathak, G., Hughes, R. M., & Tucker, C. L. (2016). Optimized second generation CRY2/CIB dimerizers and photoactivatable Cre recombinase. *Nature Chemical Biology*, *12*(6), 425–430. https://doi.org/10.1038/nchembio.2063

2. Cao, L., Coventry, B., Goreshnik, I., Huang, B., Sheffler, W., Park, J. S., Jude, K. M., Marković, I., Kadam, R. U., Verschueren, K. H. G., Verstraete, K., Walsh, S. T. R., Bennett, N., Phal, A., Yang, A., Kozodoy, L., DeWitt, M., Picton, L., Miller, L., … Baker, D. (2022). Design of protein-binding proteins from the target structure alone. *Nature*, *605*(7910), Article 7910. https://doi.org/10.1038/s41586-022-04654-9

3. Klose, A. D., & Larsen, E. W. (2006). Light transport in biological tissue based on the simplified spherical harmonics equations. *Journal of Computational Physics*, *220*(1), 441–470. https://doi.org/10.1016/j.jcp.2006.07.007

4. Duke, C. G., Savell, K. E., Tuscher, J. J., Phillips, R. A., & Day, J. J. (2020). Blue Light-Induced Gene Expression Alterations in Cultured Neurons Are the Result of Phototoxic Interactions with Neuronal Culture Media. *ENeuro*, *7*(1), ENEURO.0386-19.2019. https://doi.org/10.1523/ENEURO.0386-19.2019

5. Lehtinen, K., Nokia, M. S., & Takala, H. (2022). Red Light Optogenetics in Neuroscience. *Frontiers in Cellular Neuroscience*, *15*. https://www.frontiersin.org/articles/10.3389/fncel.2021.778900

6. Hsieh, C.-L., Goldsmith, J. A., Schaub, J. M., DiVenere, A. M., Kuo, H.-C., Javanmardi, K., Le, K. C., Wrapp, D., Lee, A. G., Liu, Y., Chou, C.-W., Byrne, P. O., Hjorth, C. K., Johnson, N. V., Ludes-Meyers, J., Nguyen, A. W., Park, J., Wang, N., Amengor, D., … McLellan, J. S. (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science*, *369*(6510), 1501–1505. https://doi.org/10.1126/science.abd0826

7. Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S., & Crispin, M. (2020). Site-specific glycan analysis of the SARS-CoV-2 spike. *Science*, *369*(6501), 330–333. https://doi.org/10.1126/science.abb9983

8. Grant, O. C., Montgomery, D., Ito, K., & Woods, R. J. (2020). Analysis of the SARS-CoV-2 spike protein glycan shield reveals implications for immune recognition. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-71748-7

9. Sztain, T., Ahn, S.-H., Bogetti, A. T., Casalino, L., Goldsmith, J. A., Seitz, E., McCool, R. S., Kearns, F. L., Acosta-Reyes, F., Maji, S., Mashayekhi, G., McCammon, J. A., Ourmazd, A., Frank, J., McLellan, J. S., Chong, L. T., & Amaro, R. E. (2021). A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nature Chemistry*, *13*(10), Article 10. https://doi.org/10.1038/s41557-021-00758-3

10. Lv, Z., Deng, Y.-Q., Ye, Q., Cao, L., Sun, C.-Y., Fan, C., Huang, W., Sun, S., Sun, Y., Zhu, L., Chen, Q., Wang, N., Nie, J., Cui, Z., Zhu, D., Shaw, N., Li, X.-F., Li, Q., Xie, L., … Wang, X. (2020). Structural basis for neutralization of SARS-CoV-2 and SARS-CoV by a potent therapeutic antibody. *Science (New York, N.y.)*, eabc5881. https://doi.org/10.1126/science.abc5881

11. Liu, H., Wu, N. C., Yuan, M., Bangaru, S., Torres, J. L., Caniels, T. G., van Schooten, J., Zhu, X., Lee, C.-C. D., Brouwer, P. J. M., van Gils, M. J., Sanders, R. W., Ward, A. B., & Wilson, I. A. (2020). Cross-Neutralization of a SARS-CoV-2 Antibody to a Functionally Conserved Site Is

Mediated by Avidity. *Immunity*, *53*(6), 1272-1280.e5. https://doi.org/10.1016/j.immuni.2020.10.023

12. Tian, W., Li, D., Zhang, N., Bai, G., Yuan, K., Xiao, H., Gao, F., Chen, Y., Wong, C. C. L., & Gao, G. F. (2021). O-glycosylation pattern of the SARS-CoV-2 spike protein reveals an "O-Follow-N" rule. *Cell Research*, *31*(10), Article 10. https://doi.org/10.1038/s41422-021-00545-2

13. Gong, Y., Qin, S., Dai, L., & Tian, Z. (2021). The glycosylation in SARS-CoV-2 and its receptor ACE2. *Signal Transduction and Targeted Therapy*, *6*(1), Article 1. https://doi.org/10.1038/s41392-021-00809-8

14. del Alamo, D., Sala, D., Mchaourab, H. S., & Meiler, J. (2022). Sampling alternative conformational states of transporters and receptors with AlphaFold2. *ELife*, *11*, e75751. https://doi.org/10.7554/eLife.75751

15. Stein, R. A., & Mchaourab, H. S. (2022). SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLOS Computational Biology*, *18*(8), e1010483. https://doi.org/10.1371/journal.pcbi.1010483

16. Mikhaylov, V., & Levine, A. J. (2023). *Accurate modeling of peptide-MHC structures with AlphaFold* (p. 2023.03.06.531396). bioRxiv. https://doi.org/10.1101/2023.03.06.531396

17. Guo, H.-B., Perminov, A., Bekele, S., Kedziora, G., Farajollahi, S., Varaljay, V., Hinkle, K., Molinero, V., Meister, K., Hung, C., Dennis, P., Kelley-Loughnane, N., & Berry, R. (2022). AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-14382-9

Appendix A: *Supporting Information for Transient light-activated gene expression in Chinese hamster ovary cells*
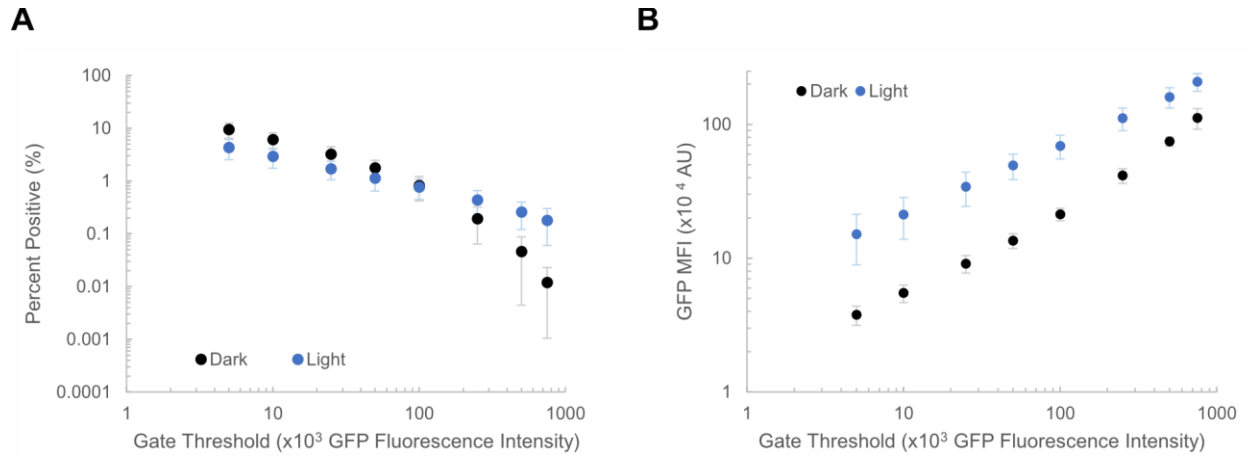

