ABSTRACT

Title of Dissertation: APPLICATION OF ADVANCED MACHINE LEARNING STRATEGIES FOR BIOMEDICAL RESEARCH

Renee Ti Chou
Doctor of Philosophy, 2023

Dissertation Directed by: Professor Michael P. Cummings
Department of Biology

Biomedical research delves deeply into understanding individual health and disease mechanisms. Recent advancements in technologies have further transformed the field with large-scale data sets, enabling data-driven approaches to identify important patterns and relationships from large data sets. However, these data sets are often noisy and unstructured. Moreover, missing values and high dimensionality further complicate the analysis processes aimed at yielding meaningful results. With examples in ocular diseases and malaria, this dissertation presents novel strategies employing machine learning to tackle some of the challenges in biomedical research.

In ocular diseases, sustained ocular drug delivery is critical to retain therapeutic levels and improve patient adherence to dosing schedules. To enhance the sustained delivery system, we engineer peptide sequences as an adapter to impart desired properties to ocular drugs. Specifically, we develop machine learning models separately for three properties–

melanin binding, cell-penetration, and non-toxicity. We employ data reduction techniques to reduce the number of features while maintaining the machine learning model performance and apply interpretable machine learning techniques to explain model predictions on the three properties. Experimental validation in rabbits show two-fold increase in drug retention time with the selected peptide candidate. The developed machine learning framework can be further tailored to engineer other properties in molecular sequences with a wide variety of potential in biomedical applications.

Malaria is an infectious disease caused by protozoan of the genus *Plasmodium* and has been a burden in global health. Developing malaria vaccines is challenging due to the diversity in parasite antigen sequences, which may lead to immune escape. To facilitate the vaccine development process, we leverage the wealth of systems data collected from various sources. For facile data management, a database is constructed to store the structured data processed from the results of the bioinformatics tools. Due to the small fraction of *Plasmodium* proteins labeled as known antigens, and the remaining proteins unknown of being antigens or non-antigens, a positive-unlabeled machine learning method is applied to identify potential vaccine antigen candidates. Beyond malaria, our approach provides a promising framework for identifying and prioritizing vaccine antigen candidates for a broad range of disease pathogens.

APPLICATION OF ADVANCED MACHINE LEARNING STRATEGIES FOR BIOMEDICAL RESEARCH

by

Renee Ti Chou

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:

Dr. Michael P. Cummings, Chair/Advisor
Dr. Najib El-Sayed, Dean's Representative
Dr. Laura M. Ensign
Dr. Philip Johnson
Dr. Brian Pierce
Dr. Shannon Takala Harrison

Acknowledgments

This endeavor would not have been possible without the individuals who played pivotal roles in shaping my Ph.D. journey.

First and foremost, I would like to express my deepest gratitude to my advisor and committee chair, Dr. Michael P. Cummings. I joined Dr. Cummings' lab at the Center for Bioinformatics and Computational Biology with a passion for learning interdisciplinary research skills and communicating with researchers from diverse scientific backgrounds, focusing on biomedical research. Dr. Cummings possesses extensive research experience in applying computational and mathematical methods to biological problems, which aligns with my interest in developing a machine learning platform for studying biomedical sciences. Dr. Cummings is also exceptionally supportive in both research and professional development. He is the best mentor, deeply caring for his students, and has taught me not to be afraid of any challenges. Whether it is about research or fellowship applications, he is always there to guide and support me.

I am extremely grateful to my two collaborators of my Ph.D. research projects, Dr. Shannon Takala-Harrison and Dr. Laura M. Ensign. Since joining Dr. Cummings' lab in June 2019, I have been working on malaria research in collaboration with Dr. Shannon Takala-Harrison's group at the Center for Vaccine Development and Global Health, University of Maryland School of Medicine. I would like to thank Dr. Takala-Harrison for

her invaluable insights into malaria, from the initial stages of the research to the process of manuscript writing. In each meeting with Dr. Takala-Harrison, I consistently learn from her unique perspectives, which enable me to improve the research further based on her helpful suggestions.

In September 2019, I began working with Dr. Laura M. Ensign's lab at the Center for Nanomedicine, Wilmer Eye Institute, which is part of the Johns Hopkins University School of Medicine. Dr. Ensign's expertise in drug delivery provides valuable insights and supplements the interdisciplinary research. She is always very supportive of my professional development and loves to share her personal stories of challenges she has encountered during her academic career. These stories have been a great source of encouragement for me when facing failures along the way. I am deeply grateful to have Dr. Ensign as my Ph.D. project collaborator.

I also want to extend my sincere thanks to my other dissertation committee members, Dr. Najib El-Sayed, Dr. Philip Johnson, and Dr. Brian Pierce, for their support over the past four years, including my initial committee meeting, qualifying examination, as well as seminars in Computational Biology, Bioinformatics, and Genomics, and the Center for Bioinformatics and Computational Biology seminars. They have been instrumental in helping my research in various ways, from teaching and commenting to supporting my fellowship applications. I cannot thank them enough for their informative and insightful advice on my dissertation.

Special thanks go to my colleagues in Dr. Cummings' lab, Alexis S. Boleda, Rana Khalil, and Yi Chen, for their helpful feedback and comments during lab meetings. I am also thankful to Jason Fan for his suggestions at the early stages of the malaria research. I would

also like to express my gratitude to all the members of the Center for Bioinformatics and Computational Biology, the Biological Sciences Graduate Program, and the Computation and Mathematics for Biological Networks (COMBINE) program who have shown their support for my research through seminars and courses. My dissertation has been supported by the COMBINE fellowship and the Ann G. Wylie Dissertation Fellowship awarded by the Graduate School.

I would also like to acknowledge my family, who have had the most significant impact on the beginning of my Ph.D. journey. My parents have been consistently supportive, no matter what decisions I have made. My sister has been my role model since I was little, and her success as a researcher in the biomedical field has inspired me greatly. Most importantly, I want to mention the unwavering support of my husband, Henry Hsueh, throughout my Ph.D. study. Since college, we have been supporting each other's dreams of becoming researchers in our respective fields of passion. He is not only my supporter but also an important collaborator on one of my dissertation projects. Thanks should also go to my two corgis, Pika and Pichu, for their unconditional love and emotional support.

Without the support and resources provided by these kind and impactful individuals around me, my Ph.D. journey at the University of Maryland, College Park, would not have been as fruitful, and my research would not have achieved such high quality.

Table of Contents

List of Tables

List of Figures

Part I

Overview of the Dissertation

Chapter 1: Introduction

1.1 Background

The advancement of biological and computational technologies has enabled the generation of large and complex data in biological sciences research, and has promoted the broad application of machine learning in various biomedical domains over the past decades [1, 2]. As the volume of data increases, multiple research fields gradually transitioned from traditional, model-focused approaches, to approaches that are more data-driven [3]. However, challenges emerge when extracting meaningful patterns and relationships from the large amount of data. Machine learning, including both statistical methods and computational algorithms, aims at learning relationships among data, which can gain important insights from complex and large-scale data by computing the underlying and inherent structures within a data set. However, in biomedical application, there is a wide variety of biological data types, such as genome sequences, gene expressions, and molecular structures [2]. Because of such diversity, the selection of representative features from the high-dimensional data set and the usage of machine learning algorithms are usually problem-specific [4]. Moreover, the rapid growth of data could lead to a lack of substantial labeling, hampering the model performance due to insufficient information [5]. Therefore, it is critical to develop adaptive and advanced strategies to solve the biomedical research problems more

effectively and efficiently.

The dissertation delves into two types of biomedical problems: sustained ocular drug delivery and malaria vaccine antigen identification. This dissertation introduces machine learning-based platforms that can be further extended to various other biomedical applications. In the research domain of sustained ocular drug delivery, patient adherence may be enhanced through maintaining the drug therapeutic levels in the eye. Utilizing melanin residing in the pigmented tissues in the eye, drugs with melanin binding and cell penetration properties can be stored and slowly released from the depot. To impart the desired properties to drugs, peptides, which are short sequences of amino acids, can be used as an adapter and be conjugated to drugs. For peptides with lengths ranging between 7 and 12 amino acids, the number of possible combinations is $\sim 4.3 \times 10^{-15}$, given 20 different amino acids. Among other methods, machine learning is an appropriate approach to rationally design peptides with desired properties. By performing interpretable machine learning techniques, the predictions of the model can be explained, leading to reproducible and transparent results.

Regarding the research of identifying malaria vaccine antigen candidates, effective malaria vaccines targeting either of the most-predominant species, *Plasmodium falciparum* and *Plasmodium vivax*, are an unmet need. Reverse vaccinology, which leverages the wealth of systemic data derived from pathogen genomes, has been adopted to facilitate the process of vaccine development. However, most methods involved filtering candidate antigens with criteria solely based on domain knowledge, and a more comprehensive, data-centric machine learning approach is less explored. Without prior assumptions about the importance of protein variables, machine learning assists in learning the variable importance through

training data sets, and, if provided, the corresponding labels, which indicate whether a protein is an antigen or non-antigen. Nevertheless, due to the fact that validating the antigenicity of a protein requires several rigorous experiments and thus is time-consuming, the antigenic labeling of the proteins is sparse, with only a few proteins labeled as antigens, and the remaining proteins being unlabeled. To overcome such challenge, an advanced approach of positive-unlabeled learning is adapted to identify potential antigen candidates with the goal to further improve the reverse vaccinology pipeline in vaccine development.

1.2 Dissertation Outline

The dissertation is structured so that each chapter corresponds to a manuscript. Part I: Overview of the Dissertation, provides the background and scope of the problem domains, as well as a brief introduction to each chapter. Part II: Multifunctional Peptide Engineering, including Chapters 2 and 3, focuses on using an ensemble machine learning method to engineer multifunctional peptides to improve sustained ocular drug delivery. Part III: Malaria Vaccine Identification, consisting of Chapters 4 and 5, emphasizes on using the positive-unlabeled learning technique to identify potential candidates for malaria vaccine antigens to facilitate the vaccine development process. The appendices in Part IV provide additional materials related to the research findings.

Chapter 2: Machine Learning-Driven Multifunctional Peptide Engineering for Sustained Ocular Drug Delivery, presents research results published in *Nature Communications* (`https://doi.org/10.1038/s41467-023-38056-w`), authored by H. T. Hsueh, R. T. Chou (co-first author), U. Rai, W. Liyanage, Y. C. Kim, M. B. Appell, J. Pejavar,

K. T. Leo, C. Davison, P. Kolodziejski, A. Mozzer, H. Kwon, M. Sista, N. M. Anders, A. Hemingway, S. V. K. Rompicharla, M. Edwards, I. Pitha, J. Hanes, M. P. Cummings, and L. M. Ensign. The research addresses the challenge of delivering drugs into the eye, which stems from the inherent ocular barriers and clearing mechanisms [6, 7], resulting in an intensive dosing schedule that discourages patient compliance. Thus, it is important to develop effective ocular drug delivery systems that can maintain sustained therapeutic drug levels. To assist in the delivery of drugs to the depot formed by the melanin inside the pigmented tissues in the eye, the research leverages machine learning models to guide the engineering of multifunctional peptide adapters, which imparts melanin binding and cell penetrating properties to ocular drugs. My contributions to this work include: **(i)** developing melanin binding peptide microarray assays; **(ii)** designing an ensemble machine learning pipeline to predict melanin binding, cell penetration, and low cytotoxicity peptides; and **(iii)** conducting interpretable machine learning analyses to understand and explain model predictions. The corresponding supplementary information is in Appendix A. H. T. Hsueh, R. T. Chou, J. Hanes, M. P. Cummings, and L. M. Ensign are named as inventors on the U.S. Provisional Patent Application No. 63/340,714, which covers aspects of this work.

Chapter 3: Engineered Peptide-Drug Conjugate Provides Sustained Protection of Retinal Ganglion Cells with Topical Administration in Rats, presents further application of the selected peptide candidate from machine learning models trained in Chapter 2 to another ocular drug, sunitinib, that protects retinal ganglion cells. The research work is published in *Journal of Controlled Release* (`https://doi.org/10.1016/j.jconrel.2023.08.058`), and is authored by H. T. Hsueh, R. T. Chou (co-first author), U. Rai,

P. Kolodziejski, W. Liyanage, J. Pejavar, A. Mozzer, C. Davison, M. B. Appell, Y. C. Kim, K. T. Leo, H. Kwon, M. Sista, N. M. Anders, A. Hemingway, S. V. K. Rompicharla, I. Pitha, D. J. Zack, J. Hanes, M. P. Cummings, and L. M. Ensign. The research focuses on improving the drug delivery system to treat chronic diseases related to the posterior segment of the eye, such as retina, choroid and optic nerve, with the ultimate goal of enhancing patient adherence for better disease management. My contributions to this manuscript include: **(i)** participating in conceptualizing, designing, and interpreting experiments and results; **(ii)** using machine learning models to predict and select peptide candidates to be conjugated to the drug; and **(iii)** applying an object detection technique to facilitate the measurement of cell survival rates to validate the effectiveness of the peptide-drug conjugate in the drug delivery system. The supplementary materials are described in Appendix B.

Chapter 4: Positive-Unlabeled Learning Identifies Vaccine Candidate Antigens in the Malaria Parasite *Plasmodium falciparum*, discusses research that studies approaches to facilitate malaria vaccine development. The manuscript is currently under review by *npj Systems Biology and Applications*, and is a collaborative work by the authors, R. T. Chou, A. Ouattara, M. Adams, A. A. Berry, S. Takala-Harrison, and M. P. Cummings. Malaria is a mosquito-borne infectious disease caused by *Plasmodium* species. The parasite has multiple life stages, and exhibits various immune evasion strategies, such as extremely variable surface antigens [8]. Thus, it is critical to identify conserved potential vaccine antigens that are less variable but with subdominant immunogenicity. The research employs a machine learning-based reverse vaccinology approach to identify potential vaccine antigen candidates for malaria. Since only a few known antigens are selected based on our stringent criteria, the data set is largely unlabeled. My contributions to this research include: **(i)**

adapting a positive-unlabeled learning algorithm to classify *P. falciparum* proteins into antigens or non-antigens while tackling the problem with sparse antigenic labeling; **(ii)** improving the machine learning model by utilizing the tree structure in the positive-unlabeled random forest model; and **(iii)** performing downstream computational analyses to characterize top antigen candidates and to further select a smaller set of antigen candidates based on desired properties for future experimental validation experiments. The supplementary details for Chapter 4 can be found in Appendix C.

Chapter 5: *Plasmodium vivax* Antigen Candidate Prediction Improves with the Addition of *Plasmodium falciparum* Data, highlights research findings of a comprehensive study conducted to improve the identification of vaccine antigen candidates for *P. vivax*, the second-most prevalent species causing malaria, by integrating data from the well-studied species, *P. falciparum*. The study also employs positive-unlabeled learning to construct a machine learning model with multiple different training sets generated by integrating the data of the two species. The research work is jointly conducted by the authors, R. T. Chou, A. Ouattara, S. Takala-Harrison, and M. P. Cummings, and will be submitted to *npj System Biology and Applications* soon. My contributions to the manuscript include: **(i)** applying the positive-unlabeled learning framework described in Chapter 4 to various combinations of training data from *P. vivax* and *P. falciparum*; **(ii)** decomposing and quantifying the effects of the addition of known antigens and/or unlabeled proteins; and **(iii)** characterizing top candidate antigens, analyzing important protein variables for identifying top candidates, and comparing important variables identified from across various machine learning models to gain insights into the proposed integration methodology. Additional information for Chapter 5 is provided in Appendix D.

Part II

Multifunctional Peptide Engineering

Chapter 2: Machine Learning-Driven Multifunctional Peptide Engineering for Sustained Ocular Drug Delivery

2.1 Abstract

Sustained drug delivery strategies have many potential benefits for treating a range of diseases, particularly chronic diseases that require treatment for years. For many chronic ocular diseases, patient adherence to eye drop dosing regimens and the need for frequent intraocular injections are significant barriers to effective disease management. Here, we utilize peptide engineering to impart melanin binding properties to peptide-drug conjugates to act as a sustained-release depot in the eye. We develop a super learning-based methodology to engineer multifunctional peptides that efficiently enter cells, bind to melanin, and have low cytotoxicity. When the lead multifunctional peptide (HR97) is conjugated to brimonidine, an intraocular pressure lowering drug that is prescribed for three times per day topical dosing, intraocular pressure reduction is observed for up to 18 days after a single intracameral injection in rabbits. Further, the cumulative intraocular pressure lowering effect increases ∼17-fold compared to free brimonidine injection. Engineered multifunctional peptide-drug conjugates are a promising approach for providing sustained therapeutic delivery in the eye and beyond.

2.2 Introduction

In many disease settings, sustained delivery of therapeutic levels of drug can improve treatment efficacy, reduce side effects, and avoid challenges with patient adherence to intensive dosing regimens [9, 10]. This is particularly critical in the management of chronic diseases, where long-term adherence to medication usage and clinical monitoring can suffer [11, 12]. In the ophthalmic setting, the leading causes of irreversible blindness and low vision are primarily age-related, chronic diseases, such as glaucoma and age-related macular degeneration [13–15]. Recent approvals of devices that provide sustained therapeutic release, such as the Durysta® intracameral implant for continuous delivery of an intraocular pressure (IOP) lowering agent, and the surgically implanted port-delivery system that provides continuous intravitreal delivery of ranibizumab, highlight the importance of these next generation approaches for ocular disease management [16–19]. Conventionally, sustained therapeutic effect is achieved by an injectable or implantable device that controls the release of the therapeutic moiety into the surrounding environment. However, these devices typically require injection through larger gauge needles or a surgery for implantation, with both procedures having associated risks [20–22]. Further, the buildup of excipient material, the need for device removal, and the potential for foreign body reaction can cause further issues [18, 23, 24].

One approach for circumventing the issues associated with sustained release devices is to impart enhanced retention time and therapeutic effect to drugs upon administration to the eye without the need for an excipient matrix/implant. Binding to melanin, a pigment present within melanosomes in multiple ocular cell types, was previously reported to affect

ocular drug biodistribution [25]. Due to the low turnover rate of ocular melanin, a drug that can bind to melanin may accumulate in pigmented eye tissues, leading to drug toxicity or drug sequestration [26, 27]. However, with the right balance of melanin-binding affinity and capacity, melanin may act as a sustained-release drug depot in the eye that results in prolonged therapeutic action [28]. Several drugs have been demonstrated to have intrinsic melanin binding properties due to particular physicochemical properties, which in some cases, prolongs the pharmacologic activity in the eye [28–30].

To impart beneficial melanin-binding properties to drugs, one approach is to engineer peptides with high melanin binding that could be conjugated to small molecule drugs through a reducible linker. Thus, the peptide would provide enhanced retention time, while the linker would ensure that drug could be released and exert its therapeutic action in a sustained manner. In addition, there are available databases describing how peptide sequence affects cell-penetration [31, 32], and separately cytotoxicity [33], enabling the potential for engineering multifunctional peptides that can be chemically conjugated to drugs. Incorporating multiple functions into one peptide sequence remains challenging, and thus multifunctional peptides are often designed by fusing peptides via a linker, thus forgoing potentially more efficient rational design, or by testing additional properties on peptides with known functions [34–36]. In contrast, machine learning could allow for designing peptide sequences that simultaneously provide multiple desired properties.

Here, we describe the development of engineered peptides informed by machine learning, which have three properties: high binding to melanin, cell-penetration (to enter cells and access melanin in the melanosomes), and low cytotoxicity. As there was no prior information for how peptide sequences affect melanin binding, we experimentally determine the

effect of peptide sequence on melanin binding using a microarray. We then apply machine learning-based analyses to identify peptide sequences that display all three desired properties. Importantly, with the Shapley additive explanation (SHAP) analysis [37] of peptide variables, the machine learning model interpretation provides additional insights and reasoning for the multifunctionality of the peptides. As a proof-of-principle, we demonstrate here that an engineered peptide, HR97, can be conjugated to the intraocular pressure (IOP) reducing drug, brimonidine tartrate. A single intracameral (ICM) injection of the HR97-brimonidine conjugate is able to provide sustained IOP reduction in normotensive rabbits compared to ICM injection of an equivalent amount of brimonidine tartrate, or a topical dose of Alphagan® P 0.1% eye drops. Further, the maximum measured change in IOP from baseline (ΔIOP) is increased with ICM injection of the HR97-brimonidine conjugate. We anticipate that engineered peptide-drug conjugates will facilitate the development of implant-free injectables for use in a variety of ophthalmic indications.

2.3 Results

2.3.1 Development of high throughput melanin binding peptide microarray methodology

To determine how peptide sequence affects melanin binding properties, we adapted a high-throughput flow-based peptide microarray system to characterize melanin binding events (Fig. 2.1a). Commercially available eumelanin was processed into nanoparticles (mNPs) to prevent sedimentation and provide reproducible surface area available for binding to peptides printed on the substrate surface. The mNPs had an mean size of 200.7$\pm$5.99

nm and ζ-potential of -23.7 ± 1.39 mV (Fig. 2.1b, c). The mNPs were further biotinylated (b-mNPs) to facilitate fluorescent labeling with streptavidin DyLight680. The b-mNPs showed slightly larger mean size of 216.0 ± 14.85 nm and ζ-potential of -21.2 ± 2.15 mV (Fig. 2.1b, c), and maintained similar spherical morphology (Fig. A.1a) and binding to small molecule drugs brimonidine tartrate and sunitinib malate (Fig. A.1b). The first microarray was printed with 119 peptides to screen flow conditions for the highest fluorescent reporter signal, which identified that the 500 μg/mL of biotinylated mNPs in pH 6.5 PBS buffer at room temperature was optimal (Fig. 2.1d and Fig. A.2). We then used the fluorescent reporter signals to construct a melanin binding classification random forest model. The prediction accuracy was 0.92. The permutation-based variable importance analysis [38] further revealed that the net charge, basic amino acids, and isoelectric point (pI) may contribute to distinguishing melanin binding and non-melanin binding peptides (Fig. 2.1e).

2.3.2 Training of the melanin binding regression model

A second larger peptide screen was implemented to generate melanin binding data to use for the additional model generation (Fig. 2.2a). Specifically, we used the trained random forest model to predict melanin binding for ~630,000 randomly generated peptides, and those classified as melanin binding were selected. A total of 5499 peptides were printed in duplicate, and the fluorescent reporter intensities were reported as the amount of the b-mNPs that bind to the printed peptides on the microarray. Surprisingly, we identified 780 peptides displaying higher levels of fluorescent reporter intensities than any

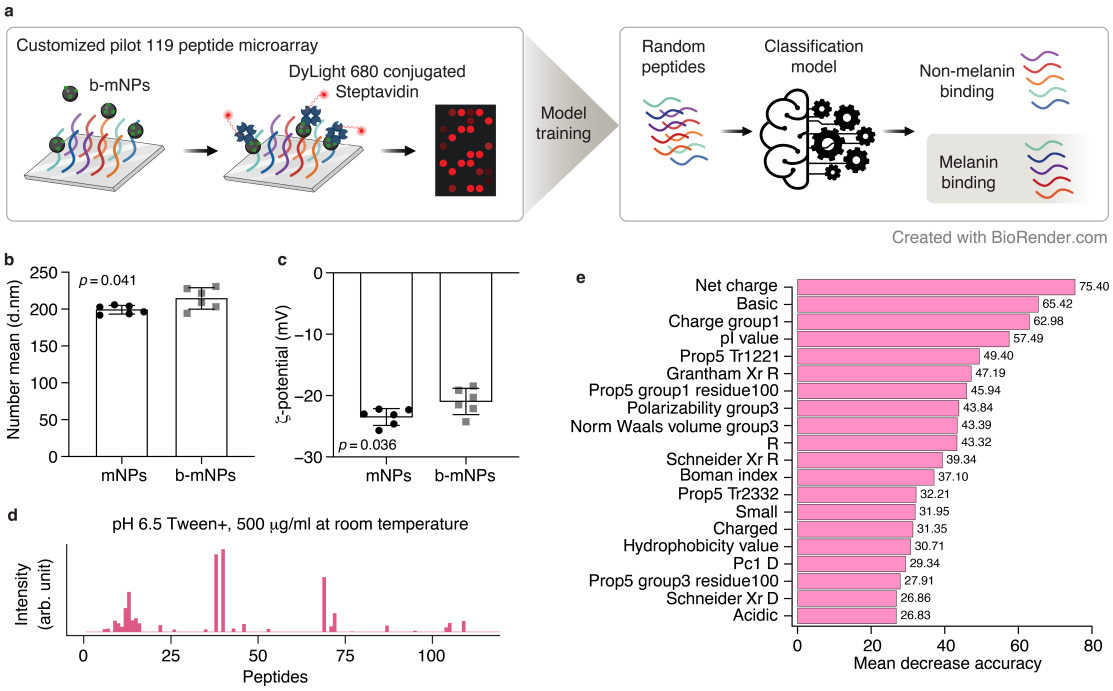


Figure 2.1 Pilot 119 melanin binding peptide microarray screening with machine learning analysis. **a** Schematic illustration of the first peptide microarray. Peptides were anchored to a microarray, and melanin nanoparticles (mNPs) with surface biotinylation (b-mNPs) were flowed over to characterize binding events. The fluorescence intensity of the biotin was detected using DyLight 680-conjugated streptavidin to quantify melanin binding for each peptide. An initial classification model was trained using the data generated. Random peptides were then classified by the model as melanin binding or non-melanin binding. Created with BioR- ender.com. **b,c** Plot showing the sizes **(b)** and ζ-potential **(c)** of mNPs (black dots, $n = 6$ and b-mNPs (gray squares, $n = 6$). Data are presented as mean ± SD. Group means were compared using Student's t tests (two-tailed). **d** The optimal interaction profiling of b-mNPs against 16 positive control peptides (peptide numbers: 1–16) and 103 random peptides (peptide numbers: 17–119). **e** Permutation-based variable importance analysis of the melanin binding classification random forest. The x-axis indicates the mean decrease in prediction accuracy after variable permutation. The values are shown at the end of the bars. The top 20 important variables ranked by mean decrease in accuracy are shown.

of the 16 peptides described in the literature that bound to human melanoma cells [39] and melanized *C. neoformans* [40], which were previously screened by the phage display technique. Furthermore, there were 758 peptides showing higher fluorescent values than the highest melanin binding peptides (661.5 arb. units) from the 119-peptide microarray, demonstrating the enrichment of melanin binding properties from training the random forest model. Next, the fluorescent reporter intensities values were used as the response variable in training a regression model. Applying a variable reduction procedure using random forest to eliminate less informative variables from the data set, reduced the number of variables from 1094 to 64 (Fig. A.3a), and model performance measured by the coefficient of determination (R^2) improved from 0.48 to 0.53. A wide array of machine learning models was explored and trained on the variable-reduced data set and were integrated with a super learning (SL) framework that combined various types of base models weighted using a meta-learner. By applying the iterative base model filtering procedure (Fig. 2.2b), the complexity of the SL was further reduced. To explore other combinations of base models in the SL ensemble, homogeneous base models consisting of models from only one algorithm family were constructed. A nested cross-validation (Fig. 2.2c) was applied to estimate an unbiased generalization performance. All SL models with base model reduction were selected as the top model in the inner loop cross-validations, and the performance evaluated in the outer loop cross-validation improved to $R^2 = 0.54 \pm 0.01$ (Table A.1). The reduced SL was selected amongst 31 competitive models as the final melanin binding regression model. When training the same set of models on the whole data set, and number of base models in the SL was reduced from 907 to 38 (Fig. 2.2d). Adversarial computational control was performed, and the generalization performance was $R^2 = -0.04 \pm 0.02$, indicating

that the machine learning was effective in learning meaningful relationships in the melanin binding data set.

2.3.3 Training of cell-penetration and cytotoxicity classification models

Engineered peptides must enter cells to reach and bind to melanin within the melanosomes and should be minimally toxic to cells. Thus, the SkipCPP-Pred [31] and the ToxinPred [33] databases were used to create SL classification ensembles to engineer tri-functional peptides. Variable reduction decreased the number of variables from 1094 to 11 for the cell-penetration data set (Fig. A.3b) and from 1094 to 56 for the cytotoxicity data set (Fig. A.3c). The prediction accuracies calculated from out-of-bag samples improved from 0.91 to 0.93 and from 0.951 to 0.958 for cell-penetration and cytotoxicity, respectively. We subsequently trained base models and SL ensembles, and the generalization performances in terms of Matthews correlation coefficient (MCC), F_1 (harmonic mean of precision and recall), and balanced accuracy for cell-penetration were 0.79±0.01, 0.90±0.01, and 0.90±0.01, respectively; and those for cytotoxicity were 0.88 ± 0.004, 0.92 ± 0.002, and 0.95 ± 0.002, respectively (Tables A.2, A.3). The number of base models in the reduced SL models trained on the whole data sets were decreased from 310 to 65 for cell-penetration, and from 311 to 22 for cytotoxicity (Fig. A.4). There were 300 competitive cell-penetration models and 175 competitive cytotoxicity models. A GBM model and the reduced SL were selected as the final predictive cell-penetration and cytotoxicity models. Similar to melanin binding, adversarial controls had decreased generalization performances, where the MCC, F_1, and balanced accuracy were -0.002 ± 0.05, 0.52 ± 0.03, and 0.50 ± 0.03 for cell-penetration, and

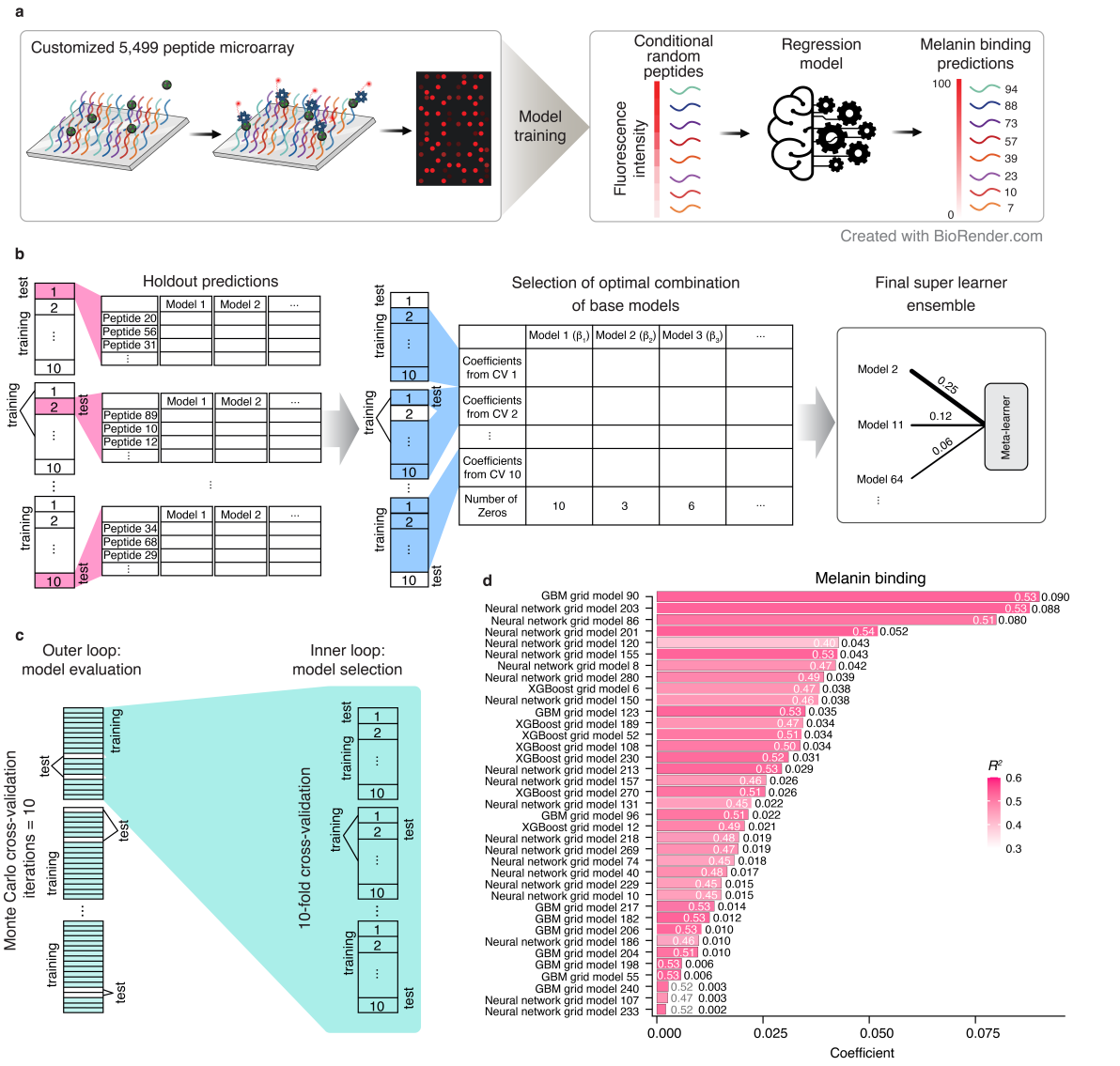


Figure 2.2 Schematic of the machine learning pipeline based on the super learner framework for the melanin binding data set. **a** Scheme of a larger microarray, which includes 5499 peptides used to train a regression super learner. Random peptides were generated based on position-dependent amino acid frequencies calculated using the second peptide array data, and the melanin binding levels were predicted. Peptides with desired melanin binding levels were selected for further experimental validation. Created with BioRender.com. **b** Scheme of the super learner complexity reduction. Holdout predictions of peptides (shown as rows) were generated for each base model (shown as columns) with tenfold cross-validation (CV) on the input data set. A meta-learner (generalized linear model) was fitted on the holdout predictions with another tenfold cross-validation. The number of base models was reduced by applying an iterative reduction procedure (see Section 2.5). The final super learner ensemble was trained on the input data set with the optimal combination of the selected base models. **c** Scheme of the machine learning pipeline for an unbiased model performance evaluation. The nested cross-validation

Figure 2.2 ***(previous page)*** includes an outer loop for model evaluation and an inner loop for model selection (cyan). The outer loop generated 10 sets of train-test splits using a Monte Carlo method, and the inner loop generated 10 sets of train-test splits using a modulo method. **d** Plot of the base models of the final melanin binding super learner. Coefficients of determination (R^2) are denoted with color and conveyed as white text on the bars or gray text adjacent bars. Base model coefficients are indicated at the bar ends. There is one model having zero coefficient and not shown. See Sections 2.5 and A.1 for information about model hyperparameter details and statistics of model performance.

0.001 ± 0.01, 0.05 ± 0.02, and 0.62 ± 0.04 for cytotoxicity.

2.3.4 Validation of predicted peptide properties *in vitro*

A position-dependent amino acid frequency matrix was used to generate 127 peptides that spanned the range of low to high predicted melanin binding. Among the 127 peptide candidates, 113 peptides were classified as cell-penetrating and 117 peptides were predicted as non-toxic. To experimentally measure melanin binding *in vitro*, biotinylated peptides were incubated with mNPs, and the bound fraction was calculated using an avidin-based fluorescent reporter (Fig. 2.3a). The Pearson correlation coefficient was computed to compare the predicted and experimental melanin binding values, and the correlation coefficient was $r = 0.84$, showing a high level of correlation between the predicted and experimental values (Fig. 2.3b). We next characterized how the predicted cell-penetrating properties of the peptides affected cell uptake in a retinal pigment epithelium cell line (ARPE-19). ARPE-19 cells were cultured using standard methods (non-induced, $n = 3$) and using culture conditions that induce melanin production (induced, $n = 3$) [28]. A positive correlation was observed between the measured *in vitro* melanin binding of the peptides and the intracellular peptide concentrations in melanin-induced cells for cell-penetrating peptides ($r = 0.77$, $p < 2.2 \times 10^{-16}$) but not non-cell-penetrating peptides ($r = 0.28$, Fig. 2.3c, d),

suggesting correlation between the two properties. Further, peptides predicted to be cell-penetrating demonstrated significantly higher intracellular concentrations (median 229.4 pmol/100 K cells) than those of non-cell-penetrating peptides (median 26.7 pmol/ 100 K cells) in the melanin-induced cells ($p = 6.9 \times 10^{-6}$, Fig. 2.3e). In contrast, the intracellular peptide concentrations were not affected by the predicted properties in non-induced cells (Fig. A.5).

2.3.5 Analysis of peptide variables that contribute to observed properties

To identify which peptide variables contributed to the properties observed *in vitro*, Shapley additive explanation (SHAP) analysis of the final predictive models was performed. The results showed that peptide property predictions were based on contribution from multiple variables. More specifically, basic peptides and higher net charge variables had higher contributions to melanin binding predictions (Fig. 2.4a), which was consistent with the top variables identified by the random forest classification model trained on the pilot peptide microarray. Similarly, higher net charge and higher isoelectric point contributed more to cell-penetration (Fig. 2.4b), and less free cysteines had more influence on non-toxic predictions (Fig. A.6). To understand how reliable the interpretable results were, adversarial controls were constructed with the final predictive models using a 10-fold cross-validation. Indeed, the distributions and levels of variable contributions changed for melanin binding, cell-penetration, and cytotoxicity (Fig. A.7). Among all the peptide candidates, HR97 (FSGKRRKRKPR) was selected based on combination of the three peptide properties (melanin binding$_{\mathrm{HR97}} = 79.1 \pm 0.7\%$, cell uptake$_{\mathrm{HR97}} = 759.9 \pm 19.6$ pmol/100 K cells, non-toxic$_{\mathrm{HR97}}$

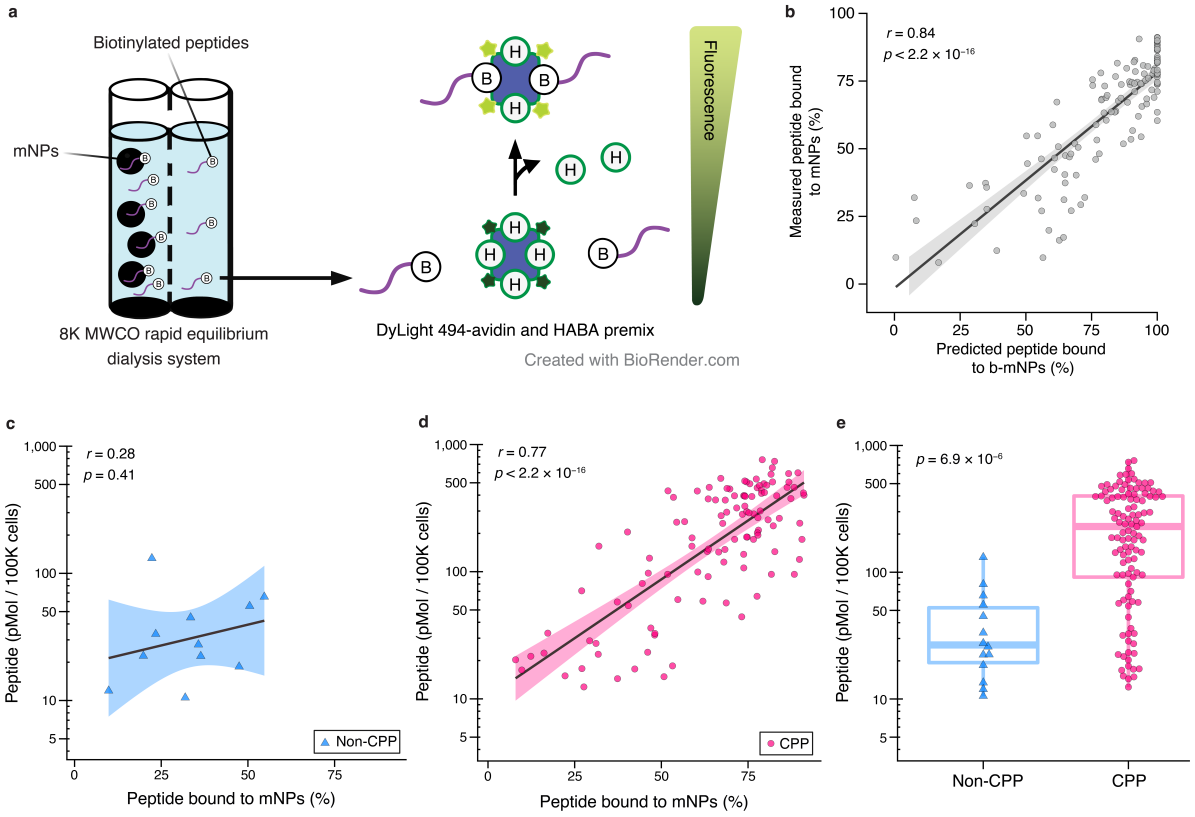


Figure 2.3 Experimental validations of final model predictions on melanin binding and cell-penetration. **a** Schematic showing an *in vitro* melanin binding assay with melanin nanoparticles (mNPs) using a biotin quantification kit. The DyLight 494-tagged avidin emitted fluorescence when the biotinylated peptides displaced the weakly interacting 4′-hydroxyazobenzene-2-carboxylic acid (HABA or H). Created with BioRender.com. **b** Plot of the relationship between predicted melanin binding and binding measured experimentally *in vitro*. The x-axis indicates melanin binding predictions from the final super learner, and the y-axis indicates the experimental melanin binding values ($n = 4$ for each peptide). Dots represent the mean value for peptides. The black linear trend line conveys the Pearson correlation relationship (two-tailed), and the gray area indicates the 95% confidence interval. **c, d** Comparison of melanin binding and cell-penetration in melanin-induced human adult retinal pigment epithelial (ARPE-19) cells. Blue triangles denote predicted non-cell-penetrating peptides (non-CPP), and magenta dots represent predicted cell-penetrating peptides (CPP). The x-axes indicate melanin binding measured *in vitro* ($n = 4$ for each peptide), and the y-axes convey intracellular peptide concentration measured from the cell uptake assay ($n = 3$ for each peptide). Black linear trend lines indicate Pearson correlation relationships, with 95% confidence intervals shown as shaded areas. The correlation coefficients

Figure 2.3 ***(previous page)*** and p-values (two-tailed) are shown. **e** Summary of CPP ($n = 113$) and non-CPP ($n = 14$) intracellular concentrations. Box plot conveys median (middle line), 25th and 75th percentiles (box), and the 1.5 × interquartile range (whiskers). The p value was calculated using a Mann–Whitney U test (two-tailed).

= 96.9%, Fig. 2.4c). HR97 had the highest intracellular concentration, which outperformed the well-characterized cell-penetrating peptide fragment of the HIV trans-activator protein (TAT_{47-57}, YGRKKRRQRRR). HR97 demonstrated increased cell uptake compared to TAT_{47-57} in both the induced ARPE-19 cells (cell $uptake_{HR97} = 759.9 \pm 19.6$ pmol/100 K cells, cell $uptake_{TAT47-57} = 457.1 \pm 34.2$ pmol/100 K cells) and the non-induced cell type (cell $uptake_{HR97} = 82.5 \pm 9.1$ pmol/100 K cells, cell $uptake_{TAT47-57} = 68.3 \pm 4.6$ pmol/100 K cells). In addition, HR97 showed no sign of cytotoxicity in ARPE-19 cells at concentrations up to 5 mg/mL (Fig. A.8). HR97 predictions embodied all the properties that were the largest contributors to each functionality, including being basic (63.64% basic amino acids), possessing a high net charge (6.98) and a high isoelectric point value (12.99), and no cysteines (Fig. 2.4d–f). By visualizing the peptide design space defined by the sequences and variables used in training the desired functional properties, the peptide candidates with high melanin binding predictions were shown up in the same cluster, showing similar sequence motifs and physiochemical properties (Fig. 2.5a, b). Further, peptides predicted to have high melanin binding were mostly predicted to be cell-penetrating, but cell-penetrating peptides may not be melanin binding (Fig. 2.5c). The results also suggest that some melanin binding peptides may be toxic (Fig. 2.5d).

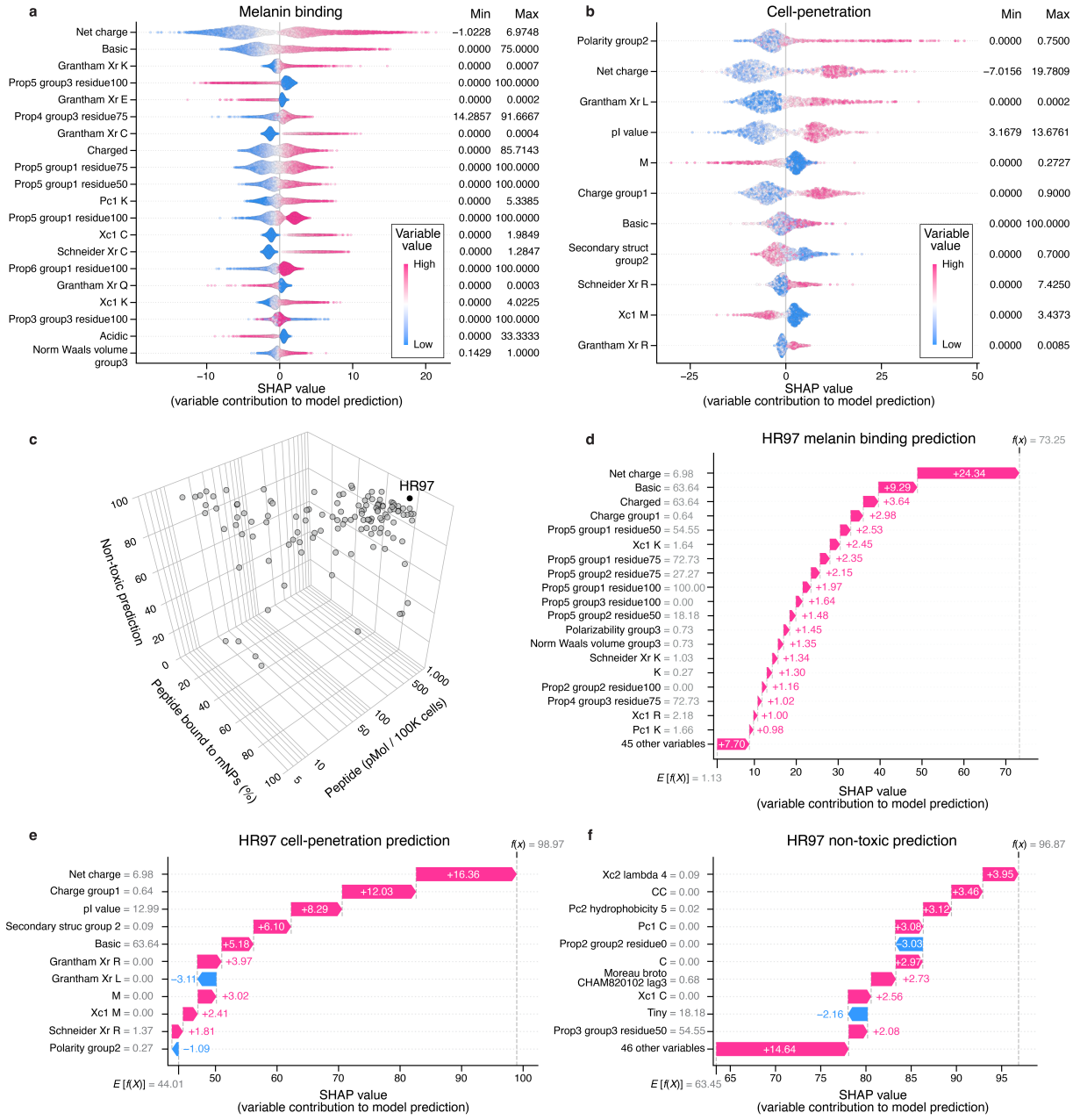


Figure 2.4 Melanin binding, cell-penetration model interpretation, and variable contributions to HR97 multifunctional peptide predictions. Overall variable contributions to model predictions for **(a)** melanin binding and **(b)** cell-penetration. The top important variables analyzed using Shapley additive explanations (SHAP) are shown. Dots represent peptides from cross-validation test sets. The x-axes indicate SHAP values, indicative of variable contributions to model prediction ranging from 0 to 100. The variables were ranked based on the difference between the maximum and minimum SHAP values. The color gradient indicates the variable values normalized by percentile ranks. Higher variable values are indicated by darker magenta color and lower values by darker blue color. The minimum and maximum variable values are

Figure 2.4 ***(previous page)*** noted on the right of each subplot. **c** Scatter plot showing the *in vitro* melanin binding, *in vitro* cell-penetration, and predicted cytotoxicity values of the 127 candidate peptides. Dots represent peptides. HR97 (black dot) was selected based on the optimal multifunctional combination. **d–f** Variable contributions to HR97 multifunctional predictions for melanin binding, cell-penetration, and cytotoxicity. The top variables ranked by absolute SHAP values are shown. Magenta bars indicate positive contributions, and blue bars are negative contributions. The y-axis labels convey variable names and their values for HR97. E[f(X)] denotes the expected prediction value, and f(x) is the final prediction, calculated from the sum of all SHAP values plus E[f(X)].

2.3.6 Characterization and validation of a peptide-drug conjugate *in vivo*

To investigate the effect of peptide conjugation on drug pharmacodynamics, we chose brimonidine tartrate, a topical IOP lowering drug prescribed for glaucoma treatment. The HR97 peptide was conjugated to brimonidine (HR97-brimonidine) via a quaternary-ammonium traceless linker system, and the structure of the intermediates and the purified conjugate were validated by NMR and MALDI-TOF (Figs. A.9–A.12). Conjugation to HR97 provided a ∼10-fold increase in the *in vitro* melanin binding capacity of brimonidine (5.9×10^{7} K_d (M) vs. 5.0×10^{-8} K_d (M)), which brought the binding capacity closer to other drugs with high intrinsic melanin binding, such as sunitinib malate (Fig. 2.6a) [28, 41–45]. When incubated in human aqueous fluid, only ∼7% of the brimonidine was released from the HR97-brimonidine conjugate over 28 days *in vitro* (Fig. 2.6b). However, upon incubation with supraphysiological concentrations of human cathepsin cocktails to enzymatically cleave the linker, ∼52% of the brimonidine was liberated within 48 h (Fig. 2.6c). The effect of the HR97-brimondine conjugate on IOP was then evaluated in normotensive Dutch Belted rabbits. A single topical dose with the commercial brimonidine eye drop ($n = 5$) was found to provide a peak reduction in IOP from baseline (ΔIOP) of -3.0 ± 0.82 mmHg that

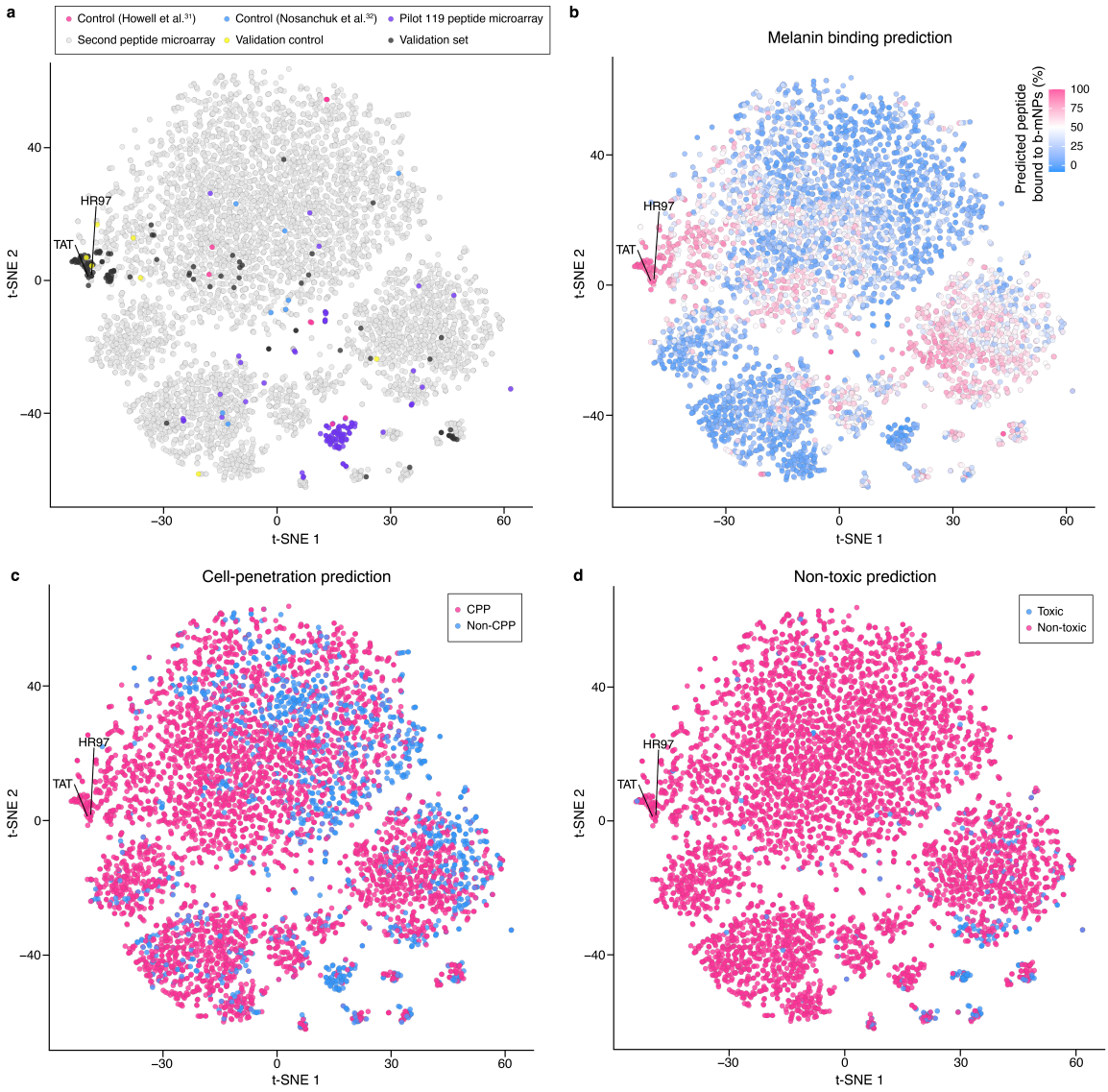


Figure 2.5 Visualization of the peptide design space based on sequences and physiochemical properties. **a** t-distributed stochastic neighbor embedding (t-SNE, used to visualize high-dimensional data) plots showing the peptide design space defined by the combination of one-hot encoded peptide sequences and variables used in melanin binding, cell-penetration, and cytotoxicity model training. Dots represent control peptides from Howell et al. [39] (magenta) and Nosanchuk et al. [40] (blue); peptides used in the pilot (purple) and second (gray and yellow) melanin binding peptide microarrays; and multifunctional peptide candidates (black and yellow) used in the validation experiments. HR97 and TAT are noted. **b** t-SNE plot of peptides colored by melanin binding prediction. Higher melanin binding values are colored by darker magenta and lower by darker blue. **c** t-SNE plot of peptides colored by cell-penetration prediction. Magenta dots represent predicted cell-penetrating peptides (CPP), and blue dots are predicted non-cell-penetrating peptides (non-CPP). **d** t-SNE plot of peptides colored by cytotoxicity prediction. Blue dots denote predicted toxic peptides, and magenta dots indicate non-toxic peptides.

recovered to baseline within 8 h (Fig. 2.6d). In contrast, a single ICM injection of the HR97-brimonidine conjugate resulted in a greater peak ΔIOP compared to an ICM injection of brimonidine solution at 2 days (-4.9 ± 0.46 mmHg vs. -2.6 ± 1.65 mmHg, $p < 0.05$, red arrow). In a separate experiment, ICM injection of saline or HR97 ($n = 5$ for each) resulted in a similar decrease in IOP that returned to baseline by day 3, and ICM injection of a physical mixture of HR97 and brimonidine tartrate ($n = 5$) resulted in a similar IOP profile to the brimonidine solution, returning to baseline by day 8 (Fig. A.13). To ensure that the dramatic decrease in IOP with the HR97-brimonidine conjugate was not due to toxicity, a board-certified ophthalmologist evaluated the eyes injected with the HR97-brimonidine conjugate on day 7. It was observed that the lids, lashes, and conjunctiva were normal, the corneas were clear, the corneal endothelium was normal without any pigment deposition, the anterior chambers were normal depth, there was no apparent inflammation or fibrin strands, the lenses were clear, and the iris pigmentation was symmetric. According to the same evaluation methods, no ocular toxicity was observed upon ICM injection of saline, HR97, or a physical mixture of HR97 and brimonidine tartrate for at least 28 days (Tables A.4–A.7). The mean ΔIOP in the HR97-brimonidine conjugate group remained significantly larger than in the rabbits dosed with brimonidine solution or the physical mixture of HR97 and brimonidine tartrate for up to 14 days (Fig. A.6d, A.13). Further, the time for the mean ΔIOP to return to baseline was 20 days in the HR97-brimonidine conjugate group compared to 8 days in both groups of rabbits dosed with brimonidine solution or the physical mixture of HR97 and brimonidine tartrate. When summing the area under the curve (AUC_{last}) for the cumulative ΔIOP over the 20-day measurement period after ICM injection, the HR97-brimonidine conjugate showed a ∼17-fold greater AUC compared

to brimonidine solution ($p < 0.001$) (Fig. A.6e). A pharmacokinetic study was conducted separately to characterize the intraocular distribution of brimonidine after ICM injection of HR97-brimonidine in Dutch Belted rabbits. The brimonidine concentration remained relatively high in the pigmented iris tissue (980 ng/g) compared to less pigmented parts of the eye, such as the aqueous humor (0.4 ng/g) and the retina (8.3 ng/g) up to 28 days after a single ICM injection (Fig. A.6f). The brimonidine concentration in the aqueous on day 7 (83.3 ng/g) was similar to what we previously reported at 2 h after a single drop of Alphagan P (0.15%) (105 ng/g), which was the time with the largest IOP reduction in that study [46]. On day 14 after ICM injection of HR97-brimonidine, the brimonidine concentration in the aqueous (3.9 ng/g) was similar to what we previously reported at 4 h after a single drop of Alphagan P (0.15%) (4 ng/g) [46].

2.4 Discussion

Chronic eye diseases such as glaucoma require continuous treatment to prevent disease progression. Eye drops are the most common dosage form of glaucoma therapy, though low adherence to intensive drop dosage schedules is a major challenge in disease management [11,47,48]. One study using an electronic monitoring device found that only 64% of patients adhered to the three-times daily dosing schedule for brimonidine eye drops over a 4-week period, even though they were aware of the monitoring [49]. Sustained drug delivery systems may be an attractive alternative for the management of chronic ocular diseases like glaucoma. The first sustained-release polymer-based implant for glaucoma treatment, Durysta®, was recently approved for sustained IOP lowering for several months with a

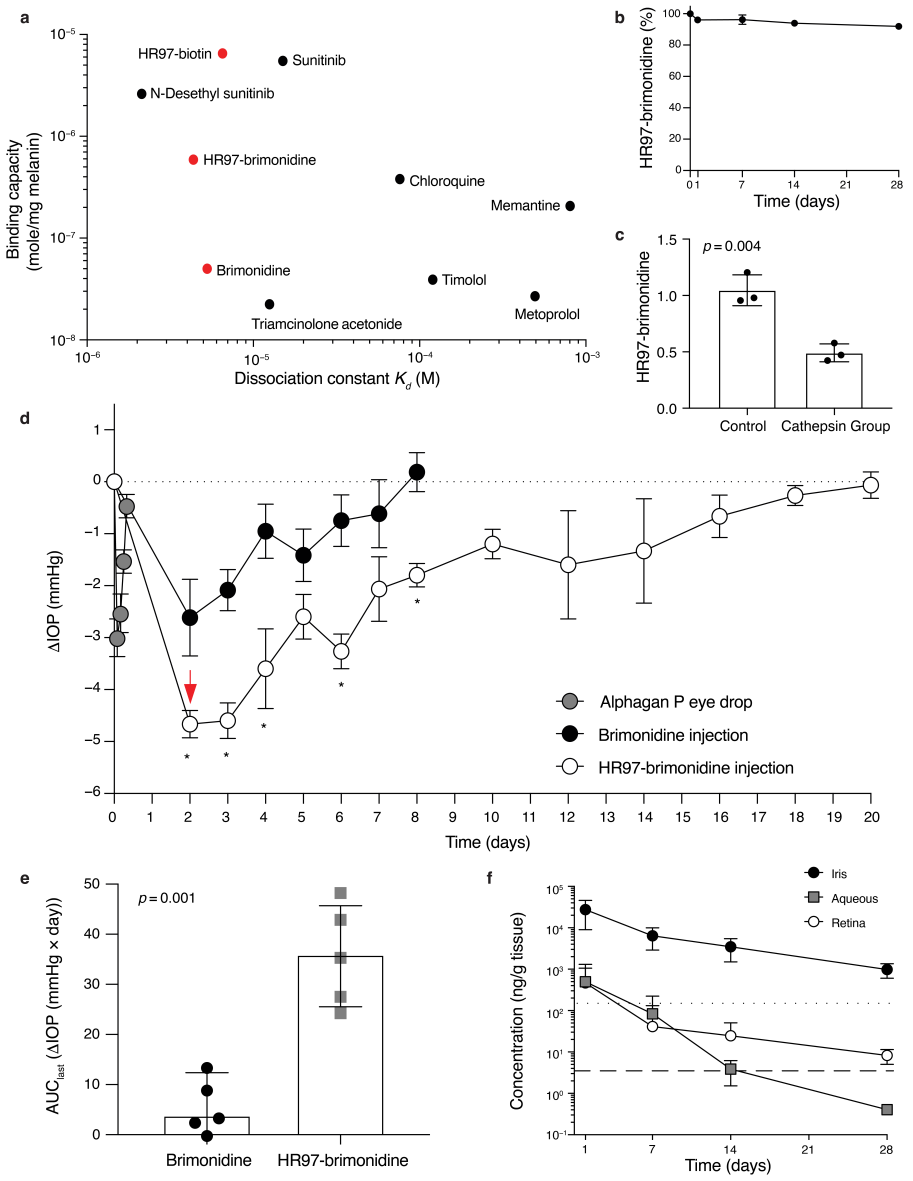


Figure 2.6 Characterization of HR97-brimonidine *in vitro* and *in vivo*. **a** *In vitro* binding capacity and dissociation constant of HR97-biotin, HR97-brimonidine, and brimonidine characterized using a melanin nanoparticle (mNP) assay (red dots, $n = 3$–5). Values shown for comparison include those we previously measured for sunitinib and N-desethyl sunitinib [28], and literature values for other ophthalmic drugs [41–45]. **b** *In vitro* stability of HR97-brimonidine conjugate in human aqueous humor for 28 days. The percent remaining was normalized to the starting concentration on day 0 ($n = 3$). Data are shown as mean ± SD. **c** Cathepsin cleavage assay of the HR97-brimonidine conjugate. HR97-brimonidine ($n = 3$) were incubated with human cathepsin cocktails or buffer only for 48 h at 37 °C (two-tailed t-test). Data are shown as mean ± SD. **d** Comparison of the intraocular pressure (IOP) change from baseline (ΔIOP) after a single ICM injection of HR97-brimonidine conjugate (white dots), brimonidine solution (black dots, 200 μg brimonidine equivalent), and a single drop of Alphagan P (gray dots, 0.15%) in normotensive

Figure 2.6 ***(previous page)*** Dutch Belted rabbits ($n = 5$ per group). The IOP was measured every 1–2 days until returning to the baseline. The red arrow highlights the further decrease in IOP provided by the HR97-brimonidine. Two-tailed t-test was used, $^*p < 0.05$ (adjusted p values for days 2, 3, 4, 6, and 8 were 0.044, 0.007, 0.038, 0.007, 0.007, respectively). Data are presented as mean ± SEM. **e** Cumulative ΔIOP of brimonidine (black dots) and HR97-brimonidine (gray squares) after ICM injection. The cumulative ΔIOP was characterized by calculating the area under the curve over the 20-day measurement period (AUC_{last}, $n = 5$). Two-tailed t-test was used. Data are presented as mean ± SD. **f** Levels of brimonidine in the iris (black dots), aqueous (gray squares), and retina (white dots, $n = 3$–4) over time after ICM injection of HR97-brimonidine (200 μg brimonidine equivalent). The concentrations of brimonidine measured in the aqueous after a single drop of Alphagan P (0.15%) as part of a previous study [46] at 2 h (maximal IOP lowering time point; dotted line) and 4 h (dashed line) after dosage are shown. Data are shown as mean ± SD.

single ICM injection [17]. However, the polymer matrix typically took longer to biodegrade than the duration of drug release, and repeated injection with additional implants was associated with increased risk of corneal endothelial cell loss and other corneal adverse reactions [50]. In contrast to conventional polymer-based sustained drug delivery systems, the approach we describe here does not require an implant or large amounts of excipients that will remain in the eye for extended periods. By utilizing short peptide sequences that impart melanin binding to the drug conjugate, a sustained intraocular drug release system was created without the need for a polymer matrix.

Ocular melanin is a biopolymer that resides within melanosomes in pigmented ocular tissues, including the iris, ciliary body, choroid, and retinal pigment epithelium (RPE) [51]. Although the amount of pheomelanin in the eye varies depending on eye color, the amount of eumelanin in ocular tissues, including the RPE, iris pigment epithelium, and pigmented ciliary epithelium is more consistent across the population [52]. It has been described that drug binding to melanin and accumulation inside cells may diminish therapeutic effect by sequestering the drug or causing ocular toxicity [26,27]. In the case of atropine, the intrinsic

melanin binding properties were shown to lead to prolonged residence time in pigmented rabbit ocular tissues [53], and a sustained miotic response in pigmented rabbits [29]. In addition, we previously demonstrated that improving the intraocular absorption of sunitinib, a drug with relatively high melanin binding capacity, with a novel gel-forming hypotonic eye drop led to prolonged therapeutic effect of up to 1 week after dosing [28]. Indeed, a recent study used machine learning methods to characterize the structural features of small molecule drugs that impact intrinsic melanin binding, leading to the development of a model that predicted intrinsic melanin binding with 91% accuracy [30]. These findings motivated us to develop engineered adaptors designed to impart tunable melanin binding properties to small molecule drugs used to treat ocular diseases. Further, as melanin is contained within cells, the engineered adaptor should additionally provide cell penetration. Here, we developed a machine learning-based methodology to engineer tri-functional peptides that displayed melanin binding, cell-penetration, and non-toxic properties. The peptide sequence that provided the optimal combination of high melanin binding, high cell-penetration, and low cytotoxicity, HR97, was then conjugated to brimonidine as a proof-of-principle. The HR97-brimonidine conjugate provided up to 18 days of IOP lowering with a single ICM injection in normotensive rabbits, which contrasts with the 8 h-effect provided by a brimonidine eye drop.

Peptides are short sequences of amino acids that can have many combinations with diverse biological functions. Compared to other aptamers and small molecule drug libraries, peptides are relatively cost effective to synthesize and are relatively easy to modify or conjugate to small molecule drugs [54]. Currently, there are more than 80 FDA-approved peptide drugs and more than 600 in clinical and pre-clinal trials [55–57]. Peptides optimized

for a single function, either exhibiting cell-penetration or cell targeting properties, have been widely exploited as drug carriers to shuttle drugs across biological barriers [58–60]. Peptides such as TAT, penetratin, PEP-1 and polyarginine (R6 or R8) and have been conjugated with various cargos for targeting the anterior and posterior segment [61–68]. For example, various fluorescein conjugated peptides were screened for the ability to cross porcine cornea *ex vivo* [68, 69]. Penetratin (PNT) showed an eightfold increase in permeability compared to PEP-1, though most of the peptide was found to be sequestered within cells rather than having crossed the cornea [68, 69]. In another study, TAT peptide was conjugated to human acidic fibroblast growth factor (aFGF) and applied topically to rat eyes [70]. They found that the conjugates reached the retina with a t_{max} of 30–60 min and with possible mechanism of conjunctival-scleral penetration route [70]. However, it is known that drugs can more easily reach the posterior segment with topical administration in rat and mouse eyes compared to larger eyes, such as rabbits [71–73].

Many peptide screening technologies have been developed for identifying novel functional peptides, including phage display, mRNA display, and peptide microarray [74–76]. Phage display and mRNA display are capable of screening a larger number of peptides ($\sim 10^{11}$–10^{13}) compared to peptide microarray ($\sim 10^{5}$). However, in phage display and mRNA display, the peptide sequences are randomly generated with fixed ratios of amino acids [75]. In contrast, coupling computationally generated peptide sequences with peptide microarrays has the advantage of rapidly improving peptide design through machine learning model refinement. Peptides can be computationally represented by physicochemical and structural descriptors [77] or encoded using various rules such as binary encoding and evolution-based encoding [78]. Since peptide sequence is the source of functionality,

a machine learning-based approach can be employed to develop predictors that learn the relationships between peptide variables derived from the sequence and the desired functional property [79–81]. Peptide databases have also been made available for data-driven functional peptide design, including cell-penetration and toxicity [32, 33]. However, there is only a limited number of studies for, and no database of, melanin binding peptides. An example here is that in the two studies that reported peptide sequences that were characterized as melanin binding, phage display was used to identify 8 peptides that bind to melanin in human melanoma cells [39] and 8 peptides that bound to melanized *C. neoformans* [40]. However, in our peptide microarray, 8 of these peptides did not demonstrate detectable melanin binding, and overall, we identified 780 peptides displaying higher levels of melanin binding than any of these peptides described in the literature. Furthermore, the second peptide microarray designed using the initial machine learning model provided more potent melanin binding peptides compared to the first peptide microarray, demonstrating the rapid improvement in design by machine learning model refinement.

