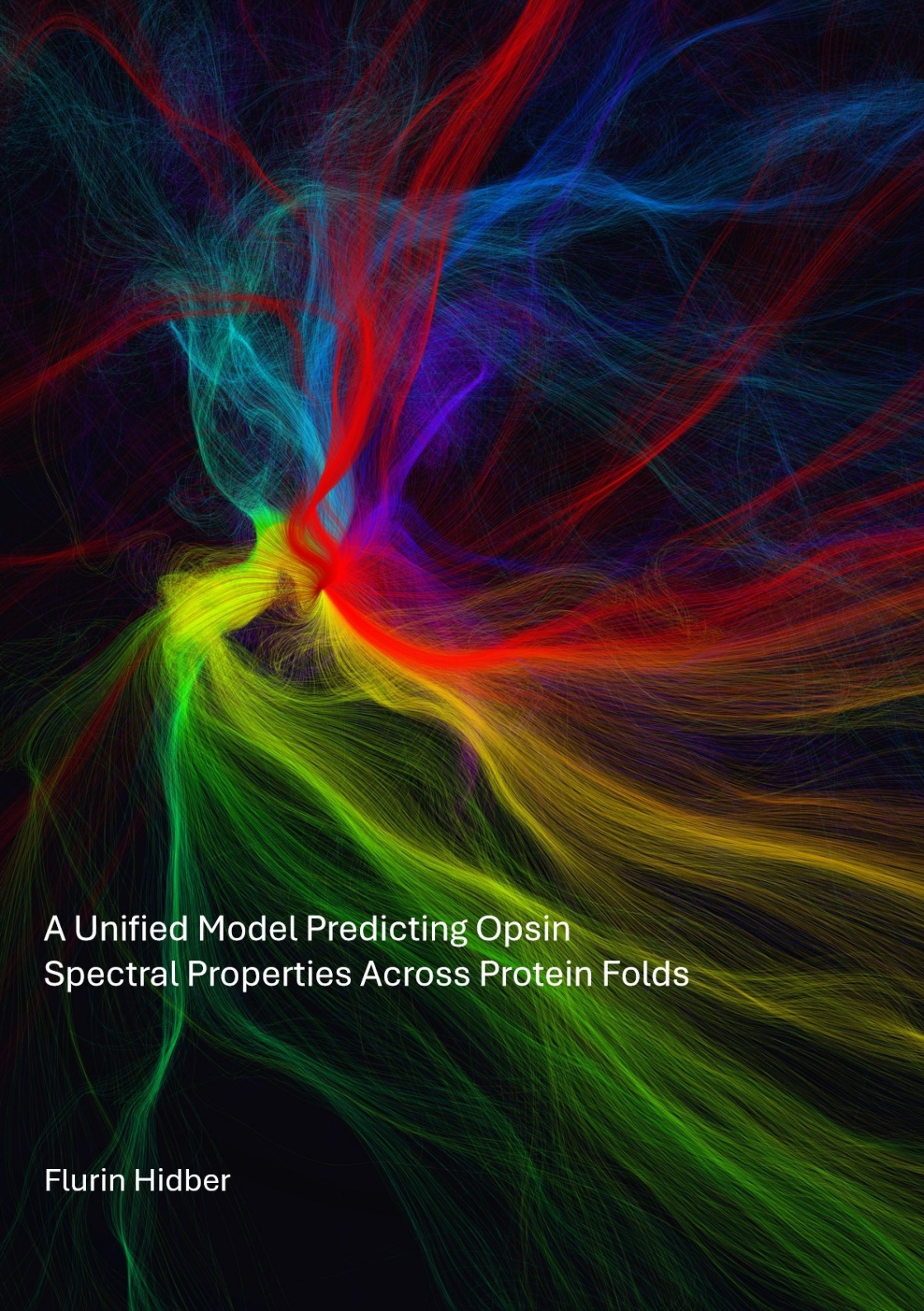


A Unified Model Predicting Opsin Spectral Properties Across Protein Folds

Flurin Hidber

DISS. ETH NO. 31926

A UNIFIED MODEL PREDICTING OPSIN SPECTRAL PROPERTIES ACROSS PROTEIN FOLDS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES
(Dr. sc. ETH Zurich)

presented by

FLURIN STEFAN HIDBER

born on 10.08.1994

accepted on the recommendation of

Prof. Dr. Volodymyr Korkhov, examiner
Dr. Xavier Deupi, co-examiner
Prof. Dr. Peter Matthias, co-examiner
Prof. Dr. Hideaki Kato, co-examiner

2026

I dedicate this thesis to my mom and dad.

ABSTRACT

Type I (microbial) and type II (animal) opsins are two independent solutions to the same problem in nature: photoreception. Both bind retinal in structurally analogous pockets whose residue composition determines spectral properties through the same physics. LAMBDA (Light Absorption Modeling through Binding Domain Analysis), the central contribution of this thesis, exploits this convergence: it is a graph neural network that predicts absorption maxima from binding pocket composition, trained across both opsin families and the engineered lipocalin hCRBPII. By representing binding pockets as graphs aligned on retinal, LAMBDA treats spectral tuning as a property of the chromophore environment rather than the surrounding fold. Trained on 2,120 sequences, it achieves 5.18 nm mean absolute error on type II opsins and 6.86 nm on type I; applied to 47,700 sequences, it produces the Opsin Atlas, a dataset of predicted spectral properties spanning the known diversity of both families. Cross-family learning requires a shared structural vocabulary. Type II opsins have Ballesteros–Weinstein numbering; type I opsins lacked an equivalent. MOGRN (Microbial Opsin Generic Residue Numbering) fills this gap—a structure-based numbering system anchored to the conserved retinal-binding site, validated against 129 structures spanning the functional diversity of microbial rhodopsins. LAMBDA uses MOGRN positions for type I and Ballesteros–Weinstein positions for type II to construct binding pocket graphs from sequence alone. ProtOS, a Python framework that integrates protein data management with access to machine learning models, provides the infrastructure that makes this feasible at scale: it handles sequences, structures, embeddings, and annotations through a consistent processor architecture, routing 47,700 sequences through the full analysis pipeline without manual intervention. ProtOS-MCP extends this infrastructure through a natural language interface via the Model Context Protocol, making spectral prediction, annotation, and protein engineering workflows accessible to structural biologists through conversation rather than code. The rhodozyme workflow demonstrates ProtOS in application: ProtOS integrates structure prediction, backbone design, and sequence design models into a single pipeline to computationally design a light-gated enzyme—a rhodopsin scaffold carrying a transplanted catalytic triad whose activation wavelength is set by the retinal binding pocket.

ZUSAMMENFASSUNG

Opsine vom Typ I (mikrobiell) und Typ II (tierisch) sind zwei unabhängige Lösungen für dasselbe Problem in der Natur: Photorezeption. Beide binden Retinal in strukturell analogen Bindetaschen, deren Aminosäurezusammensetzung die spektralen Eigenschaften durch dieselbe Physik bestimmt. LAMBDA (Light Absorption Modeling through Binding Domain Analysis), der zentrale Beitrag dieser Arbeit, nutzt diese Konvergenz: Es ist ein Graph-Neuronales-Netz, das Absorptionsmaxima aus der Zusammensetzung der Bindetasche vorhersagt, trainiert über beide Opsin-Familien und das synthetische Lipocalin hCRBPII. Durch die Darstellung von Bindetaschen als Graphen, die auf Retinal ausgerichtet sind, behandelt LAMBDA die spektrale Abstimmung als Eigenschaft der Chromophor-Umgebung und nicht der umgebenden Faltung. Trainiert auf 2120 Sequenzen erreicht es einen mittleren absoluten Fehler von 5.18 nm bei Typ-II-Opsinen und 6.86 nm bei Typ I; angewandt auf 47 700 Sequenzen erzeugt es den Opsin-Atlas, einen Datensatz vorhergesagter spektraler Eigenschaften über die bekannte Vielfalt beider Familien. Familienübergreifendes Lernen erfordert ein gemeinsames strukturelles Vokabular. Typ-II-Opsine verfügen über die Ballesteros–Weinstein-Nummerierung; Typ-I-Opsinen fehlte ein Äquivalent. MOGRN (Microbial Opsin Generic Residue Numbering) schliesst diese Lücke — ein strukturbasiertes Nummerierungssystem, verankert an der konservierten Retinal-Bindestelle, validiert gegen 129 Strukturen, die die funktionelle Vielfalt mikrobieller Rhodopsine abdecken. LAMBDA verwendet MOGRN-Positionen für Typ I und Ballesteros–Weinstein-Positionen für Typ II, um Bindetaschen-Graphen allein aus der Sequenz zu konstruieren. ProtOS, ein Python-Framework, das Proteindatenmanagement mit dem Zugang zu maschinellen Lernmodellen verbindet, stellt die Infrastruktur bereit, die dies im grossen Massstab ermöglicht: Es verarbeitet Sequenzen, Strukturen, Embeddings und Annotationen über eine konsistente Prozessorarchitektur und leitet 47 700 Sequenzen ohne manuelles Eingreifen durch die gesamte Analysepipeline. ProtOS-MCP erweitert diese Infrastruktur durch eine natürlichsprachliche Schnittstelle über das Model Context Protocol und macht spektrale Vorhersage, Annotation und Protein-Engineering-Workflows für Strukturbiologen über Konversation statt Code zugänglich. Der Rhodozym-Workflow demonstriert ProtOS in der Anwendung: ProtOS integriert Strukturvorhersage, Rückgrat-Design und Sequenz-Design-Modelle in eine einzige Pipeline zur computergestützten Entwicklung eines lichtgesteuerten Enzyms — eines Rhodopsin-Gerüsts mit einer transplantierten katalytischen Triade, deren Aktivierungswellenlänge durch die Retinal-Bindetasche bestimmt wird.

RÉSUMÉ

Les opsines de type I (microbiennes) et de type II (animales) sont deux solutions indépendantes au même problème dans la nature : la photoréception. Toutes deux lient le rétinal dans des poches de liaison structurellement analogues dont la composition en résidus détermine les propriétés spectrales par la même physique. LAMBDA (Light Absorption Modeling through Binding Domain Analysis), la contribution centrale de cette thèse, exploite cette convergence : c'est un réseau de neurones sur graphes qui prédit les maxima d'absorption à partir de la composition de la poche de liaison, entraîné sur les deux familles d'opsines et le lipocaline synthétique hCRBPII. En représentant les poches de liaison comme des graphes alignés sur le rétinal, LAMBDA traite l'accord spectral comme une propriété de l'environnement du chromophore plutôt que du repliement environnant. Entraîné sur 2120 séquences, il atteint une erreur absolue moyenne de 5.18 nm sur les opsines de type II et 6.86 nm sur le type I ; appliqué à 47 700 séquences, il produit l'Opsin Atlas, un jeu de données de propriétés spectrales prédites couvrant la diversité connue des deux familles. L'apprentissage inter-familles nécessite un vocabulaire structurel commun. Les opsines de type II disposent de la numérotation Ballesteros–Weinstein ; les opsines de type I en étaient dépourvues. MOGRN (Microbial Opsin Generic Residue Numbering) comble cette lacune — un système de numérotation structurel ancré au site conservé de liaison du rétinal, validé sur 129 structures couvrant la diversité fonctionnelle des rhodopsines microbiennes. LAMBDA utilise les positions MOGRN pour le type I et les positions Ballesteros–Weinstein pour le type II afin de construire des graphes de poche de liaison à partir de la seule séquence. ProtOS, un framework Python qui intègre la gestion des données protéiques avec l'accès à des modèles d'apprentissage automatique, fournit l'infrastructure rendant cela réalisable à grande échelle : il gère séquences, structures, embeddings et annotations via une architecture de processeurs cohérente, acheminant 47 700 séquences à travers le pipeline d'analyse complet sans intervention manuelle. ProtOS-MCP étend cette infrastructure par une interface en langage naturel via le Model Context Protocol, rendant la prédiction spectrale, l'annotation et les workflows d'ingénierie protéique accessibles aux biologistes structuraux par la conversation plutôt que par le code. Le workflow rhodozyme illustre ProtOS en application : ProtOS intègre des modèles de prédiction de structure, de conception de squelette et de conception de séquence en un seul pipeline pour la conception computationnelle d'une enzyme photo-activée — un échafaudage rhodopsine portant une triade catalytique transplantée dont la longueur d'onde d'activation est déterminée par la poche de liaison du rétinal.

RIASSUNTO

Le opsine di tipo I (microbiche) e di tipo II (animali) sono due soluzioni indipendenti allo stesso problema in natura: la fotoricezione. Entrambe legano il retinale in tasche di legame strutturalmente analoghe la cui composizione in residui determina le proprietà spettrali attraverso la stessa fisica. LAMBDA (Light Absorption Modeling through Binding Domain Analysis), il contributo centrale di questa tesi, sfrutta questa convergenza: è una rete neurale su grafi che predice i massimi di assorbimento dalla composizione della tasca di legame, addestrata su entrambe le famiglie di opsine e sulla lipocalina ingegnerizzata hCRBPII. Rappresentando le tasche di legame come grafi allineati sul retinale, LAMBDA tratta l'accordatura spettrale come una proprietà dell'ambiente del cromoforo piuttosto che del ripiegamento circostante. Addestrato su 2120 sequenze, raggiunge un errore assoluto medio di 5.18 nm sulle opsine di tipo II e 6.86 nm sul tipo I; applicato a 47 700 sequenze, produce l'Opsin Atlas, un dataset di proprietà spettrali predette che copre la diversità nota di entrambe le famiglie. L'apprendimento inter-famiglia richiede un vocabolario strutturale condiviso. Le opsine di tipo II dispongono della numerazione Ballesteros–Weinstein; le opsine di tipo I ne erano prive. MOGRN (Microbial Opsin Generic Residue Numbering) colma questa lacuna — un sistema di numerazione basato sulla struttura, ancorato al sito conservato di legame del retinale, validato su 129 strutture che coprono la diversità funzionale delle rodopsine microbiche. LAMBDA utilizza le posizioni MOGRN per il tipo I e le posizioni Ballesteros–Weinstein per il tipo II per costruire grafi della tasca di legame dalla sola sequenza. ProtOS, un framework Python che integra la gestione dei dati proteici con l'accesso a modelli di apprendimento automatico, fornisce l'infrastruttura che rende tutto ciò realizzabile su larga scala: gestisce sequenze, strutture, embeddings e annotazioni attraverso un'architettura di processori coerente, instradando 47 700 sequenze attraverso l'intera pipeline di analisi senza intervento manuale. ProtOS-MCP estende questa infrastruttura con un'interfaccia in linguaggio naturale tramite il Model Context Protocol, rendendo la predizione spettrale, l'annotazione e i workflow di ingegneria proteica accessibili ai biologi strutturali attraverso la conversazione anziché il codice. Il workflow rhodozyme dimostra ProtOS in applicazione: ProtOS integra modelli di predizione strutturale, progettazione dello scheletro e progettazione della sequenza in un'unica pipeline per la progettazione computazionale di un enzima fotoattivato — uno scaffold rodopsinico con una triade catalitica trapiantata la cui lunghezza d'onda di attivazione è determinata dalla tasca di legame del retinale.

CONTENTS

1

INTRODUCTION

1.1 A BRIEF HISTORY OF PROTEIN STRUCTURES

For most of history, proteins were invisible. Biochemists knew they existed, knew they catalyzed reactions and transmitted signals, but could not see them. X-ray crystallography changed this. In 1958, John Kendrew solved the structure of myoglobin [1]—for the first time, researchers could see a folded protein. Myoglobin is a soluble protein floating freely in the cytoplasm, a chain of amino acids wrapped into a compact globule with a heme group nestled in a hydrophobic pocket where oxygen binds. The structure revealed why myoglobin worked.

Membrane proteins proved far harder. They have hydrophobic surfaces that contact lipids and hydrophilic surfaces exposed to water on either side of the membrane. Removed from the membrane, they aggregate, and aggregated proteins cannot form the ordered crystals that X-ray crystallography requires. Membrane protein structures remained out of reach for decades, yet these proteins handle many interactions between a cell and its environment.

Despite these challenges, the first view of a membrane protein came in 1975. Richard Henderson and Nigel Unwin used electron microscopy to image bacteriorhodopsin, a light-driven proton pump from *Halobacterium salinarum* [2]. The resolution was 7 Å, too coarse to resolve individual atoms, but the map showed seven rod-like densities spanning the membrane—bacteriorhodopsin adopted a seven-transmembrane-helix fold. This architecture would prove significant: the same seven-helix fold defines a major class of animal signaling proteins.

In 1993, Gebhard F.X. Schertler, Claudio Villa, and Richard Henderson published a 9 Å projection map of bovine rhodopsin, the light-sensitive protein of vertebrate eyes [3]. The map confirmed the motif: rhodopsin, like bacteriorhodopsin, adopted a seven-transmembrane-helix fold. Bovine rhodopsin belongs to the GPCRs—G protein coupled receptors—a large family of signaling proteins named for the G proteins they activate inside the cell. A single photon triggers a structural change in rhodopsin that starts a signaling cascade, amplifying a tiny signal into a large cellular response. Rhodopsin was the first GPCR amenable to structural analysis.

Then in 2000, Krzysztof Palczewski solved bovine rhodopsin at 2.8 Å, the first GPCR at atomic resolution [4], confirming the seven-helix architecture. The much higher resolution revealed how the retinal, the light-absorbing chromophore, sits in the transmembrane core.

Later in 2007, Brian Kobilka's group solved the β_2-adrenergic receptor [5], showing that other GPCRs could be crystallized. Kobilka shared the 2012 Nobel Prize with Robert Lefkowitz for their work.

Crystallography requires proteins to form ordered crystals, a barrier that excludes many membrane proteins. Cryo-electron microscopy [6], another method to elucidate protein structures, led to many more structures being solved. The technique flash-freezes proteins in thin ice, preserving their conformations, but for decades cryo-EM produced only blurry images, useful for overall shape but not atomic detail. The breakthrough came around 2013 with direct electron detectors, which dramatically improved signal quality, and new algorithms that could

align and average millions of particle images. Proteins that resisted crystallization—large complexes, flexible assemblies, membrane proteins in near-native states—could now be resolved. Henderson shared the 2017 Nobel Prize in Chemistry with Jacques Dubochet and Joachim Frank for this work.

Together these two methods shaped structural biology. The Protein Data Bank [7], the largest experimental database for protein structures, grew from roughly 1,000 structures in 1993 to over 200,000 by 2024.

Structure determination investigates one protein at a time. Sequencing opened a window onto the full diversity of the protein world, at a fraction of the effort. UniProt [8] now contains over 200 million protein sequences. For some protein families, this revealed unexpected abundance. Microbial rhodopsins are a prime example. First identified in 2000 from Pacific Ocean samples [9], they are now recognized as among the most abundant proteins on Earth [10], with thousands of sequences spanning bacteria, archaea, algae, and fungi.

Unsurprisingly, sequence diversity far exceeds structural coverage using experimental methods. Computational models had long attempted to predict protein structures from sequence, but accuracy remained limited. In 2020, DeepMind's AlphaFold2 marked an inflection point, achieving a leap in accuracy that made structure prediction broadly useful in structural biology [11]. Within two years, DeepMind predicted structures with AlphaFold for over 200 million proteins [12]. Many models have followed. Among them, Boltz [13], an open-source system, predicts proteins bound to their ligands—small molecules, cofactors, and interaction partners—enabling computational analysis of binding pockets without requiring experimental structures of each complex.

A similarly important class of deep learning methods are protein language models. Trained on millions of protein sequences, models like ESM-2 [14, 15] and Ankh [16, 17] produce embeddings—dense numerical vectors for each amino acid in a sequence that capture evolutionary and structural context. Averaging these per-residue vectors yields a single fingerprint for the entire protein. Because proteins with diverged sequences but similar functions produce similar embeddings, these fingerprints serve as features for downstream models that predict protein properties.

Structure prediction and protein language models have transformed what is computationally accessible, but predicted structures lack context—they come without experimental conditions, functional annotations, or consistent coordinate systems, and annotating them at scale requires automation that does not yet exist. The bottleneck has shifted from generating data to understanding it. For researchers studying membrane protein families like GPCRs and microbial rhodopsins, this shift has practical consequences. These families now have hundreds of structures and are targets for drug discovery [18] and biotechnology, yet systematic studies across their members remain difficult.

1.2 OPSINS

This thesis studies opsins. They are a family of seven-transmembrane proteins that become light-sensitive upon binding retinal as their chromophore. They exist across all domains of life, where they perform functions ranging from ion transport to signal transduction.

Opsins bind retinal, a vitamin A derivative, through a covalent Schiff base linkage to a conserved lysine residue—a bond formed between retinal's aldehyde group and the lysine's

amino group. Strictly, *opsin* refers to the apoprotein alone, while *rhodopsin* refers to the protein with retinal bound; in practice, the terms are used interchangeably. Opsins detect light through retinal: a photon is absorbed by retinal's conjugated electron system, causing it to isomerize—change shape—which triggers a conformational change in the surrounding protein. The residues lining the binding pocket determine which wavelengths retinal absorbs, so different opsins, with different pocket residues, respond to different colors of light. All opsins share this basic mechanism—retinal isomerization coupled to conformational change—yet they fall into two evolutionarily distinct families that share no detectable sequence homology [19]. Both have seven transmembrane helices surrounding a retinal binding pocket, but the folds do not superimpose: the helices are arranged differently and the binding pocket is oriented differently within the transmembrane bundle.

Type I opsins, the microbial rhodopsins, are found in archaea, bacteria, algae, and fungi. They are not GPCRs. Retinal usually binds in the all-*trans* configuration, and light triggers isomerization to 13-*cis*. The protein responds directly by pumping ions, opening a channel, or activating an enzymatic domain, with no G protein involved. The first to be characterized was bacteriorhodopsin, discovered in 1971 in *Halobacterium salinarum* [20], which pumps protons across the membrane. The family has since expanded dramatically. Channelrhodopsins, which open cation channels in response to light, enabled optogenetics—the use of light to control genetically targeted cells—and transformed neuroscience [21, 22]. Despite this functional diversity, within the transmembrane helices, Type I opsins share a highly conserved fold.

Type II opsins are GPCRs found in animals [23]. Retinal binds in the 11-*cis* configuration, and light triggers isomerization to all-*trans*. The conformational change activates a bound G protein, initiating a signaling cascade; the protein itself does not pump ions or open channels. The first to be structurally characterized was bovine rhodopsin, the photoreceptor in rod cells that enables dim-light vision [24]. Cone opsins, which mediate color vision, are closely related and exist in multiple spectral variants [25]. The spectral tuning of these opsins determines which wavelengths each type detects.

These two families represent convergent evolution [10]—two separate protein lineages faced the same problem of detecting light and arrived at the same solution: binding retinal via a protonated Schiff base (one that carries a positive charge) and using light-triggered isomerization to drive conformational change.

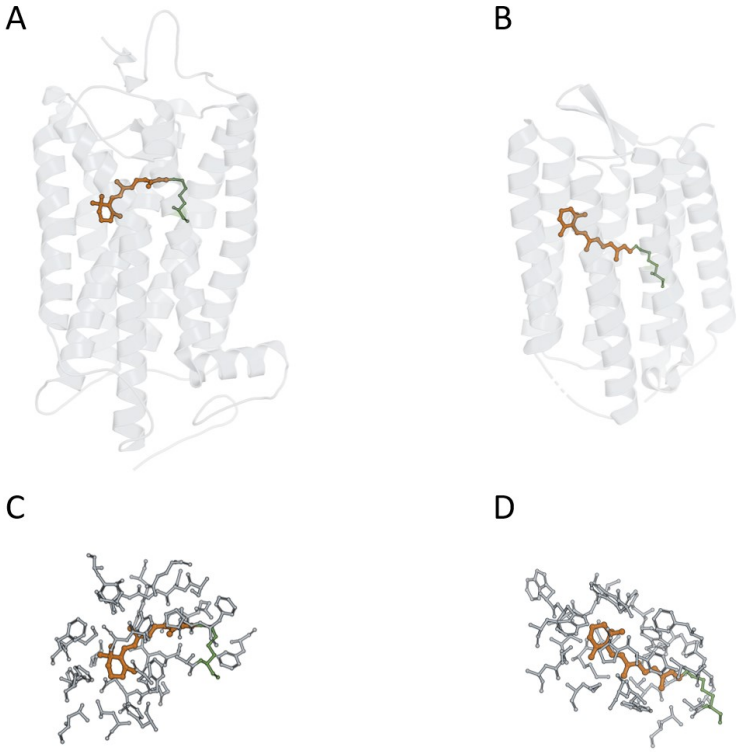


FIGURE 1.1: Structural comparison of Type I and Type II opsins. (A) Bovine rhodopsin (PDB: 1U19), an animal opsin (Type II/GPCR), shown in transparent cartoon with the retinal chromophore (rust) and Schiff base lysine K296 (green). (B) Bacteriorhodopsin (PDB: 1C3W), a microbial rhodopsin (Type I), shown in the same style. (C,D) Binding pocket close-ups for bovine rhodopsin and bacteriorhodopsin, respectively, showing all residues within 6 Å of retinal as gray sticks.

This shared solution has an important consequence. Both families use the same chromophore in geometrically similar binding pockets. In bacteriorhodopsin, retinal binds to lysine 216 (K216) on TM7 [26], while in bovine rhodopsin, retinal binds to lysine 296 (K296) on TM7 [4]. In both cases, the chromophore sits in a pocket formed by transmembrane helices, the Schiff base is protonated, and a negatively charged residue—the counterion—sits nearby to stabilize the positive charge (Figure 1.1).

The spectral properties of an opsin—the wavelengths of light it absorbs—depend on the residues surrounding retinal. The same ligand sits in structurally analogous pockets across both

families, and the same physics governs how those pockets modulate its electronic properties. This means that any retinal binding domain, regardless of which protein fold surrounds it, can in principle be studied from the same perspective: how does the local environment shape retinal's absorption? If this perspective holds, data from both opsin families—and even non-opsin retinal binders—could inform a single model of spectral tuning.

Standardized positional annotation makes this concrete. Within a family, knowing which residue in one protein corresponds to which in another enables systematic comparison. Type II opsins, as GPCRs, benefit from Ballesteros–Weinstein numbering, and researchers can compare binding pocket residues across animal opsins using standardized positions. Type I opsins lack an equivalent system. Early efforts used bacteriorhodopsin as a reference, numbering positions in other proteins by their alignment to bR. But microbial rhodopsins have diverged too far for sequence alignment to reliably identify equivalent positions across the full superfamily. A more recent approach defined positions based on conserved hydrogen bond networks, but these networks differ between functional classes—what holds for bacteriorhodopsin does not generalize to halorhodopsins, channelrhodopsins, or enzyme rhodopsins. The problem remains unsolved.

1.3 SPECTRAL TUNING

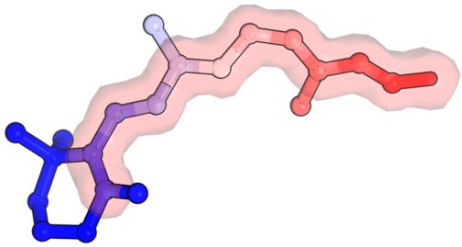


FIGURE 1.2: Charge distribution along the retinal chromophore. The retinal of bovine rhodopsin (PDB: 1U19) shown as sticks within a transparent Gaussian surface envelope, colored by a point-charge gradient from the protonated Schiff base nitrogen (red, positive) to the β-ionone ring (blue, neutral). Positive charge density decreases along the conjugated polyene chain, and binding pocket residues modulate this distribution—shifting the energy gap for light absorption and thus determining λ_{max}.

Every opsin has a characteristic absorption spectrum, and the peak of that spectrum, the wavelength at which the protein absorbs light most strongly, is called λ_{max}. This parameter determines what color of light activates the opsin. An opsin with λ_{max} at 480 nm responds best to blue light, one at 530 nm responds to green, and one at 620 nm responds to red. Across both opsin families, λ_{max} ranges from approximately 350 nm (ultraviolet) to 650 nm (far-red), a span of 300 nm achieved using the same chromophore (Figure 1.2).

Spectral properties matter for applications. Optogenetics—the use of light to control genetically targeted cells—depends on opsins with known λ_{max} values, because the wavelength of the stimulus determines which opsin responds. Red-shifted opsins ($\lambda_{max} > 580\,nm$) are particularly valuable: red light penetrates tissue more deeply than blue, enabling activation of cells in deeper brain regions or thicker tissue. Blue-shifted opsins ($\lambda_{max} < 450\,nm$) enable spectral multiplexing, allowing multiple opsins with different λ_{max} values to be activated independently in the same tissue. Engineering opsins with specific spectral properties requires understanding the molecular basis of what determines λ_{max}.

The chromophore itself provides only part of the answer. Retinal (aldehyde) in solution absorbs at approximately 380 nm [27], near ultraviolet light, yet when bound to different opsins, absorption spans from the ultraviolet to the far-red. The protein environment tunes the chromophore, and understanding how this large spectral range is achieved using a single chromophore requires examining the local environment around the retinal.

Retinal is a polyene, a chain of alternating single and double bonds. The electrons in these double bonds are delocalized across the chain, forming a π-electron system. Light absorption promotes an electron from a lower-energy state to a higher-energy state. The energy gap between these states determines the wavelength of light absorbed. A smaller gap means lower energy, which corresponds to longer wavelengths (red shift). A larger gap means higher energy, which corresponds to shorter wavelengths (blue shift).

The protonation state of the Schiff base is critical: protonated opsins absorb in the visible range (above 400 nm), while deprotonated opsins absorb in the ultraviolet (below 400 nm), as seen in UV-sensitive SWS1 cone opsins [28]. How the protein environment stabilizes the positive charge of the protonated Schiff base affects the distribution of electron density across the retinal's conjugated system, altering the energy gap and thereby determining λ_{max}.

Beyond the counterion, other binding pocket residues contribute through hydrogen bonds, local dipoles, and dispersive interactions with retinal's π-system. The cumulative effect of all these interactions determines λ_{max}.

Small changes in the binding pocket can produce large spectral shifts. Researchers interested in optogenetics exploit this to engineer opsins with novel spectral properties. The D85N mutation in bacteriorhodopsin, for example, replaces the negatively charged counterion with a neutral asparagine, producing a red shift of approximately 25 nm [29]. In bovine rhodopsin, the equivalent mutation E113Q has a different outcome: removing the counterion destabilizes the protonated Schiff base, causing it to deprotonate and shifting absorption from 500 nm to approximately 380 nm—a blue shift into the UV. The counterion thus serves a dual role: it tunes λ_{max} electrostatically and maintains the protonation state required for visible-light absorption. Other binding pocket residues contribute individual shifts of 5 nm to 30 nm.

Beyond understanding, researchers seek prediction. The ability to predict λ_{max} from sequence alone would enable rapid screening of novel sequences and guide engineering efforts. Family-specific methods exist. OPTICS [30] predicts λ_{max} for type II opsins using sequence features and phylogenetic information trained on the Visual Physiology Opsin Database (VPOD), achieving approximately 5.5 nm mean absolute error. For microbial opsins, Inoue et al. [31] developed a LASSO regression model using binding pocket residue identities, achieving 7.8 nm MAE. RhoMax [32] uses a graph neural network on type I opsin binding pocket structures to achieve 6.8 nm MAE.

These methods are family-specific by construction. OPTICS and Inoue et al. rely on multiple sequence alignments within their respective families, and RhoMax was trained exclusively on Type I data. The scientific question, however, is not family-specific. Given a retinal binding pocket—regardless of which protein fold surrounds it—how does it modulate the spectral properties of its ligand? This is the central question of the thesis: a unified model that predicts opsin spectral properties across protein folds.

1.4 THESIS CONTRIBUTIONS

LAMBDA (Light Absorption Modeling through Binding Domain Analysis) is the central contribution of this thesis. It is a spectral property prediction model for retinal-binding proteins—not restricted to type I or type II opsins individually, but trained across both families and even a non-opsin retinal binder (hCRBPII, a lipocalin fold). LAMBDA represents binding pockets as graphs, enabling a single model to learn from any retinal binding domain regardless of the surrounding protein fold. Trained on 2,120 sequences, LAMBDA achieves 5.18 nm mean absolute error on type II opsins and 6.86 nm on type I. I applied LAMBDA to 47,700 opsin sequences to create the Opsin Atlas, a dataset of predicted spectral properties spanning the known diversity of both families.

MOGRN (Microbial Opsin Generic Residue Numbering) standardizes positional annotation for microbial rhodopsins. Following the Ballesteros–Weinstein convention used for GPCRs, MOGRN introduces a structure-based numbering system anchored to the conserved retinal-binding site. Validated against 129 structures spanning the functional diversity of microbial rhodopsins, MOGRN provides a common coordinate system for analyzing binding pocket residues across the type I family. LAMBDA uses MOGRN positions for type I and Ballesteros–Weinstein positions for type II to construct binding pocket graphs from sequence alone.

