Domain insertion scanning to study and engineer ion channels

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Willow Coyote-Maestas

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Daniel Schmidt

January 2021

ACKNOWLEDGEMENTS

I am deeply grateful for all the supportive people, communities, and mentors I've been fortunate to be surrounded by during my PhD.

I am grateful for my thesis committee: Laurie Parker (chair), Aaron Goldstrohm, Mikael Elias, and Sivaraj 'Shiv' Sivaramakrishnan. Laurie, was an inspiration with how to balance being a wonderful human, excellent scientist, and maintaining other parts of one's life. Aaron, I appreciate for helping to make it possible for me to spend the end of my PhD with my family. Mikael, I enjoyed TAing for (and Romas) and am so grateful for all the wonderful science conversations. I deeply regret not being able to collaborate with Mikael though we sure talked about it! Shiv, was always rigorous yet kind and I appreciated always being near his office. I will always cherish when Shiv and I sat peacefully watching the sunset up at Itasca one year.

My primary advisor, Daniel Schmidt, gave me so much it is hard to fathom what to high light. Daniel was always willing to make time to talk science either high level abstract ideas that made my head hurt down to figuring out neat experimental pipelines. Daniel pushed me to move in new directions that neither of us had experience, which led to me learning how to pick up new science skills and delve into new fields rapidly. This also helped put me in positions to take new approaches to old fields. When I started, I had little idea what I was doing and Daniel always believed in me despite my lack of background. Daniel help keep my feet on the ground while pushing for us to try sometimes impossible goals. All this helped me grow as a scientist and thinker. Daniel is brilliant and rigorous and I am so grateful to have from this to try to do the absolute best science I can while being grounded in statistics and physical principles. I will always be grateful that he let me move across the country at the end of my PhD to be close to my family and be able to balance my personal and professional needs. He always supported me in taking time off when I needed it to process what was going on in my personal life. Daniel gave me an amazing balance of support when I asked for it and freedom to explore. We would workshop ideas passing back and forth a white board marker sketching out ideas until we got to something that made sense which was immense fun and a delightful challenge. Daniel is an amazing scientist and I feel so fortunate to have learned from him.

I thank my Bioinformatics and Computational Biology MS advisor, Chad Myers, who was essential to the success of the my project. I greatly enjoyed digging into data with Chad and his willingness to provide helpful guidance to the complete machine learning and data mining novice that I was when I started. I appreciated his generosity, kindness, and warmth. It was an immense to privilege to work with such an excellent computational biologist and human.

I also thank Margaret 'Meg' Titus who while not formally an advisor or on my committee was integral to part of my PhD. From the first week of Itasca to the last weeks of my PhD Meg has always had my back. Meg's door was always open and she would give me tissues when I needed to cry, cheer me on when I needed it, and give me rigorous but kind feedback. Meg read every fellowship application, paper, and most recently helped coach me for important talks. Her feedback was always fair, honest, and aimed to help. Meg is such a generous, compassionate, and kind mentor and advocate that I was so fortunate to be around. There were several challenging moments when I considered dropping out due to challenges in my personal life and it's due to Meg's support that I did not. I hope I can learn from Meg's dedicated mentorship and boundless generosity.

I am grateful for James 'Jaime' Fraser who is another informal yet crucial mentor from my PhD. I met Jaime at a conference, we hit it off, and when I needed to find a science home in the SF bay area to be closer to family, he offered to host me. Jaime is the mentor, human, and scientist I hope to one day be; he is a fearless advocate for his students, an amazing human for his family, and a brilliant rigorous scientist with broad interests and pursuits. In the last year, Jaime has helped me grow as a scientist, pushed me to reach for lofty ambitions, taught me how to tell a good story, and welcomed me into the UCSF and broader bay area science community. Jaime has opened so many doors for me that are letting me move a lot faster in my science and career than I thought possible. Jaime has helped me in so many ways and I hope to help others to pass it on.

I also want to thank the entire Schmidt lab for being such a wonderful community to spend the last few years. I enjoyed how we all learned to do science together and would help each other out whenever needed. We are a generous, rigorous, and brilliant group who were all striving to do the best science possible. David Nedrud and I started at the same time as the first graduate students and eventually our paths converged. I've been so grateful for the opportunity to collaborate with David without whom I wouldn’t have been able to do such expansive and cool science. Beyond that we learned (through some hiccups) how to work together and balance needs which will be integral to future teamwork. David was also essential for me moving to the bay area; he offered to carry on the experimental torch and prepare when I returned for a productive and hectic week in the lab. Overall, I am deeply indebted and grateful for everything David has done and I feel fortunate to have worked with him. Yungui He was essential to my and the lab’s success. Yungui provided the foundation and support that let us focus on the cool parts of science. Beyond that Yungui was always so generous and kind; he played a fundamental supportive role throughout my PhD. I enjoyed working with Yungui and am stunned by his generosity. Alina Zdechlik was and remains one of my closest friends throughout my PhD and beyond. I feel fortunate to have worked with her all these year’s even if we sometimes pushed each other’s buttons or we never really got to collaborate as much as I hoped. Alina was a fundamental supporting person and someone I always enjoyed sharing my and her successes with. We also each pushed each other to be more rigorous in our science though perhaps I was sometimes too nitpicky when it comes to writing. Steffan Okorafor I was grateful to work with as an undergraduate mentee. Working with him taught me how much growing I have to do in this space and hopefully helped me grow a bit. I loved seeing his successes, helping him through failures, and learning how to provide guidance as he progressed. Steffan is such a wonderful and kind person who always lit the lab up with his laughter and good spirit. I cannot wait to see where all of us land once we finish our PhD.

Nearly my entire PhD involved flow cytometry and as such my second lab home was the Flow Cytometry Core. I am grateful to Therese ‘Terri’ Martin, Jason Motl, and Rashi Arora. They were tremendously accommodating and integral to my success. I enjoyed the meandering conversation I had with Terri about life, traveling, science, and being raised by hippies. Jason was always incredibly generous, helpful, and attentive. Rashi was a wonderful addition to the Flow Core near the end of my PhD. She put up with me trying all sorts of crazy experiments – some of which were abject failures. I always came with too many samples, a little late, and somewhat bedraggled and they helped me out. I enjoyed getting to know all of them and working with them. I am scared of how I will continue to do experiments without the UMN Flow Core but hopefully I will manage.

I want to thank the entire Fraser lab for helping me feel at home in San Francisco and at UCSF. I spent much of the Pandemic on remote zoom calls with the Fraser lab. If I had to be stuck inside, only connecting over zoom, and living in fear of a deadly pandemic, there are few groups of people I would rather be stuck together-apart with. Specifically, the DMS crew has been a great group of people to bounce ideas off of. Especially Gabriella Estevam always was open to hearing about my analysis, provided me helpful comments, and I enjoyed (as hopefully she did) helping her design her deep mutational scanning experiments. The whole lab provided lots of feedback as I developed my last manuscript (and chapter of my thesis) that I feel helped me take the project to the next level. I also want to thank Roberto Efrain-Diaz for welcoming me to the Bay Area, being a JEDI (Justice, Equity, Diversity, Inclusion) role model and pushing me to be more of a JEDI, becoming a fundamental part of my support network, and becoming one of my dearest friends. Everyone else in the Fraser lab has been wonderful and I am stunned by what brilliant and wonderful scientists and humans they are.

I would never have made it to graduate school if it were not for undergraduate mentors. I thanks and acknowledge all the professors at Evergreen State College who told me I could go as far as I wanted in science. In high schools I had active discouragement from science teachers that made it challenging for me to consider pursuing science. At Evergreen I was for the first time seen as a potential scientist and told that grad school pays you! In particular, I must thank Marty Beagle for telling me I could and not to lose my style, Erin Ellis/Martin for helping me get into a lab at University of Washington, Paula Schofield for being such an amazing supportive mentor, Lydia Mckinstry for being supportive yet rigorous, and Andrew Brabban for giving me an amazing opportunity to help with a lab-based course. I also want to acknowledge everyone at University of Washington that gave me my first foray in real research science, an experience at conference where I got to be a part of ideas in development, and who showed me I could really do science. In particular, I want to thanks Katherine Heal for being an amazing graduate student to work under, Laura Truxal Carlson with whom I had so much fun working with, and Anitra Ingalls who was always supported me throughout my lab work and beyond.

I want to acknowledge the MCSB graduate students who I was fortunate to spend my time with. Especially Daniel Sorenson who is one of my closest friends who I enjoyed every deep into the night somewhat inebriated foray into diverse topics from social justice, challenging science, architecture, literature, politics, and our personal lives. Dan is one of my dearest friends and I am so grateful to have found someone who shared some many similar interests and such a sharp wit to discuss. Alina and David who I mentioned before but were integral fellow graduates students. Zac Shreiber who started out as my arch rival (shared birthday and similar research interests) but became one of my dearest friends with whom I enjoyed many meals and walks. I learned from Zac how for some the research path is not ideal and there are graceful ways to find a new path. For Zac leaving the program was an amazing step up for his life (and income) that I found inspiring. Ryan Nasti, Collette Rogers, and Elizabeth Fay were all dear friends with whom I am so happy to have spent this time in addition to the rest of the community.

I was also fortunate to help revitalize the local SACNAS chapter with many inspiring students from many fields. In particular I want to acknowledge William ‘Ty’ Frazier, Becky Rodriguez, Jonas Alvarez, and Yectli Huerta. Yectli was the intrepid and fearless leader we needed and Becky was the excitable follow up that continued growing SACNAS. Jonas (also MCSB) I enjoyed spending time with and always checking in. Ty also became one of my favorite people to spend time with; I find his path and backgrounds to be quite the inspiration. He is a sharp math mind who is an advocate for all people regardless of backgrounds and does not sacrifice any of his personality in the name of academic professionalism.

In the middle of my PhD I was fortunate to become a HHMI Gilliam fellow. This program, the community, and the doors this opened has been one of the most positive experiences of my life. The Gilliam fellow community is made up of brilliant yet fearless advocates for change in science. My eyes have been opened through this experience and I’ve been empowered to make a difference. The change that Gilliam fellows are making right now is changing the shape of science. As Gilliam fellows we take part in workshops, which empower us to be truer to ourselves, acknowledge where we can grow, and then help us give us skills to succeed. We also shared our deepest most challenging problems. It’s from these workshops that I felt empowered to make a change that let me be closer to my family. Many of the Gilliam fellows have become my closest friends and it has made me feel like I don’t have to choose between being a human with a life and being a successful academic scientist. I am so grateful to the leaders of the Gilliam program, all the students, and HHMI overall. The Gilliam community builds space and community for marginalized people in science and I love seeing how we support each other.

I want to thank the scientific fields and all the amazing people from professors to trainees that I had the good fortune to meet and spent time around. I was humbled and inspired at the Protein Science and Ion Channel Gordon Conferences especially. I made good friends, struck up collaborations (many more discussed then actualized), and was so happy to find audiences for my work. I've met many close friends at these conferences, and I've enjoyed seeing all us trainees continue to grow and move on to our next steps. There are too many people who have provided me advice, feedback, support, and inspiration to note. I cannot wait to spend my scientific career with these fantastic communities.

Outside of science I was fortunate to have several dear friendships. In particular, I want to acknowledge Julia Huggins, a friend of a friend from undergrad, who let me move in when I first moved to Minnesota and who welcomed me to Minnesota. We lived together for my first year and introduced me to all of her friends. Julia is one of my closest friends now and she has supported me through some of the most challenging chapters of my life. Patrick Hertz was also a fundamental part of my support network (who Julia introduced me to) who I spent many good beers, walks, conversations, and so much more. There are many other dear friends who were also important such as Lauren Pagel and Gina Emon.

I must thank and acknowledge Mia Owen. Mia is my former partner and one of my dearest friends. Mia gave me so much support, love, good times, and so much more through such challenging times in my life. Writing has always been a struggle for me and I was fortunate she is an excellent writer and generously gives of feedback (if somewhat sarcastically). Mia taught me the ways of the word, oxford commas, and how to organize a sentence. Without these practical skills, I would've struggled even more to communicate my science. We spent so many wonderful bike rides, meals critiquing food, making extravagant dishes that took way too long, playing games, reading, talking, exploring the world, music and ideas, and so much more. Mia and her family welcomed me and always made me feel like I had a home in Minnesota. Without Mia I would have felt completely lost in my life and PhD; she gave me an anchor and support.

Beyond Minnesota, I received a tremendous amount of support from my dearest friends from childhood, undergrad, and beyond. I am grateful for Cady Watts always picking up the phone when I was in dire straights. Cady is amongst the most supportive people and we are always there for each other and in my PhD that has been no different. Daniel Cherniske similarly for long phone conversations and giving me supportive and loving words. I thank Aidan Haberstroh for always willing to drop in and go for walks when I was back around my hometown. I also thank Lais Carvalho for being always a WhatsApp chat away as we commiserate on our lives. I acknowledge Jesse Pledger for being a supportive friend and always willing to go on an adventure either him coming to Minnesota or me going to New York City to get into shenanigans and music. I also thank Tom Finch for always checking in on me, sending me love, and playing a crucial support role from afar by phone.

I also want to acknowledge my family. I acknowledge all my siblings. Especially my little brother Tekuani Maestas who I supported and vice versa as we were in unfamiliar environments. Him in a high school on Pine Ridge and I in graduate school in Minnesota. I thank Scott Valentino for supporting me through challenging times and making me feel at home on my return to California.

These last couple of years have been such a trying time for my mother's side of my family. I am grateful that because of this we have become closer and supported each other accordingly. I must thank my mother, Tara Coyote, for well… Everything. I would literally not be here today without her. I was always taught to follow my dreams and do whatever it is I want. She always gave me immense freedom as a child free from any expectations, which turned me into a self-motivated person. She has always been my biggest champion and cheered me on from the sidelines regardless what I did. I need to thank her for so many many many many many different parts of me that has helped me to succeed. I grew up in diverse environments and moving a lot, which taught me how to quickly absorb information and find a new way of being and thinking. From this I have learned how to enter new scientific fields and quickly make community within these fields. This has made it so I have science (and otherwise) communities across disciplines and the world. Her cancer diagnosis and all the fear that came along with it taught me to so much compassion and gratitude for the people in my life. This has taught me to try to be more compassionate towards others, value them, and priotize human being wellness, which will benefit me as a mentor. I am so grateful for everything my mother gave me intentionally and otherwise that make me who I am and by extension help me my the science as I do.

I also thank my grandparents, who I talked to nearly every weekend of my PhD. They are always so excited about my progress (sometimes more than me). Their boundless support and excitement has been an inspiration and helped continue to motivate me when I am feeling challenged by my circumstances. My grandfather was the father I never had growing but and became an inspiration to me. He was an ecology teacher, which made it so I was always surrounded by the big pictures of biology. While I am interested in the molecules of life, constantly hearing about the wonders of nature helped kindle my fire.

I thank my uncle, Ted Cheeseman, who has always been an intellectual beacon in my life. He taught me long division when I was 5 and always was willing to discuss the world around us. Now we have a playful rivalry about who will get my PhD first… As I am writing this it is obviously me! In all seriousness though I am quite grateful for my uncle and his help at many different stages of my life.

Overall, I have been and continue to be supported by so many people that I can only list smallest portions of contributions and the sources from friends, mentors, family, and communities that have been integral to my success in graduate school. I've already clearly gone quite long…

DEDICATION

Over the course of my PhD, I had a series of very challenging losses and challenges. I lost one of the dearest people in my life, Grayce Forsythe, within several months of starting. This was the most challenging experience of my life thus far. My mother, Tara Coyote, was diagnosed with cancer 6 months after. The following year after my aunt Rowan Holland passed away and soon after a dear childhood friend, Dylan O'Connor, passed away. Near the end of my PhD my mother's cancer metastasized and soon after my Grandmother, Gail Cheeseman, was diagnosed with a later stage and very aggressive lymphoma. Fortunately, through modern therapeutic miracles my mother and grandmother are both stable. All this hardship and the challenges associated with living in Minnesota without close ties led me to dive deep into my research. My research was my refuge from the chaos of the rest of my life. My PhD is dedicated to all these people and the challenges that inspired me to dig into my research. This thesis is also dedicated who all the loved ones who've helped support me in these challenging years.

ABSTRACT

New proteins primarily evolve through recombining modular protein domains with discrete structure and function. Often these recombination involve combining a catalytic function with a sensing domain, so protein function can be regulated by different stimuli. This form of domain recombination-based evolution underlies the intricate signaling networks that allow our cells and by extension our bodies to sense and respond to stimuli. Protein engineers mimic nature by combining domains with desirable properties into new useful combinations never seen in nature. This approach for generating synthetic multi-domain proteins has yielded groundbreaking therapeutics and tools for biology, most notably Car-T cancer therapies and GCaMP calcium sensors. However, domain recombination is challenging and requiring years of iterative optimization but unlike evolution we don’t have millions of years to spare. Both, our basic understanding of the biophysical principles of how proteins evolve and how to better engineer proteins are limited by a lack of domain compatibility rules. In the work presented in this thesis, we sought out to apply massively parallel domain insertion experiments and learn rules for domain compatibility. As our target protein, we used ion channels as they are an attractive engineering target and ion channels evolved through extensive domain recombination. Initially we started with a small set of 3 inserted domains inserted into all amino acid positions of a potassium channel kir2.1. We successfully engineered a light-switchable potassium channel that could be used by neuroscientists, however we found a tremendous amount of variability that necessitated expanding out to a broader sample of domain recombination space. Before we could achieve this goal, we needed to improve experimental pipelines because the methods the domain insertion field used at the time were not scalable nor generalizable. We developed a new domain insertion library generation method, SPINE, that yielded near perfect libraries. SPINE allowed us to expand out to over 700 different inserted domains with which we exhaustively sampled insertional space and developed a mechanistic model of domain recombination. We then expanded outward to several additional recipient channels to benchmark our work. Overall, we made major strides towards the goal of a mechanistic model for assembling protein domains. We expect this body of work will provide a foundation that will make domain-based engineering more effective and improve our understanding of the fundamentals of how proteins evolve.

Table of contents

List of Tables

List of figures

Chapter 1

Introduction

Proteins underlie nearly every part of our biology. Proteins let us move through the world by way of myosins-actin ratcheting (1), help us see so you can read this thesis through sensory GPCRs and other receptors (2), and underlie how we think so I could conceive of this thesis through ion channels working together to create electrophysiology (3). Proteins can play diverse roles in physiology by being immensely adaptable. Protein engineers take advantage of protein adaptability and functional diversity to build useful catalysts (4), tools (5), and therapeutics (6). By building with proteins, we learn the fundamental biophysical and biochemical rules that biology follows through evolution (7). At the most basic level, proteins are polymers made up of 20 different amino acids which encode structure and function. However, the biophysical principles that govern amino acids sequence giving rise to structure and function are idiosyncratic and remain elusive (8). Recent innovations in DNA sequencing and synthesis combined with clever ways to couple sequencing to protein function enable testing massive libraries of variants to learn biophysical rules and better engineer proteins through brute force (9)(10)(11). In my PhD, I combined biophysics, biochemistry, genomics, and computer science to develop high throughput pipelines to enable massively parallel experiments and applied these pipelines to study and engineer ion channels(12)(13)(14). We studied the fundamentals of protein structure-function (14,15), developed methods for coarse-grained structural biology (14), learned rules for assembling proteins together for engineering (14), and developed switchable tools for neuroscientists (13).

Protein evolution

There is a tremendous diversity of different types of proteins that underlie the many ways we interact, perceive, and exist in the world. All these proteins evolved to allow life to fill all the ecological niches on earth and continue adapting to changing environments. New proteins rarely evolve completely *de novo* or from random sequences because neutral drift (random nondeleterious mutations) is slow and functional proteins are rare (16). For example, from a random pool of peptides only 1 in 10^{11} number of proteins were found to have function (17). Instead of *de novo* evolution, proteins evolve by repurposing existing protein sequences for new functions (16). The generally accepted model for protein evolution is gene duplication and divergence where the function of a protein can be maintained while a new protein can diversify through random mutations to perform a new biological function and fill a selectable niche (18). This divergence occurs through point mutations that make small changes to protein functions, for example mutations that tune enzymatic catalytic rates, substrate binding affinities, or stability (19).