Multifunctional peptides with dual or triple pharmacological properties have also been integrated into drug delivery systems through conjugation to drugs or drug-loaded cargos [34,82,83]. However, it is challenging to design peptides with multiple functions contained in a single sequence. Often single function peptides are fused directly or by a linker peptide [83–85], which may increase the peptide length and reduce the desired functional properties of each component. Another approach is to optimize additional functional properties by substituting amino acids on a template peptide with a known function [35, 36], which may require extensive laboratory screening and is time-consuming. Generating multifunctional peptides with the flexibility to choose the desired functional levels is a less

explored research area [86, 87]. Here, our machine learning and model interpretation approach guided the engineering of multifunctional peptides. The peptide properties were analyzed using the shared variable set, revealing mutually important variables contributing to both melanin binding and cell-penetration, where peptides with moderate to high net charge and containing more basic amino acids tend to possess both melanin binding and cell-penetrating properties. Further, we unexpectedly observed correlation between melanin binding and cell-penetrating in cell uptake *in vitro*. Thus, the highest intracellular accumulation was achieved by increasing the amount of peptide that can access intracellular melanosomes, where the peptides can then bind to melanin and provide sustained drug release.

Many machine learning models including random forest, support vector machines, and deep learning have been developed to predict how amino acid sequence governs peptide properties [88]. Super learning is an ensemble machine learning method that takes advantage of various machine learning models. The predictive performance of a super learner ensemble is assured to be at least as accurate as the best-performing base model [89, 90]. The same model types with varying hyperparameter combinations can be included in a SL ensemble. Recently, it was described that base model hyperparameter tuning could improve overall SL model performance [91]. Based on this finding, we further developed a procedure to systematically select optimal base model composition by iteratively filtering out models that have less contributions to the SL ensemble. Indeed, we obtained better SL model performance compared to the one including all base models. In this study, we explored a wide array of possible machine learning models and identified multiple competitive models through statistical analyses. SL provided a framework to integrate these

explored models. Although the meta-learner may add a layer of complexity, it demonstrated an interpretable summary of the model importance in terms of their contributions to the final predictions. In addition, the complexity of the machine learning architecture was reduced by variable reduction of the data sets and base model filtration of SL. Further, interpretable machine learning that extracts relevant information such as variable contributions to output predictions from the data relationships learned by the model is important for explaining model predictions [92, 93]. Many of the functional peptide predictors and other drug discovery tools do not have information on how and why top candidates were identified [94–96]. In this study, we showed that interpretation of machine learning models can provide insights to improve the design of multifunctional peptides. The SHAP analysis not only indicated important variables contributing most to the model prediction, but also showed the relationships between variable values and prediction outputs.

The studies described here are not without limitations. First, while the *in vitro* ARPE19 cell assay helped validate the cell-penetrating and melanin binding performance, the methodology used here did not differentiate between peptides that were free or bound to melanin or other structures within the cell. Indeed, there was a baseline level of peptide associated with non-pigmented cells, but a substantial increase in cellular localization was observed when the cells were induced to produce melanin. Second, the traceless linker conjugation yield of the HR97-brimonidine was low and requires further optimization. The cathepsin-labile linker was chosen because cathepsins are largely located intracellularly and are present in minimal amounts in extracellular fluids such as aqueous humor [97–100]. Thus, the intracamerally delivered HR97-brimonidine would be stable until it had localized within melanin-containing cells. However, the level of brimonidine measured in rabbit

iris tissue remained high, suggesting that further optimization of the linker cleavage and brimonidine release rate may also extend the duration of the therapeutic effect. Finally, the duration of IOP lowering reported here (20 days) was sufficient to demonstrate the proof-of-principle in normotensive rabbits but would not be clinically translatable. Future work with more potent drugs may increase the duration of action.

The approach we described here to apply ensemble machine learning to peptide microarray enabled the efficient design of multi-functional peptides, which in this application enhanced the intraocular pharmacokinetics and pharmacodynamics of the ophthalmic drug brimonidine. Engineered HR97 peptide demonstrated increased cell-penetrating properties compared to known cell-penetrating peptides, such as TAT, and simultaneously possessed high melanin binding capacity and low cytotoxicity. In the current context, utilizing short peptide sequences that impart melanin binding to a drug conjugate may provide an avenue for creating safe and effective implant-free sustained intraocular drug release systems. More broadly, the approach described here can be applied to generate multifunctional peptide-drug conjugates for a variety of biomedical applications.

2.5 Methods

2.5.1 Material sources

Brimonidine was purchased from TCI America. Eumelanin from *Sepia officinalis*, 0.22 μm Millex-GV PVDF filter, ferric ammonium citrate, bovine serum albumin (BSA), Tween 20, fetal bovine serum (FBS), trifluoroacetic acid (TFA), tert-Butyl methyl ether (MTBE), thionyl chloride, Tetrabutylammonium iodide, N,N-diisopropylethylamine, hu-

man cathepsins B, K, L and S, Whatman® Anotop® 0.02 μm syringe filter and Triton X-100 were purchased from Sigma Aldrich (St. Louis, MO, USA). ARPE-19 (ATCC CRL-2302, lot No. 70013110), and DMEM:F12 medium were purchased from the American Type Culture Collection (Manassas, VA, USA). EZ-Link™ Amine-PEG$_2$-Biotin, BupH MES buffer saline pack (2-(N-morpholino)ethanesulfonic acid buffer), EDC (1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride), NHS (N-hydroxysuccinimide), Pierce™ Fluorescence Biotin Quantitation Kit, rapid equilibrium dialysis (RED) 8 K device, PrestoBlue™ HS Cell Viability Reagent, DMEM with high glucose and pyruvate, Trypsin-EDTA (0.25%) with phenol, RIPA lysis buffer, Streptavidin DyLight 680, and penicillin/streptomycin were purchased from Thermo Fisher Scientific (Waltham, MA, USA). Disposable PD-10 desalting columns were purchased from VWR. Dulbecco's Phosphate Buffered Saline (DPBS), 1 × phosphate buffered saline (PBS), 10 × PBS, high-performance liquid chromatography (HPLC) grade acetonitrile, dimethylformamide (DMF), and water were purchased from Fisher Scientific (Hampton, NH, USA). Mc-Val-Cit-PAB was purchased from Cayman Chemical (Ann Arbor, MI, USA). Endotoxin-Free Ultra-pure Water were purchased from MilliporeSigma (Burlington, MA, USA). A Hamilton 1700 Series gas tight syringes (25 μL, Model 1702 RN, 27 gauge) was purchased from Hamilton Company (Reno, NV, USA). BD 1 mL TB syringe with 28 G needles were purchased from BD (San Jose, CA, USA). Isoflurane was purchased from Baxter (Deerfield, IL, USA). Reverse-action forceps were purchased from World Precision Instruments (Sarasota, FL, USA). Neomycin, polymyxin b, and bacitracin zinc ophthalmic ointment was purchased from Akorn (Lake Forest, IL, USA).

2.5.2 Melanin nanoparticle synthesis and characterization

Melanin nanoparticles (mNPs) were synthesized from the eumelanin of *Sepia officinalis*. In brief, 10 mg/mL of eumelanin was suspended in the DPBS using an ultrasonic probe sonicator (Sonics, Vibra Cell VCX-750 with model CV334 probe, Newtown, CT, USA) by pulsing 1 s on/off at 40% amplitude for 30 min in a 4 °C water bath. The suspension was then filtered through a 0.22 μm Millex-GV PVDF filter and transferred to PD-10 desalting columns. The resulting mNPs solution was lyophilized for 7 days and stored at −20 °C until further use. For mNP biotinylation (b-mNPs), mNPs were suspended in 2 mL MES buffer with 2.4 mg of EDC and 3.6 mg of NHS for 15 min at room temperature to first activate the carboxylic acid groups. To increase the buffer pH above pH 7.4 for amine reaction, 400 μL of 10 × PBS was directly added to the mixture and incubated for 5 min. Various amounts of EZ-Link^{TM} Amine-PEG$_2$-Biotin (5, 15, 20, 30 mg) were reacted with activated mNPs for 2 or 6 h at room temperature. Since all conditions led to a similar degree of mNP biotinylation, reaction conditions using 5 mg of amine-PEG$_2$-biotin with 2 h incubation at room temperature was used moving forward. The reaction mixture was then transferred to PD-10 desalting columns to further collect the b-mNPs. To transfer the b-mNPs to different solvents (water, pH 6.5 PBS, pH 7.4 PBS) for optimization of the peptide microarray, PD10 columns were first equilibrated with buffer, and then the b-mNPs were added. Particle size and ζ-potential were determined by dynamic light scattering and laser Doppler anemometry, respectively, using a Zetasizer Nano ZS90 (Malvern Instruments). Size measurements were performed at 25 °C at a scattering angle of 173°. Samples were diluted in 10 mM NaCl solution (pH 7), and measurements were performed according

to instrument instructions. Pierce™ Fluorescence Biotin Quantitation Kits were used to quantify the biotin content on the b-mNPs. B-mNPs (1 mg/mL) were diluted 1:50, 1:100, 1:200 with 1 × PBS and the standard biocytin concentration (10–60 pmol/10 μL) were freshly prepared for measuring the biotin concentration. Transmission electron microscopy (H7600; Hitachi High Technologies America) was conducted to determine the morphology of mNPs and b-mNPs.

2.5.3 Optimization of processing conditions for peptide microarray

A total of 119 peptides, including 8 peptides of length 7 amino acids (aa) and 8 peptides of length 10 aa from the literature [39, 40], and 103 random 15 aa peptides generated with a frequency of 5% for each of the 20 amino acids, were printed in duplicate on peptide microarrays by PEPperPRINT. The peptide microarrays contained hemagglutinin (HA) peptides (YPYDVPDYAG; 9 spots) as internal quality controls. Varying screening conditions of the peptide microarray were performed. A spectrum scan of melanin nanoparticles (mNPs) and biotinylated mNPs confirmed that the autofluorescence was near background levels after Em = 650 nm. Streptavidin DyLight 680, which was the highest wavelength (Ex = 675 nm, Em = 705 nm) that PEPperPRINT could use in their peptide microarray system, was selected to minimize detection of melanin. Two peptide microarray copies were first pre-stained with streptavidin DyLight680 (0.2 μg/ml) and the control antibody (manufacturer: BioxCell & PEPperPrint, catalogue numbers: #RT0268, PEPperCHIP® Mouse Monoclonal anti-HA (12CA5)-DyLight800 Control; 1:2000 dilution or 0.5 μg/ml) in incubation buffer (pH 6.5 PBS with 0.005% Tween 20 and 10% Rockland blocking buffer

MB-070) for 45min at room temperature to examine background interactions and internal quality control. No background interaction of streptavidin DyLight680 or the control antibody with the 119 different peptides were observed. To screen the optimal melanin binding condition, six different washing buffers were prepared: PBS at pH 6.5 with or without 0.005% Tween 20, PBS at pH 7.4 with or without 0.005% Tween 20, and Ultra-pure water with or without 0.005% Tween 20. The Rockland blocking buffer MB-070 was used to incubate all peptide microarrays for 30 min before the melanin binding assay. Six different incubation buffers were formulated with 10% of blocking buffer in the six different washing buffers mentioned earlier. b-mNPs (10, 100, or 500 μg/ml) in six different incubation buffers were incubated with the peptide microarray for 16 h at 4 °C or room temperature. All microarrays were subsequently washed with the same type of washing buffers and incubated with 0.2 μg/mL of streptavidin DyLight680 for 45 min in the same type of incubation buffer at room temperature for detecting the b-mNPs. The peptide microarrays were then washed for 3 $\times$ 10 s with the same type of washing buffers and proceeded to quantification of spot intensity. The pilot tests suggested that 500 μg/ mL of biotinylated mNPs in pH 6.5 PBS buffer at room temperature was optimal (optimal condition shown in Fig. 2.1d, remaining conditions shown in Fig. A.2. With the optimal flow conditions, 10 of the 16 peptides reported in the literature had detectable fluorescence intensities due to binding by b-mNPs. Quantification of spot intensities and peptide annotation were based on the 16-bit gray scale Tag Image File Format files that exhibit a higher dynamic range than the 24-bit colorized Tag Image File Format files. Microarray image analysis was done with PepSlide® Analyzer, version 1.4. The software algorithm decomposed fluorescence intensities of each spot into raw, foreground and background signal, and calculated mean

median foreground intensities and spot-to-spot deviations of spot duplicates. Based on mean median foreground intensities, intensity maps were generated and interactions in the peptide maps highlighted by an intensity color code with red for high and white for low spot intensities. The PEPperPRINT protocol tolerated a maximum spot-to-spot deviation of 40%, otherwise the corresponding intensity value was zeroed. We labeled the top 20% of peptides ranked by intensities as melanin binding (23 peptides), which included 10 literature-reported peptides with non-zero fluorescent signal. The remaining peptides were labeled as non-melanin binding (96 peptides).

2.5.4 Random forest classification model training with the pilot 119-peptide microarray

Random forest is an ensemble tree-based statistical machine learning model and is robust to variable noise and insensitive to variable scales [38]. Physiochemical variables and numerical representations of peptides were computed using the R packages *Peptides*, version 2.4.4 [101] and *protr*, version 1.6–2 [102]. The resulting 1094 variables include composition, transition, distribution, autocorrelation, conjoint triad, quasi-sequence-order descriptors, and pseudo-amino acid and amphiphilic pseudo-amino acid composition descriptors. The maximum value of lag was set to 6, so the minimum length of a peptide to be analyzed without generating a missing value is 7. A random forest classification model with 100,000 trees and balanced sampling was trained on the melanin binding data set. The model was built using the R package *randomForest*, version 4.7–1.1 [103]. For each tree in the random forest, a bootstrap sample of ∼63.2% of the melanin binding peptides and the

same amount of non-melanin binding peptides was generated to construct the tree. The remaining peptides were considered out-of-bag to the tree and were used to evaluate the performance of the random forest by calculating the aggregated out-of-bag predictions across all trees. The out-of-bag class errors were calculated and a classification threshold of 0.5 proportion of votes was used. As part of the same analysis, permutation variable importance was obtained with the *importance* function in the *randomForest* package. For each tree in the random forest, out-of-bag instances were permuted for each variable in the subset, and the decrease in accuracy was recorded. The mean decrease in accuracy for each variable was calculated over all 100,000 trees and normalized by dividing the mean by the standard error.

2.5.5 Expansion of the peptide microarray

Melanin binding candidate peptides were generated randomly with a frequency of 5% for each of the 20 amino acids. Peptides classified as melanin binding by the trained random forest model were selected, resulting in 5483 peptides of length ranging from 7 to 12 aa. Along with the 16 known melanin binding peptides from the literature, a total of 5499 peptides were printed in duplicate along with HA controls (YPYDVPDYAG; 68 spots) on peptide microarrays by PEPperPRINT. Peptide sequences were printed in duplicate of a custom peptide microarray. Pre-staining of a peptide microarray copy was done with streptavidin DyLight680 (0.2 μg/ml) and the control antibody (mouse monoclonal anti-HA (12CA5) DyLight800; 0.5 μg/ml) in incubation buffer to characterize non-specific binding. Subsequent incubation of another peptide microarray with the b-mNPs at a concentration

of 500 μg/ml in incubation buffer (PBS at pH 6.5 with 0.005% Tween 20 with 10% Rockland blocking buffer MB-070) was followed by staining with streptavidin DyLight680 (0.5 μg/mL) and the control antibody (0.5 μg/mL). The control staining of the HA epitopes was done simultaneously as internal quality control to confirm the assay quality and the peptide microarray integrity. Quantification of spot intensities were described earlier in the previous section.

2.5.6 Variable reduction of the machine learning input data

To reduce the number of variables and improve the model performance, a variable reduction procedure was applied to the machine learning input data before model training. Permutation-based variable importance was first computed on the data set with random forest (100,000 trees) using the R package *ranger*, version 0.14.1 [104], with balanced sampling for classification analyses. Variables with negative importance values were removed. Next, subsets of the machine learning data set containing cumulative top-ranked variables were used to train random forests with 1000 trees, and the models were evaluated by the Akaike information criterion (AIC). The AIC values classification models were calculated using the original formula proposed by Akaike [105]: $\mathrm{AIC} = -2\ln(\hat{L}) + 2k$, where $\hat{L}$ is the maximum likelihood value, and k is the number of parameters. For regression AIC was calculated using the likelihood of normal distribution, assuming residuals are normally distributed: $\mathrm{AIC}_{\mathrm{reg}} = N\ln(MSE) + 2k$, where N is the number of samples, and MSE is the mean squared error. The classification AIC was based on the likelihood of Bernoulli distribution, and was generalized to multi-class classification: $\mathrm{AIC}_{\mathrm{clf}} = 2 \cdot \ln 2 \cdot N \cdot H_p(q_\theta) + 2k$,

where N is the number of samples, H_p is denotes cross entropy, and q_θ is the estimated probability with parameters θ. The variable subset with the lowest AIC value was selected for each machine learning data set.

2.5.7 Machine learning model training for melanin binding predictions

Peptide variables were computed as described for the melanin binding peptides in the pilot microarray. Because the distribution of melanin binding fluorescence intensity was right-skewed, the intensity values were first normalized by $\log_{10}$-transformation for a balanced response variable. The melanin binding data set was processed using the variable reduction method. To generate the machine learning input data set, less informative peptide variables were eliminated as described above. A nested cross-validation framework was then applied to provide an unbiased estimate of the generalization performance. The framework contains two types of cross-validations. The first includes ten sets of train-test splits computed using a Monte Carlo sampling method, which is referred to as the outer loop cross-validation. For each training set in the outer loop, another ten sets of train-test splits were generated using a modulo method. These cross-validations are referred to as the inner loop cross-validations. The inner loop cross-validations were used to select the best-performing model, and the outer loop cross-validation was used to evaluate the whole machine learning training process.

A wide array of machine learning models, including neural networks [106], gradient boosting machines (GBM) [107], extreme gradient boosting (XGBoost) [108], generalized linear model (GLM) [109], (distributed) random forests (DRF) [38], and extremely ran-

domized trees (XRT) [110], were employed to train the input data. Hyperparameters for neural networks, GBM, and XGBoost were selected using the random grid search. Details about the grids used and the hyperparameters selected can be found in Section A.1 and the provided code. There were 300 neural networks, 300 GBM models, and 300 XGBoost models trained for the melanin binding data set, along with five default GBM models, three default XGBoost models, one GLM, one DRF, and one XRT. The model types and hyperparameters were defined based on the architecture of *H2O AutoML* [111]. For non-tree-based models, variables in the training set were scaled to have zero means and unit variances. Unstable neural networks with potentially large activation values were removed.

To integrate the models explored, a super learner (SL) model was built using the R interface of *H2O.ai*, version 3.38.0.2 [112]. A generalized linear model algorithm (meta-learner) was used to calculate the coefficients (weighted contributions) of the base machine learning models according to their holdout predictions generated from the tenfold cross-validation. The meta-learner was then evaluated with another tenfold cross-validation trained on the base model holdout prediction data set. Coefficient distributions were collected from the ten cross-validation meta-learner models, resulting in a $n \times m$ matrix, where n is the number of base models, and $m = 10$ is the number of cross-validation folds. The original SL algorithm used a meta-learner to calculate base model contributions and did not emphasize explicit base model selection. To reduce the complexity of SL, we developed an iterative filtering procedure to improve performance and decrease prediction run time. Specifically, base models with the number of zero coefficients >5 across cross-validation folds were removed. The filtering procedure was repeated until there were no base models or no further reduction of the base models. In addition, SL models with different compo-

sitions of base models were also constructed for comparison. Homogeneous SL ensembles were constructed with base models of the same model type (neural networks, GBM, XGBoost).

Regression models trained on the melanin binding data were evaluated in each inner loop cross-validation using multiple metrics, including coefficient of determination (R^2), percent normalized mean absolute error (MAE, less sensitive to outliers), and percent normalized root mean squared error (RMSE). A scoring scheme that calculates the sum of ranks of all metrics used was applied, and non-parametric Mann–Whitney U tests comparing the top model and the rest of the models were conducted to identify competitive models, with p values adjusted using the Benjamini–Hochberg procedure [113]. Evaluation results regarding the competitive models whose performances were not significantly different from the top model for all evaluation metrics were reported. Next, the top model was selected from each inner loop cross-validation and evaluated using the corresponding test sets in the outer loop cross-validation, and the generalization performance was computed (Table A.1).

Finally, the abovementioned model training procedure was performed on the whole data set, and the final predictive model was selected based on the same scoring scheme of the sum of all metric ranks.

2.5.8 Machine learning model training for cell-penetration predictions

Cell-penetrating and non-cell-penetrating peptides of various lengths (10–61 amino acids) were collected from the *SkipCPP-Pred* website [31], for which the redundant cell-

penetrating peptides from the *CPPsite2.0* database [32] have been removed, and non-cell-penetrating peptides were generated randomly [31]. There were 460 cell-penetrating and 462 non-cell-penetrating peptides. Peptide variables were computed as described above for classification of the melanin binding peptides from the pilot microarray. The variable reduction procedure as described above was then applied to the data set. A nested cross-validation framework was employed to generate train-test splits for outer and inner loop cross-validations. Multiple machine learning models were trained on the cell-penetration data set, including 100 neural network grid models, 100 GBM grid models, 100 grid XGBoost models, five default GBM models, three default XGBoost models, one DRF, and one XRT. Models were integrated using the SL framework, resulting in SL models separately containing all base machine learning models, reduced base models, all neural networks, all GBM models, and all XGBoost models. Balanced sampling was applied where appropriate for the machine learning algorithms.

Classification models were evaluated with logarithmic loss, Matthews correlation coefficient (MCC), F_1 (harmonic mean of precision and recall) and balanced accuracy. A scoring scheme computing the sums of all metric ranks was applied. Competitive models with no significant difference from the top model in terms of model performance, along with the means and standard errors of metrics obtained from 10-fold cross-validations were reported. The top model from each inner loop cross-validation was selected. The generalization performance (Table A.2) was evaluated in the outer loop cross-validation, using logarithmic loss, MCC, F_1, balanced accuracy, enrichment factor (EF), and Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) [114].

The final predictive model was generated by applying the same model training pro-

cedure on the whole data set. Class prediction thresholds for the final model were selected based on the maximum F_1.

2.5.9 Machine learning model training for cytotoxicity predictions

Toxic and non-toxic peptides of various lengths (4–35 aa) were collected from the *ToxinPred* website [33]. Peptides with length <7 were excluded, resulting in 1777 toxic and 3522 non-toxic peptides. Peptide variables were calculated as described in the random forest classification section, and non-toxic and toxic peptides were labeled as positive and negative, respectively. The dimensionality of the data set was reduced using the variable reduction as described in the above section. A nested cross-validation framework was applied, and the machine learning models include 100 neural networks, 100 GBM models, 100 XGBoost models, five default GBM models, three default XGBoost models, one GLM, one DRF, and one XRT. The mean number of peptides in non-toxic and toxic classes was calculated and used as the number of samples of each class for balanced sampling. Models were integrated using the SL framework, generating SL models containing all base models, reduced base models, all neural network models, all GBM models, and all XGBoost models. The models selected as the top model in each inner loop cross-validation were selected using the evaluation metrics and the scoring scheme as described for cell-penetration model training. The generalization performance was computed based on the selected models and recorded in Table A.3. The final predictive model was generated by performing the above model training procedure on the whole data set, and the class prediction threshold was determined by the maximum F_1 score.

2.5.10 Peptide generation for machine learning model validation

Amino acid frequencies at each position were calculated for the 5499 melanin binding peptides used in the expanded peptide microarray, where the peptides were grouped into 8 sets based on intensity ranges. For each intensity group, random peptides were simulated based on the position-dependent amino acid frequency. In total, 127 peptides of length ranging from 7 to 12 were selected, including the TAT_{47-57} peptide as the reference cell-penetrating peptide and 7 peptides from the expanded peptide microarray as validation controls. Melanin binding intensity values were predicted by the reduced melanin binding SL model. Selected peptide sequences were subsequently analyzed by the cell-penetration and toxicity final models for further classification.

2.5.11 Peptide synthesis

The library of 127 C-terminal biotinylated peptides used in cell culture experiments was synthesized by Gene Script using their Crude Peptide Library service. A terminal lysine was added to each peptide sequence to facilitate biotin conjugation. Peptides from the crude peptide library were further purified by being first dissolved in 50% acetonitrile (ACN) with 0.1% TFA at 10 mg/mL. Shimadzu LC20 high-performance liquid chromatography (HPLC) system with Phenomenex reverse-phase preparative HPLC column (Gemini® 10 μm C18 110 Å, LC Column 250 $\times$ 21.2 mm, AXIA™ Packed) were used to separate and collect the peptides with an elution gradient of 5/5/90/90/5/5% solvent B (TFA 0.05% in ACN) at 0/2/10/12/13.5/15 min with a flow rate of 5 mL/min with monitoring at 220 nm.

2.5.12 Melanin binding assay for machine learning model validation

The mNPs were mixed with C-terminal biotinylated peptides (10 μM) in pH 6.5 PBS solution and incubated in the rapid equilibrium dialysis (RED) 8 K device for 24 h on an orbital shaker at 900 rpm. A total of 10 μL of the solution from the rapid dialysis reservoir was collected. The concentration of unbound biotinylated peptides was analyzed with the Pierce^{TM} Fluorescence Biotin Quantitation Kit. Four sets of melanin binding assays were performed. Melanin binding was calculated as the difference in free peptide normalized with the starting peptide concentration. Experimental melanin binding values of the 127 peptide candidates were compared with the predicted melanin binding values with the Pearson correlation. Melanin binding predictions larger than 100% were cast to 100% because this was the maximum value in the melanin binding training set.

2.5.13 Cell-penetration assay with ARPE19 cell type for machine learning model validation

Three 96 well plates per ARPE19 cell type group (melanin-induced or non-melanin induced) were seeded at 0.01×10^6 cells/well. ARPE-19 cells were either cultured with DMEM:F12 medium containing 10% FBS according to protocol provided by the vendor (non-melanin induced) or cultured in DMEM high glucose, pyruvate media with 250 μM of ferric ammonium citrate [115] for 2 months (melanin-induced) [28]. The expression of melanin was confirmed visually with bright field microscopy and by measuring absorbance at 475 nm (>0.4 arb. units) [28]. Within each plate, 12 wells were randomly selected to

quantify the cell numbers with an automated cell counter (Countess 3 Automated Cell Counter, Thermo Fisher) for normalization in the cell uptake study. Next, 100 μL (100 μM in pH 6.5 PBS) of each of the 127 C-terminal biotinylated peptides was added to $n = 3$ wells for both the induced and non-induced ARPE-19 cells for 6 h at 37 °C. The cells were then washed thoroughly five times with PBS solution to remove extracellular peptide. To quantify cell-associated peptides, the cells were lysed with 100 μL of RIPA lysis buffer at 4 °C for 48 h. The concentration of intracellular biotinylated peptides was analyzed with the Pierce^{TM} Fluorescence Biotin Quantitation Kit. The mean intracellular concentration values of the three replicates were then grouped by cell types (melanin-induced or non-melanin induced), and a two-tailed Mann–Whitney U (Wilcoxon rank-sum) test was calculated using the *wilcox.test* function in R. The intracellular concentration values were also plotted against experimental melanin binding. The relationships between experimental cell-penetration and melanin binding values in the two ARPE19 cell type groups were quantified using the Pearson correlation.

2.5.14 Shapley additive explanations (SHAP) analysis of variable contributions

To better characterize variable contributions to peptide property predictions, models trained on the outer loop training sets with the same hyperparameters as the final predictive model were used to calculate SHAP values using the corresponding test sets. For each sample in the test set, the SHAP analysis calculated the additive variable attributions to the model prediction. Specifically, models were imported using the Python interface of *H2O.ai*,

version 3.38.0.2 [116], and the background data set was generated by randomly selecting 100 samples from the training set. Next, SHAP values, with the number of sampling times set as 1000, were computed using the function *KernelExplainer* in the Python package *SHAP*, version 0.41.0 [37]. The *KernalSHAP* method calculates variable contributions (SHAP values) using a local interpretable model-agnostic explanations (LIME) strategy [117]. The top 20 variables ranked by the difference between the maximum and minimum SHAP values in the aggregated test set samples were selected and visualized along with the variable values normalized by percentile ranks.

Explanations of HR97 multifunctional peptide predictions were computed using the final models trained on the whole machine learning data sets. The same SHAP analysis method as described above was performed, and the top variables ranked by absolute SHAP values were visualized as waterfall plots using the function *plots.waterfall* in the *SHAP* package.

2.5.15 Adversarial computational controls

To assess if the model performance evaluation was overly optimistic, and if the machine learning models have learned the meaningful relationships in the data sets, adversarial controls were generated by training the models on the data sets with the response variables randomly shuffled [118]. The same nested cross-validation framework and model selection procedure as described in the above model training sections were used, and the generalization performance was computed with the models selected as the top model from the inner loop cross-validations. Statistical results of the competitive models from the inner loop

cross-validations (Section A.1), and the generalization performance of the adversarial controls evaluated in the outer loop cross-validation (Tables A.1–A.3) were reported. Variable contributions of the adversarial control models having the same hyperparameters as the final predictive model of each property were computed and visualized as described in the SHAP analysis section.

2.5.16 Peptide design space visualization

Peptide sequences of the control melanin binding peptides, pilot 119 peptides, expanded 5499 peptides, and 127 peptide candidates for experimental validation were converted using one-hot encoding, and the post-padding was applied for peptides with shorter lengths. Combined with the union set of variables from the variable-reduction processed melanin binding, cell-penetration, and cytotoxicity data sets, the new data set was normalized and analyzed using t-Distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear dimensionality reduction technique, with the *Rtsne* function in the R package *Rtsne*, version 0.16 [119]. The t-SNE results were visualized along with the multifunctional predictions.

2.5.17 Traceless linker system for conjugating HR97 to brimonidine

The traceless linker system was designed for release of intact parent drug when triggered by an intracellular chemical and enzymatic event, such as protease cleavage of the amide bond [120]. Activation of the linker, MC-Val-Cit-PAB-OH (Maleimidocaproyl-L-valine-L-citrulline-p-aminobenzyl alcohol), was conducted as previously reported with minor modifications [120]. MC-Val-Cit-PAB-OH (8.68 g, 15.2 mmol) was suspended in DMF

(43.4 mL) at 0 °C with water bath sonication for 30 min. After the solids were fully dispersed, thionyl chloride (1.22 mL, 16.7 mmol) was added dropwise. Following the addition, the reaction was held at 0 °C for 45 min and then treated slowly with water (130 mL) to precipitate a yellow solid (MC-Val-Cit-PAB-Cl), which was collected by filtration. The solid was washed sequentially with water and MTBE and dried under vacuum (∼30% yield) [120]. Brimonidine base was combined with the MC-Val-Cit-PAB-Cl (1.1 eq) in DMF (0.25 M) at room temperature. Tetrabutylammonium iodide (0.5 eq) was added to the solution, followed by the addition of N,N-diisopropylethylamine (2.5 eq), and the mixture was stirred for 24 h. The mixture was diluted with 50:50 acetonitrile:water at 40-fold dilution for purifying the MC-Val-Cit-PAB-brimonidine. A Shimadzu LC20 HPLC system coupled with photodiode-array detector (PDA) and with Phenomenex reverse-phase preparative HPLC column (Gemini® 10 μm C18 110 Å, LC Column 250 × 21.2 mm, AXIA™ Packed) was used to separate and collect the conjugates with an elution gradient of 10/90/90/10% solvent B (TFA 0.05% in ACN) at 1/11/13/15 min with a flow rate of 10 mL/min. The collected fractions were then transferred to the 20 mL scintillation vials and a Biotage V-10 solvent evaporator with Volatile mode was used to remove the acetonitrile. The solution fractions were frozen and lyophilized (∼8% yield). NMR was used to confirm the presence of key functional groups in the products of each stage of the synthesis, including brimonidine, Mc-VC-PAB-Cl, and Mc-VC-PAB-brimonidine. All compounds were dissolved in deuterated DMSO and characterized with a Bruker spectrometer (500 MHz). ^{1}H chemical shifts were reported in ppm (δ) and the DMSO peak was used as an internal standard. Data were processed using TopSpin NMR Data Analysis software, version 4.1.0, from Bruker (Billerica, MA, USA). The prep-HPLC retention time (RT) of brimonidine, Mc-VC-PAB-brimonidine,

and Mc-VC-PAB-Cl was 5.1, 9.8, and 11.4 min, respectively. HR97 with cysteine at the C-terminus as the functional group for linker conjugation (FSGKRRKRKPRC, $Mw = 1519$, >97% purity) was conjugated to the quaternary-ammonium-linked brimonidine (MC-Val-Cit-PAB-brimonidine) via a thiol-maleimide reaction. The MC-Val-Cit-PAB-brimonidine was first dissolved in 1 mL of PBS at 5 mg/mL. HR97 peptide powder (0.5 eq) was added directly to the solution. The solution mixtures were adjusted to pH 7.4 and allowed to react for 2 h at room temperature. The solution mixtures were then added to 1 mL of acetonitrile and purified with the same prep-HPLC conditions. The collected fraction solutions were transferred to the 20 mL scintillation vials and the Biotage V-10 solvent evaporator with volatile mode were used to remove the acetonitrile. The solutions were lyophilized and stored at −20 °C (∼35% yield). For the sample preparation and MALDI-TOF analysis, the MALDI matrix sinapic acid (10 mg) was dissolved in 1 mL of acetonitrile in water (1:1) with 0.1% TFA, and 1 μL of sample (50 μM) was deposited on the MALDI sample plate. The matrix (2 μL, 10 mg/mL) was deposited on the air-dried sample and allowed to air dry for 10–20 min. The MALDI-TOF MS analysis was performed on a Bruker Voyager DE-STR MALDI-TOF (Mass Spectrometric and Proteomics core, Johns Hopkins University, School of Medicine) operated in linear, reflective-positive ion mode.

2.5.18 *In vitro* melanin binding assay

Brimonidine, HR97-biotin, and HR97-brimonidine at a range of concentrations (3.125, 6.25, 12.5, 25, 50, 100 μg/mL) were dissolved in pH 6.5 PBS solution. The solutions (400 μL) were then mixed thoroughly with 400 μL of 1 mg/mL mNPs in pH 6.5 PBS solution

and transferred to the inner reservoir of the rapid equilibrium dialysis (RED) device inserts (8 K MWCO). The outer reservoir was filled with 800 μL of pH 6.5 PBS solution. The samples were incubated on an orbital shaker with temperature control at 37 °C and 300 rpm for 48 h ($n = 3$). The solutions from outer reservoir (free drug) were than collected and transferred to an autosampler vial for HPLC analysis (Prominence LC2030, Shimadzu, Columbia, MD) with photodiode-array detection (PDA) system. Separation was achieved with a Luna® 5 μm C18(2) 100 Å LC column 250 × 4.6 mm (Phenomenex, Torrance, CA) at 40 °C using isocratic flow. The amount of bound drug was used to calculate the binding capacity (mol drug/mg melanin) and the dissociation constant (K_d) as previously described [28, 42].

2.5.19 *In vitro* stability test for HR97-brimonidine conjugate

Two pairs of human donor eyes were obtained from the Lions Gift of Sight under protocol IRB00056984 approved by the Johns Hopkins University School of Medicine Institutional Review Board. Both donors were male with an mean age of 74.5. The post-mortem times ranged from 35–40 h. The eyes were kept at 4 °C during transport and arrived within 48 h post-mortem. The vitreous and aqueous were first isolated and subsequently combined and filtered through the 0.02 μm syringe filter to remove cell debris. HR97-brimonidine (1 mg/mL) was incubated with human aqueous or vitreous (700 μL) at 37 °C ($n = 3$). On days 0, 1, 7, 14, 21 and 28, 100 μL of the solutions were collected, diluted with 900 μL of acetonitrile, and characterized by HPLC (Prominence LC2030, Shimadzu) with Luna® 5 μm C18(2) 100 Å LC column 250 × 4.6 mm (Phenomenex). The elution flow rate was 1

mL/min and with gradient of 10/90/90/10% solvent B (TFA 0.1% in ACN) in 1/11/13/15 min at λ_{max} = 250 nm for HR97-brimonidine (RT = 4.6 min). The area under the curve (AUC) on day 0 was used to normalize the AUC calculated on days 1, 7, 14, 21 and 28.

2.5.20 Cathepsin cleavage assay for HR97 and HR97-brimonidine conjugate

An assay to demonstrate enzymatic cleavage of the linker was used as previously described with adaptations [120]. In brief, the HR97-brimonidine conjugate solution (200 μM) was diluted with an equal volume of 100 mM citrate buffer at pH 5.5. Cysteine was added to a final concentration of 5 mM before the addition of human cathepsins B, K, L, and S to final concentrations of 150 nM each. The mixture was then incubated for 0 h (control group) or 48 h at 37 °C. The solutions were further diluted with acetonitrile to 1mL and conjugate concentration was measured using the HPLC method described above. All concentration values are normalized to the HR97-brimonidine at 0 h.

2.5.21 Cell viability assay of HR97 peptide

The PrestoBlue^{TM} HS cell viability system was used to assess cell viability. ARPE-19 cells were seeded at 0.01 $\times$ 10^6 cells/well in 96 well plates and cultured with DMEM:F12 medium containing 10% FBS according to the vendor protocol. After 7 days, 90 μL of DMEM/F12 containing 0, 1, 5, 10, or 20 mg/mL of the HR97 was added. The cells ($n = 5$) were then incubated for 12 h, and viability was measured by adding 10 μL of PrestoBlue^{TM} HS cell viability reagents at 37 °C. After 0.5, 1, 2, 3, 4 and 5 h, absorbance (570 nm and

600 nm) was measured at 37 °C and normalized according to the protocol provided by the vendor.

2.5.22 Animal studies—Animal welfare statement

Experimental animal protocol (RB21M176) was approved by the Johns Hopkins Animal Care and Use Committee. All animals were handled and treated in accordance with the Association for Research in Vision and Ophthalmology Statement for Use of Animals in Ophthalmic and Vision Research. Dutch Belted rabbits (4–5 mo) were obtained from Robinson Services, Inc. Rabbit sex was uniformly distributed and randomly assigned to each group, which consisted with either 3 male/2 female, or 2 male/3 female for IOP/safety studies and 2 male/2 female for the pharmacokinetic study.

2.5.23 Rabbit IOP measurements, topical dosing, and ICM injection

For the IOP measurements in normotensive rabbits, Dutch Belted rabbits (2–3 kg) were used ($n = 5$). IOP was measured with a hand-held rebound tonometer icareTONOVET (Vantaa, Finland) in the awake and gently restrained rabbit. Each rabbit was acclimatized to the IOP measurement procedure for at least 5 days to obtain a stable background IOP reading. A mean of three IOP measurements for an individual eye were taken every other day for 6 days (3 times in total) and used as a baseline value. For the ICM injection procedure, rabbits were anesthetized with ketamine/xylazine and received topical anesthesia with 0.5% proparacaine hydrochloride. A corneal pre-puncture was performed with a 30 G needle, followed with a single bolus ICM injection of 200 μg (mass of brimonidine) of HR97-

brimonidine or brimonidine tartrate solution in 100 μL saline using a 28 G needle. After the procedure, topical bacitracin-neomycin-polymyxin ophthalmic ointment was applied to both eyes to prevent infection and dry eyes. On day 7, an ophthalmologist masked to treatment evaluate the HR97-brimonidine injected eyes with the following items: functionality of lids, lashes, conjunctiva, cornea transparency, pigmentation of corneal endothelium, depth of anterior chambers, inflammation, fibrin strands, and symmetry of the lens [121]. The lenses were all clear and the iris pigmentation was symmetric. In a separate study, a corneal pre-puncture was performed with a 30 G needle, followed with a single bolus ICM injection of 200 μg (mass of brimonidine) containing a physical mixture of unconjugated HR97 and brimonidine tartrate (HR97 + brimonidine), the equivalent amount of HR97 peptide alone, or saline alone in 100 μL saline. On day 7, day 14, day 21, and day 28, an ophthalmologist masked to treatment performed the same safety evaluations described above. IOP was measured on days 2, 3, 4, 5, 6, 7, 8, 10, 12, 14, 16, 18, and 20 after the ICM injection, and change in IOP from the baseline (ΔIOP) was reported. The mean of three IOP measurements was taken for each eye by one observer, and then confirmed by a masked observer. Alternatively, a single topical eye drop (Alphagan® P 0.1%, 50 μL) was given ($n = 5$). The IOP were measured immediately before the topical dosing (0 h), and at 2, 4, 6 and 8 h after the eyedrop administration. For the pharmacokinetics studies, rabbits ($n = 4$ per group) received a single ICM injection with 200 μg (mass of brimonidine) HR97-brimonidine as described above. Rabbits were sacrificed 1, 7, 14, 28 days after the injection, and iris, aqueous, and retina were collected for measuring the brimonidine concentration. One of the iris tissue samples was left out of the analysis in the day 1 group due to an issue with sample collection.

2.5.24 Measurement of brimonidine in ocular tissues

Brimonidine concentrations in ocular tissues were measured by liquid chromatography-tandem mass spectrometry (LC-MS/MS) as previously described [46]. All samples were collected in pre-weighed tubes and stored at −80 °C until processing for analysis. Tissue samples were homogenized in 100–600 μL 1 × PBS using a Bullet Blender® (Next Advance, Inc, Troy, NY, USA) before extraction. Brimonidine were extracted from 15 to 50 μL of tissue homogenates with 50 μL of acetonitrile containing 50/50/2.5 ng/mL of the internal standards. The top layer was then transferred to an autosampler vial for LC-MS/MS analysis after centrifugation. All ocular tissue samples were analyzed using a 1 × PBS standard curve for brimonidine. Separation was achieved with a Waters HSS PFP (2.1 × 50 mm, 1.8 μm). The column effluent was monitored using a 5500 mass-spectrometric detector (Sciex) using electrospray ionization operating in positive mode. The mobile phase A was water containing 0.1% formic acid and mobile phase B was acetonitrile containing 0.1% formic acid. The gradient started with mobile phase B held at 20% for 0.5 min and increased to 100% over 0.5 min; 100% mobile phase B was held for 1 min and then returned to 20% mobile phase B and allowed to equilibrate for 1 min. Total run time was 3 min with a flow rate of 0.5 mL/min. The spectrometer was programmed to monitor the following multiple reaction monitoring (MRM) transition 391.9 → 295.9 for brimonidine and 295.9 → 216.1 for the internal standard, brimonidine-d4. Calibration curve for brimonidine was computed using the area ratio peak of the analysis to the internal standard by using a quadratic equation with a x-2 weighting function over the range of 0.25–500, with dilutions of up to 1:100 (v:v). Core technicians performing sample and data analysis were masked

to treatment group.

2.5.25 Statistical analysis

Statistical analyses of two groups were conducted using two-tailed parametric (Student's t test) or non-parametric (Mann–Whitney U) tests as appropriate. Correlation coefficients were computed using Pearson correlation (two-tailed). For multiple statistical testing, p values were adjusted using the Benjamini–Hochberg procedure [113]. Statistical analyses were performed using GraphPad Prism 9 or R version 4.2.2 (2022-10-31).

Chapter 3: Engineered Peptide-Drug Conjugate Provides Sustained Protection of Retinal Ganglion Cells with Topical Administration in Rats

3.1 Abstract

Effective eye drop delivery systems for treating diseases of the posterior segment have yet to be clinically validated. Further, adherence to eye drop regimens is often problematic due to the difficulty and inconvenience of repetitive dosing. Here, we describe a strategy for topically dosing a peptide-drug conjugate to achieve effective and sustained therapeutic sunitinib concentrations to protect retinal ganglion cells (RGCs) in a rat model of optic nerve injury. We combined two promising delivery technologies, namely, a hypotonic gel-forming eye drop delivery system, and an engineered melanin binding and cell-penetrating peptide that sustains intraocular drug residence time. We found that once daily topical dosing of HR97-SunitiGel provided up to 2 weeks of neuroprotection after the last dose, effectively doubling the therapeutic window observed with SunitiGel. For chronic ocular diseases affecting the posterior segment, the convenience of an eye drop combined with intermittent dosing frequency could result in greater patient adherence, and thus, improved disease management.

3.2 Introduction

Achieving drug delivery to the retina with topical eye drops is a "holy grail" of ocular drug delivery. Unfortunately, precorneal clearance, ocular tissue barriers such as the cornea and sclera, and uveal clearance reduce the amount and duration of drug delivery to posterior segment cells and tissues. Thus, there are no FDA approved eyedrops for treating diseases of the posterior segment, and patient adherence to many eye drops is limited by the necessity of dosing multiple times per day [122, 123]. To overcome these challenges, a drug delivery system that provides longer corneal residence time and increased intraocular absorption could facilitate increased drug delivery to the retina. In addition, increasing intraocular drug retention time could further improve drug accumulation and sustain the duration within the therapeutic window.

We recently described the development of a hypotonic, thermo-sensitive gel-forming eye drop that provided increased and sustained intraocular drug absorption without increased systemic drug exposure [124]. In addition, we described that sunitinib, a dual leucine zipper kinase (DLK) and leucine zipper kinase (LZK) inhibitor that promotes retinal ganglion cell (RGC) survival, could be effectively delivered to the posterior segment using the gel-forming eye drop [125–127]. The combination of enhanced intraocular absorption provided by the gel-forming eye drop (SunitiGel) and the intrinsic melanin binding properties of sunitinib led to significant protection of RGCs with only once weekly eye drop dosing [127]. Studies have shown that ocular melanin could affect drug distribution and retention in ocular pigmented cells [128], and prolong the effects of certain drugs [129]. Machine learning models have also been developed to predict the melanin binding potential

of small molecule ophthalmic drugs [130]. We recently described an approach for applying machine learning to engineer multifunctional peptide adaptors to embody high melanin binding, cell-penetration, and low cytotoxicity in the same peptide sequence [131]. The highest performing peptide, HR97, was conjugated to the intraocular pressure (IOP) lowering drug brimonidine tartrate, to impart these properties to the peptide-drug conjugate. We demonstrated that a single intracameral injection of the HR97-brimonidine conjugate provided a larger IOP reduction in normotensive rabbits that lasted for up to 18 days, which was significantly longer than injection of brimonidine alone (7 days), or topical brimonidine tartrate eye drops (Alphagan®, 8 h). Accumulation of the HR97-brimonidine conjugate in the melanin containing cells of the iris likely contributed to the magnitude and duration of the therapeutic effect in the anterior segment.

In this work, we hypothesized that conjugation of the engineered multifunctional peptide adaptors to sunitinib for delivery to the posterior segment using the gel-forming eye drop would provide even more prolong therapeutic effects in the posterior tissues. We observed that the HR97-sunitinib conjugate had increased binding capacity to ocular melanin and was cleaved by proteases to release free sunitinib *in vitro*. Rats were dosed topically with HR97-SunitiGel once daily for seven days, followed by optic nerve head crush at various times after the last dose to assess the duration of RGC protection. We observed that the HR97-SunitiGel showed prolonged neuroprotective effects for up to 2 weeks after the last topical dose, whereas the protective effect of SunitiGel was only observed at 1 week after the last dose. Our observations support the potential for improving and prolonging therapeutic delivery to the posterior segment tissues by addressing multiple barriers to drug delivery and retention in the eye.

3.3 Results

3.3.1 Conjugation of HR97 peptide to sunitinib increases melanin binding *in vitro*

We developed a scheme for conjugating HR97 to sunitinib via a quaternary-ammonium traceless linker system and structurally confirmed each intermediate product by NMR, HPLC, and MALTI-TOF (Figs. B.1–B.5). When incubated in human vitreous and aqueous fluids *ex vivo* for 28 days, only ∼15% (Fig. 3.1a) and ∼5% (Fig. 3.1b) of the sunitinib were released. In contrast, upon incubation with supraphysiological concentrations of human cathepsins to enzymatically cleave the linker, ∼72% of the sunitinib was released within 48 h (Fig. 3.1c, Fig. B.6). The conjugation of HR97 to sunitinib increased the solubility to 56-fold compared to sunitinib base and 5.5-fold compared to sunitinib malate (Fig. 3.1d). Although sunitinib already shows relatively high intrinsic melanin binding properties compared to other ophthalmic drugs, conjugation to HR97 provided a 1.3-fold increase in melanin binding capacity compared to sunitinib *in vivo* (K_d 7.30×10^{-6} M vs. 5.51×10^{-6} M) (Fig. 3.2a). Additionally, HR97-sunitinib provided a 2.2-fold increase in cell uptake compared to sunitinib in non-induced ARPE-19 cells (Fig. 3.2b). When incubated with ARPE-19 cells induced to produce melanin, HR97-sunitinib provided a 1.4-fold increase in cell uptake compared to sunitinib (Fig. 3.2b).

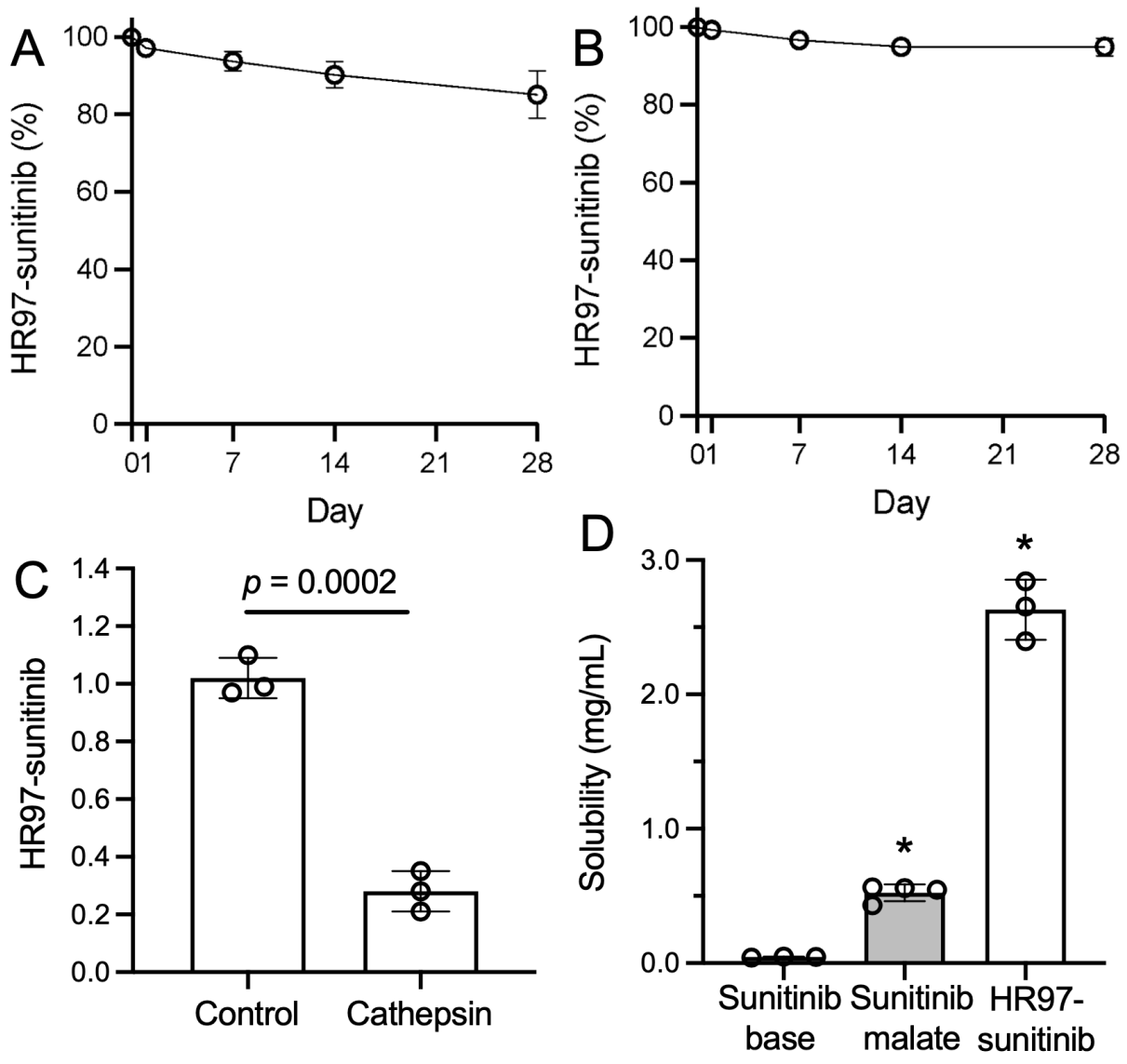


Figure 3.1 Characterization of HR97-sunitinib stability and solubility. *In vitro* stability of HR97-sunitinib conjugate in **(a)** human vitreous and **(b)** human aqueous humor for 28 days. The amount of HR97-sunitinib remaining was normalized to the starting concentration ($n = 3$). Data are presented as mean ± SD. **c** Cathepsin cleavage assay of the HR97-sunitinib conjugate. HR97-sunitinib was incubated with human cathepsin cocktails (Cathepsin) or buffer only (Control) for 48 h at 37 °C ($n = 3$). The amount of HR97-sunitinib remaining was normalized to the starting concentration ($n = 3$). Data are presented as mean ± SD, p-value = 0.0002. **d** Conjugation to HR97 increased the intrinsic solubility of sunitinib compared to sunitinib free base and sunitinib malate salt ($n = 3$). Data are presented as mean ± SD, $^*p < 0.01$ compared to sunitinib base.

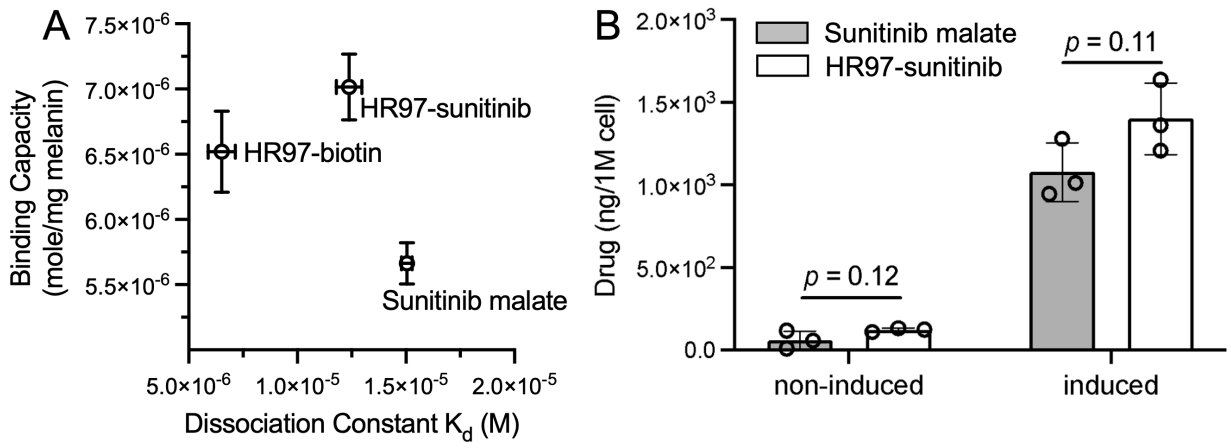


Figure 3.2 Characterization of HR97-sunitinib melanin binding and cell uptake *in vitro*. **a** *In vitro* melanin binding capacity and dissociation constant of HR97-biotin [131], HR97-sunitinib, and sunitinib malate ($n = 3$). Lower dissociation constant indicates stronger binding. Data are precented as mean ± SD. **b** ARPE-19 cells were cultured under normal conditions (ARPE) or under conditions that induce melanin production (induced ARPE) and incubated with sunitinib malate or HR97-sunitinib for 6 h. The cells were then collected and washed prior to extracting sunitinib. Drug content was normalized to per 1 million cells. Data are shown as mean ± SD, $n = 3$.

3.3.2 A deep learning object detection model was more accurate in counting RGCs

A key aspect of assessing neuroprotective capacity involves counting RGCs in different regions of flat-mounted retina tissues. Manual cell counting can be time-consuming, so we sought to develop a reliable, automated image analysis method. We used RGC images to train SSD- MobileNet (Fig. B.7) and Faster R-CNN with Inception Resnet v2 (Fig. B.8) models, both of which are often used in the object detection research [132,133]. The Faster R-CNN with Inception Resnet v2 performed well in both high (>60) and low (<20) cell density image conditions ($r = 0.993$), whereas the SSD-MobileNet slightly over-performed when the RGCs density was high in the images (Fig. 3.3a, b). The two object detection

models were then compared to CellProfiler Analyst, a well-established open-source program for cell classification and recognition. The prediction results generated by CellProfiler Analyst were more vulnerable to the quality of the images, with lower prediction accuracies for both high and low cell density images compared to the deep learning object detection models ($r = 0.947$) (Fig. 3.3c). We further demonstrated the capability of Faster R-CNN inception Resnet v2 in identifying the RGCs in various RGCs image conditions, such as high and low cell density, oversaturated, and dim image settings (Fig. 3.3d–i). We used the trained Faster R-CNN with Inception Resnet v2 model (hereafter referred to as the cell counting program) to assess retinal images collected from a time-course study of the rat ONH crush animal model to identify the optimal screening window for a neuroprotection drug delivery study (Fig. B.9a). The quantification results using the cell counting program showed that the number of surviving RGCs decreased most rapidly between days 4 and 11 after the optic nerve head crush, and the curve started to flatten 11 days after the procedure (Fig. B.9b). Thus, a period of 7 days after the crush procedure was selected as the timeframe to assess RGC protection.

3.3.3 HR97-SunitiGel showed prolonged neuroprotective effects compared to SunitiGel

We next tested the potential duration of neuroprotection after topical dosing of HR97-SunitiGel. Brown Norway rats were dosed with HR97-SunitiGel or SunitiGel daily for 7 days, the optic nerve head crush procedure was performed on day 0, 7, or 21 after the last topical dose, and the RGC survival was characterized 7 days after the injury (Fig. 3.4a).

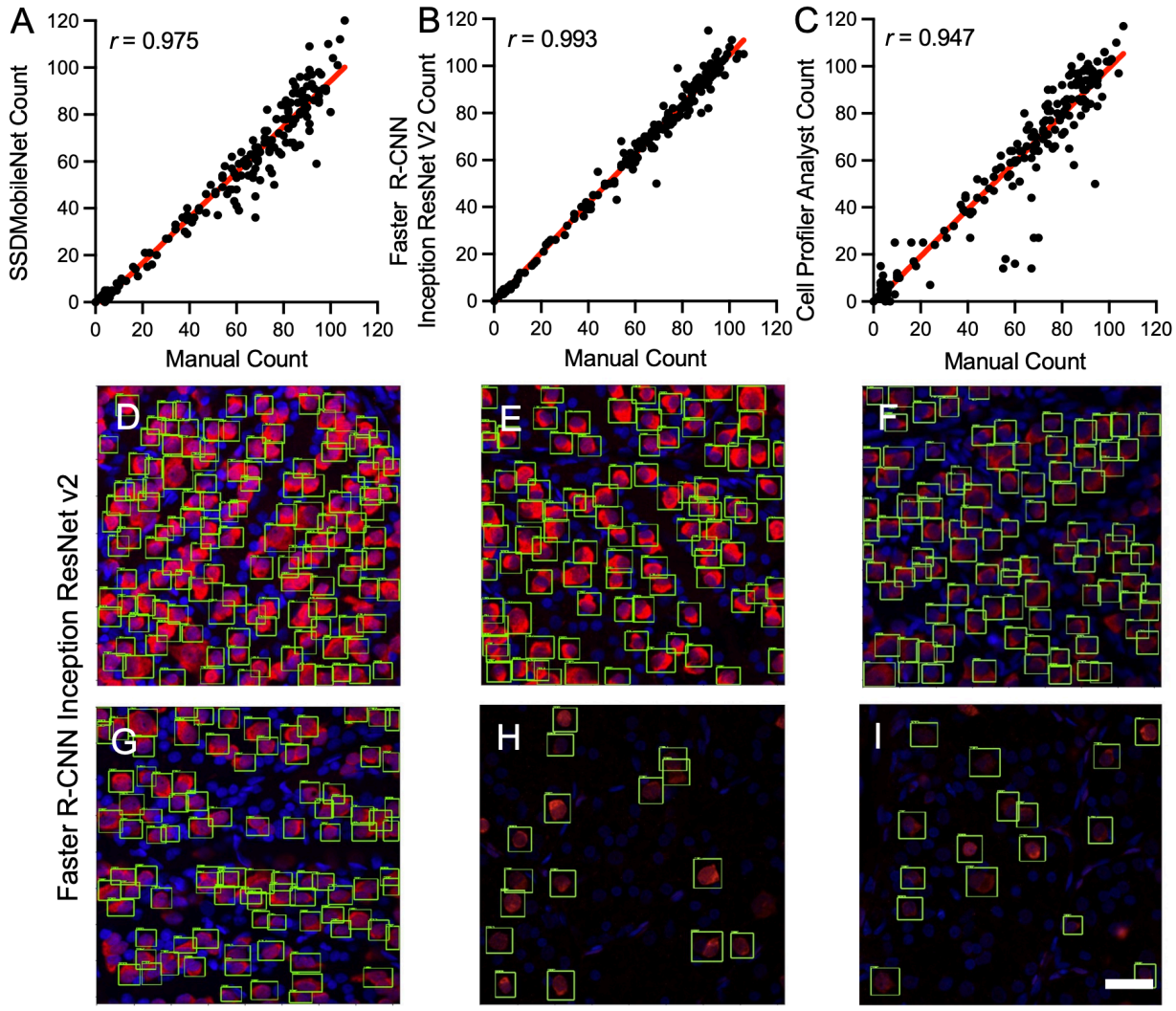


Figure 3.3 Comparison between SSD-MobileNet, Faster R-CNN Inception ResNet v2, and CellProfiler software. A total of 173 images with 4247 manual labeled cells from both healthy and ONH crushed retinas were used to train the SSD-MobileNet and Faster R-CNN Inception ResNet v2 models. The same image sets were used as inputs to CellProfiler for generating the features. Simple linear correlations between automated and manual quantification of 200 RGCs 40X images were calculated and the Pearson correlation coefficient (r) are noted. **a** SSD-MobileNet (epoch 51,350), $r = 0.975$, which is significant with the p-value threshold of 0.0001 (two-tailed). **b** Faster R-CNN Inception ResNet v2 (epoch 13,683), $r = 0.993$, which is significant with the p-value threshold of 0.0001 (two-tailed). **c** The random forest classifier was used in CellProfiler Analyst, $r = 0.947$, which is significant with the p-value threshold of 0.0001 (two-tailed). Among the three different platforms, the Faster R-CNN model more accurately quantified cells in images with **(d)** oversaturated and crowded RGCs; **(e)** mid-density RGCs; **(f)** low brightness RGCs; **(g)** mid-density, low brightness RGCs; **(h)** low density RGCs; and **(i)** low density, low brightness RGCs. Scale bar = 50 μm.

The RGC quantification results computed by the cell counting program showed that the neuroprotective effect of HR97-SunitiGel lasted for at least 2 weeks after the last dose (869.2 ± 58.86 RGCs/mm^2 compared to sham, 623.7 ± 70.39 RGCs/mm^2, Fig. 3.4b), with the effect waning 4 weeks after the last dose (692.2 ± 96.58 RGCs/mm^2, Fig. 3.4c). In contrast, SunitiGel provided significant RGC protection at 1 week (846.4 ± 125.8 RGCs/mm^2) compared to the sham group, with protection waning 2 weeks after the last dose (717.3 ± 59.94 RGCs/mm^2, Fig. 3.4d).

3.3.4 HR97-SunitiGel provided increased intraocular residence time in rats and therapeutically relevant drug delivery to the posterior segment in rabbits

Based on the improved efficacy of HR97-SunitiGel at the 2-week timepoint compared to SunitiGel, pharmacokinetic characterizations were conducted in rats to determine differences in intraocular drug concentrations. At week 2 after the last topical dose, HR97-SunitiGel provided 52-fold (362.2 ng/g vs. 7.0 ng/g), 21-fold (2430.0 ng/g vs. 116.4 ng/g), and 1.3-fold (7824.6 ng/g vs. 6093.5 ng/g) higher concentrations of combined sunitinib and N-desethyl sunitinib in the rat retina, choroid/RPE, and iris/ciliary body, respectively, compared to SunitiGel (Fig. 3.5a). To subsequently confirm that therapeutically relevant drug concentrations could be achieved in the larger eyes, rabbits were dosed once daily for seven days with HR97-SunitiGel or SunitiGel. At 2 h after the last dose, HR97-SunitiGel provided 4.5-fold (27.8 ng/g vs. 6.1 ng/g), 4.7-fold (54.5 ng/g vs. 11.5 ng/g), and 3.8-fold (182.8 ng/g vs. 48.5 ng/g) higher concentrations of combined sunitinib and N-desethyl suni-

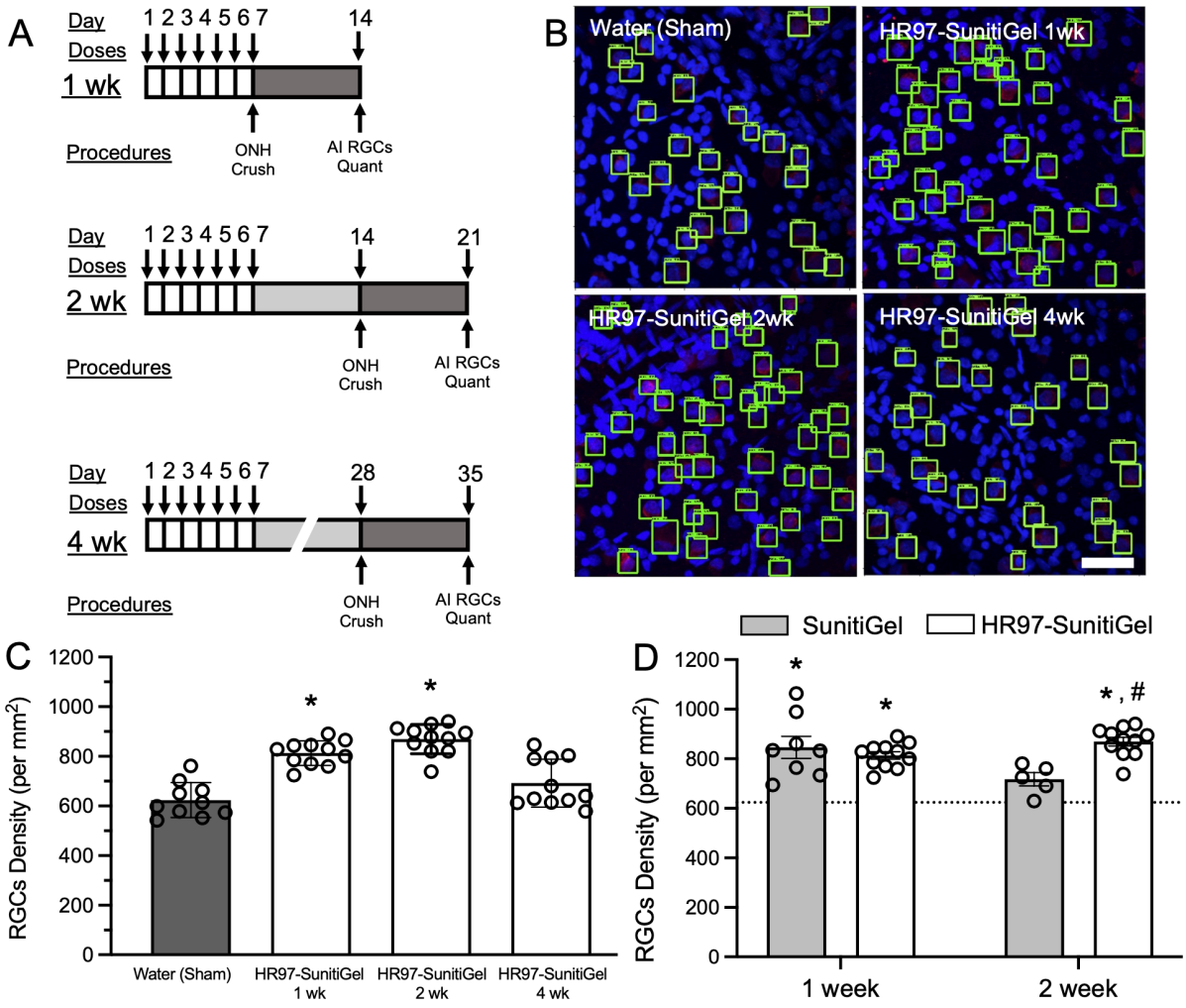


Figure 3.4 HR97-SunitiGel extended RGC protection to at least 2 weeks after the last topical dose in rat model of optic nerve injury. **a** Schematic showing the schedule for dosing HR97-SunitiGel (5 μL of 1 mg/mL equivalent sunitinib concentration) relative to the timing of the optic nerve head (ONH) crush procedure. Brown Norway rats were dosed daily for 7 days and the ONH was crushed on day 0, 7, or 21 after the last dose. After 7 days, the retinas were harvested and stained with DAPI (blue) and RBPMS (red) for RGC counting. **b** Representative images with RGCs identified by the cell counting program outlined with green bounding boxes. Scale bar = 50 μm. **c** HR97-SunitiGel provided significant neuroprotective effects for up to 2 weeks after the last dose, with the effect waning after 4 weeks. Data are presented as mean ± SD (n = 9–12 per group), $^{*}p < 0.05$. **d** SunitiGel provided neuroprotection for up to 1 week after the last dose with the effect waning after 2 weeks. The dotted line represents the mean RGC density in the Sham group. Data are presented as mean ± SD (n = 5–12 per group), $^{*}p < 0.05$ compared to the sham group, $^{\#}p < 0.05$ compared to SunitiGel.