ProtOS is a Python framework for protein data management. It integrates the handling of protein sequences, structures, embeddings, and annotations through a consistent processor architecture, providing the infrastructure needed to compute embeddings, annotate GRN positions, and construct pocket graphs at scale. ProtOS-MCP adds a natural language interface through the Model Context Protocol, making these capabilities accessible to structural biologists who do not write code.

2

PROTOS

Data curation is the quiet misery of computational biology—ninety percent of the work, invisible in the final result.

This chapter presents ProtOS before the scientific contributions (MOGRN and LAMBDA) because it is the methodological foundation: the framework through which all data in this thesis was collected, processed, and analyzed. Understanding its components—how sequences, structures, embeddings, and annotations are managed—gives the reader the vocabulary needed to follow the research chapters.

ProtOS was not planned as a research project. It began as scripts for parsing structures, fetching sequences, reconciling identifiers, and extracting binding pocket features. The same problems appeared in every analysis, solved each time with slightly different code. After redoing these workflows for GPCRs, animal opsins, and microbial opsins, the scripts became general, the interfaces became standardized, and integration produced a framework. Biological data is messy in ways that the word "data" does not convey—structure files designed for manual inspection, inconsistent annotations between databases, unreliable metadata. What started as disposable data curation scripts became a Python package that makes protein data management reproducible—and, through a natural language interface, accessible to researchers who do not write code.

Beyond curation, the scientific question requires integrating different data types—sequences, structures, embeddings, measured properties—that live in separate databases under separate identifiers. No single data type answers the question of what determines spectral tuning. Answering it requires combining all of them across thousands of proteins.

ProtOS addresses three aspects of this problem (Figure 2.1). First, a zero-configuration directory structure and entity registry ensure that all data for a given protein lives under one path, with persistent identity across sessions. Second, six processors handle the data modalities this work requires—sequences, structures, residue contact graphs, standardized positions, embeddings, and properties—with outputs from one becoming inputs to another through recorded relationships. Third, a model manager orchestrates external compute for structure prediction, embedding generation, and spectral prediction.

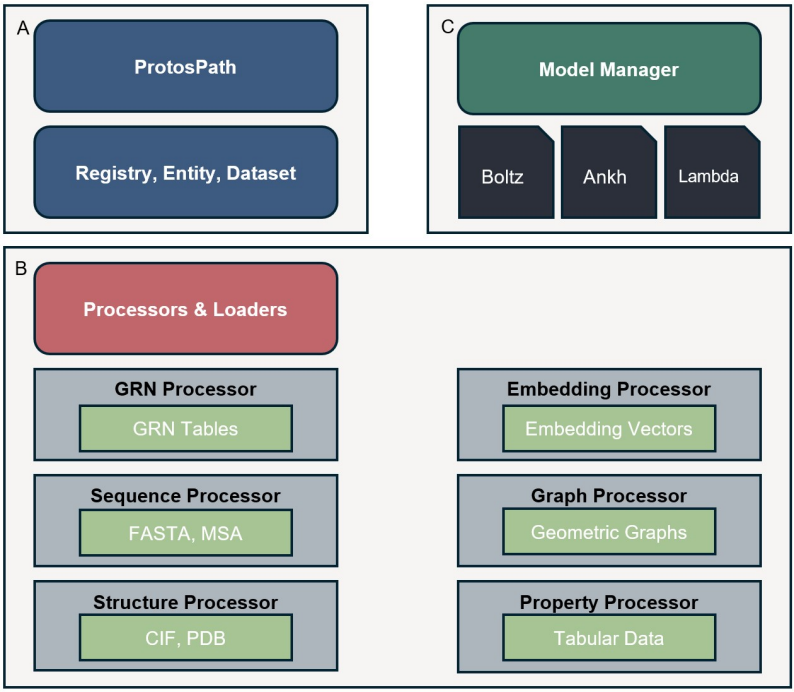


FIGURE 2.1: ProtOS architecture overview. (A) ProtosPath and the Entity/Registry/Dataset system provide persistent identity and organization. (B) Processors handle specific data modalities: GRN Processor for standardized positions, Sequence Processor for FASTA and alignments, Structure Processor for coordinates, Embedding Processor for pLM representations, Graph Processor for contact networks, and Property Processor for tabular data. (C) The Model Manager orchestrates external compute for structure prediction (Boltz), embeddings (pLM), and spectral prediction.

The sections that follow demonstrate each processor using Type II opsin examples to show how data is integrated across modalities. They are not a continuous narrative—each covers a different data modality, illustrated with a minimal code snippet. These snippets showcase ProtOS as an abstraction layer: data management happens behind the scenes, and the researcher only invokes the processors. Together, they give the reader the vocabulary needed to follow the research chapters.

2.1 SEQUENCE PROCESSOR

Sequence is the starting point of any protein analysis. At scale, this means managing thousands of sequences from UniProt, NCBI, and organism-specific repositories, each with different identifiers and curation standards.

The Sequence Processor manages retrieval, registration, and the analytical operations that follow. It routes requests to UniProt or NCBI based on identifier format, downloads sequences, and registers each in ProtOS. Registered sequences become the substrate for homology search (BLAST [33] against remote databases, MMseqs2 [34] for local high-throughput clustering), pairwise and multiple sequence alignment through BioPython [35], per-position conservation analysis, residue covariation via mutual information, and combinatorial mutant library generation at specified positions.

Sequence Processor

```
seq_proc = SequenceProcessor()
loader = SequenceLoader(processor=seq_proc)
loader.download_and_register("uniprot:P02699", name="RHO_BOVIN")

ncbi = NCBILoader(processor=seq_proc)
hits = ncbi.blast_search(
    sequence=seq_proc.load_entity("RHO_BOVIN"),
    database="swissprot", hitlist_size=5000
)
ncbi.download_batch(accessions, dataset_name="opsin_atlas")
```

Building the Type II opsin dataset required collecting sequences across all known animal opsin subfamilies. Starting from nine query sequences—one per subfamily, spanning rod opsins, cone opsins (MWS and SWS), melanopsin, neuropsin, encephalopsin, RGR, peropsin, and parapinopsin—I used iterative BLAST searches against UniProt to retrieve homologs at increasing distance thresholds. Redundancy filtering produced a dataset of 27,639 sequences across nine subfamilies.

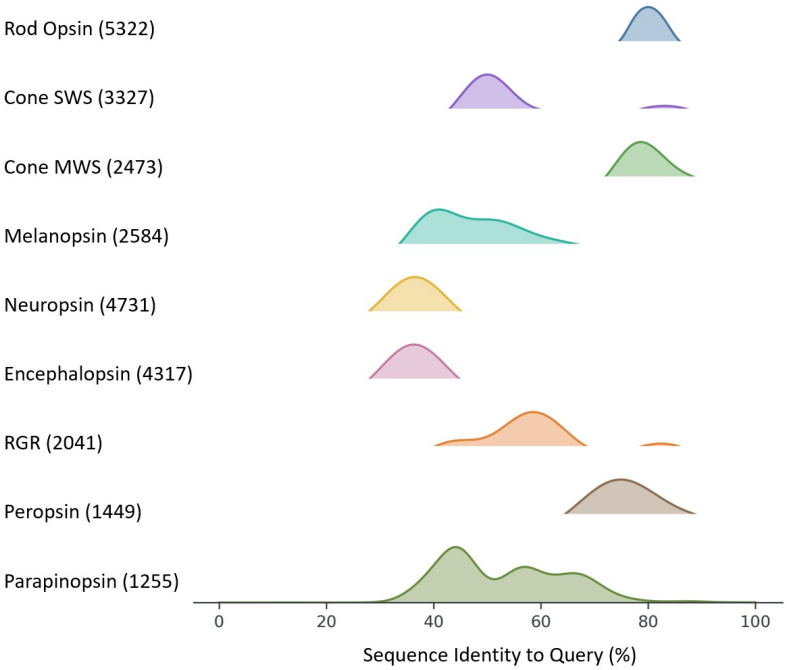


FIGURE 2.2: Per-subfamily sequence identity distributions for Type II opsins across nine subfamilies. Each ridge shows the distribution of BLAST identity (%) to the subfamily query sequence. Rod opsins form a tight peak near 70%; cone SWS opsins show broader diversity extending below 40%. Subfamily colors follow the subfamily palette used in all subsequent figures.

Figure 2.2 shows the sequence identity distributions for the nine subfamilies. Rod opsins form a tight distribution centered around 70% identity to the bovine rhodopsin query. Cone MWS opsins and cone SWS opsins show broader diversity, with identities extending below 40%. Non-visual opsins—melanopsin, neuropsin, encephalopsin—form smaller clusters. Each subfamily is internally cohesive while well-separated from others, validating the classification used throughout this thesis.

These 27,639 sequences are now registered in ProtOS. Each carries its accession, subfamily assignment, and species—and each is available to every subsequent processor. But sequence identity alone does not reveal how binding pocket residues are arranged in three-dimensional space—and it is this spatial arrangement that determines spectral properties.

2.2 STRUCTURE PROCESSOR

Binding pocket analysis requires knowing where residues sit relative to the chromophore—which residues are in contact, at what distance, and with what geometry. This information lives in coordinate files: experimental structures from the PDB [7] and predicted structures from AlphaFold DB [11, 12]. Structural data is sparse—the PDB holds roughly 200,000 experimental structures for hundreds of millions of known sequences—and predicted structures lack ligands, waters, and ions that define a binding site. The coordinate files themselves, in PDB and mmCIF formats, must be parsed into representations where spatial relationships become queryable.

The Structure Processor parses coordinate files into queryable representations where atoms, residues, and their contacts are programmatically accessible. Each structure is registered in ProtOS. Relationships to sequences arise explicitly—when a sequence is extracted from a structure, or when Boltz2 predicts a structure from a sequence, the provenance is recorded. The processor identifies ligand contacts (including water-mediated and ion-mediated contacts), filters structures by chain or residue range, applies coordinate transformations and Kabsch superpositions [36], and merges multiple structures into multi-chain complexes.

Structure Processor

```
struct_proc = StructureProcessor()
loader = StructureLoader(processor=struct_proc)
loader.download_batch(
    ["1C3W", "1U19"], dataset_name="opsins"
)

contacts = struct_proc.get_ligand_interactions(
    "1C3W", ligand_id="RET",
    chain_id="A", cutoff=4.0
)
```

Bacteriorhodopsin (PDB 1C3W [26], Type I) and bovine rhodopsin (PDB 1U19 [37], Type II) illustrate what structural comparison reveals. Extracting residues within 4 Å of retinal yields approximately 20 pocket residues in each case. Both pockets span helices 3 through 7, with the counterion (D85 in bacteriorhodopsin, E113 in bovine rhodopsin) positioned to stabilize the protonated Schiff base [38]. Kabsch alignment on retinal atoms places the two chromophores in the same frame: the pockets overlap around the Schiff base and counterion and diverge elsewhere. Global structural alignment fails because the helix topologies differ; alignment on the ligand succeeds because both proteins face the same physical constraint—binding retinal in a manner that permits isomerization and spectral tuning.

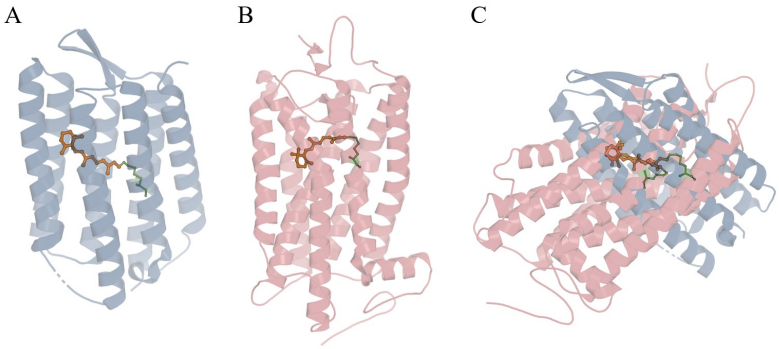


FIGURE 2.3: Retinal-aligned overlay of bacteriorhodopsin (1C3W, blue) and bovine rhodopsin (1U19, red). Retinal (orange) occupies a geometrically similar position in both folds despite independent evolutionary origins.

Both structures are now in ProtOS, each with extracted pocket residues and ligand contacts. The binding pocket residues form spatial contact networks—representing these as graphs makes the topology explicit and comparable.

2.3 GRAPH PROCESSOR

Graphs are mathematical representations of connectivity. Nodes represent entities, edges represent relationships between them. Crucially, graphs can be constructed from either sequence or structure: if you know which positions correspond to nodes, the edges encode whatever relationship you define—sequence proximity, spatial contact, or functional coupling. Residue contact graphs formalize binding pocket topology. Nodes represent residues, edges connect pairs whose Cα atoms fall within a distance threshold. The resulting network captures three-dimensional proximity independent of sequence numbering—directly comparable across proteins with different sequences, different folds, or no detectable homology. For machine learning, graph neural networks operate directly on nodes and edges, making pocket graphs a natural bridge between structural coordinates and predictive models.

The Graph Processor generates these graphs from any structure already in ProtOS, optionally restricting to residues near a specified ligand to isolate binding pockets or active sites. The resulting graph is stored and linked to its source structure.

Graph Processor

```
graph_proc = GraphProcessor(
    default_cutoff=7.0,
    default_level="residue"
)
graph_name = graph_proc.generate_graph(
    "1U19", chain="A",
    near_hetatm=("RET", 7.0)
)
graph_data = graph_proc.load_entity(graph_name)
```

Graphs can be generated at atom or residue level, individually or in batch across an entire structure dataset. Nodes can be selected by GRN position rather than residue number, producing graphs where every node carries a transferable label—comparable across proteins regardless of sequence numbering. Conversion to PyTorch Geometric [39] tensors is built in, so that a pocket graph extracted here becomes a direct input to a graph neural network.

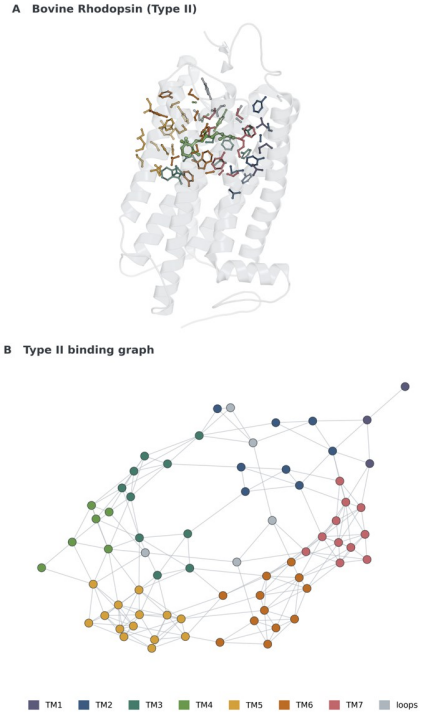


FIGURE 2.4: Binding pocket graph for bovine rhodopsin (PDB: 1U19). (A) Retinal binding pocket with residues within 9 Å of retinal. (B) The corresponding residue contact graph: nodes represent pocket residues, edges connect Cα pairs within 4 Å.

Graphs built from structures require coordinates. But if every node can be labeled with a standardized positional identifier—one that maps sequence positions to the protein fold—then graphs can be constructed from sequence alone. A binding pocket graph becomes a set of generic residue numbers and their adjacency, applicable to any sequence for which those positions are annotated. This integration between graphs and positional annotation is central to LAMBDA's design.

2.4 GRN PROCESSOR

Generic Residue Numbering (GRN) annotates sequences with positional labels that refer to the protein fold rather than the raw sequence index. For any sequence, if you know which positions carry which GRN labels, you can compare sequences not as mere alignments but as mappings

onto the shared fold. The GRN Processor is, in essence, a fold-aware positional annotation system for sequences.

For seven-transmembrane proteins, the Ballesteros–Weinstein convention [40] defines the coordinate system: each helix is numbered 1 through 7, with the most conserved residue on each helix designated X.50. Position 3.50 is the anchor on helix 3; positions 3.49 and 3.51 are its immediate neighbors. GRN position 3.28 refers to the same structural location in any GPCR, regardless of sequence length or insertion history. Databases like GPCRdb [41] maintain curated reference tables that map sequence positions to these coordinates. The GRN Processor manages these reference tables and annotates query sequences by aligning them to the references—no structure is needed for the query itself.

GRN Processor

```
grn_table = seq_proc.annotate_with_grn(
    dataset_name="opsin_atlas",
    reference_table="type II",
    protein_family="gpcr_a",
)
grn_proc = GRNProcessor()
table = grn_proc.load_table("atlas_grn")
```

Once annotated, GRN positions propagate across modalities. The Graph Processor can select nodes by GRN label rather than residue number, producing cross-family comparable graphs. The Structure Processor can annotate coordinates with positional identity. The Sequence Processor can generate mutant libraries at specific GRN positions rather than arbitrary sequence indices.

Animal opsins are GPCRs, so Ballesteros–Weinstein numbering applies. The nine query opsins—rod opsin (*B. taurus*), cone MWS and cone SWS (*H. sapiens*), parapinopsin (*I. punctatus*), encephalopsin, melanopsin, neuropsin, peropsin, and RGR (all *H. sapiens*)—span the evolutionary breadth of Type II opsins. I annotated these nine sequences at 18 functionally important GRN positions to reveal which positions are conserved and which vary between subfamilies.

Figure 2.5 shows the resulting alignment table. A researcher familiar with this notation can immediately identify which structural position is being discussed and relate it to any other GPCR—this shared vocabulary is what makes the analysis transferable. The 18 positions fall into three functional categories [42]. Conserved anchors maintain the helical scaffold: N at 1.50 and D at 2.50 coordinate a water-mediated hydrogen bond network between TM1 and TM2, while W at 4.50 stabilizes the helix bundle. Spectral tuning sites surround the chromophore: position 3.28 is the counterion (E in visual opsins, shifting to D or other residues in non-visual subfamilies), and position 3.32 varies between A, S, and T—small residues with different hydrogen-bonding capabilities that modulate the electrostatic environment around

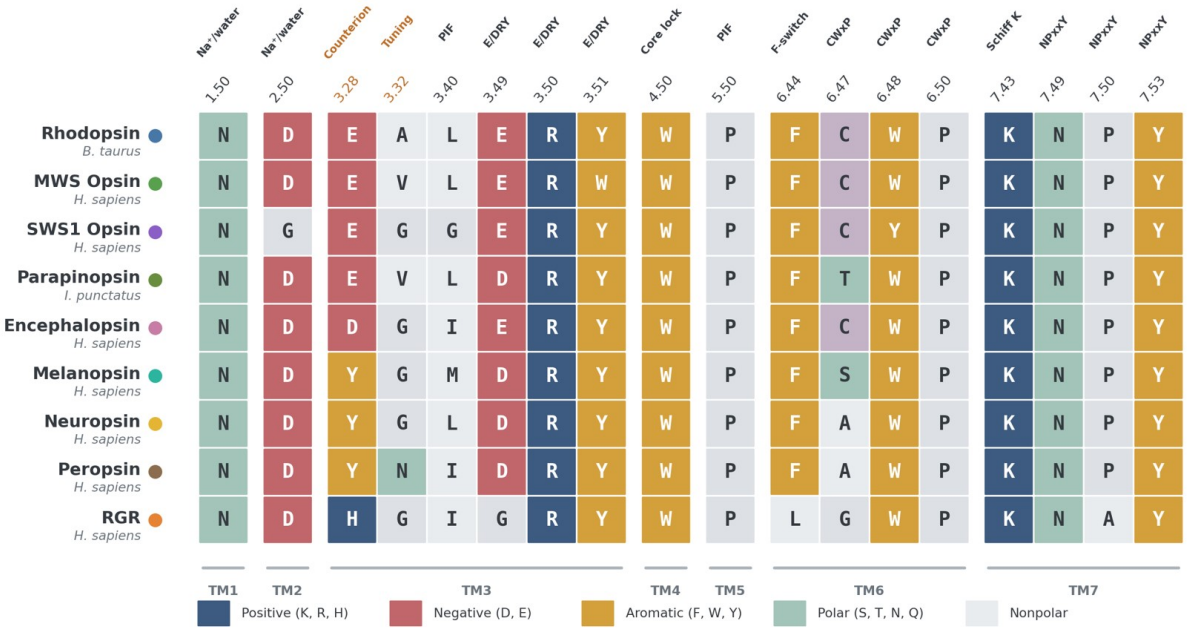


FIGURE 2.5: GRN microswitch and spectral tuning table. Nine query opsins (rows) annotated at 18 GRN positions (columns). Amino acids colored by physicochemical property: positive (slate), negative (terracotta), aromatic (ochre), polar (sage), nonpolar (gray). Columns grouped by transmembrane helix with functional labels. The Schiff base lysine at 7.43 is absolutely conserved; spectral tuning sites (3.28, 3.32) show subfamily-specific variation.

retinal. Activation microswitches [43] form the signal transduction relay: E/DRY at 3.49–3.51 (ionic lock), PIF at 3.40, 5.50, and 6.44 (inter-helix connector), CWxP at 6.47–6.50 (TM6 toggle switch), NPxxY at 7.49–7.53 (G protein coupling), and the Schiff base lysine at 7.43—absolutely conserved because every opsin requires it to bind retinal.

One departure stands out. Cone SWS opsin has glycine at position 2.50 instead of the canonical aspartate. This is confirmed across 3,237 SWS1 subfamily members in the dataset (52% G, 38% N, 8% D), indicating evolutionary loss of the sodium coordination site in this UV/blue-sensitive lineage—consistent with the known absence of sodium sensitivity in short-wavelength opsins.

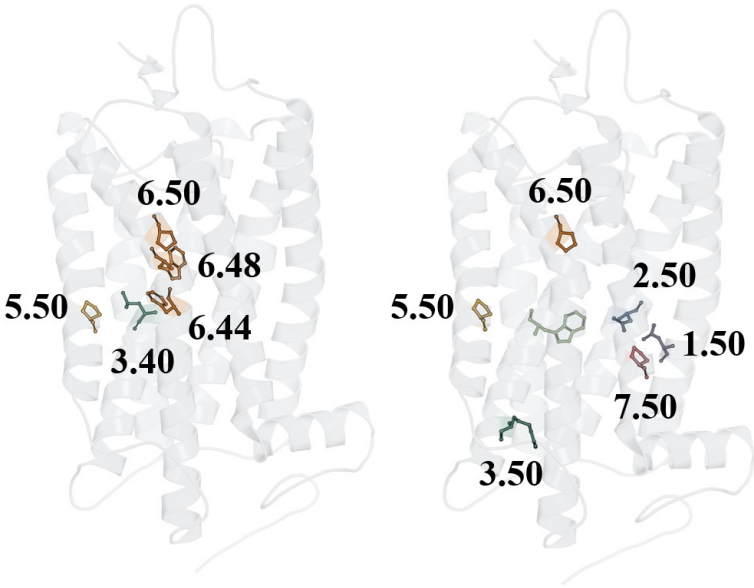


FIGURE 2.6: Selected GRN positions mapped onto bovine rhodopsin (PDB: 1U19), shown as two views of the same structure. Left: PIF connector (3.40, 5.50) and CWxP toggle switch (6.44, 6.48, 6.50). Right: conserved helix anchors (1.50, 2.50, 3.50, 7.50) and neighboring activation microswitches. Residues shown as sticks, colored by functional category. Retinal chromophore in rust.

Figure 2.6 maps a subset of these positions onto bovine rhodopsin. The microswitches form a connected relay from the extracellular retinal-binding site through the transmembrane core to the cytoplasmic G protein interface.

Type I opsins are not GPCRs and Ballesteros–Weinstein does not apply. MOGRN [**hidber2026mogrn**], described in chapter 3, fills this gap by establishing a structure-based numbering system for microbial rhodopsins anchored to the retinal-binding site. GRN positions are sparse and

curated—18 positions per sequence. Dense per-residue features, capturing patterns across entire sequences, complement them.

2.5 EMBEDDING PROCESSOR

Protein language models (pLMs) such as ESM-2 [14] and Ankh [16] are neural networks trained on millions of protein sequences. They learn not just the grammar of protein sequences—which amino acids co-occur and which substitutions are tolerated—but also capture information about the underlying structure and function, producing information-dense high-dimensional vectors called embeddings. The Embedding Processor wraps model inference, stores the resulting tensors as protein-linked datasets, and provides the pooling and extraction operations needed for downstream use.

Embedding Processor

```
emb_proc = EmbeddingProcessor(
    model_name="ankh_large"
)
embeddings = emb_proc.embed_sequences(
    sequences,
    embedding_type="per_residue"
)
```

Ankh Large produces a 1536-dimensional vector for each residue in the sequence—for a 300-residue protein, a matrix of 300×1536. Averaging across residues yields a single 1536-dimensional vector summarizing the whole protein. The per-residue matrix provides node features on contact graphs, where each node gets its own embedding. The mean-pooled vector enables similarity searches and clustering across the full dataset. The Embedding Processor supports ESM-2 and Ankh model families across multiple dimensionalities, with per-residue and mean-pooled embedding types.

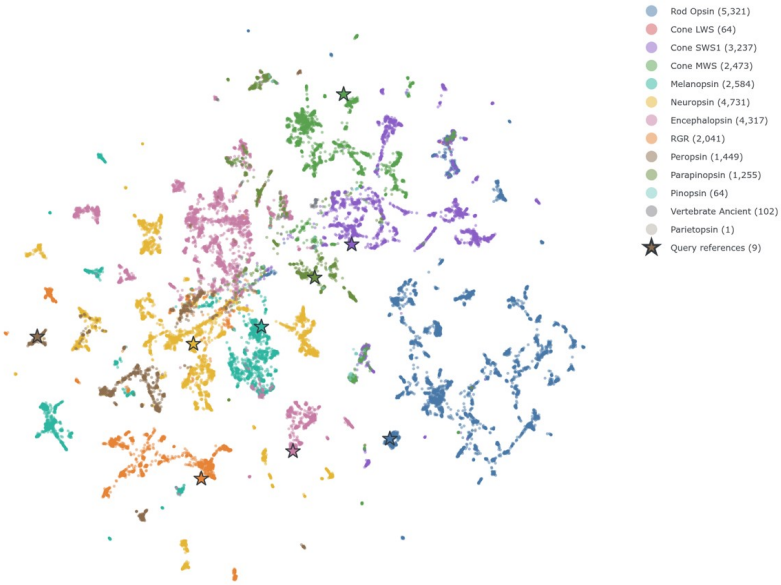


FIGURE 2.7: UMAP projection of Ankh Large mean-pooled embeddings across nine Type II opsin subfamilies. Each point is one protein, colored by subfamily. Star markers indicate the nine query references. Subfamilies cluster without supervision: rod opsins (steel blue) dominate the center; cone opsins, melanopsin, and non-visual subfamilies occupy separate regions.

I embedded all 27,639 sequences with Ankh Large. Mean-pooling across sequence length yields one vector per protein. UMAP [45] projection into two dimensions (Figure 2.7) reveals the structure of opsin embedding space without supervision—the model received no subfamily labels, spectral measurements, or structural information during training.

Subfamilies cluster: proteins within a subfamily produce more similar embeddings than proteins across subfamilies, confirming that the learned representations capture functional groupings. The nine query sequences, marked as stars, sit within their respective subfamily clouds.

At this point, ProtOS has demonstrated capabilities for handling sequences, structures, graphs, GRN annotations, and embeddings. Since ProtOS is an integrated system, a protein annotated with all of these links them: GRN positions map a sequence onto the fold, enabling construction of a binding pocket graph from sequence alone, and embeddings enrich each node with learned per-residue features. What remains is linking these to the measurements that describe each protein's biology.

2.6 PROPERTY PROCESSOR

Every protein carries associated values—numbers, labels, categories that describe what is known about it. The Property Processor stores any such tabular data in CSV tables keyed to entity identifiers. A row might hold a measured λ_{max} from a spectroscopy paper, a subfamily classification, a species name, or an experimental condition. New columns can be added as a project progresses. When LAMBDA returns a predicted λ_{max}, that prediction enters the same table alongside the experimental measurement—queryable in the same way, linked to the same protein.

Property Processor

```
prop_proc = PropertyProcessor()
prop_proc.record_properties(
    "atlas_properties", rows,
    allow_create=True
)
table = prop_proc.load_table(
    "atlas_properties"
)
```

For the opsin research in this thesis, the property table links each of the 27,639 sequences to annotations such as subfamily, sequence identity to the query, sequence length, and species—but also to opsin-specific properties like measured λ_{max}. Approximately 1,800 opsins have published absorption spectra, assembled from decades of spectroscopy across both Type I and Type II families. The 27,639 sequences have embeddings and GRN annotations, but fewer than 7% have measured absorption data. This sparse coverage—rich annotations, limited measurements—is characteristic of protein science and illustrates why linking modalities matters. With all modalities linked, the final piece is using this integrated data as input to predictive models.

2.7 MODEL MANAGER

A growing number of machine learning models serve structural biology, from structure prediction to protein design. These models consume varying types of data and produce different output formats. The ProtOS Model Manager serves as an interface between ProtOS data and these models, preparing job submissions from processor outputs and ingesting results back into the system. Each model is described by its expected inputs and outputs; adding a new model requires only this specification. Models already integrated include Boltz2 [13] for structure prediction and Ankh [16] for embeddings, as introduced in the introduction.

Model Manager

```
manager = ModelManager()
inv = manager.prepare(
    "boltz2",
    inputs={"sequence": seq,
            "name": "design_01"}
)
result = manager.run_and_ingest(inv)
```

The `prepare` call packages processor outputs into the format Boltz2 expects; `run_and_ingest` executes the job and registers the returned structure in ProtOS. The same pattern applies to any model. Results re-enter ProtOS as properties or new data modalities, available to every processor.

2.8 DISCUSSION

Starting from nine query sequences, I assembled a multi-modal dataset of 27,639 proteins. Each protein carries a sequence, GRN annotations at standardized positions, per-residue embeddings, and—where data exists—measured spectral properties stored in the property table. From the GRN annotations and sequences, binding pocket graphs can be constructed without requiring a structure for every protein. These are the core representations that the research chapters build on.

3

MOGRN — A GENERIC RESIDUE NUMBERING SYSTEM FOR MICROBIAL RHODOPSINS

I believe MOGRN will have a significant impact on the optogenetics and microbial rhodopsin communities. For decades, GPCR researchers have shared a structural vocabulary—Ballesteros–Weinstein numbering—that made results immediately transferable across laboratories. Microbial rhodopsin research has lacked this. MOGRN provides it.

3.1 INTRODUCTION

The work summarized here is presented in "A Generic Residue-Numbering System for Microbial Rhodopsins — Unifying Structural Frameworks and Functional Mapping" [**hidber2026mogrn**]. I am a co-first author on this publication.

Position 85 in bacteriorhodopsin is the counterion—the negatively charged aspartate that stabilizes the protonated Schiff base. In channelrhodopsin-2, the equivalent function sits at position 123. The numbers differ because each lineage accumulated different insertions and deletions over billions of years. Recognizing that these two residues serve the same role requires structural comparison; sequence numbering alone cannot reveal it.

GPCRs solved this problem decades ago. Ballesteros–Weinstein numbering [40] assigns each position a helix-relative identifier—helix number and offset from the most conserved residue—creating coordinates that translate across the entire superfamily. GPCRdb built on this foundation [41]. When a researcher reports that a mutation at position 3.32 affects ligand binding, every GPCR researcher knows where that is. Microbial rhodopsins—functionally and taxonomically diverse—had nothing equivalent. Laboratories developed ad hoc conventions, referencing bacteriorhodopsin numbering, channelrhodopsin-2, or the C1C2 chimera. Identity between major classes falls as low as 10–12% [46], too low for reliable sequence alignment. A numbering system for microbial rhodopsins had to be anchored to structure.

We developed MOGRN (Microbial Opsin Generic Residue Numbering), anchored to the one feature all microbial rhodopsins share: the retinal-binding pocket. The system is defined by structural alignment of 129 structures spanning all major functional classes, producing a curated reference table, and was applied to approximately 40,000 sequences. The key elements follow; the full methodology, results, and discussion are in the linked publication.

3.2 SUMMARY

MOGRN follows the Ballesteros–Weinstein convention: each position receives a two-part identifier indicating the helix (1–7) and the position relative to a helix-specific anchor. The anchor on each helix is the residue closest to retinal, designated X.50. The Schiff base-forming lysine on helix 7 becomes position 7.50 by definition. Positions toward the N-terminus receive

lower numbers; positions toward the C-terminus receive higher numbers. This ligand-centric definition ensures that the anchors fall within the most conserved region of each helix—the retinal-binding pocket—providing a stable reference even across distantly related families.