Protein domains in evolution and engineering

Mutation-based evolution is incremental as it requires the accumulation of many mutations to evolve completely new functions. A more successful strategy that emerged to allow rapid evolution of cell signaling networks is to recombine existing modular protein domains with discrete structures and functions. For example, the SH2, SH3, and PDZ peptide recognition domains are recombined with kinase and phosphatase domains for targeted phosphorylation-based regulation of other proteins (20). By combining a binding domain to different functional domains, we get multi-domain proteins that play different roles such as phosphorylating or dephosphorylating the same target proteins. It is from domain recombination-based evolution that we get the intricate and beautiful protein interaction networks that make up our diverse biology.

Protein engineers mimic nature by recombining existing protein domains to make tools and therapeutics. For example, parts of T-cell receptors and antibodies were combined to generate revolutionary CAR-T cell cancer therapies (21). In CAR-T therapies a designed protein is engineered into a patient's immune cells that targets a cancer cells based on surface exposed epitopes. Basic biology also benefits from these approaches. The most famous example are a class of fluorescent reporters for calcium, GCaMP, which combines parts of a fluorescent protein, green fluorescent protein (GFP) with a calcium sensitive protein, calmodulin (22). GCaMP enables observing neuron activation which allows whole new types of live animal experiment that combine behavioral assays with electrophysiology observation. Overall, protein domain-based engineering is a useful way to engineer useful tools and therapeutics.

There are hundreds of protein domains that can be recombined to generate new multi-domain proteins (23). Beyond natural proteins, protein designers are expanding the protein domain universe through building synthetic proteins (11)(24). These synthetic proteins are designed to bind ligands (25), other proteins (26)(27), or generate new folds (11,24). While the geometries of these in solution by themselves are becoming designable alone, assembling is not straightforward because when domains are combined protein structure and stability changes (28). This phenomenon makes it hard to know how a domain's fold in solution will behave after recombination into a multi-domain protein. Due to these gaps-in-knowledge, we are stuck with trial and error and iterative optimization to get tools that work. Understanding how to assemble protein domains would solve this problem and be massively impactful by helping multi-domain protein engineering and yield insight into the fundamental biophysical principles that guide protein evolution.

Massively parallel genotype-phenotype experiments

Since the Human Genome project, the price of DNA sequencing and synthesis rapidly decreased (9). All of biology is impacted by cheap DNA sequencing and synthesis. High throughput protein biology is a nascent field that takes advantage of synthesizing any DNA sequence we want, coupling this to a phenotype, and measuring how different variants behave in a library (10). These 'genotype-phenotype' experiments let us understand how all possible variants contribute to phenotypes. By testing all possible variants, we go beyond nature– allowing us to see which apparent constraints of life are true limits or whether life evolved as it did due to happenstance. High throughput genotype-phenotype experiments are revolutionizing protein biology because previously hard to study complex protein properties can now be studied systematically, such as identifying pathogenetic mutations (29,30), solving protein structures (31,32), protein folding (11), and allosteric regulation (13,14,33,34).

Based on genotype-phenotype experiments, others have developed mechanistic understandings of how mutations alter proteins and how amino acids contribute to disease (35-39). Despite the importance of domain recombination for engineering and evolution, domain insertions and recombination have not been studied systematically. To understand evolution and better engineer tools, we need a set of rules that explain how to assemble protein domains. We cannot base rules from bioinformatic observation because looking at how proteins evolve does not tell us which recombination were tried but didn't work verse those that nature never tested yet could work. We must experimentally test many combinations and learn protein domain compatibility rules. There have been several studies that scanned proteins with one or two domains, but before our work there was no systematic exploration of donor domains and recipient proteins.

Protein switch biology and tools

The intricate signaling networks of cells allow the cell to quickly change states in response to stimuli (40). Many protein switches make up signaling networks to allow signal amplification and propagation (20). Protein switches evolve primarily through the recombination of a sensor domain and with a catalytic domain (20). Engineers learned from nature to engineer synthetic protein switches to control and study cell phenotypes (41). The general approach is to combine a protein that one wants to control and a protein domain that switches in response to an exogenous stimulus, such as light (e.g., LOV2 (42)) or a ligand (e.g., cpDHFR (43) or Unirapr (44)). However, it is hard to know where to insert a switchable domain to generate a switchable protein. An ideal site would not cause disruption of the recipient protein and would be coupled to protein function. Sites coupled to protein function are often within the structured regions of a protein where most insertions are disruptive. However, rational based approaches to identifying switchable sites of a protein have focused on identifying sites that can switch and not on those that allow an insertion (184). As a result, these rational approaches have not found broad use. Ideally, a rational approach would incorporate finding sites that are both switchable and amenable to insertion. Due to the complexity and constraints of proteins to maintain stability and function, it is challenging from typical low throughput experiments to develop a universal engineering framework. In my PhD, we hypothesized this problem could be solved by generating massive libraries and testing for switching and disruption of the recipient protein.

Ion channels

Ion channels underlie the electrophysiology that gives rise to heartbeats and cognition among other important processes (e.g., kidney function, glucose metabolism, and immune responses) (3). Disruption in channel function or expression results in numerous diseases including neurological disorders (45)(46), cardiac diseases (47), and cancer (48). There are hundreds of ion channels which have varied tissue expression and functional properties which work in tandem to give rise to complex electrophysiology (3). Due to the immense complexity, it is hard to know exactly how each channel contributes to physiology.

Ion channel’s diverse roles in physiology and disease and complexity make them desirable targets for engineering switching. Similar to a switchable channel, archaeal rhodopsin ion pumps are widely used in neuroscience to activate or inactivate neurons with light (49)(50).

However, rhodopsin-based tools modulate neural activity in nonphysiologically relevant ways by conducting different ions (51), not having dynamic range (51), and changing baseline electrophysiology (52). It would be better if we could switch the channels that actually underlie our electrophysiology to understand our electrophysiology. For example, if we could express switchable ion channels within specific cell types, control channel function, and observe behavior, we could understand how a specific ion channel contributes to organismal level properties. This would help us develop models of animal behavior based in molecular biology and biophysics. Beyond understanding fundamental biology, switchable channels could be useful to treat ion channel disease. For instance, a channel engineered to respond to a specific small molecule could be expressed and switched on in epileptic patients to limit the overexcitability that causes seizures.

Channels are an ideal model protein to study domain recombination because the ion channels in our genomes are the result of domain recombination-based evolution (53). One superfamily of channels that contains potassium channels, sodium channels, calcium channels, and sensory transient receptor potential channels evolved through recombining a channel pore, transmembrane regions, and various sensory domains such as voltage and ligand sensors.

During my PhD, we developed and applied high throughput domain insertion scanning pipelines to engineer light-switchable ion channels and studied the fundamentals of ion channel structure function and based on our massive datasets we developed a framework to explain domain-recipient compatibility. Surprisingly, we found insertional scanning could be used as a coarse-grained structural biology experimental based method to identify which regions of a protein are involved in folding and which are involved in function. We started out using a simple potassium channel as our model protein but expanded out to four additional proteins to test the generalization of our conclusions on additional proteins. This is all covered in the next three chapters the chapters 2 (13) and 3 (12) are published and peer-reviewed and chapter 4 is on a pre-print server and submitted to a journal (14). Summaries of each chapter are below:

Chapter 2: Domain Insertion Permissibility-Guided Engineering of Allostery in Ion Channels

Allostery is a fundamental principle of protein regulation that remains hard to engineer, in particular in membrane proteins such as ion channels. Here we use human Inward Rectifier K$^+$ Channel Kir2.1 to map permissibility to insertion of domains with different biophysical properties. Here we find that permissibility is best explained by dynamic protein properties, such as conformational flexibility. Several regions in Kir2.1 that are equivalent to those regulated in homologs, such as G-protein-gated inward rectifier K$^+$ channels (GIRK), have differential permissibility; that is, for these sites permissibility depends on the structural properties of the inserted domain. Our data and the well-established link between protein dynamics and allostery led us to propose that differential permissibility is a metric of latent allosteric capacity in Kir2.1. In support of this notion, inserting light-switchable domains into sites with predicted latent allosteric capacity, renders Kir2.1 activity sensitive to light.

Chapter 3: Targeted insertional mutagenesis libraries for deep domain insertion profiling

Domain recombination is a key principle in protein evolution and protein engineering, but unlike single amino acid mutations, inserting a donor domain into every position of a target protein is not easily experimentally accessible. Most contemporary domain insertion profiling approaches rely on DNA transposons, which are constrained by sequence bias. Here we establish **S**aturated **P**rogrammable **In**sertion **E**ngineering (SPINE), an unbiased, comprehensive, and targeted domain insertion library generation technique using oligo library synthesis and multi-step Golden Gate cloning. Through benchmarking to MuA transposon-mediated library generation on four ion channel genes, we demonstrate that SPINE-generated libraries are enriched for in-frame insertions, have drastically reduced sequence bias as well as near-complete and highly-redundant coverage. Unlike transposon-mediated domain insertion that was severely biased and sparse for some genes, SPINE generated high-quality libraries for all genes tested. Using the Inward Rectifier K$^+$ channel Kir2.1, we validate the practical utility of SPINE by constructing and comparing domain insertion permissibility maps. SPINE is the first technology to enable saturated domain insertion profiling. SPINE could help explore the relationship between domain insertions and protein function, and how this relationship is shaped by evolutionary forces and can be engineered for biomedical applications.

Chapter 4: The biophysical basis of protein domain compatibility in ion channels

Understanding the biophysical mechanisms that govern the combination of protein domains into viable proteins is essential for advancing synthetic biology and biomedical engineering. Here, we use massively-parallel genotype/phenotype assays to determine cell surface expression of over 300,000 variants of inward rectifier K^+ channel Kir2.1 recombined with hundreds of protein motifs. We use machine learning to derive a quantitative biophysical model and practical rules for domain recombination. Insertional fitness depends on nonlinear interactions between the biophysical properties of inserted motifs and the recipient protein, which adds a new dimension to the rational design of fusion proteins. Insertion maps reveal a generalizable hierarchical organization of Kir2.1 and several other ion channels that balances stability needed for folding and dynamics required for function.

Chapter 2

Note this was originally published:

Coyote-Maestas W, He Y, Myers CL, Schmidt D. Domain insertion permissibility-guided engineering of allostery in ion channels. Nat Commun. 2019 Dec;10(1):290.

Domain Insertion Permissibility-Guided Engineering of Allostery in Ion Channels

Introduction

Allostery is the phenomenon in proteins where the state of proximal sites is coupled to the state of distal sites. In nature, allosteric regulation is widespread in multidomain proteins, such as plant photoreceptors (54), that arise from recombination of functionally and structurally discrete protein domains. In the lab, we recombine domains to generate synthetic proteins; for example, antibodies that are joined end-to-end with signaling domains to create chimeric T-cell receptors for immunotherapy (21). In both scenarios, how these components become allosterically coupled is essentially trial-and-error. That blind trial and error can progressively lead to optimized design through natural selection over billions of years is a central concept in the evolution of natural systems (55). However, in the lab we need to accomplish this task in less time and with greater efficiency.

One class of proteins that are challenging to engineer rationally are ion channels. Ion channels play critical roles in the biological signaling processes that determine the operation of cells and networks of the brain and the heart and are thus major drug targets (3). Virtually every aspect of ion channel gating relies on allosteric regulation, and many drugs achieve their therapeutic effect through allosteric modulation (56). Being able to engineer the allosteric regulation of ion channels *de novo,* for example as chemo- or optogenetic tools(57)(58), would enable fine-tuned control and thus exploration of how individual channels contribute to cell physiology.

Models of allostery that could aid us in this task have continued to develop since the initial description of allostery as a phenomenon in proteins and the structure of haemoglobin, the prototypical allosteric protein(59). In fact, to reconcile that intrinsically disordered proteins can facilitate long-range allosteric regulation(60), that allostery can occur without structural change(61), phenomena such as ‘negative cooperativity’(62), and agonism/antagonism switching(63), models in which allostery emerges from dynamic (entropic) mechanisms –instead of structurally distinct macroscopic conformations– have long been considered(61).

One framework, the ensemble allosteric model (EAM), unifies the classic Monod-Wyman-Changeux (MWC) and Koshland, Nemethy, Filmer (KNF) models with allostery emerging from intrinsic disorder and conformational fluctuations(64). The EAM model describes allostery as thermodynamic interdependence of a protein's cooperative structural elements, whose intrinsic stabilities are influenced by ligand binding(63)(65). The practical utility of the EAM models for engineering allostery was recently tested by guiding the engineering of a protein switch(66). A defining feature of the EAM model is that allostery does not rely on specific obligatory allosteric pathways in proteins, but instead arises from the energetic balance of all structural elements within the protein – the conformational ensemble. In this view, a protein's primary sequence not only encodes the tertiary structure, but also a protein's energy landscape, which manifests as a protein's conformational ensemble. Any perturbation that modulates the stability of a structural element will affect the stability of all other coupled elements. If allosteric coupling has such degenerate requirements, it is easy to see how a protein, in addition to its primary function, can possess hidden –latent– functions that are not under selection(67)(68). These latent functions could become exploited and facilitated by the same amino acid sequence if selection pressures change. Latent function could be considered byproducts ('spandrels', (69)) of the energy landscape topography. Latent phenotypes are co-opted in numerous biological contexts, including soluble proteins such as enzymes(70) and hormone receptors, (70). As another example, a scaffold protein (Ste5) allosterically regulates Erk-like kinases that diverged before the evolution of Ste5 itself, implying that the allosteric capacity to be regulated was already present at that point (67).

How is the notion of latent allostery relevant to ion channels? The majority (43 out of 45) of human ion channel families appeared in the early metazoan (53), so any subsequent functional diversification could conceivably be the result of leveraging latent regulatory mechanisms that existed in ancestral ion channel clades. However, it is unclear whether (1) ancestral channels used latent pathways to diversify, (2) modern ion channels still possess latent allostery, and (3) whether latent allostery can be leveraged to engineer allosteric regulation into channels.

Approaching these questions from the perspective of the EAM model, we note that the greatest allosteric coupling response is observed when at least one of the involved structural elements is 'poised' to undergo disorder to order transitions (65). A structural element is poised when its intrinsic stability is such that it undergoes local disorder/order transitions. We therefore hypothesized that a good probe for allosteric capacity would be one that is able to determine regional conformational flexibility. We furthermore hypothesized that conformational flexibility would manifest as regional structural plasticity, which can be examined by measuring the permissibility of this region to a domain insertion. That is, regions that are conformationally flexible (fully or partially disordered) are more tolerant to an in-frame insertion of a protein motif (the probe) than regions that are well-structured (ordered). Conceptually, the idea of probing regional or site-specific protein plasticity, and more generally the idea of functionally linking together protein domains through domain insertion and recombination, is well accepted. It has been applied to the engineering of enzymes(44,71,72), ion channels (73,74), sensors for cellular states (75), and ligand / metabolite sensors(76,77). While approaches vary greatly, ranging from random(71,72,76), evolution-based(78,79), and structure-aided rational design[23], domain-insertion profiling with DNA sequencing (DIP-seq)(76,80) is particularly well suited to probing latent allostery. DIP-Seq combines rapid insertion library generation, with high throughput assays that can link a functional fusion protein (phenotype) to the specific insertion-product (genotype). Because DIP-Seq is unbiased and agnostic of the underlying mechanisms that give rise to allostery, it can broadly query a protein's allosteric capacity.

In our view, the power of DIP-Seq for mapping allosteric capacity could be greatly improved if we explicitly examine how permissibility, for the same host protein, depends on the insertion of different domains (differential permissibility) – dDIP-Seq. We argue this as follows: regions of a protein that are strongly biased towards order (e.g., transmembrane helices) are non-permissive to any type of inserted domain because insertions will disrupt this region's energetic balance and break the secondary and tertiary structure elements crucial for folding, trafficking, or multimeric assembly. Conversely, regions strongly biased towards disorder (e.g. unstructured termini) are generally permissive to any type of inserted domain (because the enthalpic and entropic impact due to the domain insertion are minimal). An exception to this rule would be sites that contain trafficking and/other related signaling motifs(81,82). Lastly, regions that are conformationally flexible and undergo disorder/order transitions; these regions are energetically balanced such that both ordered and disordered states are populated. In the limit that the perturbation introduced by insertion is relatively small compared to the free energy required to unfold this region, changes to the intrinsic stability of this region –and consequently overall protein stability– will depend on the properties of the inserted domain. The EAM model provides a link between these expectations and allostery. If (a) emergence of allosteric regulation requires regions that are conformationally flexible and poised to undergo disorder/order transitions, and (b) if differential permissibility is a metric for poised conformational flexibility, then, by deduction, differential permissibility can be used to map allosteric capacity.

Here, we challenge these hypotheses and predictions, and study exploited and latent allostery in Inward Rectifier K^+ channel Kir2.1 via dDIP-Seq. Inward Rectifier K^+ channels (Kir), are tetrameric K^+ channels, with a diverse set of physiological roles(83). They regulate resting membrane potential and excitability in neuromuscular tissue and vascular tone on blood vessels. They are involved in mechanisms of drug abuse and addiction, as well as learning by modulating synaptic plasticity. In the pancreas, they regulate insulin secretion. In recent years, several crystal and cryo-EM structures have improved our understanding of their gating mechanisms. In Kir2.1 for example, the binding of the positive allosteric regulator PIP_2 induces a disorder-to-order transition of a tether helix. Because of that transition, the G-loop, located within the C-terminal domain (CTD), wedges into the transmembrane domain and forces the intracellular gate open allowing K^+ to flow (**Figure 2-1a**). Under physiological conditions, Kir channels generate an inward K^+ currents at potentials negative to reversal potential for K^+ (**Figure 2-1b**). They also permit some current at slightly more positive voltages before becoming blocked by Mg^{2+} and polyamines(84,85). Interestingly, other members of the inward rectifier family share the overall topology (**Figure 2-1c)**, but are regulated by other ligands, including Gβγ (Kir3.x) and ATP (Kir6.x) binding in distinct regions of the CTD. Together, the availability of high resolutions structures(86–89), functional studies (reviewed in (83)), and ability to express in heterologous systems make Kir ideal test cases for applying the dDIP-Seq workflow.

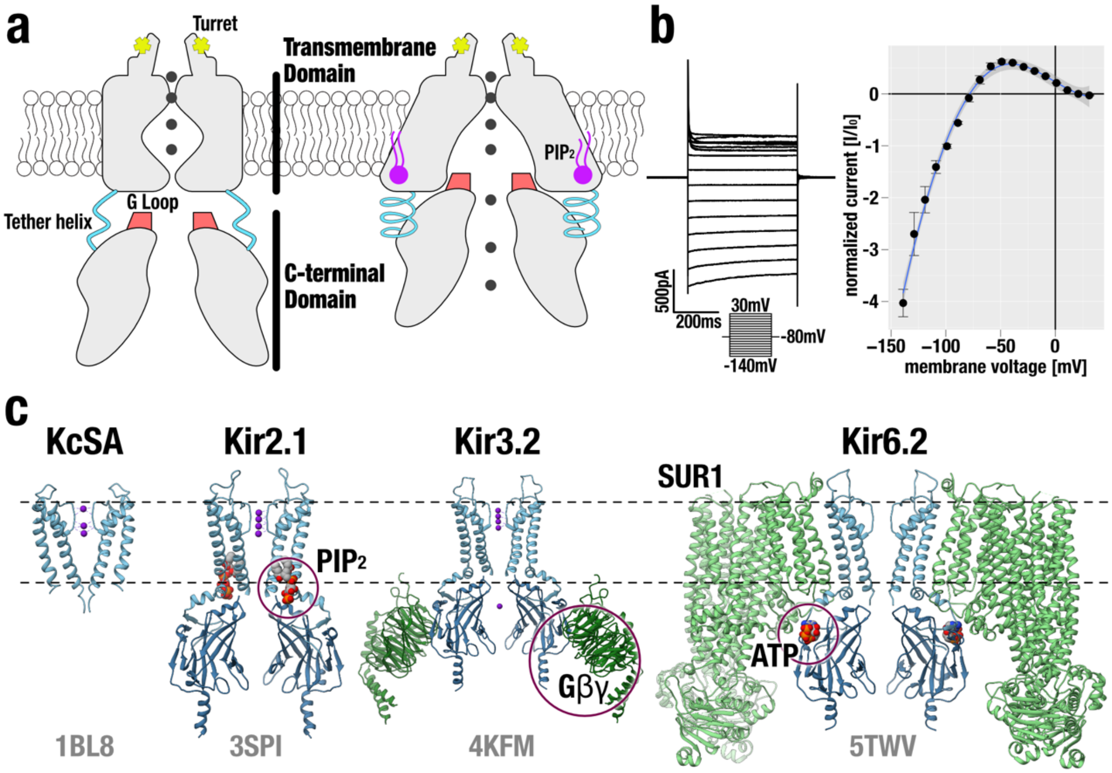


Figure 2-1: Inward Rectifier K$^+$ Channels. **(a)** Domain architecture of Kir2.1 in the closed (left) and open conformations (right). The C-terminal domain (CTD) is connected to the transmembrane domain (TMD) via a tether helix (light blue). Upon binding of PIP$_2$ (purple) at the interface between TMD and CTD, the tether helix undergoes a disorder-to-order transition and brings both domains closer together. The G-loop is wedged into the TMD causing the inner helix gate to open. Adapted from 1/5$^{/21\ 1:19:00\ PM}$**. (b)** Whole-cell patch clamp electrophysiology of WT Kir2.1 transiently expressed in HEK293 cells. A representative recording (left) and normalized currents (right) show strong inward rectification (±s.e.m., n=4). (**c**) A comparison between KcSA and representative structures of the inward rectifier K$^+$ channel family (PDB accession shown in grey) reveals that overall domain architecture is conserved. Channels are shown as blue and accessory protein as green ribbons. Allosteric modulators are indicated with purple circles.