Figure A-1. Percent positive cells and MFI plotted against gate threshold. (**A**) Number of cells crossing the eGFP-positive gate with increasing gate threshold. At high gate thresholds, the number of eGFP-positive cells are higher for the activated LACE system. At lower gate thresholds, the number of eGFP-positive cells are higher for the nonactivated LACE system. (**B**) GFP MFI of LACE cells plotted against gate thresholds. eGFP-positive cells are higher for the LACE system in the light for all gate thresholds.
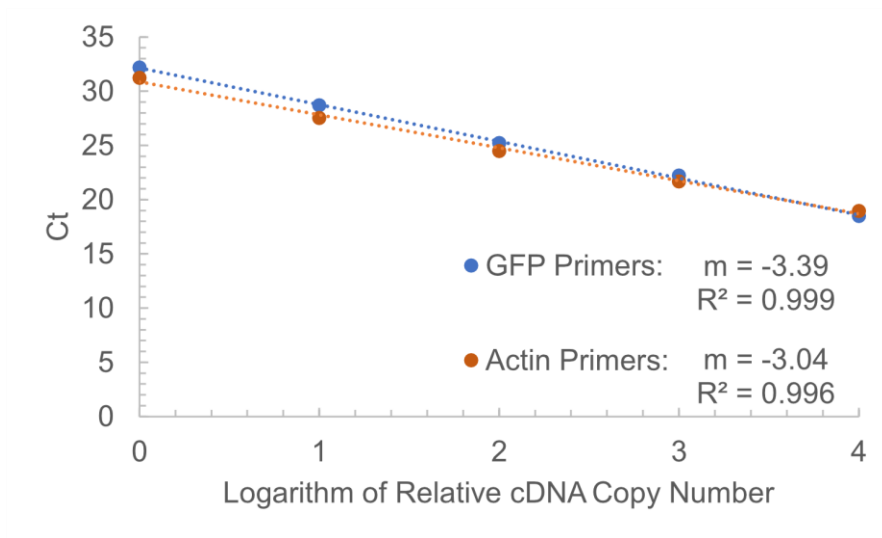
Figure A-2. qRT-PCR primer efficiencies. Primer efficiencies for eGFP and Actin. eGFP and Actin primer efficiencies calculated from the slopes (m) of the log(cDNA) plots are 97.2% and 113%, respectively. Data are averaged from duplicate dilutions.

A   Transient Plasmids

B   Lentiviral Plasmids
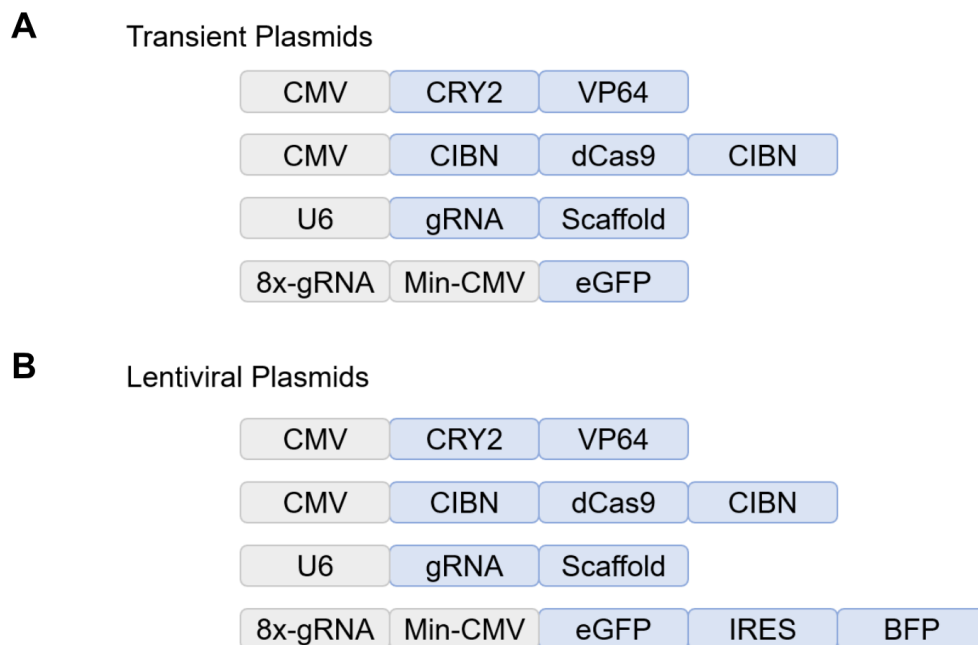
Figure A-3. Transient and lentiviral expression constructs. (**A**) Transient plasmid expression constructs. CRY2 and CIBN are expressed by CMV promoters, gRNA is expressed by a U6

promoter, and GFP is expressed by a minimal CMV promoter, whose expression is activated by VP64. (**B**) Lentiviral expression constructs. All constructs except the minimal CMV construct are identical to the transient plasmids (**A**), except in a lentiviral backbone. The minimal CMV construct was modified to contain an additional IRES sequence and BFP to identify light-activated cells once the GFP is removed.

Appendix B: *Supporting information for Computational evaluation of light propagation in cylindrical bioreactors for optogenetic mammalian cell cultures*

Calculation of mass absorption and scattering coefficients

With an absorption coefficient of 0.04 cm^{-1} at 60 million cells suspended in 3 mL phosphate buffered saline (PBS) at 450 nm is approximately 0.04 cm^{-1}. Given the high optical transparency of PBS, the mass absorption coefficient α_a of mammalian cells is estimated from the absorption coefficient κ and the cell number density ρ_c:

$$\alpha_a = \frac{\kappa}{\rho_c}.$$

Substituting values, we find

$$\alpha_a = \left(\frac{0.04}{cm}\right)\left(\frac{3cm^3}{(6 \cdot 10^7 cells)}\right)\left(\frac{m^2}{100^2 cm^2}\right) = 2 \cdot 10^{-13} \frac{m^2}{cell}.$$

For scattering, we use a reduced scattering coefficient of 2 cm^{-1} and a scattering anisotropy g of 0.98. The scattering coefficient was calculated as:

$$\sigma_s = \frac{\sigma_s'}{1 - g}.$$

The scattering coefficient is calculated to be:

$$\sigma_s = \left(\frac{2cm^{-1}}{1 - 0.98}\right)\left(\frac{100cm}{m}\right) = 10^4 m^{-1}.$$

Subsequently the mass scattering coefficient α_s is calculated as:

$$\alpha_s = \frac{\sigma_s}{\rho_c}$$

Substituting:

$$\alpha_s = \frac{10^4 m^{-1}}{10^8 \frac{cell}{mL}}\left(\frac{m^3}{10^6 mL}\right) = 10^{-10}\frac{m^2}{cell}.$$

Height-to-diameter ratio of a cylinder

For a cylinder with a height-to-diameter (HD) ratio of n, the surface area to volume ratio can be expressed as:

$$\frac{2\pi r^2 + 4\pi n r^2}{2\pi n r^3}$$

where n is the HD ratio.