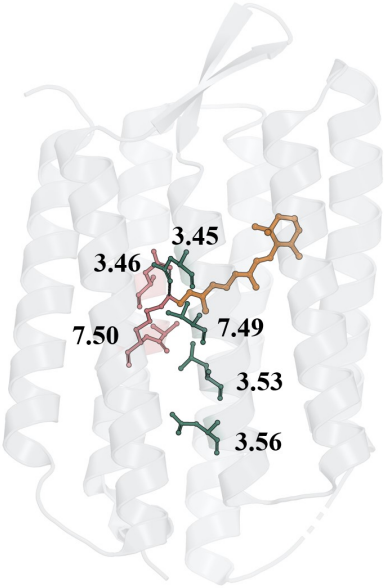


FIGURE 3.1: Bacteriorhodopsin (PDB: 1C3W) with key MOGRN positions highlighted. The Schiff base lysine K216 (7.50), counterions D85 (3.45) and D212 (7.46), and TM3 residues T89 (3.49), W86 (3.46), and D96 (3.56) are shown as sticks. Retinal in rust.

Figure 3.1 shows the system on bacteriorhodopsin (PDB: 1C3W [26]). The Schiff base lysine K216 sits at 7.50, the counterions D85 and D212 at 3.45 and 7.46, and the TM3 motif residues that determine ion pump specificity at their respective positions. These residues, traditionally referenced by protein-specific numbers that differ in every homolog, become directly comparable across the superfamily.

The TM3 motif at positions 3.45–3.49–3.56 illustrates what standardized coordinates reveal (Table 3.1). These three positions encode functional identity across ion-pumping rhodopsins. Outward proton pumps carry DTD (aspartate–threonine–aspartate), inward proton pumps carry FSE or DTA, chloride pumps carry TSA or NTQ, and sodium pumps carry NDQ. Before MOGRN, identifying these motifs meant structural alignment for each new protein. In standardized coordinates, they become queryable signatures. Additional functional positions include spectral tuning switches at 3.53, 4.51, 6.54, and 7.49, the retinal planarity determinant at 4.54 (the non-G rule), and lateral fenestration sites at 5.44 and 5.47.

Function	3.45	3.49	3.56	Example
Outward H^+ pump	D	T	D	HsBR
Inward H^+ pump (SzR)	F	S	E	SzR4
Inward H^+ pump (XeR)	D	T	A/L/S	NsXeR
Cl^- pump (archaeal)	T	S	A/D	NpHR
Cl^- pump (bacterial)	N	T	Q	NmClR
Na^+ pump	N	D	Q	KR2

TABLE 3.1: TM3 motif at positions 3.45–3.49–3.56 across microbial rhodopsin functional classes.

The reference table was built from 129 structures—69 experimental and 60 computationally predicted—spanning all major functional classes. To annotate a new sequence, the GRN Processor finds the highest-identity reference in this table, aligns the query against it, and transfers GRN positions from the reference. Predicted structures achieved sub-angstrom accuracy for the binding pocket (0.51 Å to 0.75 Å iRMSD), confirming that current structure prediction methods are reliable for MOGRN annotation. The system accommodates non-canonical architectures: heliorhodopsins with inverted membrane topology, enzyme-fused rhodopsins with an additional TM0 helix, and local structural distortions that produce gaps or insertions. Separately, approximately 40,000 non-redundant sequences were aggregated from genomic and metagenomic sources—independently of the Type II opsin dataset used in LAMBDA—and annotated with MOGRN positions, identifying 31 sequence clusters: 14 containing characterized rhodopsins and 17 containing only uncharacterized sequences, representing unexplored functional diversity.

The curated reference table enables GRN annotation for both opsin families: the GRN Processor annotates type I opsins using the MOGRN table, while type II opsins use Ballesteros–Weinstein positions. Both families now map to helix-relative coordinates anchored to the retinal pocket, so binding pocket graphs can be built for any opsin regardless of family.

4

LAMBDA — CROSS-FAMILY SPECTRAL PREDICTION

What do a Honduran yellow-shouldered bat and a beluga whale have in common?

LAMBDA (Light Absorption Modeling through Binding Domain Analysis) is a graph neural network that predicts opsin absorption maxima from binding pocket composition. It predicts λ_{max} across any retinal-binding fold—type I, type II, and engineered proteins like hCRBPII—using binding domain analysis rather than family-specific regression. It predicts multiple λ_{max} values simultaneously (11-*cis*, all-*trans*, protonated, deprotonated), enabling estimation of spectral shifts between chromophore states. And because GRN-annotated sequence is sufficient input, it scales to tens of thousands of predictions without requiring structures. These capabilities serve three applications: the Opsin Atlas (47,700 sequences), spectral engineering through binding pocket mutations, and understanding how retinal-binding pockets might be incorporated into synthetic folds.

4.1 INTRODUCTION

Optogenetics [22] depends on matching light wavelength to protein response. Type I (microbial) opsins provide the ion channels and pumps that enable neural activation and silencing [47]. Type II (animal) opsins underlie visual neuroscience and increasingly serve as light-activated GPCRs for controlling intracellular signaling [19]. In both cases, the absorption maximum λ_{max} determines which wavelengths activate the protein, and selecting or engineering opsins with specific λ_{max} values is a core challenge. Red-shifted opsins ($\lambda_{max} > 580\,\text{nm}$) enable deeper tissue penetration because longer wavelengths scatter less in biological tissue. Spectrally separated pairs—one blue-absorbing, one red-absorbing—allow multiplexed control of distinct cell populations in the same preparation [48]. Many optogenetic applications also require a distinct off-switch: bistable opsins can be both photoactivated and photoreversed, but this bidirectional control requires sufficient separation between dark-state and activated-state absorption maxima. If the two states absorb at similar wavelengths, a single light source drives both forward and reverse photoreactions, precluding selective control. Predicting λ_{max} for both chromophore states of the same protein, and the spectral gap between them, is therefore also important.

Beyond opsins, engineered systems demonstrate that retinal can be bound and spectrally tuned in protein folds entirely unrelated to opsins. Human cellular retinol binding protein II (hCRBPII) [49] is the clearest example: As protein engineering capabilities grow—through directed evolution, computational design, and de novo fold generation—the prospect of incorporating retinal-binding pockets into novel protein scaffolds creates the possibility of coupling light activation to functions beyond the ion transport, channel gating, and GPCR signaling that natural opsins provide. Understanding how binding pocket architecture determines spectral properties, independent of global fold, is a prerequisite for this direction.

Light absorption follows the relationship $E = hc/\lambda$: a photon is absorbed when its energy matches the gap between electronic states of the chromophore. The protein environment tunes this gap through electrostatic interactions with the conjugated polyene chain of retinal. The counterion [38]—a negatively charged residue 3 Å to 4 Å from the protonated Schiff base—is the primary determinant, but the entire network of binding pocket residues contributes, and single mutations at key positions can shift λ_{max} by 10 nm to 40 nm [27]. All retinal-binding proteins use the same chromophore yet span λ_{max} values from ∼350 nm (UV) to ∼650 nm (far-red), with the variation arising along two independent axes: geometric isomerism and protonation state. Type II opsins bind 11-*cis* retinal in the dark state; type I opsins and photoactivated type II opsins bind all-*trans* retinal. Independently, the Schiff base linkage to the conserved lysine can be protonated (the dominant state for most opsins, absorbing in the visible range) or deprotonated (as in UV-sensitive SWS1 opsins, which bind 11-*cis* retinal but absorb below 400 nm).

Both properties—isomer state and protonation—are determined by the same binding pocket environment (Figure 4.1). Because energy is inversely proportional to wavelength, spectral tuning becomes increasingly difficult to resolve at long wavelengths.

Tens of thousands of opsin sequences exist in databases, but fewer than 3,000 have measured λ_{max} values. Existing computational methods—OPTICS [30], the Inoue model [50], RhoMax [32]—achieve reasonable accuracy within their respective domains (chapter 1), but they share two fundamental limitations. They are fold-specific—none can predict across both opsin families or handle retinal-binding proteins with different folds. And they predict only a single λ_{max} for their target family's native chromophore state, unable to predict activated states, deprotonated states, or the spectral gap between chromophore states that determines whether bidirectional control is feasible.

LAMBDA addresses these limitations by framing spectral prediction as a question about binding pockets rather than sequences. The same pocket representation implicitly encodes the chromophore environment for different retinal conformations—11-*cis*, all-*trans*, protonated, deprotonated—so the model predicts λ_{max} for multiple states simultaneously from a single shared representation. Predicting both dark-state (11-*cis*) and activated-state (all-*trans*) absorption maxima for the same protein yields the spectral shift $\Delta\lambda$ between on and off states—a quantity that determines whether bidirectional optogenetic control is feasible and that no existing method can estimate.

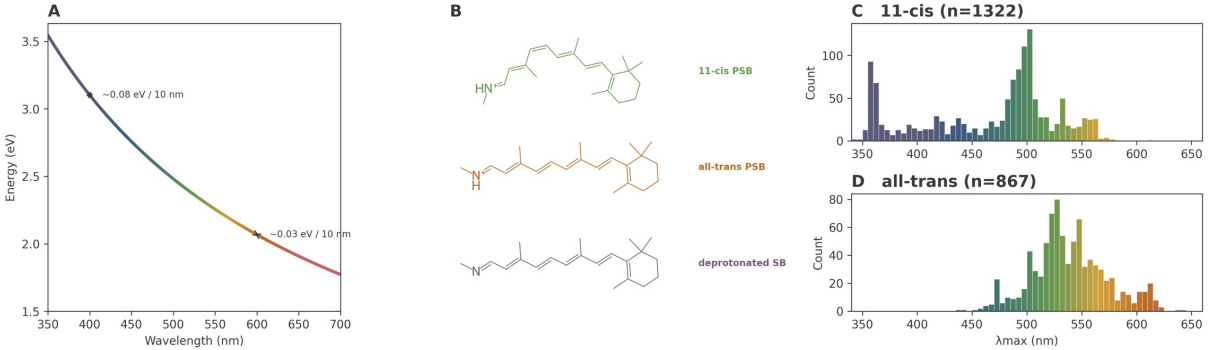


FIGURE 4.1: Spectral tuning in opsins. (A) The relationship between energy gap and absorption wavelength ($E = hc/\lambda$). Equal wavelength differences correspond to smaller energy differences at longer wavelengths (see main text). (B) Retinal Schiff base configurations: 11-*cis* protonated (dark state in type II opsins), all-*trans* protonated (type I opsins and activated type II), and deprotonated (UV-sensitive opsins). (C,D) Distribution of measured $\lambda_{\max}$ values for 11-*cis* and all-*trans* retinal across the training datasets.

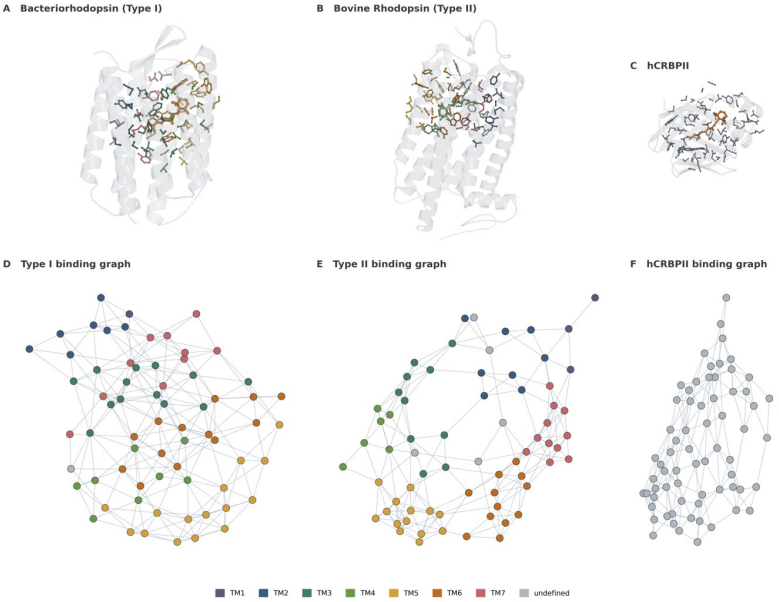


FIGURE 4.2: Binding pocket structures and graph representations for three retinal-binding protein families. Top: crystal structures with pocket residues shown as sticks. Bottom: corresponding contact graphs used as LAMBDA input. (A, D) Bacteriorhodopsin, Type I opsin. (B, E) Bovine rhodopsin, Type II opsin. (C, F) hCRBPII, a β-barrel protein. Opsin residues colored by transmembrane helix (TM1–TM7).

Since spectral tuning is determined by the binding pocket, the natural representation for prediction is the pocket itself—the residues surrounding retinal and their spatial contacts—rather than the full protein sequence or the global fold. In LAMBDA, binding pockets are represented as graphs: nodes are binding pocket residues, edges encode spatial contacts between them. This topology captures which residues interact, independent of their sequence positions or the global architecture that places them there. Figure 4.2 illustrates this across three structurally unrelated protein families. Type I opsins (bacteriorhodopsin, 1C3W) and type II opsins (bovine rhodopsin, 1U19) are both seven-transmembrane alpha-helical bundles, but with independently evolved architectures—different helix tilts, loop connectivities, and retinal binding orientation—and no detectable sequence homology. hCRBPII (4QYP) adopts an entirely different fold—a β-barrel—yet engineered variants bind retinal and span over 200 nm of spectral tuning through only nine point mutations. Despite these architectural differences, the graph representation reduces each binding pocket to the same format: nodes are residues within contact distance of retinal, edges encode spatial adjacency within 4 Å. The opsin graphs have similar sizes (68 and 69 nodes); the hCRBPII graph has comparable complexity (67 nodes). The same model

architecture processes all three graph types, learning spectral tuning determinants from binding pocket composition regardless of the surrounding protein scaffold.

A binding pocket graph derived from a crystal structure is the definitive representation, but structures exist for only a fraction of known opsins. GRN annotation (chapter 3) is what makes the approach scalable: a sequence annotated with GRN coordinates maps directly onto the family's reference binding graph topology. GRN tells us which pocket positions are present and where they sit relative to retinal, so binding pocket graphs can be constructed from sequence alone—no structure required. Aligning reference structures from both families on retinal reveals that certain positions occupy equivalent spatial locations relative to the chromophore: position 7.43 (type II) and 7.50 (type I) are both the Schiff base lysine; 3.28 and 3.57 are both in the counterion region. Treating mapped positions as equivalent graph nodes enables cross-family learning despite the independent evolutionary origins of these folds. Protein language model embeddings from Ankh-large [16], chosen for its balance of embedding quality and computational efficiency, provide 1536-dimensional per-residue feature vectors that serve as node features, encoding amino acid identity and evolutionary context. Both GRN annotation and pLM embeddings require only a protein sequence as input.

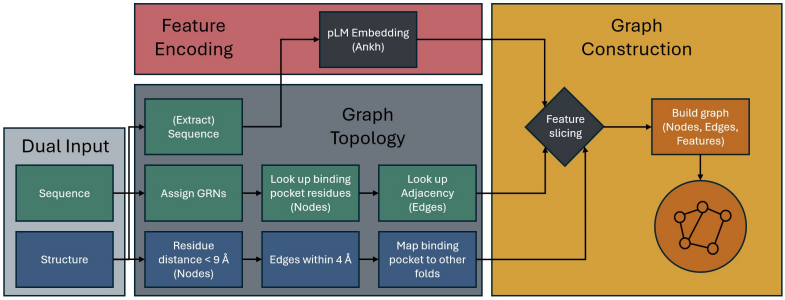


FIGURE 4.3: Data flow for the two binding pocket graph construction pathways. The structure-based path (top) processes any retinal-binding protein through Structure Processor → GRN annotation → Graph assembly. The sequence-only path (bottom) exploits conserved GRN systems to construct graphs directly from sequence via GRN annotation → graph topology lookup → embedding enrichment. Both paths produce the same graph format consumed by LAMBDA.

Two pathways construct LAMBDA input graphs, depending on whether a structure is available (Figure 4.3). The structure-based path processes any retinal-binding protein with a known or predicted structure: binding pocket residues are identified based on proximity to retinal, GRN coordinates are assigned, per-residue pLM embeddings are computed, and the contact graph is assembled. The sequence-only path annotates the sequence with GRN positions, applies the family-specific binding pocket topology directly, and uses pLM embeddings as node features—this is the path used for the Opsin Atlas predictions on 47,700 sequences. Both paths produce the same output: a binding pocket graph where each node carries a positional encoding and a pLM embedding, and edges encode spatial contacts. A graph attention network

learns the mapping from binding pocket composition to spectral properties, trained on 2,120 proteins with measured absorption values across three protein folds.

4.2 METHODS

4.2.1 *Datasets*

Training data were compiled from three sources that together span both opsin families and multiple retinal conformations.

VPOD 1.3 (Visual Physiology Opsin Database) contains 1,253 type II opsin sequences spanning vertebrate and invertebrate species. The dataset includes both wild-type and experimentally characterized mutant opsins, with λ_{max} values for 11-*cis* retinal ranging from 350 nm to 611 nm. VPOD provides the largest collection of animal opsin spectral data, spanning vertebrate visual opsins, invertebrate opsins, and non-visual opsins such as melanopsin.

The Inoue dataset expands coverage to type I (microbial) opsins, containing 785 sequences from 30 species. These sequences were collected from NCBI, metagenomic databases, and the Tara Oceans project, with experimentally determined λ_{max} values for all-*trans* retinal. The dataset spans bacteriorhodopsins, halorhodopsins, channelrhodopsins, and sensory rhodopsins, with absorption maxima ranging from 436 nm to 644 nm.

To enable prediction of the all-*trans* retinal state for type II opsins, training data were augmented with activated-state λ_{max} values collected from the literature. For 31 type II opsins with reported photoactivated or meta-state absorption maxima, the all-*trans* λ_{max} was recorded alongside the dark-state (11-*cis*) value. This approach is necessarily approximate: "activated state" encompasses different photointermediates depending on the opsin's photocycle (meta II in visual opsins, the stable photoproduct in bistable opsins). Nevertheless, these dual-state annotations are suitable as ground truth for training the all-*trans* prediction head on type II opsins. These 31 opsins were assigned exclusively to the training split to maximize the model's exposure to dual-state measurements.

hCRBPII (human cellular retinol binding protein II) represents an engineered system where Wang et al. created retinal-binding variants through targeted mutations. This curated dataset of 51 sequences demonstrates spectral tuning spanning 425 nm to 644 nm with all-*trans* retinal—a range of 219 nm achieved through combinations of nine point mutations. hCRBPII adopts a lipocalin fold unrelated to opsins [51], providing a test case for cross-fold generalization. As with type I and type II opsins, the hCRBPII binding pocket was aligned to retinal and equivalent positions mapped to the nearest GRN coordinates, enabling the same graph-based representation.

Dataset	Samples	Fold	Retinal	λ_{max} range (nm)
VPOD 1.3	1,253	Type II	11-*cis*	350–611
Inoue	785	Type I	all-*trans*	436–644
Dual-state (this work)	31	Type II	both	varies
hCRBPII	51	Lipocalin	all-*trans*	425–644
Total	**2,120**			

TABLE 4.1: Training datasets for LAMBDA.

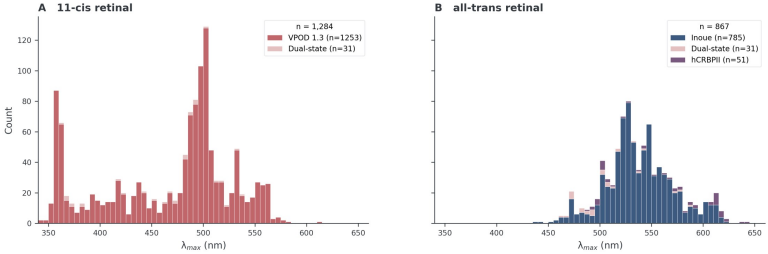


FIGURE 4.4: Distribution of measured λ_{max} values across training datasets. (A) 11-*cis* retinal: VPOD 1.3 ($n = 1{,}253$) and dual-state opsins ($n = 31$). (B) All-*trans* retinal: Inoue dataset ($n = 785$), dual-state opsins ($n = 31$), and hCRBPII ($n = 51$).

The distribution of measured λ_{max} values across these datasets is shown in Figure 4.4.

The diversity of these datasets—spanning different opsin types, species, and retinal conformations—necessitates a multi-output approach. Type II opsins bind 11-*cis* retinal in the dark state, while type I opsins bind all-*trans* retinal. The model addresses this by predicting λ_{max} for multiple retinal states simultaneously, allowing it to learn shared spectral tuning principles while capturing conformation-specific effects. This creates an inherent asymmetry: the all-*trans* prediction head is trained on type I opsins (Inoue, $n = 785$), hCRBPII ($n = 51$), and the 31 dual-state type II opsins, while the 11-*cis* head is trained exclusively on type II data (VPOD, $n = 1{,}253$). The model does produce 11-*cis* predictions for type I opsins, but these are extrapolations with no training signal—type I opsins natively bind all-*trans* retinal, and predicting their absorption with 11-*cis* or other isomers (9-*cis*, 13-*cis*) would require isomer-specific training data that does not currently exist at scale.

4.2.2 *Binding Pocket Graphs*

Type I (microbial) and type II (animal) opsins share no detectable sequence homology, yet both evolved seven-transmembrane architectures that bind retinal. LAMBDA exploits this convergence by representing binding pockets as graphs aligned on the chromophore itself. A single reference structure defines the binding pocket graph for each protein family: bovine

rhodopsin (PDB: 1U19 [37]) for type II opsins and bacteriorhodopsin (PDB: 1C3W [26]) for type I opsins. Within each reference structure, binding pocket residues are identified as those with any sidechain atom within 9 Å of retinal. These residues become nodes in the graph. Edges connect residue pairs with atoms within 4 Å of each other, encoding the network of spatial contacts through the binding pocket.

To enable cross-family learning, the two reference structures are aligned on the retinal polyene chain (atoms C7–C15) using Kabsch superposition [36]. After alignment, binding pocket residues from both families that occupy equivalent spatial locations are identified through sidechain volume overlap: for each pair of GRN positions (one from each family) whose Cα atoms fall within 2 Å, a voxelized sidechain overlap is computed, and pairs exceeding 15% overlap of the smaller sidechain volume are mapped in a greedy one-to-one matching. This procedure yields 25 cross-family position pairs out of $\sim$70 binding pocket positions per family. Positions without a cross-family partner are retained as family-specific nodes. The Schiff base lysine (7x43 in type II, 7x50 in type I) is hardcoded as a mapped pair.

Mapped positions share graph node indices during training, enabling the model to learn that residues at equivalent spatial locations serve analogous roles in spectral tuning despite belonging to different protein folds. Unmapped positions contribute family-specific nodes that capture tuning determinants unique to each fold.

4.2.3 *Preprocessing*

Generating binding pocket graphs for any opsin sequence without requiring experimental structures is achieved through Generic Residue Numbering (GRN) systems, which assign structural coordinates based on sequence alignment alone.

GRN systems label positions within transmembrane proteins using a standardized notation: the helix number followed by a position relative to a conserved reference residue (e. g. 3x50 denotes position 50 in helix 3). For type II opsins, the established Ballesteros–Weinstein [40] numbering system is used. For type I opsins, MOGRN [**hidber2026mogrn**] is used, with reference positions (Xx50) chosen as the residue closest to retinal in each helix.

The preprocessing pipeline operates as follows:

1. **Family classification:** Input sequences are classified as type I or type II based on sequence similarity to reference opsins using MMseqs2 [34].
2. **Reference selection:** The most similar reference sequence with known GRN assignments is identified.
3. **Sequence alignment:** Pairwise alignment against the reference using Biopython's [35] `pairwise2` module.
4. **GRN transfer:** Positions are labeled according to the alignment, assigning GRN coordinates to each residue.
5. **Graph construction:** Binding pocket residues (those with GRN positions in the reference graph) are extracted, and the family-specific edge topology is applied.
6. **Feature generation:** Protein language model embeddings are computed for each node using Ankh-large.

4.2.4 *Model and Training*

LAMBDA uses a graph neural network that processes binding pocket graphs to predict λ_{max} for multiple retinal conformations.

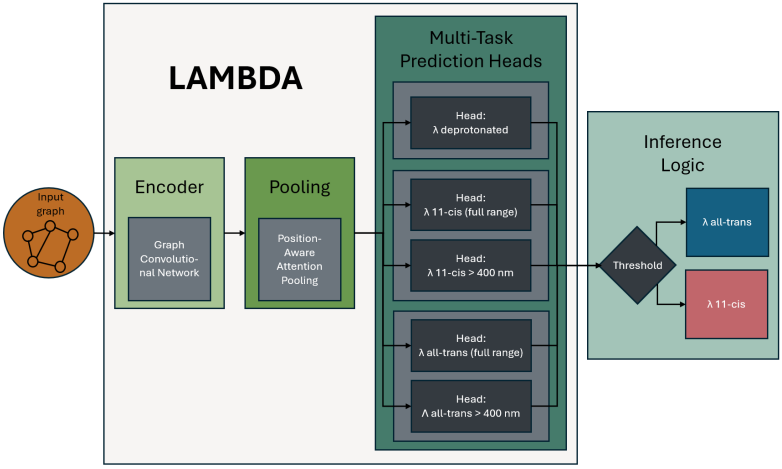


FIGURE 4.5: LAMBDA model architecture. A binding pocket graph is processed by a GCN encoder with residual connections, then aggregated through position-aware attention pooling. Five regression heads predict λ_{max} across retinal conformations and protonation states. Inference logic selects the appropriate prediction based on a 400 nm protonation threshold.

Input representation. Each node in the graph corresponds to a binding pocket residue and carries two feature types: (1) a positional encoding derived from the GRN coordinate, capturing the residue's structural location relative to retinal, and (2) a 1536-dimensional embedding from Ankh-large, capturing evolutionary and biochemical context. The positional encoding is learned during training, allowing the model to discover which binding pocket locations matter most for spectral tuning.

Message passing. The encoder uses a Graph Convolutional Network [52] (GCN) with residual connections to update node representations through neighborhood aggregation. The GCN layer propagates information between connected residues, integrating local interactions across the binding pocket.

Pooling. A position-aware attention pooling mechanism with 32 heads generates a global representation from node-level features. This mechanism uses the GRN positional encodings—one-hot encoding of the mapped binding pocket position encoded through an embedding layer—to compute attention weights over residue contributions, enabling the model to learn multiple complementary views of position-dependent importance.

Multi-task output. The pooled representation feeds into five regression heads that predict λ_{max} across retinal conformations and protonation states. For both 11-*cis* and all-*trans* retinal, two heads operate in parallel: one predicts λ_{max} for protonated Schiff base states only ($\lambda_{max} >$

400 nm), while the other predicts λ_{max} across the full spectral range including deprotonated states. A fifth head predicts λ_{max} for deprotonated Schiff base (UV-absorbing states, $\lambda_{max} <$ 400 nm). Protonation state is classified by a threshold at 400 nm, chosen based on the bimodal distribution of λ_{max} values in the training data: opsins with λ_{max} below this boundary are classified as having a deprotonated Schiff base, and the corresponding UV head prediction is used; otherwise the protonated-state head prediction is reported. All prediction heads share the same encoder—the same learned representation of binding pocket structure feeds all outputs. This shared-encoder architecture reflects the biological reality that a single binding pocket determines spectral properties across chromophore states, and enables generalization to additional retinal conformations (9-*cis*, 13-*cis*) as training data become available.

Loss function. Training uses a combined loss that sums mean squared error across all prediction tasks, with masking to handle missing labels (e. g. type I opsins lack 11-*cis* measurements). The full-range heads (which predict across both protonated and deprotonated states) receive half weight relative to the state-specific heads, as they serve primarily as a regularizer that encourages globally consistent predictions. The split at 400 nm into separate protonated and UV heads allows the model to learn distinct distributions for each protonation state rather than fitting a single non-Gaussian distribution.

Optimization. The model is trained with a learning rate of 2×10^{-4} using a ReduceLROnPlateau schedule that decays the rate to a minimum of 1×10^{-7} when validation loss plateaus. A batch size of 2 is used; larger batch sizes consistently degraded learning, likely because the small dataset and high variance across opsin families benefit from the noisier gradient estimates of very small batches.

Regularization. To prevent overfitting, dropout is applied within the GCN encoder and attention pooling layers, weight decay on all parameters, and early stopping based on validation loss with a patience of 50 epochs.

Data splitting. The combined dataset is divided into training (80%), validation (10%), and test (10%) sets using stratified sampling. Stratification is performed at two levels: by source dataset (VPOD, Inoue, hCRBPII) and by species within each dataset. The 31 dual-state opsins are assigned exclusively to the training split to maximize exposure to dual-state measurements. The remaining data are split to ensure representative sampling across opsin types and taxonomic groups. For species with fewer than 3 samples, random assignment is used to avoid stratification failures.

Evaluation metrics. Mean absolute error (MAE) in nanometers after denormalizing predictions to the original λ_{max} scale, coefficient of determination (R^2), and the fraction of predictions within 5 nm and 10 nm of experimental values are reported. For protonation state, classification accuracy is reported.

Normalization. Target values are normalized to $[0, 1]$ using the observed ranges in training data: 11-*cis* λ_{max} (401 nm to 611 nm), all-*trans* λ_{max} (401 nm to 644 nm), and UV λ_{max} (340 nm to 399 nm).

4.2.5 *Opsin Atlas*

A comprehensive sequence collection was assembled by mining the NCBI non-redundant protein database to achieve broad coverage of opsin diversity across all major subfamilies in both type I and type II families. A phylogenetically informed sampling strategy was designed

Subfamily	Query	Accession
Proton pump	Bacteriorhodopsin (*H. salinarum*)	P02945
Proton pump	Archaerhodopsin-1 (*Halorubrum* sp.)	P69051
Proton pump	Archaerhodopsin-2 (*Halobacterium* sp. aus-2)	P29563
Chloride pump	Halorhodopsin (*H. salinarum*)	P0DMH7
Sensory rhodopsin I	SRI (*H. salinarum*)	P0DMH8
Sensory rhodopsin II	SRII (*N. pharaonis*)	P42196
Cation channel	Channelrhodopsin-1 (*C. reinhardtii*)	A0A2K3CXC9
Cation channel	Channelrhodopsin-2 (*C. reinhardtii*)	Q8RUT8
Anion channel	GtACR1 (*G. theta*)	L1JRS2
Green-absorbing PR	Proteorhodopsin (SAR86)	Q9F7P4
Blue-absorbing PR	Proteorhodopsin (HOT75)	Q9AFF7
Heliorhodopsin	HeR (metagenome)	A0A2P2C3K4
Xanthorhodopsin	Xanthorhodopsin (*S. ruber*)	Q2S2F8

TABLE 4.2: Type I (microbial) opsin query sequences for the Opsin Atlas.

using BLAST searches seeded from representative sequences of each major opsin subfamily. Each query was searched against NCBI nr using BLASTp, retrieving up to 5,000 hits per query. Because queries from related subfamilies often retrieve overlapping sets of sequences, each protein was assigned to the subfamily of the query to which it had the highest sequence identity. This similarity-based classification supersedes NCBI annotations, which frequently use "bacteriorhodopsin" as a generic label for diverse microbial rhodopsins regardless of their phylogenetic placement—for example, 46% of type I sequences carry "bacteriorhodopsin" in their NCBI annotation, yet similarity-based classification distributes these across eight distinct families, reflecting their actual evolutionary relationships. All per-query results were then merged into unified type I and type II tables.

For type I (microbial) opsins, 13 query sequences covering 10 functional subfamilies were used (Table 4.2).

For type II (animal) opsins, 14 query sequences covering 13 subfamilies were used (Table 4.3).

The retrieved sequences underwent quality filtering. Sequences exceeding 500 residues (likely multi-domain proteins or fusion constructs) were removed. To eliminate redundancy while preserving phylogenetic diversity, exact sequence duplicates were removed and cases where the same gene from the same species appeared under multiple accession numbers were filtered, while retaining orthologous sequences from different species. Most critically, each sequence was validated through the GRN annotation pipeline, retaining only sequences that (1) contained the conserved Schiff base lysine at the expected GRN position (7.43 for type II, 7.50 for type I), and (2) achieved sufficient GRN coverage to generate the input graphs for LAMBDA.