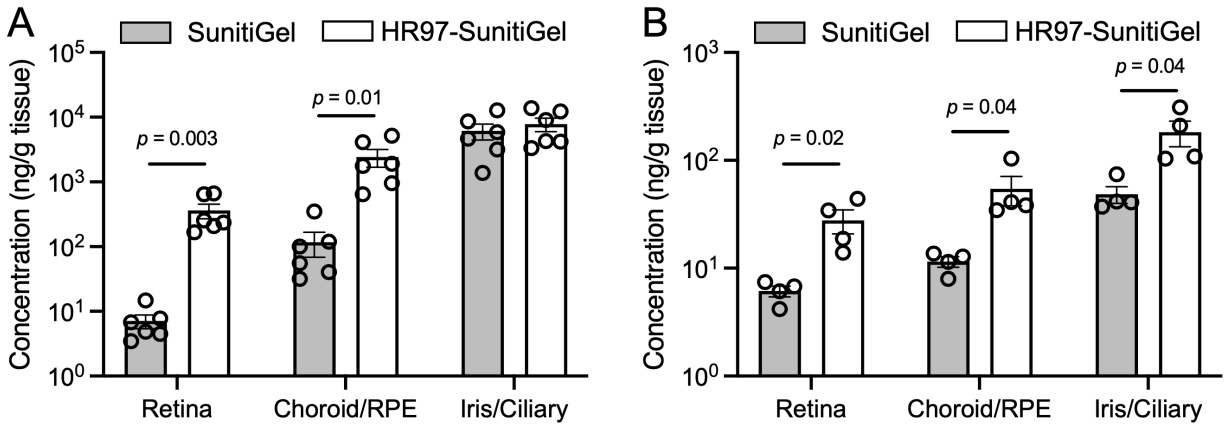


Figure 3.5 Characterization of intraocular drug concentrations after topical dosing with SunitiGel or HR97-SunitiGel in rats and rabbits. a Brown Norway rats were dosed unilaterally with SunitiGel or HR97-SunitiGel once daily for 7 days, and ocular tissues were collected 14 days after the last dose (consistent with the 2-week dosing regimen shown in Fig. 3.4a). Combined levels of sunitinib and N-desethyl sunitinib were reported per tissue sample. Data are presented as mean ± SEM ($n = 6$ per group). **b** Dutch Belted rabbits were dosed unilaterally with SunitiGel or HR97-SunitiGel once daily for 7 days, and ocular tissues were collected 2 h after the last dose. Combined levels of sunitinib and N-desethyl sunitinib were reported per tissue sample. Data are presented as mean ± SEM ($n = 4$ per group).

tinib in the rabbit retina, choroid, and iris, respectively, compared to SunitiGel (Fig. 3.5b). Importantly, the concentrations of sunitinib in the rabbit retina were comparable to concentrations found to be protective in the optic nerve crush model in rats [127].

3.4 Discussion

Patient adherence is important in treatment of chronic ocular diseases such as glaucoma, wherein patients must chronically apply IOP lowering eye drops. Typically, only 40%–75% of patients adhere to glaucoma drop therapy regimens, even in scenarios where the patients know they are being monitored and were provided free medication [134–136]. Failure to use medications as prescribed contributes to the progression of disease and can potentially lead to vision loss. Here, we investigate a complementary strategy to IOP

lowering, which is to directly target survival of the RGCs independent of IOP [125, 126]. In this scenario, there is an added challenge of achieving effective drug delivery to the posterior segment with an eye drop, which is limited by ocular tissue barriers and uveal clearance [135, 137, 138]. Designing a drug delivery system that can effectively deliver drugs to the posterior segment for neuroprotection, utilizing a non-invasive administration approach, and providing prolonged therapeutic effect to reduce dosing frequency may address several unmet needs in glaucoma management.

Melanin is a biopolymer that resides within melanosomes in pigmented ocular tissues, such as the retinal pigment epithelium (RPE), choroid, iris, and ciliary body [128, 139]. Melanin can be further classified as eumelanin and pheomelanin. Though the amount of pheomelanin in the eye is more variable in different populations, the amount of eumelanin in the RPE, pigmented ciliary epithelium, and iris pigment epithelium is relatively consistent regardless of skin and eye pigmentation [128]. There is an increasing amount of evidence indicating that binding to ocular melanin could affect the intraocular distribution and pharmacodynamics of small molecule drugs. For example, studies have shown that in pigmented rabbits atropine had prolonged residence time [140], and pilocarpine induced a sustained miotic response [141]. Further, drug-melanin interactions and the amount of free versus melanin-bound drug have been characterized in retinal pigment epithelium cells [128, 142–144], so have the correlation between *in vitro* cell uptake and *in vivo* intraocular pharmacokinetics [130, 145]. Additionally, machine learning models have been applied to melanin binding data comprised of 3400 small molecule drugs to predict how structural properties impact melanin binding [130].

Cell-penetrating peptides, including TAT, penetratin, and poly-arginine (R6 or R8),

have been utilized for delivering drugs to the anterior or posterior segments of the eye [146–150]. In a recent study, researchers developed a novel peptide-dexamethasone conjugate composed of a cell-penetrating peptide, an enzyme-cleavable linker, and dexamethasone conjugated via a hydrazone bond. Following intravitreal injection, the conjugate remained stable in the vitreous, and the dexamethasone was released via intracellular interaction with cathepsin D, thus offering a distinctive approach for sustained drug delivery to the posterior segment of the eye [150]. In our previous work, we leveraged a super learning-based methodology to engineer multifunctional peptides that are cell-penetrating, melanin binding, and have low cytotoxicity [131]. We demonstrated that conjugating the IOP lowering drug, brimonidine tartrate, to an engineered multifunctional peptide (HR97) significantly increased the IOP lowering efficacy for up to 18 days, with a 17-fold increased area under the curve compared to brimonidine solution after a single intracameral injection in rabbits [131]. We also showed that SunitiGel effectively protected RGCs in the rat optic nerve head crush model with once weekly dosing with no evidence of toxicity in healthy eyes [127]. Here, by conjugating HR97 peptide to sunitinib, topical delivery of HR97-SunitiGel effectively protected RGCs for at least 2 weeks after the last topical dose.

RGC identification and quantification are often employed in studies investigating cell and vision loss in glaucoma [125, 126, 151, 152]. The cell quantification has often been conducted manually by masked individuals hand-counting the cells [126, 127] or using the Image J software [153, 154]. Recently, the open-source CellProfiler and CellProfiler Analyst have received considerable attention for quantification of the cells because of its user-friendly interface, flexible analysis module, and integration of machine learning algorithms [155, 156]. Although CellProfiler provides an automatic pipeline for quantifying cells, the accuracy is

heavily reliant on image quality. Commercial software, such as Metamorph (BioVision, Waltham, MA), Cellomics (Thermo Fisher), or TruAI deep-learning technology (Olympus) provide a convenient user interface and pre-designed modules to process and quantify cell images, but with annual subscription fees. The RGC quantifier developed in this study is based on the open source TensorFlow deep learning object detection system with the Faster R-CNN model. Although the model was trained on a relatively small image set, our cell counting program provided a high accuracy with increased flexibility to detect cells in images with varying cell densities and image qualities. Moreover, the model could be further trained with more confocal images in the future to accommodate different RGC staining qualities or expand to various systems of glaucoma animal models via transfer learning.

Although promising, our study is not without limitations. Previously, the cell culture model system highlighted the increase in uptake and retention provided by the HR97 peptide *in vitro* [131], yet the effects were less prominent here with a drug that is intrinsically melanin binding. Similarly, the *in vitro* melanin binding assay showed significant but relatively minor increases in melanin binding capacity and affinity. However, the increased melanin binding capacity of the HR97-sunitinib conjugate paired with the increased potential for cell-penetration did sustain the pharmacokinetic and pharmacodynamic effects of sunitinib in the ONH rat model *in vivo*. Though the pharmacokinetic analysis revealed higher sunitinib concentrations in the ocular tissues of rats than in rabbits, the sunitinib concentrations achieved in the retina of rabbits with topical HR97-SunitiGel dosing were similar to what was previously found to be efficacious in the rat ONH crush model [127]. The data support the frequently described phenomenon that rodent models are not suffi-

cient for predicting drug delivery to the posterior segment with topical eye drops, and that confirmation in larger species is necessary [124]. Further, the choice of a cathepsin-labile linker was based on the fact that cathepsins are mainly found intracellularly, and they are only present in minute quantities in extracellular fluids such as vitreous and aqueous humor [157–160]. However, characterization of the linker cleavage and sunitinib release rate in cells and potential refinement of the linker chemistry may further prolong the therapeutic effect. Additionally, though we did demonstrate delivery of therapeutically relevant drug concentrations in rabbits, therapeutic efficacy has yet to be confirmed. Further, confirming the potential sustained therapeutic effects of the loading dosing approach followed by the intermittent dosing regimen (e.g., once daily dosing for seven days followed by dosing once every 2 weeks) should be confirmed in large animals. Similarly, though we did not see any evidence of toxicity with intracameral injections of HR97-brimonidine and HR97 alone in rabbits [131], more thorough studies of intraocular biocompatibility would be an important next step in the development of HR97-SunitiGel. The intrinsic melanin binding capacity of sunitinib limited the potential margin of benefit that could be observed for the duration of therapeutic effect provided by the HR97 conjugation, though we anticipate that more drastic improvements may be achievable for other drugs with low intrinsic melanin binding.

3.4.1 Conclusion

The development of innovative drug delivery systems that can overcome ocular barriers and enhance drug retention in the eye is essential for successful treatment of posterior segment diseases. Our study demonstrated that the HR97-sunitinib conjugate delivered

via gel-forming eye drops (HR97-SunitiGel) provided increased sunitinib delivery to the posterior segment of rats and rabbits and prolonged neuroprotective effect for up to two weeks after the last dose in rats. Overall, the results obtained here demonstrate the benefits of increasing the melanin binding and cell penetration of small molecule drugs in the eye. The potential for developing topical eye drop delivery systems that can not only provide effective drug delivery to the posterior segment of the eye, but also require less frequent application, would be of high value clinically and in improving patient quality of life.

3.5 Methods

3.5.1 Material sources

Sunitinib base and sunitinib malate were purchased from LC laboratories (Woburn, MA, USA). Eumelanin from *Sepia officinalis*, 0.22 μm Millex-GV PVDF filter, ferric ammonium citrate, bovine serum albumin (BSA), Tween 20, fetal bovine serum (FBS), trifluoroacetic acid (TFA), tert-Butyl methyl ether (MTBE), thionyl chloride, tetrabutylammonium iodide, N,N-diisopropylethylamine, human cathepsins B, K, L, and S, Whatman® Anotop® 0.02 μm syringe filter, Poloxamer 407, and Triton X-100 were purchased from Sigma Aldrich (St. Louis, MO, USA). ARPE-19 cells (CRL-2302) and DMEM:F12 medium were purchased from the American Type Culture Collection (Manassas, VA, USA). Rapid equilibrium dialysis (RED) 8 K device, DMEM with high glucose and pyruvate, Trypsin-EDTA (0.25% w/v) with phenol, RIPA lysis buffer, RNA binding protein, mRNA processing factor (RBPMS) rabbit anti rat polyclonal antibody, Alexa Fluor 555 conjugated goat anti-rabbit IgG (H + L) secondary antibody, penicillin/streptomycin, 4',6-diamidino-

2-phenylindole, dihydrochloride (DAPI), Fluoromount-G, Image-iT^{TM} Fixative Solution (4% w/v formaldehyde, methanol-free), and penicillin/streptomycin were purchased from Thermo Fisher Scientific (Waltham, MA, USA). Disposable PD-10 desalting columns were purchased from VWR (Radnor, PA, USA). Dulbecco's Phosphate Buffered Saline (DPBS), 1 × phosphate buffered saline (PBS), 10 × PBS, high-performance liquid chromatography (HPLC) grade acetonitrile (ACN), dimethylformamide (DMF), and water were purchased from Fisher Scientific (Hampton, NH, USA). Mc-Val-Cit-PAB was purchased from Cayman Chemical (Ann Arbor, MI, USA). Endotoxin-free ultra-pure water was purchased from Millipore Sigma (Burlington, MA, USA). Isoflurane was purchased from Baxter (Deerfield, IL, USA). Reverse-action forceps were purchased from World Precision Instruments (Sarasota, FL, USA). Neomycin, polymyxin b, and bacitracin zinc ophthalmic ointment were purchased from Akorn (Lake Forest, IL, USA).

3.5.2 Traceless linker system for conjugating HR97 to sunitinib

The traceless linker system was designed for the release of intact parent drug when triggered by an intracellular chemical and enzymatic event, such as protease cleavage of the amide bond [161]. Activation of the linker, MC-Val-Cit-PAB-OH (maleimidocaproyl-L-valine-L-citrulline-p-aminobenzyl alcohol), was conducted as previously described with minor modifications [131, 161]. MC-Val-Cit-PAB-OH (8.68 g, 15.2 mmol) was suspended in DMF (43.4 mL) at 0 °C with water bath sonication for 30 min. After solids were fully dispersed, thionyl chloride (1.22 mL, 16.7 mmol) was added dropwise. Following the addition, the reaction was held at 0 °C for 45 min and then treated slowly with water (130

mL) to precipitate a yellow solid (MC-Val-Cit-PAB-Cl), which was collected by filtration. The solid was washed sequentially with water and MTBE and dried under vacuum (∼30% yield) [161]. Sunitinib base was combined with MC-Val-Cit-PAB-Cl (1.1 eq) in DMF (0.25 M) at room temperature. Tetrabutylammonium iodide (0.5 eq) was added to the solution, followed by the addition of N,N-diisopropylethylamine (2.5 eq), and the mixture was stirred for 24 h. The mixture was diluted with 50:50 acetonitrile:water at 40-fold dilution for purifying MC-Val-Cit-PAB-sunitinib. A Shimadzu LC20 HPLC system coupled with the photodiode array detector (PDA) and the Phenomenex reverse-phase preparative HPLC column (Gemini® 10 μm C18 110 Å, LC Column 250 × 21.2 mm, AXIA^{TM} Packed) was used to separate and collect the peptide-drug conjugates with an elution gradient of 10/90/90/10% v/v solvent B (TFA 0.05% v/v in can) at 1/11/13/15 min with a flow rate of 10 mL/min. The collected fractions were then transferred to the 20 mL scintillation vials and a Biotage V-10 solvent evaporator with volatile mode was used to remove ACN. The solution fractions were frozen and lyophilized (∼7% yield). Nuclear magnetic resonance (NMR) spectroscopy was used to confirm the presence of key functional groups in the products at each stage of the synthesis, including sunitinib base, Mc-VC-PAB-Cl, and Mc-VC-PAB-sunitinib. All compounds were dissolved in deuterated DMSO and characterized with a Bruker spectrometer (500 MHz). ^{1}H chemical shifts were reported in ppm (δ) and the DMSO peak was used as an internal standard. Data were processed using TopSpin NMR Data Analysis software, version 4.1.0, from Bruker. HR97 with cysteine at the C-terminus as the functional group for linker conjugation (FSGKRRKRKPRC, MW 1.5 kDa, >97% purity from Thermo Fisher peptide custom service) was conjugated to the quaternary-ammonium-linked sunitinib (MC-Val-Cit-PAB-sunitinib) via a thiol-maleimide

reaction. The MC-Val-Cit-PAB-sunitinib was first dissolved in 1 mL of PBS at 5 mg/mL, and the HR97 peptide powder (0.5 eq) was added. The solution mixture was adjusted to pH 7.4 and allowed to react for 2 h at room temperature. The solution was then added to 1 mL of acetonitrile and purified with the same prep-HPLC conditions. The collected fractions were transferred to 20 mL scintillation vials and a Biotage V-10 solvent evaporator with volatile mode was used to remove ACN. The solutions were lyophilized and stored at −20 °C (∼29% yield). For the sample preparation and MALDI-TOF analysis, the MALDI matrix sinapic acid (10 mg) was dissolved in 1 mL of acetonitrile in water (1:1) with 0.1% TFA, and 1 μL of sample (50 μM) was deposited on the MALDI sample plate. The matrix (2 μL, 10 mg/mL) was deposited on the air-dried sample and allowed to air dry for 10–20 min. The MALDI-TOF MS analysis was performed on a Bruker Voyager DE-STR MALDI-TOF (Mass Spectrometric and Proteomics core, Johns Hopkins University, School of Medicine) operated in linear, reflective-positive ion mode.

3.5.3 *In vitro* stability test for HR97-sunitinib conjugate

Two pairs of human donor eyes were obtained from the Lions Gift of Sight under protocol IRB00056984 approved by the Johns Hopkins University School of Medicine Institutional Review Board. Both donors were male with a mean age of 74.5. The post-mortem times ranged from 35 to 40 h. The eyes were kept at 4 °C during transport and arrived within 48 h post-mortem. The vitreous from each eye was isolated, combined, and filtered through a 0.02 μm syringe filter to remove cell debris. The aqueous was similarly combined and filtered. Each fluid type was aliquoted into 3 replicates (700 μL) and HR97-sunitinib

(1 mg/mL) was added, and the mixture was incubated at 37 °C (n = 3). On day 0, 1, 7, 14, 21, and 28, 100 μL of the supernatant was collected, diluted with 900 μL of ACN, and characterized by HPLC (Prominence LC2030, Shimadzu) with Luna® 5 μm C18(2) 100 Å LC column 250 × 4.6 mm (Phenomenex, Torrance, CA). Separation was achieved with a Luna® 5 μm C18(2) 100 Å LC column 250 × 4.6 mm (Phenomenex) at 40 °C using isocratic flow (1 mL/min 60% TFA 0.1% in ACN). HR97-sunitinib was detected at λmax = 420 nm (HR97-sunitinib retention time = 1.9 min). The area under the curve (AUC) at day 0 was used to normalize the AUCs calculated at days 1, 7, 14, 21, and 28.

3.5.4 Cathepsin cleavage assay for HR97-sunitinib conjugate

An assay to demonstrate enzymatic cleavage of the linker was used as previously described with adaptations [131, 161]. In brief, the HR97-sunitinib conjugate solution (200 μM) was diluted with an equal volume of 100 mM citrate buffer at pH 5.5. Cysteine was added to a final concentration of 5 mM before the addition of human cathepsins B, K, L, and S (150 nM each). The mixture was then incubated from 0 h (control group) to 48 h at 37 °C. The solutions were further diluted with ACN to 1 mL and the conjugate concentration was measured using the HPLC method described above. The concentration of HR97-sunitinib at 0 h was used to normalize the ratio remaining at later time points.

3.5.5 *In vitro* melanin binding assay

Melanin nanoparticles (mNPs) were synthesized from the eumelanin of *Sepia officinalis* as previously described [131]. In brief, 10 mg/mL of eumelanin was suspended in

DPBS using an ultrasonic probe sonicator (Sonics, Vibra Cell VCX-750 with model CV334 probe, Newtown, CT, USA) by pulsing 1 s on/off at 40% amplitude for 30 min in a 4 °C water bath. The suspension was then filtered through a 0.22 μm Millex-GV PVDF filter and transferred to PD-10 desalting columns. The resulting mNPs solution was lyophilized for 7 days and stored at −20 °C until further use. Sunitinib malate and HR97-sunitinib at a range of concentrations (12.5, 25, 50, 100 μg/mL) were dissolved in pH 6.5 PBS solution in 3 replicates. The solutions (400 μL) were then mixed thoroughly with 400 μL of 1 mg/mL mNPs in pH 6.5 PBS solution and transferred to the inner reservoir of the rapid equilibrium dialysis (RED) device inserts (8 K MWCO). The outer reservoir was filled with 800 μL of pH 6.5 PBS solution. The samples were incubated on an orbital shaker with temperature controlled at 37 °C and 300 rpm for 48 h. The solutions from outer reservoir (free drug) were then collected and transferred to an autosampler vial for HPLC analysis (Prominence LC2030, Shimadzu, Columbia, MD) with the photodiode-array detection (PDA) system. Separation was achieved with a Luna® 5 μm C18(2) 100 Å LC column 250 × 4.6 mm (Phenomenex, Torrance, CA) at 40 °C using isocratic flow. The amount of bound drug was used to calculate the binding capacity (moles drug/mg melanin) and the dissociation constant (K_d) as previously described [127, 162]. The data point of HR97 conjugated to biotin for the high-throughput melanin binding screening assay was originally shown in our previous work [131].

3.5.6 *In vitro* cell uptake assay

The ARPE-19 cells and the induction of melanin expression in ARPE-19 cells were cultured as previously described [127]. One million ARPE-19 cells cultured for 2 months were collected for each condition and plated in 6-well plates for another 48 h. Sunitinib malate and HR97-sunitinib (25 μg/mL equivalent of sunitinib) in PBS were added to the cells for 6 h at 37 °C. The cells were then washed for 3 times with PBS using centrifugation at 3000 rcf for 5 min. After the last wash, the cells were incubated in acetonitrile at room temperature for 24 h to extract the drug. The samples were then centrifuged at 17,000 rcf at 37 °C for 30 min, and the supernatant was collected to measure the drug concentration using HPLC as described above.

3.5.7 Characterization of drug solubility

To measure solubility, 1 mg of sunitinib malate, sunitinib base, or HR97-sunitinib was placed in microcentrifuge tubes with 0.2 mL of PBS. The samples were then placed on an orbital shaker (150 rpm) in an incubator at 37 °C. After 7 days, samples were collected and centrifuged at 17,000 rcf for 30 min. The supernatants were collected, and concentrations were measured using HPLC as described above for sunitinib. Supernatant samples were mixed 1:10 with ACN containing 0.1% v/v TFA. ACN and water were used as a mobile phase at a ratio of 55:45. Samples were eluted isocratically at a flow rate of 1 mL/min through a C18-reversed phase column at 40 °C. UV absorbance was monitored at 420 nm.

3.5.8 Animal studies—Animal welfare statement

All experimental protocols were approved by the Johns Hopkins Animal Care and Use Committee. All animals were handled and treated in accordance with the Association for Research in Vision and Ophthalmology Statement for Use of Animals in Ophthalmic and Vision Research. Equivalent numbers of both male and female animals were used. Brown Norway rats (6–10 weeks) were obtained from Harlan/ Envigo. Dutch Belted rabbits (4–5 mo) were obtained from Robinson Services, Inc.

3.5.9 Rat optic nerve head (ONH) crush model

We previously described that the delivery of sunitinib malate in SunitiGel provided sustained RGC protection with once weekly dosing in an optic nerve head (ONH) crush model [127]. The model was implemented similarly to investigate the timing of quantification of RGCs in the time-course study and the potential benefit of the HR97 peptide conjugation. In the time-course study, Brown Norway rats received the optic nerve head crush on day 0, and on day 1, 4, 7, 11, 14, or 19, the retinas were harvested and stained with DAPI and RBPMS for quantifying the remaining RGCs using the AI deep learning algorithm ($n = 6$). In the HR97-SunitiGel RGC protection study, Brown Norway rats ($n = 36$, split into 3 groups of 12) received seven daily doses (5 μL) of 1 mg/mL sunitinib equivalent in HR97-SunitiGel (6.2 mg of HR97-sunitinib dissolved in 1 mL of 12% w/w F127). The ONH crush procedure was then performed on separate groups of animals on day 0, 7, and 14 after the last dose. Another group of rats ($n = 12$) received seven daily eye drops (5 μL) of water as a sham control prior to undergoing the ONH crush procedure

on day 0 (immediately after the last dose). The third group of rats ($n = 12$) received seven daily eye drops (5 μL) of 1 mg/mL equivalent sunitinib in SunitiGel (1.34 mg sunitinib malate in 1 mL of 12% w/w F127, equal to 1 mg/mL sunitinib equivalent). Rats received general anesthesia prior to topical anesthesia. Proparacaine hydrochloride (0.5% w/v) was applied topically to the right eye 1 min before the surgical process. The temporal conjunctiva of the left eye was grasped with 0.12 mm toothed forceps and incised parallel to the limbus with sharp iris scissors. The dissection was performed using two pairs of curved blunt-tipped forceps, and the orbital fat and soft tissue were retracted to expose the orbital portion of the optic nerve. The optic nerve was crushed at a position 1.5–2 mm posterior to the globe using reverse-action forceps for 10 s. The orbital soft tissue was then repositioned over the nerve and the conjunctiva was left to close by secondary intention [127]. After the procedure, topical bacitracin-neomycin-polymyxin ophthalmic ointment was applied to both eyes to prevent infection. Any animals that bled severely during the surgery were sacrificed and excluded from the study. Seven days after the optic nerve crush, rats were sacrificed for subsequent analyses.

3.5.10 Retinal ganglion cell staining and imaging

The process of harvesting retina tissues, staining, and imaging were performed as previously described [127, 163]. Rats were sacrificed by cervical dislocation under general anesthesia. The eyes were then harvested and fixed with 4% w/v paraformaldehyde for 2 h. The retinas were removed, incised for flat mounting, and post-fixed for 2 h. The retinas were then washed with PBS containing 0.5% w/v Triton-100 for 30 min and incubated

for 3 days at 4 °C in a solution containing rabbit anti-rat RBPMS antibody diluted 1:250 in PBS with 1% w/v Triton X-100 and 1% w/v BSA. The retinas were then washed for three times with PBS containing 0.5% w/v Triton-100 and incubated overnight at 4 °C in a solution containing Goat anti-Rabbit IgG H&L Secondary Antibody Alexa Fluor 555 (Thermo Fisher, Waltham, MA, USA) diluted 1:1000 in PBS with 1% w/v Triton X-100 and 1% w/v BSA. The retinas were washed again for three times and incubated overnight in DAPI diluted 1:1000 in PBS. The resulting retinal wholemount was then mounted on a slide using Fluoromount-G. The prepared retinas were imaged with a Zeiss 710 Confocal Microscope. For each retinal wholemount, 16 images were taken from the region 2–3 mm from the optic nerve per each retinal quadrants using a 40X objective. The DAPI images were pseudo-colored in blue and the RBPMS images were pseudo-colored in red.

3.5.11 Retinal ganglion cell counting

For training the TensorFlow Object Detection model (TensorFlow Version 1.2), 173 images containing 4247 RGCs, and 62 images containing 1757 RGCs were manually labeled using the *LabelImg* function as the training and testing sets, respectively. Faster R-CNN with Inception Resnet v2 and SSD-MobileNet were used as the TensorFlow Object Detection models. The weighted sigmoid (sigmoid cross-entropy loss function) was used for the classification loss, and the weighted smooth L1 (box regression in object detection) was used for the localization loss. To avoid overfitting, the training process was terminated when the total loss was <1 or when the total loss has reached a steady state. An inference graph at each epoch was created with the script export_inference_graph.py provided in

the object_detection directory. The CellProfiler, version 3.1.9, with CellProfiler Analyst, version 2.2.1, was used in this study. The *Metadata* function was used to extract the DAPI and RBPMS sub-layer information from confocal images and the names were assigned using *NamesAndTypes*. The DAPI layers were first smoothed with the Gaussian filter function, and the *IdentidyPrimaryObjects* function was used to identify the DAPI areas using global and three-class Otsu threshold methods. The RBPMS sub-layers were smoothed with the Guassian Filter and then again smoothed with the Median Filter because the RBPMS staining with the RGCs had given uneven intensities and confused the program in the subsequent steps. The *IdentifySecondaryObjects* function with Propagation strategy and adaptive, three-class Otsu, and foreground methods were used to identify the RBPMS as secondary objects based on the DAPI areas identified in the primary object session. The object shape size, object intensity, and texture were measured for both DAPI and RBPMS objects as the features for CellProfiler Analyst, version 2.2.1. An SQLite database was generated after the CellProfiler had finished the feature extraction. In CellProfiler Analyst, *RandomForestClassifier* was used and RGCs in the unclassified window were dragged to the positive area following the suggested protocols. >100 cells in each positive and negative class were then manually assigned until the classification accuracy reached over than 90% computed using the evaluation function. Finally, the output scores were used to quantify the RGCs in each image. RGCs in another set of 200 images were manually counted by three researchers masked to the sample identity and the means were calculated for each image. In a rare instance that the cell count per image varied by >10%, the images were recounted by each person until the variance was <10%. Quantification results of the RGC images predicted by Faster R-CNN with Inception Resnet v2, SSD-MobileNet, and

CellProfiler Analyst models were compared to the manual counting results using Pearson correlation.

3.5.12 Pharmacokinetic studies

Brown Norway rats ($n = 6$) received once daily eye drops (5 μL) containing 1 mg/mL sunitinib equivalent in HR97-SunitiGel (6.2 mg of HR97-sunitinib dissolved in 1 mL of 12% w/w F127) or containing 1 mg/mL equivalent sunitinib in SunitiGel (1.34 mg sunitinib malate in 1 mL of 12% w/w F127, which was equal to 1 mg/mL sunitinib equivalent) for seven days. Fourteen days after the last dose, the iris, choroid, and retina were collected and analyzed for sunitinib concentration using LC-MS/MS. Similarly, Dutch Belted rabbits ($n = 4$) received once daily eye drops (50 μL) containing 1 mg/mL sunitinib equivalent in HR97-SunitiGel or containing 1 mg/mL equivalent sunitinib in SunitiGel. Two hours after the last dose, the iris, choroid, and retina were collected for analysis of sunitinib concentration using LC-MS/MS.

3.5.13 Measurement of sunitinib in ocular tissues

Sunitinib concentrations in ocular tissues were measured by liquid chromatography-tandem mass spectrometry (LC-MS/MS) as previously described [124]. All samples were collected in pre-weighed tubes and stored at -80 °C until being processed for analysis. Tissue samples were homogenized in 100–600 μL 1 $\times$ PBS using a Next Advance Bullet Blender before extraction. Sunitinib was extracted from 15 to 50 μL of tissue homogenates with 50 μL of acetonitrile containing 50/50/2.5 ng/mL of the internal standards (sunitinib-

d10). The top layer was then transferred to an autosampler vial for LC-MS/MS analysis after centrifugation. All ocular tissue samples were analyzed using a 1 × PBS standard curve for sunitinib. Separation was achieved with Waters Cortecs C18 (2.1 × 50 mm, 2.7 μm). The column effluent was monitored using a Sciex triple quadrupol 4500 with electrospray ionization operating in the positive mode. The Mobile phase A was water containing 0.1% formic acid and the mobile phase B was acetonitrile containing 0.1% formic acid. The gradient started with the mobile phase B held at 10% for 0.5 min and increased to 100% within 0.5 min; 100% of mobile phase B was held for 1 min, and then the mobile phase B returned back to 10% and was allowed to equilibrate for 1 min. The total run time was 3 min with a flow rate of 0.3 mL/min. The spectrometer was programmed to monitor the following MRM transition 399.1 → 283.2 for sunitinib and 409.1 → 283.2 for the internal standard, sunitinib-d10. Calibration curve for sunitinib was computed using the area ratio peak of the analysis to the internal standard by using a quadratic equation with a x-2 weighting function over the range of 0.25–500, with dilutions up to 1:100 (v:v). Core technicians performing sample and data analysis were masked to the treatment groups.

3.5.14 Statistical analyses

Statistical analyses of two groups were conducted using two-tailed Student's t-test, two-tailed Mann–Whitney test, or two-way analysis of variance (ANOVA). For comparison of multiple groups, one-way ANOVA with Dunnett's multiple comparison test was used. Pearson correlation coefficients (r) and the corresponding p-values (two-tailed) were calculated to assess the relationships between model predictions and the mean values of the

manual counting.

Part III

Malaria Vaccine Antigen Identification

Chapter 4: Positive-Unlabeled Learning Identifies Vaccine Candidate Antigens in the Malaria Parasite *Plasmodium falciparum*

4.1 Abstract

Malaria vaccine development is hampered by extensive antigenic variation and complex life stages of *Plasmodium* species. Vaccine development has focused on a small number of antigens identified prior to availability of the *P. falciparum* genome. In this study, we implement a machine learning-based reverse vaccinology approach to predict potential new malaria vaccine candidate antigens. We assemble and analyze *P. falciparum* proteomic, structural, functional, immunological, genomic, and transcriptomic data, and use positive-unlabeled learning to predict potential antigens based on the properties of known antigens and remaining proteins. We prioritize candidate antigens based on model performance on reference antigens with different genetic diversity and quantify the protein properties that contribute the most to identifying top candidates. Candidate antigens are characterized by gene essentiality, gene ontology, and gene expression in different life stages to inform future vaccine development. This approach provides a framework for identifying and prioritizing candidate vaccine antigens for a broad range of pathogens.

4.2 Introduction

Artemisinin-based combination therapies and other tools have contributed to substantial reductions in the malaria burden in many endemic areas over the last decade [164]. However, progress toward malaria elimination has stalled as malaria incidence has plateaued and gains have been threatened by the emergence of resistance to interventions in the parasite and vector [164–167]. With the possible future exception of dracunculiasis caused by Guinea worm, no infectious disease has been completely eradicated without the aid of an efficacious vaccine [168, 169]. Thus, malaria vaccines are a critical tool for malaria elimination.

Plasmodium parasites are transmitted to humans when infective mosquitoes take a blood meal and inject sporozoites, which develop and multiply in the liver. Vaccines directed against this pre-erythrocytic stage is meant to block infection. After emerging from the liver, *Plasmodium* merozoites invade and replicate inside red blood cells. This erythrocytic stage of the life cycle causes malaria disease and death, which blood-stage vaccines are intended to limit. Transmission-blocking vaccines would inhibit parasite sexual recombination and development in the mosquito, preventing onward transmission [170]. Design of a broadly protective malaria vaccine has been hampered by several factors, including multiple parasite life stages that express different antigens, extensive genetic diversity within individual antigens targeted by vaccines, partial natural immunity that is short-lived and non-sterilizing, and incomplete knowledge of immune correlates of protection [171]. To date, very few malaria vaccine candidates have been evaluated in clinical trials, with most demonstrating limited efficacy [172], including the first malaria vaccine approved for use

by the World Health Organization, RTS,S, which displayed only 36% efficacy in a Phase 3 trial when given to 5–17 months old as a primary series followed by a booster dose [173]. Another promising vaccine, R21, recently showed an efficacy of 71% in phase 1/2b [174].

Malaria parasites are haploid in humans and briefly diploid in mosquitoes. Extensive genetic variation is generated through mutation during mitotic reproduction in humans and by sexual recombination in the mosquito. The first *P. falciparum* genome was published in 2002 [175], but nearly 20 years later most vaccine development efforts have focused on a small number of highly diverse vaccine candidates identified prior to the availability of the genome using traditional vaccinology approaches that identify antibody targets in immune sera. These highly immunogenic candidates have typically evolved extensive genetic diversity in response to immune pressure. Thus, many vaccines have displayed some degree of allele-specific efficacy (including RTS,S) [176–179], demonstrating greater efficacy against parasites with target alleles matching those in the vaccine formulation (i.e., vaccine allele-specific efficacy) [171].

Reverse vaccinology utilizes bioinformatics approaches to identify pathogen antigens or epitopes that could be used as vaccine candidates [180–182]. It was first proposed by Rino Rappuoli who screened the Meningococcus B proteome to identify five antigens with bactericidal activities, which were subsequently included in the licensed four-component MenB vaccine (4CMenB, Bexsero®) [183–186]. Reverse vaccinology has since been used to identify vaccine antigens for other bacterial and viral pathogens [187–192]. The wealth of systems data available for *P. falciparum* lends itself to the use of reverse vaccinology to identify new malaria vaccine antigens, which may allow identification of less immunodominant but more conserved antigens that have been missed using traditional vaccinology

approaches based strictly on immunogenicity. There has been limited use of reverse vaccinology to identify malaria candidate antigens. Singh *et al.* [193] applied the concept to identify candidate antigens with signal peptide and glycosylphosphatidylinositol (GPI) anchor motifs while Pritam *et al.* [194] also used signal peptide and GPI-anchor prediction tools along with T-cell epitope prediction to identify *P. falciparum* epitopes. Both studies focused on a limited number of protein or epitope properties. In contrast, machine learning in reverse vaccinology does not require *a priori* assumptions about the importance of specific criteria, and instead, "learns" protein properties most associated with vaccine potential based on known antigens.

Positive-unlabeled (PU) learning is applicable to many biological problems where the labeling process is often expensive or time-consuming, and only a small fraction of entities might be labeled [192, 195]. Learning from the labeled positives, PU learning identifies potential positives among the unlabeled entities based on the properties of the data [196, 197]. This approach has been used to identify genes associated with human disease based on various data types, including human protein interaction data, gene expression data, gene ontology, and phenotype-gene association data [198], yet to our knowledge, it has not been applied to identify candidate antigens. PU learning is particularly attractive for *P. falciparum*, as approximately 40% of genes in the genome encode proteins of unknown function [199, 200].

Here, we modify canonical positive-unlabeled random forest (PURF) [201] to distinguish proteins with vaccine potential (i.e., antigens) from non-antigens, based on properties of known *P. falciparum* antigens, and rank the candidates with probability scores. Variable importance is assessed to understand the protein properties contributing most to

identifying candidate antigens. The candidates are linked to other data types (e.g., gene essentiality [202], stage-specific single-cell transcriptomic data [203–205], and proximity to the known malaria vaccine antigens), to allow further characterization and prioritization in subsequent vaccine development.

4.3 Results

4.3.1 Identification of potential *P. falciparum* candidate antigens

In this study, 52 known antigens were selected from the intersection of the antigen sets obtained from the literature and from epitope information from the Immune Epitope Database (IEDB) [206], based on their ability to elicit an immune response [192]. Among the known antigens, four antigens (CSP, MSP5, P230, and RH5) representing vaccine candidates from different parasite life stages and with varying levels of genetic diversity [207–210], were selected to serve as reference points for candidate antigen prioritization. A relational database was created to organize data assembled and generated for the *P. falciparum* proteins (Figs. 4.1, C.1). The structural, proteomic, and immunological data were generated using various bioinformatic programs (Section 4.5). We also retrieved genomic, transcriptomic, and functional information from public databases such as PlasmoDB [199]. Additional variables were created by combining variables from different data types. The 272 variables comprise 28 structural variables, 121 proteomic variables, 116 immunological variables, and 7 genomic variables.

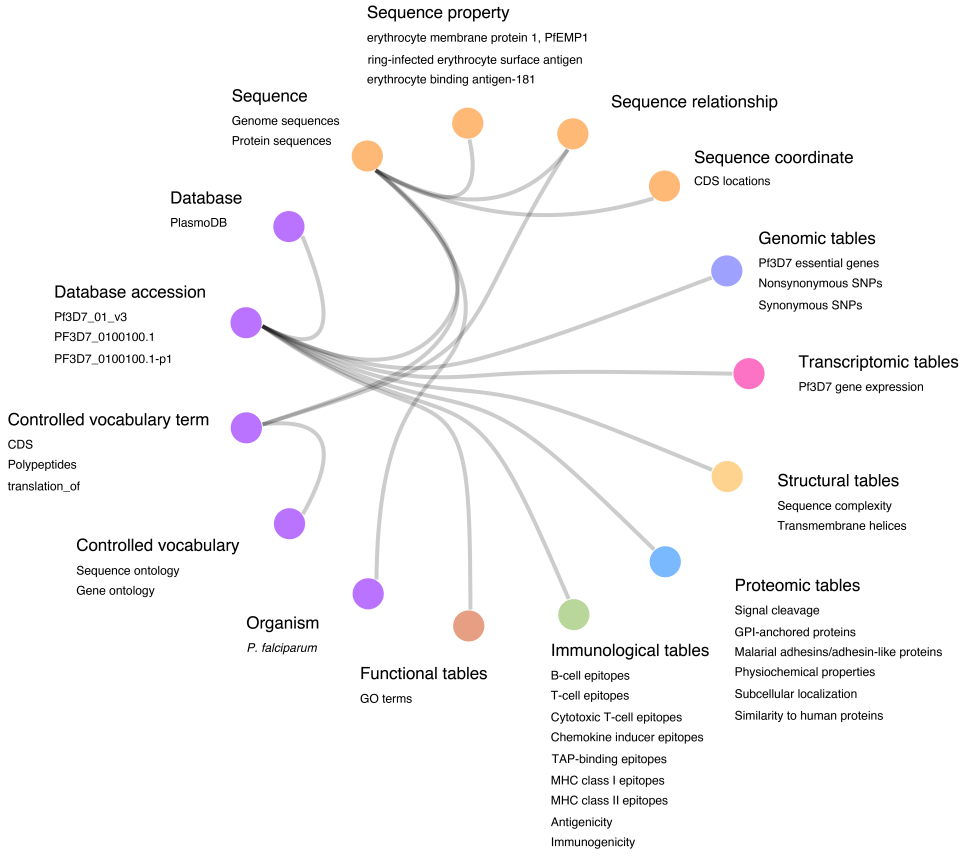


Figure 4.1 Database schema of *P. falciparum* vaccine target identification. The database is structured as a collection of data tables here represented as nodes with colors indicating different groups of tables. Part of the tables in the database are listed as examples. The lines of the hierarchical edge bundling plot show the hierarchical relationships between tables. The orders in the hierarchical structure are origin (root node), group of tables, and data table. Tables with the same type of relationship to the foreign table are collapsed to one node. Data tables generated from computational analyses are connected to sequence (purple) and basic information (orange) tables with gene accession identifiers.

4.3.2 Training positive-unlabeled random forest models

We employed tree-based PU learning (PURF), an ensemble of individual tree models. PURF incorporates a modified impurity measure (see Section 4.5) that estimates the probabilities of the positives and negatives based on observations in the tree node [201]. To evaluate the ensemble, we simulated fully labeled data, and estimated the receiver operating characteristic (ROC) curve, which was calculated using the probability scores (out-of-bag scores; see Section 4.5). The estimated ROC curve was then compared with ROC curves calculated using the probability scores against the true labels and using the PU labels (Fig. C.2). The estimated ROC curves were like those of true labels (Mann–Whitney, $q = 0.06$, $n = 5$), while the ROC curves of PU labels were different from the others ($q = 0.01$ for both comparisons). This result demonstrates that even without the true label information, the ROC curve may be recovered from the score distribution.

To select the positive level (hyperparameter for prior probability of positive samples) of PURF, we trained with positive levels from 0.1 to 0.9. The positive level of 0.5 shows the highest area under the estimated ROC curve (AUROC = 0.98) (Fig. C.3). Proteins were ranked based on probability scores, which is defined as the proportion of trees in the ensemble predicting the protein to be antigenic. The overall percentile ranks (PRs) of the known antigens were highest for the ensemble with 0.1 positive level (area under the ranking curve; $AUC = 0.83$), whereas all known antigens were predicted correctly (explicit positive recall; $EPR = 1$) by the ensembles with 0.8 and 0.9 positive levels (Fig. C.4). The AUC and EPR of the ensemble with 0.5 positive level were 0.81 and 0.83, respectively.

To improve the performance we utilized a method like the synthetic minority over-

sampling technique (SMOTE) [211] to increase representation of known antigens. The weighting made known antigens equally representative by duplicating those that are more distant from others in the variable space, which increased classification performance. The estimated ROC curve showed an increase in classification separability ($AUROC = 0.99$, $positivelevel = 0.5$, Fig. C.5). The known antigens obtained a higher percentile rank (Fig. C.6), and the EPR of the ensemble with 0.5 positive level increased to 0.92.

4.3.3 Classification tree filtering using reference antigens

To utilize the random forest structure to prioritize candidate antigens, we identified tree models that correctly predicted all reference antigens that were in the out-of-bag set of the tree (those proteins not used to build the tree). Trees that did not have reference antigens in the out-of-bag set or incorrectly predicted any of the out-of-bag reference positives were removed. PURF with tree filtering had an estimated AUROC of 0.99 (Fig. 4.2a). The evaluation of the 52 known antigens showed that 51 had percentile rank >50 and the EPR was 0.94 (Fig. 4.2b). For further characterization, we selected the top 200 candidate antigens with a probability score >0.94 because half of the known antigens had scores above this threshold.

To assess robustness, we performed an iterative validation procedure by sequentially removing the positive label from one of the 48 known antigens (excluding the four reference antigens) from each iteration as an adversarial control [212], conducted variable space weighting, and retrained our models. The results show small mean differences in scores of the remaining known antigens before and after the label removal (Fig. 4.2c), and there

was no significant difference between filtered and unfiltered ensembles (Mann–Whitney, $p = 0.32$, Fig. C.7). The top 200 candidate lists from the 48 ensembles were generated, and the cumulative numbers of candidates that agreed on 48, 47, 46 ensembles, and so on, are similar between the filtered and unfiltered ensembles (Fig. 4.2d), demonstrating that the tree filtering procedure did not affect the overall PURF structure in predicting candidate antigens.

To understand protein variables contributing to the identification of the known antigens, we investigated the mean decrease in prediction accuracy across all trees in the filtered ensemble with variable permutations. The top ten most important variables include one structural, one genomic and eight proteomic variables (Fig. 4.3a). Comparisons of the variable values between the known antigens and 52 random proteins predicted to be non-antigens by tree-filtered PURF reveal that the known antigens contain fewer amino acids with high polarizability (K, M, H, F, R, Y, W), comprise fewer amino acids with high van der Waals volume (M, H, K, F, R, Y, W), and have fewer hydrophobic amino acids (C, L, V, I, M, F, W) (Fig. 4.3b). Moreover, the known antigens have fewer positively charged amino acids (K, R) and a lower isoelectric point value (Fig. 4.3b). Known antigens also have a higher secretory signal peptide probability, a higher number of non-synonymous SNPs, and have higher flexibility and hydrophilicity for predicted epitopes (Fig. 4.3b). The importance of variables grouped by data categories showed that the proteomic variables are most important in identifying known antigens (Fig. 4.3c).

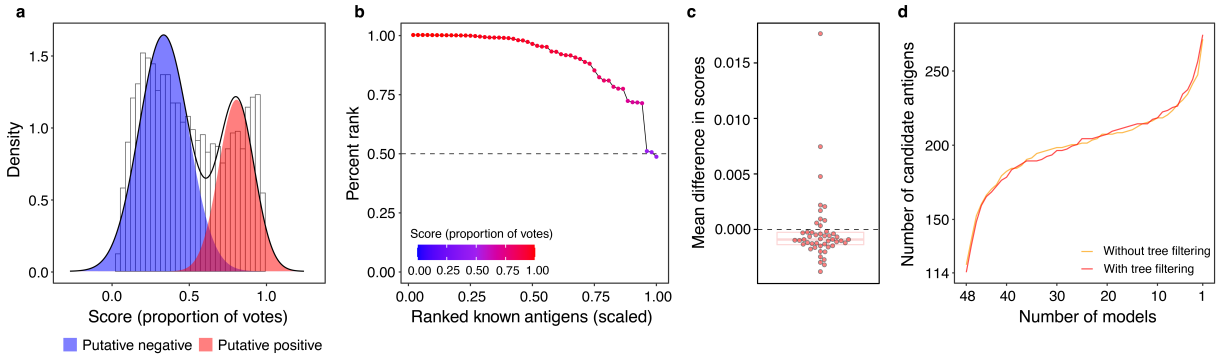


Figure 4.2 Model evaluation and validation of positive-unlabeled random forest models. **a** Score distributions of unlabeled proteins predicted by the tree-filtered model. The putative positive (red) and negative (blue) distribution groups were calculated by fitting a two-component Gaussian mixture model. A receiver operating characteristic curve (ROC) was calculated based on the putative distributions, and the area under the receiver operating characteristic curve (AUROC) was 0.99. **b** Evaluation of known antigen scores predicted by the tree-filtered model. Points represent known antigens. The x-axis shows the scaled ranks of the 52 known antigens. The y-axis notes percentile ranks (PR) of known antigens in the set containing all *P. falciparum* proteins. The dashed line indicates the 50th percentile rank. Gradient colors show probability scores. The area under the ranking curve was 0.90. **c** Distribution of mean differences in scores after known antigen label removal for the final tree-filtered ensemble. Dots represent the 48 validation iterations. The box plot shows median with first and third quartiles. The lower and upper whiskers indicate 1.5 × interquartile range from the first and third quantiles, respectively. The grey dashed line conveys a zero-mean difference in scores. **d** Plot of overlapping antigens across the top 200 candidate sets generated from the validation models. The x-axis shows the number of validation models in reverse order, and the y-axis indicates the number of candidate antigens in agreement with the corresponding number of models. Line colors show data from non-tree-filtered (yellow) and tree-filtered (red) validation models, respectively.

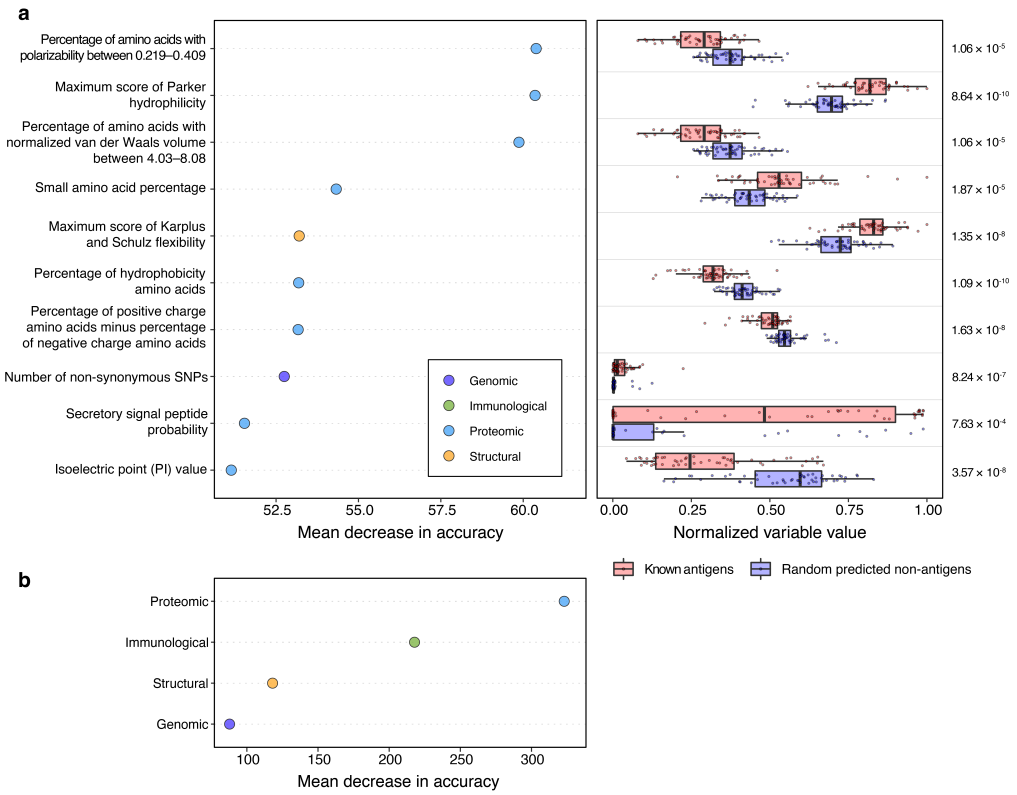


Figure 4.3 Positive-unlabeled random forest model interpretation based on known antigens. a The left panel displays permutation-based variable importance analysis of the final tree-filtered model. The x-axis shows the mean decrease in accuracy (scaled by the standard error) of the known antigen set ($n = 52$) after permuting the variables for each tree in the model. The y-axis lists the ten most important variables in predicting the known antigens. The property groups of the variables are noted by colors. The right panel shows summary of variable values of the known antigens (red) and randomly selected proteins ($n = 52$; blue) that are predicted as non-antigens by the final tree-filtered model. The ten most important variables obtained from the premutation-based variable importance analysis are shown. Points represent proteins. Boxplots show median with first and third quartiles, and the whiskers indicate the 1.5 interquartile range extended from the first and third quartiles. Numbers on the right show adjusted p-values calculated using two-sided Mann–Whitney tests. Variable values were normalized based on the entire data set. **b** Permutation-based group variable importance analysis. Variable importance was calculated on the known antigens, and the decrease in accuracy after variable permutation was recorded. Variables in the same property groups were permutated together. The mean decrease in accuracy was standardized using the standard error computed across all trees in the model.

4.3.4 Proximity of top-ranked candidates to reference antigens

To understand how tree filtering assisted in prioritizing antigen candidates based on the reference antigens, we examined the proximity space before and after tree filtering. Proximity values are the proportion of times a pair of proteins occur in the same terminal node of a tree model and represent the similarity with respect to variables used in the model. The proximity was converted to a Euclidean distance (smaller values indicate more closeness) and visualized using multidimensional scaling. The top candidate antigens were clustered into three groups (Fig. 4.4). The probability scores of candidate antigens in groups 1 and 2 increased after tree filtering (Fig. C.8), indicating that some candidates in groups 1 and 2 have been prioritized into the top candidate list after tree filtering.

To study the relationships between the candidate and reference antigens, we compared the Euclidean distances of the candidate antigens in each group to each of the four reference antigens. The distances significantly changed (FDR <0.05) in all three groups after tree filtering. Comparing the three groups, after tree filtering, group 3 had the farthest median distance to the reference antigens, group 1 had the closest median distance to CSP, MSP5, and P230, group 2 is closest to RH5 (red points in Fig. C.9). For RH5, MSP5, and P230, both groups 1 and 2 moved closer to the three reference antigens (blue and purple points in Fig. C.9) and group 3 moved further away after tree filtering (dark orange points in Fig. C.9), suggesting that reference antigens may have less effect on prioritizing group 3 antigen candidates. Overall, RH5, MSP5, and P230 may have positive influences on the prioritization of group 1 and group 2 antigen candidates. Interestingly, the median distances of group 2 antigen candidates are less than 0.5 to all reference antigens (red

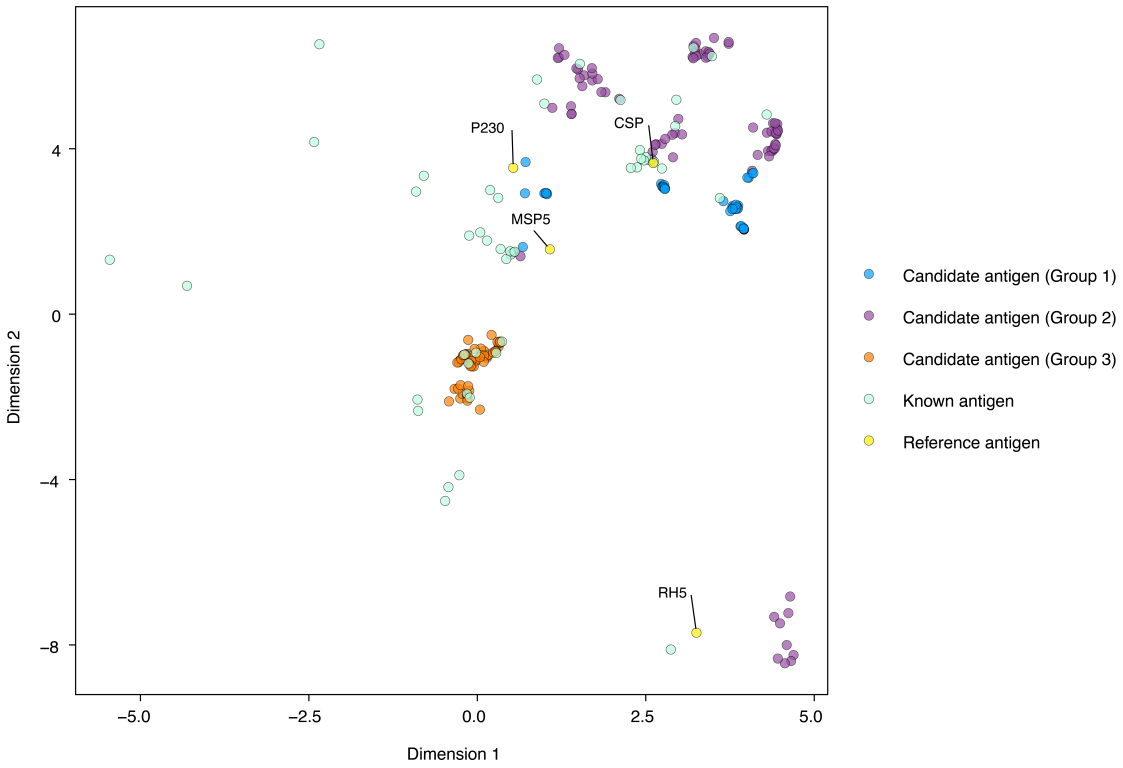


Figure 4.4 Clustering of top 200 candidate antigens based on proximity measured from tree-based model. First two dimensions of UMAP are shown. Top 200 candidate antigens from the final tree-filtered model were grouped based on k-means clustering. Points represent top 200 candidate antigens in three groups, 48 known antigens (light cyan), and four reference antigens (yellow; protein names noted by text).

points in Fig. C.9), suggesting that over half of the trees in PURF agreed on the protein similarities between group 2 and all four reference antigens.

4.3.5 Variable importance of candidate antigen groups

Permutation-based variable importance analyses were conducted for each of the three candidate antigen groups. The shared importance variables in identifying the candidates as antigens for the three groups includes higher number of non-synonymous SNPs, higher flexibility and hydrophilicity for predicted epitopes, lower probability of mitochondrial sub-

cellular localization, and fewer number of hydrophobic amino acids (Fig. C.10). The shared important properties of candidate antigens in the three groups were similar to the properties of known antigens. Among the three groups, group 2 had the most similar important variables as known antigens. The secretory signal peptide probability, which was ranked nineth in the important variable list for known antigens, was ranked 83rd, 226th, and 83rd in the results of groups 1, 2, and 3, respectively, suggesting that secretory signal peptide may be important in classifying the protein as antigens (probability score $\geq$0.5), but not as critical for higher probability score ($\geq$0.9).

In terms of top-ranked variables, the number of non-synonymous SNPs, epitope flexibility, and epitope hydrophilicity were ranked among the top three for groups 1 and 2, and among the top four for group 3 (Tables C.1–C.3). The median number of non-synonymous SNPs is lower and there was a smaller variance in the distribution for group 2 compared to groups 1 and 3 (Fig. C.10). The predicted number of B-cell epitopes in outer membrane regions was ranked as the most important variable for group 3, whereas it was the least important among the 272 variables for groups 1 and 2 (Table C.3).

4.3.6 Characteristics of identified potential vaccine antigen targets

We applied gene ontology (GO) enrichment analysis to assess annotation-associated properties of candidate antigen groups compared to the background of the *P. falciparum* proteome. Group 1 was significantly enriched for genes encoding proteins involved in cell-cell adhesion, cytoadherence to the microvasculature, erythrocyte aggregation, and antigenic variation. Similar enriched GO terms were observed for group 3 (Table 4.1). Group

2 candidate antigens were enriched in parasite nucleus and cytoplasm and not associated with antigenic variation (Table 4.1), suggesting these potential antigens may be less immunogenic or less exposed to the host immune response. Further examination of the gene products of group 1 revealed that 85% of the candidates are erythrocyte membrane proteins (PfEMP1), whereas 36% and 26% of candidates in groups 2 and 3, respectively, are conserved proteins with unknown functions.

We filtered the candidate antigen groups by gene essentiality, where genes with mutagenesis index score (MIS) <0.5 were retained [202]. We examined the expression of the genes encoding the remaining candidate antigens in different *P. falciparum* life stages based on single-cell transcriptomic data from the Malaria Cell Atlas [203–205]. Of the group 1 candidates remaining after essentiality filtering, one was expressed mainly in the blood stage, and the other was expressed in all life stages, with higher expression levels in a larger portion of cell populations in the blood and gametocyte stages (Fig. C.11). For groups 2 and 3, most candidate antigen genes were expressed primarily in the blood, gametocyte, and ookinete stages, and a smaller number of groups 2 and 3 candidates were expressed in all life stages (Fig. C.11).

4.4 Discussion

Over the past decades various malaria vaccine candidates have been developed and proceeded to clinical trials. Nevertheless, a highly efficacious and long-lasting malaria vaccine against *P. falciparum* is still an unmet need. We are now in the second wave of malaria vaccine development [213], with the goal of selecting vaccine antigens with potential

to elicit an enhanced immune memory response and a protective efficacy of at least 75% against clinical malaria [214]. With the advancement of genome sequencing of *Plasmodium* and bioinformatics tools, reverse vaccinology has become a viable vaccine development approach for this complex organism.

Reverse vaccinology has been applied using sequential filtration of protein properties or with machine learning, both of which have identified potential new vaccine antigens for *Plasmodium* species, but with some limitations [194, 215]. Approaches based on sequential filtration lack standardized filtering criteria, with thresholds often selected based on empirical evidence, and could be difficult to generalize when there are many protein variables [192, 216]. In *P. falciparum*, there are only a small number of known antigens that can be labeled as positives, and non-antigens are difficult to identify from the literature or based on reference genomes with incomplete annotation. One study using machine learning algorithms to predict potential vaccine antigens in eukaryotic pathogens only examined seven protein variables and did not consider genome properties such as sequence complexity and genetic diversity [215], both of which are relevant to malaria vaccine development and have impacted the efficacy of first-generation malaria vaccines [176–179]. Additionally, this study examined only a relatively small set of three *Plasmodium* proteomes (73 antigens and 51 non-antigens, from *P. falciparum*, *P. yoelii yoelii*, and *P. berghei*). In contrast, we performed comprehensive analyses on 5,393 *P. falciparum* proteins and computed 272 protein variables on each. To ensure a high-quality PU data set of known *P. falciparum* antigens, we took the intersection of antigen sets curated from the literature and IEDB [206].

PU learning takes advantage of unlabeled data and improves modeling when only a small portion of entities are labeled as positive [196, 217]. In this study, we chose random

forest [218] as the basis for our PU learning because of its high predictive accuracy, high interpretability, and insensitivity to outliers and predictive variable scales [219]. Additionally, PURF is amenable to the modifications we developed here. Permutation-based variable importance analysis is naturally derived from the random forest architecture and imparts a quantitative measure of the variable importance [218]. Moreover, many studies involving machine learning analyses focus primarily on the model accuracy and develop complex models that are hard to interpret. However, it is critical to understand the relationships learned by the model and whether they are biologically meaningful [212, 220]. Here, the interpretation of PURF provides helpful insights on how the models have learned in distinguishing known antigens from non-antigens, and how the previously unknown candidate antigens were identified. Although PU learning enabled us to fully harness the entire *P. falciparum* proteome, it is a data-driven approach that could be affected by the known antigens provided. Thus, in this study, efforts were made to ensure the quality of the known antigens. Further inclusion of more high-quality known antigens may improve the model performance.

To develop malaria vaccines with higher efficacy, it is critical to consider genetic variation that is immunologically relevant [171]. Four reference antigens that are actively in development as malaria vaccines were chosen to help understand our models. Circumsporozoite protein (CSP) is a surface protein expressed during the pre-erythrocytic stage and is the active component of the RTS,S vaccine [221]. Reticulocyte binding homolog 5 (RH5) is expressed in the blood stage, functions as an invasion ligand, and is currently under malaria vaccine development [222, 223]. Merozoite surface protein 5 (MSP5) is associated with natural antibody responses [224]. P230, in the 6-cysteine protein family, is expressed

and located on the surface of gametocytes [225]. These reference antigens display a range of genetic diversity, as measured by percentile rank of SNPs per Kb coding sequence over *P. falciparum* proteome (P230 0.39, MSP5 0.43, RH5 0.52, CSP 0.94).

The approach described in this study identified previously unknown vaccine candidate antigens for *P. falciparum* vaccine development. The research scheme provides a flexible framework, in which the candidate antigens can also be prioritized using a different set of reference antigens selected using other criteria, while not affecting the overall PURF structure. Candidate antigens identified in this study have been filtered based on gene essentiality, where mutations in these genes could affect parasite viability, and thus may help reduce parasite escape from vaccine-induced immune responses [202]. Most candidate antigens were expressed predominantly in a single life stage, which is consistent with the observations of previous studies [226]. For instance, group 3 antigens were mostly expressed in blood and sexual stages, which were associated with higher number of B-cell epitopes in the outer membrane regions. However, some candidates were expressed in multiple life stages, which may make them attractive vaccine antigens because they would target multiple life stages. An interactive summary report of the identified candidate antigens identified is available in an online research notebook (`https://doi.org/10.13016/me1l-1ahr`). The information about the closest reference antigens to the candidates and single-cell gene expression is also included. For future studies, further filtering criteria, such as isoelectric point, molecular weight, and folding propensity, may be applied to select candidate antigens for heterologous protein expression in other species systems to perform functional assays [227, 228].

Our approach exploits PU learning in reverse vaccinology to identify potential *P. fal-*

ciparum vaccine candidate antigens for future vaccine development, which does not assume filtering criteria of protein variables, is driven by the proteome, and leverages a small set of known antigens. The alteration of the model ensemble based on the reference antigens aids in candidate antigen prioritization. In response to the shift in species constitution in malaria endemic areas, the developed framework can be expanded to *P. vivax* and other *Plasmodium* species that cause human malaria [229]. The methodology can be further tailored and applied to other disease pathogens. More broadly, beyond vaccine development, the study may also inspire other scientific research areas, if there is only a relatively small amount of evidence collected to guide the prioritization of the study entities.

4.5 Methods

4.5.1 Known antigen protein collection

Known antigens were selected based on literature and epitope information. Covidence (`www.covidence.org`), a web-based application tool designed for systematic review and streamline the screening of literature search, was used to select, and extract literature covering malaria vaccine research. Our goal was to look for all the malaria vaccine candidates that have been already reported in the literature. The search terms include the following: "malaria vaccine", "malaria vaccine candidate", "malaria vaccine antigen", "malaria vaccine protein". In brief, the search covered papers and documents having both malaria and vaccine in any of its sections. The search generated a set of articles that discuss malaria vaccine candidates, rather than a list of each of the candidates. Overall, our search produced 7,415 articles in total. We then manually went through all these papers

to identify proteins used as malaria vaccine candidates. Non-redundant candidates were selected based on gene names, GenBank ID, or aliases.

The known antigens selected based on the epitope information were extracted from the PlasmoDB [199] immunology section. Epitopes from the Immune Epitope Database (IEDB) [206] are mapped to the PlasmoDB proteins with exact string matching; at the same time, the corresponding GenBank proteins from IEDB were aligned to PlasmoDB proteins using BLAST [230]. The similarity threshold of a best hit is percent identity $\geq$97%. We selected proteins from PlasmoDB as known antigens if the protein has a similarity score larger than or equal to the similarity threshold, or having all listed epitopes aligned exactly to the PlasmoDB protein sequence. The set based on the literature contained 177 known antigens, and the set based on the epitope information had 373 known antigens. The final known antigen list was an intersection of the two sets and included 52 antigen proteins.

4.5.2 Collection of *Plasmodium* data and bioinformatic analyses

P. falciparum 3D7 genome information and protein sequences were collected from PlasmoDB [199] release 43 (2019-04-25). An in-house database was constructed using MariaDB version 10.3.22 (`https://mariadb.com/`). The data tables are connected via table identifiers or gene accessions. Part of the Chado schema from the Generic Model Organism Database [231] was integrated into the database design to eliminate redundancies. The database contains eight categories of tables, including basic information, sequence information, genomic, transcriptomic, functional, structural, proteomic, and immunological tables.

In brief, the reference genome, coding sequences (CDS), and protein sequences were directly downloaded from PlasmoDB [199]. Proteins with stop codons within the sequence or derived from pseudo genes were removed. Protein sequences having "X" symbols were also removed. Selenocysteines in selenoproteins were replaced with cysteines for downstream bioinformatic analyses. The preprocessing resulted in 5,393 *P. falciparum* proteins. General information including genome, coding sequence locations, protein sequences, and sequence ontology terms were stored in the basic information and sequence information database tables.

For genomic data tables, single nucleotide polymorphisms (SNPs) discovered from next-generation sequencing (NGS) were directly downloaded from PlasmoDB [199] under the genetic variation section (365 genomes). The measures of SNPs include total number of SNPs, number of non-synonymous SNPs, number of synonymous SNPs, number of nonsense SNPs, number of non-coding SNPs, ratio of non-synonymous to synonymous SNPs, and number of SNPs per Kb coding sequence. Gene essentiality measured from saturation-level mutagenesis was obtained from the literature [202]. Transcriptomic data included DNA microarray [232] and bulk RNA-seq [233–236] data at various *P. falciparum* life stages retrieved via PlasmoDB, and single-cell RNA-seq data from the Malaria Cell Atlas (MCA) [203–205]. Functional data including gene ontology terms were downloaded directly from PlasmoDB [199] as a GAF file.

For structural data, transmembrane helices were predicted using the TMHMM version 2.0 web server (`http://www.cbs.dtu.dk/services/TMHMM/`) [237,238]. Sequence complexity was analyzed using the SEG [239]. Beta-turns, surface accessibility, and flexibility were analyzed using IEDB Antibody Epitope Prediction version 3.0 [240–242]. By combining

the results from TMHMM and SEG, new protein variables of sequence complexity in the outer-membrane (non-cytoplasmic), transmembrane, and inner-membrane (cytoplasmic) regions were generated.