This can be simplified to:

$$\frac{1 + 2n}{nr}.$$

Appendix C: *Supporting Information for* Production of novel SARS-CoV-2 Spike truncations in Chinese hamster ovary cells leads to high expression and binding to antibodies

Table A-1. Protein sequences of CHO-expressed Spike truncations. Signal sequences are highlighted in gray, and potential N-linked glycosylation sites are highlighted in green. Percentages indicate percent coverage of full-length Spike construct, including signal sequences and 6x His tags.

Truncation	Amino acid sequence, including secretion signal and 6x His tag
T1 (23.5%)	MFVFLVLLPLVSSQNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWN RKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDE VRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFR KSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY RVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFL PFQQFGRDIADTTDAVRDPQTLEILDITPCSHHHHHH
T2 (25.1%)	MFVFLVLLPLVSSQTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPL SETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFAS VYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF VIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYL YRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGV GYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTE HHHHHH
T3 (28.1%)	MFVFLVLLPLVSSQTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPL SETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFAS

	VYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
	VIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYL
	YRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGV
	GYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTE
	SNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSHHHHHH
T4 (27.5%)	MFVFLVLLPLVSSQNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWN
	RKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDE
	VRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFR
	KSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY
	RVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFL
	PFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQ
	DVNCTEVPVAIHADQLTPTWRVYSTGSNVHHHHHH
T5 (28.3%)	MFVFLVLLPLVSSQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSG
	WTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVE
	KGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNC
	VADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPG
	QTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPF
	ERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSF
	ELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTEHHHHHH
T6 (35.4%)	MFVFLVLLPLVSSQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSG
	WTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVE
	KGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNC
	VADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPG
	QTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPF
	ERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSF

|	
	ELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG RDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTE VPVAIHADQLTPTWRVYSTGSNVHHHHHH
T7 (35.5%)	MFVFLVLLPLVSSQNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWN RKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDE VRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFR KSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY RVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFL PFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQ DVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIP IGAGICASYQTQTNSPASVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISV TTEILPVSMTKTSVDCTMYICGDHHHHHH
T8 (36.8%)	MFVFLVLLPLVSSQKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKS WMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFK IYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSS GWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTV EKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISN CVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAP GQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKP FERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLS FELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTEHHHHHH
S1 (54.7%)	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQ DLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWI FGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEF RVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTP

	INLVRDLPQGFSALEPLVDLPIGIN ITRFQTLLALHRSYLTPGDSSSGWTAGA AAYYVGYLQPRTFLLKYNE NGTITDAVDCALDPLSETKCTLKSFTVEKGIYQT SNFRVQPTESIVRFP NITNLCPFGEVF NATRFASVYAWNRKRISNCVADYSV LYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIA DYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTE IYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPA TVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTT DAVRDPQTLEILDITPCSFGGVSVITPGT NTSNQVAVLYQDV NCTEVPVAIHA DQLTPTWRVYSTGSNVFQTRAGCLIGAEHV NNSYECDIPIGAGICASYQTQT NSPAHHHHHH