For prediction, LAMBDA was applied with output heads matched to each opsin type's native chromophore configuration. Type I opsins, which bind all-*trans* retinal, received predictions for $\lambda_{\text{max}}^{\text{AT}}$. Type II opsins received predictions for their dark-state chromophore ($\lambda_{\text{max}}^{\text{11-cis}}$), but $\lambda_{\text{max}}^{\text{AT}}$ was additionally predicted for all type II sequences. This dual prediction for type II

Subfamily	Query	Accession
Rod opsin	Rhodopsin (*Bos taurus*)	P02699
Rod opsin	Rhodopsin (*H. sapiens*)	P08100
SWS1 cone opsin	OPN1SW (*H. sapiens*)	P03999
MWS cone opsin	OPN1MW (*H. sapiens*)	P04001
LWS cone opsin	OPN1LW (*H. sapiens*)	P04000
Encephalopsin (OPN3)	OPN3 (*H. sapiens*)	Q9H1Y3
Melanopsin (OPN4)	OPN4 (*H. sapiens*)	Q9UHM6
Neuropsin (OPN5)	OPN5 (*H. sapiens*)	Q6U736
Parapinopsin	Parapinopsin (*I. punctatus*)	O42266
Pinopsin	Pinopsin (*G. gallus*)	P51475
Peropsin	Peropsin (*H. sapiens*)	O14718
RGR	RGR (*H. sapiens*)	P47804
TMT opsin	TMT opsin (*D. rerio*)	R9R6C7
VA opsin	VA opsin (*O. masou*)	O13018

TABLE 4.3: Type II (animal) opsin query sequences for the Opsin Atlas.

opsins—though training data for their all-*trans* state is limited to the 31 dual-state sequences—enables estimation of spectral shifts upon photoactivation, a property critical for optogenetic applications. All data management, sequence processing, and GRN annotation were performed using ProtOS.

4.3 RESULTS

LAMBDA was evaluated on the held-out test set (10% of data, stratified by dataset and species). Performance is reported for each prediction target alongside published results from family-specific methods where available.

Evaluation protocols differ across methods. LAMBDA reports held-out test set performance (10% stratified split). OPTICS uses k-fold cross-validation on VPOD WDS ($n = 1{,}211$) with amino acid property encoding. Inoue et al. report within-family evaluation on their full dataset of 884 type I sequences. RhoMax reports median absolute error across 4 family-aware cross-validation splits on 884 type I sequences (mean AE $= 10.45\,\mathrm{nm}$) and requires structures as input. Despite these protocol differences, all methods operate on comparable dataset sizes and the reported accuracies are broadly comparable. The hCRBPII test set contains only 6 samples, limiting statistical power.

Metric	LAMBDA	OPTICS	Inoue et al.	RhoMax
MAE ± std (nm)				
Type II (11-*cis*)	**5.18 ± 0.82**	5.49	–	–
Type I (all-*trans*)	**6.86 ± 0.89**	–	7.80	6.83[†]
hCRBPII (all-*trans*, $n = 6$)	**5.84 ± 1.13**	–	–	–
R^2				
Type II (11-*cis*)	**0.972**	0.964	–	–
Type I (all-*trans*)	**0.894**	–	–	–
hCRBPII (all-*trans*, $n = 6$)	**0.978**	–	–	–
Other				
Protonation acc.	98.3%	–	–	–
Within 5 nm (11c)	77.6%	–	–	–
Within 5 nm (AT)	52.7%	–	–	–
Within 10 nm (11c)	93.9%	–	–	–
Within 10 nm (AT)	83.9%	–	–	–

[†]Median absolute error; all other values are mean absolute error.

TABLE 4.4: LAMBDA performance on held-out test set compared with family-specific methods.

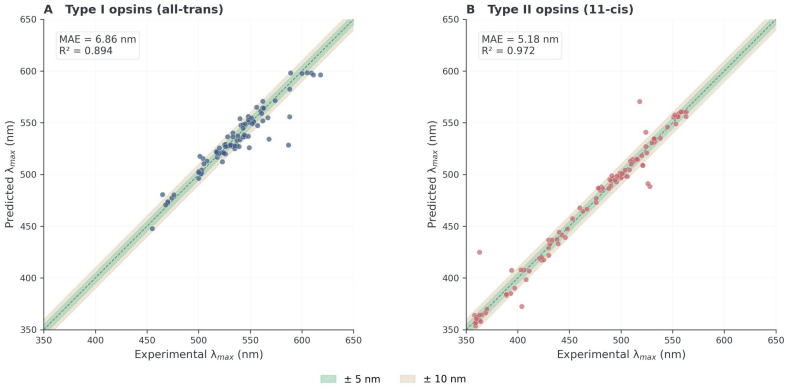


FIGURE 4.6: LAMBDA prediction accuracy on the held-out test set. Predicted vs. measured $\lambda_{\max}$ for (A) type I opsins (all-*trans*) and (B) type II opsins (11-*cis*).

LAMBDA achieved a mean absolute error of 5.18 nm ± 0.82 nm for type II opsins (11-*cis* retinal, $n = 119$) and 6.86 nm ± 0.89 nm for type I opsins (all-*trans* retinal, $n = 87$) on the held-out test set (Figure 4.6). These values reflect the model's actual output, using the protonation-state classification described in Methods to select between head predictions. The model correctly classified protonation state with 98.3% accuracy. For hCRBPII (42 training, 3

validation, 6 test samples), the model achieved an MAE of 5.84 nm ± 1.13 nm ($R^2 = 0.978$)—comparable to opsin performance despite the fundamentally different lipocalin fold. The small test set limits statistical confidence, but the result demonstrates that the binding domain graph framework accommodates non-opsin folds when training examples and appropriate GRN mappings are available.

These results place LAMBDA's accuracy in the same range as the best family-specific methods while predicting across chromophore conformations and protein families from a unified model. That joint training does not degrade within-family accuracy suggests that spectral tuning principles transfer across folds and that the cross-family position mapping captures genuine structural equivalences rather than introducing noise. A systematic underestimation of far-red absorption maxima, where the model saturates near 610 nm, represents the primary limitation and is addressed in the Discussion.

The trained model was applied to 47,700 opsin sequences assembled from NCBI to produce the Opsin Atlas, spanning 20,061 type I opsins across 10 subfamilies and 27,639 type II opsins across 12 subfamilies.

Type I opsins (Figure 4.7A). Predictions of $\lambda_{\max}^{\mathrm{AT}}$ for native all-*trans* retinal show a mean of 519.5 nm ± 28.1 nm (range: 436 nm to 644 nm) across 20,061 sequences. The aggregate distribution approximates a single Gaussian centered near 520 nm, reflecting the relatively homogeneous spectral tuning of microbial rhodopsins around their shared proton-pumping or sensory functions. Within this envelope, individual subfamilies occupy characteristic spectral ranges: proton pump: 532.4 nm ± 23.8 nm (n = 5,108, 25.5%), heliorhodopsin: 498.5 nm ± 37.6 nm (n = 4,993, 24.9%), green-absorbing proteorhodopsin: 520.7 nm ± 16.3 nm (n = 4,942, 24.6%), xanthorhodopsin: 527.3 nm ± 12.5 nm (n = 4,304, 21.5%), blue-absorbing proteorhodopsin: 537.0 nm ± 20.7 nm (n = 310, 1.5%; note that the predicted mean for this subfamily is redder than for green-absorbing proteorhodopsin—the model inverts their relative spectral order, likely reflecting the small sample size and sequence-similarity-based subfamily assignment rather than measured $\lambda_{\max}$), cation channel: 486.1 nm ± 25.6 nm (n = 120, 0.6%), chloride pump: 525.6 nm ± 16.1 nm (n = 103, 0.5%), sensory rhodopsin II: 508.5 nm ± 25.3 nm (n = 89, 0.4%), sensory rhodopsin I: 509.0 nm ± 24.5 nm (n = 86, 0.4%), and anion channel: 478.3 nm ± 15.7 nm (n = 6, <0.1%). Four subfamilies—proton pumps, heliorhodopsins, green-absorbing proteorhodopsins, and xanthorhodopsins—together account for 96.5% of all type I sequences. Proteorhodopsins show a characteristic bimodal distribution, with distinct blue-absorbing and green-absorbing sub-populations reflecting ecological adaptation to different ocean depths. Heliorhodopsins span the widest spectral range but are blue-shifted relative to the classical proton pumps.

Type II opsins (Figure 4.7B). In contrast to the unimodal type I distribution, the 11-*cis* dark-state predictions for 27,639 type II opsins (mean 467.0 nm ± 42.9 nm, range 344 nm to 583 nm) resolve into multiple distinct peaks corresponding to the spectral classes of animal vision. Per-subfamily statistics are: rod opsin: 481.7 nm ± 24.0 nm (n = 5,321, 19.3%), neuropsin: 459.8 nm ± 36.7 nm (n = 4,731, 17.1%), encephalopsin: 465.7 nm ± 21.5 nm (n = 4,317, 15.6%), cone SWS1: 413.0 nm ± 42.3 nm (n = 3,237, 11.7%), melanopsin: 495.2 nm ± 24.6 nm (n = 2,584, 9.3%), cone MWS: 532.4 nm ± 31.3 nm (n = 2,473, 8.9%), RGR: 456.0 nm ± 25.3 nm (n = 2,041, 7.4%), peropsin: 448.3 nm ± 20.5 nm (n = 1,449, 5.2%), parapinopsin: 430.1 nm ± 33.6 nm (n = 1,255, 4.5%), VA opsin: 436.6 nm ± 32.4 nm (n = 102, 0.4%), cone LWS: 486.8 nm ± 26.3 nm (n = 64, 0.2%), and pinopsin: 453.1 nm ± 34.2 nm (n = 64, 0.2%).

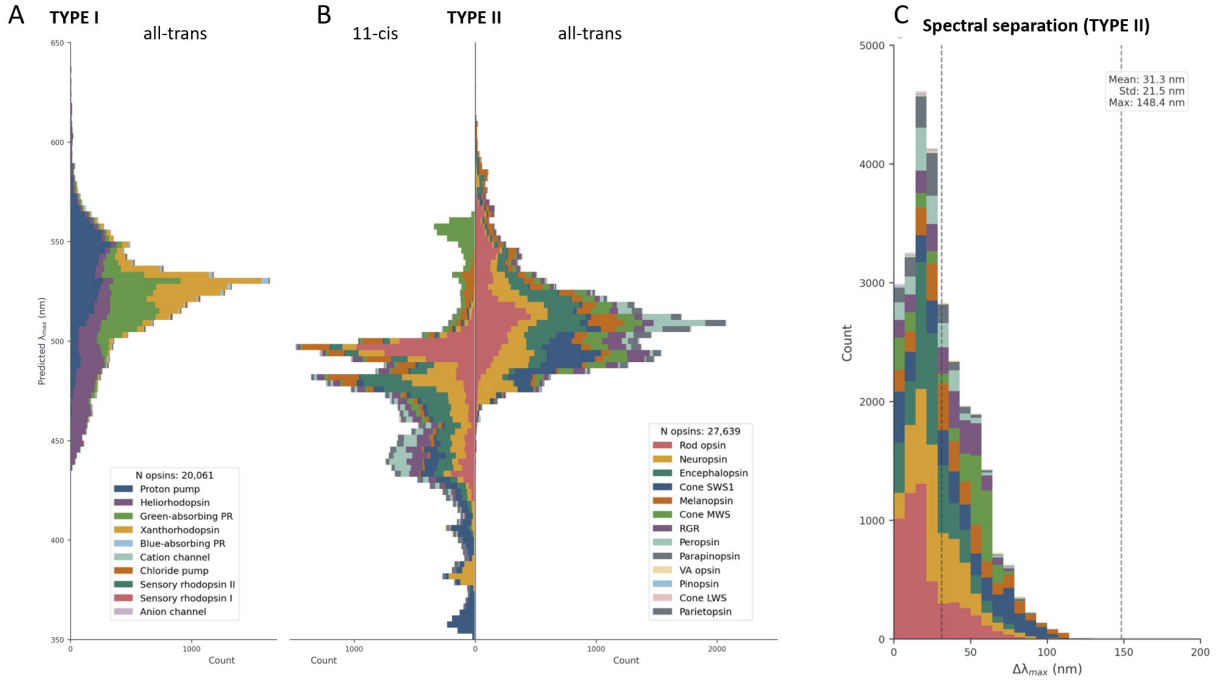


FIGURE 4.7: Opsin Atlas spectral landscape. Predicted $\lambda_{\max}$ distributions for (A) 20,061 type I opsins (all-*trans*), (B) 27,639 type II opsins (11-*cis*), and (C) absolute spectral shift $|\Delta\lambda|$ for type II opsins (protonated states only).

The dominant peak near 500 nm is formed by rod opsins, the primary dim-light photoreceptors. A second major cluster in the UV-blue range below 430 nm comprises cone SWS1 opsins and parapinopsin, which detect short wavelengths for color discrimination and circadian entrainment. Cone MWS opsins form a distinct peak near 530 nm, while melanopsin clusters near 495 nm. Non-visual opsins—encephalopsin and neuropsin—populate the intermediate blue-green range. This multimodal structure reflects the evolutionary diversification of animal opsins into specialized spectral channels for color vision, scotopic sensitivity, and non-visual photoreception.

LAMBDA also predicts $\lambda_{\text{max}}^{\text{AT}}$ for all type II opsins—the absorption maximum with all-*trans* retinal, as occurs during photoactivation or stably in bistable opsins. The $\lambda_{\text{max}}^{\text{AT}}$ distribution is systematically red-shifted relative to 11-*cis* predictions and collapses from the multimodal 11-*cis* pattern into a bimodal shape: a primary peak near 520 nm dominated by rod and long-wavelength cone opsins, and a secondary shoulder near 490 nm comprising SWS1, neuropsin, RGR, and encephalopsin. This dual-state analysis applies only to type II opsins; the model produces 11-*cis* predictions for type I sequences, but these lack biological basis and training signal and are omitted from the atlas.

For type II opsins, predictions for both chromophore states enable estimation of the spectral shift upon activation. $\Delta\lambda$ is computed using the protonated Schiff base predictions only ($\lambda_{\text{max}} > 400\,\text{nm}$), excluding UV-absorbing deprotonated states—the goal is to isolate the spectral shift arising from binding pocket interactions with retinal, not the large blue shift that accompanies Schiff base deprotonation (which moves any opsin into the UV regardless of pocket composition). These $\Delta\lambda$ values are derived quantities: the all-*trans* prediction for type II opsins is trained on only 31 dual-state sequences and has not been independently validated, so the spectral shift estimates should be interpreted as indicative rather than quantitative. The signed difference $\Delta\lambda = \lambda_{\text{max}}^{\text{AT}} - \lambda_{\text{max}}^{\text{11-cis}}$ ranges from -108 to $+148\,\text{nm}$ (mean $+18.8\,\text{nm} \pm 33.0\,\text{nm}$; Figure 4.7C), with 23% of predictions showing a blue shift (negative $\Delta\lambda$). Per-subfamily means reveal consistent differences: VA opsin ($+47.5\,\text{nm} \pm 28.4\,\text{nm}$), cone SWS1 ($+41.7\,\text{nm} \pm 30.7\,\text{nm}$), RGR ($+34.5\,\text{nm} \pm 22.1\,\text{nm}$), and peropsin ($+24.8\,\text{nm} \pm 18.6\,\text{nm}$) show the largest red shifts. Rod opsin ($+17.7\,\text{nm} \pm 20.0\,\text{nm}$), neuropsin ($+17.5\,\text{nm} \pm 28.7\,\text{nm}$), and cone LWS ($+5.2\,\text{nm} \pm 15.8\,\text{nm}$) show modest shifts. Cone MWS is the only subfamily with a negative mean ($-37.9\,\text{nm} \pm 29.9\,\text{nm}$), indicating that these opsins are predicted to blue-shift upon isomerization to all-*trans*. Melanopsin ($+21.6\,\text{nm} \pm 43.4\,\text{nm}$) shows the widest spread, spanning both directions.

The atlas is available as a supplementary resource, providing predicted λ_{max} values for both chromophore states, spectral separation estimates, GRN annotations, and taxonomic metadata for each sequence.

4.4 DISCUSSION

LAMBDA demonstrates that binding domain analysis—representing the chromophore environment as a graph aligned on retinal—enables spectral prediction across protein families from a single model. Where OPTICS and the Inoue model extract features from amino acid sequences, and RhoMax uses structure-derived graphs limited to type I opsins, LAMBDA defines graph nodes by their spatial relationship to the chromophore rather than by sequence position. GRN systems make this a structure-based model that accepts sequence as input.

Accuracy comparable to the best family-specific methods (5.18 nm MAE for type II, 6.86 nm for type I) is achieved while simultaneously predicting multiple chromophore states from a unified model—a capability that sequence-based approaches, which are inherently fold- and conformation-specific, cannot provide. The improvement over OPTICS' sequence-based approach (5.49 nm cross-validated) suggests that binding pocket structure provides more robust features than lineage-specific sequence patterns. Within this representation, pLM embeddings (Ankh-large, 1536 dimensions) describe what occupies each binding pocket position while the graph describes where those positions are relative to retinal—the model depends on both, but the graph structure is what enables cross-family learning.

The most informative limitation is LAMBDA's systematic underestimation of far-red absorption maxima. The model saturates at approximately 610 nm, compressing the 590 nm to 650 nm range even when these samples are included in training. This reflects physics rather than architecture: the relationship $E = hc/\lambda$ means that energy gaps shrink dramatically at long wavelengths, and at these small scales spectral tuning becomes dominated by subtle local effects the model cannot detect—precise positioning of partial charges, fine details of hydrogen bonding, electronic polarization—that produce disproportionately large wavelength shifts. The binding pocket graph captures residue identity and spatial arrangement, but not the electrostatic precision required to resolve 0.03 eV differences in excitation energy. I mentioned this thought about LAMBDA's and current machine learning approaches' limitation in general to the author of the type I opsin dataset, Dr. Keiichi Inoue. We agreed that additional training data are unlikely to help; the features that distinguish a 600 nm opsin from a 640 nm opsin may require explicit quantum mechanical modeling. For optogenetic engineering, LAMBDA can identify candidates as deeply red-shifted but cannot rank them within that regime—precise spectral positioning above 590 nm requires either experimental characterization or hybrid approaches combining machine learning with QM/MM refinement.

More generally, the model cannot generalize to novel folds without training examples, and properties beyond $\lambda_{\max}$ remain outside its scope.

Despite these limitations, the atlas enables computational screening across the full spectral range. LAMBDA is the first method to predict the spectral separation between dark and activated states systematically. The atlas highlights specific candidates for optogenetic applications: a heliorhodopsin from *Candidatus* Kerfeldbacteria (PIS41810) is predicted as the most red-shifted type I opsin at 644 nm, deep into the optical window for tissue penetration; an SWS1 opsin from *Hylia prasina* (NWU40863) as the most blue-shifted type II opsin at 344 nm; and for bidirectional control, a peropsin from the beluga whale *Delphinapterus leucas* (XP_022422838) shows the largest predicted spectral separation at +148 nm (458 nm dark state, 607 nm activated), followed by melanopsins from *Sturnira hondurensis* (+146 nm) and *Phyllostomus discolor* (+143 nm). The ability to predict how mutations affect both chromophore states enables engineering strategies that optimize spectral separation, not just dark-state absorption—candidates that would be impractical to discover through experimental screening alone.

Beyond opsin families, the hCRBPII results demonstrate that LAMBDA's framework extends to entirely different protein folds. Despite the fundamentally different lipocalin architecture, the model achieves accuracy comparable to opsin predictions when training examples are available—though with $n = 6$ samples in the test split this result is statistically weak. The boundary is principled: extending to a new fold requires defining its binding pocket graph and providing representative training data, rather than building a separate model.

An open limitation is the role of solvent in spectral tuning. In opsins, the binding pocket is largely occluded from bulk water, and the model may implicitly learn solvation effects through correlations with surrounding residue identities. In solvent-exposed binding pockets, water molecules directly contact the chromophore and modulate its absorption in ways that depend on occupancy and orientation rather than protein sequence. pLM embeddings may capture some of this context—residues flanking a solvent-exposed pocket differ systematically from those in a buried one—but are unlikely to fully resolve the spectral effects of specific water configurations. For proteins with open binding sites, this remains a source of prediction error that neither additional sequence data nor larger embeddings are likely to eliminate without explicit solvation modeling.

When structures are available, binding pocket graphs can be derived directly. For the sequence-only path—which enables the Opsin Atlas and any large-scale application—LAMBDA depends on standardized positional annotation: without MOGRN coordinates for microbial rhodopsins, the graph representation that enables cross-family learning from sequence would not exist; without the chromophore-centered mapping between MOGRN and Ballesteros–Weinstein positions, type I and type II opsins could not share graph nodes. This dependency is also the source of extensibility—any protein family with a retinal-binding pocket and an appropriate GRN system can be incorporated. Proteins engineered to preserve the binding pocket while redesigning other regions retain predictable spectral properties, because the same pocket determines absorption regardless of what surrounds it.

5

RHODOZYME — LIGHT-ACTIVATED ENZYME DESIGN

What follows is a showcase of ProtOS's capabilities—but more than that, it is what excites me most about my research. Protein design is the natural application of everything ProtOS provides: structure management, annotation, and model integration converge in a pipeline that takes a biological question and produces testable candidates. The validation remains outside the scope of my own work.

Type II opsins undergo a conformational change upon photon absorption. In the dark state, the seven transmembrane helices pack tightly, with the intracellular face closed. Photoisomerization of retinal triggers a cascade of sidechain rearrangements at conserved positions—the microswitch residues—that propagate the conformational change from the binding pocket to the intracellular surface. These microswitches are well characterized [42]: they include residues on TM3, TM5, TM6, and TM7 whose rotameric states differ between the inactive and active conformations. The net effect is that TM6 moves 10 Å to 14 Å outward at its cytoplasmic end, exposing a cavity that in native rhodopsin binds the G protein transducin [24]. This cavity is transient, stereochemically defined, and gated by light. The retinal binding pocket—buried in the transmembrane core—determines the absorption wavelength, while the intracellular cavity is a separate surface. If an enzymatic active site could be placed on this intracellular face, the result would be a light-gated enzyme whose activation wavelength is set by the rhodopsin scaffold.

Can such a design be assembled computationally? The core technique—scaffolding a protein backbone around a fixed arrangement of catalytic residues (a theozyme)—was established by the Baker lab in RFdiffusion2 [53]. Their protocol generates 100 backbone designs per theozyme, fits 8 sequences per design with LigandMPNN, validates all candidates with structure prediction, and picks the best by predicted confidence. We follow the same protocol. The rhodozyme application adds an earlier step and an additional difficulty: the scaffold is not a novel fold but an existing rhodopsin in its active conformation, and the mask that separates fixed from designable regions must preserve both the retinal binding pocket (for light sensitivity) and the theozyme positions (for catalysis). This means candidate selection happens at two stages rather than one—first at theozyme placement (is the geometry compatible with the rhodopsin intracellular face?) and again at validation (does the predicted structure maintain both the rhodopsin fold and the catalytic geometry?). This is where ProtOS contributes. The models are difficult; the data integration between them is not, because ProtOS manages all intermediate representations and dispatches jobs to the Model Manager. The theozyme extraction and placement code (Steps 2–3) required only a small, application-specific addition outside ProtOS's general processor framework.

The design proceeds in six steps. Each step can produce multiple outputs that fan out into the next: the geometric search yielded 8 candidate placements, of which we selected 2 based on geometric quality and helix involvement; for each we generated 50 backbone designs with

RFdiffusion, sampled 8 sequences per design with LigandMPNN, and predicted all resulting candidates with Boltz2. For placement 00, this produced 50 backbone designs (45 successful), sequences from each, and 307 Boltz2 structure predictions.

5.1 STEP 1 — STARTING STRUCTURES

Three structures are required. The first is a rhodopsin in its active conformation—metarhodopsin II (PDB: 3PQR [54]), which captures the open intracellular state with TM6 displaced outward. The dark-state structure of the same protein (PDB: 1U19 [37], bovine rhodopsin at 2.2 Å) serves as a reference for the conformational change: superimposing the two reveals the TM5/TM6 displacement that creates the intracellular cavity. The second is a reference enzyme bound to its substrate in a catalytic intermediate. We use bovine trypsin (PDB: 2AGE [55]), an acyl-enzyme intermediate at 1.15 Å resolution with succinyl-AAPR covalently bound to the catalytic serine. This structure captures the triad geometry mid-catalysis—Ser195, His57, and Asp102 in their active arrangement, with the substrate positioned at the reaction center. The enzyme provides the catalytic geometry to be transplanted; the rhodopsin provides the scaffold that gates access to it.

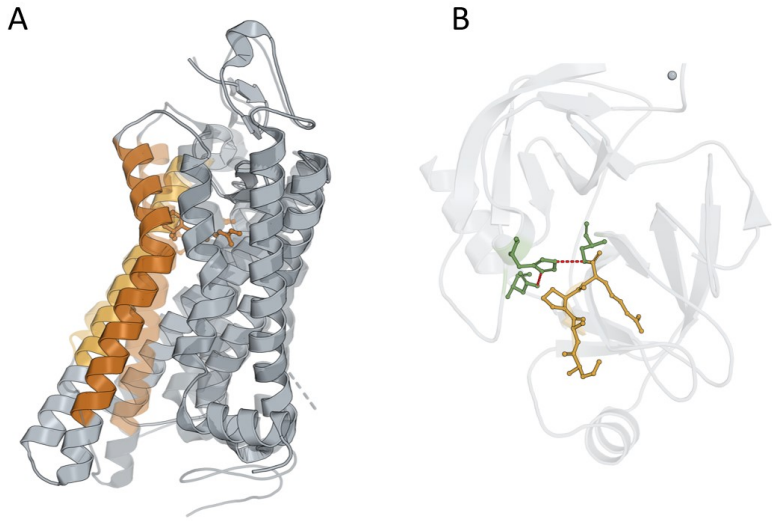


FIGURE 5.1: Input structures and the design premise. (A) Dark-state bovine rhodopsin (1U19, gray) superimposed with active-state metarhodopsin II (3PQR, terracotta), showing TM5/TM6 displacement and the intracellular cavity that opens upon activation. Retinal in rust. (B) Bovine trypsin acyl-enzyme intermediate (2AGE) with the catalytic triad as sticks (Ser195, His57, Asp102) and hydrogen-bond distances shown as dashed lines. Covalently bound substrate (succinyl-AAPR) in ochre.

5.2 STEP 2 — THEOZYME EXTRACTION

The theozyme is the minimal catalytic unit: the three sidechain positions that perform chemistry. For a serine protease, these are the nucleophilic serine, the general base histidine, and the orienting aspartate. From the trypsin structure, we extract three quantities per residue: the Cα coordinate, the Cα→Cβ vector (sidechain direction), and the residue identity. The pairwise Cα–Cα distances define a triangle; the Cβ vectors define the orientation of each sidechain within that triangle. Together, these six quantities (three positions, three directions) specify the catalytic geometry that must be reproduced in the rhodopsin scaffold.

The catalytic triad in 2AGE shows the characteristic hydrogen-bond relay: the Ser195 hydroxyl is 3.04 Å from His57 Nε2, and His57 Nδ1 is 2.75 Å from Asp102 Oδ2. These functional-atom distances define the active geometry. The Cα triangle that supports this arrangement has sides of 8.3 Å (Ser195–His57), 6.5 Å (His57–Asp102), and 10.1 Å (Ser195–Asp102).

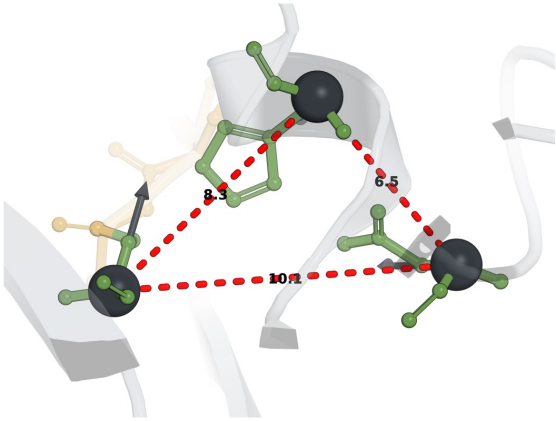


FIGURE 5.2: Theozyme extraction. The catalytic triad shown in the trypsin active site (gray cartoon) with the geometric abstraction overlaid: Cα positions as spheres, Cα→Cβ direction vectors as arrows, pairwise distances as dashed lines. Distances: Ser195–His57 = 8.3 Å, His57–Asp102 = 6.5 Å, Ser195–Asp102 = 10.1 Å. This triangle and these vectors are the input to the placement search—everything else about trypsin is discarded.

5.3 STEP 3 — THEOZYME PLACEMENT

The intracellular face of active rhodopsin is identified using GRN annotation. Residues on TM helix ends that face the cytoplasm (TM1 $\geq$ 1.60, TM3 $\geq$ 3.55, TM5 $\geq$ 5.68, TM7 $\geq$ 7.53) and intracellular loop residues (ICL1, ICL2, ICL3, H8) form the candidate region. An exhaustive search over all triplets in this region finds positions whose Cα triangle matches the theozyme triangle within 2 Å RMSD. Candidates passing the distance filter are then tested for sidechain

direction: the source Cα triangle is Kabsch-aligned onto the candidate, and the rotated Cβ vectors are compared. Matches within 30° are retained.

An additional constraint requires at least one residue on TM5 or TM6—the helices that move during activation. This ensures that the catalytic geometry depends on the active conformation: in the dark state, with TM6 packed inward, the triad distances break. The enzyme turns on with light and off without it.

This step has no equivalent in the Baker lab protocol. Ahern et al. start from a theozyme and generate a novel scaffold around it; the placement is implicit in the diffusion process. Here, the scaffold is fixed—we must find positions in an existing structure that can accommodate the catalytic geometry. The search produced 8 candidates. We selected placements 0 and 2 based on triangle RMSD, Cβ vector alignment, and the involvement of TM6 residues. For placement 00, the theozyme maps to Ser-230, His-245, and Asp-250 on the intracellular face of 3PQR. The placement reproduces the trypsin Cα triangle exactly (8.3 Å / 6.5 Å / 10.1 Å) and preserves the hydrogen-bond distances (Ser–His 3.04 Å, His–Asp 2.75 Å).

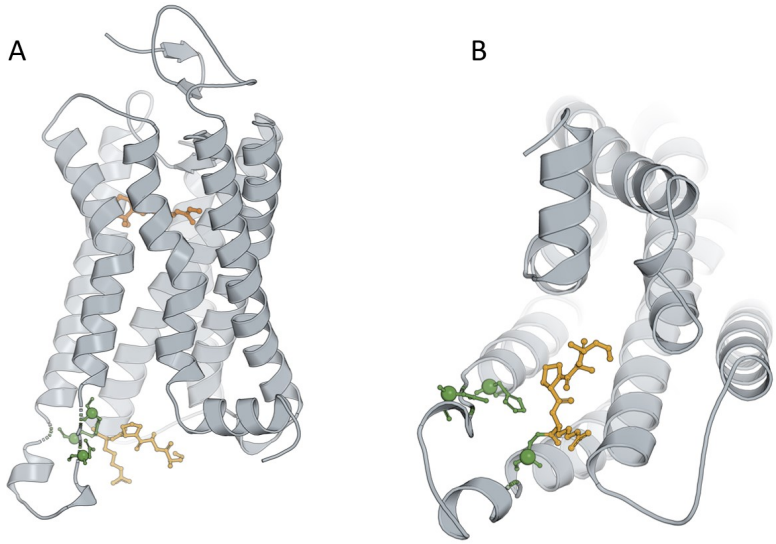


FIGURE 5.3: Theozyme placement on the rhodopsin scaffold. (A) The placement structure (3PQR with theozyme mutations) viewed from the intracellular face. TM helices in gray, theozyme residues (Ser-230, His-245, Asp-250) shown as sticks with Cα spheres in green. Retinal in rust. (B) Same structure from an alternate angle showing the theozyme sidechain arrangement relative to the transmembrane core.

5.4 STEP 4 — BACKBONE DESIGN WITH RFDIFFUSION

Point mutations at the matched positions would introduce the correct residues but not the correct backbone geometry. The surrounding loops and helix termini must accommodate the theozyme. RFdiffusion [56] generates backbone designs under constraints.

The mask defines what is fixed and what is designed. The locked regions comprise the TM helices (TM1–TM2, TM3, TM4–TM5, TM6–TM7), preserving the transmembrane core, the retinal binding pocket, and the microswitch residues that mediate the activation mechanism—210 residues in total. The three theozyme positions (Ser-230, His-245, Asp-250) are locked with their sidechain atoms explicitly constrained. Between these locked segments, 116 residues are free for RFdiffusion to redesign—these correspond to the intracellular loops (ICL1–3), the theozyme-surrounding loops, and the C-terminal region. Retinal and the tetrapeptide substrate are included as ligand context.

We generate 50 designs per placement. The number reflects a practical trade-off: each placement fans out into 50 designs $\times$ 8 sequences $\times$ structure predictions, and evaluating multiple placements multiplies this cost.