We measure how permissive each ion channel site is to insertion of three different motifs with different physiochemical properties. We confirm that permissibility can be explained by conformational flexibility, and that differential permissibility is a hallmark of sites in Kir2.1 that are involved in allosteric regulation or that are homologous to sites exploited for allosteric regulation in homologues of Kir2.1. We furthermore demonstrate that this framework of measuring differential permissibility in ion channels is useful to endow them with useful functions. Implications of differential permissibility with respect to rationalizing how function diversified during ion channel evolution are discussed.

Results

A high-throughput ion channel domain insertion pipeline

The EAM model(64) predicts that regions with allosteric capacity (exploited or latent) are poised to undergo disorder/order transitions. Because inserting domains with different physiochemical properties will perturb the energetic balance in these regions to different degrees, we hypothesized that sites with allosteric capacity would have context-dependent mutability.

To test this idea, we used the DIPSeq(76,80) workflow to insert four different motifs (PDZ, Cib81, GSAG$_{2x}$, GSAG$_{3x}$) into nearly every amino acid position of Kir2.1: DIP-Seq relies on MuA transposase to insert an antibiotic cassette into random positions of a gene (i.e., all six reading frames) (**Figure 2-2**). Upon antibiotic selection of variants with insertions, we replaced this cassette with a motif of interest using restriction sites at transposon ends. The transposition mechanism(76,90) furthermore dictated that all insertions are flanked by short linkers, Ala-Ser and Gly-Ser-Ala at N- and C-terminus, respectively. We used the 10 kDa mouse α-syntrophin PDZ domain (PDB 2PDZ) because it is well structured and has been used to study how inserting large domains with known function disrupt recipient protein activity(91). Cryptochrome-interacting basic-helix-loop-helix (Cib81) is a similarly sized 9 kDa plant protein domain that forms a two-component switchable system with its binding partner, CRY2, after blue-light illumination(92). We included flexible 0.9 kDa GSAG$_{x2}$ and 1.3 kDa GSAG$_{x3}$ linkers to establish a permissibility baseline.

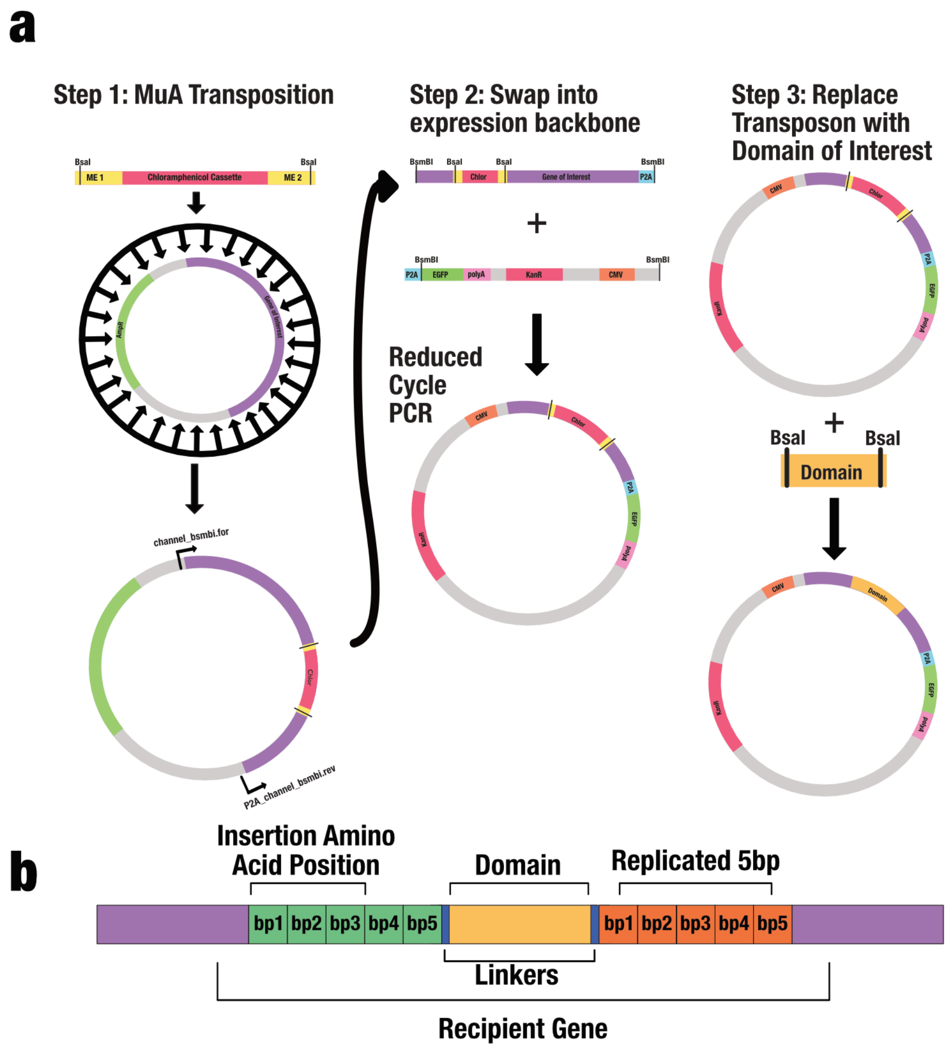


Figure 2-2: Insertion Library Construction. **(a)** Libraries were generated in three cloning and selection steps. (1) Select for a MuA transposase-delivered engineered MuA transposon with a chloramphenicol antibiotic cassette in a plasmid carrying Kir2.1. Flanking the antibiotic cassette are the beginnings and ends of flexible linkers with Golden Gate compatible BsaI type IIS restriction sites. (2) Reduced-cycle PCR amplify, add on Golden Gate compatible BsmBI type IIS restriction sites, and size separate channel genes with inserted transposons from those without transposons. Insert channel gene into a mammalian expression vector in-frame with a P2A-EGFP cassette. (3) Replace the transposon with a PCR amplified domain of interest with complementary BsaI sites and linkers using BsaI-mediated Golden Gate cloning. **(b)** Architecture of a domain insertion position: At the position of the domain insertion the five positions upstream are replicated on the other side of the transpositions; domain insertion positions are identified as the last full codon coding for an amino acid or in other words that corresponding to the amino acid

coded by bp1-3 of the replicated sequence. Domains are inserted with linkers to bring it into frame after insertion at +2 reading frame relative to the coding sequence.

After transiently expressing four insertion libraries into HEK293 along with EGFP as a transfection marker, we measured site-specific insertion ‘permissibility’. Permissibility is defined as the site-specific ability of Kir2.1 to accept an insertion without disrupting folding, homotetramer assembly, and trafficking to the cell surface. We leveraged the fact that for an inward rectifier to be expressed on the cell surface it must fold, oligomerize, and surface traffic(81,93,94). Therefore, only permissive insertion variants can be fluorescently labeled via a FLAG epitope tag we inserted into an extracellular loop of Kir2.1 (Ser116)(95). In this way, we collect two cell populations by fluorescence-activated cell sorting (FACS) (**Figure 2-3**): channel variants that express but don't surface express (EGFP^{high}/anti-FLAG^{low}) and those that do surface express (EGFP^{high}/ anti-FLAG^{high}). From both populations, we isolated and sequenced plasmid DNA, and aligned reads with the DIPSeq alignment pipeline(76,80). The complete workflow, including library generation, flow cytometry, and NGS was performed in triplicate. We calculate permissibility as site-specific enrichment between 'not surface-expressed' (NSE) to 'surface-expressed' (SE) insertion variants:

$$F(i,j) = \frac{r^i_{j_{SE}}}{t_{j_{SE}}} - \frac{r^i_{j_{NSE}}}{t_{j_{NSE}}} \qquad (1)$$

Where *r* is the number of reads at amino acid position *i*, in the *j*th dataset divided by *t*, the total number of reads in the *j*th given sample.

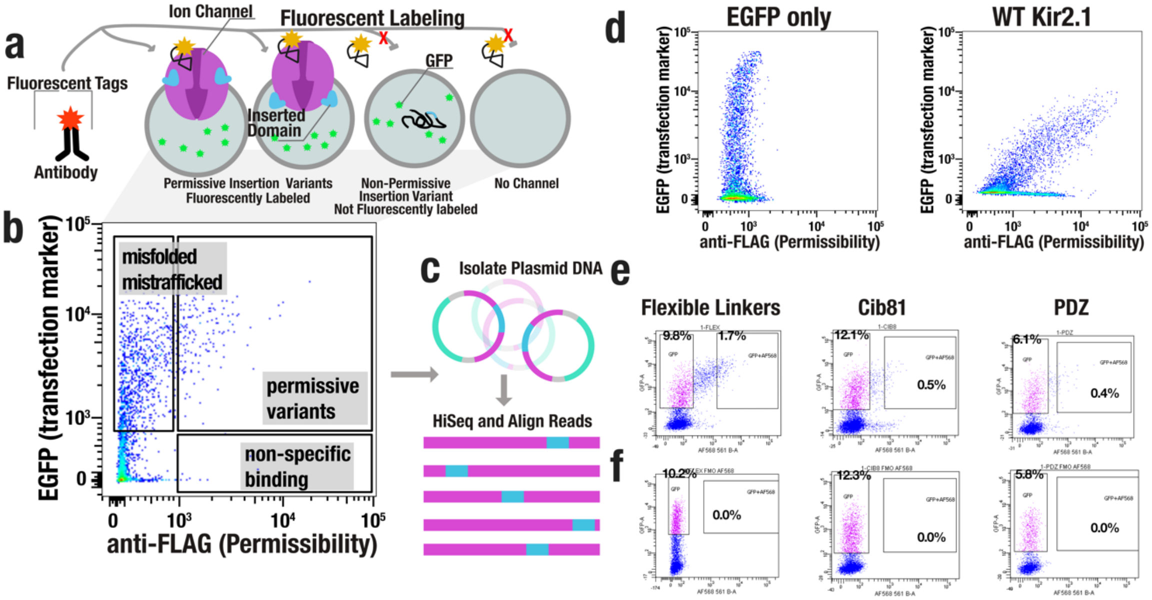


Figure 2-3: Permissibility Assay. (a) Domain insertion libraries are transiently expressed in HEK293FT cells and labeled with anti-FLAG Alexa568. (**b**) Cells are isolated by flow sorting into two populations: surface expressed (permissive insertion variants) and non-surface expressed (non-permissive) insertion variants. (**c**) Plasmid DNA is isolated from each population and subject to HiSeq. (**d**) Labeling controls (only GFP and WT Kir2.1-P2A-EGFP) expressed in HEK293FT to demonstrate antibody labeling. (**e**) Examples for each anti-FLAG labeled sorted samples (GSAG^{x2} and GSAG^{x3}, Cib81 and PDZ) expressed in HEK293FT cells with gates that were used for sorting. (**f**) Identical samples as in (**e**) without anti-FLAG labeling to guide setting gates.

Apart from some regions near the N-terminus, most notably M1, coverage in the remaining regions is near complete for all three insertion datasets (**Figure 2-4**). While scanning mutagenesis(96) of the M1 helix suggests that it likely would not allow domain insertions, the lack of data for the N-terminus, given its role in Kir2.1 gating(97) and trafficking(82), is unfortunate. That we consistently observed poor coverage in N-terminus can in part be explained by bias intrinsic to MuA transposases(98).

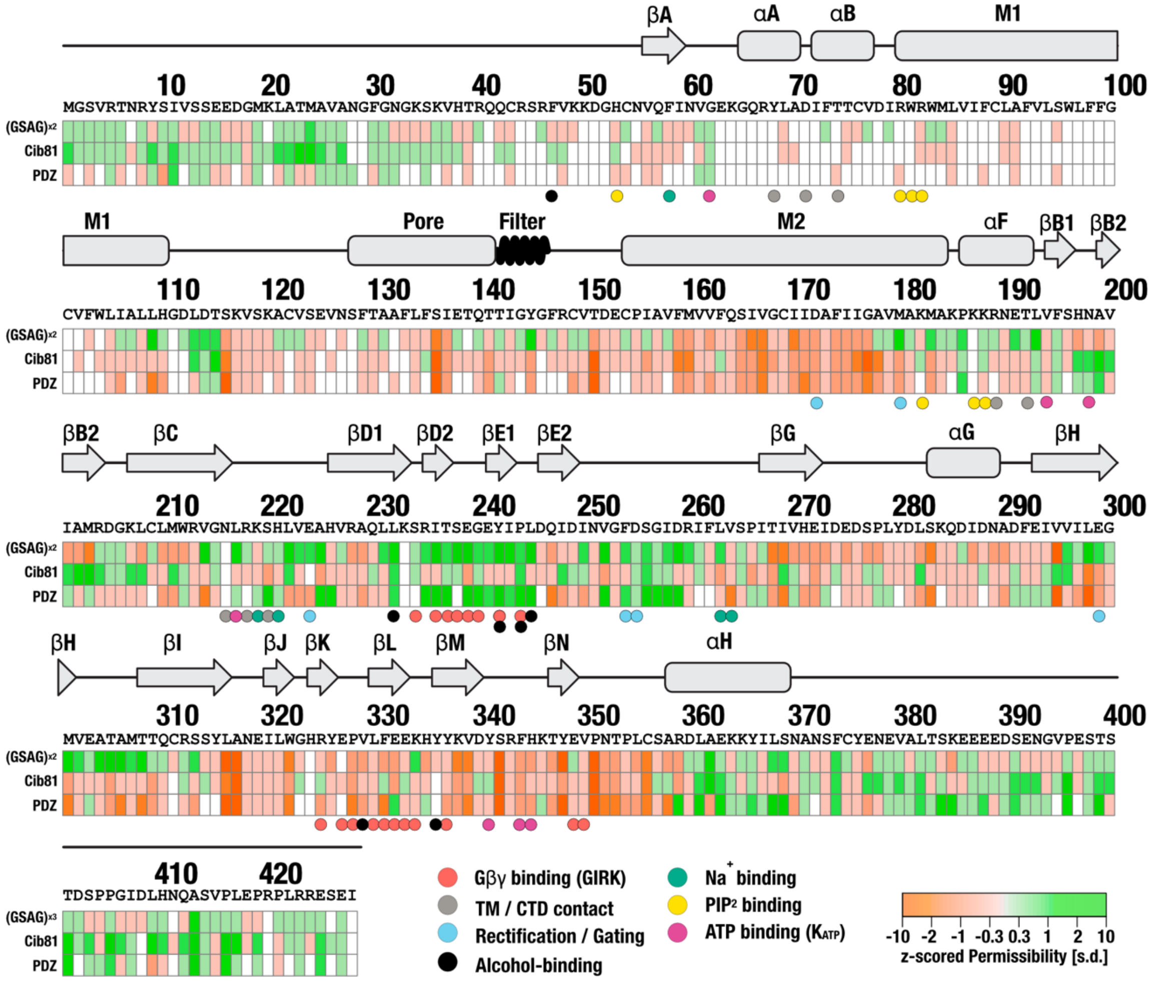


Figure 2-4: Domain Insertion Permissibility in hKir2.1. The primary sequence of human Kir2.1 (GI: 4504835) and secondary-structure elements are shown along with the permissibility score for three types of inserted domains (indicated on the left). Residues colored white refers to those for which there was insufficient data to assign a permissibility score. Key residues with functional relevance in Kir2.1 or a homolog (GIRK, K_{ATP}) are indicated by color-coded spheres below.

Permissibility is surprisingly different between domains

We then mapped permissibility onto the crystal structure of chicken Kir2.2 (PDB 3SPI)(87). Kir2.1 and Kir2.2 are nearly identical apart from an extracellular loop between M1 and the pore helix. As expected, domain insertion positions that should not allow surface expression do not (e.g., transmembrane and inter-subunit interfaces, **Figure 2-5a**, **Figures. 2-6-2-8**). Unsurprisingly, the unstructured C-terminus (which in vivo interacts with scaffolding proteins not present in HEK293 cells, such as PSD-95(99)) was highly permissive to any insertion (**Figure 2-4**). Predictably, overall flexible peptide insertions are more permissive than larger, more structured domains. Counter-intuitively, some surface-exposed, non-conserved regions (e.g., αG or βN, Kullback-Leibler divergence < 0.7, calculated for Pfam family IRK/PF01007 using MISTIC(100)) were also not permissive (**Figure 2-4**, **Figure 2-5b**). We take this as a data point suggesting that the permissibility rules, at least in Kir2.1, differ from those reported in other cytosolic proteins, such as kinases(85). Perhaps this is due selection pressures unique to membrane proteins such as the need for proper folding, assembly, surface trafficking, and membrane insertion. Lastly, a surprisingly large fraction of Kir2.1 CTD was permissive to insertion of 10 kDa domains (5.4% of CTD residues, and 3.6% of Kir2.1 residues have a permissibility score of > 2 standard deviations). Similar observations have been made in other proteins (e.g, (79,80)), and are thought to reflect that sequence continuity is not necessary for native folding(101).