114
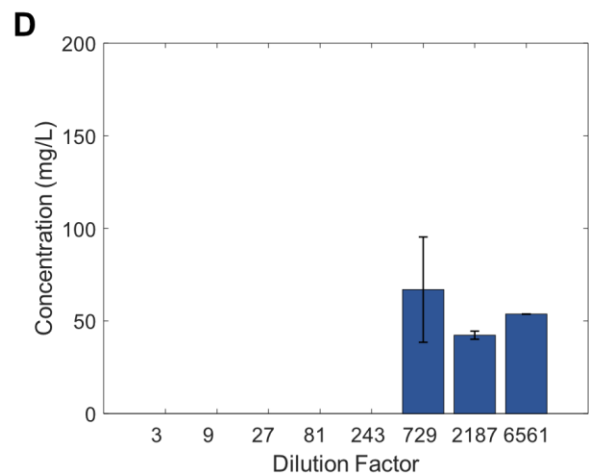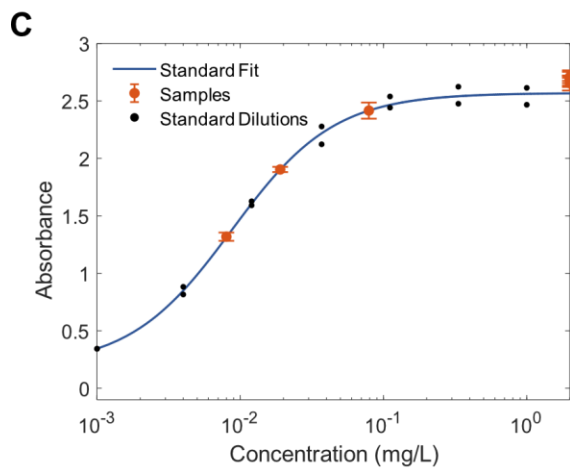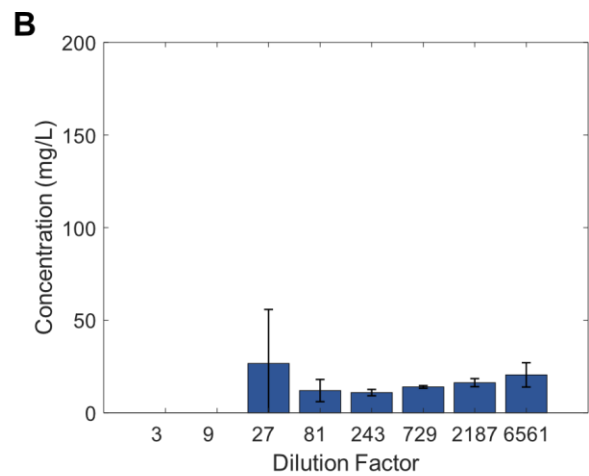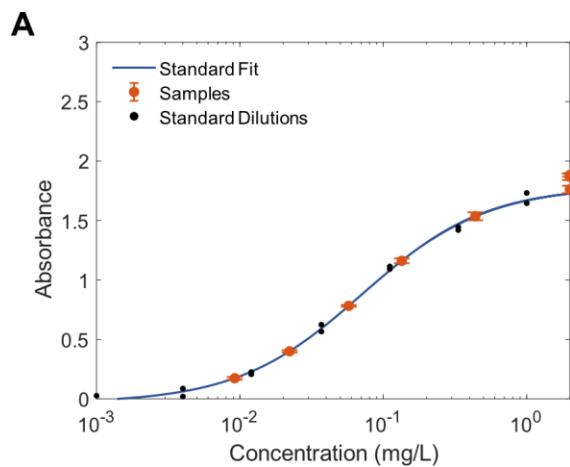
Figure A-1. Measurement of crude titers for Spike and RBD via sandwich ELISA. (**A**) Dilutions of crude Spike plotted against the Spike standard curve. (**B**) Back-calculated concentrations of Spike crude titers. (**C**) Dilutions of crude RBD plotted against the RBD standard curve. (**D**) Back-calculated concentrations of RBD crude titers. Error bars represent ± SD of technical triplicates.
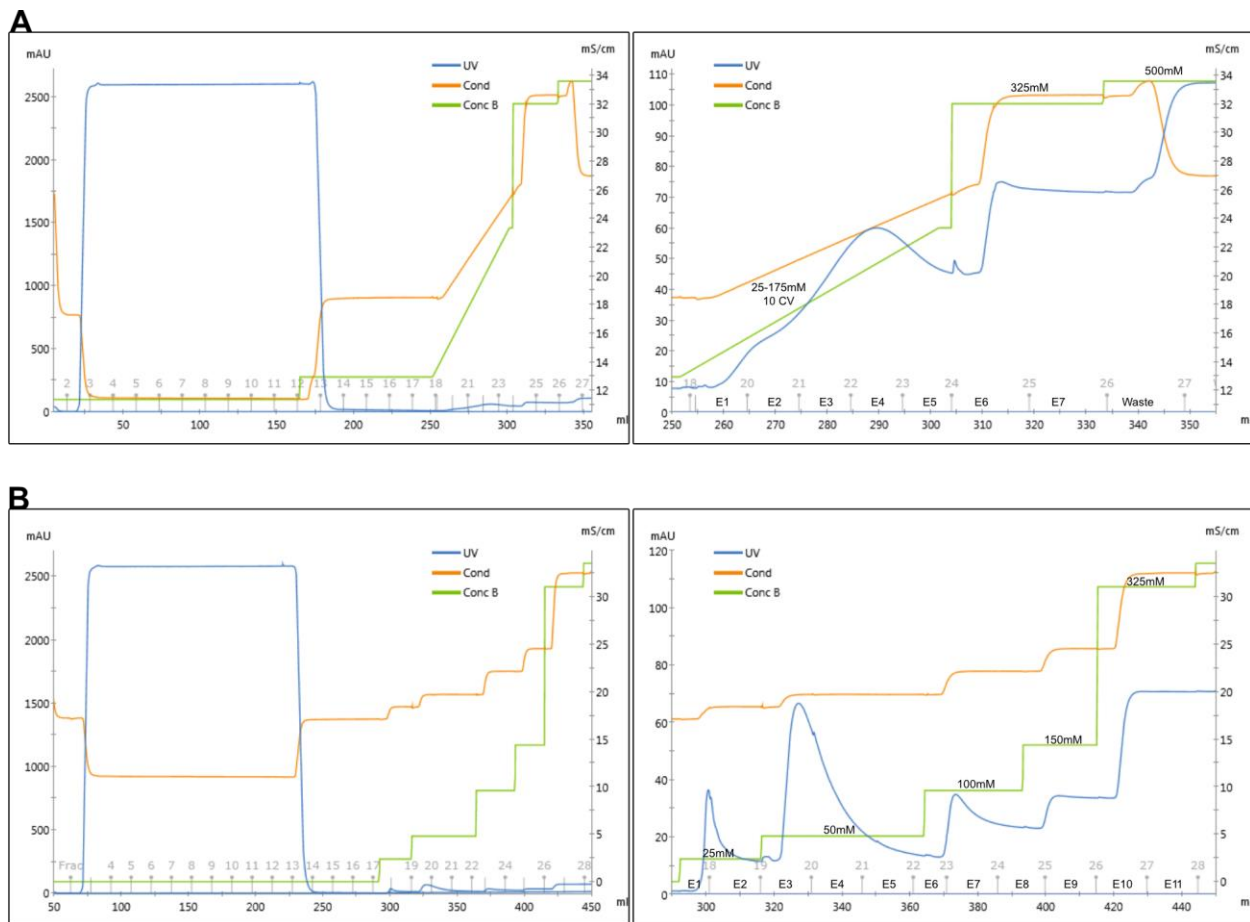
Figure A-2. Spike and RBD chromatograms plotted over volume flowed through the system. (**A**) Chromatogram of entire Spike purification (left) and of zoomed in elution fractions (right). 