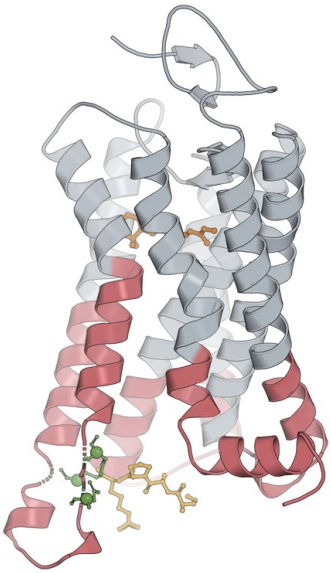


FIGURE 5.4: RFdiffusion mask. The mask applied to the rhodopsin scaffold: TM helices and theozyme positions locked (gray), intracellular loops free for design (terracotta). Theozyme residues marked as green spheres. Retinal (rust) sits in the locked transmembrane core, unchanged by design. 210 residues are locked; 116 are designed.

5.5 STEP 5 — SEQUENCE DESIGN WITH LIGANDMPNN

Each RFdiffusion backbone specifies a fold but not a sequence. LigandMPNN [57] generates amino acid sequences compatible with the designed backbone while accounting for the retinal cofactor and the substrate ligand. The theozyme residues (Ser, His, Asp at the fixed positions) are provided as constraints—LigandMPNN designs the rest.

Following Ahern et al., 8 sequences are sampled per backbone at temperature $T = 0.1$. The retinal SMILES and substrate SMILES are included in the LigandMPNN input so that the designed sequence accounts for both the cofactor that enables light activation and the substrate that the enzyme should bind. For placement 00, 45 of 50 backbone designs succeeded, yielding sequences for downstream evaluation. The top candidate has 72.7% sequence identity to wild-type rhodopsin (3PQR), with 89 mutations concentrated in the redesigned loop regions. The locked TM helices retain the native sequence almost entirely.

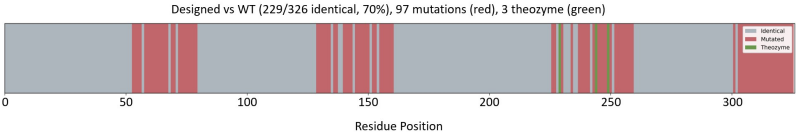


FIGURE 5.5: Sequence design. Sequence alignment of the top candidate against wild-type rhodopsin (3PQR). Identical positions in gray, mutations highlighted. Mutations cluster in the redesigned loop regions; the TM helices are largely unchanged. Theozyme positions (Ser-230, His-245, Asp-250) marked.

5.6 STEP 6 — STRUCTURE PREDICTION WITH BOLTZ2

Each designed sequence is predicted with Boltz2 [13] to evaluate whether the intended fold and catalytic geometry are maintained. The prediction includes the protein chain, retinal (as covalent cofactor), and substrate. This is the second filtering stage. Ahern et al. rank candidates by predicted confidence; we evaluate on two criteria specific to the rhodozyme constraint.

The first criterion is theozyme alignment. We superimpose each predicted structure onto the placement from Step 3, aligning only the three theozyme residues (all atoms, not just Cα), and measure the RMSD. This tests whether the catalytic geometry survived the design-predict cycle. Across all 307 predictions, we rank candidates by this theozyme all-atom RMSD. The second criterion is pLDDT of the Boltz2 prediction, reported per-residue. We examine pLDDT separately for the locked regions (which should score high, as they reproduce known structure) and the designed regions (where low confidence indicates the model is uncertain about the backbone).

The comparison between predicted and reference theozyme geometry requires more than a rigid-body superposition. The Cα positions may align well while the sidechain rotamers differ—a common outcome when Boltz2 resolves the local environment differently from the reference. We therefore allow sidechain rotation around the Cα–Cβ bond axis when assessing the match, comparing the functional-atom distances (Ser OG – His Nε2, His Nδ1 – Asp Oδ) rather than insisting on identical χ_1 angles.

An unexpected observation emerged from inspecting the validated designs. Because the theozyme was placed flatly on the ICL3 surface, the designed loops could fold in two ways: with the substrate-binding face directed inward (into the helical bundle interior, as originally intended) or outward (toward the cytosolic side of ICL3). The top-ranking prediction adopted the outward orientation—the catalytic face and substrate-binding site point toward the cytoplasm rather than into the transmembrane pocket. The theozyme geometry itself is preserved in both orientations; what differs is where the substrate approaches. The outward orientation may be more favorable, since substrate access is not occluded by the surrounding helices. Crucially, the light-gating mechanism is unaffected: the triad residues sit on positions that are only in register when TM5/TM6 adopt the active conformation, regardless of which face the binding site presents.

We selected the top candidate by balancing theozyme RMSD ($\sim$2.4 Å all-atom, $\sim$0.5 Å Cα-only, rank 6 of 307) against the highest global pLDDT in the top 10 ($\sim$92). The overall backbone RMSD to the parent rhodopsin is $\sim$0.8 Å, confirming that the fold is preserved. The pLDDT breakdown shows high confidence in the locked TM core ($\sim$95 mean) and good confidence in the designed loops ($\sim$86 mean), with the theozyme residues at $\sim$72—lower, as expected for residues at a designed interface.

The catalytic geometry in the predicted structure shows the Ser–His distance at 4.03 Å and the His–Asp distance at 3.18 Å, compared to 3.04 Å and 2.75 Å in the reference placement. These distances are longer than the ideal hydrogen-bond geometry and would require further optimization—through additional design iterations, molecular dynamics relaxation, or experimental directed evolution—to achieve catalytically competent contacts.

Because the retinal binding pocket is preserved by the mask, the Schiff base linkage to Lys-296 is intact in the predicted structure. LAMBDA can predict the spectral properties of retained candidates—because the binding pocket is unchanged, the rhodozyme is expected to absorb at the same wavelength as the parent rhodopsin.

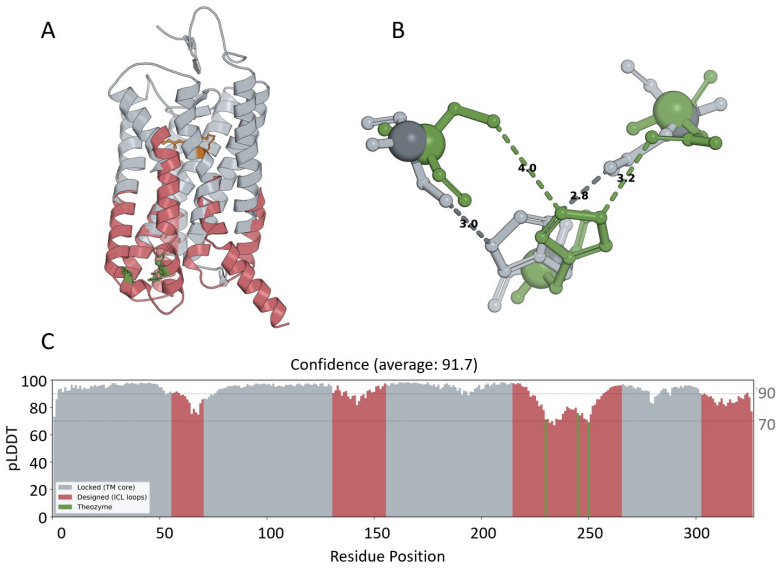


FIGURE 5.6: Boltz2 evaluation of the top candidate. (A) Predicted structure overlaid on the parent rhodopsin. Gray: reference TM core (locked regions from 3PQR). Terracotta: designed loop regions (Boltz2 prediction). Retinal and Schiff base Lys-296 in rust. Overall backbone RMSD: 0.84 Å. (B) Catalytic geometry comparison: predicted theozyme (green) overlaid on reference placement (gray), with catalytic interaction distances shown. Reference: Ser–His 3.04 Å, His–Asp 2.75 Å. Predicted: Ser–His 4.03 Å, His–Asp 3.18 Å. (C) Per-residue pLDDT confidence. Locked TM core: 94.9 mean. Designed loops: 85.6 mean. Theozyme residues: 72.1 mean. Global: 91.7.

We consider the computational design a success: Boltz2 predicts a well-folded structure (pLDDT 91.7) that preserves the rhodopsin fold (backbone RMSD 0.84 Å) and maintains the theozyme geometry within optimizable range. The most difficult step in the workflow is not the AI-driven design or validation, but the original theozyme placement. This is a combinatorial and structural biology problem that precedes the Baker lab protocol entirely. Each placement generates a full cascade of 50 backbone designs $\times$ 8 sequences $\times$ structure predictions—for two placements, this already produces over 600 candidates to evaluate. Selecting many placements without careful geometric and structural reasoning would produce a vast screening space that is computationally expensive and difficult to interpret. The placement decision requires intuition about protein geometry, knowledge of the rhodopsin conformational cycle, and judgment about which helix positions can support catalytic function. The same kind of expert judgment applies at the end of the pipeline: interpreting predicted structures, assessing whether catalytic distances are close enough, and deciding which candidates merit experimental follow-up. The AI models automate the generative steps; the structural biology reasoning that frames and interprets them remains human.

5.7 INTEGRATION

Every step in this workflow—from structure fetching through model submission to result registration—runs through ProtOS. The processors handle annotation, geometric matching, and mask construction; the Model Manager dispatches and tracks all RFdiffusion, LigandMPNN, and Boltz2 jobs, registering each output as an entity available to downstream evaluation.

RFdiffusion2, LigandMPNN, and Boltz2 are published tools; the Baker lab's protocol for combining them is established. ProtOS reproduces that protocol within a managed data framework, applied to a different and more constrained problem. When a candidate fails at validation, we can trace back to its placement, its backbone, its sequence, and ask why.

No experimental validation of the rhodozyme concept exists at this time. The designs shown here are ongoing work. The contribution is not a validated enzyme but a demonstration of what ProtOS can do: integrate multiple structure-generation models into a single pipeline with consistent data management, enabling the large screens from which candidates emerge. We show one such candidate. The rhodozyme concept itself remains speculative; the ability to explore it at scale, using the Baker lab's protocol within a managed data framework, is what this chapter highlights.

The rhodopsin scaffold offers one further possibility. Because the retinal binding pocket is separate from the intracellular catalytic face, different rhodopsins—with different absorption maxima—could carry different enzymes. In principle, a trypsin-rhodozyme built on a 500 nm rhodopsin would absorb at 500 nm; a papain-rhodozyme on a 550 nm scaffold would absorb at 550 nm. Different wavelengths could activate different enzymes, and sequential catalytic steps on the same substrate could become addressable by color. LAMBDA can predict absorption for any of these variants, because the binding pocket graph representation is independent of the intracellular design.

6

PROTOS-MCP — TALK TO YOUR PROTEINS

During my PhD, AI went from barely functional to a daily coding companion. Coding is about solving problems, not typing—but the barrier to entry remains real. ProtOS is powerful, yet the future lies in making advanced computational frameworks accessible to any structural biologist or bioinformatician. MCP is where lab scientists and computational researchers meet.

ProtOS requires Python, which limits access for researchers who cannot code. An experimentalist who asks "which residues contact retinal in this structure?" understands what they want; they cannot get it without a programmer. ProtOS-MCP removes that barrier by exposing the same processors through natural language: a researcher describes what they want, and the system executes it. The contribution is not the natural language interface itself—large language models that call tools are not new—but the domain-specific infrastructure behind it.

The Model Context Protocol (MCP) [58], developed by Anthropic, standardizes how language models discover and invoke external tools through typed schemas. A language model generates text; it does not load structures, align sequences, or predict absorption spectra. MCP provides a structured interface through which the model can perform these operations by selecting the appropriate tool, populating its parameters, and interpreting the result.

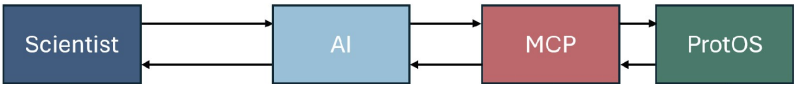


FIGURE 6.1: ProtOS-MCP architecture. A natural language question is routed through MCP tool calls to the local ProtOS server, which returns structured JSON. The language model synthesizes results into a response. All protein data remains local.

Figure 6.1 shows the ProtOS-MCP architecture. The researcher types a question in natural language. The language model interprets it and issues one or more MCP tool calls to the local ProtOS server. The server executes each call against the ProtOS installation and returns structured JSON. The model synthesizes the results into a natural language response. Critically, ProtOS-MCP runs locally—no protein data leaves the researcher's machine; only the conversational text passes through external servers.

To demonstrate the system, the rhodozyme sequence from chapter 5 served as a starting point. Its retinal-binding domain is preserved from the parent rhodopsin, so spectral tuning through binding pocket mutations remains possible without disrupting the designed catalytic site. The question posed through conversation: can the rhodozyme's predicted absorption be shifted toward longer wavelengths by transplanting binding pocket residues from the human long-wave sensitive opsin?

The complete conversation (section A.1) proceeded in four turns, involving 20 tool calls across five processors. No code was written. The model first reviewed its capabilities, then ingested the rhodozyme FASTA file, predicted its absorption at 496 nm for 11-*cis* retinal using LAMBDA, and finally executed a multi-step redshift engineering workflow: comparing binding pockets at GRN positions, screening 20 single-point mutants, and performing a greedy cumulative walk that peaked at seven mutations with a predicted shift to 514 nm. The model autonomously decomposed the complex fourth prompt into 11 sequential tool calls with correct dependencies—GRN annotation before pocket comparison before mutant design—without further prompting.

The value is not the resulting design, which awaits experimental validation. Four conversational turns compressed a multi-step spectral engineering analysis that would otherwise require detailed knowledge of the ProtOS API. The model also applied domain knowledge—such as identifying the protein family from the sequence—that was not programmed into the MCP tools. The limitations are equally clear: the workflow depends on the language model reasoning correctly about tool selection and parameter values, and a model that misunderstands one step propagates errors through the rest. ProtOS-MCP does not replace programming—analyses requiring custom logic or operations not exposed as tools still require code—but it lowers the barrier for the common case: a researcher with a biological question and data that fits the processor workflows.

7

DISCUSSION

The goal of this thesis was a unified model that predicts opsin spectral properties across protein folds. Building it required solving two problems along the way: microbial rhodopsins lacked standardized positional annotation, and no existing framework linked the sequences, structures, and measurements that the model would need at scale.

7.1 CONTRIBUTIONS

LAMBDA is the central contribution. It predicts spectral properties across both opsin superfamilies—the first model to treat color tuning as a problem of binding pocket geometry rather than family-specific sequence features. Trained on 2,120 sequences spanning type I, type II, and the lipocalin fold hCRBPII, LAMBDA achieves 5.18 nm mean absolute error on type II opsins and 6.86 nm on type I. The Opsin Atlas extends these predictions to 47,700 sequences.

LAMBDA works because of MOGRN. Type I and type II opsins diverged billions of years ago and share no detectable sequence similarity outside the binding pocket. A model that takes raw sequence as input cannot learn from both. MOGRN provides the shared vocabulary: a structure-based numbering system for microbial rhodopsins anchored to the Schiff base lysine, analogous to what Ballesteros–Weinstein [40] provides for GPCRs. With both families mapped to standardized coordinates, LAMBDA can represent any opsin binding pocket as a graph aligned on retinal, regardless of fold.

ProtOS made this feasible at scale. Routing 47,700 sequences through the full analysis pipeline requires managing data across formats and processors without manual intervention. The Opsin Atlas exists because ProtOS could execute this pipeline for every sequence without custom scripting per step.

ProtOS-MCP completes the stack by removing the programming requirement. The spectral engineering workflow demonstrated in chapter 6 ran through four conversational turns with no code written.

The deeper principle is that standardized coordinates enable cross-family learning. This is not specific to opsins. Any protein superfamily where a conserved functional site spans divergent folds could benefit from the same approach: define a coordinate system anchored to the functional site, represent the site as a graph, and train a model across families that would otherwise be treated separately.

7.2 LIMITATIONS

LAMBDA has not been experimentally validated. The Opsin Atlas contains 47,700 predictions, but predictions are not measurements. Model accuracy was evaluated against published spectral data, not against prospective characterization of novel sequences. Until atlas candidates are tested in the lab, LAMBDA remains a computational screening tool. Additionally, accuracy

degrades for deeply red-shifted opsins where training data is sparse—the physics of spectral tuning at long wavelengths remains difficult for any sequence-based approach.

MOGRN depends on a conserved structural anchor: the retinal-binding pocket. Microbial rhodopsins share this feature despite extensive sequence divergence. Other protein families may lack such an obvious invariant. The GRN Processor architecture accommodates new numbering systems, but defining them—identifying the right anchor, validating across structures, ensuring coverage of functional diversity—remains manual, expert-driven work.

ProtOS has technical boundaries. Database coverage includes UniProt, PDB, AlphaFold DB, and NCBI; additional sources require new loaders. GRN annotation requires a numbering system for the family of interest—GPCRs have Ballesteros–Weinstein, microbial rhodopsins now have MOGRN, but other families cannot use the GRN Processor until such systems are developed.

ProtOS-MCP depends on language model reasoning. For single-tool operations—loading a structure, running a prediction—it works reliably. For complex multi-step workflows, the model must chain tool calls with dependencies: annotations must complete before positions can be queried, which must complete before mutants can be designed. Errors in one step propagate. The workflow dataset in the Supplementary Materials provides a benchmark for measuring reliability, but it does not guarantee it.

More broadly, tools reduce barriers but do not replace judgment. A researcher using ProtOS-MCP can run analyses they could not run before. Interpreting the results still requires understanding the biology.

7.3 FUTURE WORK

Experimental validation of LAMBDA is the most impactful next step. Characterizing even a small subset of Opsin Atlas candidates—selected for maximal predicted spectral diversity—would test whether the model captures real spectral tuning physics or has learned correlations that do not generalize to uncharacterized sequences. Positive results would transform LAMBDA from a screening tool into a framework for understanding color tuning. Negative results would identify where the model fails and guide retraining.

For optogenetics, the atlas identifies specific design targets: red-shifted opsins for deeper tissue penetration, spectrally separated pairs for multiplexed control, and sequences with large dark-to-activated spectral separation for bistable tools. The rhodozyme concept—coupling light activation to enzymatic function—extends this further. Color-tuned rhodozymes, each responding to a different wavelength, would enable sequential catalysis triggered by different colors of light. These remain computational designs, but the infrastructure to iterate on them exists.

ProtOS is positioned to integrate the next generation of protein AI models as they arrive. RFdiffusion2 [53], Boltz2 [13], and LigandMPNN [57] are already accessible through the Model Manager. As these models mature and new ones emerge, the processor architecture and MCP interface remain stable—new capabilities become available through conversation without requiring users to learn new APIs.

7.4 CONCLUSION

This thesis started with a question about spectral tuning and ended with infrastructure for asking questions about proteins. The path was not planned. MOGRN was built because LAMBDA needed a shared coordinate system. ProtOS was built because LAMBDA needed to process 47,700 sequences without drowning in glue code. ProtOS-MCP was built because I got tired of writing the same data loading and annotation scripts for every new analysis—and because the same interface that saved me time also makes these tools accessible to experimentalists who have never written code.

MOGRN, LAMBDA, and ProtOS each address one of the gaps framed in the introduction. The underlying principle—that standardized structural coordinates enable learning across evolutionary divides—is the scientific contribution. The practical contribution is that a researcher can now go from a protein sequence to a spectral prediction through a conversation.

What remains is validation. The Opsin Atlas is a collection of hypotheses. Testing them is the next step. The same is true for protein design—the rhodozyme remains my favorite project in this thesis, and seeing those candidates tested in the lab is what I look forward to most.

The code for all projects presented in this thesis is publicly available on GitHub: LAMBDA, MOGRN, ProtOS, and ProtOS-MCP.

A

SUPPLEMENTARY MATERIALS

A.1 PROTOS-MCP CONVERSATION LOG

The complete ProtOS-MCP conversation used to redshift the rhodozyme (chapter 6) is reproduced on the following pages. The session proceeded in four turns involving 20 tool calls across five processors. No code was written.

(a) Turn 1: orientation and capability review

User

Review and list ProtOS' functionalities.

ProtOS Agent

Let me pull up the ProtOS guide to give you a comprehensive overview.

⚙ `protos_guide`
`{"topic": "overview"}`

⚙ `protos_guide`
`{"topic": "processors"}`

Here's a comprehensive review of ProtOS — a structural biology framework for managing and analyzing protein data.

Core Design Principles

ProtOS is built around zero-configuration data management with human-readable entity names (e.g., `1ubq`, `EGFR_HUMAN`). All file paths are managed internally — you only work with names. A modular processor architecture handles all data operations, and datasets organize entity collections for batch processing.

Processor Modules & Functionalities

1. Structure Processor (3D structures — PDB/mmCIF)

- Download & load structures from PDB and other sources
- Chain sequence extraction from single structures or entire datasets
- Structural alignment — CEalign, superimposition, RMSD matrices
- Binding site analysis — residue identification, pocket volume estimation, ligand interaction profiling
- Water network analysis — water-mediated residue contacts and hydrogen-bond networks
- Ligand manipulation — extract, remove, and merge structures
- GRN annotation — map Generic Residue Numbering onto structure residues
- Graph generation — atom- or residue-level contact graphs

2. Sequence Processor

- Download & register sequences from UniProt, FASTA files, or inline records
- Pairwise & multiple alignment — BLOSUM62 and MMseqs2 methods
- Conservation analysis — per-position entropy and conservation scores
- Mutant library generation — combinatorial mutant creation from a base sequence
- Sequence clustering by similarity threshold
- GRN annotation of sequences

3. Embedding Processor

- Multiple embedding models supported (ESM, ANKH, etc.)
- Generate embeddings for sequence datasets
- Cosine similarity analysis between embeddings

4. GRN Processor (Generic Residue Numbering)

- Reference tables — load curated references (e.g., GPCRdb)
- GRN assignment — align sequences to references and assign numbered positions
- Conservation comparison between entity groups at GRN positions

5. Property Processor (tabular annotations)

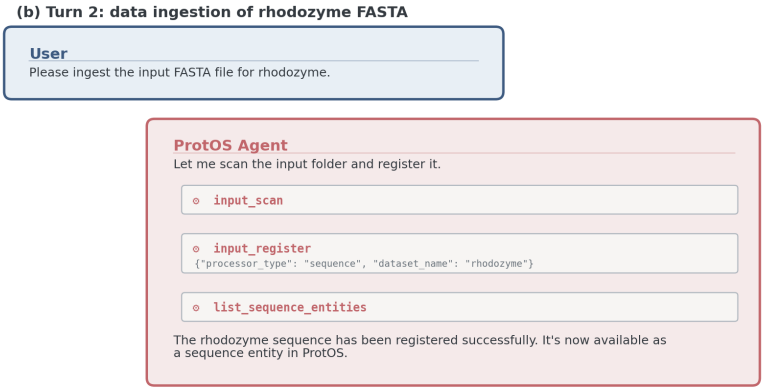
- Create, load, and export property tables
- Statistics on property distributions
- Filter entities by property values

Cross-Cutting Capabilities

- Dataset management — create, copy, merge, update, delete datasets across all processor types
- Entity registry — unified tracking of entities across formats
- Model integration — Lambda workflow (dataset → GRN → prediction), Boltz mutation screening

In short, ProtOS is a full-stack toolkit for structural biology that spans from data acquisition (PDB, UniProt, ChEMBL) through sequence/structure/ligand analysis to machine learning model integration — all managed through a consistent entity-and-dataset abstraction.

(a) Turn 1: orientation and capability review.



(b) Turn 2: data ingestion of rhodozyme FASTA.

(c) Turn 3: spectral prediction (λmax = 496 nm)

User

Now that it is registered, I want to use Lambda. Review how Lambda works, then run it for this type II opsin.

ProtOS Agent

Let me review how Lambda works by checking the model details and loading the dataset.

⚙ `describe_model`
```
{"name": "lambda"}
```

⚙ `guide_workflow_example`
```
{"workflow_type": "grn_assignment"}
```

⚙ `sequence_load_dataset`
```
{"dataset_name": "rhodozyme", "include_sequences": true}
```

The `rhodozyme` dataset contains one sequence — `rhodozyme_v1.0` (326 aa). The sequence starts with MNGTEGP..., which is clearly a type II opsin (rhodopsin family / GPCR Class A).

How Lambda Works

Lambda is ProtOS' graph-based property predictor for opsins. It predicts λmax (the peak absorption wavelength) from sequence alone. Here's what happens under the hood when you call `model_lambda_run`:

1. GRN Assignment — The sequence is aligned to a GPCR reference table (e.g., `gpcrdb_ref`) and annotated with Generic Residue Numbering (Ballesteros-Weinstein positions like 3.50, 6.48, etc.). This maps each residue to a structurally equivalent position across all Class A GPCRs.

2. Embedding Generation — A protein language model (e.g., Ankh) generates per-residue embeddings that capture evolutionary and biophysical context.

3. Graph Construction & Prediction — Lambda uses the GRN-labeled residues (especially binding pocket positions) to build a graph representation and predict λmax using a trained model.

Key parameters for `model_lambda_run`:
- `protein_family`: `"gpcr_a"` — since type II opsins are Class A GPCRs
- `sequence_dataset`: `"rhodozyme"` — your registered dataset
- `reference_table`: defaults to `"gpcrdb_ref"` — the GPCRdb numbering reference

The output is a property table with the predicted λmax for each sequence.

Let me run it now.

⚙ `model_lambda_run`
```
{"protein_family": "gpcr_a", "reference_table": "gpcrdb_ref",
"sequence_dataset": "rhodozyme"}
```

Lambda ran successfully on rhodozyme. Here are the predictions:

Rhodozyme v1.0 — Lambda Predictions

Property	Predicted λmax
11-cis retinal	496 nm
All-trans retinal	526 nm
UV (deprotonated)	367 nm

The pipeline completed all steps automatically:

- GRN Assignment: Aligned to human rhodopsin as the closest reference (normalized score 2.39, 73% coverage, 231 positions assigned). Saved as table `rhodozyme_grn`.
- Embeddings: Generated Ankh-large per-residue embeddings.
- Prediction: Stored in property table `rhodozyme_lambda_007061`.

The ~496 nm prediction for 11-cis retinal puts rhodozyme squarely in the blue-green range — consistent with a typical rhodopsin-like type II opsin. The 526 nm all-trans prediction reflects the expected redshift upon photoisomerization.

What would you like to do next — explore the GRN annotation, run mutations, or compare against other opsins?

(c) Turn 3: spectral prediction ($\lambda_{\max} = 496\,\text{nm}$).

(d) Turn 4: rational redshift engineering

User

I need you to take the LWS opsin of human. We want to adapt the binding pocket of rhodozyme to be redshifted. Pick up to 20 positions that vary and lead to color tuning at those positions, and create a mutational screen (one mutation at a time) until rhodozyme achieves the same (~550 nm) spectral absorption.

This is complex — carefully consider the steps and tools needed.

ProtOS Agent

Great approach. Here's the plan:

1. Get the human LWS opsin and annotate it with GRN
2. Load the binding domain config to identify chromophore pocket positions
3. Compare rhodozyme vs LWS at those positions to find divergent sites
4. Build a single-mutant library — one substitution at a time toward the LWS residue
5. Run Lambda on all mutants, then do a greedy additive walk to trace the spectral path from WT toward the full LWS-like pocket

Let me start.

⚙ `entity_search_entities`
{"query": "OPN1LW", "processor_types": ["sequence"]}

⚙ `load_sequence`
{"sequence_id": "human_NM_020061_LWS_OPN1LW_CDS",
"include_sequence": true}

⚙ `sequence_annotate_with_grn`
{"entity_names": ["human_NM_020061_LWS_OPN1LW_CDS"],
"protein_family": "gpcr_a", "reference_table": "gpcrdb_ref",
"output_table": "human_lws_grn"}

⚙ `grn_get_config`
{"protein_family": "gpcr_a"}

⚙ `grn_query_entity`
{"entity_id": "rhodozyme_v1.0", "grn_table": "rhodozyme_grn",
"positions": ["2.50", "2.53", "...30 binding pocket positions...",
"7.42", "7.43"]}

⚙ `grn_query_entity`
{"entity_id": "human_NM_020061_LWS_OPN1LW_CDS", "grn_table":
"human_lws_grn", "positions": ["2.50", "2.53", "...30 binding pocket
positions...", "7.42", "7.43"]}

Found 20 differing pocket positions between rhodozyme and human LWS. Registering 21 sequences (WT + 20 single mutants, each swapping one rhodozyme residue to the LWS residue) and running Lambda.

⚙ `sequence_register_records`
{"dataset_name": "rhodozyme_lws_single_mutants", "records":
[{"rhodozyme_WT + 20 individual single mutants (M86E, G90I, F91A,
A117V, T118S, E122I, V129A, F203Y, Y206V, M207L, V210T, F212C,
E257M, S260A, F261Y, A269T, F273C, F276A, F293Y, T297S)"}]}

⚙ `model_lambda_run`
{"protein_family": "gpcr_a", "reference_table": "gpcrdb_ref",
"sequence_dataset": "rhodozyme_lws_single_mutants"}

Single-mutant screen complete. The top redshifters are F261Y (+7.0 nm) and A269T (+6.9 nm), both on TM6. Five more mutations contribute modest redshifts (+0.1 to +1.5 nm), while the bottom five — including E122I and M86E — actually blueshift by 3–8 nm.

Now building the greedy cumulative walk: adding mutations one at a time in order of individual impact, running Lambda at each step to trace the spectral trajectory.

⚙ `sequence_register_records`
{"dataset_name": "rhodozyme_cumulative_walk", "records": ["21
cumulative sequences: step00_WT through step20_M86E (each adds the
next-ranked mutation)"]}

⚙ `model_lambda_run`
{"protein_family": "gpcr_a", "reference_table": "gpcrdb_ref",
"sequence_dataset": "rhodozyme_cumulative_walk"}

The greedy walk peaks at step 7 with 514.1 nm (+17.8 nm from WT), then erodes back below WT by step 19. Exporting the optimal 7-mutation variant as FASTA.

⚙ `sequence_export_entity`
{"sequence_id": "step07_F273C_grn6.56", "export_name":
"rhodozyme_LWS_optimized_7mut"}

Best design: 7 mutations → 514.09 nm (step07_F273C_grn6.56 in the walk). Not all 20 binding-pocket substitutions helped — adding the remaining 13 actually *eroded* the redshift, dragging λmax back below the WT baseline by step 19.

(d) Turn 4: rational redshift engineering through binding pocket comparison, single-mutant screening, and greedy cumulative walk (peak: 514 nm at 7 mutations).

A.2 WORKFLOW BENCHMARK SPECIFICATIONS

The ProtOS-MCP workflow collection is stratified by difficulty based on the number of tools required, the number of processors involved, and the complexity of result interpretation.

Level	Tools	Processors	Example
Beginner	1–2	1	Load structure, list ligands
Intermediate	3–5	2–3	Align structures, annotate with GRN
Advanced	5+	3+	Multi-stage functional analysis

TABLE A.1: Workflow benchmark difficulty criteria.

Each benchmark workflow is evaluated on: tool selection accuracy, parameter accuracy, execution order, result interpretation, and biological accuracy.

A.3 PROCESSOR DATA STORAGE CONVENTIONS

ProtOS organizes data in a standardized directory structure. Table A.3 describes the storage locations for each processor.

Processor	Base Directory	File Formats
Structure	`structure/`	`.cif`, `.pkl`, `.json`
Sequence	`sequence/`	`.fasta`, `.aln`, `.mmseqs`
GRN	`grn/`	`.yaml`, `.parquet`
Embedding	`embedding/`	`.npy`, `.h5`
Property	`property/`	`.parquet`, `.csv`
Molecule	`molecule/`	`.sdf`, `.mol2`, `.json`

TABLE A.3: Processor data storage conventions.

GLOSSARY

SYMBOLS

λ_{max}	absorption maximum (wavelength of peak absorption)
$\Delta\lambda$	spectral shift ($\lambda_{max}^{AT} - \lambda_{max}^{11\text{-cis}}$)
E	photon energy ($E = hc/\lambda$)
R^2	coefficient of determination
h	Planck constant
c	speed of light in vacuum

OPSIN FAMILIES AND SUBFAMILIES

bR	bacteriorhodopsin
GtACR1	*Guillardia theta* anion channelrhodopsin 1
HeR	heliorhodopsin
LWS	long-wavelength-sensitive (cone opsin)
MWS	medium-wavelength-sensitive (cone opsin)
OPN3	encephalopsin
OPN4	melanopsin
OPN5	neuropsin
PR	proteorhodopsin
RGR	retinal G protein-coupled receptor (non-visual opsin)
SRI, SRII	sensory rhodopsin I, sensory rhodopsin II
SWS1	short-wavelength-sensitive type 1 (cone opsin)
TMT	teleost multiple tissue opsin
VA	vertebrate ancient (opsin)

PROTEINS AND MODELS

Ankh	protein language model (Ankh-large, 1536-dimensional embeddings)
Boltz	biomolecular structure prediction model
ESM-2	Evolutionary Scale Modeling protein language model
hCRBPII	human cellular retinol binding protein II
LigandMPNN	ligand-aware message passing neural network for sequence design
RFdiffusion	denoising diffusion model for protein backbone design

STRUCTURAL ANNOTATION

BW	Ballesteros–Weinstein (GPCR residue numbering)
GRN	Generic Residue Numbering
ICL	intracellular loop (ICL1–ICL3)

MOGRN	Microbial Opsin Generic Residue Numbering
TM	transmembrane helix (TM1–TM7)

METHODS AND METRICS

BLAST	Basic Local Alignment Search Tool
GCN	Graph Convolutional Network
MAE	mean absolute error
MMseqs2	Many-against-Many sequence searching
MSA	multiple sequence alignment
pLDDT	predicted Local Distance Difference Test
pLM	protein language model
RMSD	root-mean-square deviation
UMAP	Uniform Manifold Approximation and Projection

DATABASES

AlphaFold DB	AlphaFold Protein Structure Database
GPCRdb	G protein-coupled receptor database
NCBI nr	National Center for Biotechnology Information non-redundant database
PDB	Protein Data Bank
UniProt	Universal Protein Knowledgebase
VPOD	Visual Physiology Opsin Database

THIS THESIS

GPCR	G protein-coupled receptor
LAMBDA	Light Absorption Modeling through Binding Domain Analysis
MCP	Model Context Protocol
ProtOS	Protein Operating System

BIBLIOGRAPHY

1. Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **181**, 662 (1958).