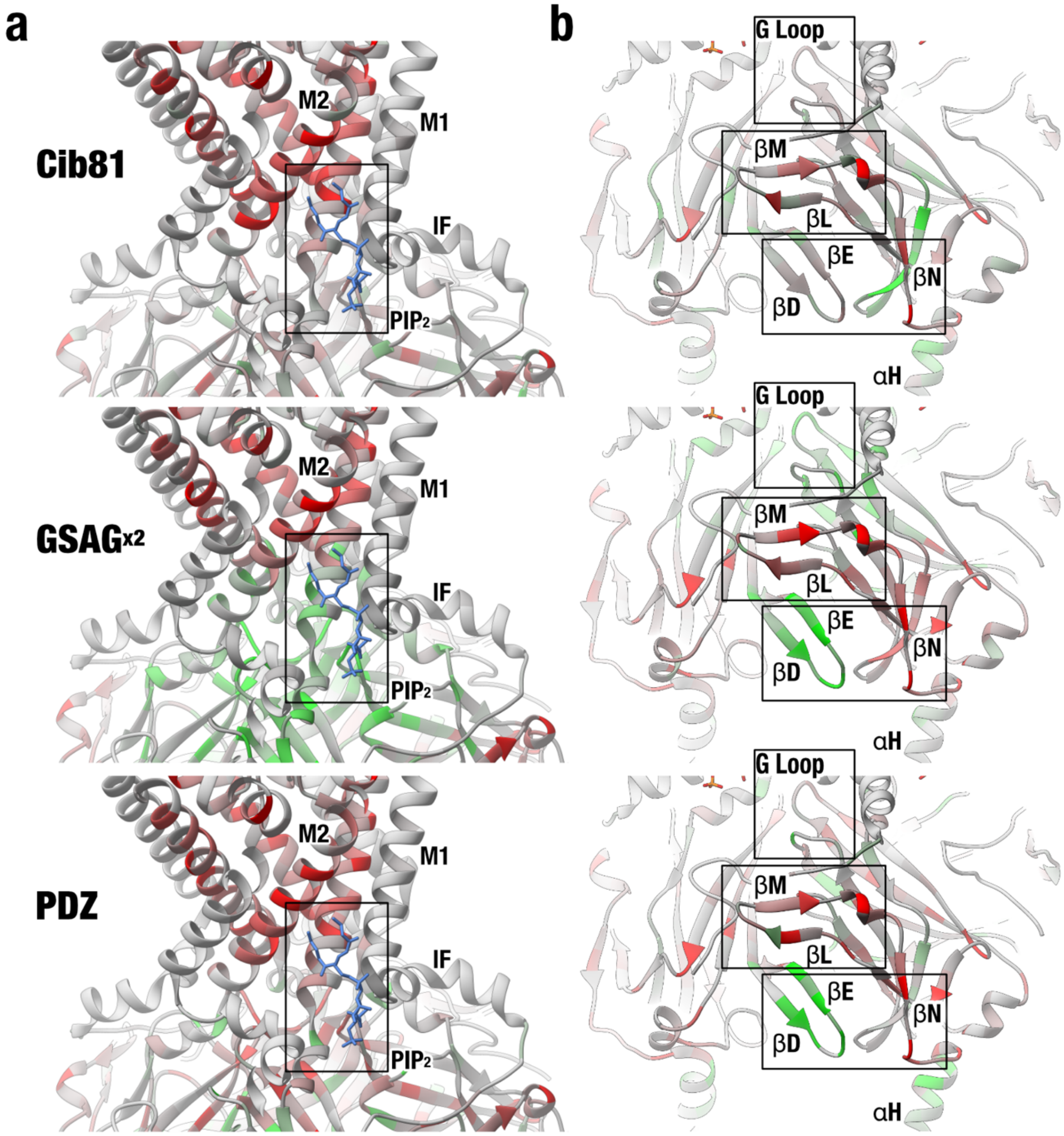


Figure 2-5: Differential Domain Insertion Permissibility. Permissibility data is mapped on the crystal structure of chicken Kir2.2 (PDB 3SPI). Domain insertion permissibility for three different domains (indicated on the left) is shown colored increasing from red-to-green. **(a)** All types of insertions into transmembrane helixes (M1 and M2) and inter-subunit interfaces (IF) are strongly selected against. Permissibility in PIP_2 binding site (boxed) depends on structural context of the insertion. (**b**) Some non-conserved, surface-exposed loops (e.g., βN) were not permissive, while the βD-βE loop (which binds Gβγ in GIRK) and the G-loop (βH-βI, the cytoplasmic gate in Kir2.1) have context-dependent permissibility (i.e. permissive for $GSAG_{x2}$ and PDZ insertion, less permissive for Cib81 insertion.

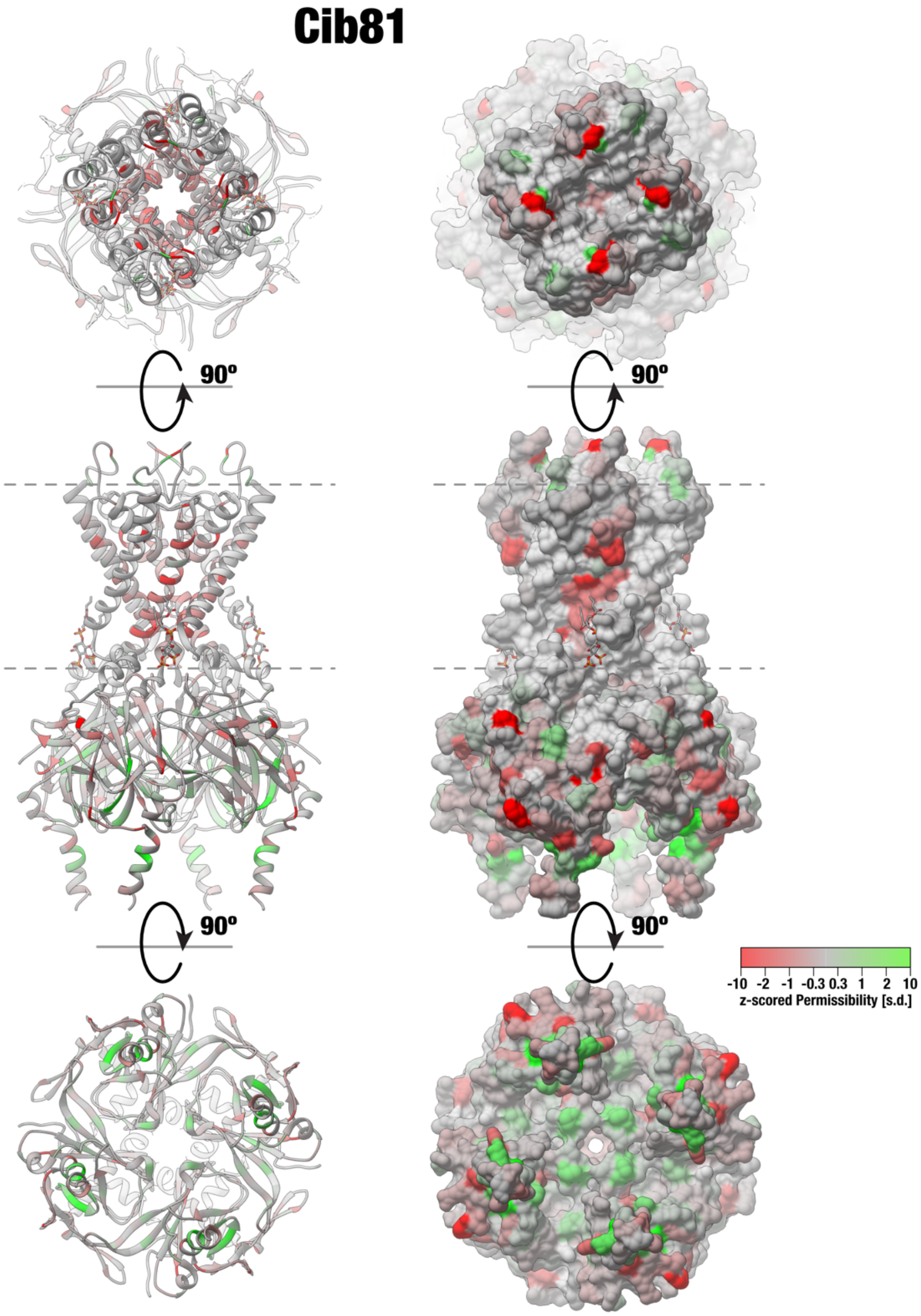


Figure 2-6: Domain Insertion Permissibility of Cib81. Permissibility data for inserting Cib81 is mapped on the crystal structure of chicken Kir2.2 (PDB 3SPI) displayed as a ribbon (left) or surface model (right). Dashed lines indicate plasma membrane boundaries.

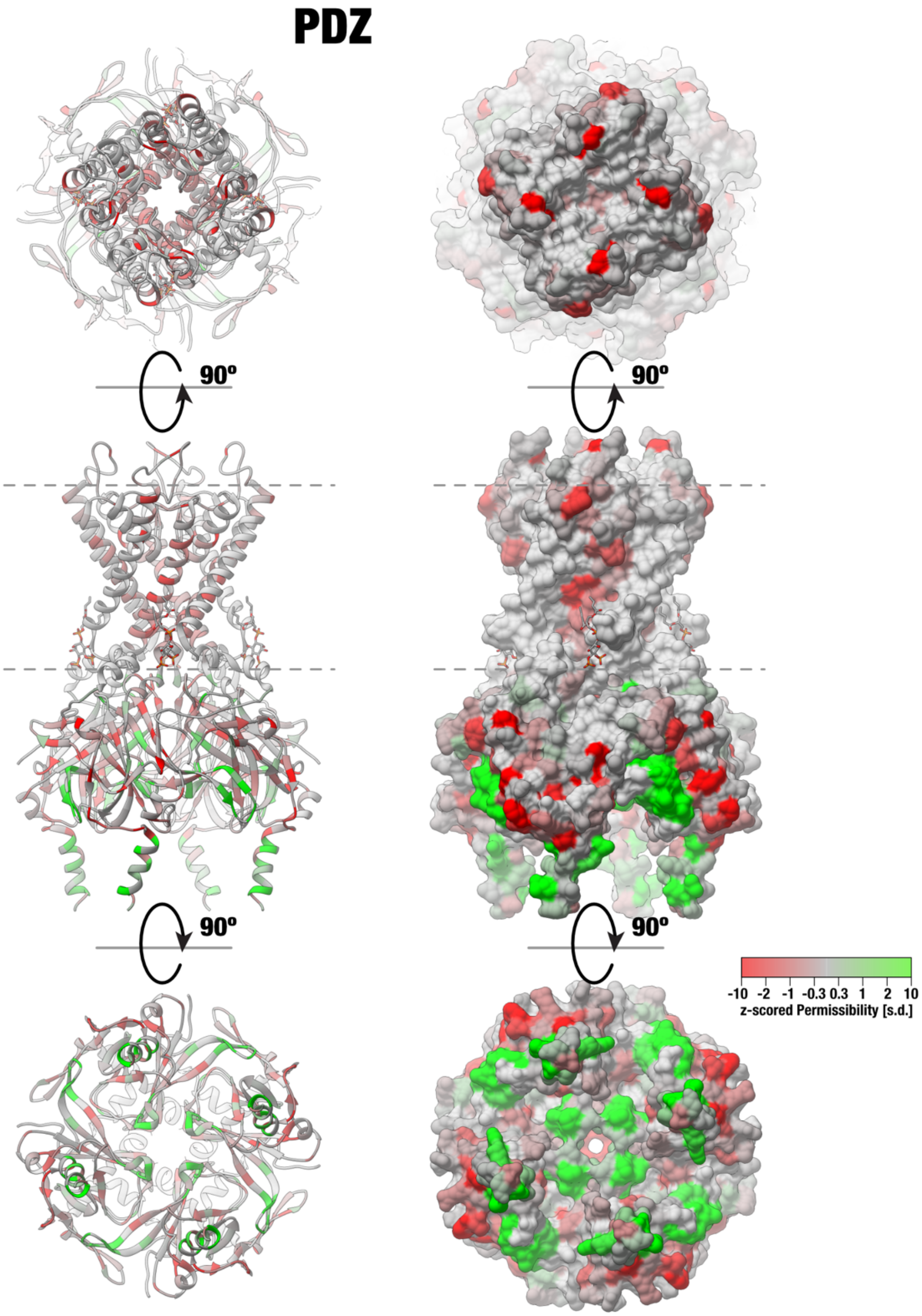


Figure 2-7: Domain Insertion Permissibility of PDZ. Permissibility data for inserting PDZ is mapped on the crystal structure of chicken Kir2.2 (PDB 3SPI) displayed as a ribbon (left) or surface model (right). Dashed lines indicate plasma membrane boundaries.

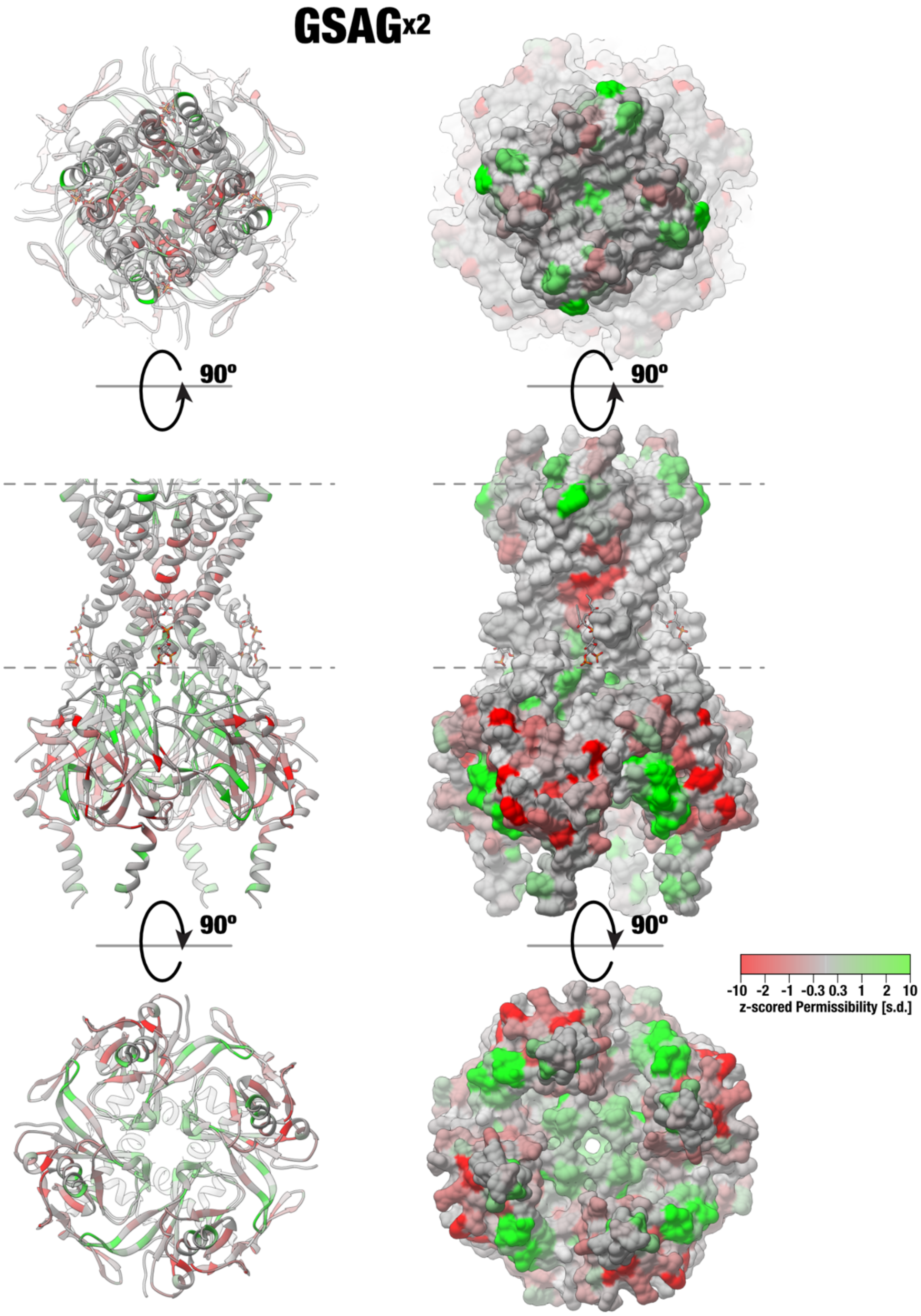


Figure 2-8: Domain Insertion Permissibility of GSAGx2. Permissibility data for inserting GSAG^{x2} linkers is mapped on the crystal structure of chicken Kir2.2 (PDB 3SPI) displayed as a ribbon (left) or surface model (right). Dashed lines indicate plasma membrane boundaries.

We more quantitatively compared permissibility profiles of biological replicates of Kir2.1 for inserted domains (structured vs. flexible) by clustering correlation matrices (Figure 2-9a). Biological replicates were overall in good agreement. This indicates that relatively little noise was introduced through the transposition, heterologous expression, and cell sorting steps. We found that domains cluster by structure (Cib81 and PDZ cluster discretely from each other, but flexible insertions do not) and size (Cib81 and PDZ cluster closer than flexible peptides) (**Figure 2-9a**).

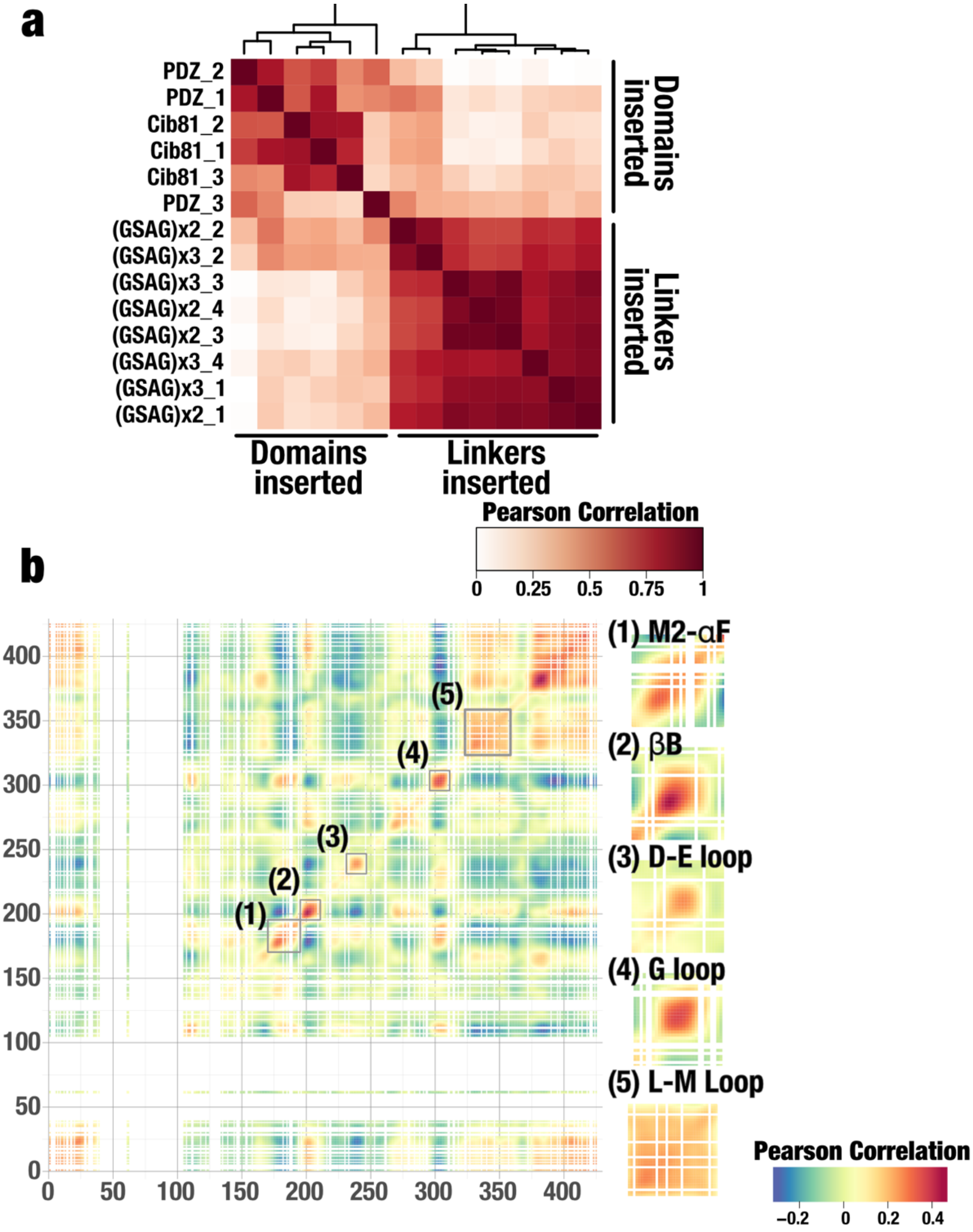


Figure 2-9: Pearson correlation of biological replicates and site-specific permissibility correlations. (a) Hierarchical clustering of all individual permissibility datasets based on domain structure and length. Biological replicates show a high degree of reproducibility. As expected, datasets for structured domains (Cib81 and PDZ) cluster discretely with themselves than with each other whereas flexible linkers cluster indiscriminately between each other. **(b)** In Kir2.1, PDZ permissibility is highly correlated in (1) M2-αF – the PIP_2 binding site, (2) the βB loop where ATP binds in Kir6.2, (3 & 5) the βD-βE and βL-βM loops where Gβγ binds in GIRK, as well as (4) the G loop involved in channel gating. White indicated sites with incomplete data.

Allosteric sites are most differentially permissive

Surprisingly, correlation matrices of permissibility profiles sorted by insertion sites reveal distinct clusters that coincide with structural features involved in allosteric regulation of Kir2.1 (**Figure 2-9b**). Permissibility in allosteric sites in Kir2.1 is more strongly correlated than in non-allosteric sites (unpaired Wilcoxon rank sum test, p-value < 2.2×10^{-16}). This included the PIP_2 binding site at the interface between the pore and cytosolic domain, and the G-loop (βH-βI), a flexible region involved in channel opening(87). The same pattern was observed in sites where allosteric modulators bind in homologs of Kir2.1. This includes the βB-βC loop (which binds ATP in Kir.6x(89)), and the βD-βE loop (which binds Gβγ in GIRK(88)).