150 mL of sample was loaded and washed with PBS containing 25 mM imidazole. A continuous gradient was applied from 25 mM-175 mM imidazole over 10 CV to elute Spike. Fractions E3 and E4 were collected. (**B**) Chromatogram of entire RBD purification (left) and of zoomed in elution fractions (right). 150 mL of sample was loaded and washed for 10 CV with PBS. A step gradient was applied for elution with 6 CV steps at 25 mM, 50 mM, 100 mM, 150 mM, and 325 mM imidazole. Fractions E3 and E4 were collected. Y-axis on the left is UV absorbance and Y-axis on the right is conductivity. A 5 mL Ni-NTA column was used with a flow rate of 5 mL/min for all steps.
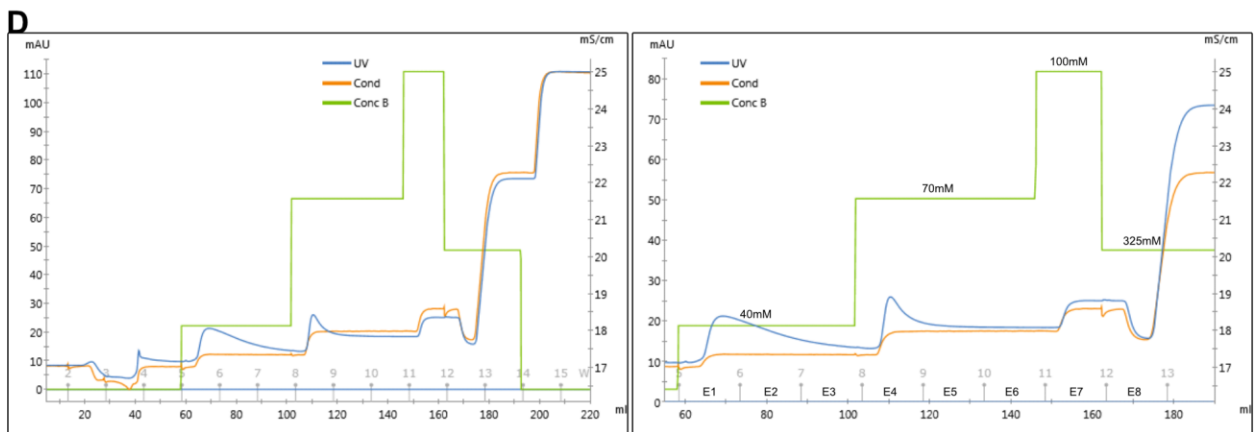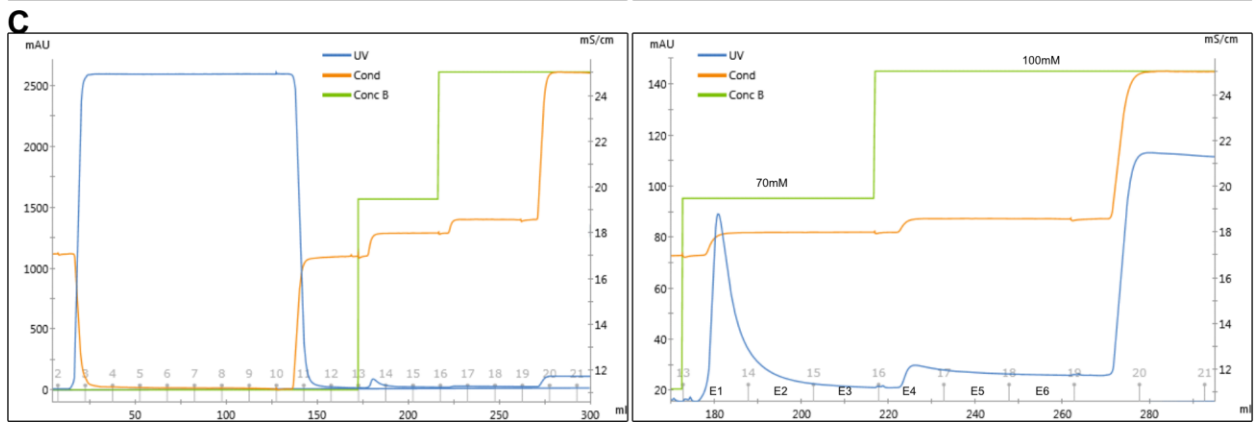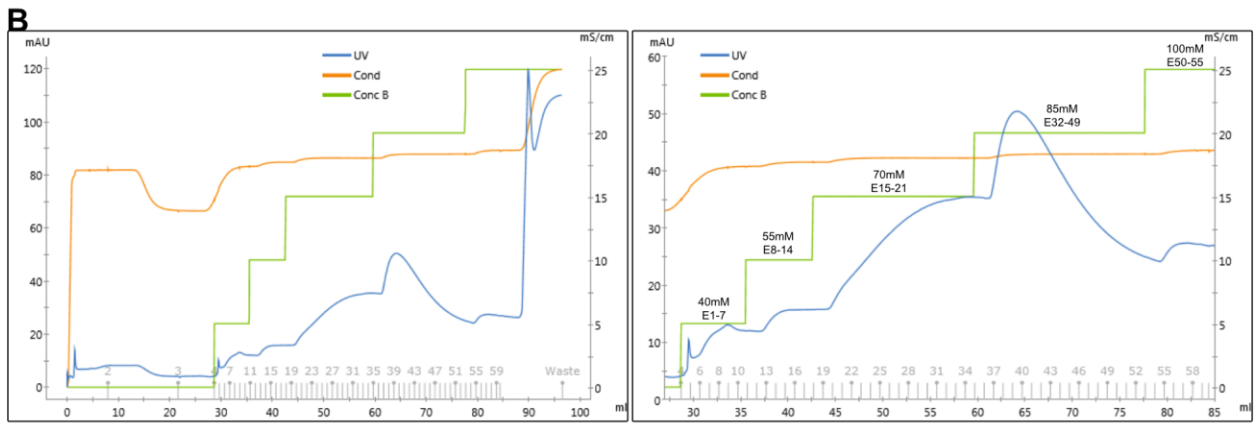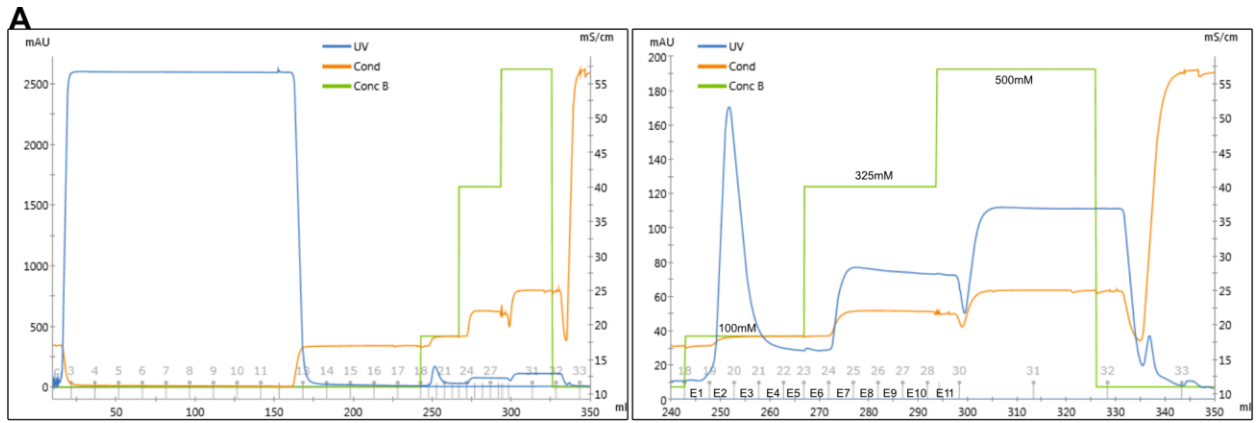
116