2. Henderson, R. & Unwin, P. N. T. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* **257**, 28 (1975).

3. Schertler, G. F. X., Villa, C. & Henderson, R. Projection structure of rhodopsin. *Nature* **362**, 770 (1993).

4. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Le Trong, I., Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M. & Miyano, M. Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science* **289**, 739 (2000).

5. Rasmussen, S. G. F., Choi, H.-J., Rosenbaum, D. M., Kobilka, T. S., Thian, F. S., Edwards, P. C., Burghammer, M., Ratnala, V. R. P., Sanishvili, R., Fischetti, R. F., Schertler, G. F. X., Weis, W. I. & Kobilka, B. K. Crystal structure of the human β_2 adrenergic G-protein-coupled receptor. *Nature* **450**, 383 (2007).

6. Kühlbrandt, W. The resolution revolution. *Science* **343**, 1443 (2014).

7. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **28**, 235 (2000).

8. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523 (2023).

9. Béjà, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P., Jovanovich, S. B., Gates, C. M., Feldman, R. A., Spudich, J. L., Spudich, E. N. & DeLong, E. F. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**, 1902 (2000).

10. Ernst, O. P., Lodowski, D. T., Elstner, M., Hegemann, P., Brown, L. S. & Kandori, H. Microbial and Animal Rhodopsins: Structures, Functions, and Molecular Mechanisms. *Chemical Reviews* **114**, 126 (2014).

11. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583 (2021).

12. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. & Velankar, S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439 (2022).

13. Wohlwend, J., Corso, G., Passaro, S., Getz, N., Reveiz, M., Leidal, K., Swiderski, W., Atkinson, L., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T. & Barzilay, R. Boltz-1: Democratizing Biomolecular Interaction Modeling. *bioRxiv*. Preprint (2024).

14. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. & Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123 (2023).

15. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J. & Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

16. Elnaggar, A., Essam, H., Salah-Eldin, W., Moustafa, W., Elkerdawy, M., Rochereau, C. & Rost, B. Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. *bioRxiv*. arXiv:2301.06568 (2023).

17. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D. & Rost, B. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 7112 (2022).

18. Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I. & Overington, J. P. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery* **16**, 19 (2017).

19. Koyanagi, M. & Terakita, A. Diversity of animal opsin-based pigments and their optogenetic potential. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1837**, 710 (2014).

20. Oesterhelt, D. & Stoeckenius, W. Rhodopsin-like protein from the purple membrane of Halobacterium halobium. *Nature New Biology* **233**, 149 (1971).

21. Nagel, G., Szellas, T., Huhn, W., Kateriya, S., Adeishvili, N., Berthold, P., Ollig, D., Hegemann, P. & Bamberg, E. Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *Proceedings of the National Academy of Sciences* **100**, 13940 (2003).

22. Deisseroth, K. Optogenetics: 10 years of microbial opsins in neuroscience. *Nature Neuroscience* **18**, 1213 (2015).

23. Shichida, Y. & Matsuyama, T. Evolution of opsins and phototransduction. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**, 2881 (2009).

24. Palczewski, K. G Protein-Coupled Receptor Rhodopsin. *Annual Review of Biochemistry* **75**, 743 (2006).

25. Nathans, J. Molecular biology of visual pigments. *Annual Review of Neuroscience* **10**, 163 (1987).

26. Luecke, H., Schobert, B., Richter, H.-T., Cartailler, J.-P. & Lanyi, J. K. Structure of bacteriorhodopsin at 1.55 Å resolution. *Journal of Molecular Biology* **291**, 899 (1999).

27. Honig, B., Dinur, U., Nakanishi, K., Balogh-Nair, V., Gawinowicz, M. A., Arnaboldi, M. & Motto, M. G. An external point-charge model for wavelength regulation in visual pigments. *Journal of the American Chemical Society* **101**, 7084 (1979).

28. Yokoyama, S. Evolution of Dim-Light and Color Vision Pigments. *Annual Review of Genomics and Human Genetics* **9**, 259 (2008).

29. Mogi, T., Stern, L. J., Marti, T., Chao, B. H. & Khorana, H. G. Aspartic acid substitutions affect proton translocation by bacteriorhodopsin. *Proceedings of the National Academy of Sciences* **85**, 4148 (1988).

30. Frazer, S. A., Baghbanzadeh, M., Rahnavard, A., Crandall, K. A. & Oakley, T. H. Discovering genotype–phenotype relationships with machine learning and the Visual Physiology Opsin Database (VPOD). *GigaScience* **13**, giae073 (2024).

31. Inoue, K., Tsunoda, S. P., Singh, M., Tomida, S., Hososhima, S., Konno, M., Nakamura, R., Watanabe, H., Bulzu, P.-A., Banciu, H. L., Andrei, A.-Ş., Uchihashi, T., Ghai, R., Béjà, O. & Kandori, H. Schizorhodopsins: A family of rhodopsins from Asgard archaea that function as light-driven inward H^+ pumps. *Science Advances* **6**, eaaz2441 (2020).

32. Sela, M., Church, J. R., Schapiro, I. & Schneidman-Duhovny, D. RhoMax: Computational Prediction of Rhodopsin Absorption Maxima Using Geometric Deep Learning. *Journal of Chemical Information and Modeling* **64**, 4630 (2024).

33. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403 (1990).

34. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026 (2017).

35. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422 (2009).

36. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **32**, 922 (1976).

37. Okada, T., Sugihara, M., Bondar, A.-N., Elstner, M., Entel, P. & Buss, V. The Retinal Conformation and its Environment in Rhodopsin in Light of a New 2.2 Å Crystal Structure. *Journal of Molecular Biology* **342**, 571 (2004).

38. Kandori, H. Retinal proteins: photochemistry and optogenetics. *Bulletin of the Chemical Society of Japan* **93**, 76 (2020).

39. Fey, M. & Lenssen, J. E. *Fast Graph Representation Learning with PyTorch Geometric* in *ICLR Workshop on Representation Learning on Graphs and Manifolds* arXiv:1903.02428 (2019).

40. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods in Neurosciences* **25**, 366 (1995).

41. Kooistra, A. J., Mordalski, S., Pándy-Szekeres, G., Esguerra, M., Mamyrbekov, A., Munk, C., Keserű, G. M. & Gloriam, D. E. GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Research* **49**, D335 (2021).

42. Venkatakrishnan, A. J., Deupi, X., Lebon, G., Tate, C. G., Schertler, G. F. & Babu, M. M. Molecular signatures of G-protein-coupled receptors. *Nature* **494**, 185 (2013).

43. Katritch, V., Cherezov, V. & Stevens, R. C. Structure-Function of the G Protein-Coupled Receptor Superfamily. *Annual Review of Pharmacology and Toxicology* **53**, 531 (2013).

44. Hidber, F. S. *et al.* A generic residue numbering system for microbial rhodopsins—Unifying structural frameworks and functional mapping. *npj Structural Biology*. Under review at npj Structural Biology (Springer Nature) (2026).

45. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*. arXiv:1802.03426 (2018).

46. De Grip, W. J. & Ganapathy, S. Rhodopsins: An Excitingly Versatile Protein Species for Research, Development and Creative Engineering. *Frontiers in Chemistry* **10**, 879609 (2022).

47. Kato, H. E., Inoue, K., Kandori, H. & Nureki, O. The light-driven sodium ion pump: a new player in rhodopsin research. *BioEssays* **38**, 1274 (2016).

48. Berndt, A., Yizhar, O., Gunaydin, L. A., Hegemann, P. & Deisseroth, K. Bi-stable neural state switches. *Nature Neuroscience* **12**, 229 (2009).

49. Wang, W., Nossoni, Z., Berbasova, T., Watson, C. T., Yapici, I., Lee, K. S. S., Vasileiou, C., Geiger, J. H. & Borhan, B. Tuning the Electronic Absorption of Protein-Embedded All-*trans*-Retinal. *Science* **338**, 1340 (2012).

50. Karasuyama, M., Inoue, K., Nakamura, R., Kandori, H. & Takeuchi, I. Understanding Colour Tuning Rules and Predicting Absorption Wavelengths of Microbial Rhodopsins by Data-Driven Machine-Learning Approach. *Scientific Reports* **8**, 15580 (2018).

51. Nossoni, Z., Assar, Z., Yapici, I., Nosrati, M., Wang, W., Berbasova, T., Vasileiou, C., Borhan, B. & Geiger, J. H. Structures of holo wild-type human cellular retinol-binding protein II (hCRBPII) bound to retinol and retinal. *Acta Crystallographica Section D: Biological Crystallography* **70**, 3226 (2014).

52. Kipf, T. N. & Welling, M. *Semi-Supervised Classification with Graph Convolutional Networks* in *International Conference on Learning Representations (ICLR)* arXiv:1609.02907 (2017).

53. Ahern, W., Yim, J., Tischer, D., Salike, S., Woodbury, S. M., Kim, D., Kalvet, I., Kipnis, Y., Coventry, B., Altae-Tran, H. R., Bauer, M. S., Barzilay, R., Jaakkola, T. S., Krishna, R. & Baker, D. Atom level enzyme active site scaffolding using RFdiffusion2. *bioRxiv*. Preprint (2025).

54. Choe, H.-W., Kim, Y. J., Park, J. H., Morizumi, T., Pai, E. F., Krauss, N., Hofmann, K. P., Scheerer, P. & Ernst, O. P. Crystal structure of metarhodopsin II. *Nature* **471**, 651 (2011).

55. Radisky, E. S., Lee, J. M., Lu, C.-J. K. & Koshland Jr., D. E. Insights into the serine protease mechanism from atomic resolution structures of trypsin reaction intermediates. *Proceedings of the National Academy of Sciences* **103**, 6835 (2006).

56. Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M. & Baker, D. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089 (2023).

57. Dauparas, J., Lee, G. R., Pecoraro, R., An, L., Anishchenko, I., Glasscock, C. & Baker, D. Atomic context-conditioned protein sequence design using LigandMPNN. *Nature Methods* **22**, 717 (2025).

58. Anthropic. *Model Context Protocol Specification* Open protocol for LLM–tool integration. Announced November 25, 2024. 2024.

ACKNOWLEDGEMENTS

Bruno and Monika—the reason a dreamer like me managed to get through university. My brothers—I learned more from them than they probably realize.

My school teachers, whom I spent many years annoying but remember every last one of. My first-grade biology teacher was a botanist who, for reasons I still don't understand, explained the hexadecimal system to a room of six-year-olds. It stuck. I hope this thesis makes up for at least some of the trouble.

Vladimir—for giving me the chance at a PhD at ETH, my home university.

Xavi—for the mentorship, and for giving me remarkable freedom in my research from an early stage. He let me explore while always keeping me on track.

Everyone in the group deserves thanks for welcoming me into the lab, teaching me to use a pipette, sharing coffee, and patiently listening to my ramblings about AI, language models, and whatever else I was excited about that week.

Twelve years at ETH brought me into contact with remarkable researchers, students, and professors. The Paul Scherrer Institute and its staff provided an environment where computational and experimental work met daily.

This work was supported by the Swiss National Science Foundation (grant no. 192780).

We acknowledge Claude Opus-4.5 and Opus-4.6 (Anthropic) for editing assistance in preparing this manuscript. Claude was also used as the LLM in the ProtOS-MCP example (chapter 6). The scientific work, analysis, and conclusions are entirely my own.

CURRICULUM VITAE

PERSONAL DATA

Name	Flurin Stefan Hidber
Date of Birth	August 10, 1994
Place of Birth	Mels, Switzerland
Citizen of	Switzerland

EDUCATION

2022 – 2026	ETH Zürich, Department of Biology (D-BIOL) Zürich, Switzerland *Doctoral studies*
2019 – 2022	ETH Zürich Zürich, Switzerland *Final degree:* MSc Interdisciplinary Sciences
2014 – 2019	ETH Zürich Zürich, Switzerland *Final degree:* BSc Interdisciplinary Sciences

PUBLICATIONS

Articles in peer-reviewed journals:

1. Hidber, F. S. *et al.* A generic residue numbering system for microbial rhodopsins—Unifying structural frameworks and functional mapping. *npj Structural Biology*. Under review at npj Structural Biology (Springer Nature) (2026).

Manuscripts in preparation:

2. Hidber, F. S. *et al. LAMBDA: Cross-family spectral prediction* Manuscript in preparation. 2025.
3. Hidber, F. S. *et al. ProtOS-MCP: A structural biology agent* Manuscript in preparation. 2025.

A GENERIC RESIDUE-NUMBERING SYSTEM FOR MICROBIAL RHODOPSINS

Flurin S. Hidber[1,2,10,#], Seiya Tajima[3,4,#], Koichiro E. Kishi[3,4,#], Yosuke Nishimura[5,6], Susumu Yoshizawa[5,7,8], Xavier Deupi[1,9,10,*], Hideaki E. Kato[3,4,11,12,13,*]

[1] Condensed Matter Theory Group, PSI Center for Scientific Computing, Theory, and Data, 5232 Villigen-PSI, Switzerland

[2] Department of Biology, ETH Zürich, Zürich, Switzerland

[3] Research Center for Advanced Science and Technology, The University of Tokyo, Meguro, Tokyo, Japan.

[4] Department of Life Sciences, School of Arts and Sciences, The University of Tokyo, Meguro, Tokyo, Japan.

[5] Atmosphere and Ocean Research Institute, The University of Tokyo, Chiba, Japan

[6] Research Center for Bioscience and Nanoscience (CeBN), Research and Development Center for Marine Biosciences, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Kanagawa, Japan

[7] Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan

[8] Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, Tokyo, Japan

[9] Laboratory of Biomolecular Research, PSI Center for Life Sciences, 5232 Villigen-PSI, Switzerland

[10] Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

[11] Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo, Tokyo, Japan.

[12] FOREST, Japan Science and Technology Agency, Kawaguchi, Saitama, Japan.

[13] CREST, Japan Science and Technology Agency, Kawaguchi, Saitama, Japan.

[#] Flurin S. Hidber, Seiya Tajima, and Koichiro E. Kishi contributed equally to this work.

*Correspondence: Xavier Deupi (xavier.deupi@psi.ch) and Hideaki E. Kato (c-hekato@g.ecc.u-tokyo.ac.jp).

ABSTRACT

Microbial rhodopsins are photoreceptors with a seven-transmembrane architecture that use the chromophore retinal—a vitamin A derivative—to convert light absorption into ion transport, signaling, and enzymatic activity. Since their discovery half a century ago, genome sequencing has uncovered thousands of these proteins across archaea, bacteria, eukaryotes, and viruses, revealing a diverse functional repertoire that includes ion pumps and channels, transducer activators, and enzyme/channel regulators. Although the structure of the seven-transmembrane bundle and the retinal-binding pocket is highly conserved, the sequence identity across major classes is low. This low sequence identity, combined with the rapidly expanding number of microbial rhodopsins identified by metagenomics, makes direct comparisons across the superfamily increasingly difficult. Here, we introduce a generic residue-numbering system for microbial rhodopsins (MO numbering system), a structure-based framework anchored to the retinal-binding pocket that enables consistent residue comparison across even the most divergent family members. Analyzing 129 structures and approximately 40,000 sequences, we demonstrate its utility by analyzing sequence–structure–function relationships across the superfamily, including previously uncharacterized members. The MO numbering system provides a common language for organizing the growing repertoire of microbial rhodopsins, integrat-

ing structural, spectroscopic, and electrophysiological data, and engineering next-generation optogenetic tools.

INTRODUCTION

The microbial rhodopsin superfamily comprises light-sensitive proteins that use the chromophore retinal—a vitamin A derivative—to convert photon absorption into concerted cellular events. Since the discovery of the proton pump bacteriorhodopsin in *Halobacterium salinarum* (*Hs*BR) over fifty years ago[1], microbial rhodopsins have revealed a striking diversity of functions. These proteins are found across all domains of life—Archaea, Bacteria, and Eukaryota—and viruses[2, 3, 4] functioning as proton, chloride, or sodium pumps, cation and anion channels, activators of downstream signal transducers, and even as enzyme/channel/transporter regulators. Their photochemical properties are equally varied: while in most microbial rhodopsins light isomerizes all-*trans* retinal to the 13-*cis* form, some produce 9-*cis* or 11-*cis* isomers as photoproducts[2, 5, 6, 7].

Despite this functional and photochemical diversity, all microbial rhodopsins share a core architecture: a tight bundle of seven transmembrane helices (TMs) surrounding the retinal-binding pocket, with retinal bound via a Schiff base to a conserved lysine on TM7. Some families form oligomers ranging from dimers to hexamers[6, 8, 9, 10, 11, 12, 13, 14], while others have an additional TM or an inverted membrane topology[6, 10, 15, 16]. Remarkably, this structural conservation persists even though sequence identity between major families can be as low as 10–12%[17], with a nearly identical orientation of retinal in the binding pocket. Such conservation despite extreme sequence divergence[18] suggests powerful evolutionary constraints on the architecture required for retinal binding and photoisomerization.

Microbial rhodopsins are often classified, annotated, and compared by tracking characteristic sequence motifs that are coupled to their function. However, the high sequence variability within the family complicates the analysis of large datasets obtained in metagenomics studies, particularly the comparison of distant subfamilies. Researchers in the field often resort to developing *ad hoc* sequence numbering systems using a reference rhodopsin, such as *Hs*BR, channelrhodopsin-2 from *Chlamydomonas reinhardtii* (CrChR2, the most common optogenetics tool)[19], or C1C2 also from *Chlamydomonas reinhardtii* (the first channelrhodopsin structure, a chimera between channelrhodopsin-1 and 2)[11]. This makes it challenging to recognize equivalent positions and conserved motifs in different rhodopsins throughout the scientific literature. Other fields have overcome this hurdle by adopting generic residue-numbering (GRN) systems. A prime example is the field of G protein-coupled receptors (GPCRs), in which the Ballesteros–Weinstein GRN system[20]—or its descendant GPCRdb system[21]—assign a unique index to equivalent positions in different proteins. Similar systems developed for microbial rhodopsins are limited to a few specific families[22] and cannot be applied across the entire superfamily.

In this work, we develop a GRN system for microbial opsins (MO numbering system) based on the structure of the retinal-binding site, rather than on sequence conservation. This numbering system provides a consistent framework for comparing residues across even the most divergent rhodopsins. Using a dataset of 129 structures and approximately 40,000 sequences, we demonstrate its utility by mapping functionally important residues across ion pumps, channels,

and sensory rhodopsins, and by identifying conserved sequence signatures in both characterized and previously uncharacterized rhodopsin families.

RESULTS

Assembly of a dataset of microbial rhodopsin structures

Comparative structural analysis of microbial rhodopsins has been somewhat hindered by limited structural coverage of certain functional classes and taxonomic groups. To address this limitation, we assembled a comprehensive dataset of 129 structures, combining 69 experimentally determined structures and 60 computationally predicted models **(Fig. 1)**. The dataset includes rhodopsins from all domains of life—archaea (27, 21%), bacteria (42, 33%), and eukaryota (50, 39%)—and also includes synthetic variants (5, 4%) and viral rhodopsins (5, 4%). Functionally, the dataset spans proton pumps (42, 33%), cation channels (28, 22%), regulatory rhodopsins with fused domains (14, 11%), anion channels (12, 9%), chloride pumps (12, 9%), transducer activators (8, 6%), sodium pumps (6, 5%), and rhodopsins of undefined function (7, 5%). Due to biases in available structural databases, proton pumps are overrepresented in the experimentally determined structures (31/69, 45%). To balance functional coverage, we enriched underrepresented families through predictive structural modeling: enzyme/channel/transporter regulators (12 predictions), cation channels (11), proton pumps (11), anion channels (8), chloride pumps (6), transducer activators (6), sodium pumps (3), and rhodopsins of undefined function (3).

The dataset reflects known taxonomic distributions of rhodopsin functions. Channelrhodopsins—both cation-conducting (23/28) and anion-conducting (10/12)—are concentrated in eukaryotes, consistent with their restriction to algae and related protists[23]. Similarly, enzyme/channel/transporter regulatory rhodopsins are predominantly eukaryotic (12/14), consistent with enzyme rhodopsins having been identified mostly in fungi, green algae, and choanoflagellates[10]. Sodium pumps appear exclusively in bacteria (6/6), consistent with their known distribution in marine flavobacteria and related lineages[24]. Chloride pumps are found in both archaea and bacteria, reflecting independent evolutionary origins in these lineages[25]. Proton pumps show the broadest distribution across Archaea, Bacteria, and Eukaryota, consistent with ancestral sequence reconstruction studies suggesting that the last common ancestor of microbial rhodopsins functioned as a light-driven proton pump[26].

Validation of the structural predictions

As nearly half of our dataset consists of predicted models, we first assessed their accuracy using two complementary validation sets of experimental structures: benchmark set A ($n = 42$), composed of structures deposited before September 2021 (which may have been used in the training of the structure prediction methods), and blind set B ($n = 27$), composed exclusively of structures deposited after the model training cutoff. We assessed the accuracy of the models by comparing the protein backbone, the retinal-binding pocket, and retinal **(Supplementary Fig. 1)**. The predicted models have high accuracy. Models in benchmark set A had a mean backbone Cα RMSD of 0.61 ± 0.23 Å, binding-pocket iRMSD of 0.51 ± 0.28 Å, and retinal L-RMSD of 0.40 ± 0.24 Å. Models in blind set B had a mean backbone Cα RMSD of 0.80 ± 0.24 Å, binding-pocket iRMSD of 0.75 ± 0.28 Å, and retinal L-RMSD of 0.62 ± 0.27 Å. The errors in

blind set B are consistently higher, but the mean differences between the predicted models and the reference structures remain below 1 Å.

To illustrate the accuracy of the predicted models, we show experimental and predicted structures for six representative rhodopsins spanning different functional classes **(Fig. 2)**. In all cases, the overall fold was accurately reproduced, with loops largely correctly positioned **(Fig. 2a)**. The retinal-binding pocket was also consistently modeled at sub-angstrom precision **(Fig. 2b)**, including accurate handling of unusual 6-*s-cis* β-ionone ring conformations (as observed in the cation channelrhodopsin *Kn*ChR), consistent with the non-G rule[27, 28, 29]. These results show that current structure prediction methods reliably capture both the global microbial rhodopsin fold and the details of the retinal-binding pocket, supporting the inclusion of predicted structures to increase the diversity of our dataset.

Overall structural conservation of the microbial rhodopsin fold

To analyze structural conservation across the dataset, we calculated pairwise RMSDs between Cα atoms of the transmembrane helices for all structure pairs. The resulting RMSD matrix **(Fig. 3)** reveals substantial structural conservation across the superfamily. However, sequence identity is low, averaging $21 \pm 9\%$ in the TM bundle (min = 7%) and $19 \pm 9\%$ in the full sequence (min = 6%). Hierarchical clustering of the RMSD matrix identified a single main structural cluster, confirming that microbial rhodopsin fold is highly conserved across the superfamily. A generic residue-numbering system for microbial rhodopsins should therefore be based on structural features rather than sequence.

Development of the generic residue-numbering system

The MO numbering system is inspired by the Ballesteros-Weinstein system for G protein-coupled receptors but adapted to the unique architecture of microbial rhodopsins. MO numbering codes consist of two numbers: the first denotes the helix (1-7) and the second the position within the helix relative to the residue closest to retinal, defined as position 50. For example, position 7.46 denotes a residue in TM7, four residues before the Schiff base-forming lysine (defined as 7.50).

The first task in developing the MO numbering system was identifying the closest residues to retinal in each helix **(Fig. 4a)**, which serve as the 'anchors' of the numbering system. The distance profiles along each helix **(Fig. 4b)** exhibit the expected systematic variations, reaching their minima at central positions and increasing toward the N- and C-termini. The binding pocket (<4 Å around retinal) is formed by residues from TM3-7. Only two residues in TM2 are part of the extended pocket (<6 Å), while TM1 is at a peripheral location within the 7-TM bundle and further from retinal. The closest residue to retinal in each helix is designated as an anchor **(Fig. 4b, vertical lines)**. The sequence variability around the anchor positions **(Supplementary Fig. 2b)** shows that they lie within or adjacent to regions of elevated sequence conservation despite the overall sequence divergence of the superfamily. These data support the choice of a ligand-centric approach to capture structurally and functionally critical positions across diverse microbial rhodopsins **(Supplementary Fig. 3)**.

Application of the generic residue-numbering system

To demonstrate the application of the MO numbering system, we first focus on residues that form the retinal-binding pocket, using *Hs*BR as a reference **(Fig. 5a)**. Retinal is covalently bound to K216$^{7.50}$ (GRNs are shown as superscripts) via a Schiff base linkage and is surrounded by 18 additional residues from TM3-7 (5 from TM3, 3 from TM4, 4 from TM5, 4 from TM6, and 2 from TM7). Two highly conserved carboxylates, D$^{3.45}$ and D$^{7.46}$, located near K$^{7.50}$ (D85$^{3.45}$ and D212$^{7.46}$ in *Hs*BR), are historically referred to as the "Schiff base counterions" and stabilize the positive charge of the protonated Schiff base. Residues at positions 3.53, 4.51, 6.54, and 7.49—known as the L/Q, N/LI, G/P, and A/TS switches, respectively—contribute to spectral tuning in microbial rhodopsins[30, 31, 32, 33]. In addition, position 4.54 (G122$^{4.54}$ in *Hs*BR) plays an important role in determining retinal planarity: when a non-glycine residue occupies this position, steric clash with the retinal β-ionone ring can induce rotation and favor a 6-*s-cis* retinal conformation (the "non-G rule")[27, 28, 29]. Two other notable positions near the β-ionone ring are 5.44 and 5.47. Although retinal is enclosed by TM3-7, many microbial rhodopsins have a small lateral opening between TM5 and TM6 that connects the retinal pocket to the lipid bilayer[11, 13, 34, 35, 36]. The shape of this opening varies among microbial rhodopsins: in some bacterial ion-pumping rhodopsins, residue 5.47 contributes to its formation, whereas in heliorhodopsins (HeR) both 5.44 and 5.47 contribute. Functionally, 5.44 has been implicated in retinal uptake from the membrane—the G171$^{5.44}$W mutation in *Ts*HeR inhibits this process[13]—while 5.47 is involved in binding the carotenoid antenna pigment, as evidenced by the G178$^{5.47}$W mutation in *Gloeobacter* rhodopsin, which inhibits carotenoid binding[37, 38].

The MO numbering system also facilitates comparative analyses across subfamilies. As an example, in ion-pumping rhodopsins, the identity of the transported substrate and the direction of transport are primarily influenced by three residues at positions 3.45, 3.49, and 3.56, which are located on one face of TM3 **(Fig. 5b)**[2, 4, 25]. Ion-pumping rhodopsins containing D$^{3.45}$, T$^{3.49}$, and D$^{3.56}$—known as the DTD motif—predominantly function as outward proton pumps. Upon photoactivation, the proton is released from the Schiff base to the extracellular bulk solvent via T$^{3.49}$, D$^{3.45}$, and the proton release group composed of E^{ECL3} and E$^{7.38}$, accompanied by a conformational change of the key arginine residue R$^{3.42}$. A new proton is subsequently supplied from the intracellular bulk solvent to the Schiff base via D$^{3.56}$, thereby accomplishing net transport of a single proton from the intracellular to the extracellular side. In contrast, inward proton-pumping rhodopsins, which comprise schizorhodopsins (SzRs) and xenorhodopsins (XeRs), lack one of the two counterions near the Schiff base. Specifically, D85$^{3.45}$ in *Hs*BR is replaced by F$^{3.45}$ in SzRs, whereas D212$^{7.46}$ in *Hs*BR is replaced by P$^{7.46}$ in XeRs. The D$^{3.45}$–T$^{3.49}$–D$^{3.56}$ motif is replaced by FSE in SzRs and by DSA, DTL, or DTS in XeRs, respectively, and the proton release group conserved in outward proton pumps is absent in both SzRs and XeRs[39, 40]. Inward chloride pumps have T$^{3.45}$-S$^{3.49}$-A/D$^{3.56}$ or N$^{3.45}$-T$^{3.49}$-Q$^{3.56}$ motifs, and the loss of negative charge caused by replacement of D$^{3.45}$ with T$^{3.45}$ or N$^{3.45}$ is compensated by binding of the chloride substrate[41, 42, 43]. Outward sodium pumps share the N$^{3.45}$-D$^{3.49}$-Q$^{3.56}$ motif, in which transient proton transfer from the Schiff base to D$^{3.49}$ enables the sodium ion to pass through the Schiff base region[24, 35, 44, 45].

The MO numbering system is equally applicable to rhodopsins with non-canonical architectures. Heliorhodopsins (HeRs), for instance, have an inverted topology, with an intracellular N-terminus and an extracellular C-terminus **(Fig. 5c)**. Despite this inversion, our structural

alignment and annotation approach remains effective. At the sequence level, HeRs retain a TM3 counterion at position 3.45, whereas the TM7 counterion at 7.46 is replaced by serine. While the molecular functions of most HeRs remain unclear, a recent study suggested that V2HeR3 functions as a proton transporter[46], and the MO numbering system may facilitate future comparative analyses as more functional data become available.

Channelrhodopsins, widely used in optogenetics, present a challenge for systematic comparison: mutagenesis and protein engineering have generated numerous functional variants, while metagenomics continues to uncover natural diversity. The MO numbering system enables mapping of this diversity onto a common structural framework, as illustrated for C1C2 structure **(Fig. 5d)**. Unlike ion-pumping rhodopsins, the TM3 residues at positions 3.45, 3.49, and 3.56 are not essential for channel function in canonical dimeric channelrhodopsins; substitutions at these and nearby positions, such as $E^{3.45}T$ or $H^{3.56}R$, primarily affect channel kinetics[47, 48]. Instead, ion selectivity and kinetics are largely determined by five glutamates (E1–E5, corresponding to $E^{2.46}$, $E^{2.47}$, $E^{2.54}$, $E^{2.61}$, and $E^{2.65}$) aligned along TM2[49, 50, 51, 52]. The DC gate, formed by $D^{4.47}$ and $C^{3.50}$, controls channel closure kinetics and light sensitivity, enabling bistable optical control of neuronal activity[53, 54, 55]. **Fig. 5d** summarizes additional positions implicated in spectral tuning, conductance, and kinetics[27, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68], and **Supplementary Fig. 4** summarizes positions that affect properties in trimeric channelrhodopsins[14, 28, 69, 70, 71, 72, 73, 74, 75].

Transducer-activating rhodopsins (sensory rhodopsins) relay light signals to downstream effectors. Structural, spectroscopic, and mutational studies have clarified the activation mechanism of *Np*SRII[76, 77, 78] **(Fig. 5e)**. *Np*SRII forms a 2:2 complex with its transducer HtrII through an interface formed by TM6 and TM7 of *Np*SRII and TM1 and TM2 of HtrII. In *Np*SRII, retinal photoisomerization induces structural changes in residues $Y174^{6.53}$, $L200^{7.45}$, $D201^{7.46}$, and $T204^{7.49}$, which are transmitted to HtrII via the interaction between $Y199^{7.44}$ and N74 of HtrII[78, 79]. Notably, the key residue $T204^{7.49}$ in *Np*SRII is not conserved in *H. salinarum* SRI (*Hs*SRI), suggesting that this signaling pathway may be specific to *Np*SRII rather than a general feature of transducer-activating rhodopsins[80].