Our data also reveals that permissibility is sensitive to the structural context of the inserted domain. Despite Cib81 and PDZ being of similar size, many sites are differentially permissive (69/229 sites, or 30%, when measured using the hamming distance criterion after binarizing permissibility data). We interpret this as context dependence for permissibility beyond simple steric effects (**Figure 2-4 & 2-5**). While flexible linkers had highest permissibly overall, in several sites only Cib81 and PDZ are tolerated (e.g., βB2), further demonstrating permissibility's context dependence. Furthermore, visual inspection of permissibility maps (**Figures. 2-4,2-6,2-7,2-8**), and a quantitative comparison of permissibility in PDZ and Cib81 datasets (using the Euclidean distance measure), shows that differential permissibility is a more common feature in functionally important regions (unpaired Wilcoxon rank sum test, p-value = $3.3*10^{-5}$, **Figure 2-10**).

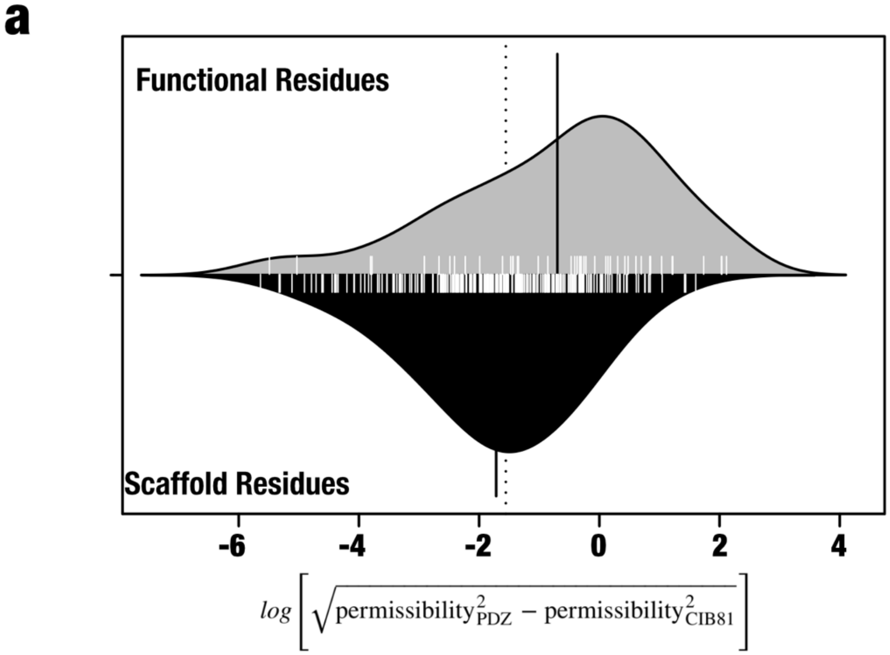


Figure 2-10: Functionally important sites are more likely to be differentially permissive. (a) Kir2.1 residues are termed functional if they are involved in Kir2.1 gating (e.g. PIP_2 binding site) or a residues that corresponds to those involved in gating of a homolog (e.g. Gβγ-binding in GIRK). Beanplots for the calculated per-residue log-transformed domain insertion permissibility difference in PDZ and CIB81 datasets. Data for residues involved in function in Kir2.1 or related homologs is shown in grey; data for scaffold residues not involved in function is shown in black. White beanlines indicate calculated differences for individual residues. Solid vertical lines indicates then bean average for each group, while the vertical dashed line indicates the total average of for both groups.

Permissibility is dependent on dynamic protein properties

Given that permissibility could not be explained by simple sterics, we explored what readily calculable and accessible protein features explain permissibility, with the goal to derive a better understanding of the underlying mechanisms that determine permissibility. To this end, we calculated structure-, conservation, and dynamic-based properties for Kir2.1 – using publicly available webservers (see methods for more detail)– and calculated correlation coefficients between these properties and different domain permissibility profiles (**Figure 2-11a**). We found that domain insertion permissibility is correlated with dynamic features and does not correlate well with static and conservation-related protein properties.

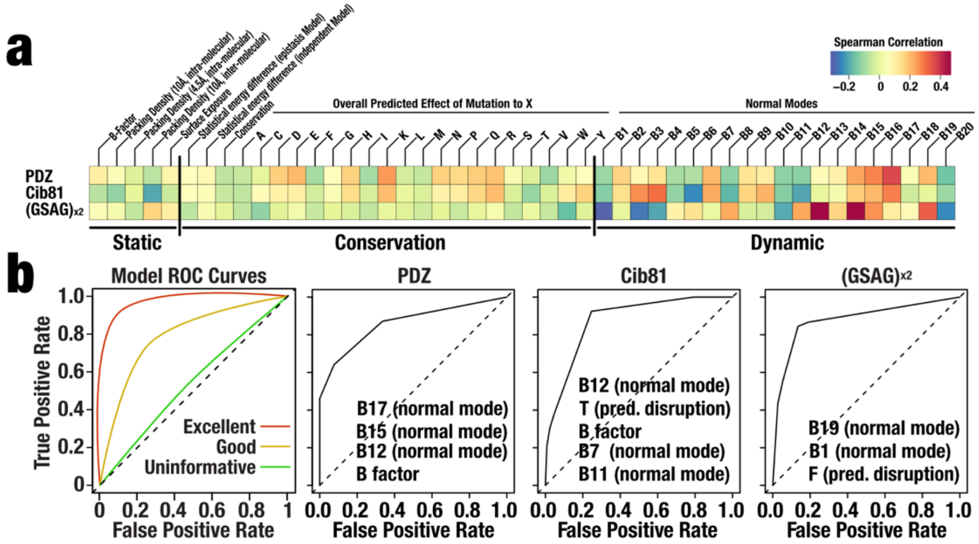


Figure 2-11: Parameter Correlation and Model Performance. **(a)** Spearman correlations between permissibility for the inserted domain indicated left with the calculate property indicated on the top. Vertical bars separate feature categories (Static, Conservation, and Dynamic). Dynamic protein properties show strong correlation, while correlation with static and conservation properties is spurious. **(b)** Model performance as measured with receiver operator characteristic (ROC) curves for each domain. Each decision tree model was evaluated using 10-fold cross-validation (see methods for details). Example of ROC curves for models with varying performance are shown on the left-most panel. For each domain, the top predictive properties used in decision trees are indicated on the ROC curve.

To further probe if correlation with dynamic features is meaningful (i.e. suggestive of mechanism) and determine how well computed properties explain permissibility, we constructed decision tree classification models. Decision trees automatically pick features and thresholds based on experimental data to build a predictive model, which is then tested on withheld data. By observing which features get picked in the best performing models, we can deduce which protein properties predict and explain permissibility best. Consistent with the result that permissibility best correlates with dynamic properties, dynamic features had the greatest predictive power across the three types of domains (**Figure 2-11b**, **Figures 2-12, 2-13, 2-14**). While some profiles' predictive models perform better than others (GSAG$_{x2}$ was best and PDZ was worse) and none fully explain permissibility, we can build models for all domains, whose performance is far better than random as assessed by receiver operating characteristic (ROC) curves and other performance criteria (**Figure 2-11b**, **Figure 2-14**). Our ability to generate predictive models for permissibility demonstrates that features used in predicting measured permissibility are meaningful and that the indefinable qualities of permissibility play minor roles. That some properties, e.g. PDZ~B factors, Cib81~B12, (GSAG)$_{x2}$~F, are most predictive in decision tree models yet are not strongly independently correlated demonstrates the superior sensitivity of non-linear models that capture interactions between features (**Figure 2-11**, **Figure 2-12**). Furthermore, the necessity for a non-linear approach (decision trees) suggests that permissibility and by extension allosteric capacity, at least in Kir2.1, is an emergent phenotype from interactions between multiple protein properties, as opposed to a linear combination of, for example, conservation and surface exposure[23].

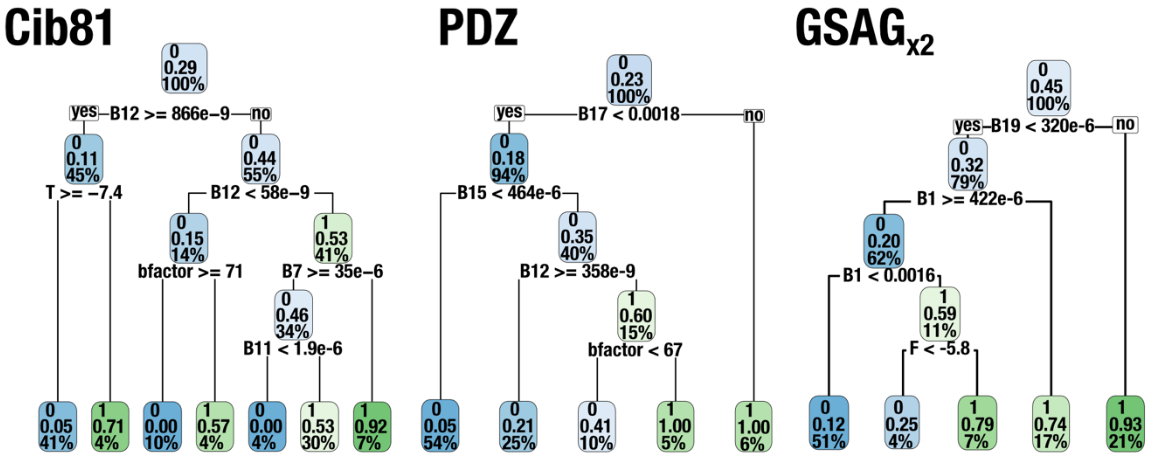


Figure 2-12: Decision Trees. Decision trees were trained on calculated conservation, static structural and dynamic protein properties to predict binarized Cib81, PDZ and GSAG$_{x2}$ permissibility. Features and cutoffs for each node are selected based on a 'complexity parameter' which penalizes uninformative complexity. All trees were restricted to a maximum depth of four and evaluated using 10-fold cross-validation. Decision tree leaves can be read as: the top-most number refers to the leaf class (0–not permissive and 1–permissive), next are percentage of permissive samples within the leaf that fall within the class, and percentage of all data that are within the leaf. Color intensity refers to purity of sample, at each leaf (fraction of non-permissive-to-permissive withh a blue-to-green

color scale). As can be seen in the models, most of the nodes use dynamic properties. This suggests that dynamic properties are more discriminative compared to other property classes.

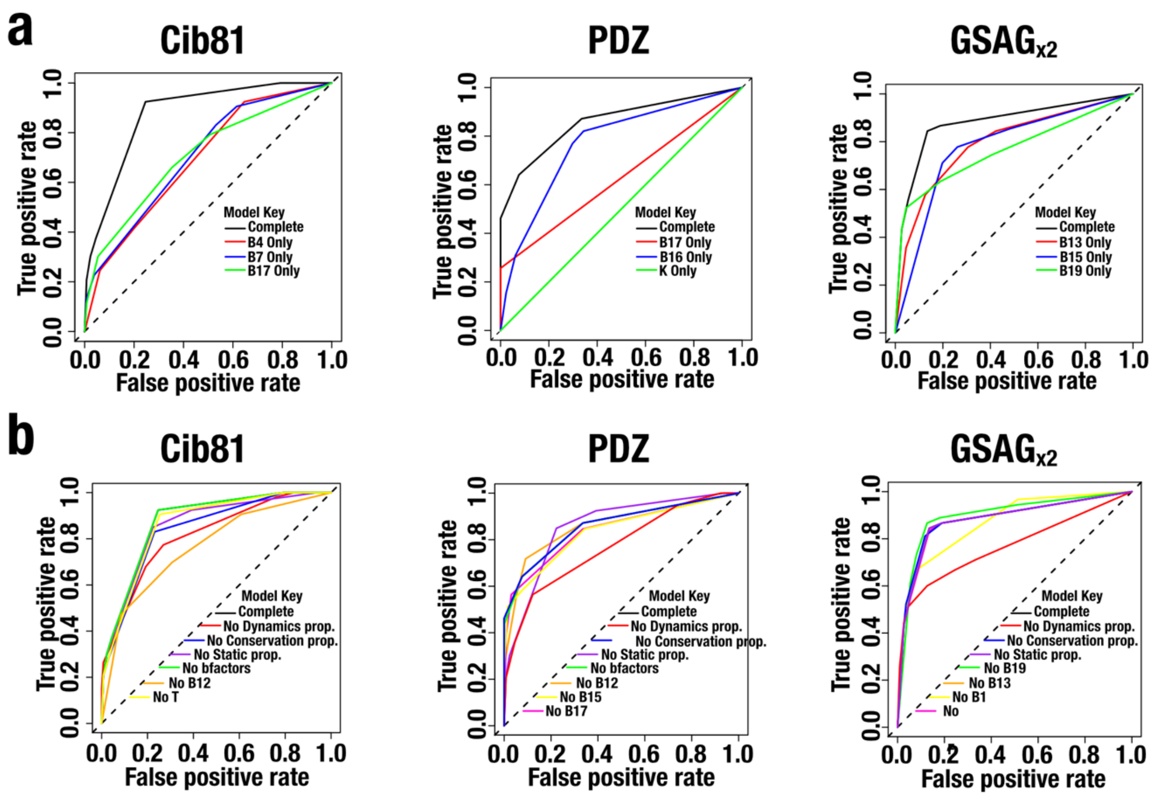


Figure 2-13: Decision trees trained on limited numbers of properties. (a) Decision trees were trained using the top three Spearman co-efficient correlated properties individually using the same model parameters as before (a maximum depth of four and cross-validated ten times). Model performance was compared using receiver operator characteristic (ROC) curves. All models trained on single properties perform worse than models trained all properties. This suggests that multiple interacting properties determine permissibility. Interestingly, the predicted effect of any amino acid mutated to lysine was highly linearly correlated with PDZ permissibility. Nevertheless, using it alone was not sufficient to predict PDZ permissibility. On the other hand, normal modes –a dynamic feature– were able to predict PDZ permissibility when used individually. **(b)** Decision trees were trained (maximum depth of four and 10-fold cross-validated) with individual properties or entire property classes withheld. Model performance was compared using ROC curves. In every case, removing all computed conservation and static structural properties had little effect on model performance. In contrast, removing dynamics based

properties was detrimental to permissibility predictions. Apart from two examples (Cib81~B12 and GSAG^{x2}~B1) removing individual properties were not substantially impactful on model performance suggesting that there is redundancy in computed properties. Overall, all decision trees reinforce the idea that (1) dynamics are the most informative and closest correlated protein properties to permissibility and (2) interactions among protein properties are important in predicting permissibility.

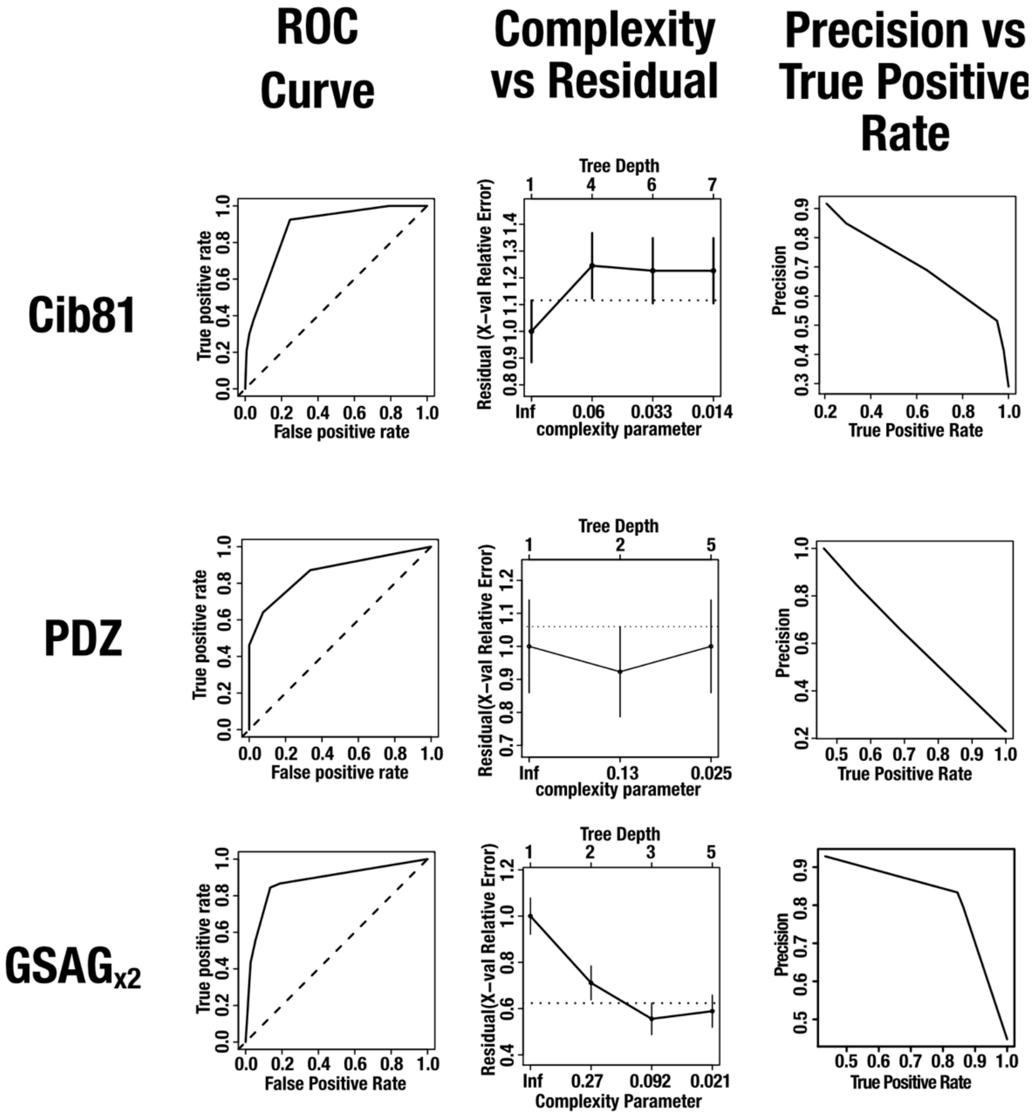


Figure 2-14: Decision Tree Model Performance. (Left Panels) Cib81, PDZ, and GSAG^{x2} decision tree performance using different criteria. All model criteria are derived by testing the model on test data withheld from training data. Receiver operator characteristic (ROC) curves also shown in **Figure 4b** are the proportion of positive data that are predicted

positive (true positive rate) plotted against the proportion of predicted negative data that are predicted positive (false positive rate) at varying model cut-offs. The dotted line represents the performance of a random model. ROC curves demonstrate that as threshold are changed, true positives increase at the expense of increasing false positives. All models perform far better than random. (Center Panels) Complexity vs. residual plots show the difference between predicted and actual data (residual) plotted against the tree depth with the amount that splitting a node improved model performance (complexity parameter) used for trimming trees also noted. The dotted line represents the ideal threshold residual for model performance and error bars are s.e.m. (n=10). These plots demonstrate that increasing tree depth, and therefore tree complexity, has diminishing impact on improving model performance. (Right panel) Precision vs. true positive rate plots show proportion of positive predicted data that are positive (precision) plotted against the true positive rate at varying model cut-offs. Precision-true positive rate curves show that as more true positives are predicted, more negative data are predicted as false positive. A random model would be a flat horizontal line. All models based on all parameters perform far better than random.