Figure A-3. T1 and T4 purification chromatograms. (**A**) Chromatogram of entire T1 purification (left) and of zoomed in elution fractions (right). First, 150 mL of sample was loaded and washed with PBS containing 25 mM imidazole. Step gradients were applied at 100 mM and 325 mM imidazole to elute T1. (**B**) Fractions E2 and E3 from the first purification were combined, dialyzed and re-purified using a step gradient with steps at 40 mM, 55 mM, 70 mM, 85 mM, and 100 mM imidazole. Fractions eluting from 40-85 mM imidazole were combined. (**C**) Chromatogram of entire T4 purification (left) and of zoomed in elution fractions (right). First, 120 mL of sample was loaded, and the resin was washed with PBS containing 25 mM imidazole. A step gradient at 70 mM and 100 mM imidazole was used to elute T4. (**D**) T4 from the first purification was dialyzed and re-purified using step gradients at 40 mM, 70 mM, 100 mM, 325 mM, and 500 mM imidazole. 40 mM imidazole fractions and the first 70 mM imidazole fraction were combined and used for further characterization. A 5 mL Ni-NTA column was used with a flow rate of 5 mL/min for all steps.

A

```
Ref_T1_Seq    NFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFST
T1_Coverage   -FRVQPTESIVR-----------------FASVYAWNRKRISNCVADYSVLYNSASFST
              **********                   ****************************

Ref_T1_Seq    FKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAW
T1_Coverage   FKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAW
              ************************************************************

Ref_T1_Seq    NSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYG
T1_Coverage   NSNNLDSKVGGNYNYLYR---KSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYG
              *****************   *************************************

Ref_T1_Seq    FQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESN
T1_Coverage   FQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESN
              ************************************************************

Ref_T1_Seq    KKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSHHHHHH
T1_Coverage   KKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSHHHHHH
              *****************************************
```

B

```
Ref_T4_Seq       NFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFST
T4_Top_Coverage  --RVQPTESIVR-----------------FASVYAWNRKRISNCVADYSVLYNSASFST
                 **********                   ****************************

Ref_T4_Seq       FKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAW
T4_Top_Coverage  FKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAW
                 ************************************************************

Ref_T4_Seq       NSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYG
T4_Top_Coverage  NSNNLDSKVGGNYNYLYR---KSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYG
                 *****************   *************************************

Ref_T4_Seq       FQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESN
T4_Top_Coverage  FQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESN
                 ************************************************************

Ref_T4_Seq       KKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVN
T4_Top_Coverage  KKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCS------------------------
                 **********************************

Ref_T4_Seq       CTEVPVAIHADQLTPTWRVYSTGSNVHHHHHH
T4_Top_Coverage  -------------------------HHHHHH
                                          *****
```