Enzyme-fused rhodopsins, such as rhodopsin-guanylyl cyclases and rhodopsin-phosphodiesterases, couple light detection to enzymatic activity. A characteristic feature of these proteins—shared with some channel-fused rhodopsins—is an additional TM0 helix at the N-terminus of the seven-transmembrane core[6, 16]. The MO numbering system annotates the seven helices forming the retinal-binding domain, excluding TM0. For NeoR, the longest-wavelength rhodopsin reported to date, functional studies have identified three acidic residues essential for activity: $E136^{3.45}$, $D140^{3.49}$, and $E262^{7.46}$ [81] **(Fig. 5f)**. As structural and spectroscopic data on enzyme-fused rhodopsins accumulate, the MO numbering system will facilitate comparative analysis of their activation mechanisms.

The MO numbering system also accommodates microbial rhodopsins that exhibit gaps or insertions arising from local structural distortions **(Supplementary Fig. 5)**. For example, ChRmine, a trimeric cation-conducting channelrhodopsin[82], contains a constricted 3_{10} helical turn within TM6 that results in a gap in the sequence alignment[14] **(Supplementary Fig. 5, salmon)**. In this case, the 'missing' position (6.48) is simply skipped. Conversely, TaraRRB-R1, a channel-regulating rhodopsin[6], contains a bulged π-helical turn in TM5 that introduces an additional residue between 5.45 and 5.46 **(Supplementary Fig. 5, green)**, which designated 5.45_1.

Together, these applications demonstrate that the MO numbering system provides a unified framework for comparing rhodopsins across functional classes, taxonomic origins, and structural variations.

Application of the MO numbering system to a large dataset of microbial rhodopsins

To test the MO numbering system at scale, we applied it to a large sequence dataset that includes both characterized and uncharacterized rhodopsins. We combined the 129 functionally characterized rhodopsins in our curated dataset with a large collection of uncharacterized rhodopsin-like sequences, yielding an initial set of over 300,000 sequences from archaea, bacteria, eukaryotes, and viruses (Nishimura et al., in preparation). After curation, we obtained approximately 40,000 non-redundant sequences. From pairwise sequence similarities, we constructed a protein similarity network in which nodes represent individual sequences and edges connect pairs exceeding a defined similarity threshold **(Fig. 6a)**. The network resolved 14 clusters containing functionally characterized rhodopsins and 17 clusters containing only uncharacterized rhodopsins **(Supplementary Fig. 6)**. The 14 characterized clusters comprise six major groups—archaeal/eukaryotic, bacterial, viral (VirR)[83, 84], canonical channelrhodopsins (ChRs)[85, 86], canonical histidine kinase-fused rhodopsins (HKRs)[85, 86], and canonical heliorhodopsins (HeRs)[87]—and eight minor groups: schizorhodopsins (SzRs)[88], cryptophyte anion-conducting channelrhodopsins (ACRs)[89], alveolate ChRs[23], pump-like channelrhodopsins (PLCRs)[90], guanylate cyclase-/phosphodiesterase-fused rhodopsins (RhoGC/RhoPDEs)[16], bestrophin-fused rhodopsins (BestRs)[6], noncanonical HKRs[91], and eukaryotic/viral HeRs[46] **(Supplementary Fig. 6)**.

To validate the clustering and explore uncharacterized rhodopsins, we predicted structures for 11 sequences from the periphery of 5 of the 14 functionally characterized clusters, along with one representative member from each of the 17 uncharacterized clusters **(Supplementary Fig. 7a)**. In total, we generated structural models for 28 rhodopsin-like proteins **(Supplementary Fig. 7b)**. All 11 characterized and 13 of the 17 uncharacterized sequences (uncharacterized clusters 1–13) adopted a rhodopsin fold and were retained for further analysis. Of the 13 uncharacterized rhodopsins, two lacked the conserved Lys$^{7.50}$ and one exhibited an incorrectly predicted retinal-binding pose. Several uncharacterized clusters exhibited distinctive structural features: for example, the rhodopsin in cluster 2 has an unusually long TM1, whereas the rhodopsin in cluster 10 contains β-strands in ECL2 and ECL3 that form an extracellular β-sheet **(Supplementary Fig. 7b)**.

To examine sequence conservation, we generated sequence logos in selected clusters **(Fig. 6b,c; Supplementary Fig. 8 and 9)**. Conserved residues grouped near the X.50 anchor positions at approximately four-residue intervals, consistent with enrichment on the retinal-facing side of each helix; lipid-exposed positions showed lower conservation. Across all clusters, residues in TM3, TM6, and TM7 were more conserved than those in other helices. The arginine at position 3.42 is highly conserved across all functionally characterized clusters—with the exception of the PLCR cluster—consistent with its known key functional roles[92, 93, 94, 95, 96]. Residues forming the retinal-binding pocket (positions 3.46, 6.50, 6.53, and 6.54) and the Schiff base counterions (positions 3.45 and 7.46) are well conserved within characterized clusters; however, position 7.46 is more conserved than 3.45. Most characterized clusters have a highly conserved Asp at 7.46, but the alveolata ChR, bacterial/archaeal HeR, and eukaryotic/viral HeR clusters

show highly conserved Asn, Ser, and Ser, respectively. Positions 3.50 and 3.51 vary markedly between characterized clusters—while remaining conserved within each—suggesting that these positions are linked to cluster-specific functions and serve as characteristic sequence signatures for these rhodopsin families. Similar conservation patterns were observed in uncharacterized clusters, although position 3.50 in cluster 1 and position 7.46 in cluster 4 showed reduced conservation.

DISCUSSION

The repertoire of microbial rhodopsins continues to expand. With estimates of global microbial diversity ranging from 10^{6}[97] to 10^{12} species[98], and over two million bacterial and archaeal genomes now sequenced[99], the number of identified rhodopsin genes—already exceeding 10,000[100]—will undoubtedly grow further. This expanding diversity, combined with ongoing optogenetics engineering efforts, has made a unified residue-numbering system increasingly necessary.

The limitations of existing numbering approaches become apparent when comparing across functional classes. Ion-pumping rhodopsins have historically been described using BR-based residue numbering, whereas channelrhodopsins are typically numbered according to ChR2- or C1C2-based systems. The three-residue TM3 motif introduced in 2013-2014 provides a useful signature for classifying ion-pump functions[24, 25]; however, in channelrhodopsins, functional properties are determined primarily by other residues, including five glutamates in TM2, rather than by the TM3 motif. These differences underscore the need for a broadly applicable numbering system, which motivated the development of the MO numbering system.

The sequence similarity network provides insight into the evolutionary relationships among microbial rhodopsins **(Fig. 6a, Supplementary Fig. 6)**. Rhodopsins from archaea and eukaryotes form a single major cluster (the archaeal/eukaryotic rhodopsin cluster), whereas bacterial rhodopsins occupy a distinct cluster (the bacterial rhodopsin cluster). This pattern indicates that eukaryotic microbial rhodopsins are more closely related to archaeal than to bacterial rhodopsins, consistent with previous phylogenetic and structural analyses[101, 102, 103, 104].

Rhodopsins within the same cluster but with different functions likely diverged from a common ancestor, whereas functionally similar rhodopsins in different clusters likely arose through convergent evolution. For example, both the archaeal/eukaryotic and bacterial clusters include chloride-pumping rhodopsins, yet their TM3 motifs differ ($T^{3.45}$-$S^{3.49}$-A/$D^{3.56}$ versus $N^{3.45}$-$T^{3.49}$-$Q^{3.56}$, respectively), suggesting that chloride-transport function evolved independently in these lineages. Within clusters, oligomeric assembly and sequence motifs are conserved despite functional diversification. This conservation likely explains why researchers have been able to interconvert functions through relatively few mutations, even among rhodopsins sharing only ~30% sequence identity[51, 52, 79, 105, 106, 107, 108].

The conservation of oligomeric assembly within clusters raises a question: what determines oligomeric state? All characterized microbial rhodopsins form specific oligomers, ranging from dimers to hexamers, and rhodopsins within the same clade consistently adopt the same oligomeric state[8, 9, 13, 14, 88, 109, 110]. Yet residues on the outer, lipid-exposed surface are generally poorly conserved, even in helices that participate in oligomeric interfaces. This suggests that oligomeric state is not dictated by a fixed set of conserved interface residues. In some rhodopsins, such as *Np*HR, outer-facing residues do contribute to interprotomer contacts and

stabilize oligomerization[111], but these residues are not conserved across clusters, implying that they fine-tune oligomer stability within a family rather than specifying oligomeric state. Instead, we propose that the type of oligomer formed is largely determined by helix-packing geometry, specifically how the transmembrane helices are tilted and arranged relative to one another. Helix tilt can be influenced by glycine and proline residues within helices, as well as loop architecture. In particular, the length and orientation of the extracellular loop 1 (ECL1) have been proposed to distinguish trimeric from pentameric or hexameric assemblies[101]. However, pentameric rhodopsins such as GR and TR form trimers in DDM micelles[101, 109], suggesting that the ECL1 may bias helix tilt toward particular arrangements without rigidly fixing protomer number. This observation implies that the membrane or detergent environment also contributes to stabilizing a given assembly.

Beyond evolutionary and oligomeric insights, our study also assessed the utility of structure prediction for rhodopsin analysis. Structure prediction methods performed well for rhodopsins. Predictions captured the overall fold and the detailed architecture of the retinal-binding pocket, often correctly reproducing the retinal configuration and orientation, including whether the chromophore adopts an all-*trans* geometry or a 6-*s*-*cis* conformation consistent with the non-G rule[28, 29, 35]. Our backbone RMSDs (0.61 Å for Set A and 0.80 Å for Set B) are comparable to those reported for Boltz-1 in the Runs N' Poses benchmark[112] (0.56 Å) and lower than AlphaFold2 predictions of GPCRs[113] (1.12–1.55 Å). The accuracy of retinal placement is particularly notable: our mean retinal RMSDs of 0.41 Å (Set A) and 0.62 Å (Set B) are substantially lower than the 2.30 Å mean ligand RMSD that Boltz-1 achieved on diverse protein-ligand complexes in the Runs N' Poses benchmark. This likely reflects the strong structural constraints imposed by the retinal-binding pocket, including the covalent Schiff base linkage to the conserved lysine.

However, inaccuracies were frequently observed in fine structural details, including the placement of loop regions and side-chain rotamers. For example, in the predicted structure of *Kn*ChR, the retinal chromophore is not covalently linked to the Schiff base-forming lysine ($K^{7.50}$) **(Fig. 2b and Supplementary Fig. 10a)**. In *Gt*ACR1, the β-ionone ring of the retinal adopts an incorrectly twisted conformation[12] **(Supplementary Fig. 10b)**. In Tara-RRB R2, the cytoplasmic ends of TM0, TM1, TM2, and TM7 show marked deviations from the experimentally determined structure **(Fig. 2a and Supplementary Fig. 10c)**. In *Hc*KCR2, the architecture of the potassium selectivity filter is incorrectly predicted[28] **(Supplementary Fig. 10d)**. These issues may become problematic when predicted structures are used e.g. as inputs for molecular dynamics simulations or for interpreting spectroscopic measurements, where accurate loop conformations and side-chain rotamers are critical. Thus, while current structure prediction methods are sufficiently accurate for comparative analyses and for capturing the overall architecture and retinal environment, they do not replace experimental structure determination when precise atomic detail or characterization of conformational dynamics is required.

The MO numbering system offers the microbial rhodopsin field a common vocabulary. As the repertoire of rhodopsins continues to expand through metagenomics and structural prediction, we anticipate that it will become a standard tool for comparing residues, interpreting functional data, and guiding optogenetic engineering.

METHODS

Structural Dataset

We assembled a comprehensive structural dataset of 129 microbial rhodopsins spanning all major functional classes—proton pumps, chloride pumps, sodium pumps, cation channels, anion channels, sensory rhodopsins, and enzyme rhodopsins—with representatives from archaea, bacteria, and eukaryotes. Experimentally determined structures were obtained from the Protein Data Bank, supplemented by eight recently solved channelrhodopsin structures (Kishi et al., in preparation) ($n = 69$). The experimental structures were partitioned into two validation cohorts based on PDB deposition date relative to the Boltz-1[114] training data cutoff (September 2021): a benchmark set (Set A, $n = 42$) potentially overlapping with training data, and a blind test set (Set B, $n = 27$) comprising structures deposited after the cutoff. For each experimentally characterized protein, we generated Boltz-1 predictions to enable systematic assessment of prediction accuracy. We additionally predicted structures for 60 functionally characterized rhodopsins lacking experimental structures. All analyses were performed on chain A, with retinal chromophores identified by residue name and validated by proximity ($\leq$6.0 Å to protein atoms).

Structure Prediction and Validation

For Boltz-1 predictions, input sequences were trimmed to transmembrane domain boundaries when exceeding 400 residues. Retinal was specified as all-trans-retinal without the aldehyde oxygen (SMILES code: CC=C(C)C=CC=C(C)C=CC1=C(C)CCCC1(C)C), allowing Boltz-1 to predict the covalent Schiff base linkage to the conserved lysine. Prediction accuracy was assessed by comparing each model to its experimental counterpart using the combinatorial extension alignment algorithm (CEalign)[115] (window size = 8, maximum gap = 30) with a two-pass refinement procedure: an initial global alignment using all Cα atoms, followed by refinement restricted to residue pairs with inter-Cα distance $\leq$3 Å after the initial superposition to exclude flexible regions with large structural divergence. Accuracy was quantified using three metrics: backbone Cα RMSD (overall fold accuracy), binding pocket RMSD (Cα atoms within 6.0 Å of retinal; local accuracy in the chromophore environment), and ligand RMSD (mean closest-atom distance between experimental and predicted retinal coordinates).

Structure Conservation Analysis

We first aligned the shared TM bundle of all members of our dataset. Pairwise structural similarities (calculated using CEalign) were computed as Cα RMSD across TM 1–7 to generate a symmetric distance matrix. Our structure analysis leverages the conserved microbial rhodopsin fold by using the closest residue in each helix (TM1-TM6) and the Schiff-base lysine in TM7 as anchors. We then performed hierarchical clustering. For each GRN position, two distance profiles were calculated: (i) side-chain distances, from heavy atoms to the nearest retinal atom, and (ii) Cα distances, from backbone Cα to the nearest retinal atom. Equivalent residue pairs for all pairwise alignments were stored to create the blueprint of the MO numbering system.

Generic Residue-Numbering System

We developed the MO numbering system by assigning GRNs based on structural equivalence relative to a reference structure, 7BMH, selected based on its low mean RMSD to all other structures. In the MO numbering system, residues are numbered relative to the helix-specific anchors using the format ⟨Helix⟩.⟨Position⟩ (e.g., 3.49, 3.50, 3.51). After annotating the reference structure, we used the stored residue pair equivalences to assign GRNs to all other structures. The resulting preliminary MO numbering system was curated to correct alignment artifacts, including helix truncations and register shifts arising from the repetitive nature of α-helical structures. Then, the less conserved outer parts of the helices were annotated up until the end of the helices by determining helical residues based on phi-psi angles of the backbone. The non-helical loop regions and tails received systematic identifiers.

Microbial Rhodopsin Sequence Search

To construct Hidden Markov Models (HMMs) for rhodopsin sequence exploration, known rhodopsins and their homologs were collected from NCBI, the UniProt Archive[116], and the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP)[117]. One collection (subAEV: 3,173 sequences) was constructed mainly from archaeal-, eukaryotic-, and viral-derived rhodopsins[118]. Another collection (subB: 2,275 sequences) was mainly composed of bacterial-derived rhodopsins. A third collection (subH: 720 sequences) was composed of heliorhodopsins and their homologs. Profile HMMs for the subAEV, subB, and subH collections were constructed from sequence alignments computed using MAFFT (v7.453)[119] with the "--genafpair" and "--maxiterate 1000" options. Using these HMMs, we searched the UniParc database (as of November 2020), the marine eukaryotic reference catalog[120], the 1000 Plant Transcriptomes Initiative dataset[121], MMETSP, and 10,032 metagenomic assemblies derived from various environments (Nishimura et al., in preparation). Rhodopsin homologs were identified using hmmsearch (HMMER v3.3)[122] with an e-value threshold of $<1e^{-5}$ and a length threshold of $\geq$174 amino acids, corresponding to the HMM length. We clustered 309,570 protein sequences using CD-HIT[123] (v.4.8.1) at 90% global identity (-c 0.90, -G 1) without additional coverage constraints (no -aS/-aL), resulting in 41,136 non-redundant sequences.

Network Analysis Visualization

We collected reference rhodopsins ($n = 147$) that were functionally and/or structurally characterized, along with the 41,136 non-redundant, uncharacterized rhodopsin sequences. Because enzymatic rhodopsins contain additional domains, the rhodopsin regions of both reference and uncharacterized sequences were extracted using aligned regions based on hmmsearch (HMMER v3.3)[122] with the three HMMs (subAEV, subB, and subH). To further remove redundancy within the uncharacterized dataset, protein clustering was performed using MMseqs2 (ver. 15-6f452)[124] at 60% identity and 50% coverage, with options "--min-seq-id 0.6 -c 0.5 --cluster-mode 2 --cov-mode 0". Cluster representatives ($n = 5{,}168$) were combined with the reference rhodopsins (total $n = 5{,}315$).

Normalized pairwise similarity within the resultant sequences was computed using CPAP (`https://github.com/yosuken/cpap`), which performed all-against-all BLASTp with op-

tions "-evalue 1e-2 -dbsize 100000000" and quantified normalized similarity scores between all possible protein pairs based on high-scoring segment pairs (HSPs) that have ≥20% identity and are ≥20 amino acids in length, as follows:

Let $n_{AB,i}$ be the length-normalized bit score (i.e., bit score per position) of the HSP covering position i of protein A, where BLASTp is performed using A as the query and B as the subject. If multiple HSPs cover position i, the maximum value is used.

$$M_{AB,i} = \max(n_{AB,i})$$

Let L_A be the length of protein A. Let N_{AB} ($0 \leq N_{AB} \leq 1$) be the normalized similarity score between proteins A and B.

$$U_{AB} = \frac{\sum_i M_{AB,i}}{\sum_i M_{AA,i}}, \quad 0 \leq i \leq L_A$$

$$N_{AB} = \frac{1}{2}(U_{AB} + U_{BA})$$

The resulting network was visualized using Gephi (ver. 0.10.1). Edges were drawn between protein pairs with normalized similarity scores ≥ 0.2.

Sequence Logo Generation

Sequence logos were generated using WebLogo3[125]. Sequences were aligned according to their MO numbering positions, and logos were computed for each transmembrane helix. For cluster-specific analyses, sequences were grouped based on their assignment in the sequence similarity network, and separate logos were generated for each TM regions (TM1-7). Amino acid frequencies are displayed as information content in bits.

Statistical Analysis and Data Availability

Correlations between structural and sequence metrics were computed using Pearson's correlation with two-tailed significance testing. All analyses (structural alignment, superposition) were performed in Python 3.10 using Biopython. Dataset management and CIF parsing were performed using ProtOS (Hidber et al., manuscript in preparation, `https://github.com/flurinh/protos`). Structure coordinates, alignments, RMSD matrices, GRN tables and all analysis code are provided at `https://github.com/flurinh/mogrn`. The structure analysis data is available at `https://doi.org/10.5281/zenodo.18147121`.

ACKNOWLEDGMENTS

We would like to thank Y. Oheda and A. Kojima (UTokyo) for technical and K. Hasegawa and A. Ohira (UTokyo) for administrative support. We acknowledge ChatGPT, a multimodal large language model created by OpenAI, for providing guidance to improve the readability of this manuscript. After using this tool, we reviewed and edited the content as needed and took full responsibility for the content of the publication. This work was supported by the Nakajima Foundation to H.E.K., Takeda Science Foundation to H.E.K., Japan Agency for Medical Research and Development (AMED) (grant no. 24bm1123057h0001 to H.E.K.), Japan

Society for the Promotion of Science (JSPS) KAKENHI (grant no. 24KJ0788 to S.T., 22K15089 to Y.N., 22H00557 to S.Y. and 22K19265, 22H00400, and 25H01338 to H.E.K.), JST FOREST (grant no. JPMJFR204S to H.E.K.), JST CREST (grant no. JPMJCR21P3 and JPMJCR23B1 to H.E.K.), European Union's Horizon 2020 research and innovation programme (grant agreement no. 951644, supporting F.S.H.), and Swiss National Science Foundation (grant no. 192780 to X.D.). The authors dedicate this paper to G.F.X. Schertler, H. Kandori, and P. Hegemann: many years of insightful and exciting scientific discussions partly inspired this work.

AUTHOR CONTRIBUTIONS

F.S.H., S.T., K.E.K., X.D., and H.E.K. discussed and defined the MO numbering system. Under the supervision of X.D., F.S.H. performed the structure comparison analysis and built the MO numbering system framework. Under the supervision of H.E.K., S.T. analyzed sequence alignments and structures. Y.N. and S.Y. mined in-house metagenome and transcriptome databases to identify rhodopsin-like genes. Under the supervision of S.Y. and H.E.K., S.T. and Y.N. constructed and analyzed the sequence similarity network. F.S.H., S.T., X.D., and H.E.K. wrote the manuscript and prepared the figures with input from all authors. X.D. and H.E.K. supervised the project.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1. Oesterhelt D, Stoeckenius W. Rhodopsin-like Protein from the Purple Membrane of Halobacterium halobium. *Nat N Biol* **233**, 149-152 (1971).
2. Ernst OP, Lodowski DT, Elstner M, Hegemann P, Brown LS, Kandori H. Microbial and Animal Rhodopsins: Structures, Functions, and Molecular Mechanisms. *Chem Rev* **114**, 126-163 (2014).
3. Kato Y, Hasunuma T. Carotenoids: Biosynthetic and Biofunctional Approaches. *Adv Exp Med Biol* **1261**, 121-135 (2021).
4. Rozenberg A, Inoue K, Kandori H, Béjà O. Microbial Rhodopsins: The Last Two Decades. *Annu Rev Microbiol* **75**, 1-21 (2021).
5. Sudo Y, *et al*. A Microbial Rhodopsin with a Unique Retinal Composition Shows Both Sensory Rhodopsin II and Bacteriorhodopsin-like Properties*. *J Biol Chem* **286**, 5967-5976 (2011).
6. Rozenberg A, *et al*. Rhodopsin-bestrophin fusion proteins from unicellular algae form gigantic pentameric ion channels. *Nat Struct Mol Biol* **29**, 592-603 (2022).
7. Sugiura M, *et al*. Unusual Photoisomerization Pathway in a Near-Infrared Light Absorbing Enzymerhodopsin. *J Phys Chem Lett* **13**, 9539-9543 (2022).
8. Shibata M, *et al*. Oligomeric states of microbial rhodopsins determined by high-speed atomic force microscopy and circular dichroic spectroscopy. *Sci Rep* **8**, 8262 (2018).

9. Mannen K, *et al*. Multiple Roles of a Conserved Glutamate Residue for Unique Biophysical Properties in a New Group of Microbial Rhodopsins Homologous to TAT Rhodopsin. *J Mol Biol* **436**, 168331 (2024).
10. Ikuta T, *et al*. Structural insights into the mechanism of rhodopsin phosphodiesterase. *Nat Commun* **11**, 5605 (2020).
11. Kato HE, *et al*. Crystal structure of the channelrhodopsin light-gated cation channel. *Nature* **482**, 369-374 (2012).
12. Kim YS, *et al*. Crystal structure of the natural anion-conducting channelrhodopsin GtACR1. *Nature* **561**, 343-348 (2018).
13. Shihoya W, *et al*. Crystal structure of heliorhodopsin. *Nature* **574**, 132-136 (2019).
14. Kishi KE, *et al*. Structural basis for channel conduction in the pump-like channelrhodopsin ChRmine. *Cell* **185**, 672-689.e623 (2022).
15. Pushkarev A, *et al*. A distinct abundant group of microbial rhodopsins discovered using functional metagenomics. *Nature* **558**, 595-599 (2018).
16. Mukherjee S, Hegemann P, Broser M. Enzymerhodopsins: novel photoregulated catalysts for optogenetics. *Curr Opin Struct Biol* **57**, 118-126 (2019).
17. Grip WJd, Ganapathy S. Rhodopsins: An Excitingly Versatile Protein Species for Research, Development and Creative Engineering. *Front Chem* **10**, 879609 (2022).
18. Olivella M, Gonzalez A, Pardo L, Deupi X. Relation between sequence and structure in membrane proteins. *Bioinformatics* **29**, 1589-1592 (2013).
19. Nagel G, *et al*. Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *Proc Natl Acad Sci* **100**, 13940-13945 (2003).
20. Ballesteros JA, Weinstein H. [19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci* **25**, 366-428 (1995).
21. Isberg V, *et al*. Generic GPCR residue numbers – aligning topology maps while minding the gaps. *Trends Pharmacol Sci* **36**, 22-31 (2015).
22. Bertalan Ev, *et al*. Hydrogen-Bonding and Hydrophobic Interaction Networks as Structural Determinants of Microbial Rhodopsin Function. *J Phys Chem B* **128**, 7407-7426 (2024).
23. Govorunova EG, *et al*. Cation and Anion Channelrhodopsins: Sequence Motifs and Taxonomic Distribution. *mBio* **12**, 10.1128/mbio.01656-01621 (2021).
24. Inoue K, *et al*. A light-driven sodium ion pump in marine bacteria. *Nat Commun* **4**, 1678 (2013).
25. Yoshizawa S, *et al*. Functional characterization of flavobacteria rhodopsins reveals a unique class of light-driven chloride pump in bacteria. *Proc Natl Acad Sci* **111**, 6732-6737 (2014).
26. Sephus CD, Fer E, Garcia AK, Adam ZR, Schwieterman EW, Kacar B. Earliest Photic Zone Niches Probed by Ancestral Microbial Rhodopsins. *Mol Biol Evol* **39**, msac100 (2022).
27. Kato HE, *et al*. Atomistic design of microbial opsin-based blue-shifted optogenetics tools. *Nat Commun* **6**, 7177 (2015).

28. Tajima S, *et al*. Structural basis for ion selectivity in potassium-selective channelrhodopsins. *Cell* **186**, 4325-4344.e4326 (2023).
29. Wang YZ, *et al*. Cryo-EM structure of a blue-shifted channelrhodopsin from Klebsormidium nitens. *Nat Commun* **16**, 5297 (2025).
30. Man D, *et al*. Diversification and spectral tuning in marine proteorhodopsins. *EMBO J* **22**, 1725-1731 (2003).
31. Inoue K, *et al*. Red-shifting mutation of light-driven sodium-pump rhodopsin. *Nat Commun* **10**, 1993 (2019).
32. Sugiura M, Tsunoda SP, Hibi M, Kandori H. Molecular Properties of New Enzyme Rhodopsins with Phosphodiesterase Activity. *ACS Omega* **5**, 10602-10609 (2020).
33. Wang WW, Sineshchekov OA, Spudich EN, Spudich JL. Spectroscopic and photochemical characterization of a deep ocean proteorhodopsin. *J Biol Chem* **278**, 33985-33991 (2003).
34. Luecke H, *et al*. Crystallographic structure of xanthorhodopsin, the light-driven proton pump with a dual chromophore. *Proc Natl Acad Sci* **105**, 16561-16565 (2008).
35. Kato HE, *et al*. Structural basis for Na+ transport mechanism by a light-driven Na+ pump. *Nature* **521**, 48-53 (2015).
36. Tzlil G, *et al*. Structural insights into light harvesting by antenna-containing rhodopsins in marine Asgard archaea. *Nat Microbiol* **10**, 1484-1500 (2025).
37. Imasheva ES, Balashov SP, Choi AR, Jung K-H, Lanyi JK. Reconstitution of Gloeobacter violaceus Rhodopsin with a Light-Harvesting Carotenoid Antenna. *Biochemistry* **48**, 10948-10955 (2009).
38. Chuon K, *et al*. Carotenoid binding in Gloeobacteria rhodopsin provides insights into divergent evolution of xanthorhodopsin types. *Commun Biol* **5**, 512 (2022).
39. Shevchenko V, *et al*. Inward H+ pump xenorhodopsin: Mechanism and alternative optogenetic approach. *Sci Adv* **3**, e1603187 (2017).
40. Higuchi A, *et al*. Crystal structure of schizorhodopsin reveals mechanism of inward proton pumping. *Proc Natl Acad Sci* **118**, e2016328118 (2021).
41. Kolbe M, Besir Hs, Essen L-O, Oesterhelt D. Structure of the Light-Driven Chloride Pump Halorhodopsin at 1.8 Å Resolution. *Science* **288**, 1390-1396 (2000).
42. Kim K, *et al*. Crystal structure and functional characterization of a light-driven chloride pump having an NTQ motif. *Nat Commun* **7**, 12677 (2016).
43. Yun J-H, *et al*. Structure-Based Functional Modification Study of a Cyanobacterial Chloride Pump for Transporting Multiple Anions. *J Mol Biol* **432**, 5273-5286 (2020).
44. Kato HE, Inoue K, Kandori H, Nureki O. The light-driven sodium ion pump: A new player in rhodopsin research. *Bioessays* **38**, 1274-1282 (2016).
45. Skopintsev P, *et al*. Femtosecond-to-millisecond structural changes in a light-driven sodium pump. *Nature* **583**, 314-318 (2020).
46. Hososhima S, *et al*. Proton-transporting heliorhodopsins from marine giant viruses. *eLife* **11**, e78416 (2022).

47. Nagel G, Brauner M, Liewald JF, Adeishvili N, Bamberg E, Gottschalk A. Light Activation of Channelrhodopsin-2 in Excitable Cells of Caenorhabditis elegans Triggers Rapid Behavioral Responses. *Curr Biol* **15**, 2279-2284 (2005).

48. Gunaydin LA, Yizhar O, Berndt A, Sohal VS, Deisseroth K, Hegemann P. Ultrafast optogenetic control. *Nat Neurosci* **13**, 387-392 (2010).

49. Sugiyama Y, *et al*. Photocurrent attenuation by a single polar-to-nonpolar point mutation of channelrhodopsin-2. *Photochem Photobiol Sci* **8**, 328-336 (2009).

50. Eisenhauer K, *et al*. In Channelrhodopsin-2 Glu-90 Is Crucial for Ion Selectivity and Is Deprotonated during the Photocycle*. *J Biol Chem* **287**, 6904-6911 (2012).

51. Berndt A, Lee SY, Ramakrishnan C, Deisseroth K. Structure-Guided Transformation of Channelrhodopsin into a Light-Activated Chloride Channel. *Science* **344**, 420-424 (2014).

52. Wietek J, *et al*. Conversion of Channelrhodopsin into a Light-Gated Chloride Channel. *Science* **344**, 409-412 (2014).

53. Berndt A, Yizhar O, Gunaydin LA, Hegemann P, Deisseroth K. Bi-stable neural state switches. *Nat Neurosci* **12**, 229-234 (2009).

54. Nack M, *et al*. The DC gate in Channelrhodopsin-2: crucial hydrogen bonding interaction between C128 and D156. *Photochem Photobiol Sci* **9**, 194-198 (2010).

55. Yizhar O, Fenno Lief E, Davidson Thomas J, Mogri M, Deisseroth K. Optogenetics in Neural Systems. *Neuron* **71**, 9-34 (2011).

56. Berndt A, *et al*. High-efficiency channelrhodopsins for fast neuronal stimulation at low light levels. *Proc Natl Acad Sci* **108**, 7595-7600 (2011).

57. Kleinlogel S, *et al*. Ultra light-sensitive and fast neuronal activation with the Ca2+-permeable channelrhodopsin CatCh. *Nat Neurosci* **14**, 513-518 (2011).

58. Lin JY, Knutsen PM, Muller A, Kleinfeld D, Tsien RY. ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nat Neurosci* **16**, 1499-1508 (2013).

59. Klapoetke NC, *et al*. Independent optical excitation of distinct neural populations. *Nat Methods* **11**, 338-346 (2014).

60. Dawydow A, *et al*. Channelrhodopsin-2-XXL, a powerful optogenetic tool for low-light applications. *Proc Natl Acad Sci U S A* **111**, 13972-13977 (2014).

61. Rajasethupathy P, *et al*. Projections from neocortex mediate top-down control of memory retrieval. *Nature* **526**, 653-659 (2015).

62. Mager T, *et al*. High frequency neural spiking and auditory signaling by ultrafast red-shifted optogenetics. *Nat Commun* **9**, 1750 (2018).

63. Oda K, *et al*. Crystal structure of the red light-activated channelrhodopsin Chrimson. *Nat Commun* **9**, 3949 (2018).

64. Sridharan S, *et al*. High-performance microbial opsins for spatially and temporally precise perturbations of large neuronal networks. *Neuron* **110**, 1139-1155.e1136 (2022).

65. Scholz N, *et al*. Mechano-dependent signaling by Latrophilin/CIRL quenches cAMP in proprioceptive neurons. *eLife* **6**, (2017).