Allosteric site insertions contextually impact function

Permissibility reports whether a Kir2.1 insertion variant can fold and traffic to the cell surface. We speculated that a significant fraction of insertion variant retain function in the presence of a large domain. To further explore Kir2.1's differential sensitivity to domain insertion in allosteric regions, we focused on a representative sample of insertion positions –drawn from known allosteric sites in Kir2.1 and homologs, and sites with qualitatively different permissibility patterns– to assess whether they remain functional (i.e., able to conduct K^+) upon domain insertion. We subjected this subset to a flow cytometry-based optical activity assay that measures population-level resting membrane potential (RMP) in HEK293FT cells using an oxonol voltage-sensor, DIBAC**4**(**3**) (**Figure 2-15a**)(56,102). Since Kir2.1 drives the RMP towards the reversal potential of K^+, cells expressing functional Kir2.1 are more hyperpolarized compared to 'empty' (RFP only) cells (**Figure 2-15b**). We note that by measuring RMP in many thousands of cells we can bypass cell-to-cell variability that can make determination of RMP for transiently transfected cells by patch clamp electrophysiology burdensome, particularly for poor-expressing insertion variants. In fact, even distributions of DiBAC**4**(**3**) fluorescence for WT Kir2.1 transfected cells are not uniform (**Figure 2-15a**), indicating a continuum of more and less hyperpolarized cells.

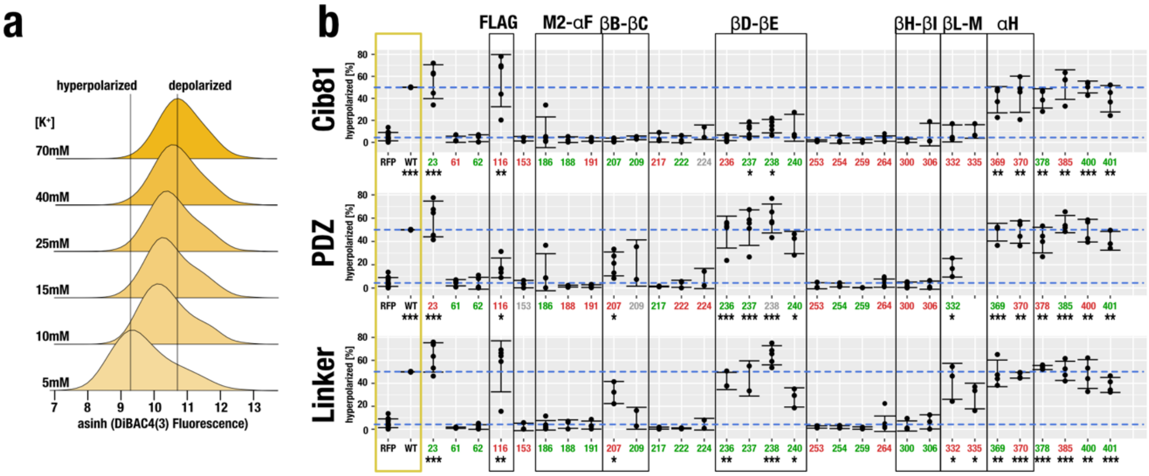


Figure 2-15: Domain Insertion Impact on Kir2.1 Function. (a) Stacked histograms of population-level DiBAC**4**(**3**) fluorescence (hyperbole arcsine transformed) in HEK293FT cells expressing WT Kir2.1 as a function of external K^+ (concentration indicated on the left). Increasing K^+ depolarizes the cells, resulting in less membrane-partitioning of the dye thus increasing measured fluorescence. **(b)** Shown are the percent hyperpolarized cells expressing the indicated Kir2.1 variant and inserted domain. Permissibility (**Figure 2**) of that variant is indicated in color (green, permissive; red, non-permissive; gray, no data). Higher percent hyperpolarized indicates function, and lower percent hyperpolarized indicates disruption. Reference measurements are provided for HEK239FT expressing miRFP670 alone and WT Kir2.1 co-expressed with miRFP670 (yellow box). Reference levels of WT and no channel are indicated by blue dashed lines. Replicates for insertion mutants are plotted with bars representing standard deviations and centered at the mean (each insertion variant n=3-5, wildtype n=21, RFP670 n=14). Significance of differences for means of each variant and WT with respect to channel (RFP) was tested using a one-side t-test. Significance levels are *** p<0.001, ** p<0.01, and * p<0.05, respectively. Variants without a mark are not significant. Regions discussed in the text are indicated by black boxes.

What is immediately apparent is that permissibility and impact on function, differ in many sites. Some permissive sites do not produce functional protein, while –perhaps more interesting– non-permissive site can produce functional protein. We expect this lack of correlation because permissibility solely tests for folding, assembly, and trafficking to the cell surface. A variant might traffic worse (thus appear non-permissive) but ultimately be functional, while a well-trafficked variant can be functionally compromised. An insertion's impact on function nevertheless follows a pattern similar to permissibility: As expected, insertions into flexible and highly permissive regions (e.g., c-terminal residues) of Kir2.1 have little impact on function. All insertions into regions critical for gating, (e.g., the G-loop βH-βI), break channel function. The effect of insertions into the exposed extracellular loop (Ser116) are subtler. Here, both Cib81 and a flexible linker are well tolerated, while PDZ impairs function significantly. The emerging theme of differential impact on function continues in other regions of Kir2.1, including the PIP_2 binding site (M2-IF). Here, Pro186 is permissive to all insertions, but partial function remains with large domain insertions (PDZ & Cib81) while a flexible linker completely breaks channel function. Conversely, permissibility and impact on function tracked quite well for insertions into the βD-βE loop. Here, both PDZ and flexible insertions allowed near wildtype channel function, while Cib81 greatly reduced channel function. Overall, as with permissibility, the functional assay shows that domain insertions impact function in a context dependent-manner that cannot be explained with simple sterics.

Allosteric site insertions enable control of channels

What permissibility assays tell us is that several sites in Kir2.1 appear to be sensitive to the structure of the inserted domain. Sites that are involved in allosteric signaling in Kir2.1 or are equivalent to functional sites in homologs (GIRK and Kir6.2) are more likely to be differentially permissive (**Figure 2-10**). Many of these sites retain partial function upon domain insertion (**Figure 2-15**). In aggregate, we interpreted these data as sites with differential permissibility being more likely to possess allosteric capacity. We speculated that introducing light-switchable domains into allosteric sites might affect Kir2.1 with light and create an optogenetic reagent. To test this idea, we assayed Cib81 insertion variants for light-dependent modulation. Since it was not feasible to test all residues, we focused on those from allosterically regulated regions in Kir2.1 (M2-IF; Pro186) and in homologs such as Kir6.x (βB-βC; Lys207) and GIRK (βD-βE, βL-βM, αH; Thr237, Ser238, Glu241, Glu332, His335, Ser369, Asn370). Controls included insertion into the unstructured C-terminus (Thr401), the sterically inaccessible extracellular loop Ser116, wildtype Kir2.1, and a pore gating mutant, V302M[55]). Initially, there was no optimization of flanking linkers. We reasoned that if Cib81 is sterically accessible and the recipient site has allosteric capacity, then a light-mediated association of the channel with co-expressed Cry2 (size

70 kDa) would modulate channel gating even if binding interfaces are not optimized. We adopted the flow cytometry RMP assay to measure Kir2.1 activity after challenge with varying concentration of extracellular K^+, with and without blue light illumination. As expected, wildtype channel and a gating mutant have no light-dependent modulation (**Figure 2-16**). Furthermore, when Cib81 is inserted into the extreme C-terminus (Thr401) or an inaccessible extracellular loop (Ser116), there was no light-dependent effect on Kir2.1 activity. Remarkably, even though channel function was severely impaired (**Figure 2-15b**), when Cib81 was inserted in the PIP_2 binding site, illumination markedly decreased the remaining Kir2.1 activity (**Figure 2-16b**). We validated this with patch clamp electrophysiology, which shows that the open probability of Kir2.1(Pro186CIB), which is low to begin with (**Figure 2-17a**), is further decreased with blue light illumination (**Figure 2-17c**).

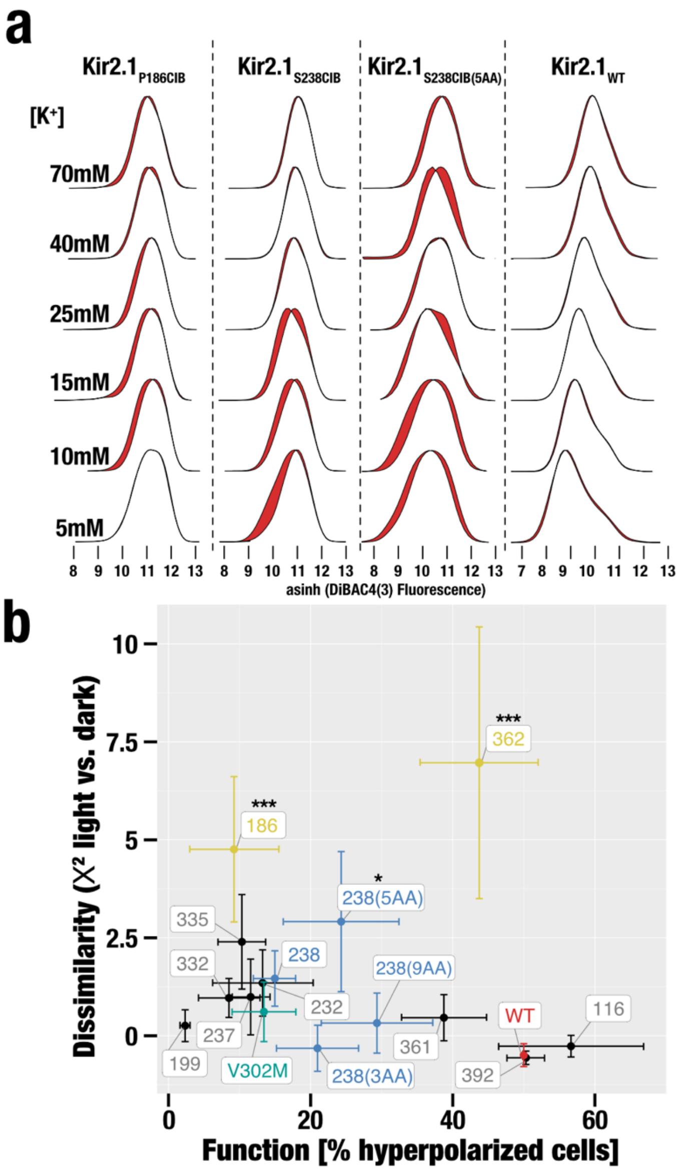


Figure 2-16: Light-modulated Kir2.1 Variants. (a) Moving average (window size: 15 residues) of permissibility z-score difference between PDZ and Cib81 datasets. Regions in which differential permissibility exceeds one standard deviation (dashed blue line) are shaded in red. Individual residues subjected to the light-switching assay are shown as labelled red dots. **(b)** Dissimilarity (X^2) of K$^+$-induced depolarization with and without illumination plotted against insertion variant function. Highlighted are mutants after linker optimization (blue), gating mutant V302M (green), and wildtype Kir2.1 (red). Significance of light modulation is tested by pairwise comparisons using Dunnett's test for multiple comparisons with wild type as control and post-hoc multiple comparison adjustment. Error bars are standard deviations on the x-axis (each insertion variant n=3-5, and wildtype

n=21) and s.e.m. on the y-axis (n=3). Significance levels: *** p < 0.001; ** p < 0.01, *, p< 0.05, all others not significant.

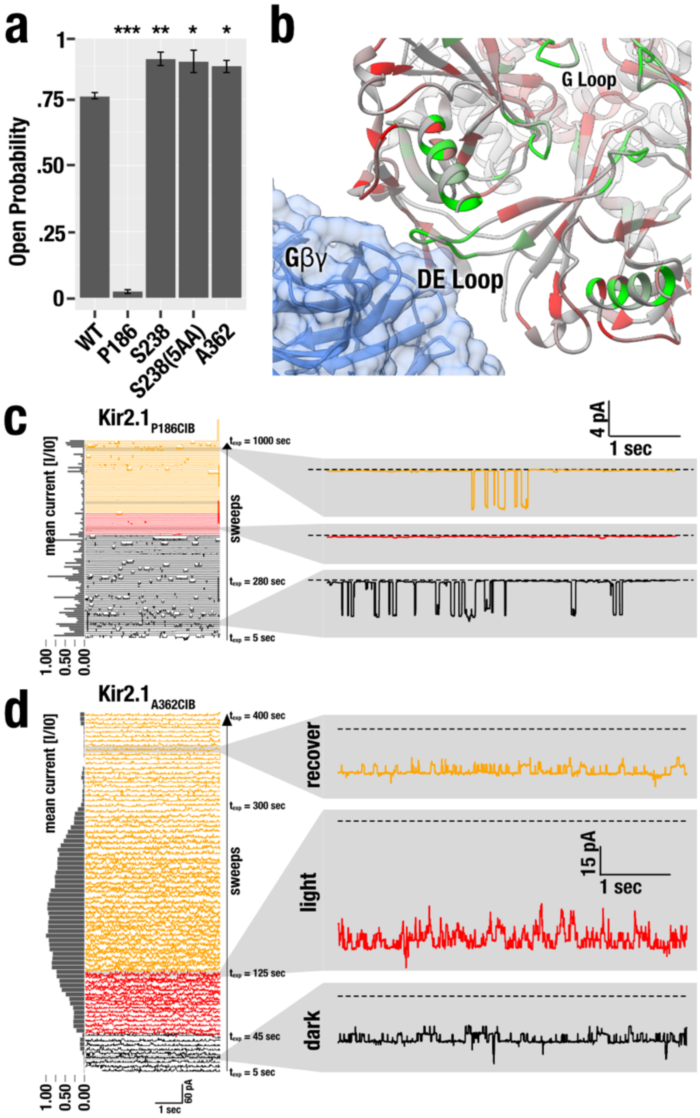


Figure 2-17: Electrophysiology of Light-switched Kir2.1 Variants. **(a)** Open Probability determined by on-cell patch clamp electrophysiology for WT Kir2.1 or the indicated insertion mutant. Boxes indicate standard deviation, thick crossbar indicates mean, and dots indicate individual measurements (n = 3-12). Significance is tested by pairwise comparisons using Dunnett's test for multiple comparisons with wild type as control and post-hoc multiple comparison adjustment. Significance levels: *** p < 0.001; ** p < 0.01, *, p< 0.05, all others not significant. **(b)** Domain Insertion Permissibility mapped increasing from red-to-green onto the crystal structure of GIRK2 (Kir3.2) in complex Gβγ (PDB 4KFM). Many highly permissive sites in Kir2.1 are homologous to those that interact

with Gβγ in GIRK2. (**c**) Representative examples of on-cell patch clamp recordings of indicated insertion variant (P186CIB top, N370CIB bottom). Shown are consecutive 5 second sweeps spread out vertically, bottom (beginning of experiment) to top (end of experiment). Black indicates sweeps before illumination, red during illumination, and yellow after illumination. Holding voltage is -100mV (P186CIB) and -80mV(N370CIB). Channels open inward (negative current). Normalized mean current for each sweep is shown on the left as a bar graph. Magnified individual sweeps before (black), during (red), and after illumination (yellow) are shown on the right. While Kir2.1 P186CIB responds to illumination with further decreasing the already small open probability, Kir2.1N370CIB is activated with illumination.

We also observed Kir2.1 light-modulation when Cib81 was inserted into the pore-facing side of the αH Helix (N370), but not the outward-facing side (Ser369) (**Figure 2-16b**). Furthermore, patch clamp validation of this insertion showed that open probability is higher than wild-type channel in the absence of illumination (**Figure 2-17a**) and further increased with illumination (**Figure 2-17d**). The αH Helix is a potential Gα binding site in GIRK based on several NMR structures; however, this interaction hasn't been fully explored[56].

We noticed weak light modulation when Cib81 was inserted into the βD-βE and βL-βM loops (e.g., Ser238, Leu332 & His335) (**Figure 2-16b**). The weak impact of Cry2 recruitment in this region could be due to none-optimized binding interfaces in contrast to those Gβγ encounters in GIRK's βD-βE and βL-βM loops (**Figure 2-17b**)[40]. When we patched cell expressing these insertion mutants, we observed higher Kir2.1 activity in one (Ser238) even in the absence of illumination (**Figure 2-17a**), suggesting that insertions into the βD-βE loop are activating. No channel activity was observed for Leu332 & His335 (**Figure 2-16b**). We explored linker optimization for a cleaner photoswitching phenotype. Flanking Cib81 by five amino acids, but not three or nine, improved light modulation of Ser238CIB (**Figure 2-16b**).

Discussion

Our findings reveal that permissibility in Kir2.1 is correlated with protein dynamics, but not structural features or conservation. There is broad support for the idea that the dynamics of structural elements in proteins, when poised, provide the mechanistic basis for allosteric coupling(64,65). Protein dynamics are influenced by the stabilities of participating structural elements, which can be altered by in-frame domain insertions. The magnitude and sign (i.e., more ordered vs. more disordered) depends on the physicochemical properties of the inserted domain. This context dependence of altered local dynamics manifests in our assays as differential permissibility.

Translated into categories of permissibility, we expect generally permissive regions, where insertions are accepted and the overall functional phenotype of the channels remains mostly unchanged. Structurally important regions are expected to cause the misfolding and a loss of function phenotype regardless of the inserted domain's properties. Regions that have conformational plasticity that depends on the context of a perturbation are expected to have differential permissibility. This means that the effect of a domain insertion on both folding, assembly, and trafficking will depend on the biophysical properties of the inserted domain. In a similar vein, depending on the context of the inserted domain, different functional phenotypes are expected.

We have found regions belonging to each of these three categories in Kir.2.1. Examples of the first category, generally permissive regions, are many sites toward the C-terminus, which have high permissibility to any domain insertion and where the impact on function is minimal, irrespective of insertion type. Universal permissibility likely means that these regions play minor roles in surface trafficking, oligomerization, or channel gating. This is consistent with their known role for interacting with binding partners, many of which are not present in HEK293 cells(99).

Examples for the second category, scaffold sites, include transmembrane helices, which have low permissibility to any insertion, and where any domain insertion severely impacts function. The universal disruptiveness of mutations within these regions is likely due to these regions being essential for folding, oligomerization, surface trafficking or membrane insertions. This is consistent with many scaffold sites occurring within transmembrane domains or at interfaces between channel monomers.

Sites in the third category have differential permissibility that depends on both the structural context of the insertion position as well as the biophysical properties of the inserted domain. We postulate that these sites have latent allosteric capacity. Indeed, significant (> 1 standard deviation) differential permissibility was measured in the βD-βE loop and the αH helix (**Figure 2-16a**). While the βD-βE loop has no known endogenous regulatory function in Kir2.1, it is part of an alcohol-binding pocket conserved in both Kir2.1 and GIRK(93,103). Functional analysis in GIRK revealed that alcohol modulates PIP_2-mediated channel activation in a G-protein independent way(104). Furthermore, the βD-βE loop is critical for mediating activating interactions between Gβγ and GIRK (**Figure 2-17b**)(88). That inserting PDZ or Cib81 domains, whose potential interaction surface is roughly similar to that of Gβγ's, into this loop resulted in an activating phenotype suggests that Gβγ modulation of GIRK can perhaps be thought of as two different processes: One process is specific binding mediated by residues that exist in GIRK but not in Kir2.1, and the second process being a mechanism for coupling this binding to channel opening that exists both in GIRK, and to a lesser extent in Kir2.1, because of the shared architecture of the C-terminal domain (CTD). This type of division of labor, where one set of sites encodes affinity, while the other set encodes a filter for efficacy has been described for several types of allosteric regulation, including Gα activation of GPCR[59] and ligand binding to bioamine receptors(105).

Our data also suggests that determining differentially permissive sites, as opposed to simply permissive sites(80), is useful to predict engineerable allosteric capacity. Inserting light-switchable domains into regions with significant differential permissibility rendered Kir2.1 activity sensitive to light (2 out of 2 regions; 2 out of 5 sites after linker optimization). Outside of the PIP_2 binding site –where Cib81 insertions resulted in almost non-functional channels– we did not observe this in regions in which permissibility was less sensitive to context of the inserted domain (0 out 4 sites), nor in other controls (0 out of 3 sites/controls). Our interpretation of this result is that mapping differential permissibility might represent a generalizable method to inform the *de novo* engineering of allosteric regulation in any protein. Further experiments are needed, of course, to fully test this idea. For example, by using permissibility mapping to render several endogenous ion channels sensitive to bio-orthogonal stimuli such as drugs and light. Such tools have great utility for understanding how ion channels sculpt the function and adaptation of neuromuscular tissues, in both normal and pathological contexts. To this end, our ability to build predictive models of permissibility also means that if models are trained on a sufficiently large experimental dataset, it might be possible to derive generalized predictive models that can predict permissibility on potentially any ion channel, thus rendering case-by-case mapping of permissibility superfluous.