C

```
Ref_T4_Seq       NFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFST
T4_Bottom_Cover  -FRVQPTESIVR-----------------FASVYAWNRKRISNCVADYSVLYNSASFST
                 **********                   ****************************

Ref_T4_Seq       FKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAW
T4_Bottom_Cover  FKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAW
                 ************************************************************

Ref_T4_Seq       NSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYG
T4_Bottom_Cover  NSNNLDSKVGGNYNYLYR---KSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYG
                 *****************   *************************************

Ref_T4_Seq       FQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESN
T4_Bottom_Cover  FQPTNGVGYQPYRVVVLSFELLHAPATVCGPKK--------CVNFNFNGLTGTGVLTESN
                 ********************************         ******************

Ref_T4_Seq       KKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVN
T4_Bottom_Cover  KKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCS------------------------
                 **********************************

Ref_T4_Seq       CTEVPVAIHADQLTPTWRVYSTGSNVHHHHHH
T4_Bottom_Cover  -------------------------HHHHHH
                                          *****
```

Figure A-4. Shotgun proteomics on T1 and T4. Coverage of (**A**) T1, (**B**) T4 top band, and (**C**) T4 bottom band against full sequences. Tandem mass spectra were extracted by MS Convert (ProteoWizard). Charge state deconvolution and deisotoping were not performed. All MS/MS samples were analyzed using X! Tandem (The GPM, thegpm.org; version X! Tandem Alanine (2017.2.1.4)). X! Tandem was set up to search the Uniprot human database and known T1 and T4 sequences assuming the digestion enzyme trypsin. X! Tandem was searched with a fragment ion mass tolerance of 20 PPM and a parent ion tolerance of 20 PPM. Carbamidomethyl of cysteine and selenocysteine was specified in X! Tandem as a fixed modification. Glu->pyro-Glu of the n-terminus, ammonia-loss of the n-terminus, gln->pyro-Glu of the n-terminus, deamidated of asparagine and glutamine, oxidation of methionine and tryptophan and dioxidation of methionine and tryptophan were specified in X! Tandem as variable modifications. Scaffold (version Scaffold_4.9.0, Proteome Software Inc., Portland, OR) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 98.0% probability by the Scaffold Local FDR algorithm. Peptide