For proteomic data, subcellular localizations were predicted using the CELLO version 2.5 web server (`http://cello.life.nctu.edu.tw`) [243]. Malarial adhesins/adhesin-like proteins were predicted using the MAAP web server (`http://maap.igib.res.in/index.php`) [244]. Physicochemical properties were analyzed using the R packages *Peptides* version 2.4.1 [245] and *protr* version 1.6.2 [246], and IEDB Antibody Epitope Prediction version 3.0 [247]. Glycosylphosphatidylinositol (GPI)-anchored proteins were predicted using the PredGPI web server (`http://gpcr.biocomp.unibo.it/predgpi/pred.htm`) [248]. Protein signal cleavage prediction was analyzed using the SignalP version 5.0 web server (`http://www.cbs.dtu.dk/services/SignalP/index.php`) [249]. Protein solubility information was obtained using the protein-sol abpred [250]. N- and O-linked glycosylation sites were predicted using GlycoEP [251]. The results of glycosylation sites were combined with the transmembrane predictions to generate additional variables of glycosylation sites in the outer-membrane, transmembrane, and inner-membrane regions. Similarity to human proteins was analyzed using BLASTP version 2.8.1+ [252].

For immunological data, T cell epitopes were predicted using the PREDIVAC web server (`http://predivac.biosci.uq.edu.au/cgi-bin/population.py`) [253], which predicted epitopes specifically for sets of HLA class II allelic variants from ten population regions. B cell epitopes were analyzed using BepiPred version 2.0, BepiPred version 1.0, and ABCpred [254–256]. Additional variables of B cell epitopes in the outer-membrane, transmembrane, and inner-membrane regions were computed using the transmembrane in-

formation from TMHMM. Cytotoxic T cell epitopes were analyzed using CTLPred [257]. Chemokine inducer epitopes were analyzed using IL-10Pred [258]. Transporter associated with antigen processing (TAP)-binding peptides were predicted using TAPPred [259]. MHC class I and class II epitopes were predicted using IEDB MHC-I Binding Predictions version 2.22.3 [260] and IEDB MHC-II Binding Predictions version 2.22.3 [261], respectively. Epitope antigenicity was analyzed using IEDB Antibody Epitope Prediction version 3.0 [262], and epitope immunogenicity was predicted using IEDB Class I Immunogenicity version 1.1 [263]. In general, the epitope information was summarized for each protein with the total number of epitopes passed the default threshold, and the maximum, mean, and minimum scores of the epitopes.

4.5.3 Data set assembly

The data set contains the predictor variables, and the response variable labels. The variables were assembled by retrieval from the database. Antigen labeling information was added as the response variable, where proteins selected as known antigens were labeled as positive and the other proteins were unlabeled. The number of proteins was 5,393, and the number of known antigens as labeled positives was 52. In total, 272 predictor variables were retrieved from the database. All predictor variables are of numeric type, and missing values in the variables were imputed by replacement with variable medians.

4.5.4 Positive-unlabeled simulation

The simulated data were generated using the function *make_classification* from the Python *scikit-learn* package [264]. The number of proteins was 5,000 and the number of predictor variables was 300, comprising 250 informative variables, 40 redundant variables, and 10 repeated variables. The response variable contained two classes (positive and negative) and was treated as true labels. Because the *P. falciparum* data set had 52 labeled positives (known antigens) out of 5,393 proteins, the data set was 99% unlabeled. To convert true labels to positive-unlabeled (PU) labels, a regular random forest classifier with 1,000 trees was trained to obtain probability (out-of-bag) scores for all proteins. We then randomly selected 50 proteins that were predicted to be positive by the regular random forest. We retained the positive labels of these 50 proteins and made the remaining 4,950 proteins unlabeled.

4.5.5 Positive-unlabeled random forest algorithm implementation

The positive-unlabeled random forest (PURF) framework is based on a modified splitting criterion called positive-unlabeled Gini index (PUGini) [201], which is derived from the Gini criterion ($Gini = 1 - \sum_j p_j^2$, where p_j is the probability of being classified a class j) [265]. The new splitting criterion estimated probabilities of positive and negative proteins according to the numbers of labeled positives and unlabeled proteins in the tree node. The probabilities of positive (p_1) and negative (p_0) proteins were respectively estimated by the following equations [201, 266], $p_1 = min(|POSnode| \times PosLevel \times |UNL|, 1)|POS||UNLnode|$, and $p_0 = 1 - p_1$, where |POSnode| and |UNLnode| are, re-

spectively, the numbers of labeled positives and unlabeled proteins in the node, and |POS| and |UNL| are, respectively, the numbers of labeled positives and unlabeled proteins in the data. Because PURF is based on random forest [218], it inherits the properties of robustness to outliers and variable errors, insensitivity to monotonic transformation of variables, and high predictive power. In this study, we implemented the PURF algorithm by extending the ensemble and tree modules in the Python scikit-learn package [264] and developed a lightweight Python package. The framework proposed by Li and Hua [201] was slightly modified where the positive level (PosLevel) has become a hyperparameter that can be explicitly tuned by the user. We also added class functions that take tree weights as an argument to calculate probability scores with the tree filtering procedure. For the initial modeling, positive levels were set to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. The forest size was 100,000 trees.

4.5.6 Positive-unlabeled random forest evaluation

Because in a PU learning problem we do not know the true state for the unlabeled proteins, we cannot calculate the traditional evaluation metrics such as those involving true negative and false positive rates. Further, the metrics based on PU labels could be affected by the proportion of labeled positives [267]. Thus, in this study we used the following two criteria, which utilize the probability score distribution from PURF and the percentile rank of labeled positives, respectively, to examine the model performance.

The first criterion involves estimating the putative true and false positive rates from the probability score distribution to calculate the receiver operating characteristic (ROC)

curve. As the probability score distribution of unlabeled proteins is bimodal, the distribution can be described using a two-component mixture with the formula: $h(x) = \pi h_1(x) + (1 - \pi)h_0(x)$, where $\pi \in (0, 1)$ and $x \in X$, X being the set of all possible proteins, h_1 is the score distribution of putative positive proteins, and h_0 is the score distribution of putative negative proteins. In this study, the two-component Gaussian mixture was computed using the R package *mixR*, version 0.2.0 [268], and the area under the receiver operating characteristic curve (AUROC) was calculated using the R package *pracma*, version 2.3.8 [269].

The second criterion calculates the percentile rank of labeled positives (known antigens) among all protein proteins based on the probability scores. The criterion also reports the proportion of labeled positives that are correctly predicted (explicit positive recall; EPR) [217, 270]. The area under the percentile rank curve was computed using the R package *pracma*, version 2.3.8 [269].

4.5.7 Variable space weighting

Because only ∼1% of the data were labeled as positive, the scarcity of the labeled positive may not well represent the positive (antigen) population. To make all known antigens equally representative for learning the antigen properties, a variable space weighting procedure was performed before training. A vector of variable medians was generated according to the variable set of the known antigens. The vector represents the center point [271], which is a generalized geometric median in higher-dimensional data, of the known antigens in the variable space. The Euclidean distance between each known antigen and the center

point was then calculated. The distances were scaled to rounded integer values from 1–10. A new data set was generated by duplicating the known antigens with the transformed distances. The number of labeled positives after variable space weighting was 122.

4.5.8 Ensemble constituent filtering

The tree filtering was conducted using four selected reference antigens (CSP, MSP5, P230, RH5) to prioritize top-scored unlabeled proteins. To select trees that correctly predicted the reference antigens in the out-of-bag set, trees having no references, or which incorrectly predicted any of the out-of-bag antigens were removed, resulting in 74,089 trees filtered from the original 100,000 trees. The probability scores were recalculated using the function *_set_oob_score_with_weights*, where the removed trees were assigned with a weight of zero.

4.5.9 Positive-unlabeled random forest validation

Known antigens, excluding the four reference antigens, were converted to unlabeled proteins iteratively. For each of the 48 iterations, a variable-space-weighted data set was generated, and an ensemble with a positive level of 0.5 determined through hyperparameter tuning and 100,000 trees was trained. The model was subsequently processed using the ensemble constituent filtering procedure. The probability scores of the remaining known antigen predicted by both unfiltered and filtered models were recorded. The differences in scores compared to the ensembles with no antigen label removal were calculated, and the mean of these differences were then computed for each iteration. Finally, the mean

differences in scores from unfiltered and filtered models were compared using a two-sided pairwise Mann–Whitney test. Additionally, top 200 unlabeled proteins ranked based on probability scores were selected for each validation model, and the numbers of proteins showed up in n, $n-1$, $n-2$, … rank lists were reported ($n = 48$).

4.5.10 Candidate antigen clustering and comparisons

To calculate the proximity matrix [218] for the final tree-filtered forest with 74,089 trees, a matrix was computed using the Python function *apply*. The matrix is symmetric with rows and columns being the proteins, and a cell value of 1 indicates the paired proteins end up in the same terminal nodes of a tree. The proximity matrix was then computed by dividing the number of trees for which the paired proteins were in the out-of-bag set. The proximity matrix was converted to a Euclidean distance matrix by subtracting the proximity value from 1. The distance matrix was further converted to a (5,393 – 1)-dimensional space using multi-dimensional scaling (MDS) with the R function *cmdscale*. The variance explained for each dimension was calculated by dividing the eigenvalue by the sum of all positive eigenvalues.

The top 200 candidate antigens were selected from the final ensemble. A k-means clustering analysis was performed on the subset of the multi-dimensional data set containing the top 200 candidate antigens. To select the optimal number of clustering groups, the Gap statistic [272] with the Tibshirani criterion [273], Silhouettes [274], and Elbow (or total within sum of square) methods were used. The number of clusters selected by the three methods were 3, 2, 3, respectively. Thus, the top 200 candidates were clustered into

three groups, and visualized along with the known antigens and reference antigens on the first two dimensions of the uniform manifold approximation and projection (UMAP) [275] matrix.

For the three candidate antigen groups, we quantified three measures comparing candidate antigens between non-tree-filtered and tree-filtered ensembles: 1) probability scores; 2) Euclidean distances from the candidate antigens to each of the four reference antigens; and 3) differences in distances. For these comparisons, we used multiple pairwise Mann–Whitney tests (probability scores and Euclidean distances), and Mann–Whitney test (differences in distances), with p-values adjusted by the Benjamini–Hochberg method [276].

4.5.11 Variable importance analyses

Permutation-based variable importance [218] was calculated for the 52 known antigens, 61 group 1 antigen candidates, 83 group 2 antigen candidates, and 56 group 3 antigen candidates. For each tree in the forest, the prediction accuracy was recorded for the out-of-bag target proteins (e.g., the 52 known antigens). For each of the 272 variables, the variable values were permuted for all 5,393 proteins, the tree was then used to predict the response of the permuted data set, and the prediction accuracy for the out-of-bag target proteins was calculated. The difference in prediction accuracy before and after variable permutation was recorded for each variable permutation. After iterating through all trees in the forest, the results from each tree were weighted according to ensemble constituent filtering (filtered trees have a weight of zero), and the weighted average of decrease in accuracy and the corresponding standard error were calculated for each variable across all trees. The final mean

decrease in accuracy was scaled by dividing the values by the standard error. For the importance analysis for variables grouped by data categories, the variables were grouped based on data properties (genomic, structural, proteomic, and immunological). When calculating the importance of each data category, the grouped variables were permuted together, and the decrease in prediction accuracy was measured after permutation.

4.5.12 Variable value comparisons of top important variables

To compare variable values, a set of non-antigens predicted by the final tree-filtered ensemble with the same size as the target proteins (known antigens or candidate group antigens) were randomly selected. The variable values of the target proteins and randomly selected non-antigens were compared using a two-tailed Mann–Whitney test for all 272 variables. The p-values were adjusted for multiple tests using the Benjamini–Hochberg procedure [276]. The variable values were normalized to be between 0–1 based on the original data set with 5,393 proteins for better visualization. The top ten most important variables based on the permutation-based variable importance analysis were visualized.

4.5.13 Gene ontology enrichment analysis

Candidate antigen groups were analyzed separately using the function *GOEnrichmentStudyNS* in the *GOATOOLS* Python package [277]. The GAF files containing associated gene ontology terms of *P. falciparum* 3D7 genes was retrieved from PlasmoDB [199] release 59 (2022-08-30). The directed acyclic graph file of gene ontology was downloaded from the Gene Ontology website (`http://geneontology.org/docs/download-ontology/`) [278,

279]. The argument propagate_counts was set to false for more conservative results. The p-values generated from multiple Fisher's exact tests were adjusted using the Benjamini–Hochberg method (or false discovery rate; FDR) [276]. The significance cut-off was set at 0.05.

4.5.14 Candidate antigen characterization

Candidate antigens in each of the three groups were further filtered based on gene essentiality that measured from saturation-level mutagenesis of *P. falciparum*, the threshold of MIS <0.5 was chosen as described in the original paper [202]. After filtering, there were 2, 26, and 14 candidates in group 1, group 2, and group 3, respectively. The candidate antigens were further characterized using the single-cell transcriptomic data from the Malaria Cell Atlas [203–205] that contained 12 life stages, including five sporozoite stages, three blood stages, three gametocyte stages, and one ookinete stage. The gene counts were normalized by size factors and $\log_2$-transformed. The proportion of cells at each stage having gene counts larger than zero, and the median and mean gene counts in the cell populations were reported. Further, the closest reference antigen to each candidate antigen based on the proximity matrix was identified. The final data set contained probability scores, clustering groups, gene products from PlasmoDB [199] release 59 (2022-08-30), closest reference antigen and the corresponding Euclidean distance.

4.5.15 Statistical analyses

R version 4.2.1 (2022-06-23) and RStudio were used to perform statistical analyses. For comparing the scores and Euclidean distances of antigen proteins and candidate antigens from models with or without tree filtering, a pairwise two-tailed Mann–Whitney test were used. For comparisons of variable values between target proteins (known or candidate antigens) and randomly selected non-antigens, or comparisons of difference in distances across the three candidate antigen groups, a regular two-tailed Mann–Whitney test was conducted. Where appropriate, the p-values for multiple tests were adjusted using the Benjamini–Hochberg procedure [276].

Table 4.1 Significantly enriched gene ontology terms with false discovery rate (FDR) <0.05 in gene ontology enrichment analysis of candidate antigen groups with the background proteome of *P. falciparum* 3D7.

	GO terms		Number of genes	$-\text{Log}_{10}$ FDR
Group 1 (61 candidates)	Biological process	Cell-cell adhesion	45	4.12
		Cytoadherence to microvasculature, mediated by symbiont protein	43	3.55
		Modulation by symbiont of host erythrocyte aggregation	42	3.53
		Antigenic variation	43	3.50
	Cellular component	Host cell plasma membrane	44	4.74
		Infected host cell surface knob	44	4.74
		Integral component of membrane	54	3.91
		Maurer's cleft	5	1.34
	Molecular function	Cell adhesion molecule binding	44	4.45
		Host cell surface receptor binding	51	3.71
		Protein binding	8	2.65
Group 2 (83 candidates)	Biological process	Chromatin remodeling	4	3.95
		Regulation of transcription, DNA-templated	7	3.19
		Positive regulation of transcription, DNA-templated	2	1.53
	Cellular component	Nucleus	45	3.57
		Cytoplasm	18	3.51
		Membrane	9	2.87
		Extracellular region	3	2.08
		Chromosome	2	1.44
		Rhoptry neck	2	1.44
		P-body	2	1.34
		Vesicle	2	1.34
	Molecular function	DNA-binding transcription factor activity	7	4.65
		ATP binding	12	4.06
		DNA binding	9	4.06
		Sequence-specific DNA binding	6	4.06
		Protein binding	21	3.90
		Actin binding	3	2.52
		Chromatin binding	3	2.19
		Protein phosphatase regulator activity	2	2.01
		Histone-lysine N-methyltransferase activity	2	1.69
		Calcium ion binding	3	1.67
Group 3 (56 candidates)	Biological process	Cell-cell adhesion	6	4.35
		Entry into host	5	3.54
		Protein phosphorylation	4	2.09
		Response to xenobiotic stimulus	4	1.81
		Cytoadherence to microvasculature, mediated by symbiont protein	4	1.46
		Modulation by symbiont of host erythrocyte aggregation	4	1.41
		Cell motility	2	1.37
		Antigenic variation	4	1.37
	Cellular component	Integral component of membrane	47	4.26
		Nucleus	13	4.26
		Membrane	16	4.23
		Infected host cell surface knob	4	3.56
		Host cell plasma membrane	5	2.72
		Apicoplast	5	1.79
		Rhoptry neck	2	1.79
		P-body	2	1.66
		Cytoplasm	7	1.43
	Molecular function	Heparin binding	4	3.84
		Host cell surface receptor binding	7	3.84
		Cell adhesion molecule binding	4	3.29
		Protein kinase activity	4	2.36

Chapter 5: *Plasmodium vivax* Antigen Candidate Prediction Improves with the Addition of *Plasmodium falciparum* Data

5.1 Abstract

Intensive malaria control and elimination efforts have led to substantial reductions in malaria incidence over the past two decades. However, the reduction in *Plasmodium falciparum* malaria cases has led to a species shift in some geographic areas, with *P. vivax* predominating in many areas outside of Africa. Despite its wide geographic distribution, *P. vivax* vaccine development has lagged far behind that for *P. falciparum*, in part due to the inability to cultivate *P. vivax in vitro*, hindering traditional approaches for antigen identification. In a prior study, we have used a positive-unlabeled random forest (PURF) machine learning approach to identify *P. falciparum* antigens for consideration in vaccine development efforts. Here we integrate systems data from *P. falciparum* (the better-studied species) to improve PURF models to predict potential *P. vivax* vaccine antigen candidates. We further show that inclusion of known antigens from the other species is critical for model performance, but the inclusion of unlabeled proteins the other species can result in misdirection of the model toward predictors of species classification, rather than antigen identification. Beyond malaria, incorporating antigens from a closely related species may

aid in vaccine development for emerging pathogens having few or no known antigens.

5.2 Introduction

Malaria is an infectious disease caused by protozoan parasites of the *Plasmodium* genus, which exhibit a multi-staged, complex life cycle in the host and the vector [280]. Despite considerable reductions in the malaria burden over the past two decades, malaria incidence has plateaued, or even increased, in the past 5-7 years [281]. The hard-won progress is now in jeopardy due to the emergence of resistance in both the parasite and the vector, alongside a decrease in investments in malaria control/eradication activities and research [280–283]. Furthermore, with the reduction of *Plasmodium falciparum* in some endemic areas, a shift in species composition has been reported, with *Plasmodium vivax* predominating in many areas outside of Africa [284, 285]. There are several factors likely contributing to this shift, including the ability of *P. vivax* to cause relapsing infections from dormant liver stages (hypnozoites), low parasite densities that can escape standard diagnostic tests, the early emergence of infective gametocytes prior to clinical symptom onset, as well as a shorter development cycle in the mosquito vector [286, 287]. These factors will likely make elimination of *P. vivax* malaria more challenging than elimination of *P. falciparum* malaria [286]. Currently, there are only a few *P. vivax* vaccine candidates in the clinical development stage [287]. Vaccine development for *P. vivax* faces some similar challenges as *P. falciparum* vaccine development, including a complex parasite life cycle with multiple stages, where different antigens are expressed at each stage [288], as well as genetic diversity, which is greater in *P. vivax* than in *P. falciparum* [289]. For immunogenic

surface proteins, this genetic diversity may lead to vaccine escape, as observed in *P. falciparum* [290, 291]. Additionally, the lack of *in vitro* culture capabilities for *P. vivax* further complicates the development of vaccines against this parasite [288].

Systems data, including whole genome sequence data, has become increasingly available for many pathogens, including *Plasmodium* [292–294]. Leveraging these data, Rino Rappuoli and colleagues proposed a reverse vaccinology approach to identify potential vaccine antigen candidates targeting the B strains of *Neisseria meningitidis* (meningococci), which resulted in the licensed MenB vaccine [295, 296]. Reverse vaccinology has been employed to a lesser extent to identify potential vaccine antigens in *Plasmodium* species [297]. However, these studies have primarily focused on *P. falciparum*, and the protein or epitope selection criteria have been limited [298, 299]. Recently, we described a machine learning-based approach designed to learn the properties of a limited set of known antigens using a large number of protein variables [300]. The machine learning algorithm we used, known as positive-unlabeled random forest (PURF) [301], is particularly useful for many classification problems where some labels are missing, and thus, only portion of the positive class is labeled [302, 303]. In this recent study, we trained PURF on *P. falciparum* with a limited number of high-quality known antigens and prioritized top-ranking candidate antigens from the unlabeled proteins [300].

Here, we utilize the PURF algorithm to train a machine learning model to identify potential vaccine antigen candidates for *P. vivax*. We further improve the model accuracy by adding data from *P. falciparum*. The impact of incorporating the heterologous data is then analyzed based on two data types: heterologous known antigens and heterologous unlabeled proteins. Our results demonstrate that the inclusion of known antigens from a

different species slightly improves the accuracy of vaccine antigen predictions. However, the integration of unlabeled proteins from another species could inadvertently amplify effects related to species distinction, potentially misdirecting the classification algorithm to focus more on protein variables that differentiate the two species over the task of antigen identification. Thus, it is critical to include only labeled data from another species in the final model. To understand variables that are important for *P. vivax* antigen identification, we then conduct variable importance analysis on the final model. Top-ranking candidate antigens are clustered into three groups, which undergo further characterization. Our methodology demonstrates potential for prioritizing and accelerating malaria vaccine development for *P. vivax* and other minority *Plasmodium* species, presenting a promising solution for addressing the global burden of malaria.

5.3 Results

5.3.1 Data engineering and model training

P. vivax protein variables were derived from publicly available genome assemblies, as well as various bioinformatics analyses, including genomic, immunological, proteomic, and structural data types (refer to Section 5.5 for further details). In this study, our existing database [300] containing solely *P. falciparum* protein variables was expanded to include data from *P. vivax*. The data set contains 6,491 *P. vivax* proteins and 272 protein variables. The selection of antigen labels for training the machine learning models was based on the union set of antigens identified in the literature and those identified using the immune epitope database (IEDB) [304], resulting in 38 known *P. vivax* antigens labeled

as positives, with remaining proteins designated as unlabeled. Positive-unlabeled random forest (PURF) [301] models with different hyper-parameter settings were constructed for the *P. vivax* data. We compared models having different positive levels representing the prior proportion of potential antigens within the whole proteome. Model performance was assessed using two metrics: the area under the receive operating characteristic curve (AUROC), which measures the separation of antigen and non-antigens based on prediction scores; and the area under the curve (AUC) for percentile ranks of known antigens, quantifying how these known antigens are ranked amongst all proteins. The evaluation of the predicted probability score distributions showed that the *P. vivax* model with a positive level of 0.5 had the highest AUROC, 0.99, indicating the model capability in identifying antigens from non-antigens (Fig. D.1). Further investigation of known antigen predictions revealed that the model, with positive level of 0.5, identified 34 known antigens with percentile ranks exceeding 0.5, and a moderately high AUC of 0.84, showing the model was able to identify known antigens. The explicit positive recall (EPR; percentage of correctly predicted labeled positives) [303,305] for known antigen prediction accuracy was 87% (Fig. D.2 and Table 5.1). The *P. falciparum* data were then added into the training data set to explore whether these data would improve the current model. Thus, PURF models were trained using a combined data set including data from both species. The combined model had an AUROC of 0.995 based on the probability score distribution, higher than that of the model including *P. vivax* data only (Fig. 5.1a, b and Fig. D.3). All 90 known antigens (38 *P. vivax*, 52 *P. falciparum*) were correctly predicted (EPR = 100%) by the combined model (Table 5.1). Moreover, all known antigens had percentile ranks above 0.5 across the entire combined protein set, with an AUC of 0.94, suggesting an improvement

in antigen predictions for both species (Fig. D.4 and 5.1b, d).

Table 5.1 *P. vivax* and *P. falciparum* known antigen prediction accuracies of PURF models trained separately on *P. vivax*, *P. falciparum*, and combined data sets.

PURF model	Prediction accuracy of *P. vivax* known antigens	Prediction accuracy of *P. falciparum* known antigens
P. vivax model	33 / 38 = 0.87	47 / 52 = 0.90
P. falciparum model	34 / 38 = 0.89	49 / 52 = 0.94
Combined model	38 / 38 = 1.00	52 / 52 = 1.00

5.3.2 Comparison of single-species models and the combined model

In this study, we defined the data from the focus species as autologous and referred to the data from the other species as heterologous. We further compared the *P. vivax* and *P. falciparum* single-species models trained on the individual species data sets against the combined model by making heterologous predictions based on the single-species models. The *P. falciparum* single-species model accurately identified 89% of the heterologous *P. vivax* known antigens, and the *P. vivax* single-species model correctly predicted 90% of the heterologous *P. falciparum* known antigens. The combined species model predicted all *P. falciparum* and *P. vivax* known antigens correctly, resulting in a 100% accuracy (Table 5.1). We further focused on the antigen prediction results of the two single-species models to assess the predicting performance of merely merging the outputs of both models, instead of training on a combined data set. For *P. falciparum* known antigen predictions, the *P. falciparum* and *P. vivax* models together correctly identified all 52 antigens. However, for the *P. vivax* known antigens, only 35 out of 38 antigens were detected across both single-species models. To validate the single and combined models, we conducted an iterative validation process involving adversarial control [306], where we removed one antigen label at a time

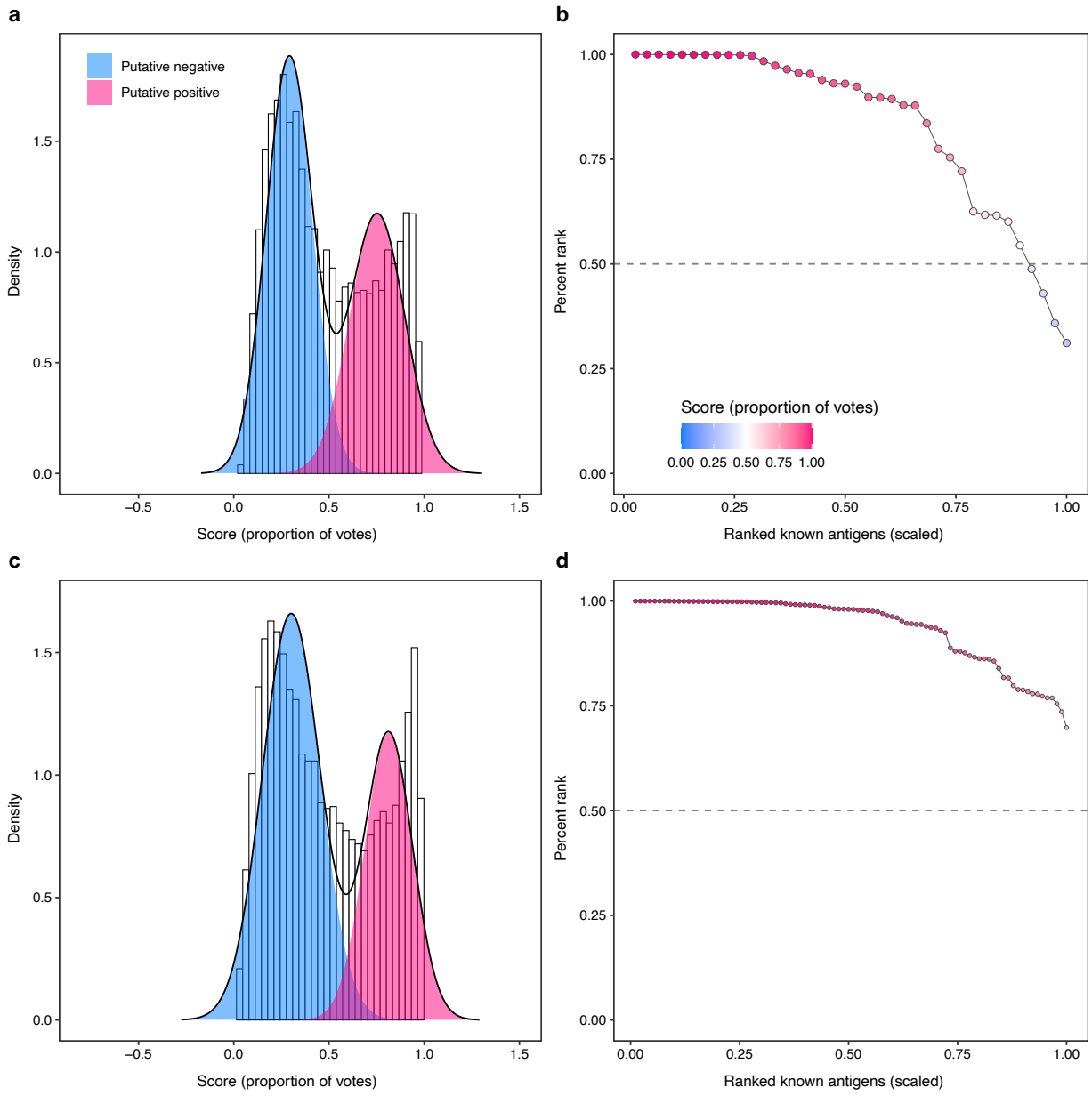


Figure 5.1 Performance of PURF models with the optimal hyper-parameter setting. **a,c** Probability score (proportion of votes) distributions of the *P. vivax* model **(a)** and the combined model **(b)**. The magenta and blue shaded areas are respectively putative positives (antigens) and putative negatives (non-antigens). The black curve was computed using a two-component Gaussian mixture model. The areas under the receiver operating characteristic curves (AUROC) for the *P. vivax* and combined models were 0.99 and 0.995, respectively. **b,d** Evaluation of known antigen scores of the *P. vivax* **(b)** and the combined model **(d)**. Dots represent known antigens. The x-axis shows scaled ranks of the known antigens, and the y-axis denotes percentile ranks (the higher the better) of the known antigens across all proteins in the data set. The respective areas under the curves for the *P. vivax* and the combined models were 0.84 and 0.94.

and trained the model, and computed the probability scores for the remaining known antigens using the trained model. The mean difference in scores calculated by subtracting original model score from the score from the adversarial model, were between ±0.1 for both single-species models and the combined model (Fig. D.5). We observed two modes in the score difference distributions of the *P. vivax* and the combined models. To understand the underlying factors influencing the bimodal distribution, we examined the association between the two distribution modes and either one of the antigen sources, including labeling source (literature, IEDB, or both) and species type (*P. vivax* or *P. falciparum*; applicable to the combined model only) by compiling contingency tables and calculating the odds ratios. The results indicated that there was a significant association between the two distribution modes and labeling source, where the p-values for the *P. vivax* model (p-value $= 2.60 \times 10^{-4}$) and the combined models (p-value $= 1.18 \times 10^{-5}$). However, there was no significant association found between the two modes and species type (p-value $=$ 0.66), suggesting that the model is robust to labeling of different species. Furthermore, the adversarial control experiments showed predicted mean accuracies of 94%, 90%, and 89% for identification of known antigens based on the *P. vivax*, *P. falciparum*, and combined models, respectively, suggesting the robustness of the antigen prediction results.

5.3.3 Effects of heterologous positives and unlabeled proteins on combined model performance

To explore how the addition of *P. falciparum* data improved the accuracy of prediction of both *P. vivax* and *P. falciparum* antigens in the combined model, we further

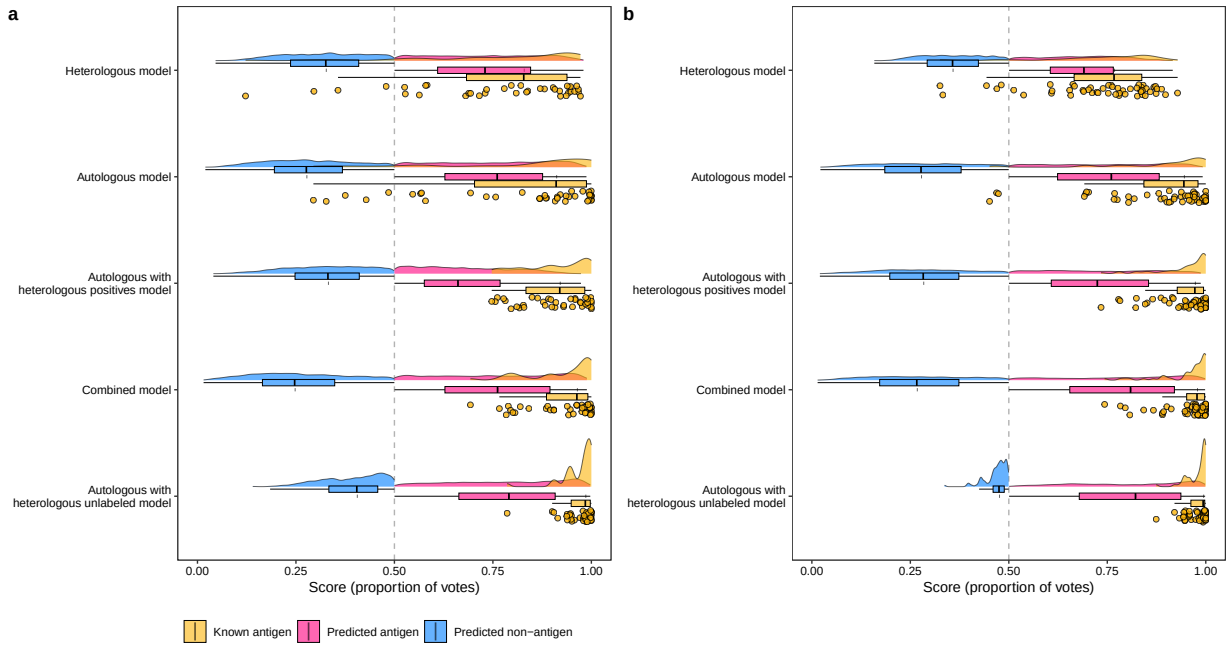


Figure 5.2 Probability score distributions of PURF models. Plots showing the score distributions for *P. vivax* **(a)** and *P. falciparum* **(b)** proteins. The results were generated from PURF models trained on different combinations of autologous and heterologous data. The amber, magenta, and blue colors represent known antigens, predicted antigens, and predicted non-antigens, respectively. Grey vertical dashed lines show the probability score of 0.5. Boxplots show the median with first and third quartiles, with whiskers denoting the extension of the 1.5 interquartile range from the first and third quartiles.

Table 5.2 Different combinations of data from *P. vivax* and *P. falciparum* and their corresponding model types.

Autologous data	Heterologous data	Model type	Note
All *P. vivax* proteins	NA	Autologous model	Heterologous model with respect to *P. falciparum*
All *P. vivax* proteins	*P. falciparum* known antigens	Autologous with heterologous positives model	
All *P. vivax* proteins	*P. falciparum* unlabeled proteins	Autologous with heterologous unlabeled model	
All *P. falciparum* proteins	NA	Autologous model	Heterologous model with respect to *P. vivax*
All *P. falciparum* proteins	*P. vivax* known antigens	Autologous with heterologous positives model	
All *P. falciparum* proteins	*P. vivax* unlabeled proteins	Autologous with heterologous unlabeled model	
Proteins from both species	NA	Combined model	

investigated the individual effects of the positive and unlabeled proteins. In addition to the autologous single-species models and their predictions for heterologous proteins mentioned in the previous section, models were trained by either incorporating heterologous positives or heterologous unlabeled proteins (Table 5.2). The probability score distributions of *P. vivax* and *P. falciparum* proteins were analyzed separately for PURF models that were trained using different combinations of autologous and heterologous data. Compared to autologous and heterologous model predictions, the autologous model with heterologous positives and autologous model with heterologous unlabeled proteins predicted all known antigens correctly (EPR = 1) for both species (Fig. 5.2). However, among these models, the combined model had the largest difference in medians between the predicted antigen and non-antigen groups (Fig. 5.2), indicating the ability of the model to distinguish antigens more clearly from non-antigens among the unlabeled proteins. Moreover, it was observed that in the autologous models including heterologous unlabeled proteins, the predicted autologous non-antigens consistently received higher prediction scores compared to other models, suggesting a possible confounding of antigen identification and species classification (Fig. 5.2). Next, we focused on the labeled positives (known antigens) and

the unlabeled proteins for each species. The percentile ranks of the known antigen predictions from the combined model, the autologous model with heterologous positives, and the autologous model with heterologous unlabeled proteins all had AUCs values >0.9 for both species (Fig. D.6). Antigen predictions of unlabeled proteins from both species (antigen or non-antigen) were further analyzed for their association with their corresponding species (*P. vivax* or *P. falciparum*). When using Cramér's V to assess the strength of the association with species, we found that the single-species and combined models had a relatively weak association, <0.10. In contrast, the autologous model with heterologous positives had a strong association with species (>0.61), and the autologous model with heterologous unlabeled proteins displayed an even stronger association above 0.80 (Table D.1). Together with the observed the score distributions of predicted antigens and non-antigens described above, this suggests that solely adding the heterologous unlabeled proteins from another species may misdirect the model to classify species rather than antigens. Additionally, it was noted that there is a significant relationship (F-statistic test; p-value = 0.012) between the mean tree depth in the model and the proportion of positives in the training data set (Fig. D.7), suggesting labeled data are important for a model to learn the comprehensive patterns among the data.

5.3.4 Analysis of model prediction space and species effect

To gain insights into the antigen prediction of the combined model, we computed the prediction space derived from the tree-based structures within the PURF model. The subsequent visualization revealed distinct clusters of predicted antigens for each species,

whereas the predicted non-antigens appeared within a single, larger cluster irrespective of species (Fig. 5.3). To elucidate the effect of species on the predicted antigens, we conducted hierarchical clustering of the predicted antigens. Starting from the root of the dendrogram, the predicted antigens were iteratively divided into two groups, and the group with the higher mean predicted probability score was selected for the next iteration and continued for four interactions (Fig. D.8). For the first three iterations, the statistical analysis showed a significant association between the two clustering groups and species (χ^2 test, p-values for the four iterations: 1.11×10^{-10}, 6.50×10^{-46}, $<2.23 \times 10^{-308}$, and 0.33). Notably, the third iteration had the strongest association between cluster group and species (Cramér's V = 0.94; 95% CI: 0.93, 0.95), whereas the remaining three iterations demonstrated weaker association strengths (Cramér's V for the first, second, and fourth iterations: 0.09, 0.24, and 0.02). The results indicated that there might be species-specific effects, with predicted antigens having higher scores. Thus, species-specific antigens may be identified with a higher score threshold. To gain a deeper understanding of whether the observed species effects were attributable to differences in amino acid composition of protein sequences in each species, we performed an association analysis between the amino acid frequencies and species. The results, based on Cramér's V, showed there was a weak association of 0.20 across all proteins, with a slight increase in association strength to 0.24 for predicted antigens, and a lower association strength of 0.10 for predicted non-antigens, suggesting that amino acid composition may not be the primary driver of the observed species effect.

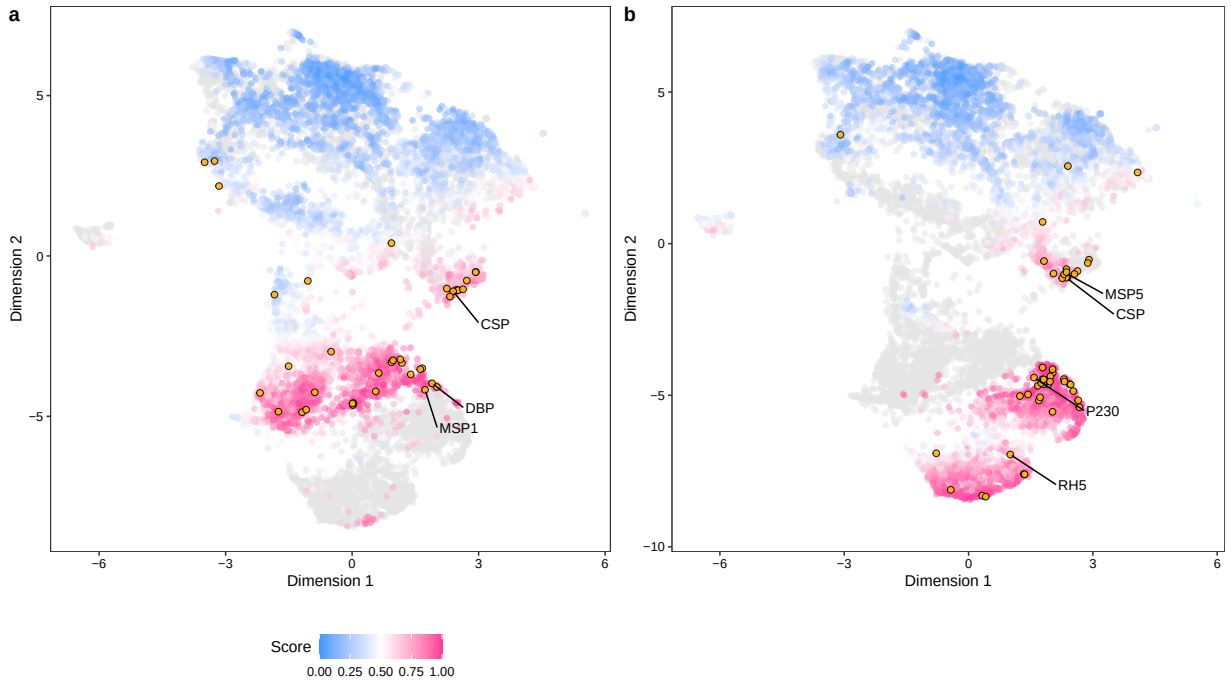


Figure 5.3 Visualization of the prediction space of the combined PURF model. a,b Uniform manifold approximation and projection (UAMP) plots highlighting the *P. vivax* **(a)** and *P. falciparum* **(b)** proteins in the prediction space derived from the Euclidean distance matrix of the combined PURF model. Dots represent proteins in the combined set with *P. vivax* and *P. falciparum* data. Autologous proteins with higher probability scores are shown by darker magenta color and lower scores by darker blue color. Grey dots display *P. vivax* and *P. falciparum* heterologous proteins in **(a)** and **(b)**, respectively. Dots with amber color indicate known antigens for *P. vivax* and *P. falciparum* respectively in **(a)** and **(b)**, with corresponding reference antigens annotated by protein names.

5.3.5 Variables contributing to Plasmodium antigen prediction

To understand the most important variables for *Plasmodium* antigen prediction using the combined models, we conducted a permutation-based variable importance analysis [307]. All four data types, including genomic, immunological, proteomic, and structural data, were represented in the top 10 most important variables (Fig. 5.4a). Among these 10 variables, the following exhibited higher values in known antigens compared to randomly selected non-antigens: secretory signal peptide probability, glycosylphosphatidylinositol (GPI)-anchor specificity score, number of non-synonymous single nucleotide polymorphisms (SNPs), total length of low complexity regions, small amino acid percentage, number of interferon (IFN)-gamma inducing epitopes, and maximum score of Parker hydrophilicity for predicted epitopes (Fig. 5.4a). In contrast, known antigens had a decreased percentage of amino acids with high normalized van der Waals volume (between 4.03–8.08), and a reduced percentage of amino acids with high polarizability (between 0.219–0.409) (Fig. 5.4a). The analysis of variable importance in the *P. vivax* single-species model showed that among the top 10 variables, five were immunological variables associated with B-cell epitopes, IFN-inducing epitopes, and antigenicity predictions (Fig. D.9a). Further analysis of the group variable importance, categorized by data types, indicated that proteomic and immunological variables may contribute more to the accuracy of antigen predictions in the combined model (Fig. 5.4b), which was consistent with the *P. vivax* group variable importance analysis results (Fig. D.9b). The comparison of top 10 important variables independently identified from the two single-species models and the combined model revealed that two variables, namely secretory signal peptide probability and number of non-synonymous SNPs, were

identified by all three models (Fig. 5.5). Both the *P. vivax* and the combined model showed concordance on three additional variables, and the *P. falciparum* and the combined model showed agreement on an additional four variables (Fig. 5.5). We performed comparative analysis to examine the changes in importance values for the top 10 variables determined by the combined model. The results showed that in the combined model, all 10 variables had greater importance values compared to the two single-species models, suggesting the high influence of these variables regarding prediction accuracy in the combined model. When comparing the two single-species models, six variables had higher importance in *P. vivax*, and four variables demonstrated a higher importance value in *P. falciparum* (Fig. D.10), indicating the contributions of both species in the combined model.

5.3.6 Characterization of top vaccine antigen candidates

From the combined model, 190 top candidate antigens were selected based on their probability scores, where the score threshold was set above the median ranking of the 90 known antigens labeled positive in the combined PURF model. The top candidates comprised 35 proteins from *P. vivax* and 145 proteins from *P. falciparum*. We performed hierarchical clustering analysis to further characterize the top candidates. The Silhouette [308] method identified two distinct groups, and the Elbow (or total within sum of square) method identified three groups in the dendrogram. We further visualized the groups in the prediction space and found that one of the two groups identified by the Silhouette method exclusively consisted of 35 candidate antigens from *P. vivax* (Fig. D.11a, group 1 in blue). The other group was further divided into two based on the Elbow method, and one

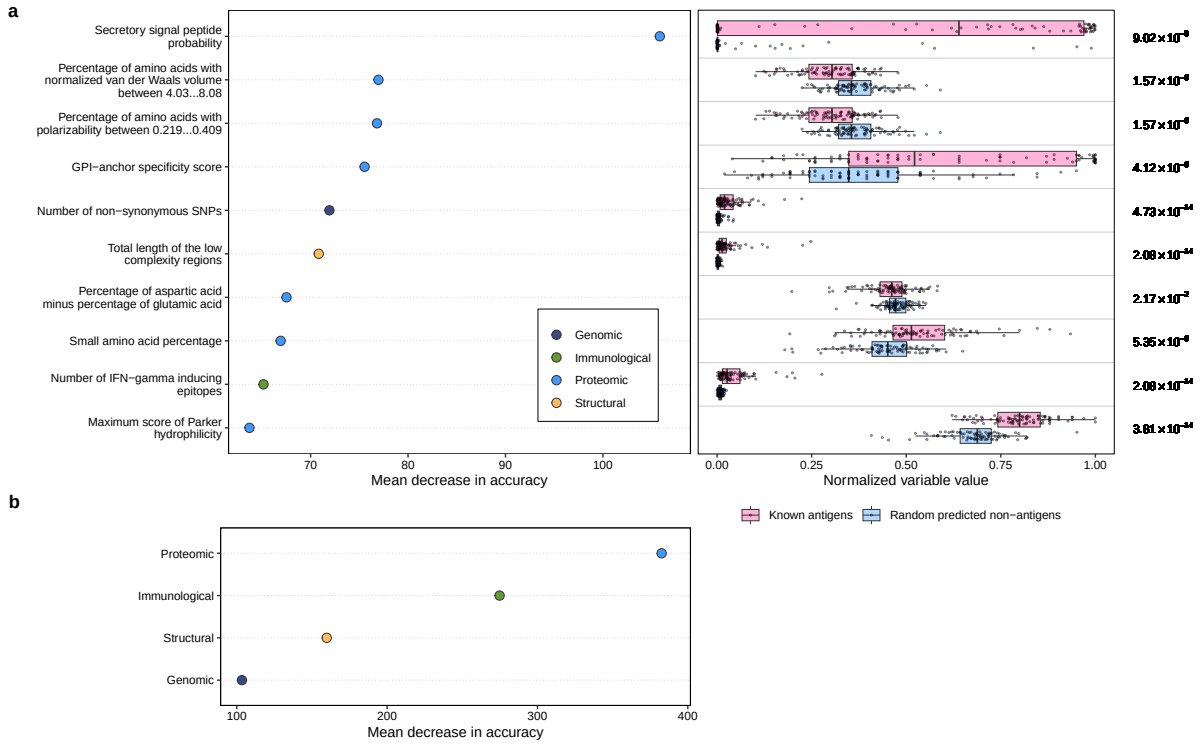


Figure 5.4 Model interpretation of the combined PURF model on the prediction of known antigens. **a** Top 10 important variables computed using permutation-based variable importance analysis on the tree-based PURF model. The left panel shows variable importance values in terms of mean decrease in prediction accuracy after variable permutation. The accuracy was calculated for the 90 known antigens from both Plasmodium species. Variables are categorized into genomic (dark blue), immunological (green), proteomic (blue), and structural (amber) data types. The right panel displays corresponding variable values normalized to range between 0 and 1. Magenta dots represent the 90 known antigens, and the blue dots show 90 predicted non-antigens that were randomly selected. Boxplots convey median with first and third quartiles, and the whiskers indicate the 1.5 interquartile range extended from the first and third quartiles. Two-sided Mann–Whitney tests were conducted with p-values adjusted using the Benjamini–Hochberg procedure, and p-values are shown on the right of the panel. **b** Permutation-based group variable importance analysis. Variables were grouped by data types and permuted together to calculate the mean decrease in accuracy across all trees in the model.

subcluster contains one candidate antigen from *P. vivax* and 38 candidate antigens from *P. falciparum* (Fig. D.11b, group 3 in orange). Gene ontology (GO) enrichment analysis was performed to identify significantly enriched GO terms within the three clusters. Group 1, containing only *P. vivax* candidate antigens, was associated with intracellular activities and locations (Fig. D.12a). Group 2, comprising nine *P. vivax* and 107 *P. falciparum* antigens, had enriched GO terms related to immunological processes, host-pathogen interactions, and cell membranes (Fig. D.12b). Group 3 was characterized by its association with integral components of the membrane as well as the nucleus (Fig. D.12c). Moreover, there were five and four *P. vivax* candidates respectively in group 1 and group 2 not having orthologs in *P. falciparum*, indicating potential candidates for species-specific vaccine. A summary table was generated for the top candidate antigens, providing detailed information regarding gene products, the closest known antigens, and the respective species corresponding to the known antigens. Interestingly, for the nine *P. vivax* candidates in group 2, six were closest in the variable space to known antigens from *P. falciparum*, demonstrating that the inclusion of *P. falciparum* data aided in the identification of potential vaccine candidate antigens for *P. vivax*.

5.4 Discussion

Proteomes are limited in size, and this is often the case in other machine learning problems. The inclusion of external data, such as crowdsourcing and synthetic data generation, has become one of the strategies to improve machine learning models [309]. Here, we showed that augmenting the *P. vivax* training data set with the *P. falciparum* pro-

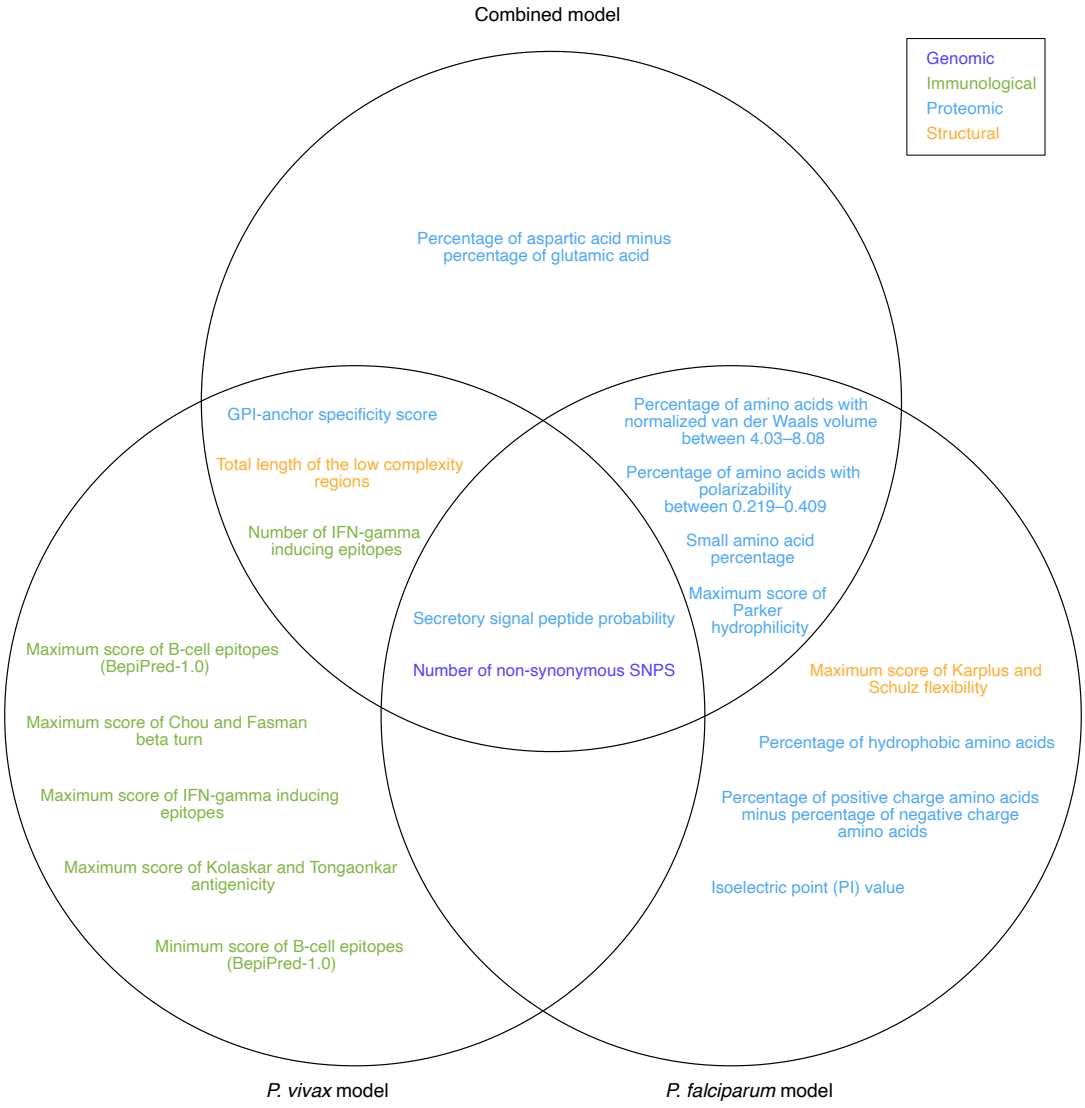


Figure 5.5 Venn diagram of top 10 important variables from different PURF models. Top important variables were identified separately from the combined, *P. vivax*, and *P. falciparum* models. Variable names are colored by the corresponding categories.

teome assisted in identification of antigens for both species. We further decomposed the effect of adding heterologous species data with the inclusion of either known antigens or unlabeled proteins. We demonstrated that inclusion of both heterologous known antigens and unlabeled proteins is important in improving the model performance. We also noticed that adding the heterologous unlabeled proteins alone may potentially transform the antigen identification problem into a species classification problem, leading to identification of proteins based on variables relevant to differentiating between *P. vivax* and *P. falciparum* rather than identification of proteins representing antigens from either species. We conducted a variable importance analysis to understand which protein variables have the greatest influence on antigen identification in the final combined model, and observed that variables related to secretory signal peptides and GPI-anchors, hallmarks of surface proteins exposed to the host immune response, were the best predictors of antigens. Candidate antigens with the highest prediction scores clustered into three groups corresponding to intracellular activities, immune responses, and cellular membrane along with microtubules, respectively. Moreover, one of the closest reference antigens to group 1 is *P. vivax* MSP1, and *P. falciparum* P230 is the closest reference antigen to all three groups.

Although *P. falciparum* is responsible for most malaria cases and deaths, *P. vivax* is the most geographically pervasive species, predominating in regions of Southeast Asia and South America. This wider geographic distribution and the unique biological properties of *P. vivax* that hinder elimination highlight the need for vaccines against this species. Most vaccine development efforts have focused on *P. falciparum* and have led to one WHO-recommended vaccine (i.e., Mosquirix^{TM} (RTS,S/AS01) [310]), as well as other promising vaccine candidates, such as the R21/MatrixM vaccine [311]. However, challenges remain

for vaccine development against other human malaria parasites, such as *P. vivax*. To date, there have been only a few *P. vivax* vaccine candidates evaluated in clinical trials, including those targeting the circumsporozoite protein (PvCSP), Duffy-binding protein (PvDBP), and ookinete surface protein (Pvs25) [312]. Moreover, *P. vivax* cannot be maintained in continuous *in vitro* culture, posing a challenge for identification and preclinical evaluation of vaccine candidates [313]. Reverse vaccinology has been applied to malaria vaccine development to identify potential antigens or epitopes, with most studies focusing on *P. falciparum* [298, 299, 314, 315]. One study of *P. vivax* performed computational analyses to identify B-cell epitopes in the merozoite surface protein-9 only [316]. Similarly, in another study, only a set of 39 *P. vivax* merozoite proteins were investigated, however, the selection criteria in terms of protein properties were not explicitly specified [317]. By employing a machine learning-based reverse vaccinology approach, we explored the whole proteome of *P. vivax* and a large number of protein properties. By including data from *P. falciparum*, we improved the single-species model trained solely on *P. vivax* data and identified potential *P. vivax* vaccine antigen candidates *in silico*. Although future work would include experimental validation of the top antigen candidates, our approach helped facilitate the selection of candidate antigens compared to traditional approaches that could be time-consuming. To our knowledge, none of the *P. vivax* candidate antigens in group 1 have been identified before, and only two out of nine *P. vivax* candidate antigens in group 2 were discussed in the literature, with one being a potential invasion-related ligand (PVP01_0534300) [318] and the other predicted as a secretory protein (PVP01_0948700) [319]. The remaining unknown *P. vivax* antigen candidates are mostly annotated as putative proteins, with functions related to accessibility to the host immune system, such as TRAP-like protein,

surface protein P113, and secreted ookinete protein.

In this study, various bioinformatics tools have been developed to compute protein properties, which are essential in reverse vaccinology [320]. To be recognized by the host immune system, it is critical to include indicators showing that the proteins are exposed to the extracellular environment [297, 320]. In our study, among all 272 protein variables, secretory signal peptide probability, computed using SignalP [321], was ranked as the most important variable in model prediction for the combined and single species models. The results also showed the importance of non-synonymous SNPs in antigen prediction. Such non-synonymous polymorphisms are common in highly immunogenic, know antigens that have evolved diversity under immune pressure. Although non-synonymous SNPs are important predictors of antigens based on the known antigens used to train the models, genetic diversity can also contribute to vaccine escape, which has posed a problem for malaria vaccines [291] and vaccine against other pathogens. Thus, further filtration of the predicted antigens can be applied to obtain potentially less immunogenic, but more genetically conserved candidate antigens, whose immunogenicity may be improved using an adjuvant. The *P. vivax* model in this study also identified more immunological variables among the most important predictor variables compared to the *P. falciparum* model. This result could stem from differences in amino acid composition between the species and the extremely AT-biased *P. falciparum* genome [322], possibly affecting the quality of epitope predictions. Finally, for the candidate antigen groups containing *P. vivax* antigens, group 1 was associated with intracellular activities, and group 2 proteins had gene ontology terms consistent with protein exposure to the immune system, including host cell plasma membrane, infected host cell surface knob, and integral component of membrane. As antigens

need to be exposed to be recognized by the immune system [292, 323], *P. vivax* proteins in group 2 may be better candidates for inclusion in a vaccine. Additionally, there were four *P. vivax* proteins in group 2 without corresponding *P. falciparum* orthologs, indicating the model did not identify antigens based purely on protein sequence homology, as well as demonstrating the selection of potential species-specific vaccine targets.

In this study, data from a well-characterized species, *P. falciparum*, was included in machine learning models to inform vaccine antigen identification of a less well-characterized species, *P. vivax*, with reference antigens from both species being utilized to instruct the selection of *P. vivax* candidate antigens. The approach described here identified and prioritized candidate antigens from the *P. vivax* proteome, of which about 78.9% are proteins with putative or unknown functions. In addition to *P. falciparum* and *P. vivax*, clinical malaria cases have been reported for *P. ovale* and *P. malariae* [324–327], and more recently, *P. knowlesi*, a simian *Plasmodium* species that causing an increased incidence of human clinical infections in Malaysia and areas of Southeast Asia [328–331]. The genomes of these minor *Plasmodium* species have also been sequenced, and the machine learning-based analytical methodology developed in this study can also be applied to identify vaccine candidate antigens for these species. Beyond malaria, our approach can be applied to other emerging pathogens having few known antigens, where data from a related, well-studied species can contribute to improved antigen identification by machine learning models.

5.5 Methods

5.5.1 Data collection

Protein sequences from the *P. vivax* P01 and *P. falciparum* 3D7 strains were extracted from PlasmoDB [332] release 45 (2019-09-05) and release 43 (2019-04-25), respectively. Proteins with stop codons or "X" symbols in their sequences or those derived from pseudogenes were filtered out. Selenocysteines were replaced with cysteines to support downstream bioinformatic analyses. The resulting 6,491 *P. vivax* and 5,393 *P. falciparum* proteins were subsequently analyzed. The protein variables, to be used as input for the machine learning model, were gathered either from public databases or analyzed using various bioinformatics programs, as detailed below. The resulting 272 variables were categorized into four groups: genomic, immunological, proteomic, and structural. These variables were stored in an in-house database [300] (MariaDB version 10.3.22, `https://mariadb.com/`) for facile data manipulation.

The genomic variables included data related to single nucleotide polymorphisms (SNPs), which were analyzed using whole genome sequencing and were directly downloaded from the genetic variation section in PlasmoDB [332] (235 for *P. vivax* and 365 for *P. falciparum*). Various measurements of SNPs were considered, such as the total number of SNPs, the numbers of non-synonymous, synonymous, nonsense, and non-coding SNPs, the ratio of non-synonymous to synonymous SNPs, and the number of SNPs per kilobase of the coding sequence. Immunological variables comprised predictions of various epitopes such as T-cell epitopes [333], B-cell epitopes [334–336], cytotoxic T-cell epitopes [337],

chemokine inducer epitopes [338, 339], and transporter associated with antigen processing (TAP) binding peptides [340]. Additionally, they contained major histocompatibility complex (MHC) class I epitopes [341], MHC class II epitopes [342], as well as assessments of antigenicity [343] and immunogenicity [344]. The epitope-related data were summarized based on numbers, maximum, mean, and minimum scores of epitopes within the protein sequence. Proteomic variables consisted of predictions regarding subcellular localization [345], malarial adhesins/adhesin-like proteins [346], and a set of physicochemical properties [347–349], such as length, weight, isoelectric point, percentage of hydrophobic amino acids. Further, predictions of glycosylphosphatidylinositol (GPI)-anchored proteins [350], signal cleavage [351], protein solubility [352], N-linked or O-linked glycosylation sites [353], and similarity to human proteins [354] were also included. Structural variables contained transmembrane helix predictions [355], sequence complexity [356], and predictions of beta turn [357], surface accessibility [358], and flexibility [359]. Relevant variables were subsequently integrated to construct additional variables. For instance, epitope predictions were combined with transmembrane predictions to derive variables that represent epitopes in outer, inner, or transmembrane regions.

5.5.2 Known antigen labeling

Known antigens were collected from both literature and the Immune Epitope Database (IEDB). The web-based tool Covidence (`www.covidence.org`) was used to select relevant papers or documents using the search terms “malaria vaccine”, “malaria vaccine candidate”, “malaria vaccine antigen”, and “malaria vaccine protein”. To ensure the quality of

known antigen labeling, each of the resulting papers and documents was manually reviewed to select the malaria vaccine candidates for inclusion as known antigens in this study. For IEDB-based known antigen labeling, data were retrieved from PlasmoDB under the immunology section. Known antigens were selected if the PlasmoDB proteins exhibited a similarity score of 97% or higher with GenBank proteins in the IEDB, and if all corresponding epitopes exactly matched the PlasmoDB protein sequence. For *P. falciparum*, 177 known antigens were selected from the literature and 373 from IEDB. For *P. vivax*, 24 known antigens were chosen from literature and 20 from IEDB. The known antigens for *P. falciparum* were determined by the intersection of the two sources, containing 52 known antigens. To get a comparable number of known antigens for *P. vivax*, the union of the two sources was computed, resulting in 38 known antigens.

5.5.3 Machine learning data assembly and data combinations

Protein variable data for *P. vivax* were retrieved from our in-house database, with antigen labels appended as an additional column. In this label column, known antigens were assigned a value of one, and the remaining proteins were marked with a value of zero. The data from *P. vivax* and *P. falciparum* were merged to train a combined model. Additionally, two variants of combined data were created by incorporating subsets of data from each species. First, data containing both autologous proteins and heterologous positives were generated either by combining *P. falciparum* known antigens with the *P. vivax* data set or vice versa. Second, data comprising autologous proteins and heterologous unlabeled proteins were obtained either by adding *P. falciparum* unlabeled proteins to *P. vivax* data

set or vice versa. Refer to Table 5.2 for a detailed overview of models and their corresponding input data.

5.5.4 Positive-unlabeled random forest training

The positive-unlabeled random forest (PURF) algorithm [301] has been optimized and tailored specifically to tackle the antigen identification problem [300]. PURF has the advantages inherited from conventional random forest, such as resilience to errors, insensitive to outliers, and high predictive power. Given the scarcity of known antigens, we enforced learning by implementing a variable space weighting process. In this process, known antigens were weighted based on the variable space, thereby enhancing their representations. Specifically, the center point of the known antigens in the variable space was computed and the Euclidean distances from each known antigen to this center point were scaled into integer values ranging between 1 and 10. Through this approach, known antigens were duplicated based on the integer weights, resulting in a total of 83 known antigens labeled as positive in the *P. vivax* data set. Regarding the combined data set, which includes known antigens from both *P. vivax* and *P. falciparum*, the total number of known antigens was 181.

PURF models, each composed of 100,000 trees, were independently trained on the *P. vivax* and combined data sets, including the variants of the combined data. The ensemble constituent filtering procedure was subsequently applied to the trained PURF models to further prioritize the top-scored unlabeled proteins, guided by the selected reference antigens. The reference antigens for *P. vivax* included the circumsporozoite protein (CSP,

PVP01_0835600.1-p1), the Duffy binding protein (DBP, PVP01_0623800.1-p1), and the merozoite surface protein-1 (MSP-1, PVP01_0728900.1-p1). For P. falciparum, the reference antigens were the circumsporozoite protein (CSP, PF3D7_0304600.1-p1), the merozoite surface protein-5 (MSP-5, PF3D7_0206900.1-p1), the transmission-blocking target antigen s230 (P230, PF3D7_0209000.1-p1), and the reticulocyte binding homolog 5 (RH5, PF3D7_0424100.1-p1). Briefly, any trees that either lacked all reference antigens in the out-of-bag set or incorrectly predicted the reference antigens in the out-of-bag set were discarded. For the *P. vivax* model, the resulting number of trees was 93,102, and for the combined model, 86,254 trees remained after the filtering process.

5.5.5 Positive-unlabeled random forest evaluation

To optimize the *P. vivax* and the combined models, a fine-tuning process on the positive level hyper-parameter was conducted. The positive level describes the prior proportion of potential antigens in the proteome. The process involved training the models across a spectrum of positive level values, ranging from 0.1 and incremented by 0.1 until 0.9. The models were then evaluated with two defined criteria [300]. The first criterion relied on the probability score distribution generated from the PURF model. A two-component Gaussian mixture model was utilized to estimate the putative true and false positive rates from the distribution, and the area under the receiver operating characteristic (AUROC) curve was subsequently computed. The second criterion examined the percentile ranks of the known antigens within the entire proteome based on the probability scores, and the explicit positive recall (EPR) [303,305] was calculated as well. The percentile ranks were also

visualized with respect to the antigens ranked by probability scores, and the area under the curve (AUC) was computed. Based on the two criteria, a positive level of 0.5 was selected for both the *P. vivax* and the combined models.

5.5.6 Adversarial control analysis

To assess and validate the robustness of the models, adversarial controls [306] were generated by changing the positive label to unlabeled for each known antigen in turn, excluding the reference antigens. In each iteration, after assigning a zero value to a known antigen, variable space weighting was applied to the altered data set. Subsequently, a model was trained, and the trees in the model were filtered using the ensemble constituent filtering procedure. Out-of-bag probability scores were then computed for the remaining known antigens that did not have their labels removed. As the reference antigens were excluded from the analysis, there were 35 adversarial control models generated for the *P. vivax* data set, and 83 for the combined data set. The resulting data were further analyzed in two ways. First, the scores of known antigens in each adversarial control had the baseline scores derived from the original models subtracted from them, following which the mean difference was calculated. The mean differences in scores from all adversarial controls were then compared across the *P. vivax*, *P. falciparum*, and the combined data sets. Second, for each adversarial control model, the accuracy of the remaining known antigens was computed using a probability score threshold of 0.5. The results of all adversarial control models were summarized as a mean and standard deviation, and compared across the *P. vivax*, *P. falciparum*, and the combined data sets. To further investigate the two

modes observed in the distribution of differences in scores, a two-sided Fisher's exact test was performed using the function *fisher.test* in R stats, version 4.2.3. This test was utilized to compare the two modes with either species types or known antigen source, where the odds ratio and p-value were subsequently calculated.

5.5.7 Comparison of models trained with different data combinations

As described in the above section in Methods, variants of combined data were generated by integrating portions of data from *P. vivax* and *P. falciparum*. Multiple PURF models were compared to understand the impact of autologous and heterologous data. For each species, these models included the single species model (which can be viewed as either an autologous or heterologous model, depending on the species for which the prediction score were generated), the combined model, the autologous with heterologous positives model, and the autologous with unlabeled model. Models were then compared based on scores of known antigens and predictions of unlabeled proteins. For each species, the score distribution for each model was visualized by plotting the densities and boxplots of the known antigens, along with predicted antigens and non-antigens, as shown in Figure 5.2. Known antigens were further quantified using EPR and percentile ranks, where the AUC values were computed. Subsequently, the predictions for unlabeled proteins were compared by examining the association between antigen predictions and species types using a χ^2 test, utilizing the function *chisq.test* in R *stats*, version 4.2.3. The strength of the association was further analyzed using Cramér's V, using the function *cramerV* in the R package *rcompanion*, version 2.4.30 [360]. The 95% confidence interval of Cramér's V was calculated using

a bootstrap approach with 1,000 replications. Additionally, the mean tree depths across all trees in each type of model were calculated and compared to the proportion of positives in the training data set. A linear model was fitted using the *lm* function in R *stats*, version 4.2.3. The dependent variable mean tree depth was $\log_2$-transformed, and the *logit* function was applied to the independent variable proportion of positives. The regression function was $log(y) = 6.8 + 0.87 \cdot logit(x)$, and the adjusted R^2 was 0.7. The p-value derived from the F-test was 0.01.