It will be interesting to see if permissibility can be predicted from the same set of calculated properties for any ion channel (indicating permissibility is universal), or whether it is a function of phylogenetic distance (indicating permissibility is path dependent). Interpreted broadly, mapping and building models of permissibility–and by extension allostery–as it changes through phylogeny may be useful in explaining how specific ion channel families evolved. Much of the core functionality and architecture of ion channel families had evolved by the time the metazoan lineage appeared[21]. Subsequent diversification, driven by adaptive pressure to develop specialized neuromuscular tissue, can be considered fine-tuning biophysical properties and evolving regulation. We can observe this in K^+ channels (Figure 2-1). After the inward rectifier architecture (represented by Kir2.1) evolved from the simpler pore-only architecture (KcSA) by the addition of the CTD, the same overall architecture is utilized for different modes of allosteric regulation, including different small molecule ligands (PIP_2, ATP, Na^+) and proteins ($G\beta\gamma$ and SUR1). The notion of latent allosteric capacity can explain how this came to be. Dynamic features for the most part arise from global architecture fine-tuned by local interactions[61]. Allosteric capacity is an emergent by-product of these dynamics features[17]. It is likely that allosteric regulation schemes leverage pre-existing intrinsic properties of a protein's structural elements, since this is the path that requires the fewest mutations to implement this function to come under selection.

Methods

MuA domain insertion library generation

Transposition libraries were generated using 100ng MuA-BsaI engineered transposon and 1:2 molar ratio transposition target DNA in 20ul reactions with 4ul 5x MuA reaction buffer and 1ul 0.22 ug/ul MuA transposon (Thermo Fisher). MuA-BsaI engineered transposon propagation plasmid or pUCKanR-Mu-BsaI was a gift from David Savage (Addgene plasmid # 79769)(91). MuA-BsaI engineered transposon was digested with BglII and HindIII Fastdigest enzymes (Thermo Fisher) and gel purified using gel purification kit (Zymo Research).

The transposition target, human Kir2.1 (GI: 4504835, [https://www.ncbi.nlm.nih.gov/protein/NP_000882]) including a porcine teschovirus ribosomal skipping sequence (P2A)(106), was codon-optimized for mouse, synthesized

(Gen9) and subcloned into pATT-Dest using NEB BamHI and HindIII. pATT-Dest was a gift from David Savage (Addgene plasmid # 79770)^{43}. A FLAG tag was inserted after T115 using Q5 site-directed mutagenesis. MuA transposition reactions were incubated at 30degC for 18 hours for transposition, followed by 75degC for 10 minutes for heat inhibition. DNA from reactions was cleaned up (Zymo Research) and eluted in 10ul water. All 10ul were transformed into 30ul electrocompetent 10G ELITE E. coli (Lucigen) in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10uF, 600 Ohms, 1.8 kV). Cells were rescued and grown without antibiotics for 1 hour at 37degC. Aliquots were then serially diluted and plated on LB agar plates containing carbenicillin (100 ug/ml) and chloramphenicol (25 ug/ml) to assess library coverage. The remaining transformation mix was grown in 50 ml LB containing carbenicillin (100 ug/ml) and chloramphenicol (25 ug/ml). All transformed libraries yielded greater than 10^5 colonies, which for Kir2.1-P2A (1369bp) is >35x coverage. Plasmid DNA was purified by midi-prep kit (Zymo Research).

Transposition-inserted Kir2.1 variants were subcloned into an expression vector by amplifying channel variant genes adding on BsmbI sites, using 10 cycles of PCR using Primestar GXL (Takara Clontech) and run on a 1% agarose gel. The larger band was cut out and gel purified (Zymo Research) to isolate channels with inserted transposons. A mammalian expression vector (pcDNA3.1) with EGFP was amplified to add on BsmbI sites complementary to those on Kir2.1-P2A. The Kir2.1-P2A (BsaI-transposon) variants where subcloned into this vector by BsmbI-mediated Golden Gate cloning(107). Reactions were cleaned (Zymo Research) and eluted with 10ul water. All 10ul were transformed into 30 ul Lucigen electrocompetent 10G ELITE E. coli and electroporated in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 uF, 600 Ohms, 1.8 kV). Cells were rescued and grown without antibiotics for 1 hour at 37degC then with an aliquot serially diluted plated on LB agar plates containing kanamycin (50 ug/ml) and chloramphenicol (25 ug/ml) to assess library coverage. The remaining transformation mix was grown in LB containing kanamycin (50 ug/ml) and chloramphenicol (25 ug/ml). All transformed libraries yielded greater than 10^5 colonies so for Kir2.1 (1369bp) there is >35x coverage. Plasmid DNA was purified by midi-prep kit (Zymo Research).

Inserted Transposons were replaced with domains in individual reactions using BsaI-mediated Golden Gate cloning. Domains (PDZ (GI: 404931, [https://www.ncbi.nlm.nih.gov/protein/404931]), Cib81, GSAG$_{x2}$, GSAG$_{x3}$) for insertions were ordered as gblocks (IDT DNA), and PCR amplified to add on BsaI and linkers (Ala-Ser and Ser-Ala-Gly, preceding and following the domain insertion) sites complementary to MuA-BsaI transposon sites for Golden Gate cloning. Domain amplicons were gel purified (Zymo Research). The product was further digested with AgeI-HF (NEB) and Plasmid-Safe ATP-dependent DNase (Epicentre) to remove any undigested transposon, then cleaned up (Zymo Research) and eluted with 10ul water. All 10ul were transformed into 30 ul Lucigen electrocompetent 10G ELITE E. coli and electroporated in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 uF, 600 Ohms, 1.8 kV). Cells were rescued and grown without antibiotics for 1 hour at 37degC. An aliquot was serially diluted and plated LB agar plates containing kanamycin (50 ug/ml) to assess library coverage. The remaining transformation mix was grown in LB containing kanamycin (50 ug/ml). All transformed libraries yielded greater than 10^5 colonies so for Kir2.1 (1369bp) there is >35x coverage. Plasmid DNA was purified by midi-prep kit (Zymo Research).

Domain insertion permissibility cell sorting assay

100 ng of each domain insertion library was transfected with 36 ul of turbofect (Thermo Fisher) into 50% confluent HEK293FT (Invitrogen, R70007) with 11.9 ug of dummy plasmid (pATT Dest) divided across a single 6 well dish (9.6 cm^2 / well).

Cells from each well were detached using 1 ml accutase (Stemcell Technologies) and twice spun down at 450xg and resuspended in FACS buffer (2% of FBS, 0.1% NaN3, 1xPBS). Cells were incubated with 1:200 anti-flag mouse antibody (Sigma, F1804) 1 hour rocking at 4degC, washed twice with FACS buffer, covered with aluminum foil, and then incubated with 1:400 anti-mouse Alexa Fluorophore 568 (Thermo Fisher, A-11004) for 30 minutes rocking at 4degC. Cells were washed twice, resuspended in 3 ml FACS buffers, and filtered using cell strainer 5 ml tubes (Falcon). Cells were kept on ice and protected from light in the transfer to the flow cytometry core. Before cell sorting, a small aliquot of cells was saved as a control sample for sequencing.

Cells were sorted into EGFP high / Alexa568 low (transfected cells without surface expression) and EGFP high / Alexa Fluorophore 568 high (transfected cells with surface expression) on a BD FACSAria II P69500132 flow cytometer. EGFP fluorescence was excited using a 488 nm laser, recorded with a 525/50 bandpass filter and a 505 long pass filter. Alexa fluorophore 568 fluorescence was excited using a 561 nm laser and recorded with a 610/20 bandpass filter. Cells were gated on side scattering and forward scattering area to separate out whole HEK293FT cells, gated on forward scattering height and width to separate single cells, then gated on co-expressed EGFP to gate out cells that received a plasmid, then gated on cells that were labeled using the anti-flag antibody for surface expressed channels. Gates were determined using single wildtype, EGFP only and unstained library samples. A representative example of this gating scheme is shown in Fig 2-18. EGFP high / Label low and EGFP high / Label high cells were collected into catch buffer (20% of FBS, 0.1% NaN3, 1xPBS). Between 2,000-100,000 cells were collected for each sample/library pair which is ~4-250x coverage of all potentially productive (i.e., in-frame and forward) domain insertions.

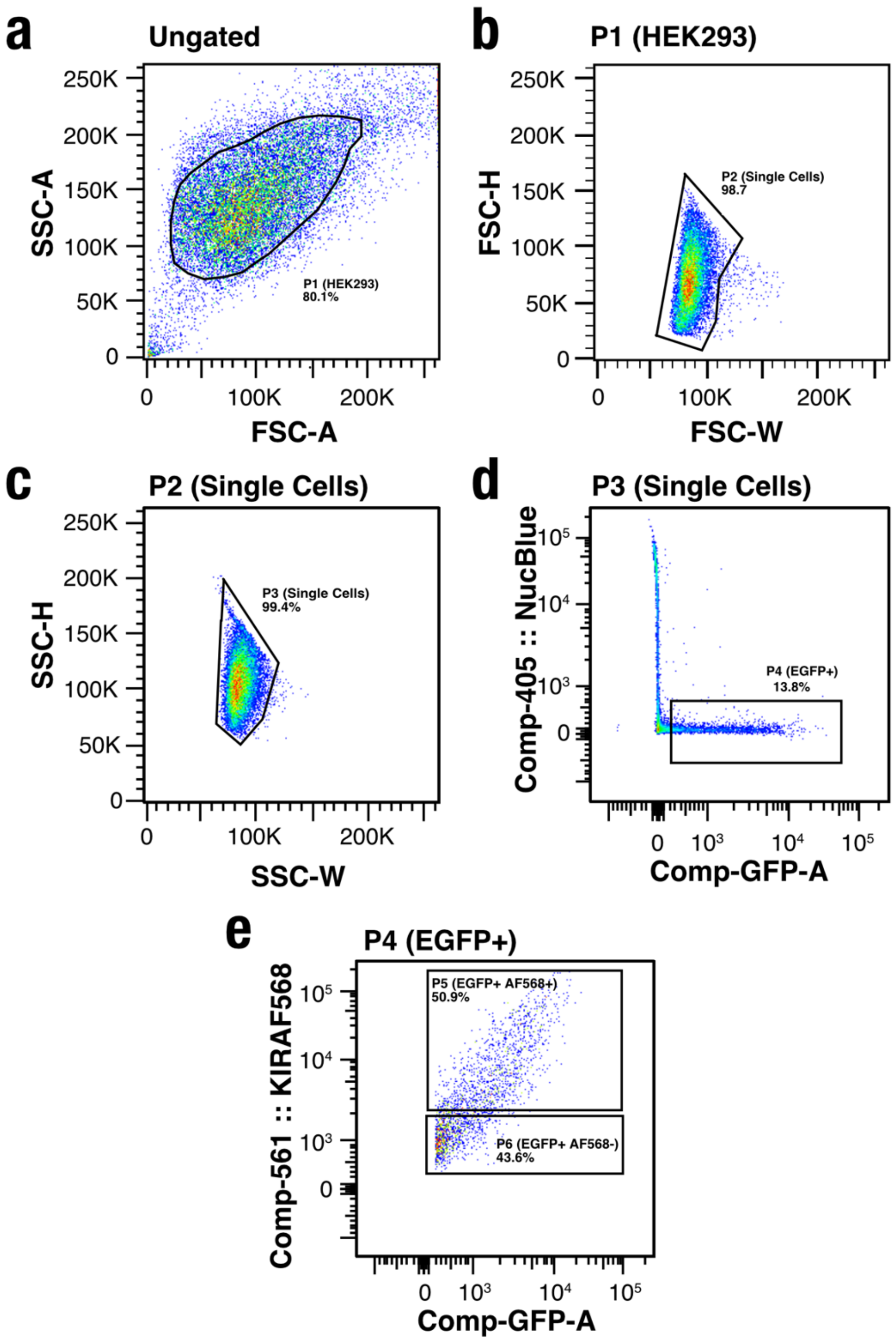


Figure 2-18: Permissibility Gating Scheme. (a) Cells are gated on side and forward scattering area to select whole HEK293 cells. **(b-c)** Forward and side scattering height and width are gated to select single cells. (**d**) Transfected cells are gated based on EGFP signal (Comp-GFP-A). **(e)** Finally, EGFP high / Label low and EGFP high / Label high populations are gated based on EGFP (Comp-GFP-A) and Alexa fluorophore 568 fluorescence (secondary antibody, Comp-561).DNA from Control, EGFP high / Label low, and EGFP high / Label high cells for each library were extracted using a Microprep DNA

kit (Zymo Research) and triple eluted with water. To remove chromosomal DNA, samples were digested with Plasmid-Safe ATP-dependent DNase (Epicentre). The resulting plasmid DNA was further purified and concentrated using (Zymo Research). The product was used as a template for 12 cycles of PCR using Primestar GXL (Takara Clontech), run on a 1% agarose gel, and gel purified (Zymo Research) to remove any primer dimers or none amplicon DNA. Purified DNA was quantified using Picogreen DNA concentration at the University of Minnesota Genomics Core. Equal amounts of each domain insertion sample were pooled by cell sorting category (control, EGFP high / Label low, EGFP high / Label high were pooled for sequencing library generation and sequencing.

Sequencing

Libraries were generated at University of Minnesota Genomics Core using Nextera XT or Nano Truseq library generation (Illumina) to fragment and add on Illumina sequencing adaptors and sequenced using either HISEQ or MISEQ sequencing platforms.

Domain insertion permissibility alignment and enrichment

Alignments were done on both forward and reverse reads using a DIP-Seq pipeline[32] developed by David Savage and coworkers that we slightly modified for compatibility with updated python packages. Reads with duplicate domain insertion calls in both forward and reverse reads were removed. This pipeline results in plaintext files indicating a domain insertion positions and whether that insertion is in-frame and in the forward direction. Enrichment was calculated by comparing the change in EGFP high /Label low to EGFP high / Label high cells. Only positions with reads in both samples were used in enrichment calculations. All these positions are treated as ‘NA’ and not considered in downstream analysis and structure mappings, with the exception of calculating correlations between datasets and correlations between sites. In these correlation calculations treat ‘NA’s as ‘0’s so removing all the data will introduce more noise when comparing between datasets due to limits from sampling.

Permissibility function for individual datasets comparing surface expressed (SE) and non-surface expressed (NSE) insertion variants:

$$F(i,j) = \frac{r^{i}_{j_{SE}}}{t_{j_{SE}}} - \frac{r^{i}_{j_{NSE}}}{t_{j_{NSE}}} \qquad (2)$$

Where *r* is the number of reads at amino acid position *i*, in the *j*th dataset divided by *t*, the total number of reads in the *j*th given sample. The resulting data from individual sequencing reads are only used to calculate correlations between domains and amino acid positions.

For structure mappings and predictive model training means of permissibility for a given domain insertion variant are used. So, the resulting mean permissibility function is:

$$G(i,j) = \frac{\sum \frac{r^i_{j_{SE}}}{t_{j_{SE}}} - \frac{r^i_{j_{NSE}}}{t_{j_{NSE}}}}{n} \qquad (3)$$

Where *r* is the number of reads at amino acid position *i*, in the *j*th dataset divided by *t,* the total number of reads in that given sample, summed for all replicates of that domain-channel combination, and divided by *n*, the number of datasets.

Mean permissibility was z-scored and mapped onto the structure of chicken Kir2.2 (PDB 3SPI) using Chimera[64]. Mapped dataset for Cib81, PDZ and GSAG$_{x2}$ linker had adequate coverage: 76.6% Cib81(333/435), 68.9% PDZ (300/435), and 76.7 % (334/435) of potential amino acid positions.

Dataset Correlations

Pearson correlations were used to calculate correlations between domain insertion datasets. Pearson correlations were also used to calculate correlations between amino acid positions across all datasets. Both these correlation matrices were calculated using the dataset that was trimmed to avoid sampling problems such that no more than 0.375 (6/16) datasets are where raw fitness was +10E-4 <x< -10E-4. These correlations were calculated with 63% of possible positions (277/435).

Spearman correlations were used to compare mean domain insertion datasets and calculated protein properties because Spearman correlations are often better at handling non-parametric correlations. Based on lack of structural and conservational data at various amino acid positions many sites had to be trimmed. Data was trimmed from positions where more than half datasets had a raw mean fitness was +10E-4 <x< -10E-4. This resulted in datasets that contained 70% of possible positions (207/293).

Computed protein properties

Static protein properties (B-factor, 10 Angstrom intra-monomeric packing density, 4.5 Angstrom intra-monomeric packing density, 4.5 Angstrom inter-monomeric packing density, and surface exposure) were calculated using the SWIFT web server(108) on chicken Kir2.2 (PDB: 3SPI)(87), conservation based properties (overall predicted disruptive effect of a mutation, conservation, and individual predicted disruptive effect of a mutation to specific amino acids (A, C, D, E, F, G, H, I, K, L, M, P, Q, R, S, T, V, W, Y)) were acquired from the EVmutation data server(109), and dynamic properties (first 20 normal modes of the B monomer) were calculated on the iGNM 2.0 normal mode webserver(110). After trimming, computed protein properties were calculated for all parameters for 67% (293/435) of possible positions.

Decision tree models

We chose decision trees for building predictive models due to their utility in handling and determining non-linear interactions. Prior to training models, data was binarized such that 0 was not permissive and 1 permissive for a given domain insertion. After trimming any data for a given mean raw fitness between +10E-4 <x< -10E-4 datasets models were trained on 62.5% (183/293) Cib81, 58.0% (170/293) PDZ, and 68.6% (201/293) of possible positions. Models were limited to a depth of 4 to minimize model overfitting, trained on the computed protein properties to predict Cib81, PDZ, and GSAG$_{x2}$ using the rpart package in R, and cross-validated 10 times. Model performance was determined using commonly used criteria: receiver operating characteristic (ROC) curves, the complexity parameter (used in minimizing tree size)/tree depth and model residuals, precision vs. recall, accuracy vs. cutoff, precision vs. cutoff, and recall vs. cutoff (**Figure 2-14**). As further validation, models were trained only the four most significant protein properties based on Spearman correlations to demonstrate necessity and utility of using decision trees vs correlation calculations (**Figure 2-13A**), by withholding the most important properties determined with decision tree and by withholding whole classes of protein properties (static, conservation, dynamic, **Figure 2-13B**).

Resting membrane potential functional assay

Single mutants were generated by inserting a BsaI site and a 5 basepair replication identical to those created by transposons that replicated the beginnings and ends of transposon-mediated domain insertions using a Q5 site-directed mutagenesis kit (NEB).

Single insertion mutants were created for 32 sites (Amino Acid positions: 23, 61, 62, 116, 153, 186, 188, 191, 207, 209, 217, 222, 224, 236, 237, 238, 239, 240, 253, 259, 264, 300, 306, 332, 335, 369, 370, 378, 385, 400, 401) and then replaced with the domains for which libraries were previously generated. Subsequently, using BsrGI and PstI sites, EGFP was replaced with miRFP670 for all mutants. miRFP670 was amplified from pmiRFP670-N1 which was a gift from Vladislav Verkhusha (Addgene plasmid # 79987)[69]. The same cloning approach was used to add 3-9 amino acid GSG linkers on either side of Cib81 in the 238 position.

The resting membrane potential assay(102) was conducted on all aforementioned domain insertion mutants in addition to miRFP670 alone (negative control) and wildtype Kir2.1 (positive control). 400 ng of each mutant was transfected with 6 ul Polyethyleneimine (Polysciences) along with 600 ng of dummy plasmid (pATT-Dest) across 2 wells of a 24 well dish. For each experiment, wildtype Kir2.1 was transfected as a benchmark and for comparison for mutant function. Cells from each well were detached using 300 ul accutase (Stemcell Technologies), spun down at 450xg three times, and re-suspended in 200 ul Tyrode (125mM NaCl, 2mM KCl, 3mM CaCl2, 1mM MgCl, 10mM HEPES, 30mM glucose, pH 7.3). Bis-[**1**,**3**-dibutylbarbituric acid] trimethine oxonol (DiBAC**4**(**3**), Thermo Fisher) was added to a final concentration of 950 nM, and cells were filtered in 5 ml cell strainer tubes (Falcon). DiBAC**4**(**3**) was diluted every day to exchange buffers from DMSO to Tyrode. Cells were kept on ice and protected from light in the transfer to the flow cytometry core.