identifications were also required to exceed specific database search engine thresholds. Protein identifications were accepted if they could be established at greater than 5.0% probability to achieve an FDR less than 5.0% and contained at least 2identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm (Nesvizhskii et al., 2003) Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. Proteins sharing significant peptide evidence were grouped into clusters.
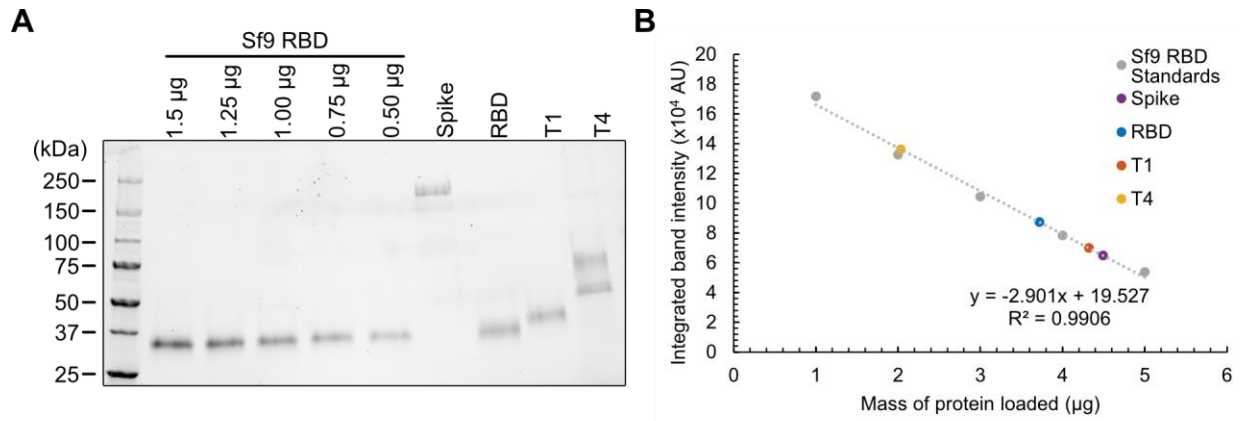
Figure A-5. Quantification of purified proteins. (**A**) SDS-PAGE and (**B**) quantification for purified proteins and serial dilutions of Sf9 RBD obtained from BEI Resources. A standard curve was prepared using a linear fit to serial dilutions of the standard protein.
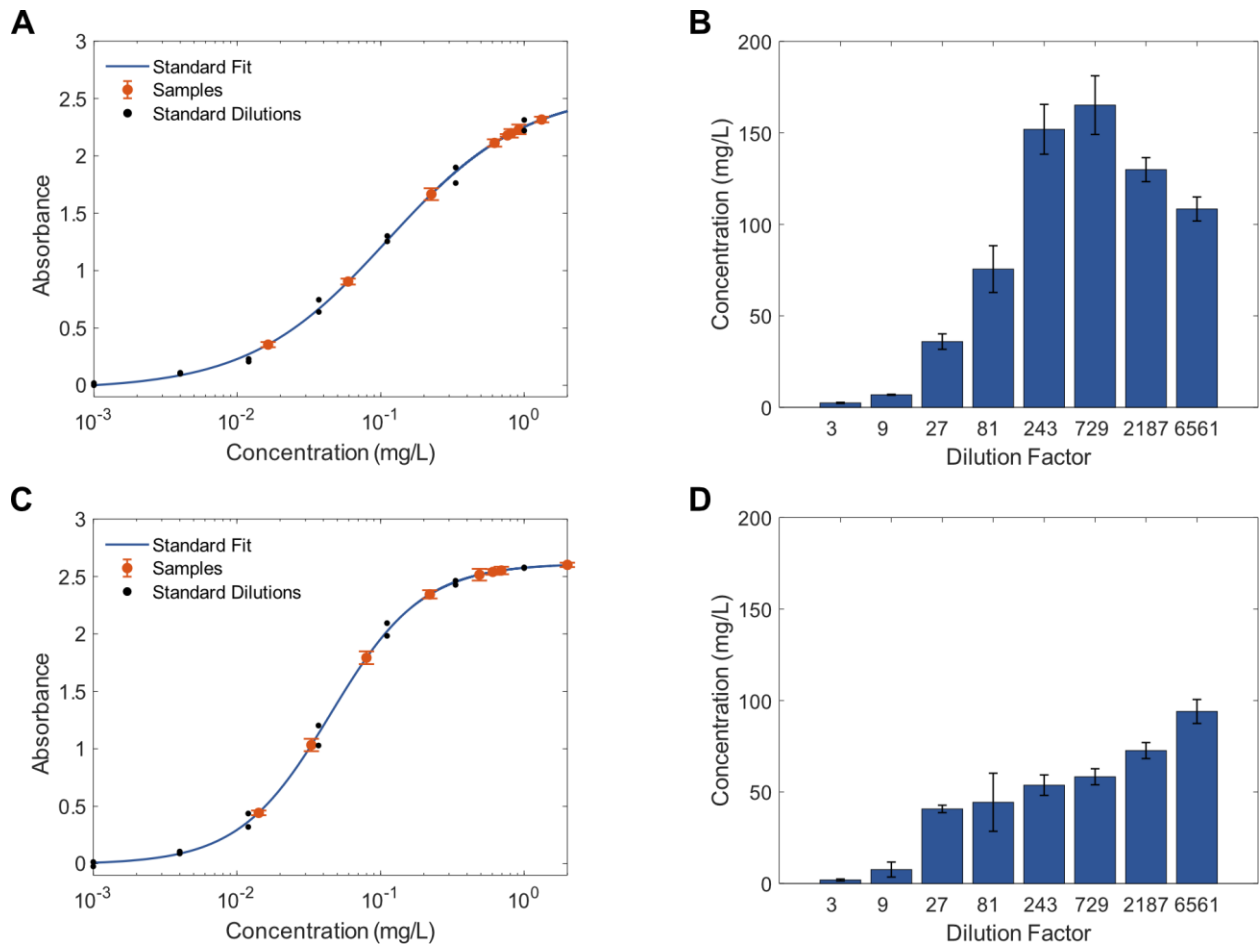
Figure A-6. Measurement of crude titers for T1 and T4 via sandwich ELISA. (**A**) Dilutions of crude T1 plotted against the T1 standard curve. (**B**) Back-calculated concentrations of T1 crude titers. (**C**) Dilutions of crude T4 plotted against the T4 standard curve. (**D**) Back-calculated concentrations of T4 crude titers. Error bars represent ± SD of technical triplicates.
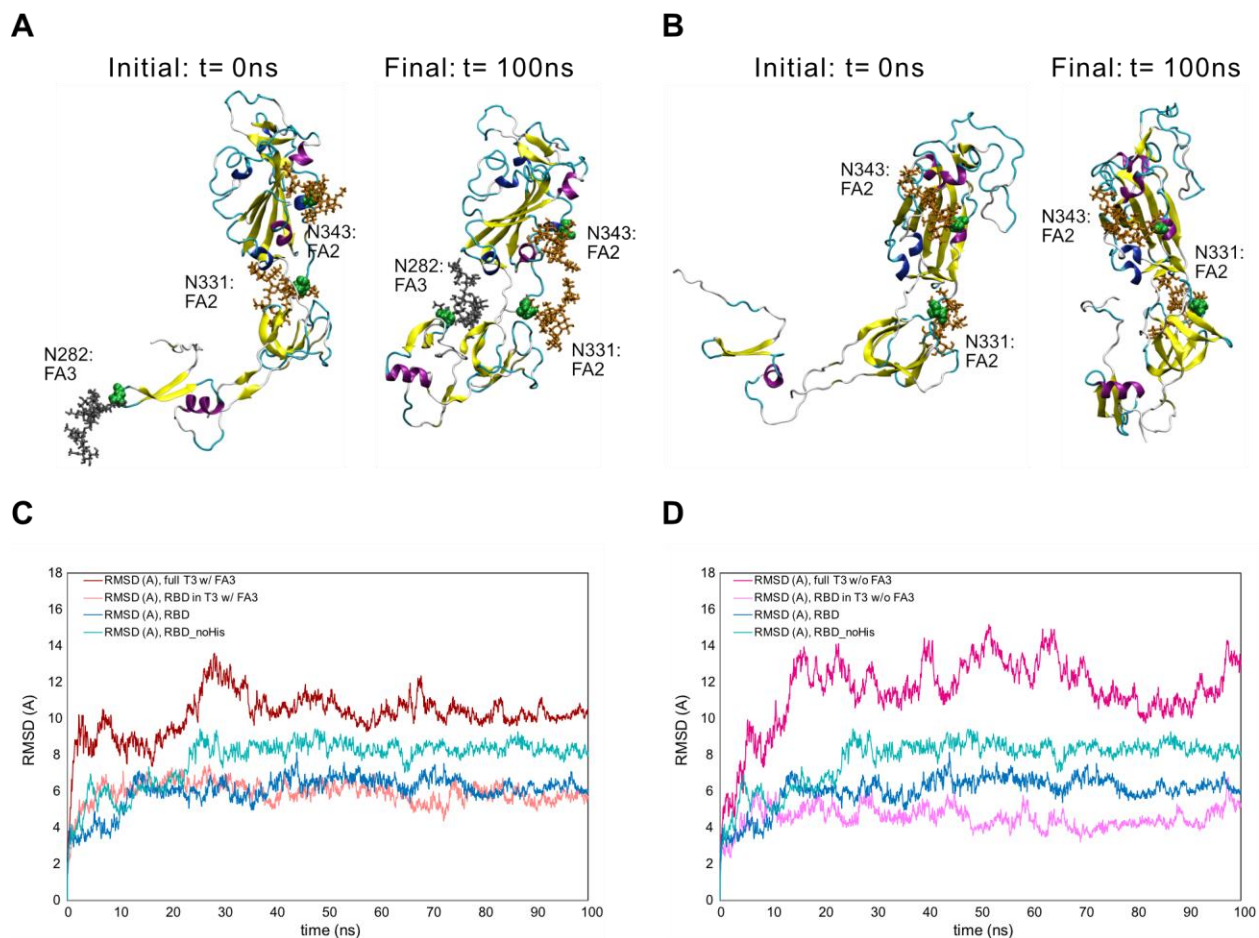
Figure A-7. Molecular dynamics structural stability snapshots and analysis of T3. MD snapshots are visualized for (**A**) T3 with FA3 glycan and (**B**) T3 without FA3 glycan. (**C**) Backbone RMSD profiles of full T3 and T3 RBD subdomain with FA3 glycan. (D) Backbone RMSD profiles of full T3 and T3 RBD subdomain without FA3 glycan. Profiles of (C) and (D) include RBD with and without the 6x His tag referenced to initial configurations for comparison. Green spheres represent the glycosylated asparagine residue, brown sticks represent FA2 glycans, and gray sticks represent FA3 glycans.
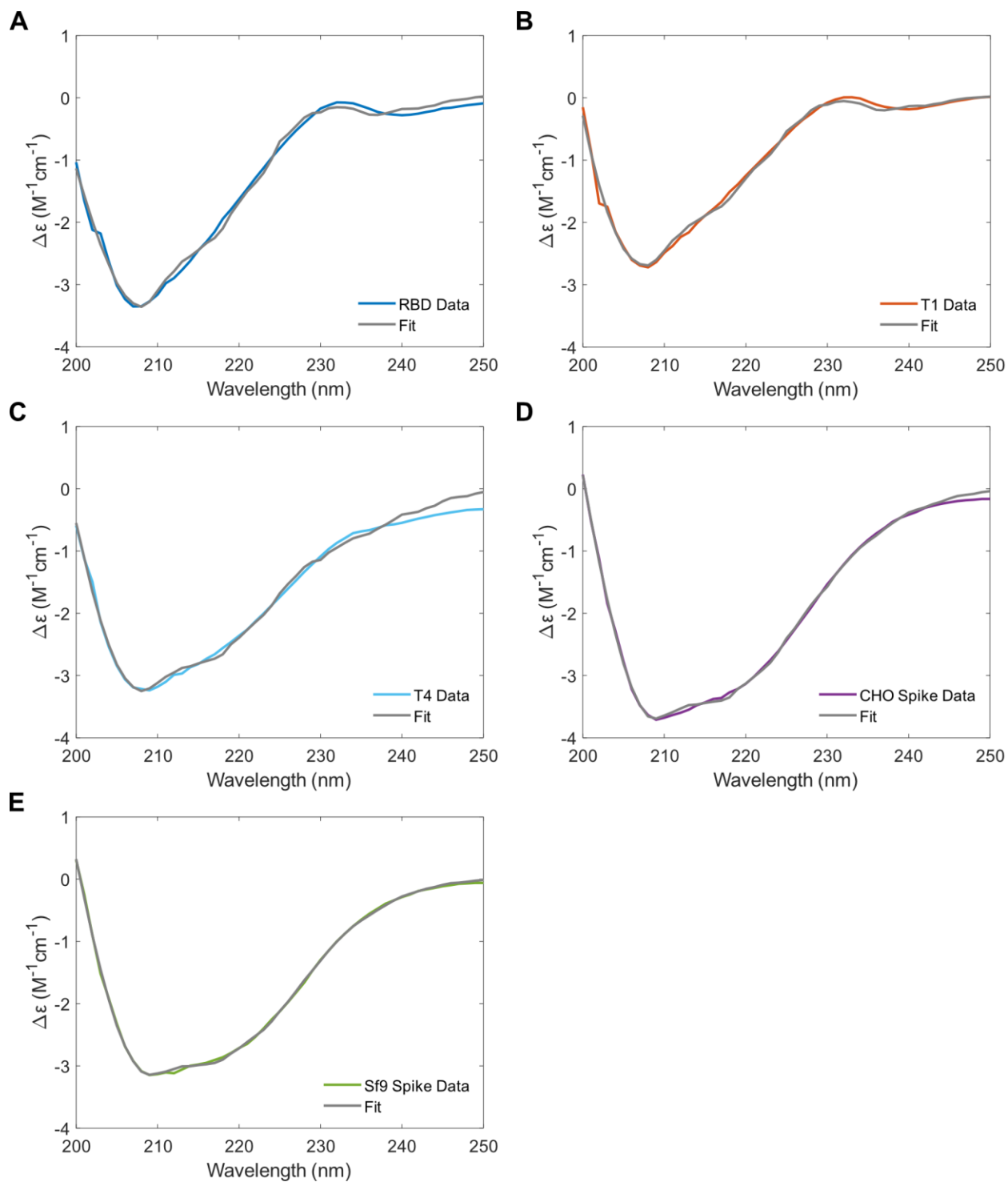
Figure A-8. Raw spectral data on proteins analyzed via circular dichroism. Δε is plotted against wavelength for (**A**) RBD, (**B**) T1, (**C**) T4, (**D**) CHO Spike, and (**E**) Sf9 Spike.
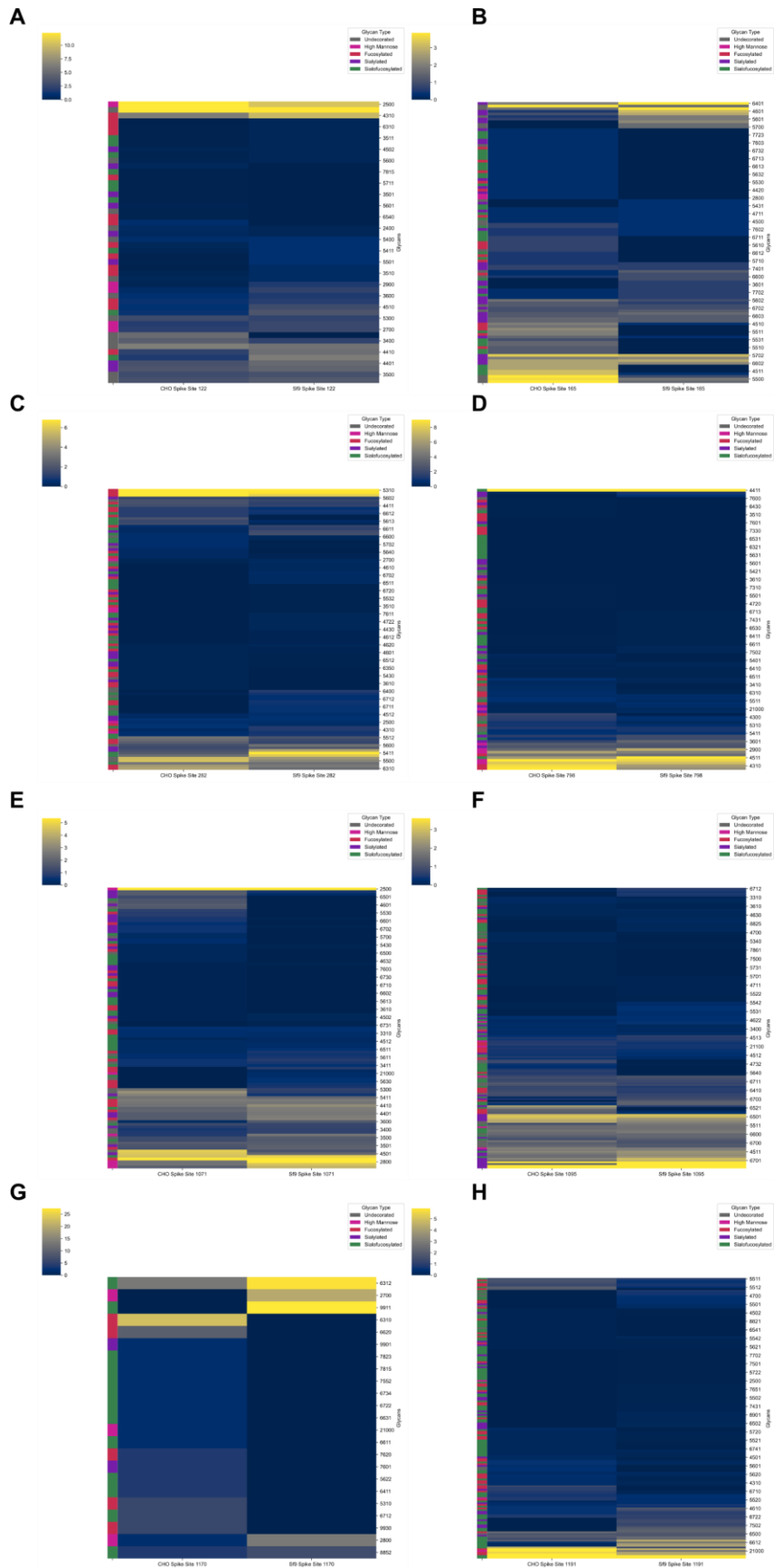
Figure A-9. Heatmaps comparing CHO and Sf9 Spike for sites (**A**) 122, (**B**) 165, (**C**) 282, (**D**) 798, (**E**) 1071, (**F**) 1095, (**G**) 1170, and (**H**) 1191. Sites 331 and 343 are omitted here and are shown in the main text.

References

Nesvizhskii, A. I., Keller, A., Kolker, E., & Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, *75*(17), 4646–4658. https://doi.org/10.1021/ac0341261