5.5.8 Model interpretation of the combined model

A proximity matrix quantifying closeness between proteins in the model was computed. Specifically, for any pair of proteins in the training data set, the proximity value was calculated by counting the times the pair was in the out-of-bag set and ended in the same leaf across all trees. The proximity value was then normalized by dividing the value by the number of times the pair of proteins was in the out-of-bag set. The proximity value ranges from zero to one, which was then transformed into a Euclidean distance by subtracting its value from one, resulting in a dissimilarity matrix. To further reduce the dimensions of the dissimilarity matrix, multidimensional scaling, also known as principal coordinates analysis, was applied using the *cmdscale* function in R *stats*, version 4.2.3. For enhanced visualization, a two-dimensional representation was constructed by employing the uniform manifold approximation and projection (UMAP) [361] method with two components.

5.5.9 Clustering and amino acid composition analyses of model predictions

To further explore the separation of predicted antigens observed in the aforementioned visualization of the model predictions based on the dissimilarity matrix, the Ward's hierarchical agglomerative clustering method [362] was utilized to analyze the data, using the R *stats* function *hclust*, along with the "ward.D2" implementation [363]. A dendrogram was generated through the hierarchical clustering analysis, and an iterative process was initiated to segregate predicted antigens into groups. In each iteration, the predicted antigens were split into two groups according to the dendrogram. To assess the association between the groups and the species types, a χ^2 test and Cramér's V were computed. A sub-dendrogram was then created from one of the two groups having the higher average probability score, and the iterative process continued. A total of four iterations were generated, yielding χ^2 p-values of 1.11×10^{-10}, 6.50×10^{-46}, $<2.23 \times 10^{-308}$, and 0.33. The Cramér's V values, along with the 95% confidence interval (CI) values, were 0.09 (95% CI: 0.06, 0.11), 0.24 (95% CI: 0.21, 0.27), 0.94 (95% CI: 0.93, 0.95), and 0.02 (95% CI: 0.00, 0.08). The resulting groups were further visualized on the UMAP representation of the dimension-reduced dissimilarity matrix, as detailed previously in the model interpretation section. Finally, to evaluate the association between amino acid composition and species types, an association analysis was conducted to analyze the two variables in three groups: the whole proteome, the predicted antigens, and the predicted non-antigens. For the whole proteome group, the frequencies of the 20 amino acids were independently computed for the proteomes of *P. vivax* and *P. falciparum*. For the predicted antigen and non-antigen groups, the amino acid frequencies of the two species were computed for each group as

well. In each comparison group, the association between the amino acid frequencies and the species types was evaluated using a χ^2 test and Cramér's V.

5.5.10 Variable importance analysis

To understand the important variables in model predictions, a permutation-based variable importance analysis was performed on the trained PURF models using the method proposed by Breiman [307]. To explore patterns in variable values, the values were first normalized to range between 0 and 1 for each variable. A set of predicted non-antigens with the same size as the known antigens in the training data set was then randomly selected. Next, to compare the variable values between the known antigens and the predicted non-antigens that were randomly selected, a two-sided Mann–Whitney test was performed. The Benjamini–Hochberg [364] method was applied to adjust the p-values for comparisons of all 272 variables. Additionally, a permutation-based group variable importance analysis was conducted for the four variable groups: genomic, immunological, proteomic, and structural. The process was the same as described earlier except that variables within the same group were permuted together to compute their collective impact on prediction accuracy. Finally, top 10 important variables in the *P. vivax*, *P. falciparum*, and combined models were compared using a Venn diagram. To understand how the top 10 important variables of the combined model influence the prediction accuracy of the known antigens in the single-species models, importance values of this identical variable set were compared across all three models.

5.5.11 Clustering of top candidate antigens

Top candidate antigens were selected based on a probability score threshold, where half of the known antigens were scored above this threshold, resulting in 190 potential vaccine antigen candidates. A dissimilarity matrix of these top candidate antigens, computed from the tree-based structures in the model, was then analyzed using Ward's hierarchical agglomerative clustering method [363], as previously described. The number of clustering groups was determined using Gap statistic with the Tibshirani criterion [365, 366], Silhouette [308], and Elbow (or total within sum of square) methods. These methods identified 1, 2, and 3 groups, respectively. As a result, top candidate antigens were clustered into either two or three groups. The groups were then visualized on a two-dimensional UMAP representation. For the clustering with two groups based on the Silhouette method, group 1 contained 35 *P. vivax* and 0 *P. falciparum* candidate antigens, while group 2 had 10 *P. vivax* and 145 *P. falciparum* candidates. For the clustering with three groups based on the Elbow method, group 1 contained 35 *P. vivax* and 0 *P. falciparum* candidates, group 2 comprised 9 *P. vivax* and 107 *P. falciparum* candidates, and group 3 consisted of 1 *P. vivax* and 38 *P. falciparum* candidates. Orthologs were identified through searching in PlasmoDB, which was based on the data set generated using the OrthoMCL algorithm [367, 368]. A summary table was subsequently generated to display information of the top candidate antigens, which include the associated clustering groups, probability scores, gene products (retrieved from PlasmoDB [332], release 62), the closest known antigens and their source, as well as the Euclidean distance (ranging from 0 to 1) to the closest known antigen.

5.5.12 Gene ontology enrichment analysis

To better understand gene ontology (GO) terms associated with the three groups of candidate antigens, an enrichment analysis was conducted using the Python package *GOATOOLS* [369]. The gene ontology terms for *P. vivax* and *P. falciparum* were downloaded directly from PlasmoDB [332], release 62 (2023-03-09). The file containing the directed acyclic graph of gene ontology was retrieved from the Gene Ontology website (`http://geneontology.org/docs/download-ontology/`) [370, 371]. For more conservative results, the argument *propagate_counts* in the function *GOEnrichmentStudyNS* was set as false. GO enrichment analysis was performed with the background proteomes from both *P. vivax* and *P. falciparum* species. The p-values from the multiple Fisher's exact tests were adjusted using the Benjamini–Hochberg [364] method. Enriched GO terms, identified based on the significant cut-off of 0.5, were categorized into biological process, cellular component, and molecular function, and were further visualized for each of the three candidate groups.

Part IV

Appendices

Appendix A: Supplementary Information for Machine Learning-Driven Multifunctional Peptide Engineering for Sustained Ocular Drug Delivery

A.1 Supplementary Notes

Supplementary materials, including the research notebook containing the code of the machine learning algorithms, have been deposited as a compressed folder (∼2.74 GB in total) in the Digital Repository at the University of Maryland (DRUM) with the identifier `https://doi.org/10.13016/0jck-hnnv`. To open the research notebook, download and decompress the folder, go to the subfolder `main_notebook`, and click on `index.html` to open the HTML document in a web browser, or click on `main_notebook.pdf` to open the PDF document. Data generated in this research have been stored in the subfolders `data` and `other_data`. The following shows the descriptions of the data files and their locations.

A.1.1 Machine learning input data sets

The `peptide_variable_descriptions.csv` file contains descriptions of peptide variables calculated in the input data sets for machine learning. The other data files are the machine learning input data sets, including pilot and second melanin binding peptide mi-

croarray data for training classification and regression machine learning models, as well as cell-penetration and cytotoxicity peptide data for training classification models.

`data/peptide_variable_descriptions.csv`

`data/mb_pilot_peptide_array_ml_input.csv`

`data/mb_second_peptide_array_ml_input.csv`

`data/cpp_ml_input.csv`

`data/tx_ml_input.csv`

A.1.2 Machine learning cross-validation results

A nested cross-validation framework was applied in this study, where the inner loop cross-validation is used to select the best performing subset of models, and the outer loop cross-validation is used estimate generalization performance. The `cv_res_statistical_testing.csv` files in the subfolders `outer_1` through `outer_10` correspond to each inner loop cross-validation fold, and in subfolder `whole_data_set` corresponds to the final outer loop cross-validation results. Models were scored and ranked based on multiple metrics. For regression the metrics were mean absolute error, root mean squared error and coefficient of determination (R^2). For classification the metrics were log loss, Matthews correlation coefficient, F_1 (harmonic mean of precision and recall), and balanced accuracy. The cross-metric rank was determined by summing the ranks of the individual metrics. Multiple statistical tests were performed by comparing the metric scores ($n = 10$) of the best model to all other models. Adjusted p-values were reported. The best model and models with no significant difference in all metrics from the best model were included in the files. The

numbers of those competitive models were indicated in the parentheses below. The files containing grid search model parameters were also provided for melanin binding, cell-penetration, and cytotoxicity models.

A.1.2.1 Melanin binding models

`other_data/melanin_binding/neural_network_grid_params.csv`

`other_data/melanin_binding/gbm_grid_params.csv`

`other_data/melanin_binding/xgboost_grid_params.csv`

`other_data/melanin_binding/outer_1/cv_res_statistical_testing.csv` (2 competitive models)

`other_data/melanin_binding/outer_2/cv_res_statistical_testing.csv` (50 competitive models)

`other_data/melanin_binding/outer_3/cv_res_statistical_testing.csv` (15 competitive models)

`other_data/melanin_binding/outer_4/cv_res_statistical_testing.csv` (43 competitive models)

`other_data/melanin_binding/outer_5/cv_res_statistical_testing.csv` (41 competitive models)

`other_data/melanin_binding/outer_6/cv_res_statistical_testing.csv` (49 competitive models)

`other_data/melanin_binding/outer_7/cv_res_statistical_testing.csv` (19 competitive models)

`other_data/melanin_binding/outer_8/cv_res_statistical_testing.csv` (33 competitive models)

other_data/melanin_binding/outer_9/cv_res_statistical_testing.csv (32 competitive models)

other_data/melanin_binding/outer_10/cv_res_statistical_testing.csv (26 competitive models)

other_data/melanin_binding/whole_data_set/cv_res_statistical_testing.csv (31 competitive models)

A.1.2.2 Cell-penetration models

other_data/cell_penetration/neural_network_grid_params.csv

other_data/cell_penetration/gbm_grid_params.csv

other_data/cell_penetration/xgboost_grid_params.csv

other_data/cell_penetration/outer_1/cv_res_statistical_testing.csv (272 competitive models)

other_data/cell_penetration/outer_2/cv_res_statistical_testing.csv (227 competitive models)

other_data/cell_penetration/outer_3/cv_res_statistical_testing.csv (277 competitive models)

other_data/cell_penetration/outer_4/cv_res_statistical_testing.csv (303 competitive models)

other_data/cell_penetration/outer_5/cv_res_statistical_testing.csv (300 competitive models)

other_data/cell_penetration/outer_6/cv_res_statistical_testing.csv (303 competitive models)

other_data/cell_penetration/outer_7/cv_res_statistical_testing.csv (304 competitive models)

other_data/cell_penetration/outer_8/cv_res_statistical_testing.csv (303 competitive models)

other_data/cell_penetration/outer_9/cv_res_statistical_testing.csv (122 competitive models)

other_data/cell_penetration/outer_10/cv_res_statistical_testing.csv (304 competitive models)

other_data/cell_penetration/whole_data_set/cv_res_statistical_testing.csv (300 competitive models)

A.1.2.3 Cytotoxicity models

other_data/toxicity/neural_network_grid_params.csv

other_data/toxicity/gbm_grid_params.csv

other_data/toxicity/xgboost_grid_params.csv

other_data/toxicity/outer_1/cv_res_statistical_testing.csv (193 competitive models)

other_data/toxicity/outer_2/cv_res_statistical_testing.csv (49 competitive models)

other_data/toxicity/outer_3/cv_res_statistical_testing.csv (194 competitive models)

other_data/toxicity/outer_4/cv_res_statistical_testing.csv (74 competitive models)

other_data/toxicity/outer_5/cv_res_statistical_testing.csv (180 competitive models)

other_data/toxicity/outer_6/cv_res_statistical_testing.csv (197 competitive models)

other_data/toxicity/outer_7/cv_res_statistical_testing.csv (159 competitive models)

other_data/toxicity/outer_8/cv_res_statistical_testing.csv (179 competitive models)

other_data/toxicity/outer_9/cv_res_statistical_testing.csv (163 competitive models)

other_data/toxicity/outer_10/cv_res_statistical_testing.csv (153 competitive models)

`other_data/toxicity/whole_data_set/cv_res_statistical_testing.csv` (175 competitive models)

A.1.3 Adversarial control machine learning cross-validation results

To understand whether the whole machine learning procedure (including model selection) has learned meaningful relationships in the data sets, adversarial control models were trained on the data sets with the response variables randomly shuffled, and cross-validation results were reported for the inner loop iterations. Next, a best-performing model was selected in each inner loop cross-validation, and the generalize performance of these top one models were estimated in the outer loop cross-validation. There was no final predictive model trained on the whole data set in this experiment because it is unnecessary to use adversarial control models for future prediction, and thus there was no `whole_data_set` subfolders included. In the file list below, the values in the parentheses following the file names showed the numbers of competitive models filtered based on the statistical analyses of the model performance. See subsection A.1.2 for detailed information. The metadata files containing the grid search model parameters were also included for reference.

A.1.3.1 Melanin binding adversarial control models

`other_data/melanin_binding_adversarial/neural_network_grid_params.csv`

`other_data/melanin_binding_adversarial/gbm_grid_params.csv`

`other_data/melanin_binding_adversarial/xgboost_grid_params.csv`

`other_data/melanin_binding_adversarial/outer_1/cv_res_statistical_testing.csv` (296 competitive models)

other_data/melanin_binding_adversarial/outer_2/cv_res_statistical_testing.csv (277 competitive models)

other_data/melanin_binding_adversarial/outer_3/cv_res_statistical_testing.csv (2 competitive models)

other_data/melanin_binding_adversarial/outer_4/cv_res_statistical_testing.csv (118 competitive models)

other_data/melanin_binding_adversarial/outer_5/cv_res_statistical_testing.csv (285 competitive models)

other_data/melanin_binding_adversarial/outer_6/cv_res_statistical_testing.csv (120 competitive models)

other_data/melanin_binding_adversarial/outer_7/cv_res_statistical_testing.csv (86 competitive models)

other_data/melanin_binding_adversarial/outer_8/cv_res_statistical_testing.csv (66 competitive models)

other_data/melanin_binding_adversarial/outer_9/cv_res_statistical_testing.csv (53 competitive models)

other_data/melanin_binding_adversarial/outer_10/cv_res_statistical_testing.csv (146 competitive models)

A.1.3.2 Cell-penetration adversarial control models

other_data/cell_penetration_adversarial/neural_network_grid_params.csv

other_data/cell_penetration_adversarial/gbm_grid_params.csv

other_data/cell_penetration_adversarial/xgboost_grid_params.csv

other_data/cell_penetration_adversarial/outer_1/cv_res_statistical_testing.csv (82 competitive models)

other_data/cell_penetration_adversarial/outer_2/cv_res_statistical_testing.csv (23 competitive models)

other_data/cell_penetration_adversarial/outer_3/cv_res_statistical_testing.csv (41 competitive models)

other_data/cell_penetration_adversarial/outer_4/cv_res_statistical_testing.csv (24 competitive models)

other_data/cell_penetration_adversarial/outer_5/cv_res_statistical_testing.csv (26 competitive models)

other_data/cell_penetration_adversarial/outer_6/cv_res_statistical_testing.csv (24 competitive models)

other_data/cell_penetration_adversarial/outer_7/cv_res_statistical_testing.csv (28 competitive models)

other_data/cell_penetration_adversarial/outer_8/cv_res_statistical_testing.csv (27 competitive models)

other_data/cell_penetration_adversarial/outer_9/cv_res_statistical_testing.csv (27 competitive models)

other_data/cell_penetration_adversarial/outer_10/cv_res_statistical_testing.csv (34 competitive models)

A.1.3.3 Cytotoxicity adversarial control models

other_data/toxicity_adversarial/neural_network_grid_params.csv

other_data/toxicity_adversarial/gbm_grid_params.csv

other_data/toxicity_adversarial/xgboost_grid_params.csv

other_data/toxicity_adversarial/outer_1/cv_res_statistical_testing.csv (141 competitive models)

other_data/toxicity_penetration_adversarial/outer_2/cv_res_statistical_testing.csv (126 competitive models)

other_data/toxicity_penetration_adversarial/outer_3/cv_res_statistical_testing.csv (118 competitive models)

other_data/toxicity_penetration_adversarial/outer_4/cv_res_statistical_testing.csv (130 competitive models)

other_data/toxicity_penetration_adversarial/outer_5/cv_res_statistical_testing.csv (80 competitive models)

other_data/toxicity_penetration_adversarial/outer_6/cv_res_statistical_testing.csv (78 competitive models)

other_data/toxicity_penetration_adversarial/outer_7/cv_res_statistical_testing.csv (131 competitive models)

other_data/toxicity_penetration_adversarial/outer_8/cv_res_statistical_testing.csv (131 competitive models)

other_data/toxicity_penetration_adversarial/outer_9/cv_res_statistical_testing.csv (136 competitive models)

other_data/toxicity_penetration_adversarial/outer_10/cv_res_statistical_testing.csv (127 competitive models)

A.2 Supplementary Figures

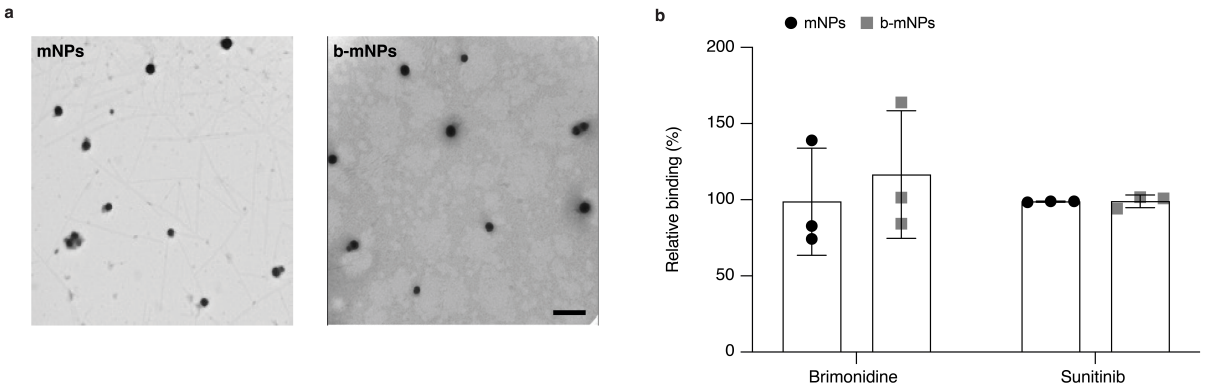


Figure A.1 Characterization of melanin nanoparticles (mNPs) and biotinylated-melanin nanoparticles (b-mNPs). a Representative transmission electron microscopes (TEM) images of mNPs and b-mNPs. The scale bar indicates 600 nm. **b** Relative binding of brimonidine tartrate and sunitinib malate to mNPs (black dots, $n = 3$ per drug group) and b-mNPs (gray squares, $n = 3$ per drug group). Data are shown as mean ± SD. No significant difference in relative binding was observed for either drug (Student's t-tests, two-tailed).

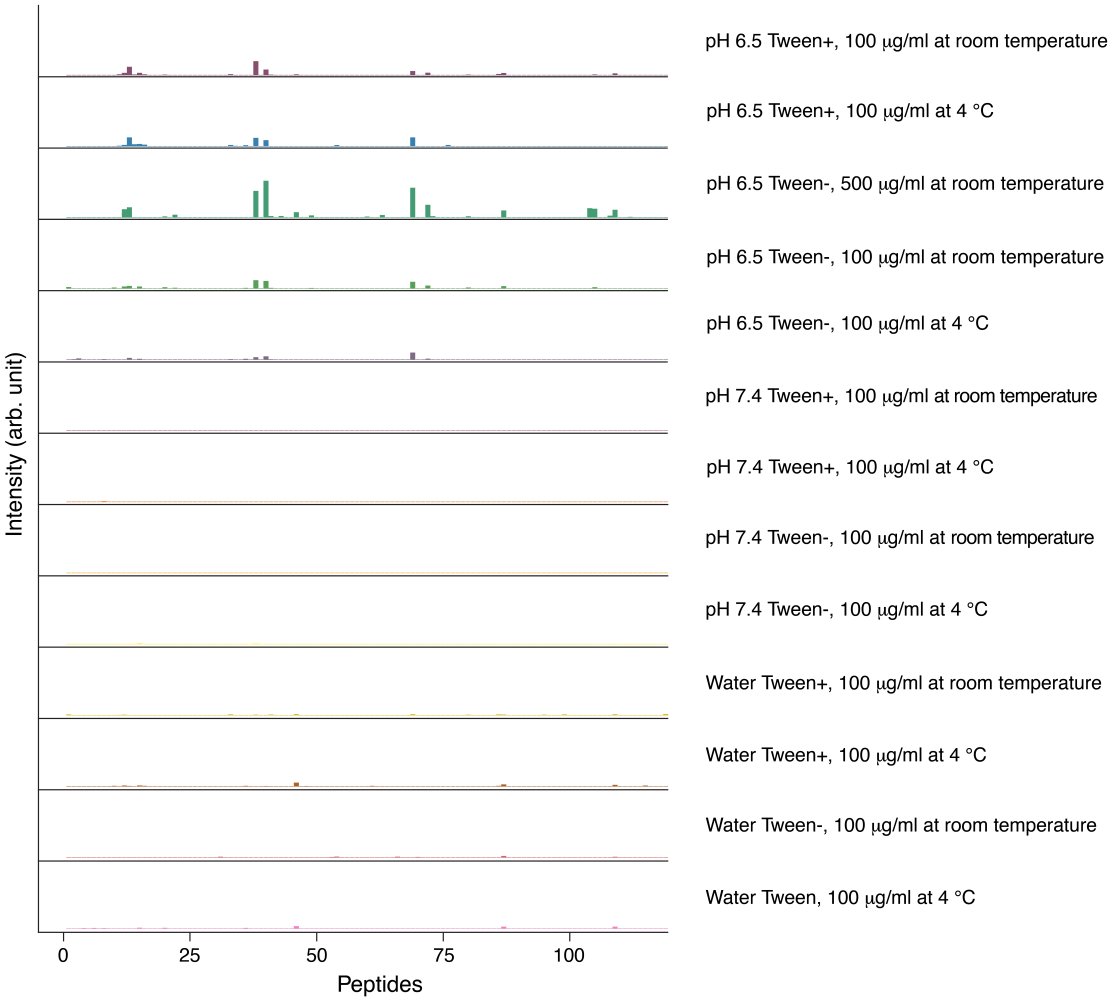


Figure A.2 Interaction profilings of b-mNPs against peptides in the pilot 119 microarray. Sparklines showing fluorescence intensities in varying washing buffer conditions (see **Section 2.5**), plotted on the same scale in arbitrary unit (arb. unit). The first 16 peptides are positive control peptides, and the remaining are 103 random peptides.

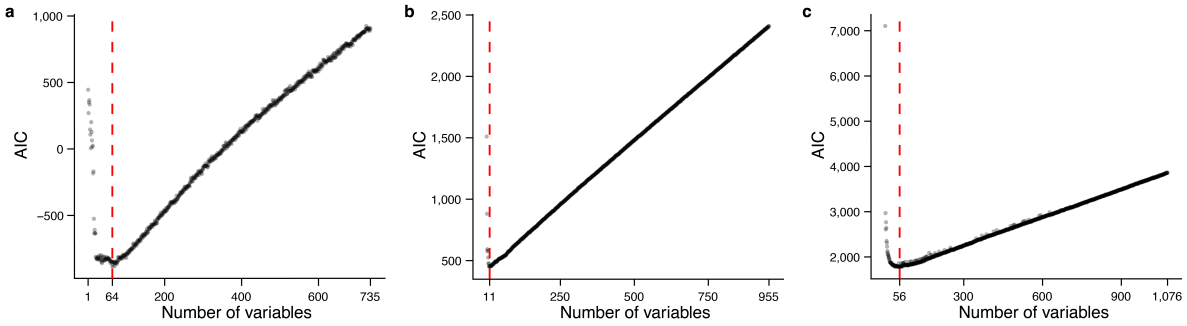


Figure A.3 Variable reduction of peptide data sets with random forests. **a**–**c** show performances of random forest models (black dots) trained on variable subsets ranked using permutation-based variable importance values for melanin binding (**a**), cell-penetration (**b**), and cytotoxicity (**c**) data sets. Akaike information criterion (AIC), a metric that penalizes complex models, was calculated for all models. The red dashed lines indicate the number of variables used in subsequent analyses.

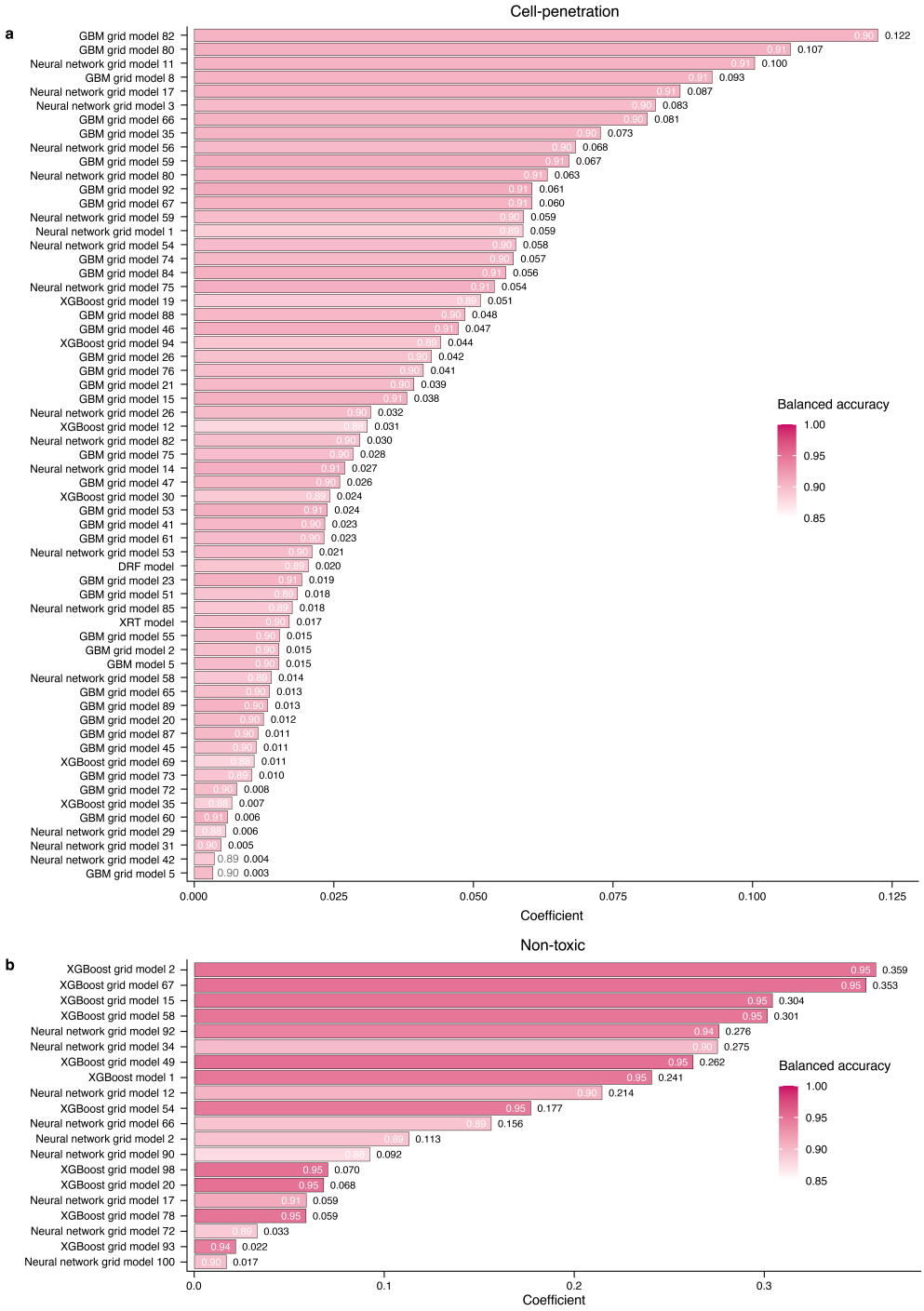


Figure A.4 Base model coefficients in final super learners. **a** Cell-penetration. **b** Cytotoxicity (with non-toxic peptides labeled as positive samples). Balanced accuracy is denoted with color and conveyed as white text on the bars or gray text adjacent bars. Values at the bar ends indicate base model coefficients. See **Sections 2.5** and **A.1** for more details regarding base model hyperparameters.

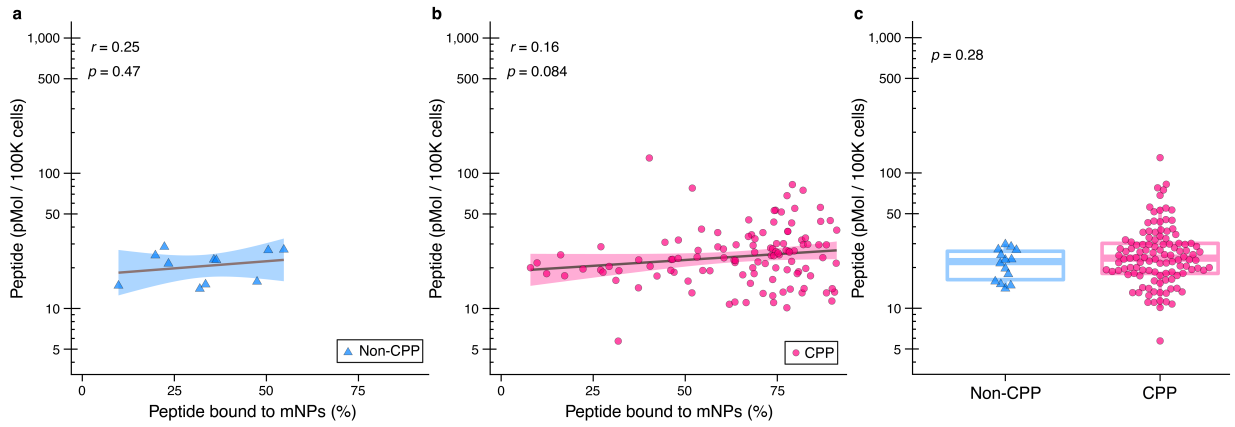


Figure A.5 Comparison of melanin binding and cell-penetration of candidate peptides in non-induced ARPE-19 cells. **a, b** Cyan triangles denote non-cell-penetrating peptides (non-CPP), and magenta dots represent cell-penetrating peptides (CPP). The x-axes indicate melanin binding measured from the mNP assay ($n = 4$), and the y-axes indicate intracellular concentration measured from the cell uptake assay with non-induced ARPE-19 cells ($n = 3$). Black linear trend lines indicate the Pearson correlation relationships, and the shaded areas convey 95% confidence intervals. The correlation coefficient and the corresponding p-values (two-tailed) are shown. **c** Intracellular concentrations of CPP ($n = 113$) and non-CPP ($n = 14$). Box plot indicates median (middle line), 25th and 75th percentiles (box), and the 1.5 × interquartile range (whiskers). The p-value was determined using a Mann–Whitney U test (two-tailed).

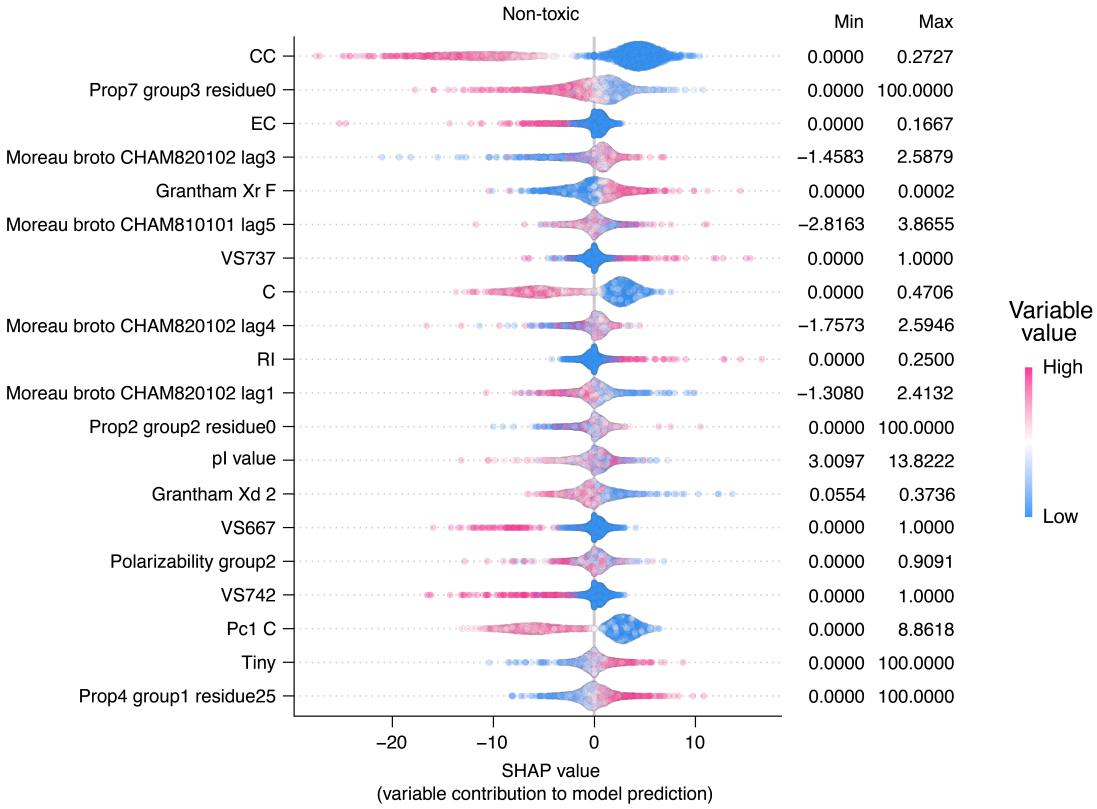


Figure A.6 Cytotoxicity model interpretation. Shapley additive explanation (SHAP) values for the top 20 variables ranked based on the SHAP value range. Dots represent peptides, and color indicates percentile ranks. The minimum and maximum variable values are listed on the right.

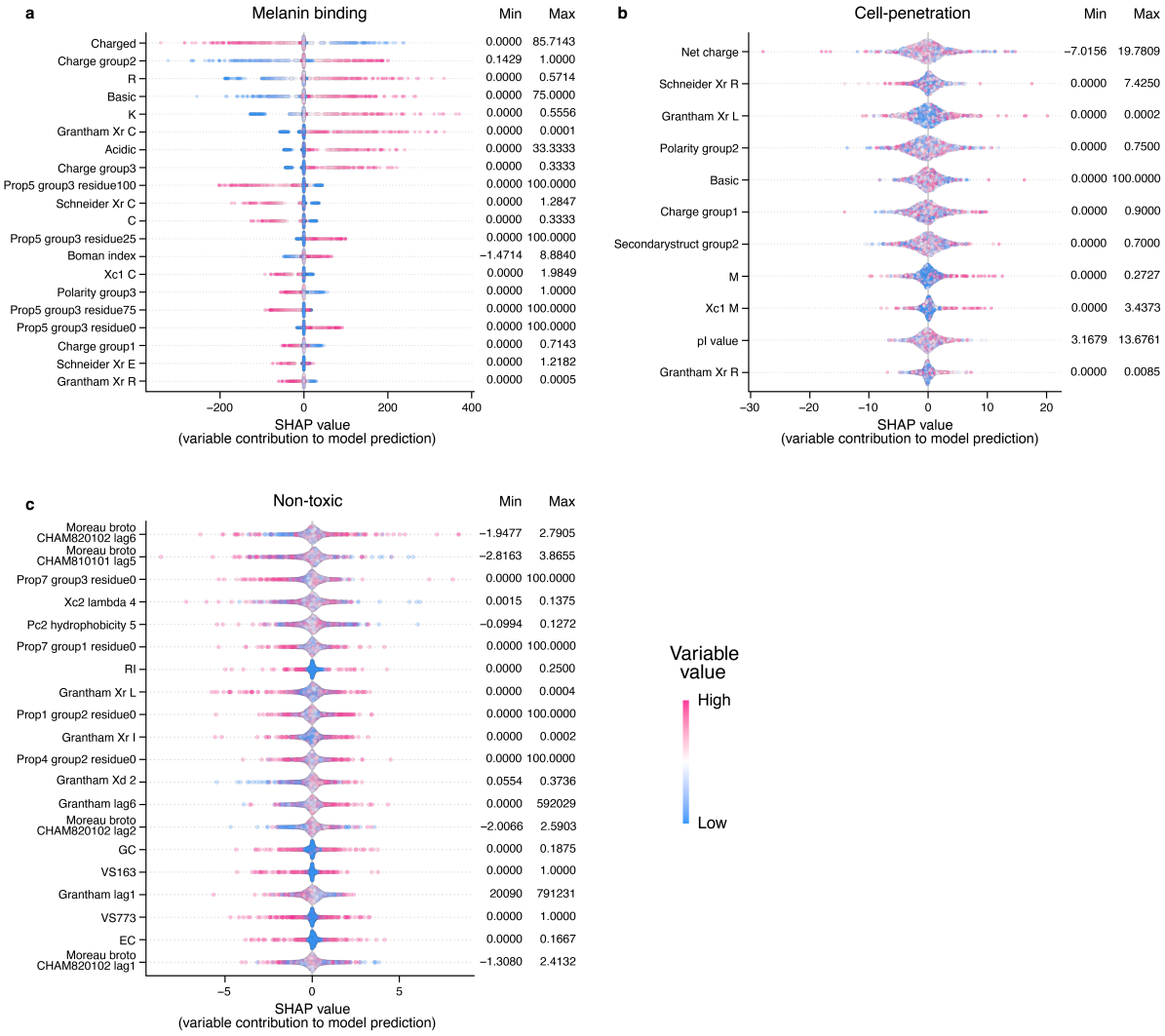


Figure A.7 Variable contributions to the prediction of the adversarial models. Top 20 important variables contributing to (**a**) melanin binding, (**b**) cell-penetration, and (**c**) cytotoxicity adversarial models. The variables were ranked based on the range of the SHAP values. Dots represent peptide samples. The color gradient shows the values of the corresponding variables, calculated as percent ranks. The minimum and maximum variables are shown on the right of each subfigure.

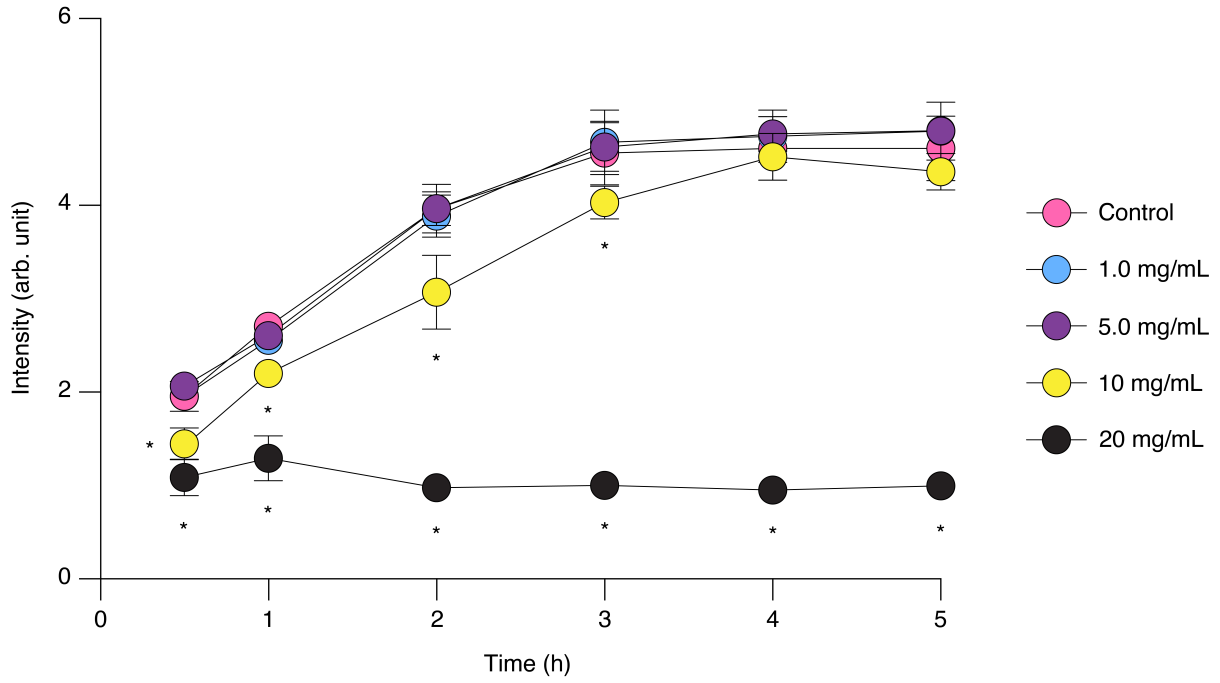


Figure A.8 Cytotoxicity validation of the HR97 peptide. Cell viability assay of the HR97 peptide. ARPE-19 cells were incubated with varying concentrations of HR97 for 12 h, and the cell viability was measured with the PrestoBlueTM HS cell viability system at 0.5, 1, 2, 3, 4, and 5 h after reagent addition ($n = 5$ per group). Data are presented as mean ± SD. HR97 concentration groups (1.0 mg/mL, cyan; 5.0 mg/mL, purple; 10 mg/mL, yellow; 20 mg/mL, black) were compared to the control group (magenta) with Student's t-tests (two-tailed). $*$ denotes $p < 0.05$. Adjusted p-values for 10 mg/mL vs. control at hours 0.5, 1, 2, and 3 were respectively 1.36×10^{-3}, 8.66×10^{-5}, 3.92×10^{-3}, and 1.73×10^{-2}; and those for 20 mg/mL vs. control at hours from 0.5 to 5 were 8.66×10^{-5}, 6.48×10^{-6}, 4.64×10^{-8}, 4.64×10^{-8}, 4.64×10^{-8}, and 4.80×10^{-8}, respectively.

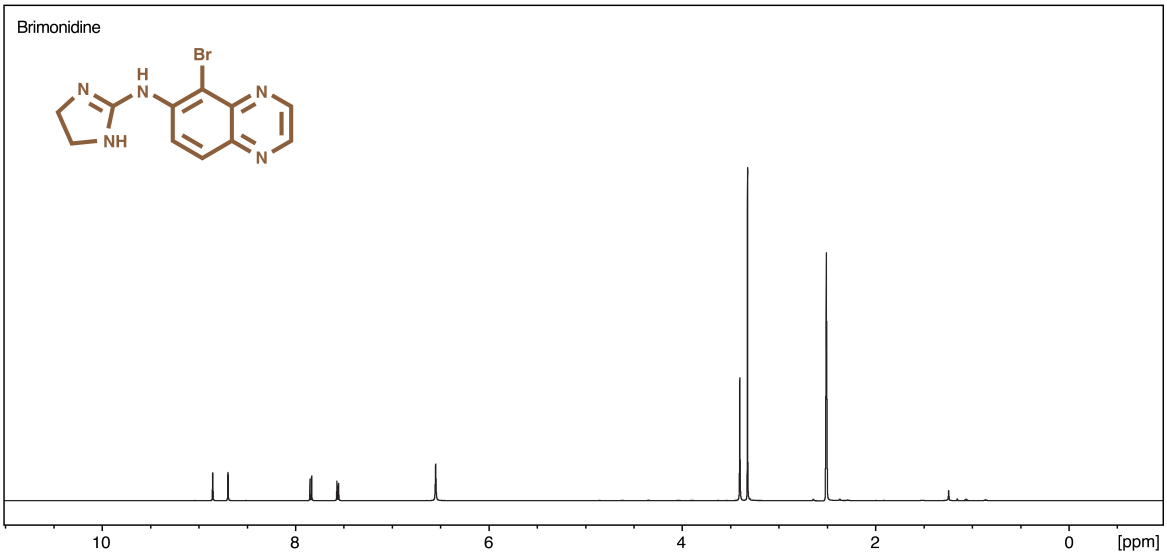


Figure A.9 NMR spectrum of brimonidine. The prep-HPLC retention time (RT) of brimonidine was 5.1 min. The molecular structure of brimonidine is shown in the upper left corner. Peak location and associated information are ^{1}H NMR (500 MHz, DMSO-d_6) 8.84 (d, $J = 5$ Hz, 1H), 8.68 (d, $J = 5$ Hz, 1H), 7.83 (d, $J = 10$Hz, 1H), 7.55 (d, $J = 10$Hz, 1H) 6.54 (s, 2H), 3.40 (s, 4H).

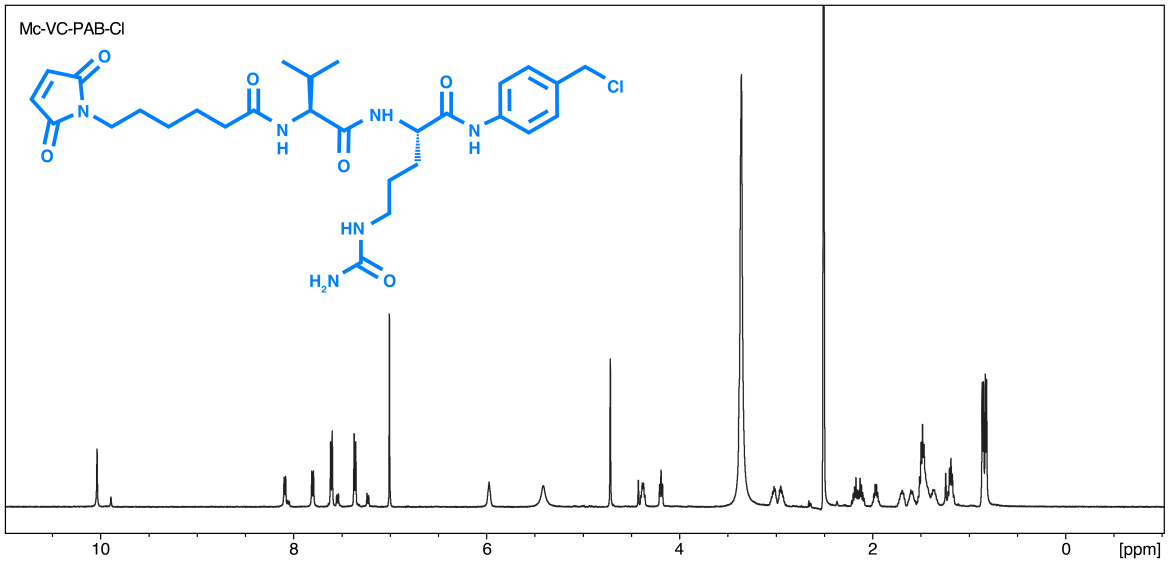


Figure A.10 NMR spectrum of Mc-VC-PAB-Cl (Maleimidocaproyl-L-valine-L-citrulline-p-aminobenzyl chloride). The prep-HPLC RT of Mc-VC-PAB-Cl was 11.4 min. The molecular structure of Mc-VC-PAB-Cl is shown in the upper left corner. Peak location and associated information of Mc-VC-PAB-Cl are ^{1}H NMR (500 MHz, DMSO-d$_6$) 10.03 (s, 1H), 8.08 (d, J = 7 Hz, 1H), 7.80 (d, J = 8.5 Hz, 1H), 7.60 (d, J = 8 Hz, 2H), 7.36 (d, J = 8.5 Hz, 2H), 7.01 (s, 2H), 5.97 (bs, 1H), 5.41 (vbs, 1H), 4.71 (s, 2H), 4.38 (t, J = 7.5Hz, 1H), 4.18 (dd, J = 1, 8 Hz 1H), 3.06–2.89 (m, 2H), 2.21–2.08 (m, 2H), 1.99–1.92 (m, 1H), 1.75–1.65 (m, 1H), 1.52–1.42(m, 5H), 1.38–1.31 (m, 1H), 1.19 (pen, J = 7.5 Hz, 2H), 0.86 (d, J = 6.5 Hz, 3H) 0.85 (d, J = 7 Hz, 3H).

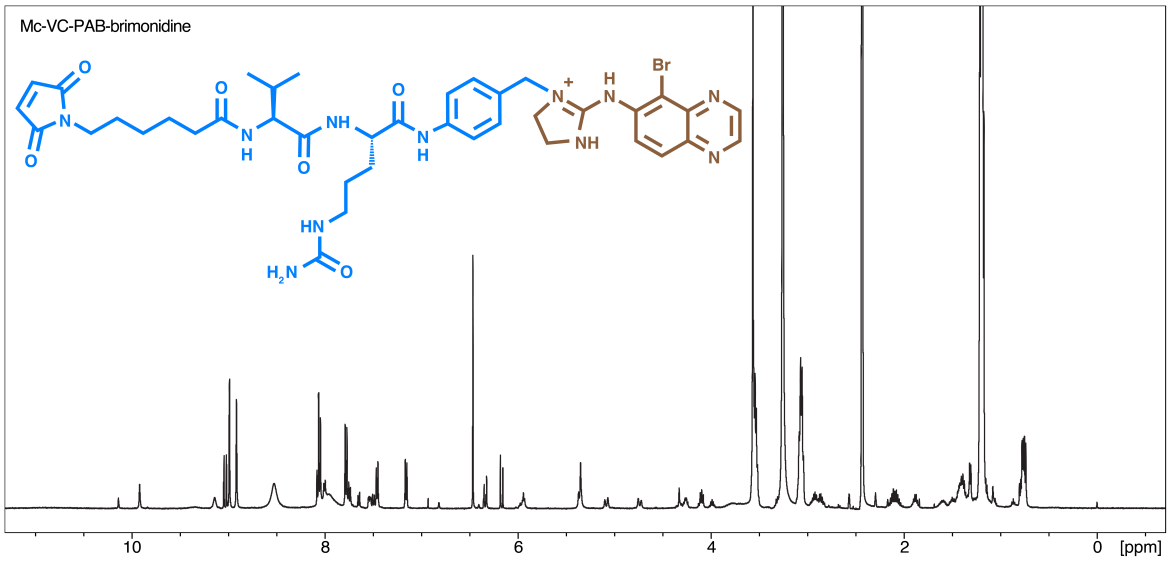


Figure A.11 NMR spectrum of Mc-VC-PAB-brimonidine. The prep-HPLC RT of Mc-VC-PAB-brimonidine was 9.8 min. The molecular structure of Mc-VC-PAB-brimonidine is shown in the upper left corner. Peak location and associated information are ^{1}H NMR (500 MHz, DMSO-d$_6$) 9.98 (s, 1H), 9.21 (bs, 1H), 9.11 (d, J = 1.5 Hz, 1H), 9.08 (d, J = 2 Hz, 1H) 9.05 (d, J = 1.5 Hz, 3H), 8.98 (s, 1H), 8.59 (bs, 5H), 8.15 (s, 1H), 8.12 (d, J = 9 Hz, 5H), 8.07(d, J = 7.5 Hz, 2H), 7.84 (d, J = 9Hz, 4H), 7.81 (d, J = 9 Hz, 1H), 7.71 (d, J = 8.5 Hz, 1H), 7.60 (dd, J = 4.5, 8.5 Hz, 1H), 7.52 (d, J = 8.5 Hz, 2H), 7.22 (d, J = 8.5 Hz, 2H), 6.53 (s, 3H), 6.40 (dd, J = 3.5, 27 Hz, 2H), 6.25 (s, 1H), 6.22 (s, 1H), 6.06–6.00 (m, 3H), 5.42 (d, J = 10.5 Hz, 5H), 5.15 (d, J = 14.5 Hz, 2H), 4.81 (d, J = 18.5 Hz, 2H), 4.33 (s, 1H), 4.25 (m, 2H), 4.13–4.07 (m, 2H), 4.01–3.96 (m, 1H), 2.96–2.82 (m, 4H), 2.17–2.03 (m, 4H), 1.87 (dd, J = 6.5 Hz, 2H), 1.64–1.56 (m, 2H), 1.53–1.48 (m, 2H), 1.44–1.36 (m, 4H), 1.32–1.31 (d, J = 6.5 Hz, 4H), 0.76 (d, J = 6.5 Hz, 3H), 0.74 (d, J = 6.5 Hz, 3H).

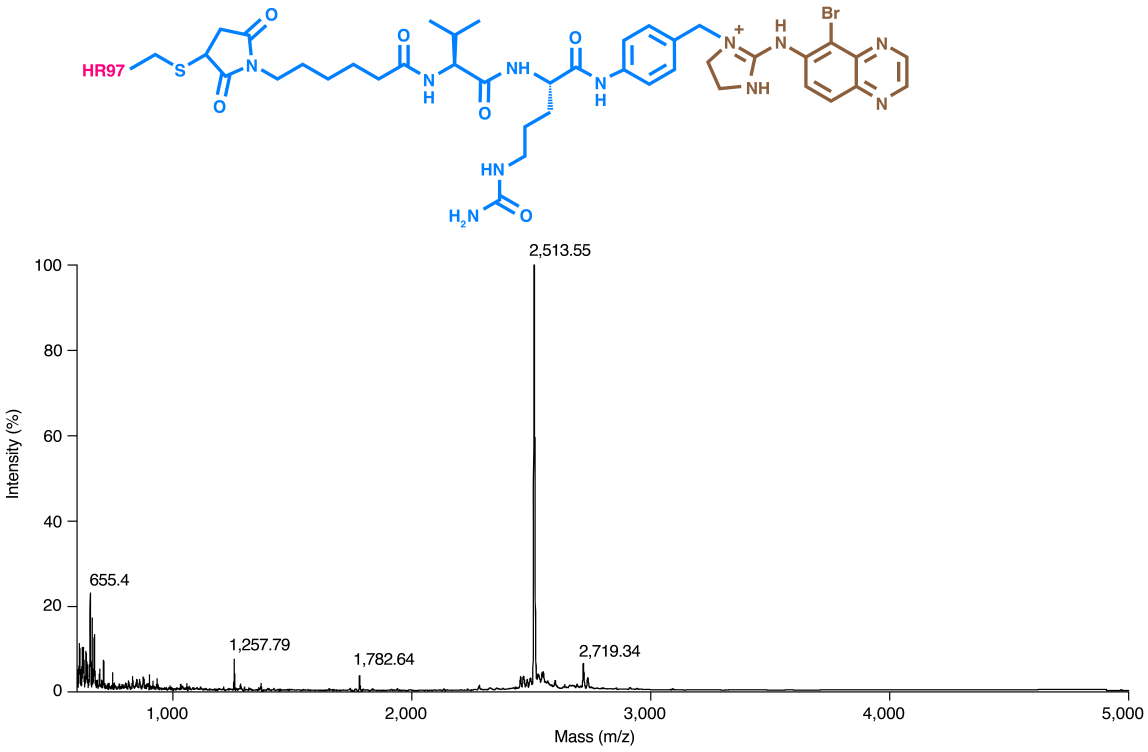


Figure A.12 MALDI-TOF spectrum of the HR97-brimonidine conjugate. The molecular structure of HR97-(quaternary-ammonium-linked)-brimonidine conjugate is shown in the upper left corner. The *m/z* calculated for $C_{103}H_{162}BN_{38}O_{20}S^+$ was 2,513.19, and 2,513.55 [M − $5H^+$ + $5Na^+$ + K^+] was found.

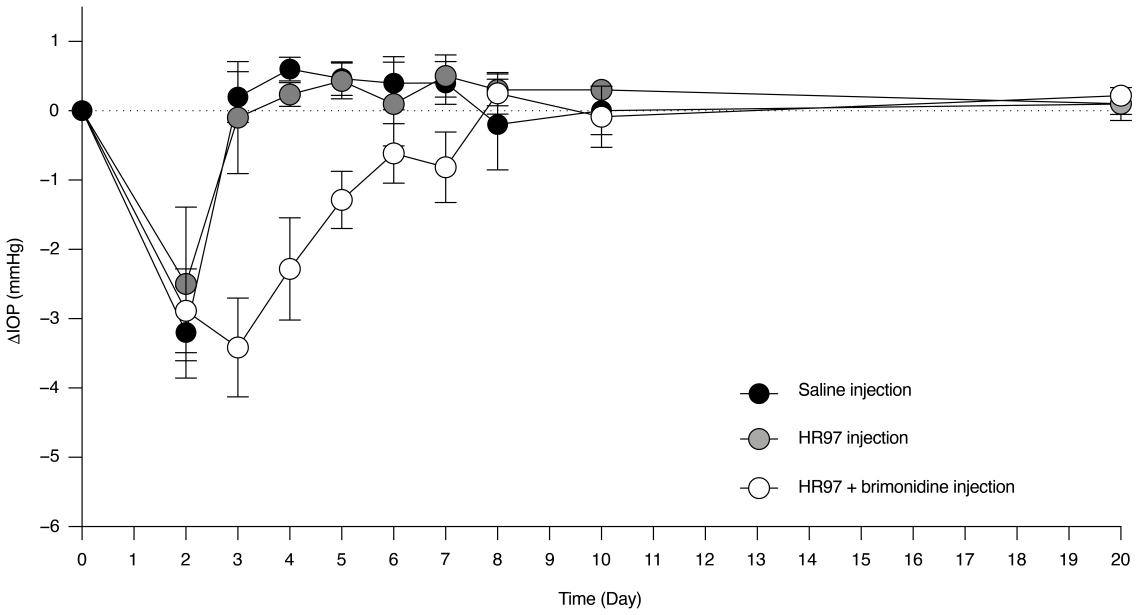


Figure A.13 Comparison of intraocular pressure (IOP) change from baseline. Line plot showing the IOP change from baseline (ΔIOP) after a single ICM injection of saline (black dots), HR97 (gray dots, equivalent to the amount of HR97 in HR97-brimonidine conjugate), and a physical mixture of HR97 and brimonidine tartrate in solution (white dots, HR97 + brimonidine, 200 μg brimonidine equivalent) in normotensive Dutch Belted rabbits ($n = 5$ per group). The IOP was measured every 1–2 days post-injection until measured value reached or exceeded baseline, and again at day 20 post-injection. Data are shown as mean ± SEM.

A.3 Supplementary Tables

Table A.1 Cross-validation performance (mean ± SEM) of the melanin binding general and adversarial control models.

Metric	General model	Adversarial control
Mean absolute error[a]	16.812 ± 0.166	32.857 ± 0.173
Root mean squared error[a]	22.782 ± 0.222	34.344 ± 0.226
Coefficient of determination (R^2)	0.543 ± 0.007	-0.038 ± 0.015

[a]Percent normalized values presented.

Table A.2 Cross-validation performance (mean ± SEM) of the cell-penetration general and adversarial control models.

Metric	General model	Adversarial control
Log loss	0.259 ± 0.017	0.715 ± 0.013
Matthews correlation coefficient	0.794 ± 0.014	-0.002 ± 0.054
F_1[a]	0.901 ± 0.005	0.522 ± 0.031
Balanced accuracy	0.897 ± 0.007	0.502 ± 0.028
Enrichment factor	2.081 ± 0.065	1.090 ± 0.159
BEDROC[b]	0.999 ± 0.000	0.529 ± 0.058

[a]Harmonic mean of precision and recall.

[b]Boltzmann-enhanced discrimination of receiver operating characteristic.

Table A.3 Cross-validation performance (mean ± SEM) of the cytotoxicity general and adversarial control models.

Metric	General model	Adversarial control
Log loss	0.172 ± 0.008	0.654 ± 0.005
Matthews correlation coefficient	0.879 ± 0.004	0.001 ± 0.012
F_1[a]	0.919 ± 0.002	0.047 ± 0.023
Balanced accuracy	0.947 ± 0.002	0.619 ± 0.035
Enrichment factor	1.519 ± 0.017	0.956 ± 0.057
BEDROC[b]	0.993 ± 0.005	0.620 ± 0.023

[a]Harmonic mean of precision and recall.

[b]Boltzmann-enhanced discrimination of receiver operating characteristic.

Table A.4 Ocular grading 7 days after a single ICM injection of saline, HR97 (equivalent to the amount of HR97 in HR97-brimonidine conjugate), or a physical mixture of HR97 and brimonidine tartrate in solution (HR97 + brimonidine, 200 μg brimonidine equivalent) in Dutch Belted rabbits ($n = 5$ per group).

Day 7	**Saline**					**Peptide**					**Peptide+brimonidine mixture**				
Rabbit ID	275	261	262	263	264	265	269	270	271	272	273	274	266	267	268
Pupillary light reflex	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival hypermia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival swelling	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival discharge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal opacity (severity)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal opacity (area)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal vascularization	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Aqueous flare	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anterior chamber cells	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Iris involvement	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anterior vitreous cells	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fluorescein staining (severity)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fluorescein staining (area)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eyelid discharge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eyelid swelling	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eyelid vascularity	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Meibomian gland function	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

All values are zero.

Table A.5 Ocular grading 14 days after a single ICM injection of saline, HR97 (equivalent to the amount of HR97 in HR97-brimonidine conjugate), or a physical mixture of HR97 and brimonidine tartrate in solution (HR97 + brimonidine, 200 μg brimonidine equivalent) in Dutch Belted rabbits ($n = 5$ per group).

Day 14	**Saline**					**Peptide**					**Peptide+brimonidine mixture**				
Rabbit ID	275	261	262	263	264	265	269	270	271	272	273	274	266	267	268
Pupillary light reflex	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival hypermia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival swelling	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival discharge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal opacity (severity)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal opacity (area)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal vascularization	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Aqueous flare	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anterior chamber cells	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Iris involvement	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anterior vitreous cells	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fluorescein staining (severity)	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0
Fluorescein staining (area)	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0
Eyelid discharge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eyelid swelling	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eyelid vascularity	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Meibomian gland function	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

All non-zero values are noted in bold.

Table A.6 Ocular grading 21 days after a single ICM injection of saline, HR97 (equivalent to the amount of HR97 in HR97-brimonidine conjugate), or a physical mixture of HR97 and brimonidine tartrate in solution (HR97 + brimonidine, 200 μg brimonidine equivalent) in Dutch Belted rabbits ($n = 5$ per group).

Day 21	**Saline**					**Peptide**					**Peptide+brimonidine mixture**				
Rabbit ID	275	261	262	263	264	265	269	270	271	272	273	274	266	267	268
Pupillary light reflex	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival hypermia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival swelling	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival discharge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal opacity (severity)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal opacity (area)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal vascularization	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Aqueous flare	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anterior chamber cells	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Iris involvement	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anterior vitreous cells	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fluorescein staining (severity)	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0
Fluorescein staining (area)	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0
Eyelid discharge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eyelid swelling	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eyelid vascularity	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Meibomian gland function	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

All non-zero values are noted in bold.

Table A.7 Ocular grading 28 days after a single ICM injection of saline, HR97 (equivalent to the amount of HR97 in HR97-brimonidine conjugate), or a physical mixture of HR97 and brimonidine tartrate in solution (HR97 + brimonidine, 200 μg brimonidine equivalent) in Dutch Belted rabbits ($n = 5$ per group).

Day 28	**Saline**					**Peptide**					**Peptide+brimonidine mixture**				
Rabbit ID	275	261	262	263	264	265	269	270	271	272	273	274	266	267	268
Pupillary light reflex	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival hypermia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival swelling	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Conjunctival discharge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal opacity (severity)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal opacity (area)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corneal vascularization	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Aqueous flare	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anterior chamber cells	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Iris involvement	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anterior vitreous cells	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fluorescein staining (severity)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fluorescein staining (area)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eyelid discharge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eyelid swelling	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eyelid vascularity	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Meibomian gland function	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

All values are zero.

Appendix B: Supplementary Information for Engineered Peptide-Drug Conjugate Provides Sustained Protection of Retinal Ganglion Cells with Topical Administration in Rats

B.1 Supplementary Figures

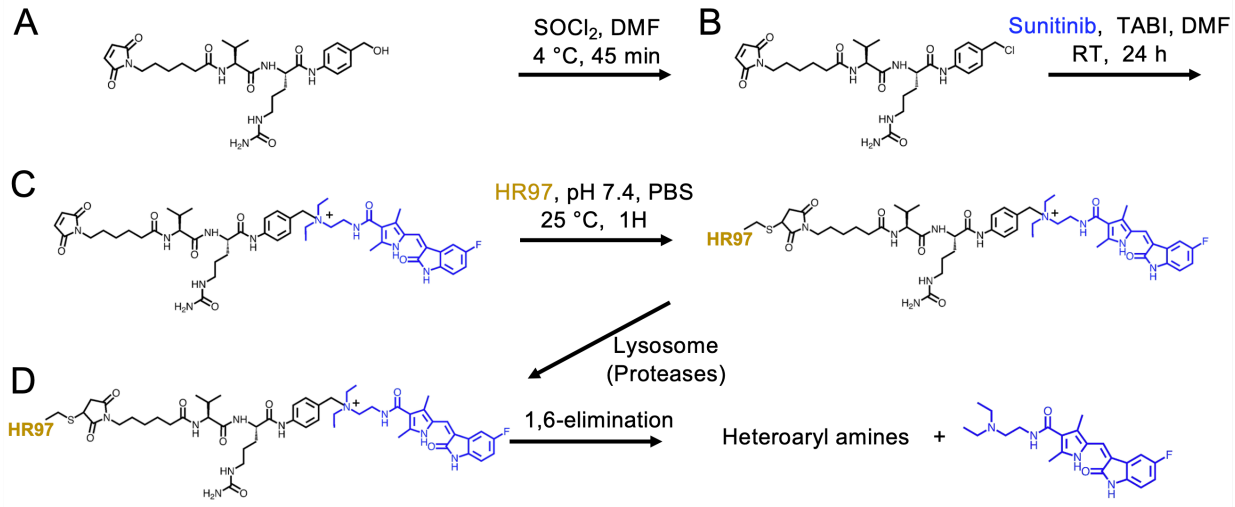


Figure B.1 Synthesis scheme for HR97-sunitinib. a MC-Val-Cit-PAB-OH was suspended in DMF and activated with thionyl chloride ($SOCl_2$) at 4 °C for 30 minutes. **b** The purified MC-Val-Cit-PAB-Cl were then conjugated to sunitinib base in the presence of tetrabutylammonium iodide (TABI) and N,N-diisopropylethylamine in DMF at room temperature for 24 hours. **c** HR97 with a terminal cysteine was conjugated to MC-Val-Cit-PAB-sunitinib via the thiol-maleimide reaction in PBS solution. **d** The HR97-sunitinib was designed for release of intact parent drug when triggered by an intracellular chemical and enzymatic event, such as protease cleavage of the amide bond. Sunitinib is shown in blue.

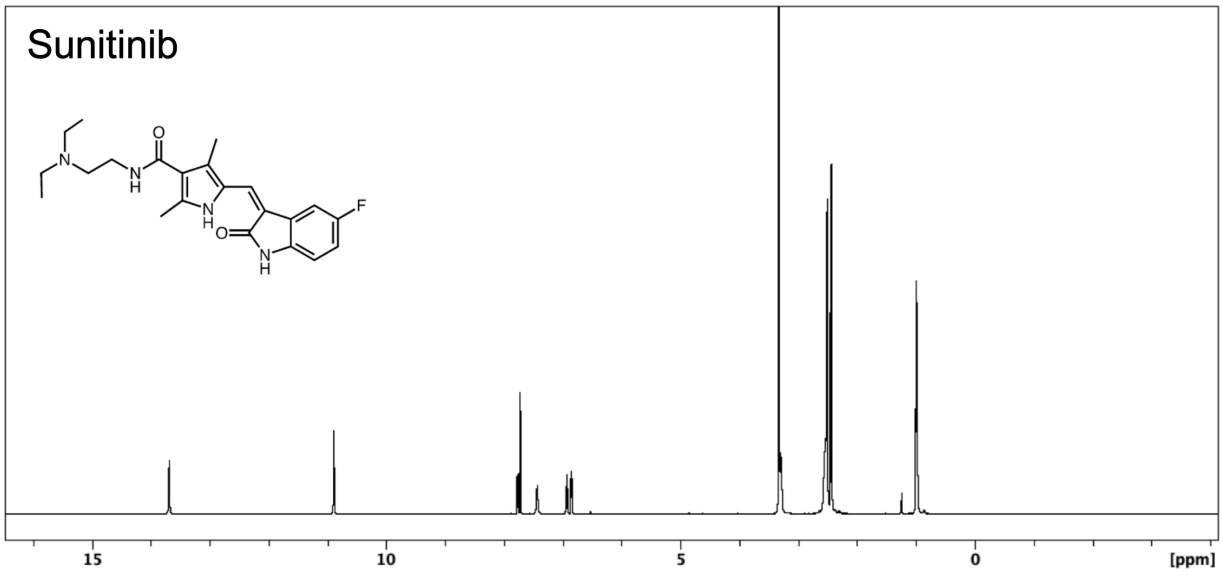


Figure B.2 NMR spectrum of sunitinib base. The prep-HPLC retention time of sunitinib base was 6.0 min. Peak location and associated information: ^{1}H NMR (500 MHz, DMSO-d_6) 13.68 (s, 1H), 10.88 (s, 1H), 7.76 (d, J = 9 Hz, 1H), 7.71 (s, 1H) 7.43 (bs, 1H), 6.92 (t, J = 9 Hz, 1H), 6.85–6.83 (m, 1H), 3.30–3.25 (m, 2H), 2.60–2.52 (m, 6H), 2.41 (d, J = 9.5Hz, 6H), 0.98 (t, J = 7 Hz, 6H).