66. Kato HE, *et al*. Structural mechanisms of selectivity and gating in anion channelrhodopsins. *Nature* **561**, 349-354 (2018).
67. Fernandez Lahore RG, *et al*. Calcium-permeable channelrhodopsins for the photocontrol of calcium signalling. *Nat Commun* **13**, 7844 (2022).
68. Mardinly AR, *et al*. Precise multimodal optical control of neural ensemble activity. *Nat Neurosci* **21**, 881-893 (2018).
69. Vierock J, *et al*. WiChR, a highly potassium-selective channelrhodopsin for low-light one- and two-photon inhibition of excitable cells. *Sci Adv* **8**, eadd7729 (2022).
70. Morizumi T, *et al*. Structures of channelrhodopsin paralogs in peptidiscs explain their contrasting K(+) and Na(+) selectivities. *Nat Commun* **14**, 4365 (2023).
71. Tucker K, Sridharan S, Adesnik H, Brohawn SG. Cryo-EM structures of the channelrhodopsin ChRmine in lipid nanodiscs. *Nat Commun* **13**, 4842 (2022).
72. Duan X, *et al*. Stabilized Ion Selectivity Corrects Activation Drift in Kalium Channelrhodopsins. *Adv Sci (Weinh)*, e09180 (2025).
73. Duan X, *et al*. Suppression of epileptic seizures by transcranial activation of K(+)-selective channelrhodopsin. *Nat Commun* **16**, 559 (2025).
74. Alekseev A, *et al*. Efficient and sustained optogenetic control of sensory and cardiac systems. *Nat Biomed Eng*, (2025).
75. Tose AJ, Nava AA, McGrath SN, Mardinly AR, Naka A. WAChRs are excitatory opsins sensitive to indoor lighting. *bioRxiv*, 2025.2009.2012.675947 (2025).
76. Gordeliy VI, *et al*. Molecular basis of transmembrane signalling by sensory rhodopsin II–transducer complex. *Nature* **419**, 484-487 (2002).
77. Moukhametzianov R, *et al*. Development of the signal in sensory rhodopsin and its transfer to the cognate transducer. *Nature* **440**, 115-119 (2006).
78. Ishchenko A, *et al*. New Insights on Signal Propagation by Sensory Rhodopsin II/Transducer Complex. *Sci Rep* **7**, 41811 (2017).
79. Sudo Y, Furutani Y, Kandori H, Spudich JL. Functional Importance of the Interhelical Hydrogen Bond between Thr204 and Tyr174 of Sensory Rhodopsin II and Its Alteration during the Signaling Process*. *J Biol Chem* **281**, 34239-34245 (2006).
80. Matsunami-Nakamura R, *et al*. Key determinants for signaling in the sensory rhodopsin II/transducer complex are different between Halobacterium salinarum and Natronomonas pharaonis. *FEBS Lett* **597**, 2334-2344 (2023).
81. Broser M, *et al*. NeoR, a near-infrared absorbing rhodopsin. *Nat Commun* **11**, 5682 (2020).
82. Marshel JH, *et al*. Cortical layer–specific critical dynamics triggering perception. *Science* **365**, (2019).
83. Filosof A, Beja O. Bacterial, archaeal and viral-like rhodopsins from the Red Sea. *Environ Microbiol Rep* **5**, 475-482 (2013).
84. Needham DM, *et al*. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proc Natl Acad Sci U S A* **116**, 20574-20583 (2019).
85. Kateriya S, Nagel G, Bamberg E, Hegemann P. "Vision" in single-celled algae. *News Physiol Sci* **19**, 133-137 (2004).

86. Luck M, *et al*. A photochromic histidine kinase rhodopsin (HKR1) that is bimodally switched by ultraviolet and blue light. *J Biol Chem* **287**, 40083-40090 (2012).
87. Pushkarev A, *et al*. A distinct abundant group of microbial rhodopsins discovered using functional metagenomics. *Nature* **558**, 595-599 (2018).
88. Inoue K, *et al*. Schizorhodopsins: A family of rhodopsins from Asgard archaea that function as light-driven inward H(+) pumps. *Sci Adv* **6**, eaaz2441 (2020).
89. Govorunova EG, Sineshchekov OA, Janz R, Liu X, Spudich JL. NEUROSCIENCE. Natural light-gated anion channels: A family of microbial rhodopsins for advanced optogenetics. *Science* **349**, 647-650 (2015).
90. Kishi KE, Kato HE. Pump-like channelrhodopsins: Not just bridging the gap between ion pumps and ion channels. *Curr Opin Struct Biol* **79**, 102562 (2023).
91. Luck M, *et al*. Photoreactions of the Histidine Kinase Rhodopsin Ot-HKR from the Marine Picoalga Ostreococcus tauri. *Biochemistry* **58**, 1878-1891 (2019).
92. Balashov SP, *et al*. Effect of the arginine-82 to alanine mutation in bacteriorhodopsin on dark adaptation, proton release, and the photochemical cycle. *Biochemistry* **32**, 10331-10343 (1993).
93. Ikeura Y, Shimono K, Iwamoto M, Sudo Y, Kamo N. Arg-72 of pharaonis Phoborhodopsin (Sensory Rhodopsin II) is Important for the Maintenance of the Protein Structure in the Solubilized State. *Photochem Photobiol* **77**, 96-100 (2003).
94. Vogt A, Guo Y, Tsunoda SP, Kateriya S, Elstner M, Hegemann P. Conversion of a light-driven proton pump into a light-gated ion channel. *Sci Rep* **5**, 16450 (2015).
95. Buhl E, *et al*. Assessing the Role of R120 in the Gating of CrChR2 by Time-Resolved Spectroscopy from Femtoseconds to Seconds. *J Am Chem Soc* **145**, 21832-21840 (2023).
96. Kubo M, *et al*. Role of Arg123 in light-driven anion pump mechanisms of pharaonis halorhodopsin. *Photochem Photobiol* **85**, 547-555 (2009).
97. Louca S, Mazel F, Doebeli M, Parfrey LW. A census-based estimate of Earth's bacterial and archaeal diversity. *PLoS Biol* **17**, e3000106 (2019).
98. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci* **113**, 5970-5975 (2016).
99. Shen W, Lees JA, Iqbal Z. Efficient sequence alignment against millions of prokaryotic genomes with LexicMap. *Nat Biotechnol*, 1-8 (2025).
100. Nagata T, Inoue K. Rhodopsins at a glance. *J Cell Sci* **134**, (2021).
101. Morizumi T, *et al*. X-ray Crystallographic Structure and Oligomerization of Gloeobacter Rhodopsin. *Sci Rep* **9**, 11283 (2019).
102. Brown LS. Fungal rhodopsins and opsin-related proteins: eukaryotic homologues of bacteriorhodopsin with unknown functions. *Photochem Photobiol Sci* **3**, 555-565 (2004).
103. Sharma AK, Spudich JL, Doolittle WF. Microbial rhodopsins: functional versatility and genetic mobility. *Trends Microbiol* **14**, 463-469 (2006).
104. Zabelskii D, *et al*. Structure-based insights into evolution of rhodopsins. *Commun Biol* **4**, 821 (2021).

105. Sasaki J, *et al*. Conversion of bacteriorhodopsin into a chloride ion pump. *Science* **269**, 73-75 (1995).

106. Hasemi T, Kikukawa T, Kamo N, Demura M. Characterization of a Cyanobacterial Chloride-pumping Rhodopsin and Its Conversion into a Proton Pump. *J Biol Chem* **291**, 355-362 (2016).

107. Inoue K, Nomura Y, Kandori H. Asymmetric Functional Conversion of Eubacterial Light-driven Ion Pumps. *J Biol Chem* **291**, 9883-9893 (2016).

108. Takeno Y, *et al*. Structural basis for color tuning and passive ion conductance in red-shifted pump-fold channelrhodopsin ChRO24. *bioRxiv*, 2025.2009.2016.675330 (2025).

109. Shionoya T, *et al*. High Thermal Stability of Oligomeric Assemblies of Thermophilic Rhodopsin in a Lipid Environment. *J Phys Chem B* **122**, 6945-6953 (2018).

110. Ikuta T, *et al*. Structural insights into the mechanism of rhodopsin phosphodiesterase. *Nat Commun* **11**, 5605 (2020).

111. Sasaki T, *et al*. Halorhodopsin from Natronomonas pharaonis Forms a Trimer Even in the Presence of a Detergent, Dodecyl-β-d-maltoside. *Photochem Photobiol* **85**, 130-136 (2009).

112. Škrinjar P, Eberhardt J, Durairaj J, Schwede T. Have protein-ligand co-folding methods moved beyond memorisation? *bioRxiv*, 2025.2002.2003.636309 (2025).

113. Heo L, Feig M. Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins* **90**, 1873-1885 (2022).

114. Wohlwend J, *et al*. Boltz-1 Democratizing Biomolecular Interaction Modeling. *bioRxiv*, (2025).

115. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **11**, 739-747 (1998).

116. UniProt C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523-D531 (2023).

117. Keeling PJ, *et al*. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol* **12**, e1001889 (2014).

118. Hasegawa-Takano M, *et al*. Cyanorhodopsin-II represents a yellow-absorbing proton-pumping rhodopsin clade within cyanobacteria. *ISME J* **18**, (2024).

119. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772-780 (2013).

120. Steinegger M, Mirdita M, Soding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods* **16**, 603-606 (2019).

121. Carpenter EJ, *et al*. Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). *Gigascience* **8**, (2019).

122. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. *Nucleic Acids Res* **46**, W200-W204 (2018).

123. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

124. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017).

125. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: A Sequence Logo Generator. *Genome Res* **14**, 1188-1190 (2004).

Figure 1

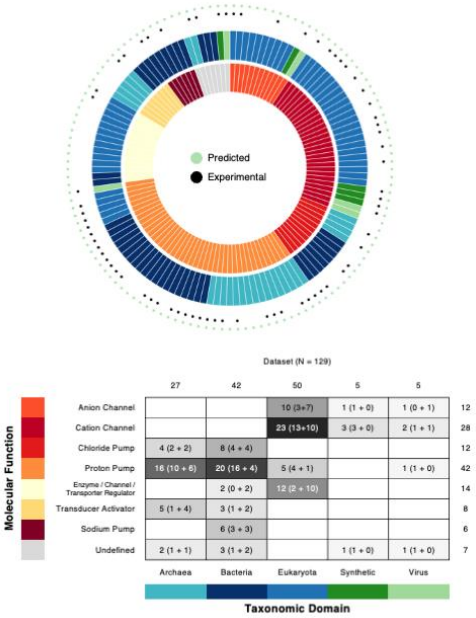


Fig. 1 | **Dataset of microbial rhodopsin structures.** The colors in the inner and outer rings indicate molecular function and taxonomic domain respectively (as shown in the table). Experimental structures and computational models are marked with black and cyan circles respectively. The table shows the composition of the dataset; each cell shows "total number of structures (experimental + predictions)" in that category.

Figure 2

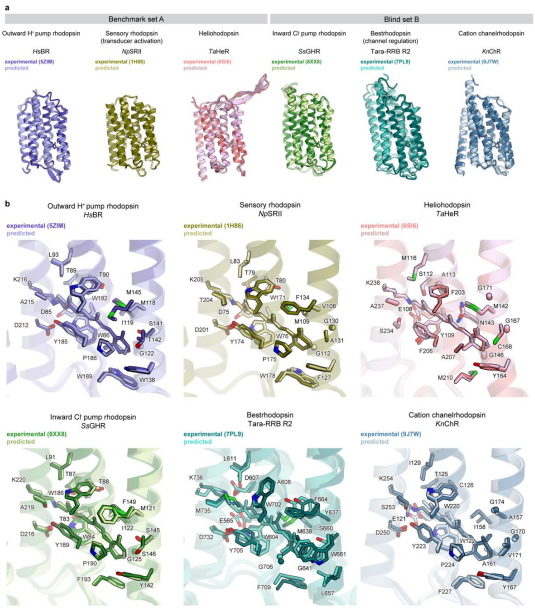


Fig. 2 | Accuracy of the predicted models. a Structural superposition between experimental (solid colors) and predicted (light colors) structures of representative rhodopsins: from benchmark set A, the outward proton pump rhodopsin *Hs*BR (PDB ID: 5ZIM), the transducer-activating rhodopsin *Np*SRII (PDB ID: 1H86), and the functionally uncharacterized heliorhodopsin *Ta*HeR (PDB ID: 6SI6); and from blind set B, the inward chloride pump rhodopsin *Ss*GHR (PDB ID: 8XX8), the channel-regulating rhodopsin Tara-RRB (PDB ID: 7PL9), and the cation-conducting channelrhodopsin *Kn*ChR (PDB ID: 9J7W). **b** Detailed view of the retinal-binding pockets. Transmembrane segments are depicted as translucent ribbons and retinal and protein side chains as solid sticks. TM 5, 6 and 7 are omitted for clarity.

Figure 3

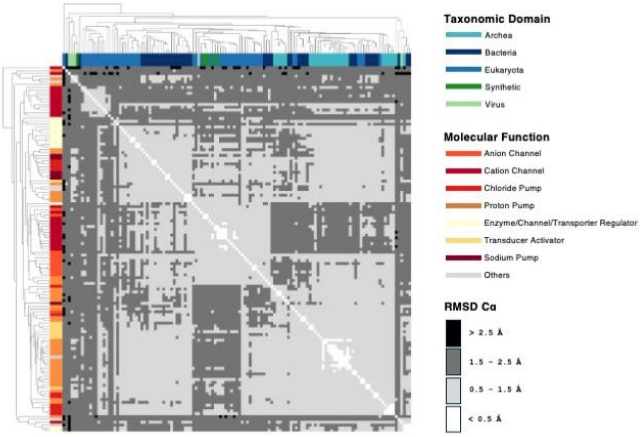


Fig. 3 | Structural landscape of microbial rhodopsins. The heatmap displays pairwise RMSD Cα values for all 129 dataset members. Rhodopsins are sorted according to hierarchical clustering based on structural similarity. Annotation bars display taxonomic domain (x-axis; green/blue) and molecular function (in the y-axis; orange/red). The greyscale bar indicates the RMSD values (light grey = 0.5-1.5 Å, dark grey = 1.5–2.5 Å, black ≥2.5 Å).

Figure 4

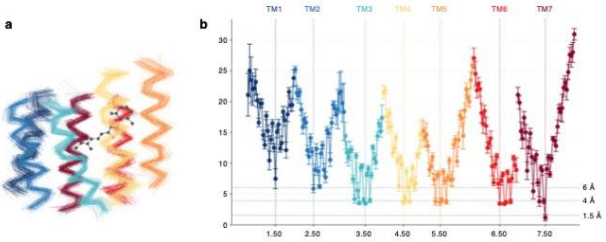


Fig. 4 | Location of the MO numbering anchor positions. a Structural alignment of the 129 dataset members allows selecting the consensus residues closest to retinal in each helix. **b** Average minimum distance from each structurally equivalent residue position to the retinal chromophore calculated across the dataset (closest side-chain atom, excluding hydrogen; an analogous graph for Cα distances is shown in **Supplementary Fig. 2a**). Vertical lines indicate the defined X.50 anchor position for each helix (1.50–7.50). Horizontal lines indicate key interaction distances: 1.5 Å approximates a covalent bond length and is relevant for the Schiff base linkage; 4 Å represents a typical cutoff for non-covalent interactions, delineating the direct binding pocket; and 6 Å encompasses the broader binding-site environment and is the radius used for iRMSD calculations.

Figure 5

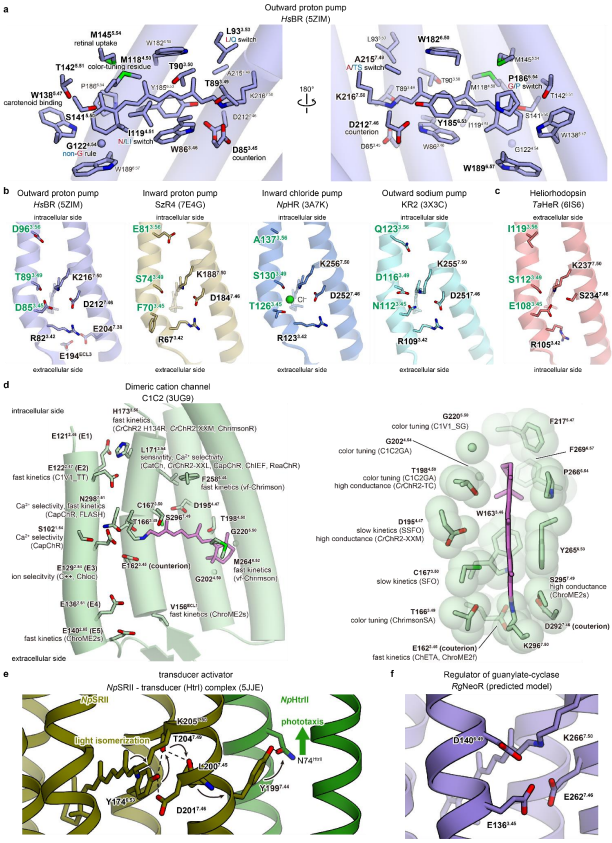


Fig. 5 | Demonstration of the MO numbering system. a Retinal-binding pocket of the outward proton pump *Hs*BR (PDB ID: 5ZIM), shown from two different orientations. Residues constituting the retinal-binding pocket are depicted as sticks. Bold labels indicate residues positioned on the front side of the retinal chromophore. Annotations below each label indicate the functional role of each residue. Blue and red annotations denote shifts toward longer and shorter wavelengths, respectively. **b** Schiff-base vicinity of representative ion-pumping rhodopsins: the outward proton pump *Hs*BR (PDB ID: 5ZIM), the inward proton pump SzR4 (PDB ID: 7E4G), the inward chloride pump *Np*HR (PDB ID: 3A7K), and the outward sodium pump KR2 (PDB ID: 3X3C) (left to right). Only TM3 and TM7 are shown as ribbons for clarity. The residues displayed correspond to functionally important positions. **c** Schiff-base vicinity of the heliorhodopsin *Ta*HeR (PDB ID: 6IS6). In panels b and c, the residues of the DTD motif are highlighted in green. **d** Distribution of functionally important residues identified across various dimeric channelrhodopsin variants, mapped onto the overall structure (left) and the retinal-binding pocket (right) of C1C2 (PDB ID: 3UG9). Variant names associated with each mapped residue are shown adjacent to the corresponding positions. Annotations below each label indicate the functional properties or engineered mutations reported for that variant. **e** Transducer-activating rhodopsin *Np*SRII (khaki) and its cognate transducer *Np*HtrII (green) from the archaeon *Natronomonas pharaonis* (PDB ID: 5JJE). Residues critical for signal activation are shown in stick representation. TM1 is omitted for clarity. **f** Three key residues implicated in long-wavelength absorption in the guanylate cyclase–fused rhodopsin *Rg*NeoR. The structure shown is a model predicted using Boltz1.

Figure 6

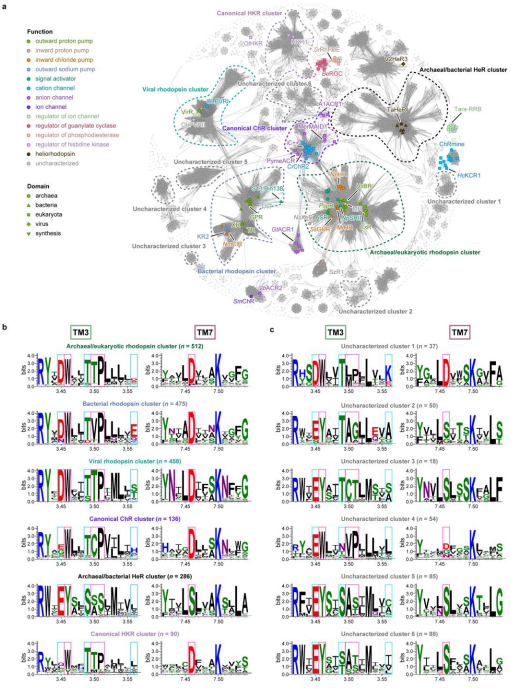
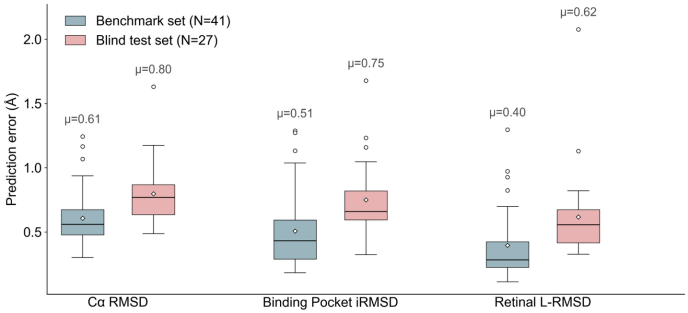


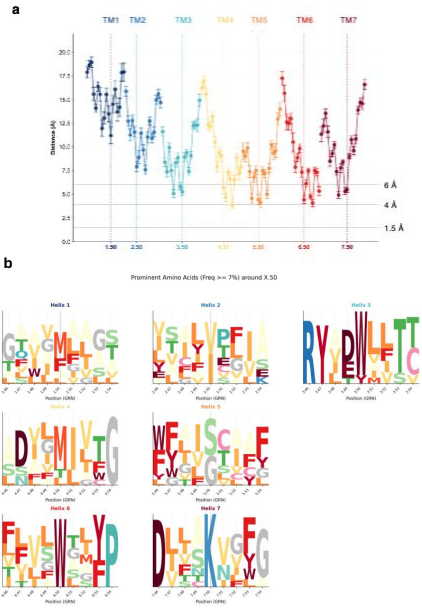
Fig. 6 | Application of the MO numbering system to a large dataset of microbial rhodopsins. a Sequence similarity network composed of approximately 40,000 non-redundant microbial rhodopsin sequences, including a large collection of uncharacterized rhodopsin-like sequences. Six major clusters containing functionally characterized rhodopsins and six of the 17 clusters containing only uncharacterized rhodopsins are highlighted with dashed circles. Representative rhodopsins are mapped onto the clusters, and the colors and shapes of the labels indicate function and domain, respectively. **b** Sequence logos of TM3 and TM7 for six major clusters containing functionally characterized rhodopsins and six of the 17 clusters containing only uncharacterized rhodopsins. Cyan rectangles indicate residues of the DTD motif and magenta rectangles indicate highly conserved residues within each cluster.

Supplementary Figure 1

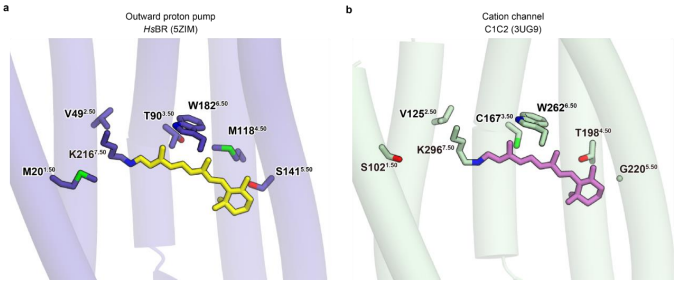


Supplementary Fig. 1 | **Accuracy of the predicted models.** Prediction errors for the two validation sets: benchmark set A (blue; n=42) and blind set B (red; n=27 for backbone, n=26 for the binding pocket). The prediction errors are presented separately for the entire structure (global Cα RMSD), for the binding pocket (iRMSD), and for retinal (L-RMSD). Means are shown over each box. Only four predicted models in benchmark set A showed an error larger than 1 Å in the binding-pocket residues (iRMSD): TR (PDB ID: 5AZD); *Ns*XeR (PDB ID: 6EYU); Chrimson (PDB ID: 5ZIH), and XR (PDB ID: 3DDL). Four models in blind set B also exceeded this threshold: *Er*NaR (PDB ID: 8QLF) and three potassium channelrhodopsins (8H86, 8H87, and 8JXO). The most challenging cases were in blind set B: B1ChR2 (PDB ID: 8JXO, pocket iRMSD 1.68 Å, L-RMSD 2.08 Å) and Tara-RRB (PDB ID: 7PL9), a rhodopsin functionally coupled to the bestrophin chloride channel (excluded from pocket analysis).

Supplementary Figure 2

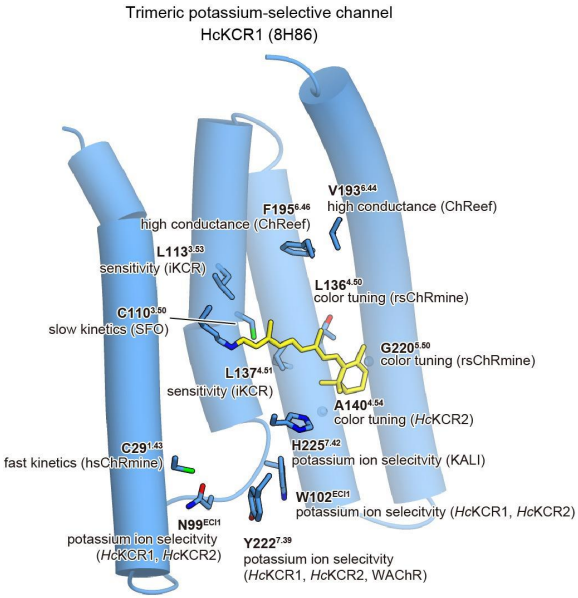


Supplementary Figure 3



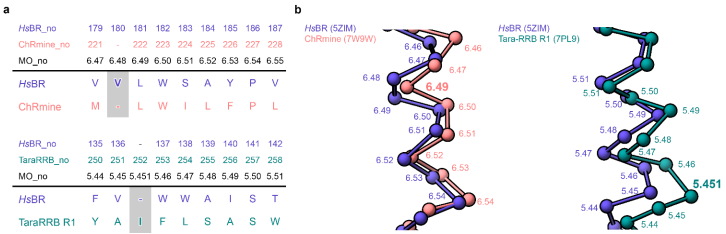
Supplementary Fig. 3 | Examples of x.50 residue positions. The x.50 residues are mapped onto the structures of **a** the representative outward proton-pumping rhodopsin *Hs*BR and **b** the cation-conducting channelrhodopsin C1C2.

Supplementary Figure 4

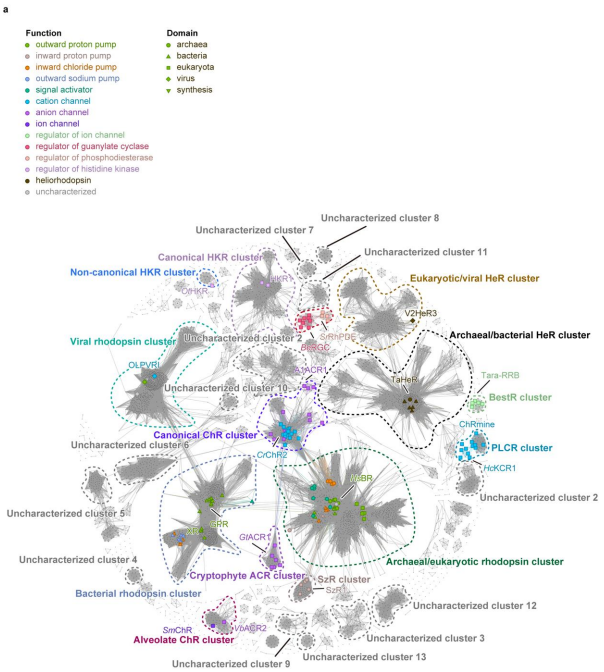


Supplementary Fig. 4 | Distribution of functionally important residues identified across various trimeric channelrhodopsin variants. Residues are mapped onto the overall structure of the potassium-selective channelrhodopsin *Hc*KCR1 (PDB ID: 8H86).

Supplementary Figure 5

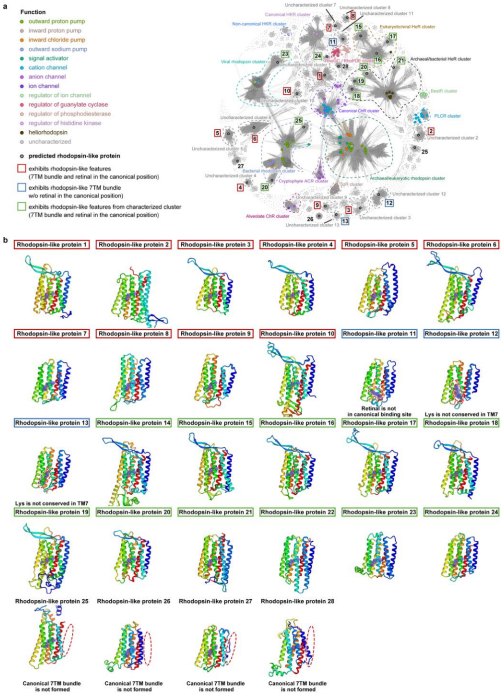


Supplementary Figure 6



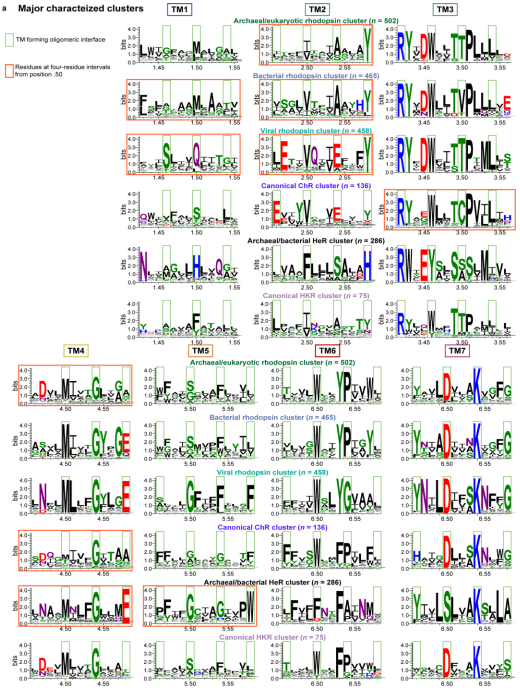
Supplementary Fig. 6 | Sequence similarity network of microbial rhodopsins. Sequence similarity network composed of approximately 40,000 non-redundant microbial rhodopsin sequences, including a large collection of uncharacterized rhodopsin-like sequences. Six major and eight minor clusters containing functionally characterized rhodopsins, as well as 13 clusters containing only uncharacterized rhodopsins, are highlighted with dashed circles. Representative rhodopsins are mapped onto the clusters, and the colors and shapes of the labels indicate function and domain (biological taxonomy), respectively.

Supplementary Figure 7



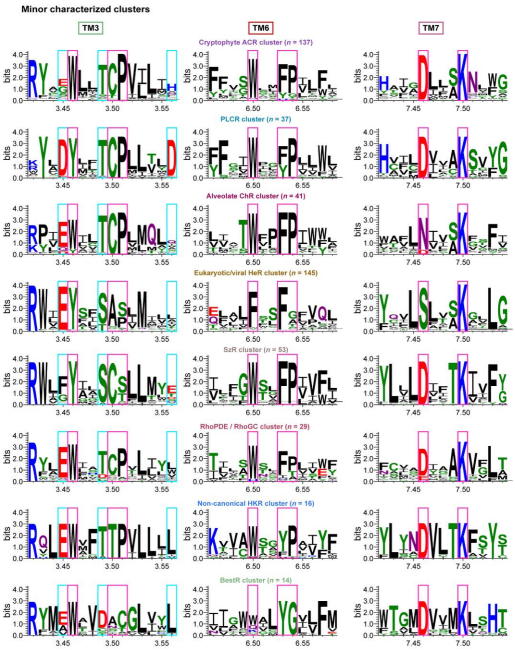
Supplementary Fig. 7 | a Sequence similarity network composed of approximately 40,000 non-redundant microbial rhodopsin sequences, including a large collection of uncharacterized rhodopsin-like sequences. Six major and eight minor clusters containing functionally characterized rhodopsins, as well as 13 clusters containing only uncharacterized rhodopsins, are highlighted with dashed circles. Representative rhodopsins are mapped onto the clusters, and the colors and shapes of the labels indicate function and domain (biological taxonomy), respectively. In addition, 28 rhodopsin-like sequences selected from the peripheral regions of characterized and uncharacterized clusters are highlighted. **b** Predicted structures of the 28 rhodopsin-like sequences selected from the peripheral regions of characterized and uncharacterized clusters. Four of the 28 sequences (25–28) exhibit a non-rhodopsin fold, one (11) shows non-canonical binding of the retinal chromophore, and two (12 and 13) lack the conserved lysine on TM7 that covalently binds the retinal chromophore.

Supplementary Figure 8



Supplementary Fig. 8 | Sequence logos of TM1–TM7 for the six major clusters containing functionally characterized rhodopsins. Transmembrane helices forming oligomeric interfaces and residues at four-residue intervals from position x.50 are highlighted with orange and green rectangles, respectively.

Supplementary Figure 9



Supplementary Fig. 9 | Sequence logos of TM3, TM6, and TM7 for eight minor clusters containing functionally characterized rhodopsins. Cyan rectangles indicate residues of the DTD motif, and magenta rectangles indicate highly conserved residues within each cluster.

Supplementary Figure 10