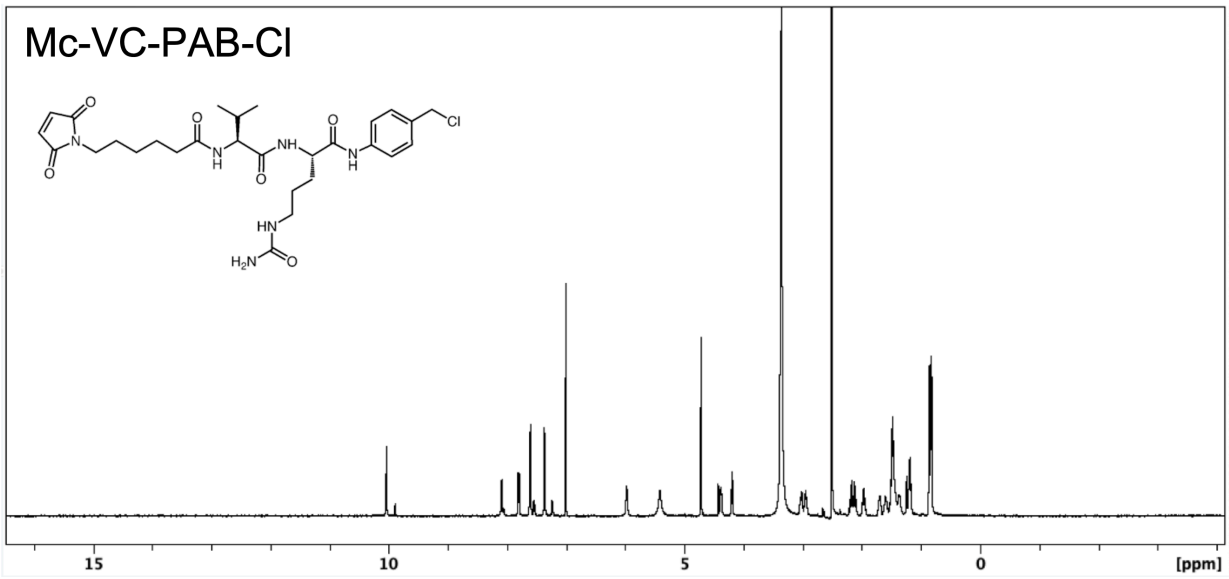


Figure B.3 NMR spectrum of Mc-VC-PAB-Cl. The prep-HPLC retention time of Mc-VC-PAB-Cl was 10.1 min. Peak location and associated information: ^{1}H NMR (500 MHz, DMSO-d_6) 10.03 (s, 1H), 8.08 (d, J = 7 Hz, 1H), 7.80 (d, J = 8.5 Hz, 1H), 7.60 (d, J = 8 Hz, 2H), 7.36 (d, J = 8.5 Hz, 2H), 7.01 (s, 2H), 5.97 (bs, 1H), 5.41(vbs, 1H), 4.71(s, 2H), 4.38 (t, J = 7.5 Hz, 1H), 4.18 (dd, J = 1, 8 Hz 1H), 3.06–2.89 (m, 2H), 2.21–2.08 (m, 2H), 1.99–1.92 (m, 1H), 1.75–1.65 (m, 1H), 1.52–1.42(m, 5H), 1.38–1.31 (m, 1H), 1.19 (pen, J = 7.5 Hz, 2H), 0.86 (d, J = 6.5 Hz, 3H) 0.85 (d, J = 7 Hz, 3H).

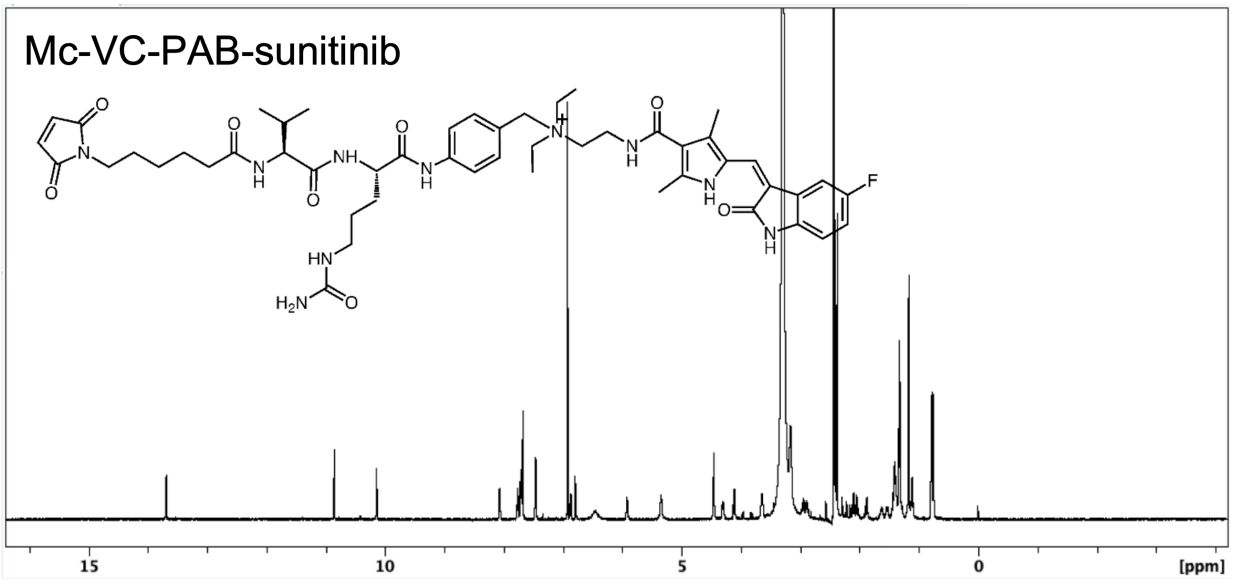
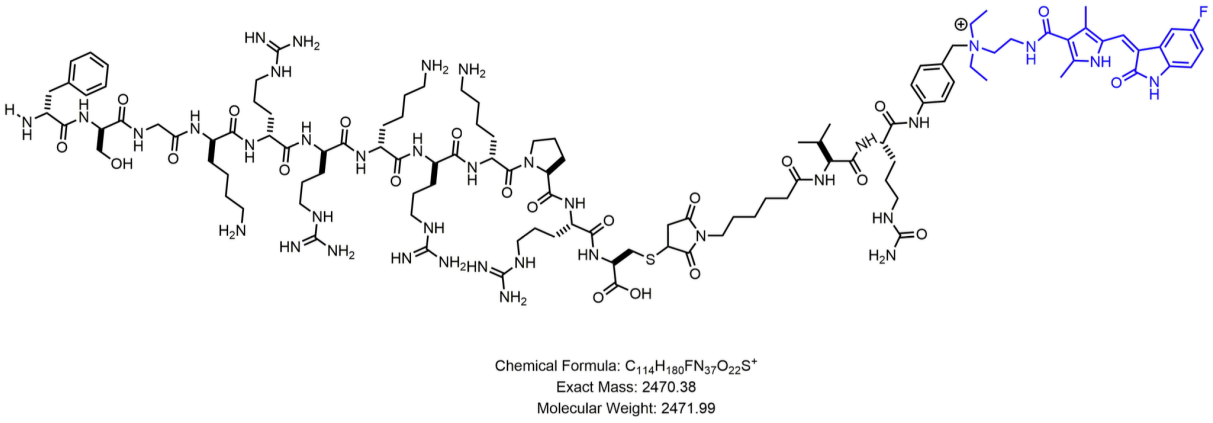


Figure B.4 NMR spectrum of Mc-VC-PAB-sunitinib. The prep-HPLC RT of Mc-VC-PAB-sunitinib was 8.3 min. Peak location and associated information: ^{1}H NMR (500 MHz, DMSO-d_6) 13.69 (s, 1H), 10.86 (s, 1H), 10.14 (s, 1H), 8.06 (d, J = 5 Hz, 1H), 7.76 (t, J = 6.0 Hz, 1H), 7.72–7.71 (m, 1H), 7.69 (s, 1H), 7.67 (d, J = 3 Hz, 2H), 6.93 (s, 2H), 6.88 (m, 1H), 6.80–6.77 (m, 1H), 6.46 (bs, 2H), 5.92 (t, J = 6 Hz, 1H), 5.34 (s, 2H), 4.45 (s, 2H), 4.32–4.28 (m,1H), 4.11 (dd, J = 7, 8.5 Hz, 1H), 3.67–3.63(m, 2H), 2.98–2.85 (m, 3H), 2.40(s, 3H), 2.38(s, 3H), 2.15–2.01(m, 2H), 1.92–1.85 (m, 1H), 1.68–1.58 (m, 1H), 1.59–1.49 (m, 1H), 1.47–1.37(m, 6H), 1.33(t, J = 7 Hz, 7H), 1.17(s, 5H), 1.144–1.080 (m, 3H), 0.78 (d, J = 7 Hz, 3H), 0.75 (d, J = 7 Hz, 3H).



Chemical Formula: $C_{114}H_{180}FN_{37}O_{22}S^+$
Exact Mass: 2470.38
Molecular Weight: 2471.99

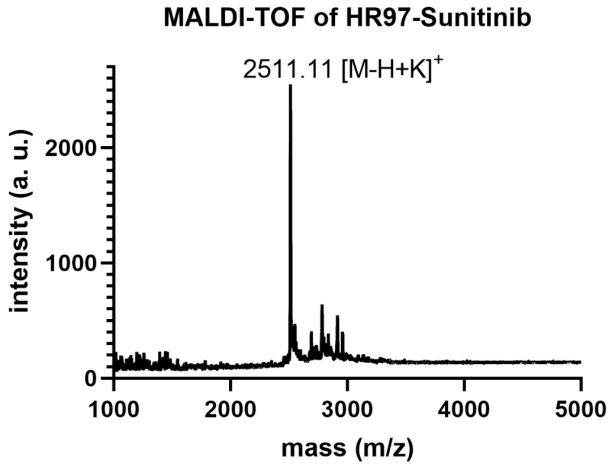


Figure B.5 Molecular structure of HR97-sunitinib conjugate and the MALDI-TOF spectrum. The molecular structure of HR97-(quaternary-ammonium-linked)-sunitinib conjugate is shown in the upper figure. Sunitinib is shown in blue. The m/z calculated for $C_{114}H_{180}FN_{37}O_{22}S+$ was 2,470.38, and 2,511.11 $[M-H+K]^+$ was found.

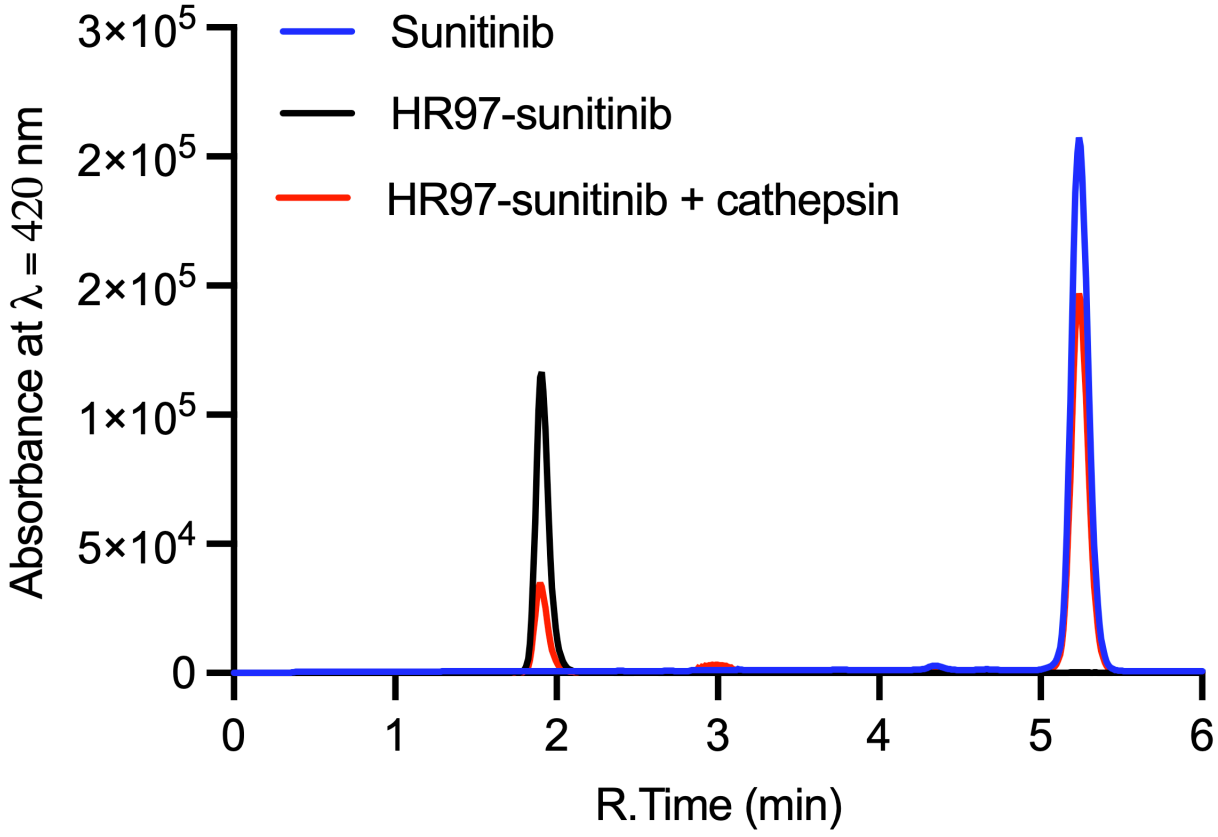


Figure B.6 HPLC analysis of cathepsin cleavage assay of the HR97-sunitinib conjugate. HR97-sunitinib was incubated with human cathepsin cocktails (cathepsin). Peak separation was visualized for HR97-sunitinib + cathepsin (red line) along with HR97-sunitinib (black line) and sunitinib (blue line). HPLC was conducted with a Luna® 5 μm C18(2) 100 ÅLC column 250 × 4.6 mm (Phenomenex, Torrance, CA) at 40 °C using isocratic flow (1 mL/min 60% TFA 0.1% in ACN). R.Time denotes retention time.

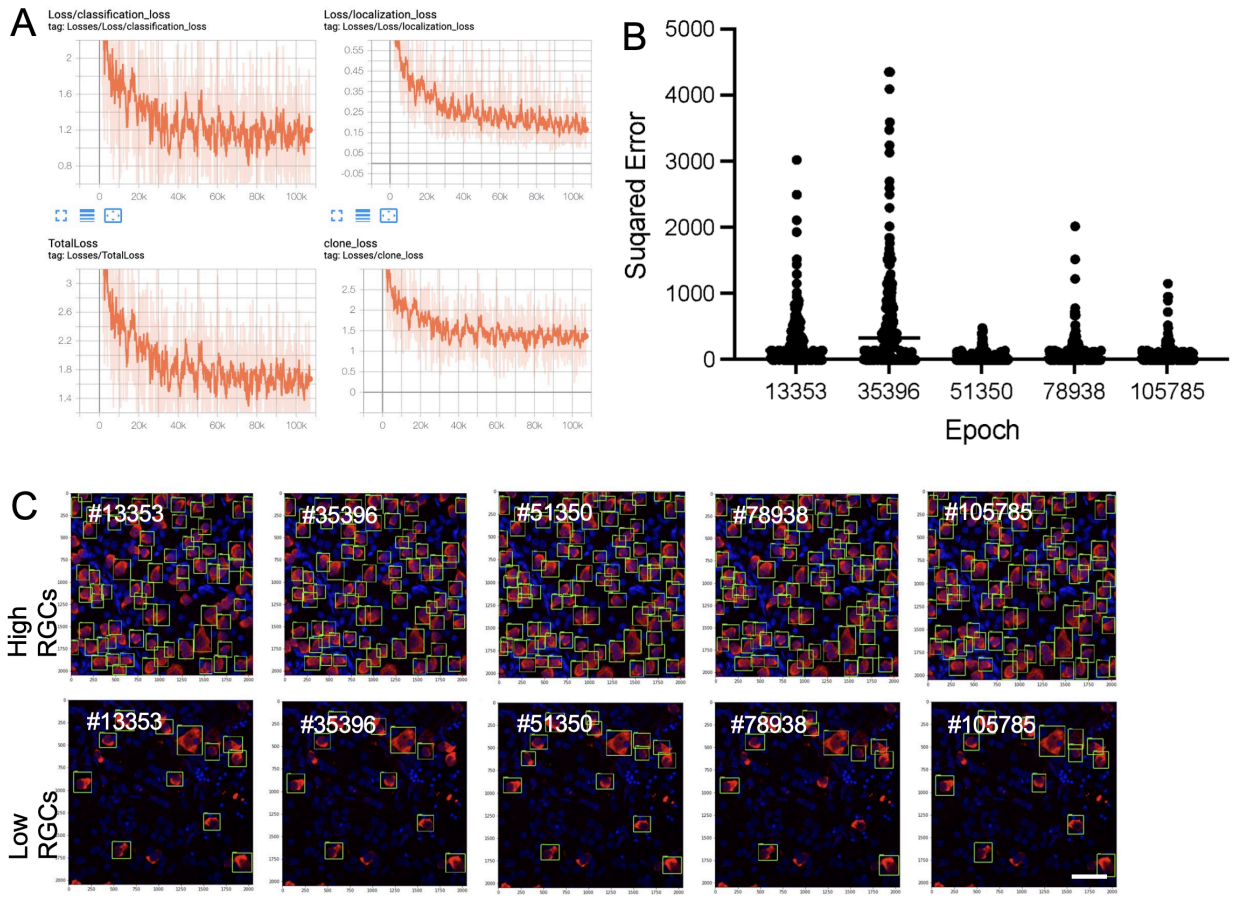


Figure B.7 RGC quantification using SSD Mobile-Net. **a** Classification, localization, total, and clone loss of training epochs monitored using Tensorboard. **b** Comparisons between predicted RGC numbers and counts by masked observers at different epochs. Squared error was used to evaluate model performance. Data are presented as mean ± SD ($n = 200$ per group). **c** Representative images with high or low RGC density trained in various epochs. Scale bar = 50 μm.

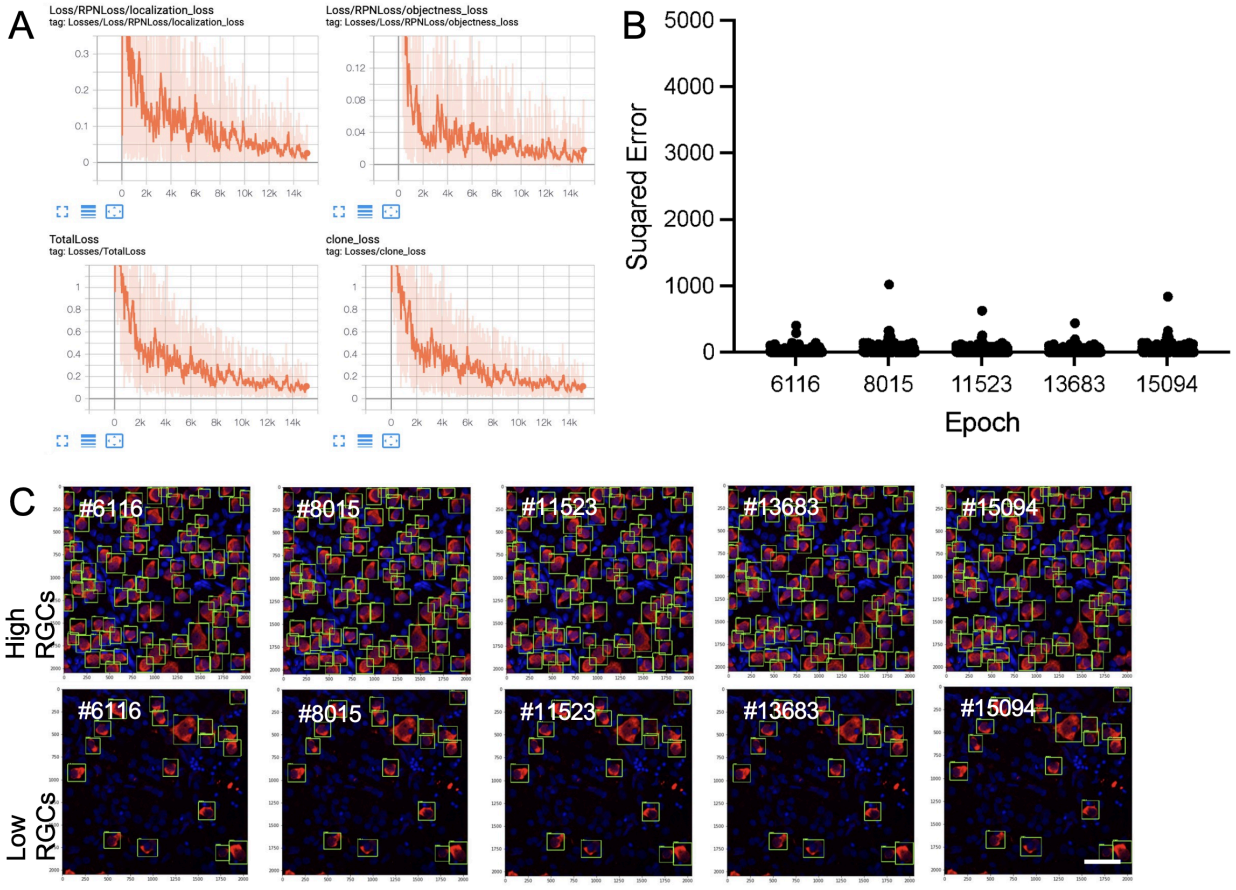


Figure B.8 RGC quantification using the Faster R-CNN Inception Resnet v2 model. **a** Localization, objectness, total, and clone loss of training epochs monitored using Tensorboard. **b** Comparisons between predicted RGC numbers and counts by masked observers at different epochs. Squared error was used to evaluate model performance. Data are presented as mean ± SD (n = 200 per group). **c** Representative images with high or low RGC density trained in different epochs. Scale bar = 50 μm.

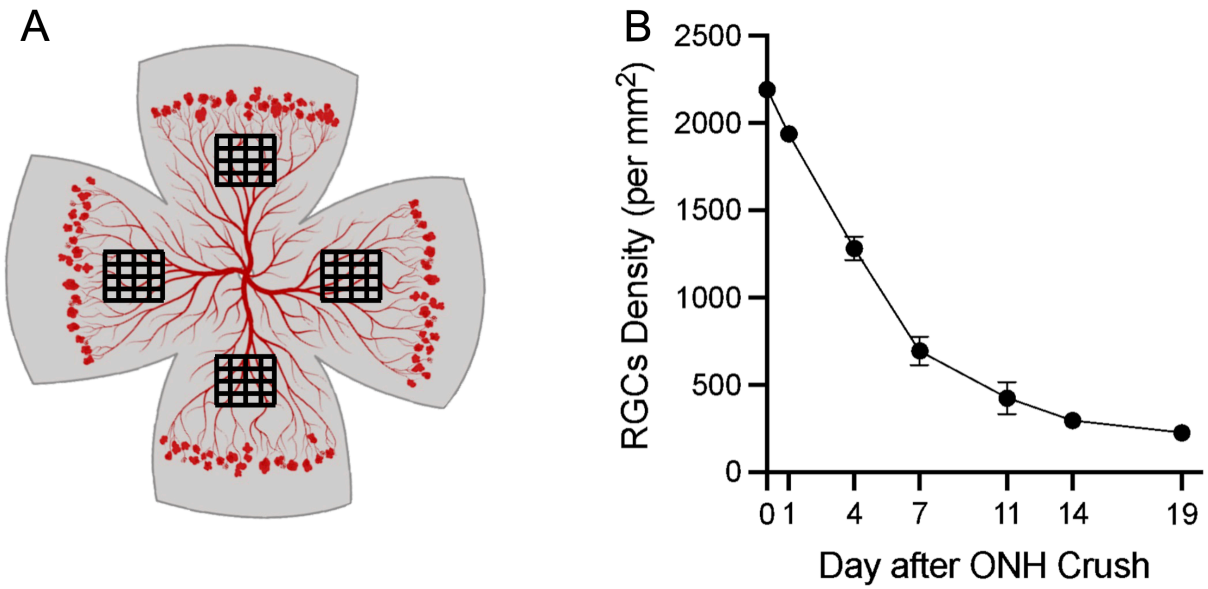


Figure B.9 Time course of RGC loss in the rat optic nerve head crush model. a Schematic showing the approximate locations of where 16 confocal images (40X objective) were obtained from the flat-mounted retina tissues ($n = 6$ animals per group). **b** RGC density quantified by the Faster R-CNN Inception Resnet v2 cell counting program. RGC numbers were converted to numbers per mm^2 tissue area. Data are presented as mean ± SD.

Appendix C: Supplementary Information for Positive-Unlabeled Learning Identifies Vaccine Candidate Antigens in the Malaria Parasite *Plasmodium falciparum*

C.1 Supplementary Figures

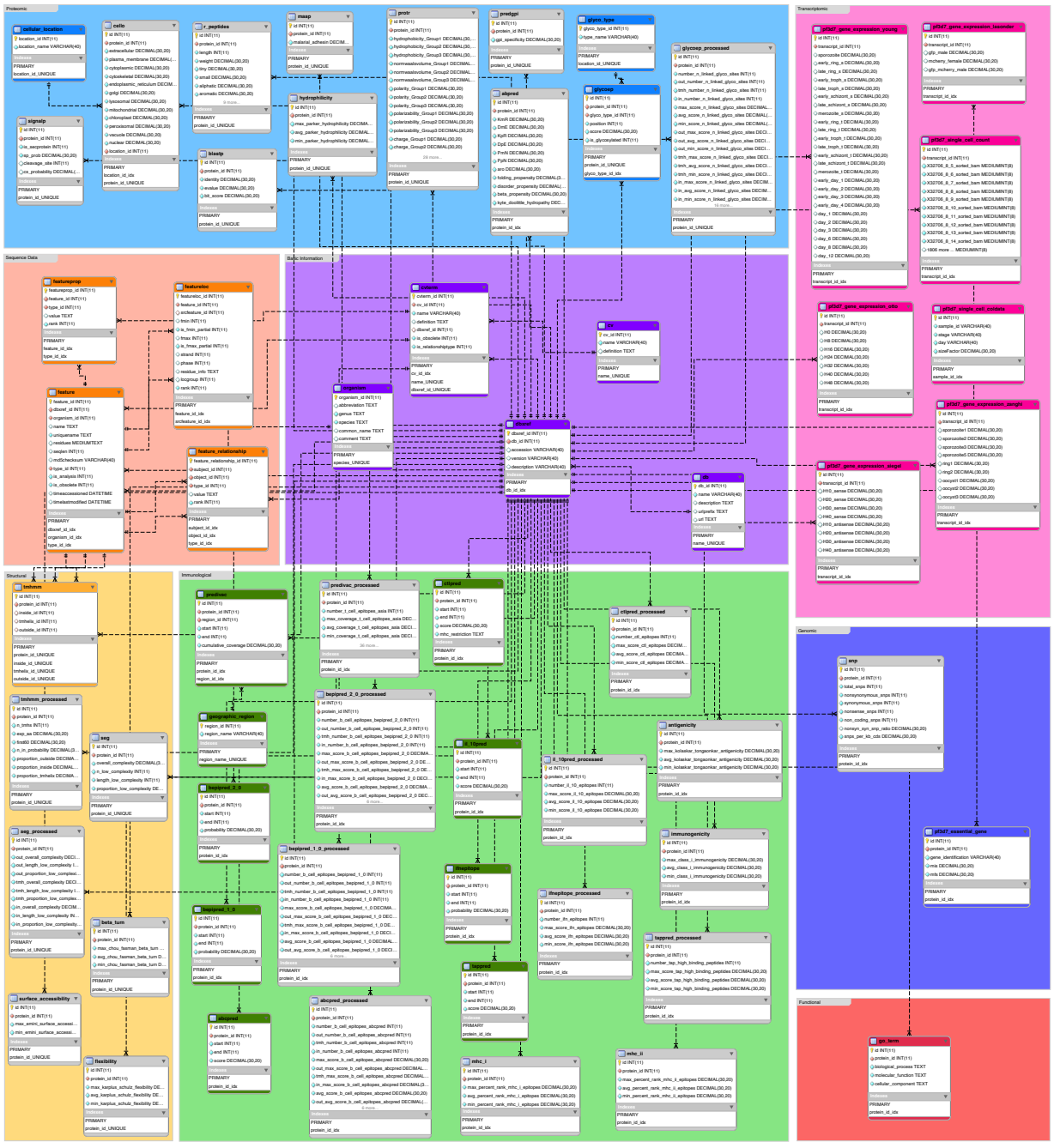


Figure C.1 Database schema of *P. falciparum* reverse vaccinology data. Data tables were grouped based on data properties and shown in different colors (light blue: proteomic; orange: sequence data; purple: basic information; yellow: structural; green: immunological; pink: transcriptomic; dark blue: genomic). Lines indicate relationship between data tables. Each table shows the variables stored and their corresponding data types in the database context. Data tables used to assemble the input data set used in the study are shown in grey.

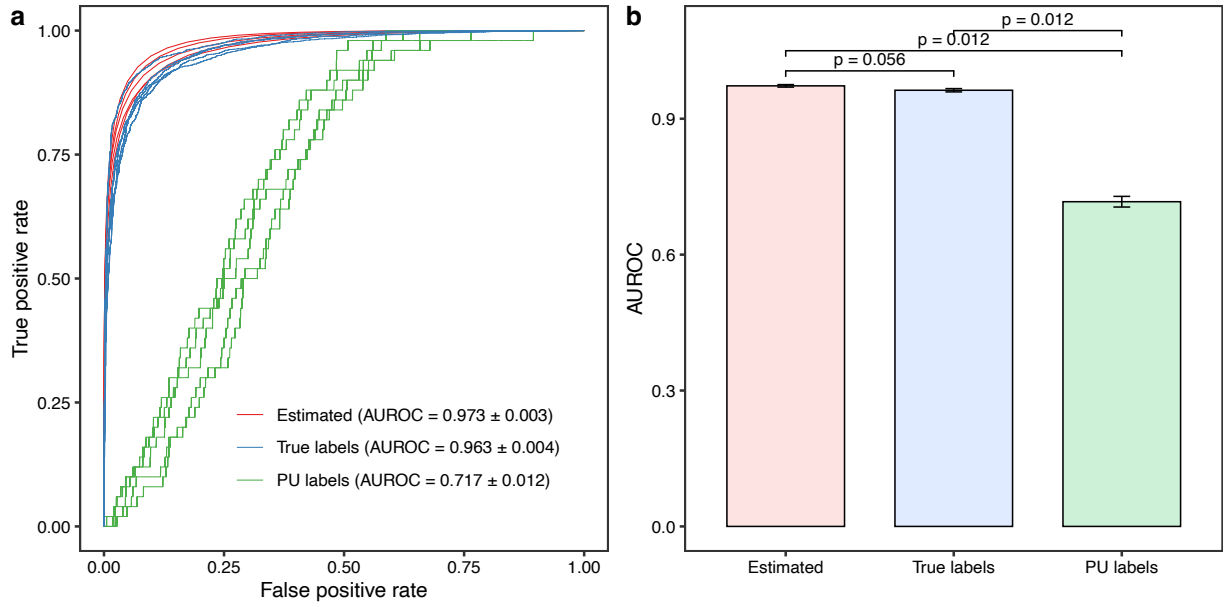


Figure C.2 Evaluation of model performance on simulated data set. a Lines represent receiver operating characteristic (ROC) curves estimated from the prediction score distribution (red), as well as computed regarding true labels and positive-unlabeled (PU) labels (blue and light green, respectively). The areas under the receiver operating characteristic curve (AUROC; mean ± SEM; $n = 5$) are noted in the parentheses in the legend. **b** Barplots showing the AUROC of ROC curves in **a**. Error bars indicate standard errors. Mann–Whitney test (two-sided) was performed, and the adjusted p values are shown on the top of the bars.

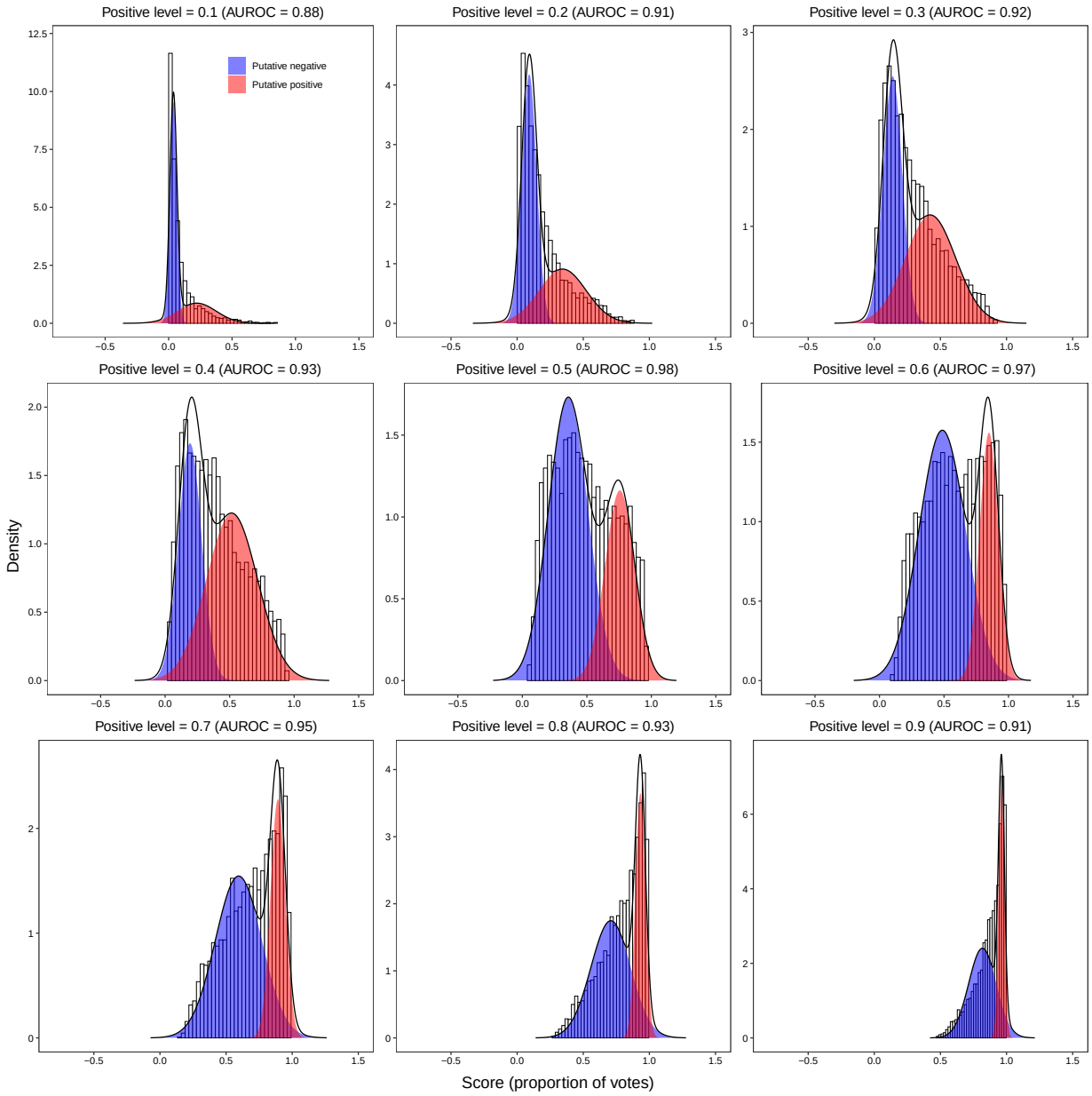


Figure C.3 Hyper-parameter tuning before variable space weighting. Subplots show prediction score distributions of the unlabeled proteins. The putative positive (red) and negative (blue) groups were computed using a two-component Gaussian mixture model. Receiver operating characteristic (ROC) curves were calculated based on the estimated distribution groups, and the areas under the receiver operating characteristic curves (AUROC) are indicated in the parentheses in the subplot titles.

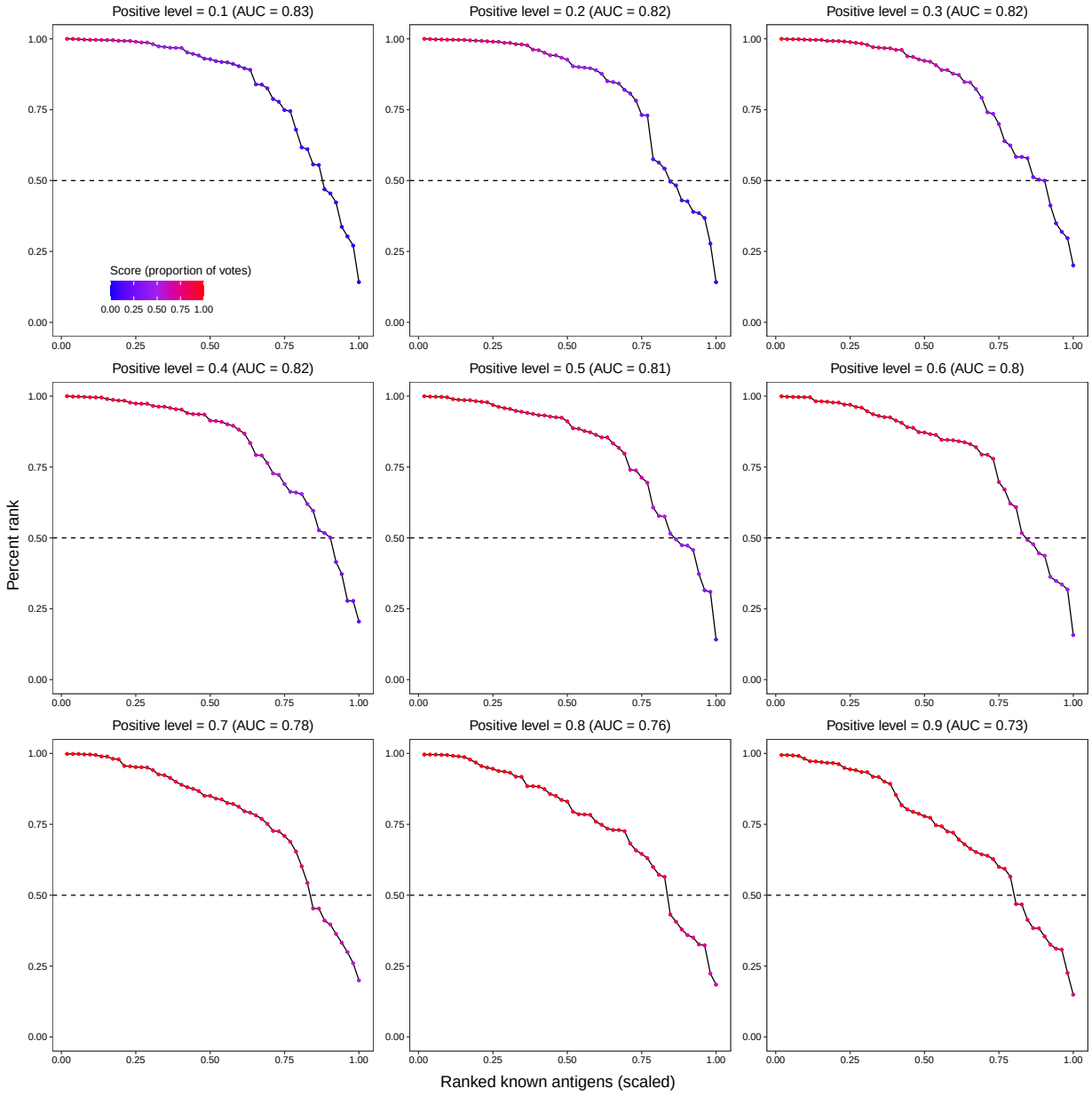


Figure C.4 Evaluation of known antigen predictions before variable space weighting. known antigens ($n = 52$) and unlabeled proteins were ranked based on probability scores from ensembles with different positive levels (0.1–0.9; shown as subplots). The x-axes show scaled ranks of known antigens, and the y-axes indicate precentile rank of known antigens among all *P. falciparum* proteins. Gradient colors represent prediction scores. Dashed lines indicated percentile ranks of 0.5. The areas under the ranking curves (AUC) are shown in the parenthesis in the subplot titles.

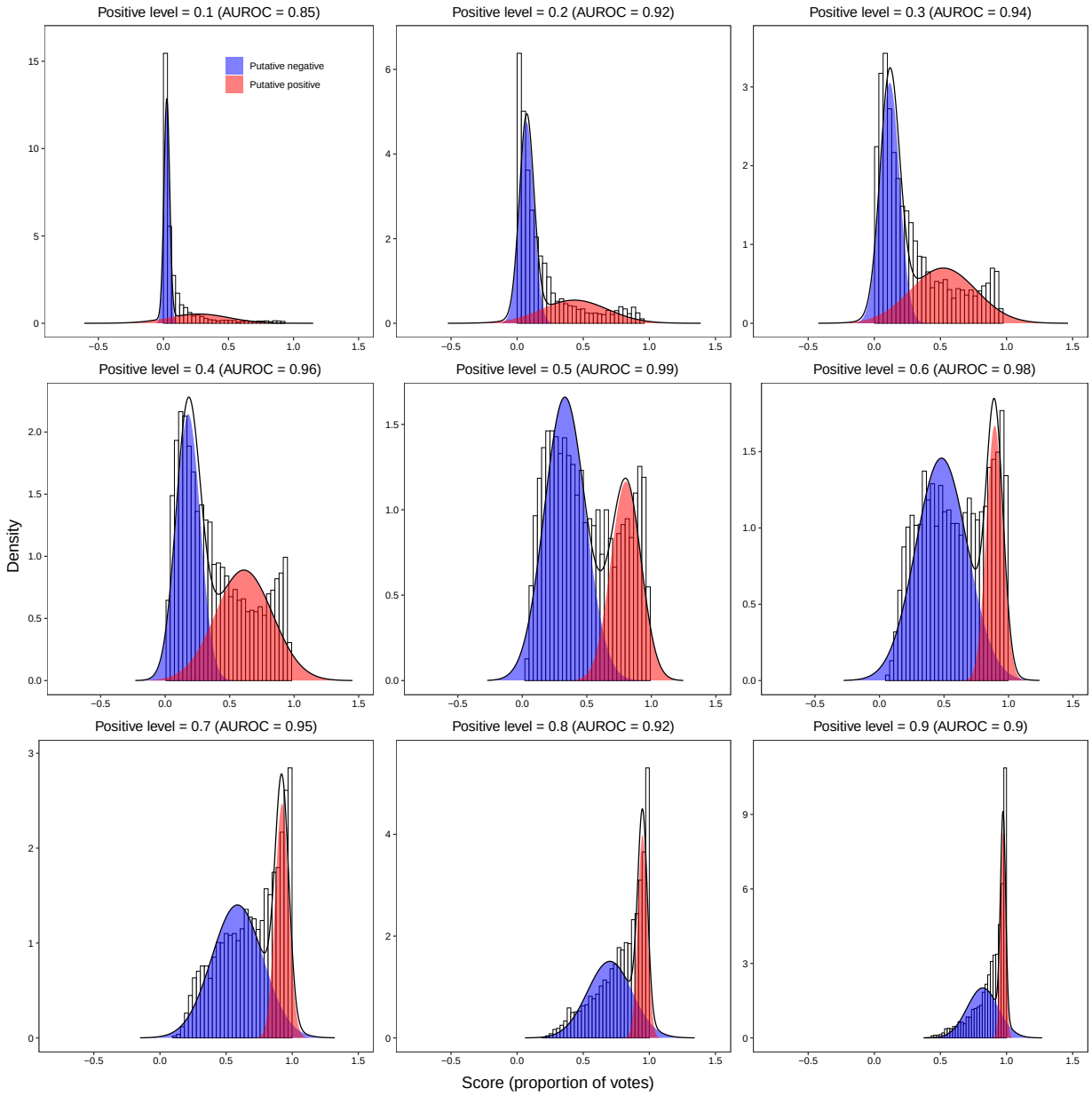


Figure C.5 Hyper-parameter tuning after variable space weighting. Probability score distributions of unlabeled proteins predicted by ensembles with different positive levels are shown in subplots. Score distributions were fitted using a two-component Gaussian mixture model to estimate the putative positive (red) and negative (blue) groups. Receiver operating characteristic curves (ROC) were calculated from the estimated distributions. The areas under the receiver operating characteristic curves (AUROC) are noted in the parentheses following the subplot titles.

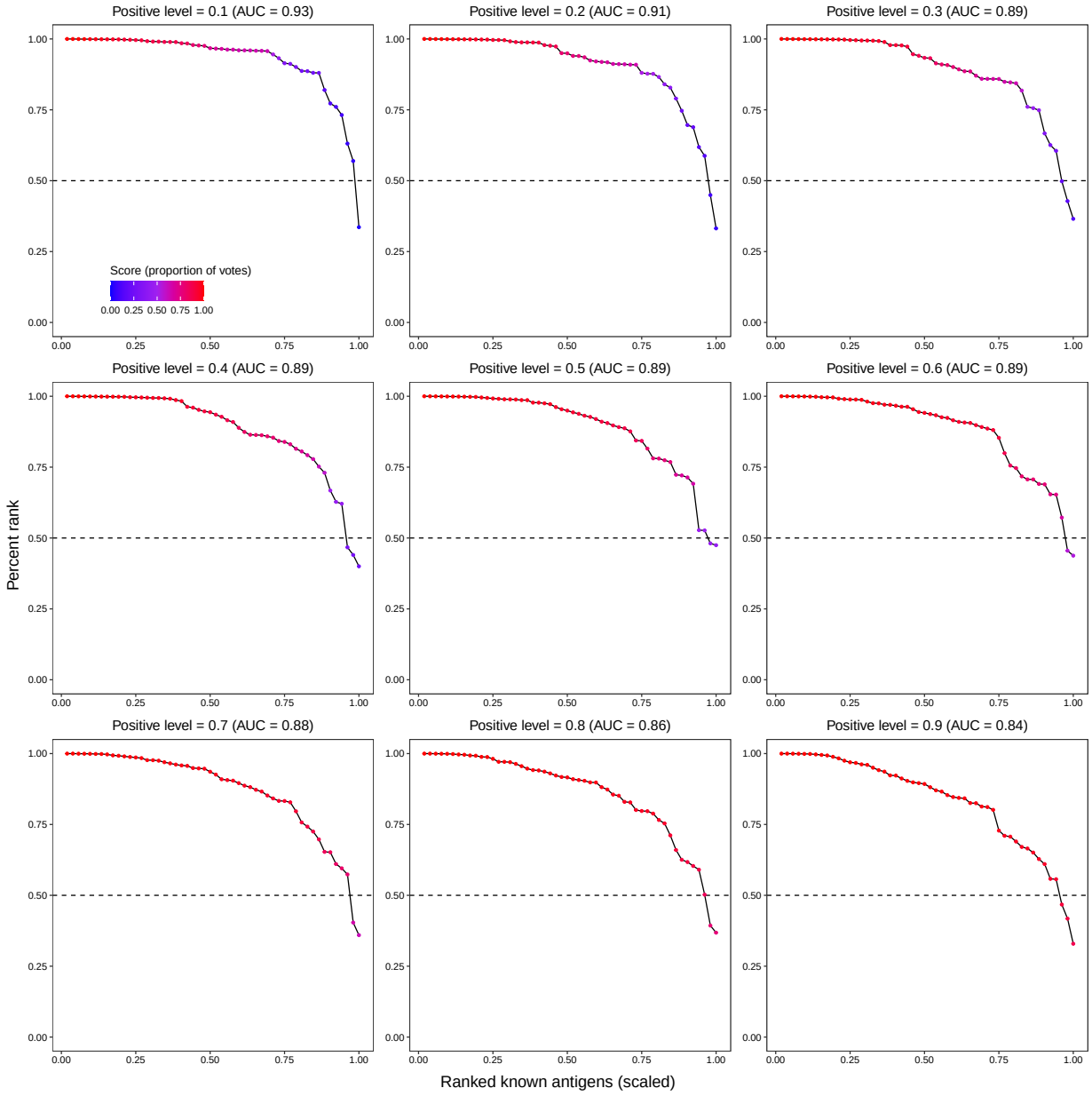


Figure C.6 Evaluation of known antigen predictions after variable space weighting. Dots in the subplots represent known antigens ($n = 52$). The x-axes show scaled ranks of known antigens only. The y-axes represent percentile ranks of known antigens among all *P. falciparum* proteins. Probability scores are noted by gradient colors. Dashed lines show 0.5 percent ranks, and the areas under the curves (AUC) are shown in the subplot title parentheses.

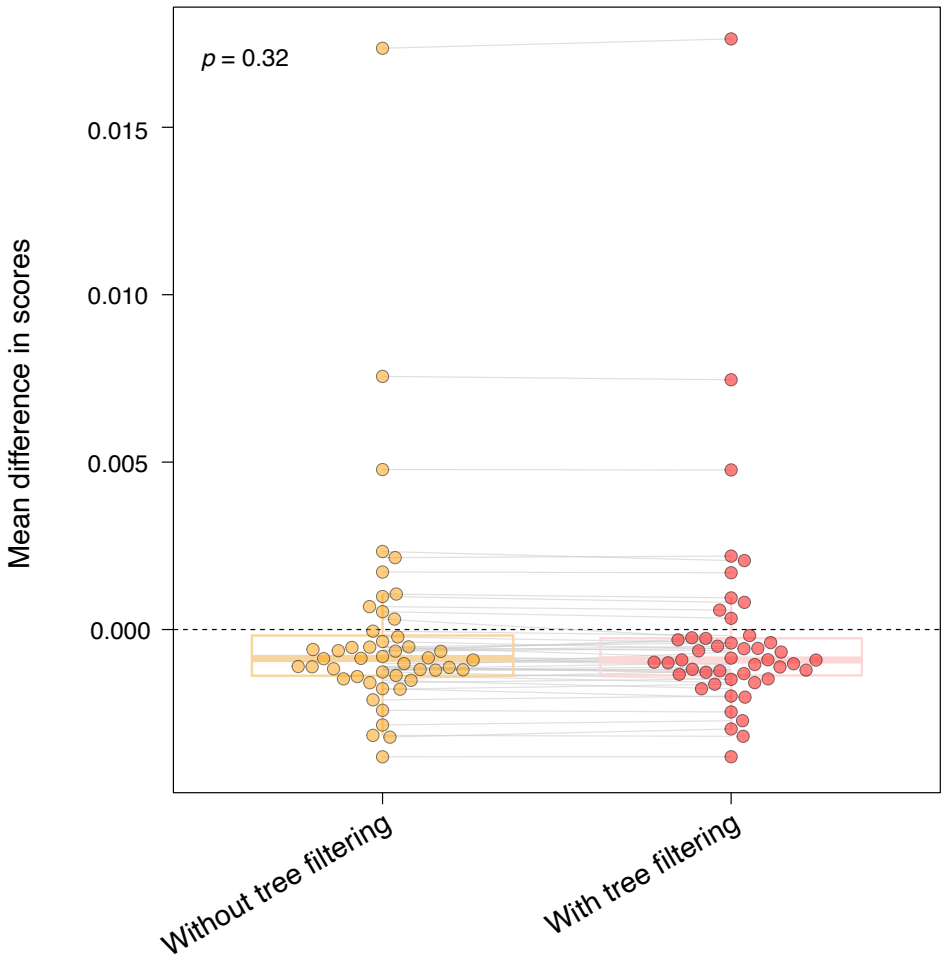


Figure C.7 Comparison of mean differences in probability scores after known antigen label removal. Labels of known antigens were removed iteratively, and the mean differences in scores for the remaining known antigens were calculated. The validation procedure was performed for ensembles with (red points, $n = 48$) and without (yellow points, $n = 48$) tree filtering. The box plots indicate medians with first and third quartiles. The lower and upper whiskers show 1.5 times the interquartile range extended from the first and third quartiles, respectively. The grey lines connect the same label removal iteration in both distributions. The dashed line shows a zero-mean difference in scores. The p-value was calculated using a pairwise two-sided Mann–Whitney test.

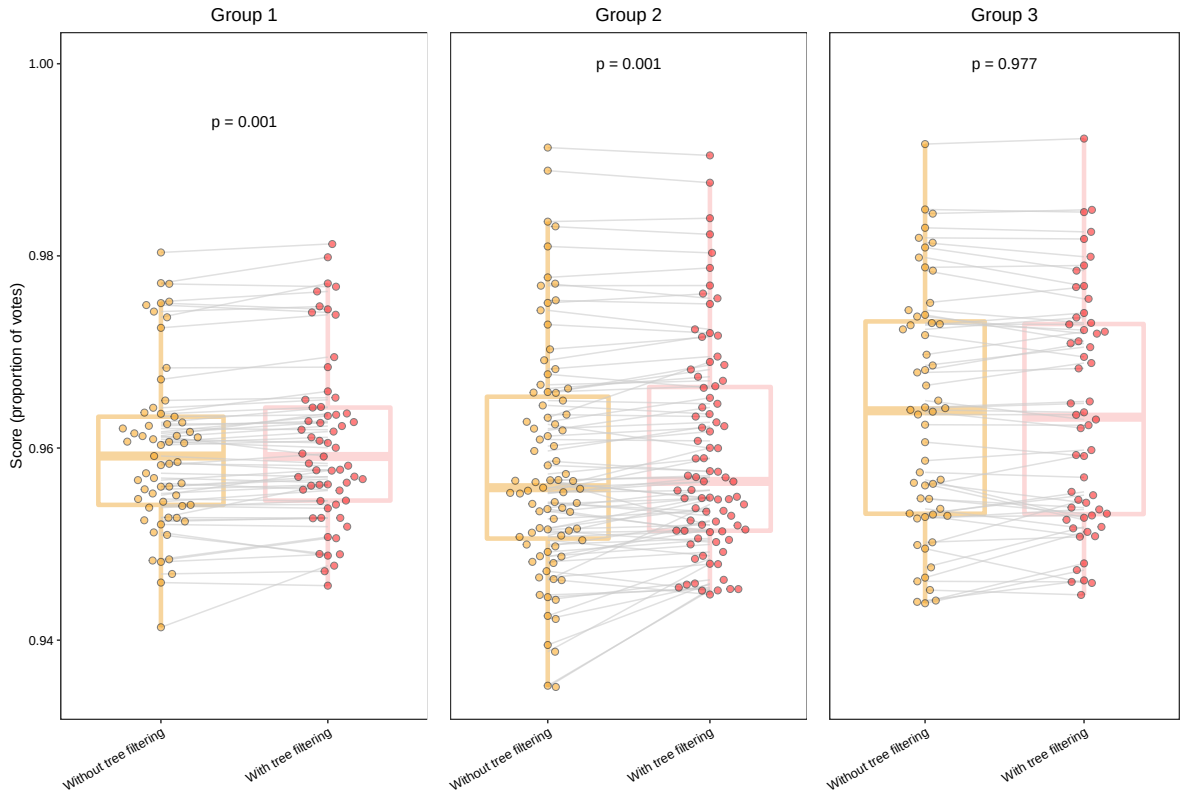


Figure C.8 Probability scores of candidate antigen groups. Comparison of scores predicted by non-tree-filtered (yellow points) and tree-filtered (red points) models for the three candidate groups (samples sizes: 61, 83, and 56). Points represent candidate antigens. Boxplots show the medians with the first and third quartiles. The lower and upper whiskers indicate 1.5 times the interquartile range extended from the first and third quartiles, respectively. Grey lines connect pairs of the same candidate antigens in the group, and adjusted p-values from two-sided pairwise Mann–Whitney tests are noted.

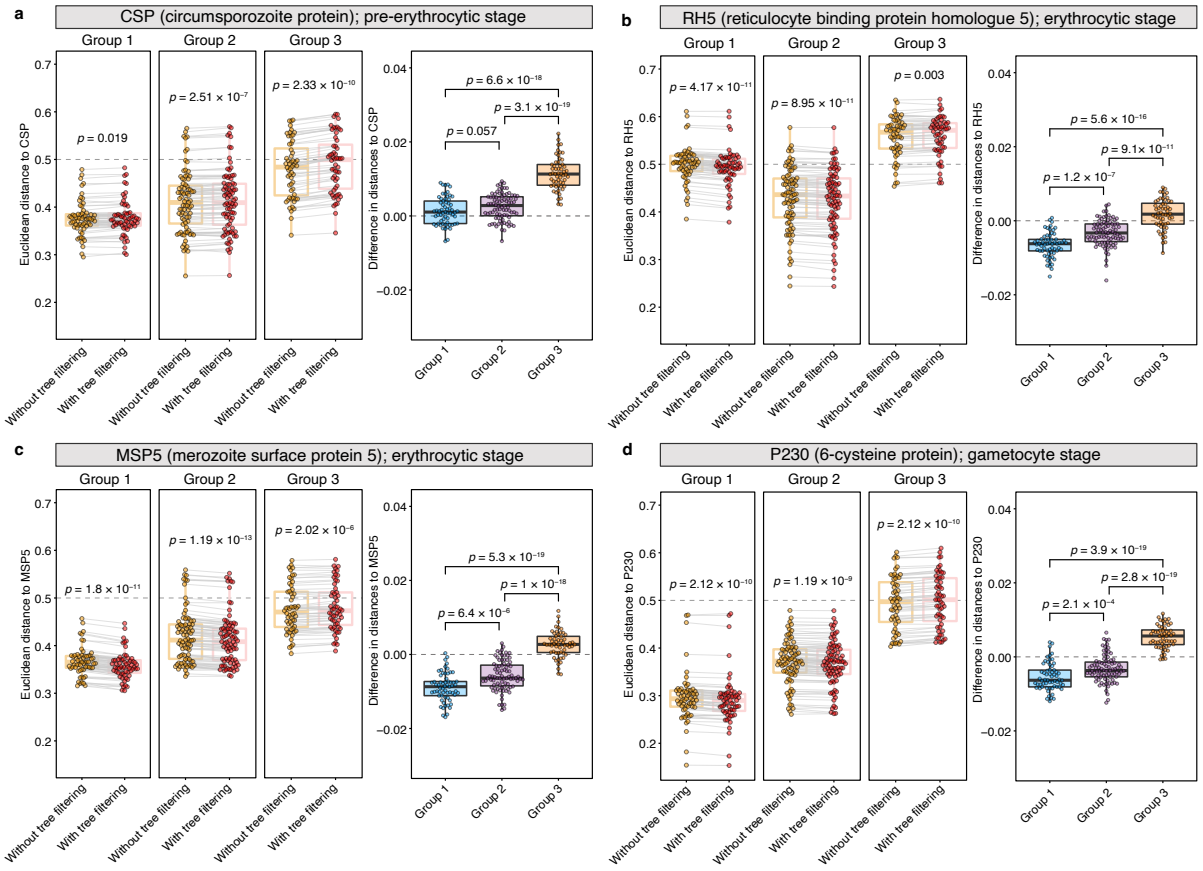


Figure C.9 Statistical comparisons of distances between candidate and reference antigens. Euclidean distances (ranging from 0–1) were calculated from the proximity matrix from the tree-based models. The summaries of distances of candidate antigens to the reference antigens CSP, RH5 MSP5, and P230 are shown in **a**, **b**, **c**, and **d**, respectively. For each plot, the left three panels show comparisons (two-sided pairwise Mann–Whitney test; p-values adjusted using the Benjamini–Hochberg procedure) of distances computed from the non-tree-filtered (yellow points) and tree-filtered (red points) models. The rightmost panel shows the comparisons (two-sided Mann–Whitney test with p-values adjusted) of distance differences between the three candidate antigen groups before and after tree filtering. Points represent candidate antigens. The corresponding p-values are noted above the compared groups. Dashed lines in the left and right panels indicate 0.5 distance and 0 difference in distances, respectively.

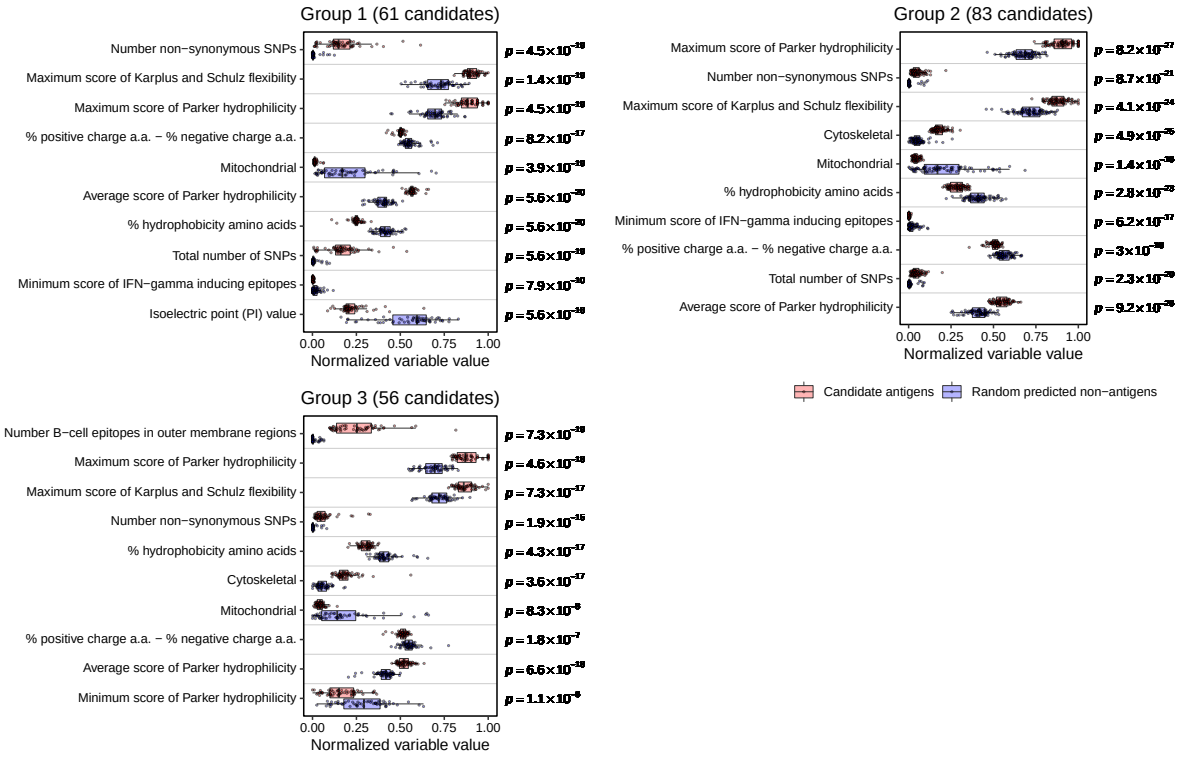


Figure C.10 Statistical comparisons of variable values of top important variables between the candidate antigen groups and randomly selected non-antigens. Top ten important variables analyzed using permutation-based variable importance based on the candidates in each group are shown (see also Tables C.1–C.3). For each candidate antigen group, the variable values were compared with a same number of randomly selected non-antigens predicted (probability score <0.5) using a two-sided Mann–Whitney test adjusted by the Benjamini-Hochberg procedure. The x-axes show normalized variable values based on the whole data set. Red dots and blue dots indicate candidates and predicted non-antigens, respectively. Boxplots display the first quartile, median, and third quartile values. The left and right whiskers indicate 1.5 times the interquartile range extended from the first and third quartiles, respectively.

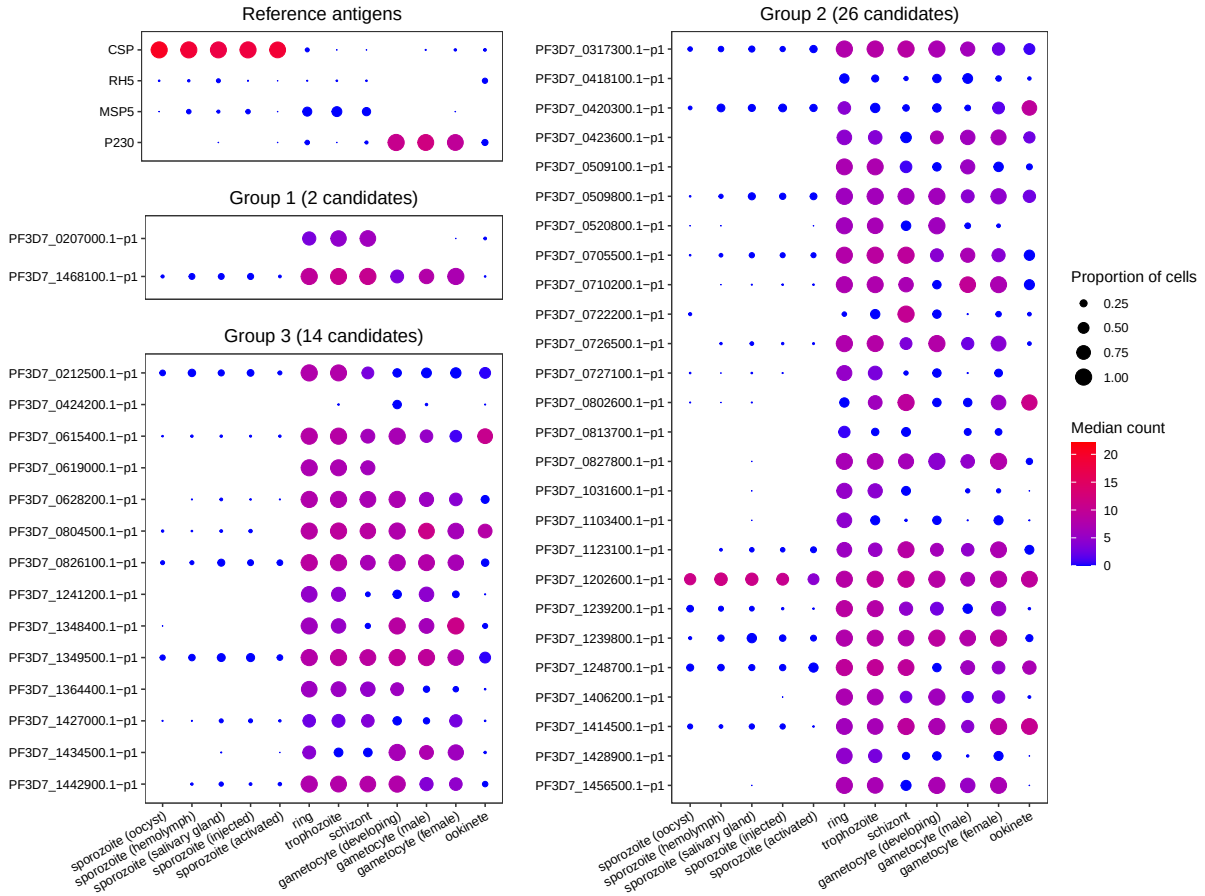


Figure C.11 Candidate antigen characterization across various *P. falciparum* life stages. Normalized gene counts of cells in each stage for the four reference antigens and the filtered candidate antigens with single-cell transcriptomic data from the Malaria Cell Atlas [203–205] are shown. Dot size represents proportion of cells having gene count larger than zero. Gradient colors indicate median count of the cell population in each life stage. The numbers of antigens in each candidate antigen group are noted in the parentheses in the subplot titles.

C.2 Supplementary Tables

Table C.1 Top important variables (upper part) and variable categories (lower part) in group 1 candidate antigens. Ranks in groups 2 and 3 individual variable and variable category importance are also shown (MDA: Mean Decrease Accuracy).

#	Variable	MDA	Group 2 rank	Group 3 rank	Group
1	Number non-synonymous SNPs	46.94	2	4	Genomic
2	Maximum score of Karplus and Schulz flexibility	46.52	3	3	Structural
3	Maximum score of Parker hydrophilicity	39.88	1	2	Proteomic
4	% positive charge a.a. − % negative charge a.a.	35.75	8	8	Proteomic
5	Mitochondrial	35.42	5	7	Proteomic
6	Average score of Parker hydrophilicity	35.29	10	9	Proteomic
7	% hydrophobicity amino acids	32.59	6	5	Proteomic
8	Total number of SNPs	32.34	9	12	Genomic
9	Minimum score of IFN-γ inducing epitopes	30.07	7	11	Immunological
10	Isoelectric point (PI) value	29.31	21	15	Proteomic
#	**Group variable**	**MDA**	**Group 1 rank**	**Group 2 rank**	
1	Proteomic group variables	174.69	1	1	
2	Immunological group variables	106.54	2	2	
3	Structural group variables	64.64	3	3	
4	Genomic group variables	58.52	4	4	

Table C.2 Top important variables (upper part) and variable categories (lower part) in group 2 candidate antigens. Ranks in groups 1 and 3 variable and variable category importance are also shown (MDA: Mean Decrease Accuracy).

#	Variable	MDA	Group 1 rank	Group 3 rank	Group
1	Maximum score of Parker hydrophilicity	52.05	3	2	Proteomic
2	Number non-synonymous SNPs	51.85	1	4	Genomic
3	Maximum score of Karplus and Schulz flexibility	51.45	2	3	Structural
4	Cytoskeletal	43.65	11	6	Proteomic
5	Mitochondrial	41.69	5	7	Proteomic
6	% hydrophobicity amino acids	40.30	7	5	Proteomic
7	Minimum score of IFN-γ inducing epitopes	38.95	9	11	Immunological
8	% positive charge a.a. − % negative charge a.a.	38.20	4	8	Proteomic
9	Total number of SNPs	38.01	8	12	Genomic
10	Average score of Parker hydrophilicity	35.37	6	9	Proteomic
#	**Group variable**	**MDA**	**Group 1 rank**	**Group 2 rank**	
1	Proteomic group variables	195.21	1	1	
2	Immunological group variables	124.44	2	2	
3	Structural group variables	76.17	3	3	
4	Genomic group variables	73.78	4	4	

Table C.3 Top important variables (upper part) and variable categories (lower part) in group 3 candidate antigens. Ranks in groups 1 and 2 variable and variable category importance are also shown (MDA: Mean Decrease Accuracy).

#	Variable	MDA	Group 1 rank	Group 2 rank	Group
1	Number B-cell epitopes in outer membrane regions	59.60	272	272	Immunological
2	Maximum score of Parker hydrophilicity	50.61	3	1	Proteomic
3	Maximum score of Karplus and Schulz flexibility	47.96	2	3	Structural
4	Number non-synonymous SNPs	43.77	1	2	Genomic
5	% hydrophobicity amino acids	41.91	7	6	Proteomic
6	Cytoskeletal	41.48	11	4	Proteomic
7	Mitochondrial	38.95	5	5	Proteomic
8	% positive charge a.a. − % negative charge a.a.	37.23	4	8	Proteomic
9	Average score of Parker hydrophilicity	36.71	6	10	Proteomic
10	Minimum score of Parker hydrophilicity	36.46	257	11	Proteomic
#	**Group variable**	**MDA**	**Group 1 rank**	**Group 2 rank**	
1	Proteomic group variables	177.86	1	1	
2	Immunological group variables	157.07	2	2	
3	Structural group variables	88.26	3	3	
4	Genomic group variables	62.82	4	4	

Appendix D: Supplementary Information for *Plasmodium vivax* Antigen Candidate Prediction Improves with the Addition of *Plasmodium falciparum* Data

D.1 Supplementary Figures

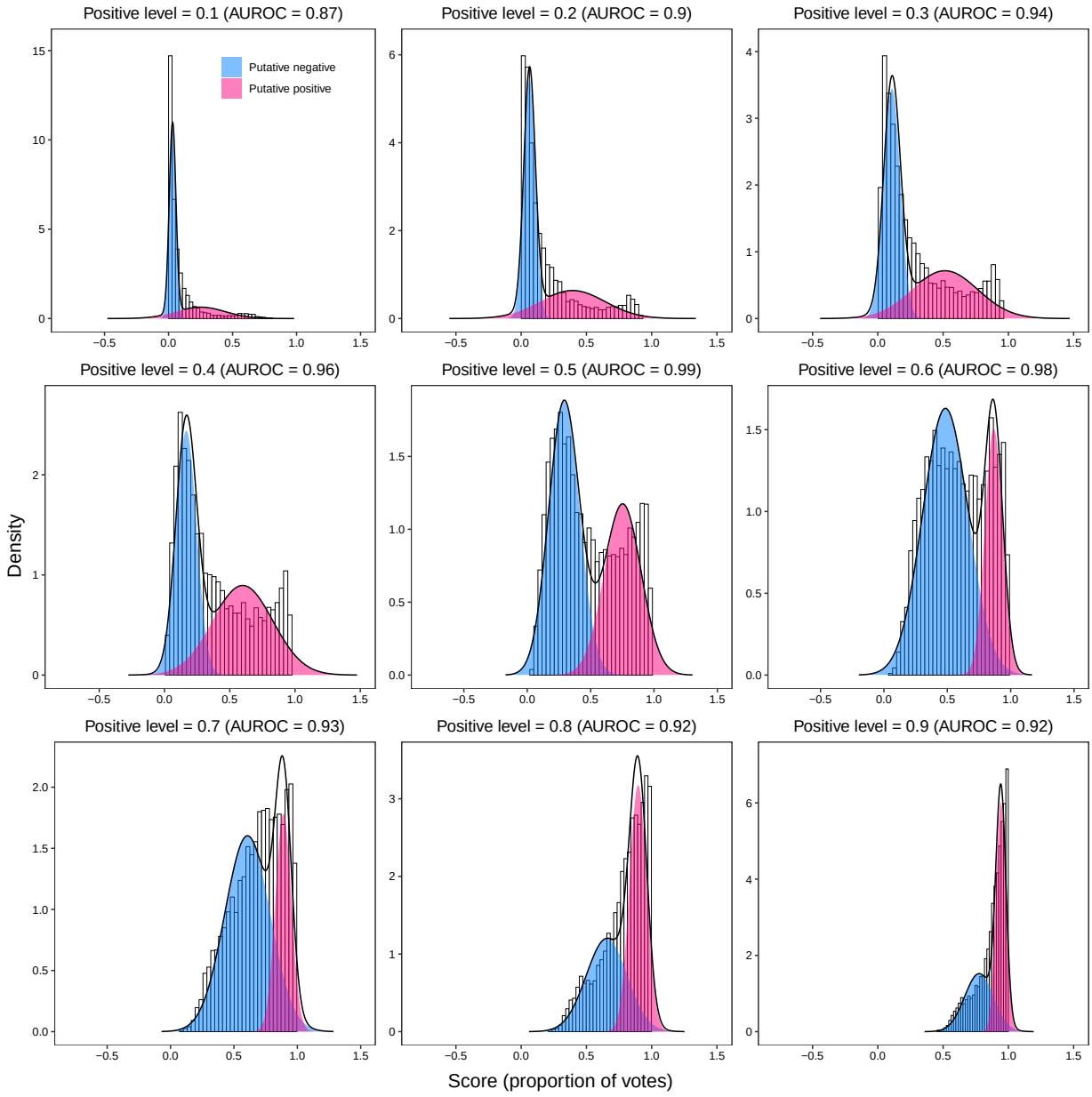


Figure D.1 Hyper-parameter tuning for PURF model trained on the *P. vivax* data set. Subplots showing probability score (proportion of votes) distributions of unlabeled proteins predicted by *P. vivax* models trained on different positive level (model hyper-parameter) settings ranging from 0.1 to 0.9. Magenta indicates putative positive distribution and blue represents putative negative distribution computed from a two-component Gaussian mixture model. Receiver operating characteristic (ROC) curves were computed based on the estimated distributions. The areas under the receiver operating characteristic curves (AUROC) are noted in the subplot titles.

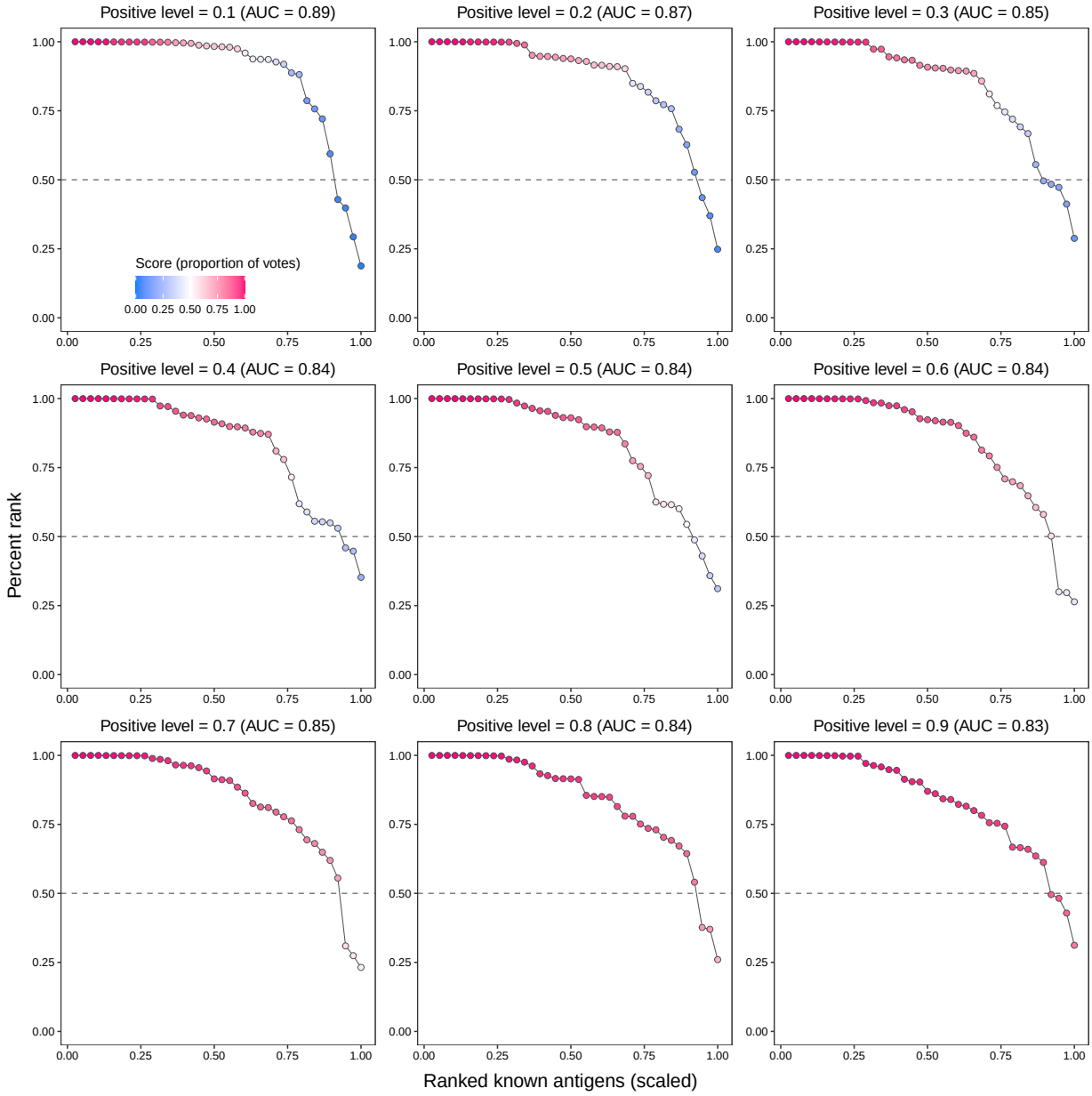


Figure D.2 Evaluation of known antigen predictions of the *P. vivax* model. Subplots of known antigen predictions from *P. vivax* models trained with different positive level (model hyper-parameter) settings, ranging from 0.1 to 0.9. Dots represent the 38 *P. vivax* known antigens, and the x-axes show the scaled ranks of these known antigens. The percentile ranks (the higher the better) calculated across all *P. vivax* proteins are indicated by the y-axes. The grey dashed lines show the percentile rank of 0.5, and the gradient colors represent probability scores (proportion of votes), with darker magenta color showing higher scores and darker blue showing lower scores. The areas under the curves (AUC) are noted in the subplot titles.

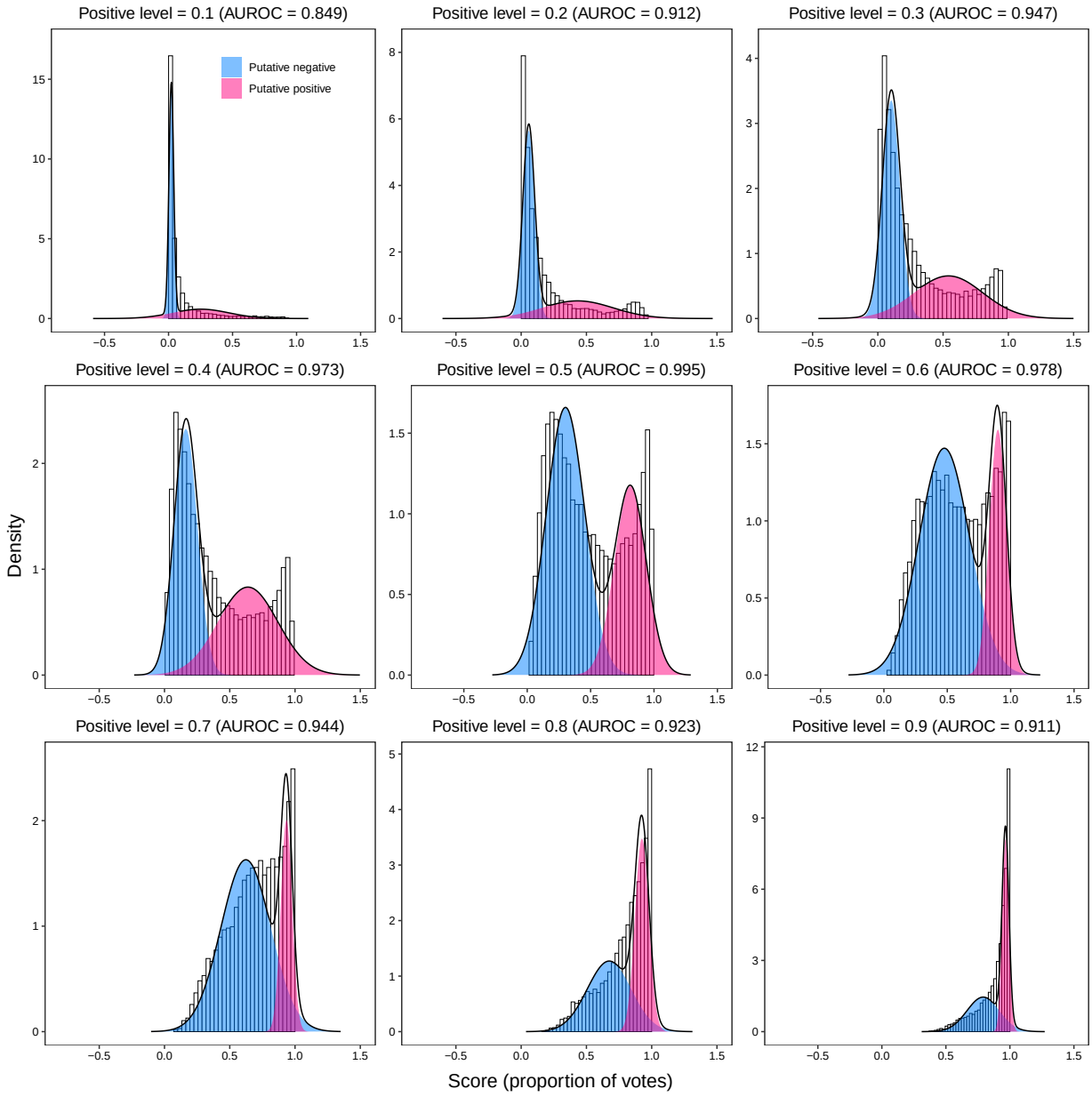


Figure D.3 Hyper-parameter tuning for PURF model trained on the combined data set. Subplots displaying distributions of probability scores (proportion of votes) of unlabeled proteins predicted by combined models trained on varying positive levels (hyper-parameters). The score distributions were modeled using a two-component Gaussian mixture to estimate the putative positive (magenta) and negative (blue) distributions. Receiver operating characteristic curves (ROC) were generated based on these estimated distributions. The areas under the receiver operating characteristic curves (AUROC) are indicated in the subplot titles.

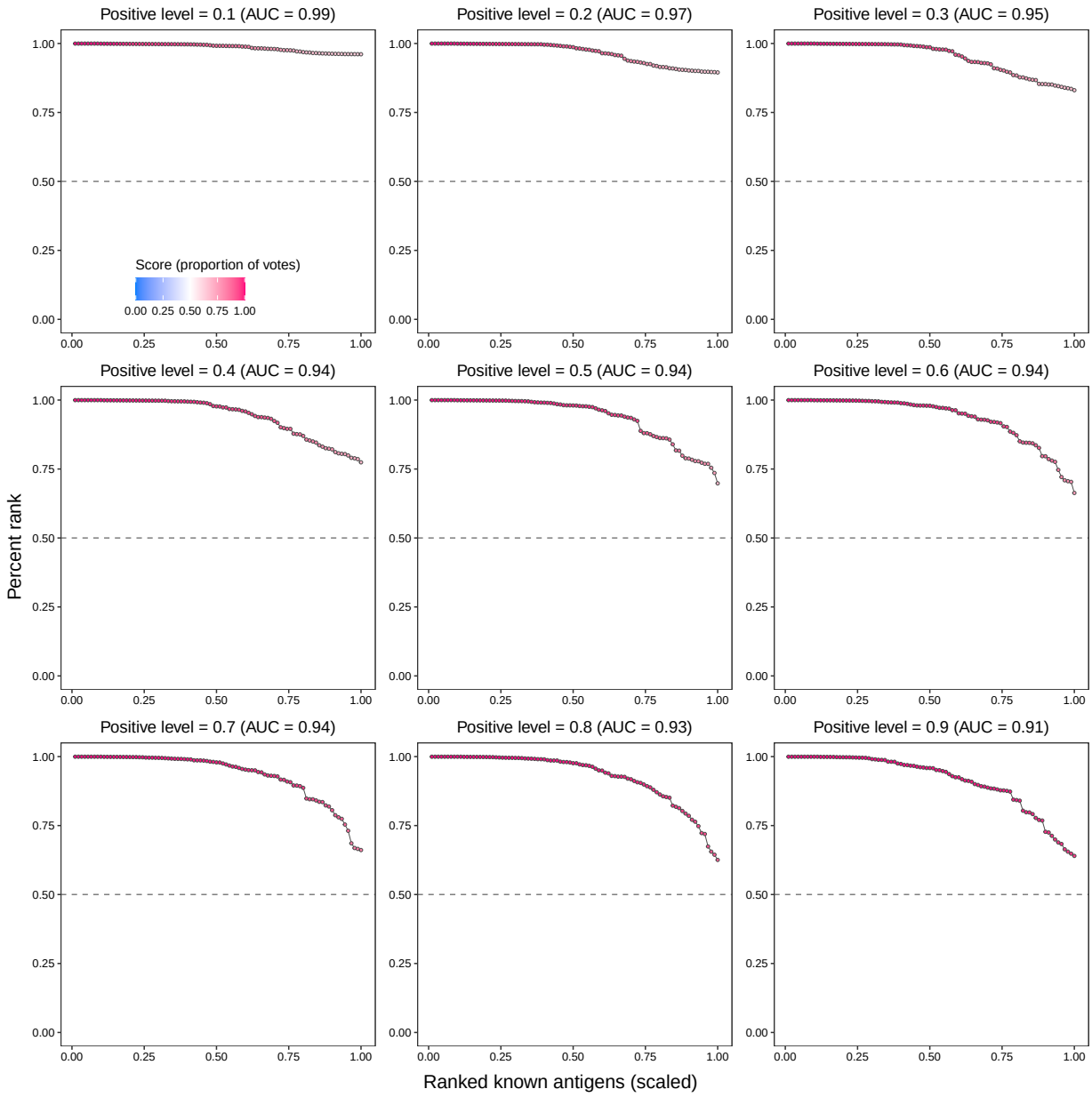


Figure D.4 Evaluation of known antigen predictions of the combined model. The subplots show percentile ranks of known antigen predictions of combined models trained with different positive level (model hyper-parameter) settings from 0.1 to 0.9. The x-axes show scaled ranks of the 90 known antigens from both *Plasmodium* species, and the y-axes indicate percentile ranks (the higher the better) of the known antigens across all proteins from both species. The grey dashed lines show the percentile rank of 0.5. Gradient colors convey probability scores (proportion of votes), with higher scores represented by darker magenta color and lower scores by darker blue color. The areas under the curves (AUC) are noted in the subplot titles.

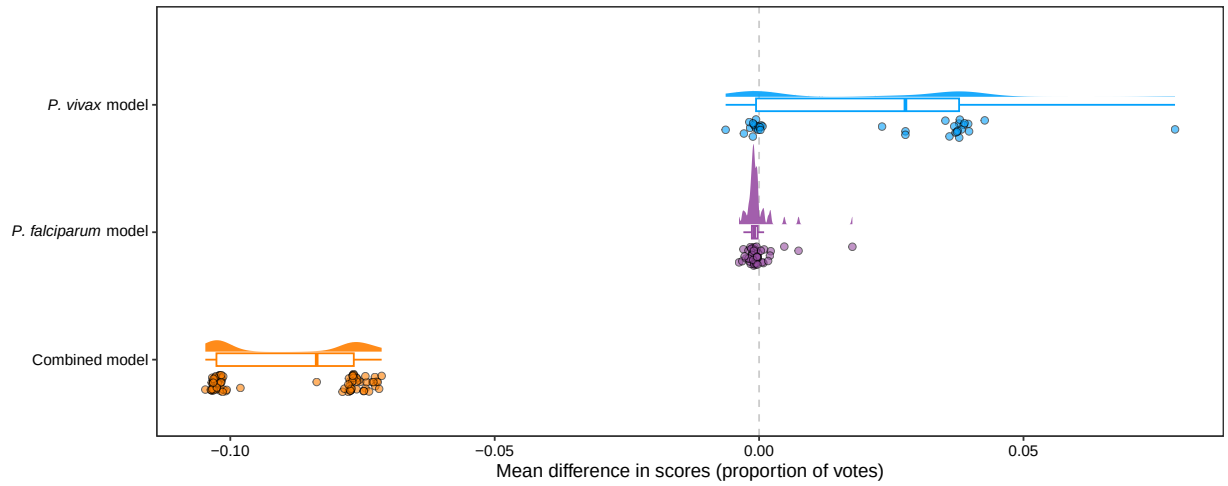


Figure D.5 Validation of PURF models. Mean differences in probability scores (proportion of votes) of the known antigens after removing the label of one of the known antigens in the input data sets for training the *P. vivax* (blue), *P. falciparum* (purple), and combined (orange) models. Boxplots show the median with first and third quartiles, and the whiskers display the extension of the 1.5 interquartile range from the first and third quartiles. The grey dashed line indicates zero mean difference in scores.

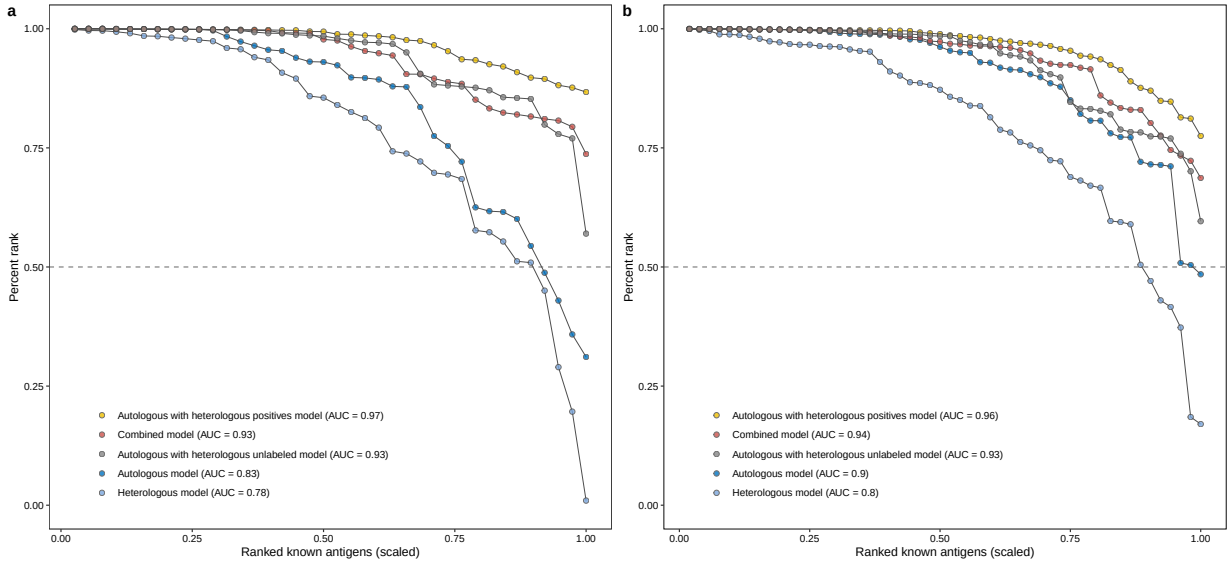


Figure D.6 Evaluation of known antigen predictions of PURF models. Line plots showing the percentile ranks of known antigens in the sets of *P. vivax* (**a**) and *P. falciparum* (**b**) proteins. Dots represent known antigens, which are connected by lines indicating antigen predictions from PURF models trained on different combinations of autologous and heterologous data. The x-axis shows the scaled ranks of the known antigens only, and the y-axis indicates the percentile ranks (the higher the better) of the known antigens across the entire *P. vivax* (**a**) or *P. falciparum* (**b**) data sets. The areas under the curves (AUC) are noted in the legend text. The grey horizontal dashed lines indicate the percentile rank of 0.5.

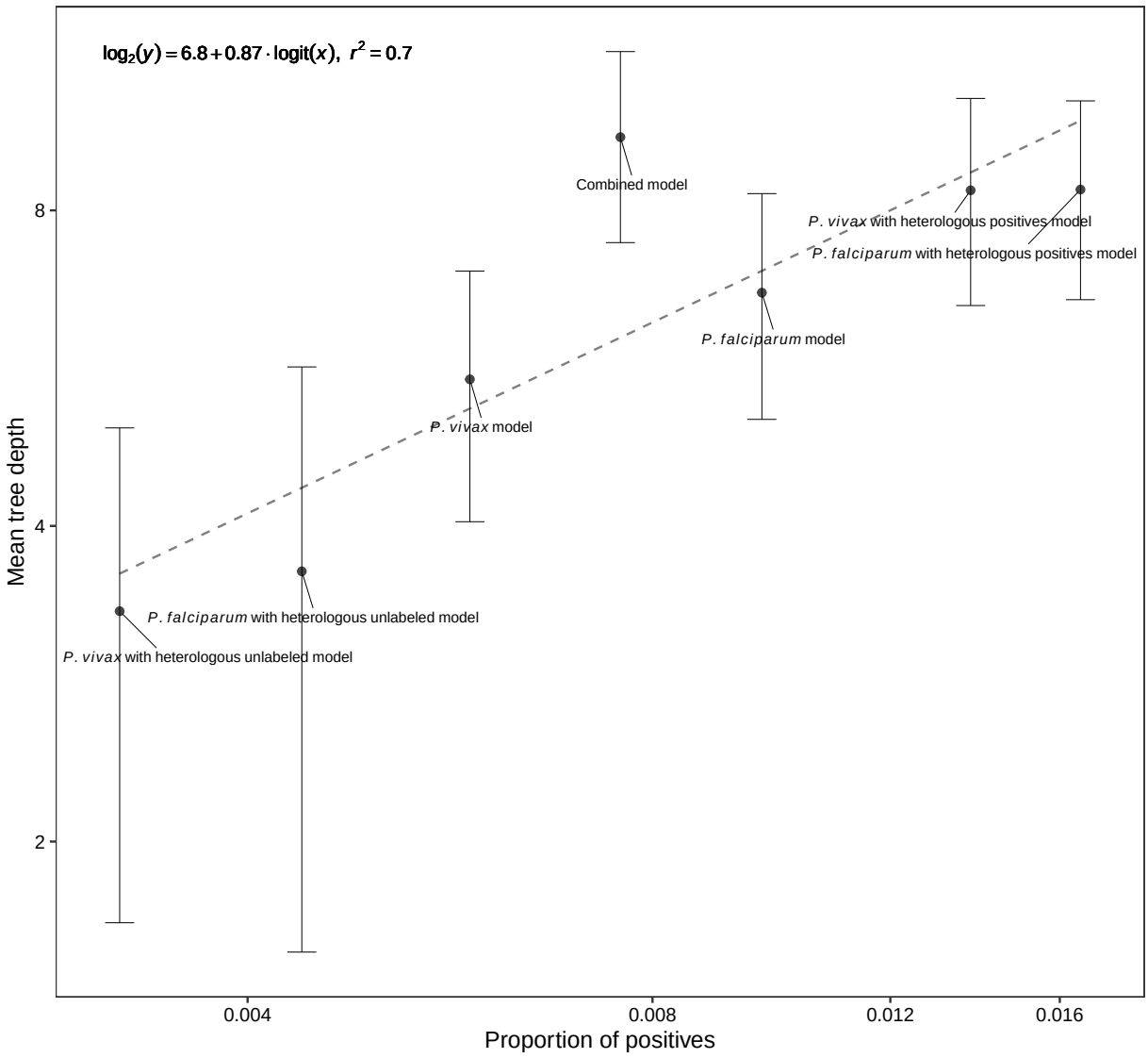


Figure D.7 Relationship between proportion of labeled positives in the data set and mean tree depth in the PURF model. The x-axis is logit-transformed and indicates the proportion of labeled positives in the data set. The y-axis is log_2-transformed and shows the mean depth across all trees in the PURF model. Dots represent PURF models ($n = 7$) with different combinations of autologous and heterologous data, and the model names are noted. Data are shown as mean ± SD. The grey dashed trend line conveys the linear regression model, where the formula and adjusted R^2 are indicated on the upper left corner. The p-value associated with the F-statistic of the linear regression is 0.012.

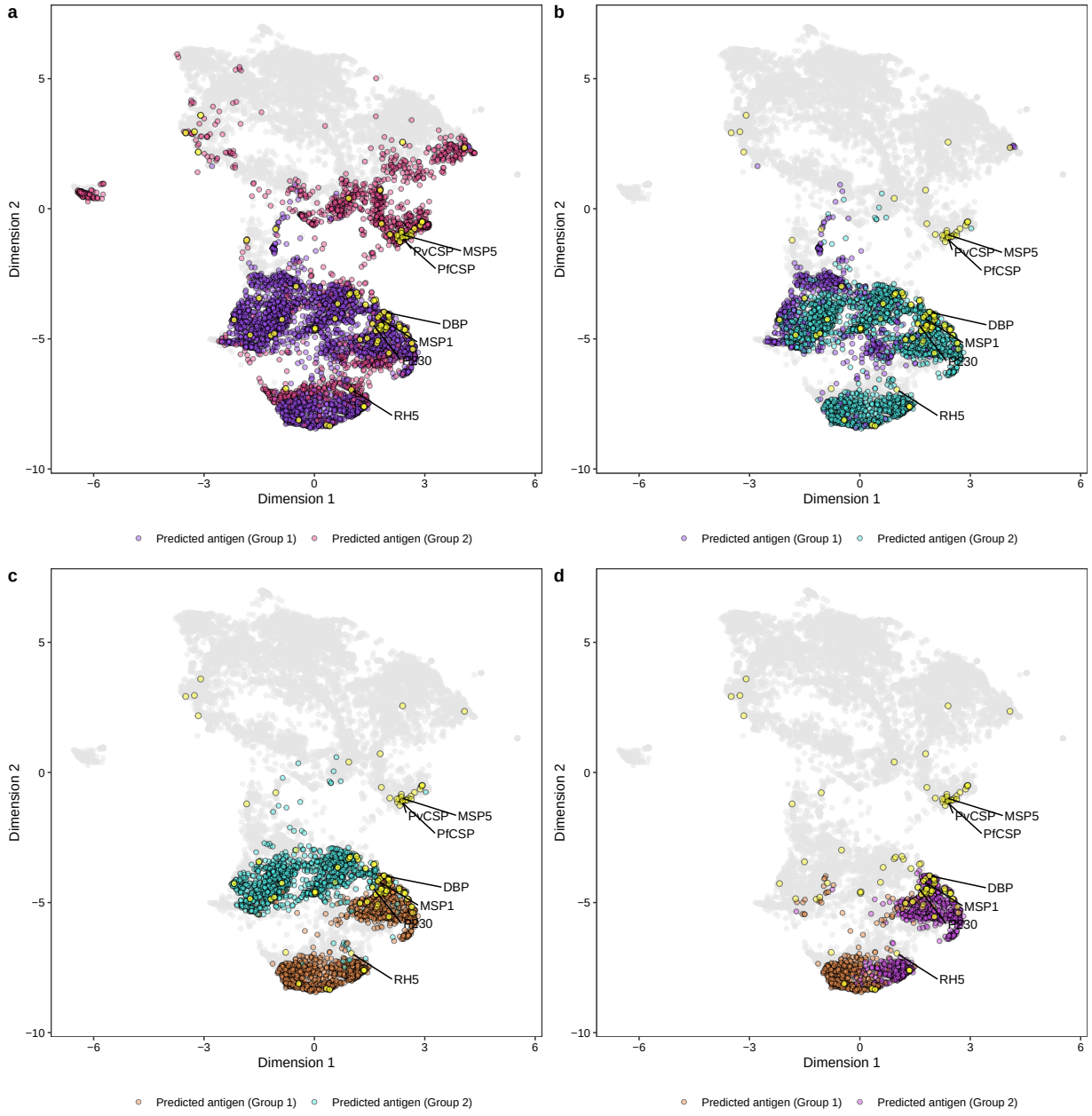


Figure D.8 Visualization of hierarchical clustering dendrogram investigation. Uniform manifold approximation and projection (UAMP) plots showing the iterative investigation of the dendrogram computed from the Euclidean distance matrix of predicted antigens derived from the combined PURF model. **a** The dendrogram was first cut into two groups, where the purple dots and pink dots show group 1 and group 2 predicted antigens, respectively. **b** Group 1 from **(a)** with a higher mean probability score was further divided into two groups, where purple and green dots indicate the new group 1 and group 2 predicted antigens, respectively. **c** The iteration continued and group 2 from **(b)** was selected because of the higher mean probability score, and further divided into another two groups separately represented by green and orange dots. **d** Group 1 from **(c)** was selected and another two groups of predicted antigens were generated based on the sub-dendrogram structure. The respective dot colors for the new group 1 and group 2 are orange and purple. Yellow dots are known antigens from both *Plamodium* species and the reference antigens are noted by text. Grey dots represent other unlabeled proteins.

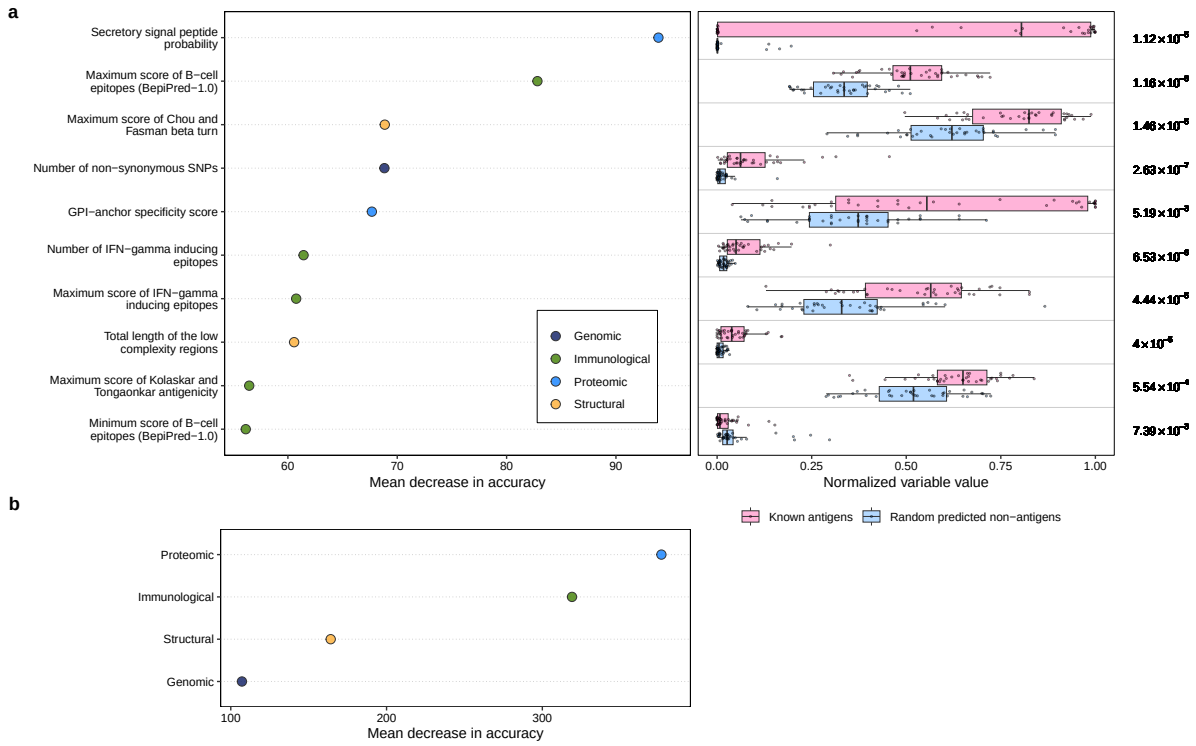


Figure D.9 Variable importance for the *P. vivax* model. **a** Top 10 important variables are shown on the left panel and categorized into genomic (dark blue), immunological (green), proteomic (blue), and structural (amber) variables. The x-axis indicate importance values in terms of mean decrease in prediction accuracy (scaled by the standard error) of the known antigens after variable permutation. The right panel displays comparisons of normalized variables values between the 38 known antigens (magenta dots) and the same number of randomly selected predicted non-antigens (blue dots). Boxplots show median with first and third quartiles, and the whiskers are the 1.5 interquartile range extended from the first and third quartiles. Two sided Mann–Whitney tests were computed, and the p-values were adjusted using the Benjamini–Hochberg procedure and noted on the right of the panel. **b** The importance of grouped variables by data types, where variables in the same variable data type were permuted together to calculate the mean decrease in accuracy of the known antigens.

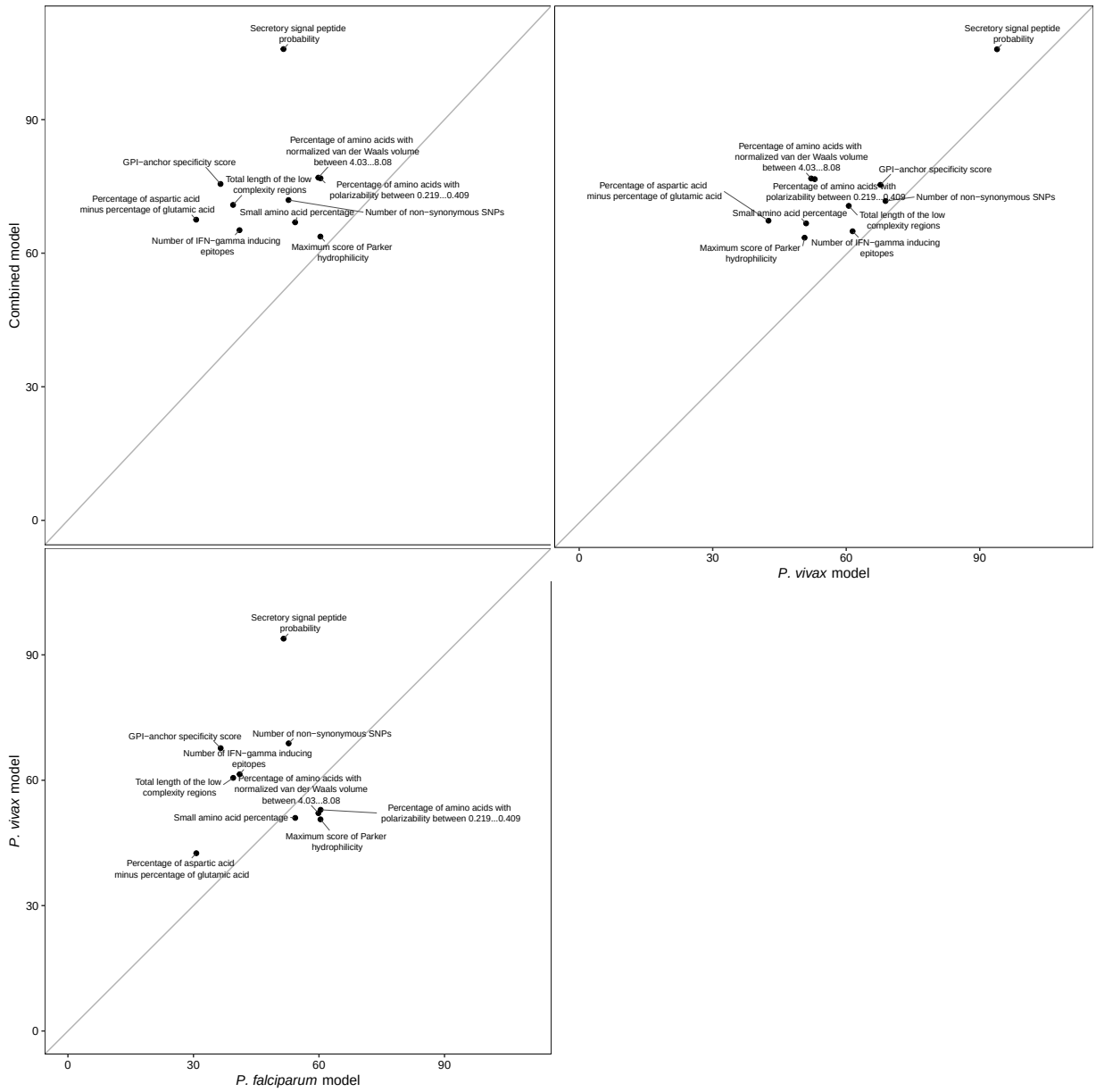


Figure D.10 Comparison of variable importance values between PURF models. Top 10 important variables were identified from the combined models, and the corresponding variable importance values are presented and compared for the combined, *P. vivax*, and *P. falciparum* model. Variable names are noted by text. The grey diagonal lines indicate where the importance values from both compared models are the same.

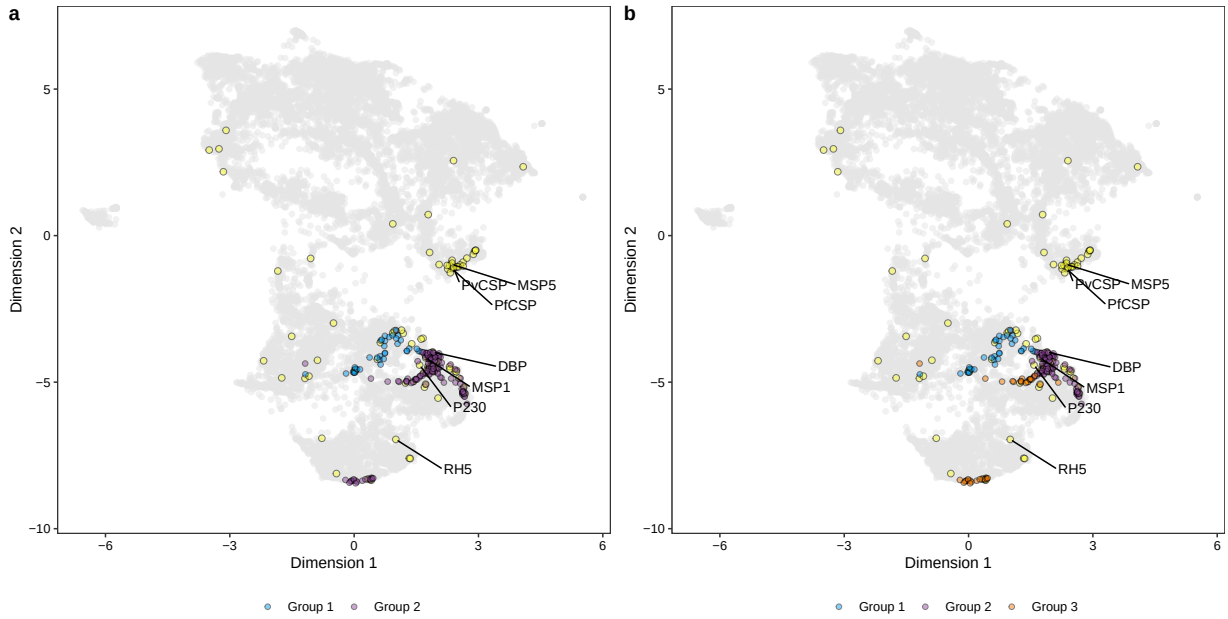


Figure D.11 Clustering analysis of top candidate antigens. Hierarchical clustering was performed on the top candidate Euclidean distance matrix derived from the combined PURF model. The dendrogram was cut into two (**a**) and three (**b**) groups and visualized on uniform manifold approximation and projection (UMAP) plots computed from the Euclidean distance matrix of the combined *P. vivax* and *P. falciparum* proteomes. **a** Blue and purple dots respectively represent group 1 and group 2 candidate antigens when cutting the dendrogram into two groups. **b** Blue, purple, and orange dots show group 1, group 2, and group 3 candidate antigens, respectively, when generating three groups from the dendrogram. Yellow dots are known antigens from both *Plasmodium* species with protein names of reference antigens annotated. Grey dots are other unlabeled proteins from both species.

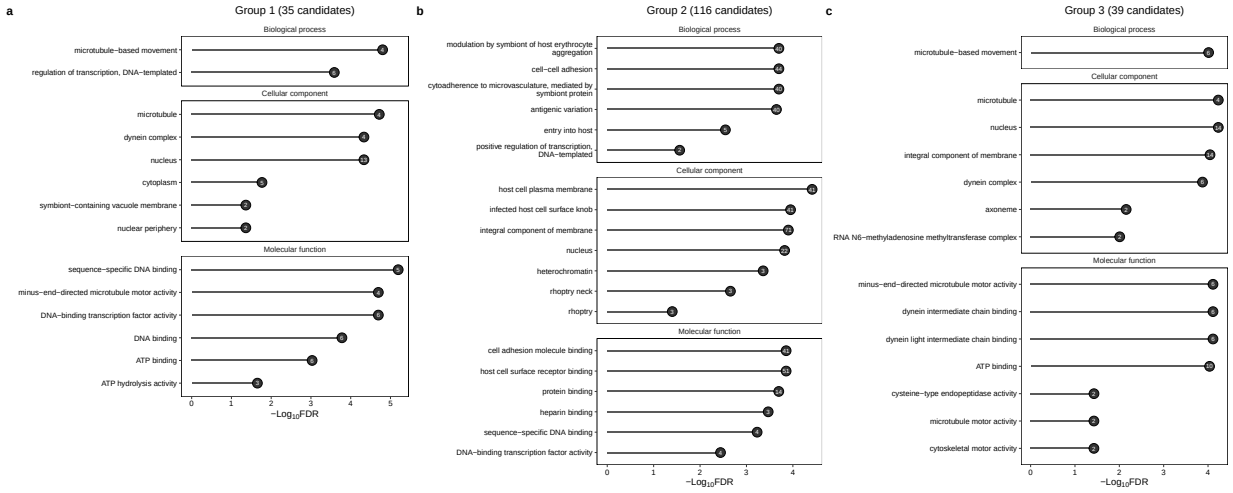


Figure D.12 Gene ontology (GO) enrichment analysis of candidate antigen groups. Plots showing enriched GO terms for group 1 (**a**), group 2 (**b**), and group 3 (**c**) candidate antigen genes, compared to the combined background proteomes of *P. vivax and P. falciparum*. GO terms with false discovery rate (FDR) <0.05 are shown on the *y*-axes and categorized into biological process, cellular component, and molecular function. The *x*-axes show log_{10}FDR, which are indicated by the grey bars of the corresponding GO terms. The values in the black dots indicate the number of antigen genes associated with the GO terms.

D.2 Supplementary Tables

Table D.1 Associations between *Plasmodium* species and antigen predictions from models trained on different combinations of autologous and heterologous data (CI: confidence interval).

PURF model	**Cramér's V**	χ^2 **test *p*-value**
Combined	0.08 (95% CI: 0.06, 0.10)	4.95×10^{-19}
P. vivax	0.10 (95% CI: 0.08, 0.12)	3.52×10^{-28}
P. falciparum	0.06 (95% CI: 0.04, 0.08)	7.08×10^{-12}
P. vivax with heterologous positives	0.61 (95% CI: 0.60, 0.62)	~ 0
P. falciparum with heterologous positives	0.62 (95% CI: 0.60, 0.63)	~ 0
P. vivax with heterologous unlabeled	0.80 (95% CI: 0.79, 0.81)	~ 0
P. falciparum with heterologous unlabeled	0.96 (95% CI: 0.95, 0.96)	~ 0

p-values $< 2.225074 \times 10^{-308}$ are reported as ~ 0.

Bibliography

[1] Hashem Koohy. The rise and fall of machine learning methods in biomedical research. *F1000Res*, 6:2012, 2017.

[2] Joe G Greener, Shaun M Kandathil, Lewis Moffat, and David T Jones. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol*, Sep 2021.

[3] Conor John Cremin, Sabyasachi Dash, and Xiaofeng Huang. Big data: historic advances and emerging trends in biomedical research. *Current Research in Biotechnology*, 4:138–151, 2022.

[4] David T Jones. Setting the standards for machine learning in biology. *Nat Rev Mol Cell Biol*, 20(11):659–660, 11 2019.

[5] Kristen Jaskie and Andreas Spanias. Positive and unlabeled learning algorithms and applications: a survey. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–8, 2019.

[6] Ashaben Patel, Kishore Cholkar, Vibhuti Agrahari, and Ashim K Mitra. Ocular drug delivery systems: an overview. *World J Pharmacol*, 2(2):47–64, 2013.

[7] Ripal Gaudana, Hari Krishna Ananthula, Ashwin Parenky, and Ashim K Mitra. Ocular drug delivery. *AAPS J*, 12(3):348–60, Sep 2010.

[8] Hajime Hisaeda, Koji Yasutomo, and Kunisuke Himeno. Malaria: immune evasion by parasites. *Int J Biochem Cell Biol*, 37(4):700–6, Apr 2005.

[9] R. Gaudana, H. K. Ananthula, A. Parenky, and A. K. Mitra. Ocular drug delivery. *AAPS J*, 12(3):348–60, 2010.

[10] A. Patel, K. Cholkar, V. Agrahari, and A. K. Mitra. Ocular drug delivery systems: an overview. *World J Pharmacol*, 2(2):47–64, 2013.

[11] B. L. Nordstrom, D. S. Friedman, E. Mozaffari, H. A. Quigley, and A. M. Walker. Persistence and adherence with topical glaucoma therapy. *Am J Ophthalmol*, 140(4):598–606, 2005.

[12] C. O. Okeke, H. A. Quigley, H. D. Jampel, G. S. Ying, R. J. Plyler, Y. Jiang, and D. S. Friedman. Adherence with topical glaucoma medication monitored electronically the travatan dosing aid study. *Ophthalmology*, 116(2):191–9, 2009.

[13] R. N. Weinreb, T. Aung, and F. A. Medeiros. The pathophysiology and treatment of glaucoma: a review. *JAMA*, 311(18):1901–11, 2014.

[14] Y. C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C. Y. Cheng. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*, 121(11):2081–90, 2014.

[15] W. L. Wong, X. Su, X. Li, C. M. Cheung, R. Klein, C. Y. Cheng, and T. Y. Wong. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob Health*, 2(2):e106–16, 2014.

[16] R. A. Lewis, W. C. Christie, D. G. Day, E. R. Craven, T. Walters, M. Bejanian, S. S. Lee, M. L. Goodkin, J. Zhang, S. M. Whitcup, M. R. Robinson, and S. R. Study Group Bimatoprost. Bimatoprost sustained-release implants for glaucoma therapy: 6-month results from a phase i/ii clinical trial. *Am J Ophthalmol*, 175:137–147, 2017.

[17] M. Shirley. Bimatoprost implant: first approval. *Drugs Aging*, 37(6):457–462, 2020.

[18] S. S. Lee, P. Hughes, A. D. Ross, and M. R. Robinson. Biodegradable implants for sustained drug release in the eye. *Pharm Res*, 27(10):2043–53, 2010.

[19] P. A. Campochiaro, D. M. Marcus, C. C. Awh, C. Regillo, A. P. Adamis, V. Bantseev, Y. Chiang, J. S. Ehrlich, S. Erickson, W. D. Hanley, J. Horvath, K. F. Maass, N. Singh, F. Tang, and G. Barteselli. The port delivery system with ranibizumab for neovascular age-related macular degeneration: results from the randomized phase 2 ladder clinical trial. *Ophthalmology*, 126(8):1141–1154, 2019.

[20] A. Malcles, C. Dot, N. Voirin, A. L. Vie, E. Agard, D. Bellocq, P. Denis, and L. Kodjikian. Safety of intravitreal dexamethasone implant (ozurdex): the safodex study. incidence and risk factors of ocular hypertension. *Retina*, 37(7):1352–1359, 2017.

[21] R. D. Jager, L. P. Aiello, S. C. Patel, and Jr. Cunningham, E. T. Risks of intravitreous injection: a comprehensive review. *Retina*, 24(5):676–98, 2004.

[22] N. Celik, R. Khoramnia, G. U. Auffarth, S. Sel, and C. S. Mayer. Complications of dexamethasone implants: risk factors, prevention, and clinical management. *Int J Ophthalmol*, 13(10):1612–1620, 2020.

[23] E. A. Thackaberry, C. Farman, F. Zhong, F. Lorget, K. Staflin, A. Cercillieux, P. E. Miller, C. Schuetz, D. Chang, A. Famili, A. L. Daugherty, K. Rajagopal, and V. Bantseev. Evaluation of the toxicity of intravitreally injected plga microspheres

and rods in monkeys and rabbits: rffects of depot size on inflammatory response. *Invest Ophthalmol Vis Sci*, 58(10):4274–4285, 2017.

[24] G. G. Giordano, P. Chevez-Barrios, M. F. Refojo, and C. A. Garcia. Biodegradation and tissue reaction to intravitreous biodegradable poly(d,l-lactic-co-glycolic)acid microspheres. *Curr Eye Res*, 14(9):761–8, 1995.

[25] A. K. Rimpela, M. Reinisalo, L. Hellinen, E. Grazhdankin, H. Kidron, A. Urtti, and E. M. Del Amo. Implications of melanin binding in ocular drug delivery. *Adv Drug Deliv Rev*, 126:23–43, 2018.

[26] E. Buszman and R. Rozanska. Interaction of quinidine, disopyramide and metoprolol with melanin in vitro in relation to drug-induced ocular toxicity. *Pharmazie*, 58(7):507–11, 2003.

[27] L. Mecklenburg and U. Schraermeyer. An overview on the toxic morphological changes in the retinal pigment epithelium after systemic compound administration. *Toxicol Pathol*, 35(2):252–67, 2007.

[28] Y. C. Kim, H. T. Hsueh, M. D. Shin, C. A. Berlinicke, H. Han, N. M. Anders, A. Hemingway, K. T. Leo, R. T. Chou, H. Kwon, M. B. Appell, U. Rai, P. Kolodziejski, C. Eberhart, I. Pitha, D. J. Zack, J. Hanes, and L. M. Ensign. A hypotonic gel-forming eye drop provides enhanced intraocular delivery of a kinase inhibitor with melanin-binding properties for sustained protection of retinal ganglion cells. *Drug Deliv Transl Res*, 12(4):826–837, 2022.

[29] A. Urtti, L. Salminen, H. Kujari, and V. Jäntti. Effect of ocular pigmentation on pilocarpine pharmacology in the rabbit eye. ii. drug response. *Int J Pharm*, 19(1):53–61, 1984.

[30] P. Jakubiak, M. Reutlinger, P. Mattei, F. Schuler, A. Urtti, and R. Alvarez-Sanchez. Understanding molecular drivers of melanin binding to support rational design of small molecule ophthalmic drugs. *J Med Chem*, 61(22):10106–10115, 2018.

[31] L. Wei, J. Tang, and Q. Zou. Skipcpp-pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. *BMC Genomics*, 18(Suppl 7):742, 2017.

[32] P. Agrawal, S. Bhalla, S. S. Usmani, S. Singh, K. Chaudhary, G. P. Raghava, and A. Gautam. Cppsite 2.0: a repository of experimentally validated cell-penetrating peptides. *Nucleic Acids Res*, 44(D1):D1098–103, 2016.

[33] S. Gupta, P. Kapoor, K. Chaudhary, A. Gautam, R. Kumar, Consortium Open Source Drug Discovery, and G. P. Raghava. In silico approach for predicting toxicity of peptides and proteins. *PLoS One*, 8(9):e73957, 2013.

[34] Y. J. Cheng, A. Q. Zhang, J. J. Hu, F. He, X. Zeng, and X. Z. Zhang. Multifunctional peptide-amphiphile end-capped mesoporous silica nanoparticles for tumor targeting drug delivery. *ACS Appl Mater Interfaces*, 9(3):2093–2103, 2017.

[35] M. Drexelius, A. Reinhardt, J. Grabeck, T. Cronenberg, F. Nitsche, P. F. Huesgen, B. Maier, and I. Neundorf. Multistep optimization of a cell-penetrating peptide towards its antimicrobial activity. *Biochem J*, 478(1):63–78, 2021.

[36] M. R. Felicio, O. N. Silva, S. Goncalves, N. C. Santos, and O. L. Franco. Peptides with dual antimicrobial and anticancer activities. *Front Chem*, 5:5, 2017.

[37] S. M. Lundberg and S. I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[38] L. Breiman. Random forests. *Mach Learn*, 45(1):5–32, 2001.

[39] R. C. Howell, E. Revskaya, V. Pazo, J. D. Nosanchuk, A. Casadevall, and E. Dadachova. Phage display library derived peptides that bind to human tumor melanin as potential vehicles for targeted radionuclide therapy of metastatic melanoma. *Bioconjug Chem*, 18(6):1739–48, 2007.

[40] J. D. Nosanchuk, P. Valadon, M. Feldmesser, and A. Casadevall. Melanization of cryptococcus neoformans in murine infection. *Mol Cell Biol*, 19(1):745–50, 1999.

[41] M. Laster and K. C. Norris. Lesson learned in mortality and kidney transplant outcomes among pediatric dialysis patients. *J Am Soc Nephrol*, 28(5):1334–1336, 2017.

[42] N. P. Cheruvu, A. C. Amrite, and U. B. Kompella. Effect of eye pigmentation on transscleral drug delivery. *Invest Ophthalmol Vis Sci*, 49(1):333–41, 2008.

[43] B. Speed, H. Z. Bu, W. F. Pool, G. W. Peng, E. Y. Wu, S. Patyna, C. Bello, and P. Kang. Pharmacokinetics, distribution, and metabolism of [14c]sunitinib in rats, monkeys, and humans. *Drug Metab Dispos*, 40(3):539–55, 2012.

[44] A. K. Rimpela, M. Schmitt, S. Latonen, M. Hagstrom, M. Antopolsky, J. A. Manzanares, H. Kidron, and A. Urtti. Drug distribution to retinal pigment epithelium: studies on melanin binding, cellular kinetics, and single photon emission computed tomography/computed tomography imaging. *Mol Pharm*, 13(9):2977–86, 2016.

[45] W. Du, S. Sun, Y. Xu, J. Li, C. Zhao, B. Lan, H. Chen, and L. Cheng. The effect of ocular pigmentation on transscleral delivery of triamcinolone acetonide. *J Ocul Pharmacol Ther*, 29(7):633–8, 2013.

[46] Y. C. Kim, M. D. Shin, S. F. Hackett, H. T. Hsueh, E. Silva R. Lima, A. Date, H. Han, B. J. Kim, A. Xiao, Y. Kim, L. Ogunnaike, N. M. Anders, A. Hemingway, P. He, A. S. Jun, P. J. McDonnell, C. Eberhart, I. Pitha, D. J. Zack, P. A. Campochiaro,

J. Hanes, and L. M. Ensign. Gelling hypotonic polymer solution for extended topical drug delivery to the eye. *Nat Biomed Eng*, 4(11):1053–1062, 2020.

[47] S. Bloch, A. R. Rosenthal, L. Friedman, and P. Caldarolla. Patient compliance in glaucoma. *Br J Ophthalmol*, 61(8):531–4, 1977.

[48] P. L. Kaufman and C. A. Rasmussen. Advances in glaucoma treatment and management: outflow drugs. *Invest Ophthalmol Vis Sci*, 53(5):2495–500, 2012.

[49] M. M. Hermann, D. Papaconstantinou, P. S. Muether, G. Georgopoulos, and M. Diestelhorst. Adherence with brimonidine in patients with glaucoma aware and not aware of electronic monitoring. *Acta Ophthalmol*, 89(4):e300–5, 2011.

[50] F. A. Medeiros, T. R. Walters, M. Kolko, M. Coote, M. Bejanian, M. L. Goodkin, Q. Guo, J. Zhang, M. R. Robinson, R. N. Weinreb, and Artemis Study Group. Phase 3, randomized, 20-month study of bimatoprost implant in open-angle glaucoma and ocular hypertension (artemis 1). *Ophthalmology*, 127(12):1627–1641, 2020.

[51] D. N. Hu, J. D. Simon, and T. Sarna. Role of ocular melanin in ophthalmic physiology and pathology. *Photochem Photobiol*, 84(3):639–44, 2008.

[52] A. K. Rimpela, M. Hagstrom, H. Kidron, and A. Urtti. Melanin targeting for intracellular drug delivery: quantification of bound and free drug in retinal pigment epithelial cells. *J Control Release*, 283:261–268, 2018.

[53] M. Salazar, K. Shimada, and P. N. Patil. Iris pigmentation and atropine mydriasis. *J Pharmacol Exp Ther*, 197(1):79–88, 1976.

[54] A. Henninot, J. C. Collins, and J. M. Nuss. The current state of peptide drug discovery: back to the future? *J Med Chem*, 61(4):1382–1414, 2018.

[55] A. A. Kaspar and J. M. Reichert. Future directions for peptide therapeutics development. *Drug Discov Today*, 18(17-18):807–17, 2013.

[56] M. Erak, K. Bellmann-Sickert, S. Els-Heindl, and A. G. Beck-Sickinger. Peptide chemistry toolbox–transforming natural peptides into peptide therapeutics. *Bioorg Med Chem*, 26(10):2759–2765, 2018.

[57] M. Muttenthaler, G. F. King, D. J. Adams, and P. F. Alewood. Trends in peptide drug discovery. *Nat Rev Drug Discov*, 20(4):309–325, 2021.

[58] D. Ghosh, X. Peng, J. Leal, and R. Mohanty. Peptides as drug delivery vehicles across biological barriers. *J Pharm Investig*, 48(1):89–111, 2018.

[59] A. Komin, L. M. Russell, K. A. Hristova, and P. C. Searson. Peptide-based strategies for enhanced cell uptake, transcellular transport, and circulation: mechanisms and challenges. *Adv Drug Deliv Rev*, 110-111:52–64, 2017.

[60] D. J. Begley. The blood-brain barrier: principles for targeting peptides and drugs to the central nervous system. *J Pharm Pharmacol*, 48(2):136–46, 1996.

[61] L. N. Johnson, S. M. Cashman, and R. Kumar-Singh. Cell-penetrating peptide for enhanced delivery of nucleic acids and drugs to ocular tissues including retina and cornea. *Mol Ther*, 16(1):107–14, 2008.

[62] L. N. Johnson, S. M. Cashman, S. P. Read, and R. Kumar-Singh. Cell penetrating peptide pod mediates delivery of recombinant proteins to retina, cornea and skin. *Vision Res*, 50(7):686–97, 2010.

[63] G. G. Jose, I. V. Larsen, J. Gauger, E. Carballo, R. Stern, R. Brummel, and C. R. Brandt. A cationic peptide, tat-cd degrees , inhibits herpes simplex virus type 1 ocular infection in vivo. *Invest Ophthalmol Vis Sci*, 54(2):1070–9, 2013.

[64] Y. Li, L. Li, Z. Li, J. Sheng, X. Zhang, D. Feng, X. Zhang, F. Yin, A. Wang, and F. Wang. Tat ptd-endostatin-rgd: a novel protein with anti-angiogenesis effect in retina via eye drops. *Biochim Biophys Acta*, 1860(10):2137–47, 2016.

[65] C. Liu, L. Tai, W. Zhang, G. Wei, W. Pan, and W. Lu. Penetratin, a potentially powerful absorption enhancer for noninvasive intraocular drug delivery. *Mol Pharm*, 11(4):1218–27, 2014.

[66] F. de Cogan, L. J. Hill, A. Lynch, P. J. Morgan-Warren, J. Lechner, M. R. Berwick, A. F. A. Peacock, M. Chen, R. A. H. Scott, H. Xu, and A. Logan. Topical delivery of anti-vegf drugs to the ocular posterior segment using cell-penetrating peptides. *Invest Ophthalmol Vis Sci*, 58(5):2578–2590, 2017.

[67] L. Tai, C. Liu, K. Jiang, X. Chen, L. Feng, W. Pan, G. Wei, and W. Lu. A novel penetratin-modified complex for noninvasive intraocular delivery of antisense oligonucleotides. *Int J Pharm*, 529(1-2):347–356, 2017.

[68] S. Pescina, C. Ostacolo, I. M. Gomez-Monterrey, M. Sala, A. Bertamino, F. Sonvico, C. Padula, P. Santi, A. Bianchera, and S. Nicoli. Cell penetrating peptides in ocular drug delivery: state of the art. *J Control Release*, 284:84–102, 2018.

[69] S. Pescina, M. Sala, C. Padula, M. C. Scala, A. Spensiero, S. Belletti, R. Gatti, E. Novellino, P. Campiglia, P. Santi, S. Nicoli, and C. Ostacolo. Design and synthesis of new cell penetrating peptides: diffusion and distribution inside the cornea. *Mol Pharm*, 13(11):3876–3883, 2016.

[70] Y. Wang, H. Lin, S. Lin, J. Qu, J. Xiao, Y. Huang, Y. Xiao, X. Fu, Y. Yang, and X. Li. Cell-penetrating peptide tat-mediated delivery of acidic fgf to retina and protection against ischemia-reperfusion injury in rats. *J Cell Mol Med*, 14(7):1998–2005, 2010.

[71] V. H. Lee and J. R. Robinson. Topical ocular drug delivery: recent developments and future challenges. *J Ocul Pharmacol*, 2(1):67–108, 1986.

[72] E. Cone-Kimball, C. Nguyen, E. N. Oglesby, M. E. Pease, M. R. Steinhart, and H. A. Quigley. Scleral structural alterations associated with chronic experimental intraocular pressure elevation in mice. *Mol Vis*, 19:2023–39, 2013.

[73] L. R. Schopf, A. M. Popov, E. M. Enlow, J. L. Bourassa, W. Z. Ong, P. Nowak, and H. Chen. Topical ocular drug delivery to the back of the eye by mucus-penetrating particles. *Transl Vis Sci Technol*, 4(3):11, 2015.

[74] G. P. Smith. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–7, 1985.

[75] D. S. Wilson, A. D. Keefe, and J. W. Szostak. The use of mrna display to select high-affinity protein-binding peptides. *Proc Natl Acad Sci U S A*, 98(7):3750–5, 2001.

[76] L. C. Szymczak, H. Y. Kuo, and M. Mrksich. Peptide arrays: development and application. *Anal Chem*, 90(1):266–282, 2018.

[77] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, II Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, and A. Tropsha. Qsar modeling: where have you been? where are you going to? *J Med Chem*, 57(12):4977–5010, 2014.

[78] X. Jing, Q. Dong, D. Hong, and R. Lu. Amino acid encoding methods for protein sequences: a comprehensive review and assessment. *IEEE/ACM Trans Comput Biol Bioinform*, 17(6):1918–1931, 2020.

[79] A. Gautam, K. Chaudhary, R. Kumar, A. Sharma, P. Kapoor, A. Tyagi, consortium Open source drug discovery, and G. P. Raghava. In silico approaches for designing highly effective cell penetrating peptides. *J Transl Med*, 11:74, 2013.

[80] M. S. Khatun, M. M. Hasan, and H. Kurata. Preaip: computational prediction of anti-inflammatory peptides by integrating multiple complementary features. *Front Genet*, 10:129, 2019.

[81] J. Yan, P. Bhadra, A. Li, P. Sethiya, L. Qin, H. K. Tai, K. H. Wong, and S. W. I. Siu. Deep-ampep30: improve short antimicrobial peptides prediction with deep learning. *Mol Ther Nucleic Acids*, 20:882–894, 2020.

[82] W. Ke, Z. Zha, J. F. Mukerabigwi, W. Chen, Y. Wang, C. He, and Z. Ge. Matrix metalloproteinase-responsive multifunctional peptide-linked amphiphilic block copolymers for intelligent systemic anticancer drug delivery. *Bioconjug Chem*, 28(8):2190–2198, 2017.

[83] X. Deng, R. Mai, C. Zhang, D. Yu, Y. Ren, G. Li, B. Cheng, L. Li, Z. Yu, and J. Chen. Discovery of novel cell-penetrating and tumor-targeting peptide-drug conjugate (pdc) for programmable delivery of paclitaxel and cancer treatment. *Eur J Med Chem*, 213:113050, 2021.

[84] J. Shi, J. G. Schellinger, and S. H. Pun. Engineering biodegradable and multifunctional peptide-based polymers for gene delivery. *J Biol Eng*, 7(1):25, 2013.

[85] K. Fosgerau and T. Hoffmann. Peptide therapeutics: current status and future directions. *Drug Discov Today*, 20(1):122–8, 2015.

[86] K. K. Yang, Z. Wu, and F. H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nat Methods*, 16(8):687–694, 2019.

[87] D. Brookes, H. Park, and J. Listgarten. Conditioning by adaptive sampling for robust design. *Proc of the 36th Int Conf on Mach Learn*, 97:773–782, 2019.

[88] S. Basith, B. Manavalan, T. Hwan Shin, and G. Lee. Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med Res Rev*, 40(4):1276–1314, 2020.

[89] M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Stat Appl Genet Mol Biol*, 6:Article 25, 2007.

[90] M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and example. *Technical Report*, 2003.

[91] J. Wong, T. Manderson, M. Abrahamowicz, D. L. Buckeridge, and R. Tamblyn. Can hyperparameter tuning improve the performance of a super learner?: a case study. *Epidemiology*, 30(4):521–531, 2019.

[92] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci U S A*, 116(44):22071–22080, 2019.

[93] R. Dybowski. Interpretable machine learning as a tool for scientific discovery in chemistry. *New J Chem*, 44:20914–20920, 2020.

[94] L. Wei, P. Xing, R. Su, G. Shi, Z. S. Ma, and Q. Zou. Cppred-rf: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *J Proteome Res*, 16(5):2044–2053, 2017.

[95] X. Fu, L. Ke, L. Cai, X. Chen, X. Ren, and M. Gao. Improved prediction of cell-penetrating peptides via effective orchestrating amino acid composition feature representation. *IEEE Access*, 7:163547–163555, 2019.

[96] X. Qiang, C. Zhou, X. Ye, P. F. Du, R. Su, and L. Wei. Cppred-fl: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. *Brief Bioinform*, 2018.

[97] J. Wasselius, H. Wallin, M. Abrahamson, and B. Ehinger. Cathepsin b in the rat eye. *Graefes Arch Clin Exp Ophthalmol*, 241(11):934–42, 2003.

[98] H. Appelqvist, P. Waster, K. Kagedal, and K. Ollinger. The lysosome: from waste bag to potential therapeutic target. *J Mol Cell Biol*, 5(4):214–26, 2013.

[99] P. E. Rakoczy, S. H. Sarks, N. Daw, and I. J. Constable. Distribution of cathepsin d in human eyes with or without age-related maculopathy. *Exp Eye Res*, 69(4):367–74, 1999.

[100] M. Goel, R. G. Picciani, R. K. Lee, and S. K. Bhattacharya. Aqueous humor dynamics: a review. *Open Ophthalmol J*, 4:52–9, 2010.

[101] D. Osorio and P. Rondón-Villarrea. Peptides: a package for data mining of antimicrobial peptides. *R Journal*, 7(1), 2015.

[102] N. Xiao, D. S. Cao, M. F. Zhu, and Q. S. Xu. protr/protrweb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, 31(11):1857–9, 2015.

[103] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[104] M. N. Wright and A. Ziegler. ranger: a fast implementation of random forests for high dimensional data in c++ and r. *J Stat Softw*, 77(1):1–17, 2017.

[105] H. Akaike. A new look at the statistical model identification. *IEEE Trans Automat Contr*, 19(6):716–723, 1974.

[106] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–44, 2015.

[107] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Ann Statist*, 29(5):1189–1232, 2001.

[108] T. Chen and C. Guestrin. Xgboost: a scalable tree boosting system. *Proc ACM SIGKDD Int*, 2016.

[109] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *J R Stat Soc Ser A Stat Soc*, 135(3):370–384, 1972.

[110] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Mach Learn*, 63:3–42, 2006.

[111] E. LeDell and S. Poirier. H2o automl: scalable automatic machine learning. *7th ICML AutoML Workshop*, 2020.

[112] H2O.ai. h2o: R interface for h2o. 2020.

[113] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B*, 57(1):289–300, 1995.

[114] J. F. Truchon and C. I. Bayly. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Chem. Inf. Model.*, 47:488–508, 2007.

[115] N. Wolkow, Y. Li, A. Maminishkis, Y. Song, O. Alekseev, J. Iacovelli, D. Song, J. C. Lee, and J. L. Dunaief. Iron upregulates melanogenesis in cultured retinal pigment epithelial cells. *Exp Eye Res*, 128:92–101, 2014.

[116] H2O.ai. h2o: Python interface for h2o. 2021.

[117] Riberio M. T., S. Singh, and C. Guestrin. "why should i trust you?": explaining the predictions of any classifier. *Proc ACM SIGKDD Int*, 2016.

[118] K. V. Chuang and M. J. Keiser. Adversarial controls for scientific machine learning. *ACS Chem Biol*, 13(10):2819–2821, 2018.

[119] van der Maaten L. Accelerating t-sne using tree-based algorithms. *J Mach Learn Res*, 15(1):3221–3245, 2014.

[120] L. R. Staben, S. G. Koenig, S. M. Lehar, R. Vandlen, D. Zhang, J. Chuh, S. F. Yu, C. Ng, J. Guo, Y. Liu, A. Fourie-O'Donohue, M. Go, X. Linghu, N. L. Segraves,

T. Wang, J. Chen, B. Wei, G. D. Phillips, K. Xu, K. R. Kozak, S. Mariathasan, J. A. Flygare, and T. H. Pillow. Targeted drug delivery through the traceless release of tertiary and heteroaryl amines from antibody-drug conjugates. *Nat Chem*, 8(12):1112–1119, 2016.

[121] K. R. Wilhelmus. The draize eye test. *Surv Ophthalmol*, 45(6):493–515, 2001.

[122] A. P. Tse, M. Shah, N. Jamal, and A. Shaikh. Glaucoma treatment adherence at a united kingdom general practice. *Eye (Lond)*, 30(8):1118–22, 2016.

[123] A. Chawla, J. N. McGalliard, and M. Batterbury. Use of eyedrops in glaucoma: how can we help to reduce non-compliance? *Acta Ophthalmol Scand*, 85(4):464, 2007.

[124] Y. C. Kim, M. D. Shin, S. F. Hackett, H. T. Hsueh, E. Silva R. Lima, A. Date, H. Han, B. J. Kim, A. Xiao, Y. Kim, L. Ogunnaike, N. M. Anders, A. Hemingway, P. He, A. S. Jun, P. J. McDonnell, C. Eberhart, I. Pitha, D. J. Zack, P. A. Campochiaro, J. Hanes, and L. M. Ensign. Gelling hypotonic polymer solution for extended topical drug delivery to the eye. *Nat Biomed Eng*, 4(11):1053–1062, 2020.

[125] D. S. Welsbie, Z. Yang, Y. Ge, K. L. Mitchell, X. Zhou, S. E. Martin, C. A. Berlinicke, Jr. Hackler, L., J. Fuller, J. Fu, L. H. Cao, B. Han, D. Auld, T. Xue, S. Hirai, L. Germain, C. Simard-Bisson, R. Blouin, J. V. Nguyen, C. H. Davis, R. A. Enke, S. L. Boye, S. L. Merbs, N. Marsh-Armstrong, W. W. Hauswirth, A. DiAntonio, R. W. Nickells, J. Inglese, J. Hanes, K. W. Yau, H. A. Quigley, and D. J. Zack. Functional genomic screening identifies dual leucine zipper kinase as a key mediator of retinal ganglion cell death. *Proc Natl Acad Sci U S A*, 110(10):4045–50, 2013.

[126] D. S. Welsbie, K. L. Mitchell, V. Jaskula-Ranga, V. M. Sluch, Z. Yang, J. Kim, E. Buehler, A. Patel, S. E. Martin, P. W. Zhang, Y. Ge, Y. Duan, J. Fuller, B. J. Kim, E. Hamed, X. Chamling, L. Lei, I. D. C. Fraser, Z. A. Ronai, C. A. Berlinicke, and D. J. Zack. Enhanced functional genomic screening identifies novel mediators of

dual leucine zipper kinase-dependent injury signaling in neurons. *Neuron*, 94(6):1142–1154 e6, 2017.

[127] Y. C. Kim, H. T. Hsueh, M. D. Shin, C. A. Berlinicke, H. Han, N. M. Anders, A. Hemingway, K. T. Leo, R. T. Chou, H. Kwon, M. B. Appell, U. Rai, P. Kolodziejski, C. Eberhart, I. Pitha, D. J. Zack, J. Hanes, and L. M. Ensign. A hypotonic gel-forming eye drop provides enhanced intraocular delivery of a kinase inhibitor with melanin-binding properties for sustained protection of retinal ganglion cells. *Drug Deliv Transl Res*, 12(4):826–837, 2022.

[128] A. K. Rimpela, M. Hagstrom, H. Kidron, and A. Urtti. Melanin targeting for intracellular drug delivery: quantification of bound and free drug in retinal pigment epithelial cells. *J Control Release*, 283:261–268, 2018.

[129] A. K. Rimpela, M. Reinisalo, L. Hellinen, E. Grazhdankin, H. Kidron, A. Urtti, and E. M. Del Amo. Implications of melanin binding in ocular drug delivery. *Adv Drug Deliv Rev*, 126:23–43, 2018.

[130] P. Jakubiak, M. Reutlinger, P. Mattei, F. Schuler, A. Urtti, and R. Alvarez-Sanchez. Understanding molecular drivers of melanin binding to support rational design of small molecule ophthalmic drugs. *J Med Chem*, 61(22):10106–10115, 2018.

[131] H. T. Hsueh, R.T. Chou, U. Rai, W. Liyanage, Y. C. Kim, M. Appell, J. Pejavar, K. T. Leo, C. Davison, P. Kolodziejski, A. Mozzer, H. Kwon, M. Sista, N. M. Anders, A. Hemingway, S.V.K. Rompicharla, M. Edward, I. Pitha, J. Hanes, M.P. Cummings, and L. M. Ensign. Machine learning-driven multifunctional peptide engineering for sustained ocular drug delivery. *Nat Commun*, 14:2509, 2023.

[132] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: a survey. *Int J Comput Vis*, 128(2):261–318, 2020.

[133] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst*, 28, 2015.

[134] B. L. Nordstrom, D. S. Friedman, E. Mozaffari, H. A. Quigley, and A. M. Walker. Persistence and adherence with topical glaucoma therapy. *Am J Ophthalmol*, 140(4):598–606, 2005.

[135] C. O. Okeke, H. A. Quigley, H. D. Jampel, G. S. Ying, R. J. Plyler, Y. Jiang, and D. S. Friedman. Interventions improve poor adherence with once daily glaucoma medications in electronically monitored patients. *Ophthalmology*, 116(12):2286–93, 2009.

[136] C. O. Okeke, H. A. Quigley, H. D. Jampel, G. S. Ying, R. J. Plyler, Y. Jiang, and D. S. Friedman. Adherence with topical glaucoma medication monitored electronically the travatan dosing aid study. *Ophthalmology*, 116(2):191–9, 2009.

[137] G. A. Rodrigues, D. Lutz, J. Shen, X. Yuan, H. Shen, J. Cunningham, and H. M. Rivers. Topical drug delivery to the posterior segment of the eye: addressing the challenge of preclinical to clinical translation. *Pharm Res*, 35(12):245, 2018.

[138] R. Singh, J. I. Wurzelmann, L. Ye, L. Henderson, M. Hossain, T. Trivedi, and D. S. Kelly. Clinical evaluation of pazopanib eye drops in healthy subjects and in subjects with neovascular age-related macular degeneration. *Retina*, 34(9):1787–95, 2014.

[139] D. N. Hu, J. D. Simon, and T. Sarna. Role of ocular melanin in ophthalmic physiology and pathology. *Photochem Photobiol*, 84(3):639–44, 2008.

[140] M. Salazar, K. Shimada, and P. N. Patil. Iris pigmentation and atropine mydriasis. *J Pharmacol Exp Ther*, 197(1):79–88, 1976.

[141] A. Urtti, L. Salminen, H. Kujari, and V. Jantti. Effect of ocular pigmentation on pilocarpine pharmacology in the rabbit eye. ii. drug response. *Int J Pharm*, 19(1):53–61, 1984.

[142] S. Bahrpeyma, M. Reinisalo, L. Hellinen, S. Auriola, E. M. Del Amo, and A. Urtti. Mechanisms of cellular retention of melanin bound drugs: experiments and computational modeling. *J Control Release*, 348:760–770, 2022.

[143] A. K. Rimpela, M. Schmitt, S. Latonen, M. Hagstrom, M. Antopolsky, J. A. Manzanares, H. Kidron, and A. Urtti. Drug distribution to retinal pigment epithelium: studies on melanin binding, cellular kinetics, and single photon emission computed tomography/computed tomography imaging. *Mol Pharm*, 13(9):2977–86, 2016.

[144] L. Hellinen, M. Hagstrom, H. Knuutila, M. Ruponen, A. Urtti, and M. Reinisalo. Characterization of artificially re-pigmented arpe-19 retinal pigment epithelial cell model. *Sci Rep*, 9(1):13761, 2019.

[145] P. Jakubiak, C. Cantrill, A. Urtti, and R. Alvarez-Sanchez. Establishment of an in vitro-in vivo correlation for melanin binding and the extension of the ocular half-life of small-molecule drugs. *Mol Pharm*, 16(12):4890–4901, 2019.

[146] L. N. Johnson, S. M. Cashman, and R. Kumar-Singh. Cell-penetrating peptide for enhanced delivery of nucleic acids and drugs to ocular tissues including retina and cornea. *Mol Ther*, 16(1):107–14, 2008.

[147] G. G. Jose, I. V. Larsen, J. Gauger, E. Carballo, R. Stern, R. Brummel, and C. R. Brandt. A cationic peptide, tat-cd degrees , inhibits herpes simplex virus type 1 ocular infection in vivo. *Invest Ophthalmol Vis Sci*, 54(2):1070–9, 2013.

[148] F. de Cogan, L. J. Hill, A. Lynch, P. J. Morgan-Warren, J. Lechner, M. R. Berwick, A. F. A. Peacock, M. Chen, R. A. H. Scott, H. Xu, and A. Logan. Topical delivery

of anti-vegf drugs to the ocular posterior segment using cell-penetrating peptides. *Invest Ophthalmol Vis Sci*, 58(5):2578–2590, 2017.

[149] S. Pescina, C. Ostacolo, I. M. Gomez-Monterrey, M. Sala, A. Bertamino, F. Sonvico, C. Padula, P. Santi, A. Bianchera, and S. Nicoli. Cell penetrating peptides in ocular drug delivery: state of the art. *J Control Release*, 284:84–102, 2018.

[150] M. Bhattacharya, A. Sadeghi, S. Sarkhel, M. Hagstrom, S. Bahrpeyma, E. Toropainen, S. Auriola, and A. Urtti. Release of functional dexamethasone by intracellular enzymes: a modular peptide-based strategy for ocular drug delivery. *J Control Release*, 327:584–594, 2020.

[151] Mary Ellen Pease, Donald J. Zack, Cynthia Berlinicke, Kristen Bloom, Frances Cone, Yuxia Wang, Ronald L. Klein, William W. Hauswirth, and Harry A. Quigley. Effect of cntf on retinal ganglion cell survival in experimental glaucoma. *Investig Ophthalmol Vis Sci*, 50(5):2194–2200, 2009.

[152] Ronald S. Harwerth, Joe L. Wheat, and Nalini V. Rangaswamy. Age-related losses of retinal ganglion cells and axons. *Invest Ophthalmol Vis Sci*, 49(10):4437–4443, 2008.

[153] Joel B. Sheffield. Imagej, a useful tool for biological image processing and analysis. *Microsc Microanal*, 13(S02):200–201, 2007.

[154] Michael D. Abràmoff, Paulo J. Magalhães, and Sunanda J. Ram. Image processing with imagej. *Biophotonics Int*, 11(7):36–42, 2004.

[155] Ana C. Dordea, Mark-Anthony Bray, Kaitlin Allen, David J. Logan, Fei Fei, Rajeev Malhotra, Meredith S. Gregory, Anne E. Carpenter, and Emmanuel S. Buys. An open-source computational tool to automatically quantify immunolabeled retinal ganglion cells. *Exp Eye Res*, 147:50–56, 2016.

[156] Anne E. Carpenter, Thouis R. Jones, Michael R. Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A. Guertin, Joo Han Chang, Robert A. Lindquist, and Jason Moffat. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol*, 7(10):R100, 2006.

[157] J. Wasselius, H. Wallin, M. Abrahamson, and B. Ehinger. Cathepsin b in the rat eye. *Graefes Arch Clin Exp Ophthalmol*, 241(11):934–42, 2003.

[158] H. Appelqvist, P. Waster, K. Kagedal, and K. Ollinger. The lysosome: from waste bag to potential therapeutic target. *J Mol Cell Biol*, 5(4):214–26, 2013.

[159] P. E. Rakoczy, S. H. Sarks, N. Daw, and I. J. Constable. Distribution of cathepsin d in human eyes with or without age-related maculopathy. *Exp Eye Res*, 69(4):367–74, 1999.

[160] M. Goel, R. G. Picciani, R. K. Lee, and S. K. Bhattacharya. Aqueous humor dynamics: a review. *Open Ophthalmol J*, 4:52–9, 2010.

[161] L. R. Staben, S. G. Koenig, S. Lehar, R. Vandlen, D. L. Zhang, J. Chuh, S. F. Yu, C. Ng, J. Guo, Y. Z. Liu, A. Fourie-O'Donohue, M. Go, L. H. Xin, N. L. Segraves, T. Wang, J. H. Chen, B. Q. Wei, G. D. L. Phillips, K. Y. Xu, K. R. Kozak, S. Mariathasan, J. A. Flygare, and T. H. Pillow. Targeted drug delivery through the traceless release of tertiary and heteroaryl amines from antibody-drug conjugates. *Nat Chemistry*, 8(12):1112–1119, 2016.

[162] N. P. Cheruvu, A. C. Amrite, and U. B. Kompella. Effect of eye pigmentation on transscleral drug delivery. *Investig Ophthalmol Vis Sci*, 49(1):333–41, 2008.

[163] H. T. Hsueh, Y. C. Kim, I. Pitha, M. D. Shin, C. A. Berlinicke, R. T. Chou, E. Kimball, J. Schaub, S. Quillen, K. T. Leo, H. Han, A. Xiao, Y. Kim, M. Appell, U. Rai, H. Kwon, P. Kolodziejski, L. Ogunnaike, N. M. Anders, A. Hemingway, J. L. Jefferys,

A. A. Date, C. Eberhart, T. V. Johnson, H. A. Quigley, D. J. Zack, J. Hanes, and L. M. Ensign. Ion-complex microcrystal formulation provides sustained delivery of a multimodal kinase inhibitor from the subconjunctival space for protection of retinal ganglion cells. *Pharmaceutics*, 13(5), 2021.

[164] WHO. World malaria report. *World Health Organization*, 2022.

[165] B. Balikagala, N. Fukuda, M. Ikeda, O. T. Katuro, S. I. Tachibana, M. Yamauchi, W. Opio, S. Emoto, D. A. Anywar, E. Kimura, N. M. Q. Palacpac, E. I. Odongo-Aginya, M. Ogwang, T. Horii, and T. Mita. Evidence of artemisinin-resistant malaria in africa. *N Engl J Med*, 385(13):1163–1171, 2021.

[166] C. L. Moyes, D. K. Athinya, T. Seethaler, K. E. Battle, M. Sinka, M. P. Hadi, J. Hemingway, M. Coleman, and P. A. Hancock. Evaluating insecticide resistance across african districts to aid malaria control decisions. *Proc Natl Acad Sci U S A*, 117(36):22042–22050, 2020.

[167] J. Nass and T. Efferth. Development of artemisinin resistance in malaria therapy. *Pharmacol Res*, 146:104275, 2019.

[168] C. V. Plowe, P. Alonso, and S. L. Hoffman. The potential role of vaccines in the elimination of falciparum malaria and the eventual eradication of malaria. *J Infect Dis*, 200(11):1646–9, 2009.

[169] D. A. Henderson. Lessons from the eradication campaigns. *Vaccine*, 17 Suppl 3:S53–5, 1999.

[170] P. E. Duffy. Current approaches to malaria vaccines. *Curr Opin Microbiol*, 70:102227, 2022.

[171] S. L. Takala and C. V. Plowe. Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. *Parasite Immunol*, 31(9):560–73, 2009.

[172] J. G. Beeson, L. Kurtovic, C. Dobano, D. H. Opi, J. A. Chan, G. Feng, M. F. Good, L. Reiling, and M. J. Boyle. Challenges and strategies for developing efficacious and long-lasting malaria vaccines. *Sci Transl Med*, 11(474), 2019.

[173] S Clinical Trials Partnership. RTS. Efficacy and safety of rts,s/as01 malaria vaccine with or without a booster dose in infants and children in africa: final results of a phase 3, individually randomised, controlled trial. *Lancet*, 386(9988):31–45, 2015.

[174] M. S. Datoo, H. M. Natama, A. Some, D. Bellamy, O. Traore, T. Rouamba, M. C. Tahita, N. F. A. Ido, P. Yameogo, D. Valia, A. Millogo, F. Ouedraogo, R. Soma, S. Sawadogo, F. Sorgho, K. Derra, E. Rouamba, F. Ramos-Lopez, M. Cairns, S. Provstgaard-Morys, J. Aboagye, A. Lawrie, R. Roberts, I. Valea, H. Sorgho, N. Williams, G. Glenn, L. Fries, J. Reimer, K. J. Ewer, U. Shaligram, A. V. S. Hill, and H. Tinto. Efficacy and immunogenicity of r21/matrix-m vaccine against clinical malaria after 2 years' follow-up in children in burkina faso: a phase 1/2b randomised controlled trial. *Lancet Infect Dis*, 22(12):1728–1736, 2022.

[175] M. J. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell. Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, 419(6906):498–511, 2002.

[176] M. A. Thera, O. K. Doumbo, D. Coulibaly, M. B. Laurens, A. Ouattara, A. K. Kone, A. B. Guindo, K. Traore, I. Traore, B. Kouriba, D. A. Diallo, I. Diarra, M. Daou, A. Dolo, Y. Tolo, M. S. Sissoko, A. Niangaly, M. Sissoko, S. Takala-Harrison, K. E. Lyke, Y. Wu, W. C. Blackwelder, O. Godeaux, J. Vekemans, M. C. Dubois, W. R. Ballou, J. Cohen, D. Thompson, T. Dube, L. Soisson, C. L. Diggs, B. House, D. E. Lanar, S. Dutta, Jr. Heppner, D. G., and C. V. Plowe. A field trial to assess a blood-stage malaria vaccine. *N Engl J Med*, 365(11):1004–13, 2011.

[177] A. Ouattara, S. Takala-Harrison, M. A. Thera, D. Coulibaly, A. Niangaly, R. Saye, Y. Tolo, S. Dutta, D. G. Heppner, L. Soisson, C. L. Diggs, J. Vekemans, J. Cohen, W. C. Blackwelder, T. Dube, M. B. Laurens, O. K. Doumbo, and C. V. Plowe. Molecular basis of allele-specific efficacy of a blood-stage malaria vaccine: vaccine development implications. *J Infect Dis*, 207(3):511–9, 2013.

[178] B. Genton, I. Betuela, I. Felger, F. Al-Yaman, R. F. Anders, A. Saul, L. Rare, M. Baisor, K. Lorry, G. V. Brown, D. Pye, D. O. Irving, T. A. Smith, H. P. Beck, and M. P. Alpers. A recombinant blood-stage malaria vaccine reduces plasmodium falciparum density and exerts selective pressure on parasite populations in a phase 1-2b trial in papua new guinea. *J Infect Dis*, 185(6):820–7, 2002.

[179] D. E. Neafsey, M. Juraska, T. Bedford, D. Benkeser, C. Valim, A. Griggs, M. Lievens, S. Abdulla, S. Adjei, T. Agbenyega, S. T. Agnandji, P. Aide, S. Anderson, D. Ansong, J. J. Aponte, K. P. Asante, P. Bejon, A. J. Birkett, M. Bruls, K. M. Connolly, U. D'Alessandro, C. Dobano, S. Gesase, B. Greenwood, J. Grimsby, H. Tinto, M. J. Hamel, I. Hoffman, P. Kamthunzi, S. Kariuki, P. G. Kremsner, A. Leach, B. Lell, N. J. Lennon, J. Lusingu, K. Marsh, F. Martinson, J. T. Molel, E. L. Moss, P. Njuguna, C. F. Ockenhouse, B. R. Ogutu, W. Otieno, L. Otieno, K. Otieno, S. Owusu-Agyei, D. J. Park, K. Pelle, D. Robbins, C. Russ, E. M. Ryan, J. Sacarlal, B. Sogoloff, H. Sorgho, M. Tanner, T. Theander, I. Valea, S. K. Volkman, Q. Yu, D. Lapierre,

B. W. Birren, P. B. Gilbert, and D. F. Wirth. Genetic diversity and protective efficacy of the rts,s/as01 malaria vaccine. *N Engl J Med*, 373(21):2025–2037, 2015.

[180] Rino Rappuoli. Reverse vaccinology. *Current opinion in microbiology*, 3(5):445–450, 2000.

[181] R. Rappuoli and A. Covacci. Reverse vaccinology and genomics. *Science*, 302(5645):602, 2003.

[182] R. Moxon, P. A. Reche, and R. Rappuoli. Editorial: reverse vaccinology. *Front Immunol*, 10:2776, 2019.

[183] M. Pizza, V. Scarlato, V. Masignani, M. M. Giuliani, B. Arico, M. Comanducci, G. T. Jennings, L. Baldi, E. Bartolini, B. Capecchi, C. L. Galeotti, E. Luzzi, R. Manetti, E. Marchetti, M. Mora, S. Nuti, G. Ratti, L. Santini, S. Savino, M. Scarselli, E. Storni, P. Zuo, M. Broeker, E. Hundt, B. Knapp, E. Blair, T. Mason, H. Tettelin, D. W. Hood, A. C. Jeffries, N. J. Saunders, D. M. Granoff, J. C. Venter, E. R. Moxon, G. Grandi, and R. Rappuoli. Identification of vaccine candidates against serogroup b meningococcus by whole-genome sequencing. *Science*, 287(5459):1816–20, 2000.

[184] H. Tettelin, N. J. Saunders, J. Heidelberg, A. C. Jeffries, K. E. Nelson, J. A. Eisen, K. A. Ketchum, D. W. Hood, J. F. Peden, R. J. Dodson, W. C. Nelson, M. L. Gwinn, R. DeBoy, J. D. Peterson, E. K. Hickey, D. H. Haft, S. L. Salzberg, O. White, R. D. Fleischmann, B. A. Dougherty, T. Mason, A. Ciecko, D. S. Parksey, E. Blair, H. Cittone, E. B. Clark, M. D. Cotton, T. R. Utterback, H. Khouri, H. Qin, J. Vamathevan, J. Gill, V. Scarlato, V. Masignani, M. Pizza, G. Grandi, L. Sun, H. O. Smith, C. M. Fraser, E. R. Moxon, R. Rappuoli, and J. C. Venter. Complete genome sequence of neisseria meningitidis serogroup b strain mc58. *Science*, 287(5459):1809–15, 2000.

[185] R. Rappuoli. Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine*, 19(17-19):2688–91, 2001.

[186] A. Sette and R. Rappuoli. Reverse vaccinology: developing vaccines in the era of genomics. *Immunity*, 33(4):530–41, 2010.

[187] M. I. Rashid, A. Naz, A. Ali, and S. Andleeb. Prediction of vaccine candidates against pseudomonas aeruginosa: an integrated genomics and proteomics approach. *Genomics*, 109(3-4):274–283, 2017.

[188] S. Talukdar, S. Zutshi, K. S. Prashanth, K. K. Saikia, and P. Kumar. Identification of potential vaccine candidates against streptococcus pneumoniae by reverse vaccinology approach. *Appl Biochem Biotechnol*, 172(6):3026–41, 2014.

[189] A. Huffman, E. Ong, J. Hur, A. D'Mello, H. Tettelin, and Y. He. Covid-19 vaccine design using reverse and structural vaccinology, ontology-based literature mining and machine learning. *Brief Bioinform*, 23(4), 2022.

[190] D. Maione, I. Margarit, C. D. Rinaudo, V. Masignani, M. Mora, M. Scarselli, H. Tettelin, C. Brettoni, E. T. Iacobini, R. Rosini, N. D'Agostino, L. Miorin, S. Buccato, M. Mariani, G. Galli, R. Nogarotto, V. Nardi-Dei, F. Vegni, C. Fraser, G. Mancuso, G. Teti, L. C. Madoff, L. C. Paoletti, R. Rappuoli, D. L. Kasper, J. L. Telford, and G. Grandi. Identification of a universal group b streptococcus vaccine by multiple genome screen. *Science*, 309(5731):148–50, 2005.

[191] L. Bruno, M. Cortese, R. Rappuoli, and M. Merola. Lessons from reverse vaccinology for viral vaccine design. *Curr Opin Virol*, 11:89–97, 2015.

[192] M. Dalsass, A. Brozzi, D. Medini, and R. Rappuoli. Comparison of open-source reverse vaccinology programs for bacterial vaccine antigen discovery. *Front Immunol*, 10:113, 2019.

[193] S. P. Singh, D. Srivastava, and B. N. Mishra. Genome-wide identification of novel vaccine candidates for plasmodium falciparum malaria using integrative bioinformatics approaches. *3 Biotech*, 7(5):318, 2017.

[194] M. Pritam, G. Singh, S. Swaroop, A. K. Singh, and S. P. Singh. Exploitation of reverse vaccinology and immunoinformatics as promising platform for genome-wide screening of new effective vaccine candidates against plasmodium falciparum. *BMC Bioinformatics*, 19(Suppl 13):468, 2019.

[195] A. I. Heinson, C. H. Woelk, and M. L. Newell. The promise of reverse vaccinology. *Int Health*, 7(2):85–9, 2015.

[196] J. Bekker and J. Davis. Learning from positive and unlabeled data: a survey. *Mach Learn*, 109:719–760, 2020.

[197] E. Sansone, F. G. B. De Natale, and Z. H. Zhou. Efficient training for positive unlabeled learning. *IEEE Trans Pattern Anal Mach Intell*, 41(11):2584–2598, 2019.

[198] P. Yang, X. Li, H. N. Chua, C. K. Kwoh, and S. K. Ng. Ensemble positive unlabeled learning for disease gene identification. *PLoS One*, 9(5):e97079, 2014.

[199] C. Aurrecoechea, J. Brestelli, B. P. Brunk, J. Dommer, S. Fischer, B. Gajria, X. Gao, A. Gingle, G. Grant, O. S. Harb, M. Heiges, F. Innamorato, J. Iodice, J. C. Kissinger, E. Kraemer, W. Li, J. A. Miller, V. Nayak, C. Pennington, D. F. Pinney, D. S. Roos, C. Ross, Jr. Stoeckert, C. J., C. Treatman, and H. Wang. Plasmodb: a functional genomic database for malaria parasites. *Nucleic Acids Res*, 37(Database issue):D539–43, 2009.

[200] U. Bohme, T. D. Otto, M. Sanders, C. I. Newbold, and M. Berriman. Progression of the canonical reference malaria parasite genome from 2002-2019. *Wellcome Open Res*, 4:58, 2019.

[201] C. Li and X.-L. Hua. Towards positive unlabeled learning for parallel data mining: a random forest framework. *Int Conf Adv Comput Appl*, pages 573–587, 2014.

[202] Min Zhang, Chengqi Wang, Thomas D Otto, Jenna Oberstaller, Xiangyun Liao, Swamy R Adapa, Kenneth Udenze, Iraad F Bronner, Deborah Casandra, and Matthew Mayho. Uncovering the essential genes of the human malaria parasite plasmodium falciparum by saturation mutagenesis. *Science*, 360(6388):eaap7847, 2018.

[203] V. M. Howick, A. J. C. Russell, T. Andrews, H. Heaton, A. J. Reid, K. Natarajan, H. Butungi, T. Metcalf, L. H. Verzier, J. C. Rayner, M. Berriman, J. K. Herren, O. Billker, M. Hemberg, A. M. Talman, and M. K. N. Lawniczak. The malaria cell atlas: single parasite transcriptomes across the complete plasmodium life cycle. *Science*, 365(6455), 2019.

[204] E. Real, V. M. Howick, F. A. Dahalan, K. Witmer, J. Cudini, C. Andradi-Brown, J. Blight, M. S. Davidson, S. K. Dogga, A. J. Reid, J. Baum, and M. K. N. Lawniczak. A single-cell atlas of plasmodium falciparum transmission through the mosquito. *Nat Commun*, 12(1):3196, 2021.

[205] A. J. Reid, A. M. Talman, H. M. Bennett, A. R. Gomes, M. J. Sanders, C. J. R. Illingworth, O. Billker, M. Berriman, and M. K. Lawniczak. Single-cell rna-seq reveals hidden transcriptional variation in malaria parasites. *Elife*, 7, 2018.

[206] R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, and B. Peters. The immune epitope database (iedb) 3.0. *Nucleic Acids Res*, 43(Database issue):D405–12, 2015.

[207] K. Gandhi, M. A. Thera, D. Coulibaly, K. Traore, A. B. Guindo, O. K. Doumbo, S. Takala-Harrison, and C. V. Plowe. Next generation sequencing to detect variation in the plasmodium falciparum circumsporozoite protein. *Am J Trop Med Hyg*, 86(5):775–81, 2012.

[208] A. Ouattara, T. M. Tran, S. Doumbo, M. Adams, S. Agrawal, A. Niangaly, S. Nelson-Owens, D. Doumtabe, Y. Tolo, A. Ongoiba, S. Takala-Harrison, B. Traore, J. C.

Silva, P. D. Crompton, O. K. Doumbo, and C. V. Plowe. Extent and dynamics of polymorphism in the malaria vaccine candidate plasmodium falciparum reticulocyte-binding protein homologue-5 in kalifabougou, mali. *Am J Trop Med Hyg*, 99(1):43–50, 2018.

[209] T. Wu, C. G. Black, L. Wang, A. R. Hibbs, and R. L. Coppel. Lack of sequence diversity in the gene encoding merozoite surface protein 5 of plasmodium falciparum. *Mol Biochem Parasitol*, 103(2):243–50, 1999.

[210] K. C. Williamson, H. Fujioka, M. Aikawa, and D. C. Kaslow. Stage-specific processing of pfs230, a plasmodium falciparum transmission-blocking vaccine candidate. *Mol Biochem Parasitol*, 78(1-2):161–9, 1996.

[211] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *J Artif Intell Res*, 16:321–357, 2002.

[212] K. V. Chuang and M. J. Keiser. Adversarial controls for scientific machine learning. *ACS Chem Biol*, 13(10):2819–2821, 2018.

[213] S. J. Draper, B. K. Sack, C. R. King, C. M. Nielsen, J. C. Rayner, M. K. Higgins, C. A. Long, and R. A. Seder. Malaria vaccines: recent advances and new horizons. *Cell Host Microbe*, 24(1):43–56, 2018.

[214] V. S. Moorthy, R. D. Newman, and J.-M. Okwo-Bele. Malaria vaccine technology roadmap. *Lancet*, 382(9906):1700–1701, 2013.

[215] S. J. Goodswen, P. J. Kennedy, and J. T. Ellis. A novel strategy for classifying the output from an in silico vaccine discovery pipeline for eukaryotic pathogens using machine learning algorithms. *BMC Bioinformatics*, 14:315, 2013.

[216] C. Aguttu, B. A. Okech, A. Mukisa, and G. W. Lubega. Screening and characterization of hypothetical proteins of plasmodium falciparum as novel vaccine candidates in the fight against malaria using reverse vaccinology. *J Genet Eng Biotechnol*, 19(1):103, 2021.

[217] F. Li, S. Dong, A. Leier, M. Han, X. Guo, J. Xu, X. Wang, S. Pan, C. Jia, Y. Zhang, G. I. Webb, L. J. M. Coin, C. Li, and J. Song. Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Brief Bioinform*, 23(1), 2022.

[218] L. Breiman. Random forests. *Mach Learn*, 45(1):5–32, 2001.

[219] R. L. Marchese Robinson, A. Palczewska, J. Palczewski, and N. Kidley. Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. *J Chem Inf Model*, 57(8):1773–1792, 2017.

[220] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci U S A*, 116(44):22071–22080, 2019.

[221] J. Molina-Franky, L. Cuy-Chaparro, A. Camargo, C. Reyes, M. Gomez, D. R. Salamanca, M. A. Patarroyo, and M. E. Patarroyo. Plasmodium falciparum preerythrocytic stage vaccine development. *Malar J*, 19(1):56, 2020.

[222] A. M. Minassian, S. E. Silk, J. R. Barrett, C. M. Nielsen, K. Miura, A. Diouf, C. Loos, J. K. Fallon, A. R. Michell, M. T. White, N. J. Edwards, I. D. Poulton, C. H. Mitton, R. O. Payne, M. Marks, H. Maxwell-Scott, A. Querol-Rubiera, K. Bisnauthsing, R. Batra, T. Ogrina, N. J. Brendish, Y. Themistocleous, T. A. Rawlinson, K. J. Ellis, D. Quinkert, M. Baker, R. Lopez Ramon, F. Ramos Lopez, L. Barfod, P. M. Folegatti, D. Silman, M. Datoo, I. J. Taylor, J. Jin, D. Pulido, A. D. Douglas, W. A. de Jongh, R. Smith, E. Berrie, A. R. Noe, C. L. Diggs, L. A. Soisson, R. Ashfield, S. N. Faust, A. L. Goodman, A. M. Lawrie, F. L. Nugent, G. Alter, C. A. Long, and

S. J. Draper. Reduced blood-stage malaria growth and immune correlates in humans following rh5 vaccination. *Med (N Y)*, 2(6):701–719 e19, 2021.

[223] R. L. Ord, M. Rodriguez, and C. A. Lobo. Malaria invasion ligand rh5 and its prime candidacy in blood-stage malaria vaccine design. *Hum Vaccin Immunother*, 11(6):1465–73, 2015.

[224] M. M. Medeiros, W. L. Fotoran, R. C. dalla Martha, T. H. Katsuragawa, L. H. Pereira da Silva, and G. Wunderlich. Natural antibody response to plasmodium falciparum merozoite antigens msp5, msp9 and eba175 is associated to clinical protection in the brazilian amazon. *BMC Infect Dis*, 13:608, 2013.

[225] C. Marin-Mogollon, M. van de Vegte-Bolmer, G. J. van Gemert, F. J. A. van Pul, J. Ramesar, A. S. Othman, H. Kroeze, J. Miao, L. Cui, K. C. Williamson, R. W. Sauerwein, C. J. Janse, and S. M. Khan. The plasmodium falciparum male gametocyte protein p230p, a paralog of p230, is vital for ookinete formation and mosquito transmission. *Sci Rep*, 8(1):14902, 2018.

[226] R. E. Sinden. A biologist's perspective on malaria vaccine development. *Hum Vaccin*, 6(1):3–11, 2010.

[227] L. M. Birkholtz, G. Blatch, T. L. Coetzer, H. C. Hoppe, E. Human, E. J. Morris, Z. Ngcete, L. Oldfield, R. Roth, A. Shonhai, L. Stephens, and A. I. Louw. Heterologous expression of plasmodial proteins for structural studies and functional annotation. *Malar J*, 7:197, 2008.

[228] R. G. Higbee, A. M. Byers, V. Dhir, D. Drake, H. G. Fahlenkamp, J. Gangur, A. Kachurin, O. Kachurina, D. Leistritz, Y. Ma, R. Mehta, E. Mishkin, J. Moser, L. Mosquera, M. Nguyen, R. Parkhill, S. Pawar, L. Poisson, G. Sanchez-Schmitz, B. Schanen, I. Singh, H. Song, T. Tapia, W. Warren, and V. Wittman. An immuno-

logic model for rapid vaccine assessment – a clinical trial in a test tube. *Altern Lab Anim*, 37 Suppl 1:19–27, 2009.

[229] K. A. Twohig, D. A. Pfeffer, J. K. Baird, R. N. Price, P. A. Zimmerman, S. I. Hay, P. W. Gething, K. E. Battle, and R. E. Howes. Growing evidence of plasmodium vivax across malaria-endemic africa. *PLoS Negl Trop Dis*, 13(1):e0007140, 2019.

[230] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

[231] C. J. Mungall, D. B. Emmert, and Consortium FlyBase. A chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, 23(13):i337–46, 2007.

[232] J. A. Young, Q. L. Fivelman, P. L. Blair, P. de la Vega, K. G. Le Roch, Y. Zhou, D. J. Carucci, D. A. Baker, and E. A. Winzeler. The plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol*, 143(1):67–79, 2005.

[233] E. Lasonder, S. R. Rijpma, B. C. van Schaijk, W. A. Hoeijmakers, P. R. Kensche, M. S. Gresnigt, A. Italiaander, M. W. Vos, R. Woestenenk, T. Bousema, G. R. Mair, S. M. Khan, C. J. Janse, R. Bartfai, and R. W. Sauerwein. Integrated transcriptomic and proteomic analyses of p. falciparum gametocytes: molecular insight into sex-specific processes and translational repression. *Nucleic Acids Res*, 44(13):6087–101, 2016.

[234] T. D. Otto, D. Wilinski, S. Assefa, T. M. Keane, L. R. Sarry, U. Bohme, J. Lemieux, B. Barrell, A. Pain, M. Berriman, C. Newbold, and M. Llinas. New insights into the blood-stage transcriptome of plasmodium falciparum using rna-seq. *Mol Microbiol*, 76(1):12–24, 2010.

[235] T. N. Siegel, C. C. Hon, Q. Zhang, J. J. Lopez-Rubio, C. Scheidig-Benatar, R. M. Martins, O. Sismeiro, J. Y. Coppee, and A. Scherf. Strand-specific rna-seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in plasmodium falciparum. *BMC Genomics*, 15:150, 2014.

[236] G. Zanghi, S. S. Vembar, S. Baumgarten, S. Ding, J. Guizetti, J. M. Bryant, D. Mattei, A. T. R. Jensen, L. Renia, Y. S. Goh, R. Sauerwein, C. C. Hermsen, J. F. Franetich, M. Bordessoulles, O. Silvie, V. Soulard, O. Scatton, P. Chen, S. Mecheri, D. Mazier, and A. Scherf. A specific pfemp1 is expressed in p. falciparum sporozoites and plays a role in hepatocyte infection. *Cell Rep*, 22(11):2951–2963, 2018.

[237] E. L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6:175–82, 1998.

[238] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol*, 305(3):567–80, 2001.

[239] J. C. Wootton and S. Federhen. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol*, 266:554–71, 1996.

[240] P. Y. Chou and G. D. Fasman. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol*, 47:45–148, 1978.

[241] E. A. Emini, J. V. Hughes, D. S. Perlow, and J. Boger. Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol*, 55(3):836–9, 1985.

[242] P. Karplus and G. Schulz. Prediction of chain flexibility in proteins. *Naturwissenschaften*, 72:212–213, 1985.

[243] C. S. Yu, Y. C. Chen, C. H. Lu, and J. K. Hwang. Prediction of protein subcellular localization. *Proteins*, 64(3):643–51, 2006.

[244] F. A. Ansari, N. Kumar, M. Bala Subramanyam, M. Gnanamani, and S. Ramachandran. Maap: malarial adhesins and adhesin-like proteins predictor. *Proteins*, 70(3):659–66, 2008.

[245] D. Osorio and P. Rondón-Villarrea. Peptides: a package for data mining of antimicrobial peptides. *R Journal*, 7(1), 2015.

[246] N. Xiao, D. S. Cao, M. F. Zhu, and Q. S. Xu. protr/protrweb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, 31(11):1857–9, 2015.

[247] J. M. Parker, D. Guo, and R. S. Hodges. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites. *Biochemistry*, 25(19):5425–32, 1986.

[248] A. Pierleoni, P. L. Martelli, and R. Casadio. Predgpi: a gpi-anchor predictor. *BMC Bioinformatics*, 9:392, 2008.

[249] J. J. Almagro Armenteros, K. D. Tsirigos, C. K. Sonderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne, and H. Nielsen. Signalp 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*, 37(4):420–423, 2019.

[250] M. Hebditch and J. Warwicker. Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. *PeerJ*, 7:e8199, 2019.

[251] J. S. Chauhan, A. Rao, and G. P. Raghava. In silico platform for prediction of n-, o- and c-glycosites in eukaryotic protein sequences. *PLoS One*, 8(6):e67008, 2013.

[252] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. Blast+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009.

[253] P. Oyarzun, J. J. Ellis, M. Boden, and B. Kobe. Predivac: Cd4+ t-cell epitope prediction for vaccine design that covers 95 *BMC Bioinformatics*, 14:52, 2013.

[254] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili. Bepipred-2.0: improving sequence-based b-cell epitope prediction using conformational epitopes. *Nucleic Acids Res*, 45(W1):W24–W29, 2017.

[255] J. E. Larsen, O. Lund, and M. Nielsen. Improved method for predicting linear b-cell epitopes. *Immunome Res*, 2:2, 2006.

[256] S. Saha and G. P. Raghava. Prediction of continuous b-cell epitopes in an antigen using recurrent neural network. *Proteins*, 65(1):40–8, 2006.

[257] M. Bhasin and G. P. Raghava. Prediction of ctl epitopes using qm, svm and ann techniques. *Vaccine*, 22(23-24):3195–204, 2004.

[258] G. Nagpal, S. S. Usmani, S. K. Dhanda, H. Kaur, S. Singh, M. Sharma, and G. P. Raghava. Computer-aided designing of immunosuppressive peptides based on il-10 inducing potential. *Sci Rep*, 7:42851, 2017.

[259] M. Bhasin and G. P. Raghava. Analysis and prediction of affinity of tap binding peptides using cascade svm. *Protein Sci*, 13(3):596–607, 2004.

[260] M. Nielsen, C. Lundegaard, P. Worning, S. L. Lauemoller, K. Lamberth, S. Buus, S. Brunak, and O. Lund. Reliable prediction of t-cell epitopes using neural networks with novel sequence representations. *Protein Sci*, 12(5):1007–17, 2003.

[261] H. H. Bui, J. Sidney, B. Peters, M. Sathiamurthy, A. Sinichi, K. A. Purton, B. R. Mothe, F. V. Chisari, D. I. Watkins, and A. Sette. Automated generation and evaluation of specific mhc binding predictive tools: Arb matrix applications. *Immunogenetics*, 57(5):304–14, 2005.

[262] A. S. Kolaskar and P. C. Tongaonkar. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett*, 276(1-2):172–4, 1990.

[263] J. J. Calis, M. Maybeno, J. A. Greenbaum, D. Weiskopf, A. D. De Silva, A. Sette, C. Kesmir, and B. Peters. Properties of mhc class i presented peptides that enhance immunogenicity. *PLoS Comput Biol*, 9(10):e1003266, 2013.

[264] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau. Scikit-learn: machine learning in python. *J Mach Learn Res*, 12:2825–2830, 2011.

[265] C. Gini. On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, 208(1):73–79, 1936.

[266] F. De Comité, F. Denis, R. Gilleron, and F. Letouzey. Positive and unlabeled examples help learning. *Int Conf Algo Learn Theory*, pages 219–230, 1999.

[267] S. Jain, M. White, and P. Radivojac. Recovering true classifier performance in positive-unlabeled learning. *AAAI Conf Artif Intel*, 2017.

[268] Y Yu. mixr: an r package for finite mixture modeling for both raw and binned data. *J Open Source Softw*, 7(69):4031, 2022.

[269] H. W. Borchers. pracma: practical numerical math functions (version 2.3.8). *R Package*, 2022.

[270] Z. Cheng, S. Zhou, and J. Guan. Computationally predicting protein-rna interactions using only positive and unlabeled examples. *J Bioinform Comput Biol*, 13(3):1541005, 2015.

[271] S. Jadhav and A. Mukhopadhyay. Computing a centerpoint of a finite planar set of points in linear time. *Discrete Comput Geom*, 12:291–312, 1994.

[272] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Technical Report, Stanford*, 2000.

[273] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of data clusters via the gap statistic. *J R Stat Soc Series B*, 63:411–423, 2001.

[274] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*, 20:53–65, 1987.

[275] L. McInnes, J. Healy, and J. Melville. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv*, 2018.

[276] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B*, 57(1):289–300, 1995.

[277] D. V. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramirez, A. Warwick Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, W. Dampier, C. Dessimoz, P. Flick, and H. Tang. Goatools: a python library for gene ontology analyses. *Sci Rep*, 8(1):10872, 2018.

[278] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, 2000.

[279] Consortium Gene Ontology. The gene ontology resource: enriching a gold mine. *Nucleic Acids Res*, 49(D1):D325–D334, 2021.

[280] WHO. World malaria report 2022. *World Health Organization, Geneva*, 2022.

[281] B. Balikagala, N. Fukuda, M. Ikeda, O. T. Katuro, S. I. Tachibana, M. Yamauchi, W. Opio, S. Emoto, D. A. Anywar, E. Kimura, N. M. Q. Palacpac, E. I. Odongo-Aginya, M. Ogwang, T. Horii, and T. Mita. Evidence of artemisinin-resistant malaria in africa. *N Engl J Med*, 385(13):1163–1171, 2021.

[282] C. L. Moyes, D. K. Athinya, T. Seethaler, K. E. Battle, M. Sinka, M. P. Hadi, J. Hemingway, M. Coleman, and P. A. Hancock. Evaluating insecticide resistance across african districts to aid malaria control decisions. *Proc Natl Acad Sci U S A*, 117(36):22042–22050, 2020.

[283] J. Nass and T. Efferth. Development of artemisinin resistance in malaria therapy. *Pharmacol Res*, 146:104275, 2019.

[284] C. V. Plowe, P. Alonso, and S. L. Hoffman. The potential role of vaccines in the elimination of falciparum malaria and the eventual eradication of malaria. *J Infect Dis*, 200(11):1646–9, 2009.

[285] D. A. Henderson. Lessons from the eradication campaigns. *Vaccine*, 17 Suppl 3:S53–5, 1999.

[286] I. Mueller, A. R. Shakri, and C. E. Chitnis. Development of vaccines for plasmodium vivax malaria. *Vaccine*, 33(52):7489–95, 2015.

[287] J. G. Beeson, L. Kurtovic, C. Dobano, D. H. Opi, J. A. Chan, G. Feng, M. F. Good, L. Reiling, and M. J. Boyle. Challenges and strategies for developing efficacious and long-lasting malaria vaccines. *Sci Transl Med*, 11(474), 2019.

[288] M. R. Galinski and J. W. Barnwell. Plasmodium vivax: who cares? *Malar J*, 7 Suppl 1(Suppl 1):S9, 2008.

[289] D. E. Neafsey, K. Galinsky, R. H. Jiang, L. Young, S. M. Sykes, S. Saif, S. Gujja, J. M. Goldberg, S. Young, Q. Zeng, S. B. Chapman, A. P. Dash, A. R. Anvikar, P. L. Sutton, B. W. Birren, A. A. Escalante, J. W. Barnwell, and J. M. Carlton. The malaria parasite plasmodium vivax exhibits greater genetic diversity than plasmodium falciparum. *Nat Genet*, 44(9):1046–50, 2012.

[290] D. E. Neafsey, M. Juraska, T. Bedford, D. Benkeser, C. Valim, A. Griggs, M. Lievens, S. Abdulla, S. Adjei, T. Agbenyega, S. T. Agnandji, P. Aide, S. Anderson, D. Ansong, J. J. Aponte, K. P. Asante, P. Bejon, A. J. Birkett, M. Bruls, K. M. Connolly, U. D'Alessandro, C. Dobano, S. Gesase, B. Greenwood, J. Grimsby, H. Tinto, M. J. Hamel, I. Hoffman, P. Kamthunzi, S. Kariuki, P. G. Kremsner, A. Leach, B. Lell, N. J. Lennon, J. Lusingu, K. Marsh, F. Martinson, J. T. Molel, E. L. Moss, P. Njuguna, C. F. Ockenhouse, B. R. Ogutu, W. Otieno, L. Otieno, K. Otieno, S. Owusu-Agyei, D. J. Park, K. Pelle, D. Robbins, C. Russ, E. M. Ryan, J. Sacarlal, B. Sogoloff, H. Sorgho, M. Tanner, T. Theander, I. Valea, S. K. Volkman, Q. Yu, D. Lapierre, B. W. Birren, P. B. Gilbert, and D. F. Wirth. Genetic diversity and protective efficacy of the rts,s/as01 malaria vaccine. *N Engl J Med*, 373(21):2025–2037, 2015.

[291] S. L. Takala and C. V. Plowe. Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. *Parasite Immunol*, 31(9):560–73, 2009.

[292] R. Rappuoli and A. Covacci. Reverse vaccinology and genomics. *Science*, 302(5645):602, 2003.

[293] R. Moxon, P. A. Reche, and R. Rappuoli. Editorial: reverse vaccinology. *Front Immunol*, 10:2776, 2019.

[294] M. Pizza, V. Scarlato, V. Masignani, M. M. Giuliani, B. Arico, M. Comanducci, G. T. Jennings, L. Baldi, E. Bartolini, B. Capecchi, C. L. Galeotti, E. Luzzi, R. Manetti, E. Marchetti, M. Mora, S. Nuti, G. Ratti, L. Santini, S. Savino, M. Scarselli, E. Storni, P. Zuo, M. Broeker, E. Hundt, B. Knapp, E. Blair, T. Mason, H. Tettelin, D. W. Hood, A. C. Jeffries, N. J. Saunders, D. M. Granoff, J. C. Venter, E. R. Moxon, G. Grandi, and R. Rappuoli. Identification of vaccine candidates against serogroup b meningococcus by whole-genome sequencing. *Science*, 287(5459):1816–20, 2000.

[295] H. Tettelin, N. J. Saunders, J. Heidelberg, A. C. Jeffries, K. E. Nelson, J. A. Eisen, K. A. Ketchum, D. W. Hood, J. F. Peden, R. J. Dodson, W. C. Nelson, M. L. Gwinn, R. DeBoy, J. D. Peterson, E. K. Hickey, D. H. Haft, S. L. Salzberg, O. White, R. D. Fleischmann, B. A. Dougherty, T. Mason, A. Ciecko, D. S. Parksey, E. Blair, H. Cittone, E. B. Clark, M. D. Cotton, T. R. Utterback, H. Khouri, H. Qin, J. Vamathevan, J. Gill, V. Scarlato, V. Masignani, M. Pizza, G. Grandi, L. Sun, H. O. Smith, C. M. Fraser, E. R. Moxon, R. Rappuoli, and J. C. Venter. Complete genome sequence of neisseria meningitidis serogroup b strain mc58. *Science*, 287(5459):1809–15, 2000.

[296] R. Rappuoli. Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine*, 19(17-19):2688–91, 2001.

[297] A. Sette and R. Rappuoli. Reverse vaccinology: developing vaccines in the era of genomics. *Immunity*, 33(4):530–41, 2010.

[298] S. P. Singh, D. Srivastava, and B. N. Mishra. Genome-wide identification of novel vaccine candidates for plasmodium falciparum malaria using integrative bioinformatics approaches. *3 Biotech*, 7(5):318, 2017.

[299] M. Pritam, G. Singh, S. Swaroop, A. K. Singh, and S. P. Singh. Exploitation of reverse vaccinology and immunoinformatics as promising platform for genome-wide

screening of new effective vaccine candidates against plasmodium falciparum. *BMC Bioinformatics*, 19(Suppl 13):468, 2019.

[300] Renee Ti Chou, Amed Ouattara, Matthew Adams, Andrea A. Berry, Shannon Takala-Harrison, and Michael P. Cummings. Positive-unlabeled learning identifies vaccine candidate antigens in the malaria parasite plasmodium falciparum. *Manuscript under review by NPJ Syst Biol Appl*, 2023.

[301] C. Li and X.-L. Hua. Towards positive unlabeled learning for parallel data mining: a random forest framework. *Int Conf Adv Comput Appl*, pages 573–587, 2014.

[302] J. Bekker and J. Davis. Learning from positive and unlabeled data: a survey. *Mach Learn*, 109:719–760, 2020.

[303] F. Li, S. Dong, A. Leier, M. Han, X. Guo, J. Xu, X. Wang, S. Pan, C. Jia, Y. Zhang, G. I. Webb, L. J. M. Coin, C. Li, and J. Song. Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Brief Bioinform*, 23(1), 2022.

[304] R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, and B. Peters. The immune epitope database (iedb) 3.0. *Nucleic Acids Res*, 43(Database issue):D405–12, 2015.

[305] Z. Cheng, S. Zhou, and J. Guan. Computationally predicting protein-rna interactions using only positive and unlabeled examples. *J Bioinform Comput Biol*, 13(3):1541005, 2015.

[306] K. V. Chuang and M. J. Keiser. Adversarial controls for scientific machine learning. *ACS Chem Biol*, 13(10):2819–2821, 2018.

[307] L. Breiman. Random forests. *Mach Learn*, 45(1):5–32, 2001.

[308] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*, 20:53–65, 1987.

[309] Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: a big data - ai integration perspective. *IEEE Trans Knowl Data Eng*, 33(4):1328–1347, 2019.

[310] A. Bjorkman, C. S. Benn, P. Aaby, and A. Schapira. Rts,s/as01 malaria vaccine-proven safe and effective? *Lancet Infect Dis*, 23(8):e318–e322, 2023.

[311] M. S. Datoo, H. M. Natama, A. Some, D. Bellamy, O. Traore, T. Rouamba, M. C. Tahita, N. F. A. Ido, P. Yameogo, D. Valia, A. Millogo, F. Ouedraogo, R. Soma, S. Sawadogo, F. Sorgho, K. Derra, E. Rouamba, F. Ramos-Lopez, M. Cairns, S. Provstgaard-Morys, J. Aboagye, A. Lawrie, R. Roberts, I. Valea, H. Sorgho, N. Williams, G. Glenn, L. Fries, J. Reimer, K. J. Ewer, U. Shaligram, A. V. S. Hill, and H. Tinto. Efficacy and immunogenicity of r21/matrix-m vaccine against clinical malaria after 2 years' follow-up in children in burkina faso: a phase 1/2b randomised controlled trial. *Lancet Infect Dis*, 22(12):1728–1736, 2022.

[312] G. T. S. da Veiga, M. R. Moriggi, J. F. Vettorazzi, M. Muller-Santos, and L. Albrecht. Plasmodium vivax vaccine: what is the best way to go? *Front Immunol*, 13:910236, 2022.

[313] M. Bermudez, D. A. Moreno-Perez, G. Arevalo-Pinzon, H. Curtidor, and M. A. Patarroyo. Plasmodium vivax in vitro continuous culture: the spoke in the wheel. *Malar J*, 17(1):301, 2018.

[314] C. Aguttu, B. A. Okech, A. Mukisa, and G. W. Lubega. Screening and characterization of hypothetical proteins of plasmodium falciparum as novel vaccine candidates in the fight against malaria using reverse vaccinology. *J Genet Eng Biotechnol*, 19(1):103, 2021.

[315] S. J. Goodswen, P. J. Kennedy, and J. T. Ellis. A novel strategy for classifying the output from an in silico vaccine discovery pipeline for eukaryotic pathogens using machine learning algorithms. *BMC Bioinformatics*, 14:315, 2013.

[316] R. N. Rodrigues-da Silva, J. H. Martins da Silva, B. Singh, J. Jiang, E. V. Meyer, F. Santos, D. M. Banic, A. Moreno, M. R. Galinski, J. Oliveira-Ferreira, and C. Lima-Junior Jda. In silico identification and validation of a linear and naturally immunogenic b-cell epitope of the plasmodium vivax malaria vaccine candidate merozoite surface protein-9. *PLoS One*, 11(1):e0146951, 2016.

[317] J. B. Hostetler, S. Sharma, S. J. Bartholdson, G. J. Wright, R. M. Fairhurst, and J. C. Rayner. A library of plasmodium vivax recombinant merozoite proteins reveals new vaccine candidates and protein-protein interactions. *PLoS Negl Trop Dis*, 9(12):e0004264, 2015.

[318] S. V. Siegel, L. Chappell, J. B. Hostetler, C. Amaratunga, S. Suon, U. Bohme, M. Berriman, R. M. Fairhurst, and J. C. Rayner. Analysis of plasmodium vivax schizont transcriptomes from field isolates reveals heterogeneity of expression of genes involved in host-parasite interactions. *Sci Rep*, 10(1):16667, 2020.

[319] Prasun Kundu, Deboki Naskar, Shannon McKie, Sheena Dass, Usheer Kanjee, Viola Introini, Marcelo U. Ferreira, Manoj Duraisingh, Janet Deane, and Julian C. Rayner. The structure of a plasmodium vivax tryptophan rich antigen suggests a lipid binding function for a pan-plasmodium multi-gene family. *bioRxiv*, 2022.

[320] S. J. Goodswen, P. J. Kennedy, and J. T. Ellis. A guide to current methodology and usage of reverse vaccinology towards in silico vaccine discovery. *FEMS Microbiol Rev*, 47(2), 2023.

[321] José Juan Almagro Armenteros, Konstantinos D Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and

Henrik Nielsen. Signalp 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*, 37(4):420, 2019.

[322] W. L. Hamilton, A. Claessens, T. D. Otto, M. Kekre, R. M. Fairhurst, J. C. Rayner, and D. Kwiatkowski. Extreme mutation bias and high at content in plasmodium falciparum. *Nucleic Acids Res*, 45(4):1889–1901, 2017.

[323] Rino Rappuoli. Reverse vaccinology. *Curr Opin Microbiol*, 3(5):445–450, 2000.

[324] K. Hayashida, K. Kajino, H. Simukoko, M. Simuunza, J. Ndebe, A. Chota, B. Namangala, and C. Sugimoto. Direct detection of falciparum and non-falciparum malaria dna from a drop of blood with high sensitivity by the dried-lamp system. *Parasit Vectors*, 10(1):26, 2017.

[325] T. G. Woldearegai, A. Lalremruata, T. T. Nguyen, M. Gmeiner, L. Veletzky, G. B. Tazemda-Kuitsouc, P. B. Matsiegui, B. Mordmuller, and J. Held. Characterization of plasmodium infections among inhabitants of rural areas in gabon. *Sci Rep*, 9(1):9784, 2019.

[326] S. M. Taylor, J. P. Messina, C. C. Hand, J. J. Juliano, J. Muwonga, A. K. Tshefu, B. Atua, M. Emch, and S. R. Meshnick. Molecular malaria epidemiology: mapping and burden estimates for the democratic republic of the congo, 2007. *PLoS One*, 6(1):e16420, 2011.

[327] L. Sitali, J. M. Miller, M. C. Mwenda, D. J. Bridges, M. B. Hawela, B. Hamainza, E. Chizema-Kawesha, T. P. Eisele, J. Chipeta, and B. Lindtjorn. Distribution of plasmodium species and assessment of performance of diagnostic tools used during a malaria survey in southern and western provinces of zambia. *Malar J*, 18(1):130, 2019.

[328] N. J. White. Plasmodium knowlesi: the fifth human malaria parasite. *Clin Infect Dis*, 46(2):172–3, 2008.

[329] A. Z. Chin, M. C. M. Maluda, J. Jelip, M. S. B. Jeffree, R. Culleton, and K. Ahmed. Malaria elimination in malaysia and the rising threat of plasmodium knowlesi. *J Physiol Anthropol*, 39(1):36, 2020.

[330] D. J. Cooper, G. S. Rajahram, T. William, J. Jelip, R. Mohammad, J. Benedict, D. A. Alaza, E. Malacova, T. W. Yeo, M. J. Grigg, N. M. Anstey, and B. E. Barber. Plasmodium knowlesi malaria in sabah, malaysia, 2015-2017: ongoing increase in incidence despite near-elimination of the human-only plasmodium species. *Clin Infect Dis*, 70(3):361–367, 2020.

[331] T. Pongvongsa, R. Culleton, H. Ha, L. Thanh, P. Phongmany, R. P. Marchand, S. Kawai, K. Moji, S. Nakazawa, and Y. Maeno. Human infection with plasmodium knowlesi on the laos-vietnam border. *Trop Med Health*, 46:33, 2018.

[332] C. Aurrecoechea, J. Brestelli, B. P. Brunk, J. Dommer, S. Fischer, B. Gajria, X. Gao, A. Gingle, G. Grant, O. S. Harb, M. Heiges, F. Innamorato, J. Iodice, J. C. Kissinger, E. Kraemer, W. Li, J. A. Miller, V. Nayak, C. Pennington, D. F. Pinney, D. S. Roos, C. Ross, Jr. Stoeckert, C. J., C. Treatman, and H. Wang. Plasmodb: a functional genomic database for malaria parasites. *Nucleic Acids Res*, 37(Database issue):D539–43, 2009.

[333] P. Oyarzun, J. J. Ellis, M. Boden, and B. Kobe. Predivac: Cd4+ t-cell epitope prediction for vaccine design that covers 95 *BMC Bioinformatics*, 14:52, 2013.

[334] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili. Bepipred-2.0: improving sequence-based b-cell epitope prediction using conformational epitopes. *Nucleic Acids Res*, 45(W1):W24–W29, 2017.

[335] J. E. Larsen, O. Lund, and M. Nielsen. Improved method for predicting linear b-cell epitopes. *Immunome Res*, 2:2, 2006.

[336] S. Saha and G. P. Raghava. Prediction of continuous b-cell epitopes in an antigen using recurrent neural network. *Proteins*, 65(1):40–8, 2006.

[337] M. Bhasin and G. P. Raghava. Prediction of ctl epitopes using qm, svm and ann techniques. *Vaccine*, 22(23-24):3195–204, 2004.

[338] G. Nagpal, S. S. Usmani, S. K. Dhanda, H. Kaur, S. Singh, M. Sharma, and G. P. Raghava. Computer-aided designing of immunosuppressive peptides based on il-10 inducing potential. *Sci Rep*, 7:42851, 2017.

[339] S. K. Dhanda, P. Vir, and G. P. Raghava. Designing of interferon-gamma inducing mhc class-ii binders. *Biol Direct*, 8:30, 2013.

[340] M. Bhasin and G. P. Raghava. Analysis and prediction of affinity of tap binding peptides using cascade svm. *Protein Sci*, 13(3):596–607, 2004.

[341] M. Nielsen, C. Lundegaard, P. Worning, S. L. Lauemoller, K. Lamberth, S. Buus, S. Brunak, and O. Lund. Reliable prediction of t-cell epitopes using neural networks with novel sequence representations. *Protein Sci*, 12(5):1007–17, 2003.

[342] H. H. Bui, J. Sidney, B. Peters, M. Sathiamurthy, A. Sinichi, K. A. Purton, B. R. Mothe, F. V. Chisari, D. I. Watkins, and A. Sette. Automated generation and evaluation of specific mhc binding predictive tools: Arb matrix applications. *Immunogenetics*, 57(5):304–14, 2005.

[343] A. S. Kolaskar and P. C. Tongaonkar. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett*, 276(1-2):172–4, 1990.

[344] J. J. Calis, M. Maybeno, J. A. Greenbaum, D. Weiskopf, A. D. De Silva, A. Sette, C. Kesmir, and B. Peters. Properties of mhc class i presented peptides that enhance immunogenicity. *PLoS Comput Biol*, 9(10):e1003266, 2013.

[345] C. S. Yu, Y. C. Chen, C. H. Lu, and J. K. Hwang. Prediction of protein subcellular localization. *Proteins*, 64(3):643–51, 2006.

[346] F. A. Ansari, N. Kumar, M. Bala Subramanyam, M. Gnanamani, and S. Ramachandran. Maap: malarial adhesins and adhesin-like proteins predictor. *Proteins*, 70(3):659–66, 2008.

[347] D. Osorio and P. Rondón-Villarrea. Peptides: a package for data mining of antimicrobial peptides. *R Journal*, 7(1), 2015.

[348] N. Xiao, D. S. Cao, M. F. Zhu, and Q. S. Xu. protr/protrweb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, 31(11):1857–9, 2015.

[349] J. M. Parker, D. Guo, and R. S. Hodges. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites. *Biochemistry*, 25(19):5425–32, 1986.

[350] A. Pierleoni, P. L. Martelli, and R. Casadio. Predgpi: a gpi-anchor predictor. *BMC Bioinformatics*, 9:392, 2008.

[351] J. J. Almagro Armenteros, K. D. Tsirigos, C. K. Sonderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne, and H. Nielsen. Signalp 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*, 37(4):420–423, 2019.

[352] M. Hebditch and J. Warwicker. Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. *PeerJ*, 7:e8199, 2019.

[353] J. S. Chauhan, A. Rao, and G. P. Raghava. In silico platform for prediction of n-, o- and c-glycosites in eukaryotic protein sequences. *PLoS One*, 8(6):e67008, 2013.

[354] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. Blast+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009.

[355] E. L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6:175–82, 1998.

[356] J. C. Wootton and S. Federhen. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol*, 266:554–71, 1996.

[357] P. Y. Chou and G. D. Fasman. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol*, 47:45–148, 1978.

[358] E. A. Emini, J. V. Hughes, D. S. Perlow, and J. Boger. Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol*, 55(3):836–9, 1985.

[359] P. Karplus and G. Schulz. Prediction of chain flexibility in proteins. *Naturwissenschaften*, 72:212–213, 1985.

[360] S Mangiafico. rcompanion: functions to support extension education program evaluation. *Rutgers Cooperative Extension*, 2023.

[361] L. McInnes, J. Healy, and J. Melville. Umap: uniform manifold approximation and projection for dimension reduction. *J Open Source Softw*, 2018.

[362] J.H Ward. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*, 1963.

[363] F. Murtagh and P. Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J Classif*, 31:274–295, 2014.

[364] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B*, 57(1):289–300, 1995.

[365] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Technical Report, Stanford*, 2000.

[366] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of data clusters via the gap statistic. *J R Stat Soc Series B*, 63:411–423, 2001.

[367] L. Li, Jr. Stoeckert, C. J., and D. S. Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–89, 2003.

[368] F. Chen, A. J. Mackey, Jr. Stoeckert, C. J., and D. S. Roos. Orthomcl-db: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*, 34(Database issue):D363–8, 2006.

[369] D. V. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramirez, A. Warwick Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, W. Dampier, C. Dessimoz, P. Flick, and H. Tang. Goatools: a python library for gene ontology analyses. *Sci Rep*, 8(1):10872, 2018.

[370] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, 2000.

[371] Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. *Nucleic Acids Res*, 49(D1):D325–D334, 2021.