Each sample was run in entirety on a BD Fortessa H0081 flow cytometer. DiBAC**4**(**3**) was excited at 488 nm and recorded at 525/50 bandpass, and miRFP670 fluorescence was excited at 640 nm and recorded with a 670/30 nm bandpass filter. Cells were gated on side scattering and forward scattering area to separate out whole HEK293FT cells, gated on forward scattering area and height to separate single cells, then gated on co-expressed miRFP670 to gate out cells that received a plasmid, then a gate was set on the lower 50% of a histogram of wildtype Kir2.1 function, all mutants percentage of cells in this gate are reported. The analysis was performed in FlowJo 10 (FlowJo, LLC). A representative example of this gating scheme is shown in **Figure 2-19**.

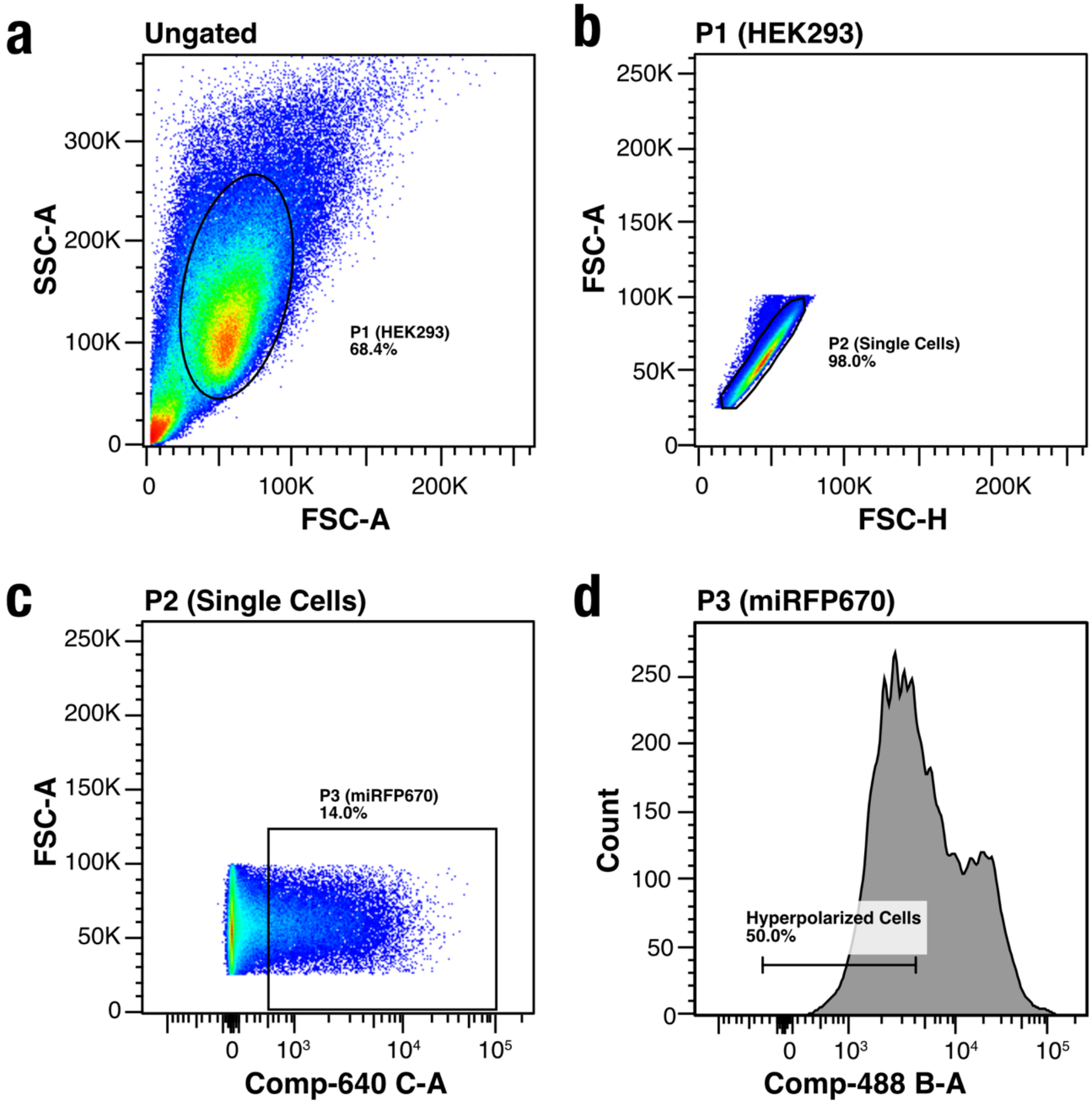


Figure 2-19: Voltage Assay Gating Scheme. (a) Cells are gated on side and forward scattering area to select whole HEK293 cells. (b) Next, cells are gated on forward scattering area and height to select single cells. **(c)** Transfected cells are gated based on miRFP670 signal (Comp-640 B-A). **(d)** Finally, a gate was set on the lower 50% of cells (corresponding to more hyperpolarized cells) based on DiBAC4(3) fluorescence (Comp-488 B-A) in wildtype Kir2.1 samples. The same gate was applied to all mutants and percentage of cells in this gate are reported.

Flow Cytometry Assay for Light-modulation of Kir2.1 function

The generation of all single mutants used in the optogenetic switching assay was previously described in the resting membrane potential assay methods section. A Cry2-

P2A-mKate2 domain was generated using gene fragments (Gen9) amplified with and assembled into the expression vector pEGFPN3 (Invitrogen) using BsmbI-mediated Golden Gate cloning. The Kir2.1(V302M) mutant was generated using Q5 site-directed mutagenesis (NEB).

The light-modulation assay was conducted for Cib81 mutants chosen as representative examples for the various permissibility and functional phenotypes we had observed. In addition, negative controls such as wildtype Kir2.1 and a pore dead mutant V302M were included. 4 ug of each mutant, 3ug of dummy plasmid (pATT-Dest), and 100 ug of Cry2-P2A-mKate2 were transiently transfected using 6ul PEI across 16 wells of a 24 well dish at 20% confluency. Cells from each well were detached using 300 ml accutase, washed three times, and resuspended in 4ml Tyrode. DiBAC**4**(**3**) was added to a final concentration of 950 nM, and cells were filtered in cell strainer 5 ml tubes (Falcon). Filtered cells were divided into twelve 5 ml tubes (300ul each) and kept on ice and protected from light in the transfer to the flow cytometry core.

Cells expressing each mutant, WT Kir2.1, Kir2.1(V302M) were challenged by the addition of K- gluconate at different concentrations (5 mM, 10 mM, 15 mM, 25 mM, 40 mM, and 70 mM), with and without illumination (455 nm LED (Thorlabs), 30 seconds, 100% duty cycle, 100uW/mm^2). Each sample was run in entirety on a BD Fortessa H0081 flow cytometer. DiBAC**4**(**3**) was excited at 488 nm and recorded with 525/50 bandpass and 502 long pass filters, miRFP670 was excited at 640 nm and recorded with a 670/30 bandpass filter, and mKate2 was excited at 561 nm and recorded at 610/20 bandpass and 595 long pass filters.

Each sample was recorded for 5 minutes or until completion. Cells were gated on Side Scattering and Forward Scattering to separate out whole HEK293FT cells, gated on forward scattering area and width to separate single cells, then gated on co-expressed miRFP670 (Kir2.1 mutant) to gate out cells that received a mutant plasmid. For each paired sample (dark and light) a custom gate was created in the non-illuminated sample to include the 15% most hyperpolarized cells (using the flowStats package). The number of events falling into this gate were then compared to the corresponding illuminated sample using the Chi-Squared test and reported as Dissimilarity (X^2, light vs. dark). Dissimilarities at different K$^+$ challenges were normalized to correct for photobleaching and averaged. A representative example of this gating scheme is shown in **Figure 2-20**.

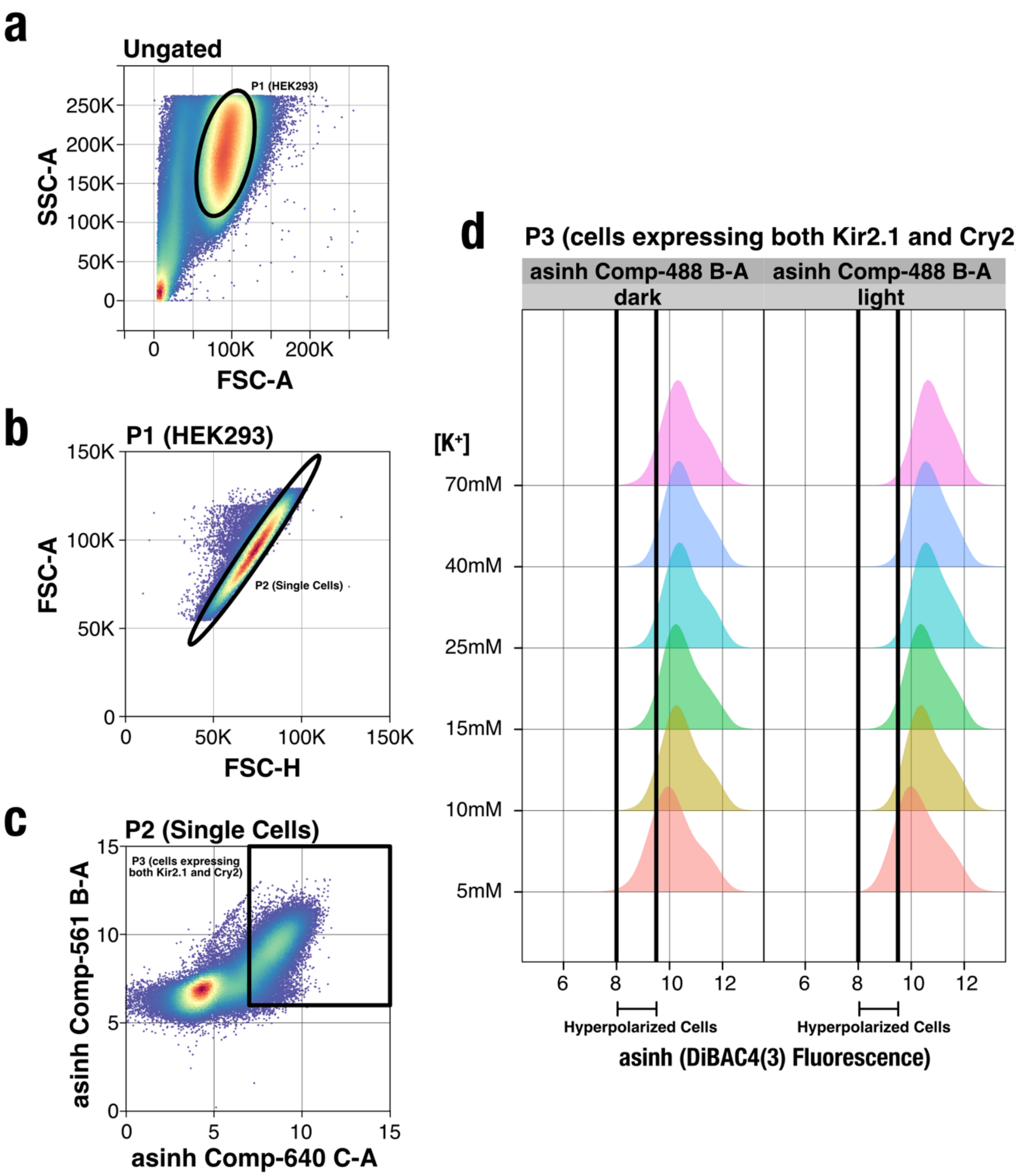


Figure 2-20: Light-Switching Gating Scheme. (a) Cells are gated on side and forward scattering area to select whole HEK293 cells. **(b)** Next, cells are gated on forward scattering area and height to select single cells. **(c)** Double-transfected cells (expressing both Cry2 and Kir2.1) are gated based on mKate2 (Comp-561 B-A) and miRFP670 signal (Comp-640 C-A), respectively. **(d)** Finally, for a given mutant (e.g., wildtype Kir2.1) and a given K^+ challenge, a custom gate was created in the non-illuminated (dark) sample corresponding to the 15% most hyperpolarized cells (reported by DiBAC**4(3)**, Comp-488

B-A). The number of events falling into this gate were compared to the corresponding illuminated sample using the Chi-Squared test and reported as Dissimilarity (χ^2, light vs. dark).

Patch Clamp Electrophysiology

HEK293FT cells were transiently transfected with Kir2.1 (WT) or Kir2.1 insertion mutant and Cry2-P2A-mKate using PEI. Cells were screened for mKate2 expression using a 565nm high-power LED (Thorlabs) filtered by a 560±40nm bandpass filter (Semrock) through a 40X lens. K^+ currents were recorded 36-48 hours post-transfection using on-cell patch clamp electrophysiology. Patches with clear channel activity were stimulated with blue (455nm) light delivered by a LED (Thorlabs) at 100 uW/mm^2 for 50 seconds at 100% duty cycle. Analog signals were filtered (2-5 kHz) using the built-in 4-pole Bessel filter of a Sutter Instrument IPA patch clamp amplifier, digitized and stored. Bath solution contained 125mM NaCl, 2mM KCl, 3mM $CaCl_2$, 1mM $MgCl_2$, 10mM HEPES, 30mM glucose, adjusted to pH 7.3 with NaOH. The pipette solution contained: 125mM K-Gluconate, 8mM NaCl, 0.1mM $CaCl_2$, 0.6mM $MgCl_2$, 1mM EGTA, 10mM HEPES, 4mM Mg-ATP, 0.4mM Na-GTP, adjusted to pH 7.3 with KOH. Osmolarity was adjusted to 295 - 300 mOsm with sucrose. Electrodes were drawn from borosilicate patch glass (Warner Instruments) to a resistance of 2-6 MΩ. Data analysis was done using custom R scripts.

Chapter 3

Note this was originally published

Coyote-Maestas W, Nedrud D, Okorafor S, He Y, Schmidt D. Targeted insertional mutagenesis libraries for deep domain insertion profiling. Nucleic Acids Research. 2020 Jan 24;48(2):e11–e11.

Author Contributions
W.C.M., D.N., and D.S. conceived the study. D.N. wrote custom software for OLS pool design. W.C.M. and D.N. generated insertion libraries. W.C.M. and D.N. conducted domain insertion permissibility experiments and flow cytometry functional assays. S.O. and Y.H. provided technical support for all experiments. W.C.M., D.N., and D.S. co-wrote the manuscript. This paper is referenced in the theses of both W.C.M. and D.N.

Targeted insertional mutagenesis libraries for deep domain insertion profiling

INTRODUCTION

Up to 80% of metazoan proteins consist of multiple protein domains (111,112). Protein domains are independent units that retain their structure and function (113) as the 'words' of the protein universe (114). Domain recombination is, therefore, an essential process in protein evolution (115,116). Ion channels are a good example of how domain recombination helped rapidly expand functional diversity in the metazoan lineage (53). Inward rectifier K^+ channels, for example, arose early in cellular life from the combination of a pore domain and a phylogenetically ancient immunoglobulin (Ig)-like domain (53,117), to which different allosteric ligand can bind and affect gating of the pore domain (83).

In biomedical engineering, domain recombination is used to generate synthetic proteins. Many biosensors (118,119) are made by functionally coupling domains that sense a stimulus (e.g., ligand binding, voltage, aberrant protein activity) and domains that report these events (e.g., emitting photons, alter gene expression, induce apoptosis). Antibodies are joined end-to-end with signaling domains to create chimeric T-cell receptors for immunotherapy (21). Domain recombination enables the design of programmable circuits from multi-domain proteins in living cells (120,121). We recently discovered that domain insertion provides a window into protein dynamics and allostery in ion channels, and it allowed us to generate a light-switchable Inward Rectifier K^+ channel (13).

Despite the significance of domain recombination in biology and biomedical engineering, saturated domain recombination remains an unsolved problem. By saturated we mean an unbiased approach that redundantly samples all possible insertions of a donor domain into a target protein. To see why saturated approaches are crucial, we should consider that both single amino acid mutations and domain insertions can alter protein structure / function relationships. By comprehensively mapping the impact of these variations, using deep scanning mutagenesis (10) or differential domain insertion profiling (13), we may reveal intrinsic protein properties (31,109,122), improve our understanding of the mechanistic basis of protein function (30,123), and guide protein engineering (76,124,125).

Many pioneering contributions have been made to this field but none enable saturated domain recombination. Random insertion approaches include overlap PCR (126,127) and limited nuclease digest and nonhomologous recombination (128–130). However, both approaches are inefficient and endonuclease-assisted approaches result in numerous tandem duplications and deletions at insertion sites. Another approach, transposon-mediated domain insertion (72,131–133), is useful for probing the structure and function of proteins (134,135) (including ion channels (136)), generating new fluorescent proteins (137), or circularly permutating proteins (138,139). The current state of the art is Domain Insertion Profiling through Sequencing (DIP-seq) (76), which combines MuA transposase-assisted library generation with high throughput assays for linking genotype (insertion position) to a phenotype (protein folding, abundance, localization, etc.). DIP-seq has been used to engineer a ligand-sensitive Cas9 (80), a light-switchable ion channel (80), and transcription factors (140).

Transposases, including MuA, have sequence bias (98,135,141–145), which results in domain insertion libraries with inconsistencies in insertion frequencies and regions without insertions (12,80,139). Additionally, transposases target random DNA sequences, causing 5 in 6 insertions to be in the incorrect reading frame or wrong direction, and the MuA transposition mechanism results in an unavoidable 5 bp replication at the insertion site (90,146). Similar to sequence coverage and depth in genomic analyses (147,148), insertion bias, incomplete coverage, and low redundancy of domain insertion libraries lead to sampling errors that decrease the quality of downstream functional data (149).

Here, we developed a method for domain insertion called **S**aturated **P**rogrammable **In**sertion **E**ngineering (SPINE). Unlike existing insertional mutagenesis approaches, which rely on the randomness of recombination or transposition, SPINE is a programmed method. It works by dividing a targeted gene into fragments and replacing each fragment with a microarray-synthesized oligo library (150,151). Each oligo in this library contains a genetic handle that can be replaced with a domain of interest by Golden Gate cloning (152). SPINE overcomes many constraints of previous approaches and generates unbiased, saturated, and targeted domain insertion libraries. These improved libraries result in less missing data and improve the dynamic range of assays that measure domain insertion permissibility, which measures the impact of domain insertion on target protein expression.

MATERIALS AND METHODS

OLS in silico design

Oligo sequences are generated using a custom algorithm (written for Python 3.7.3. and available at https://github.umn.edu/schmidt-lab/SPINE) as follows:

Target Gene Fragmentation (**Figure 3-1A**): Target gene sequences are submitted in FASTA format. Gene start and end positions within the plasmid are entered manually or calculated from a selected open reading frame. Each gene is divided into evenly distributed fragments to the nearest codon such that the length of each gene fragment does not exceed the length limitations of the synthesized oligo pool (in our case 230bp) minus additional required components: subpool amplification barcodes (2x12 bp), restriction sites (2x7 bp), and the domain insertion handle (24 bp). Each fragment break site is adjusted to create unique cut site overhangs for Golden Gate cloning. If adjusting one fragment position causes any fragment to exceed the maximal length, the other fragments are adjusted to equalize fragment distribution below this length threshold.

Target gene primer design for inverse PCR (**Figure 3-1B**): Forward and reverse plasmid primers are designed to amplify the backbone for each target gene fragment. Additional non-annealing sequences are added to the primer's 5 prime end encoding for inward-facing BsmBI recognition sites with the cut site including the first and last codon of the fragment (3 bases) plus one base extension for the 4 base cut site. These primers are optimized for melting temperature and specificity by adjusting the length of the 3 prime end. Melting temperatures are set between 55˚C and 61˚C based on calculations from both Sugimoto et al. (1996) (153) and SantaLucia & Hicks (2004) (154) (58). A primer is flagged as non-specific if annealing temperatures are greater than 35˚C at any other position in the plasmid. Non-specific primers are made specific by extending the primer or, if max melting temperatures are exceeded, the fragmented site is adjusted.

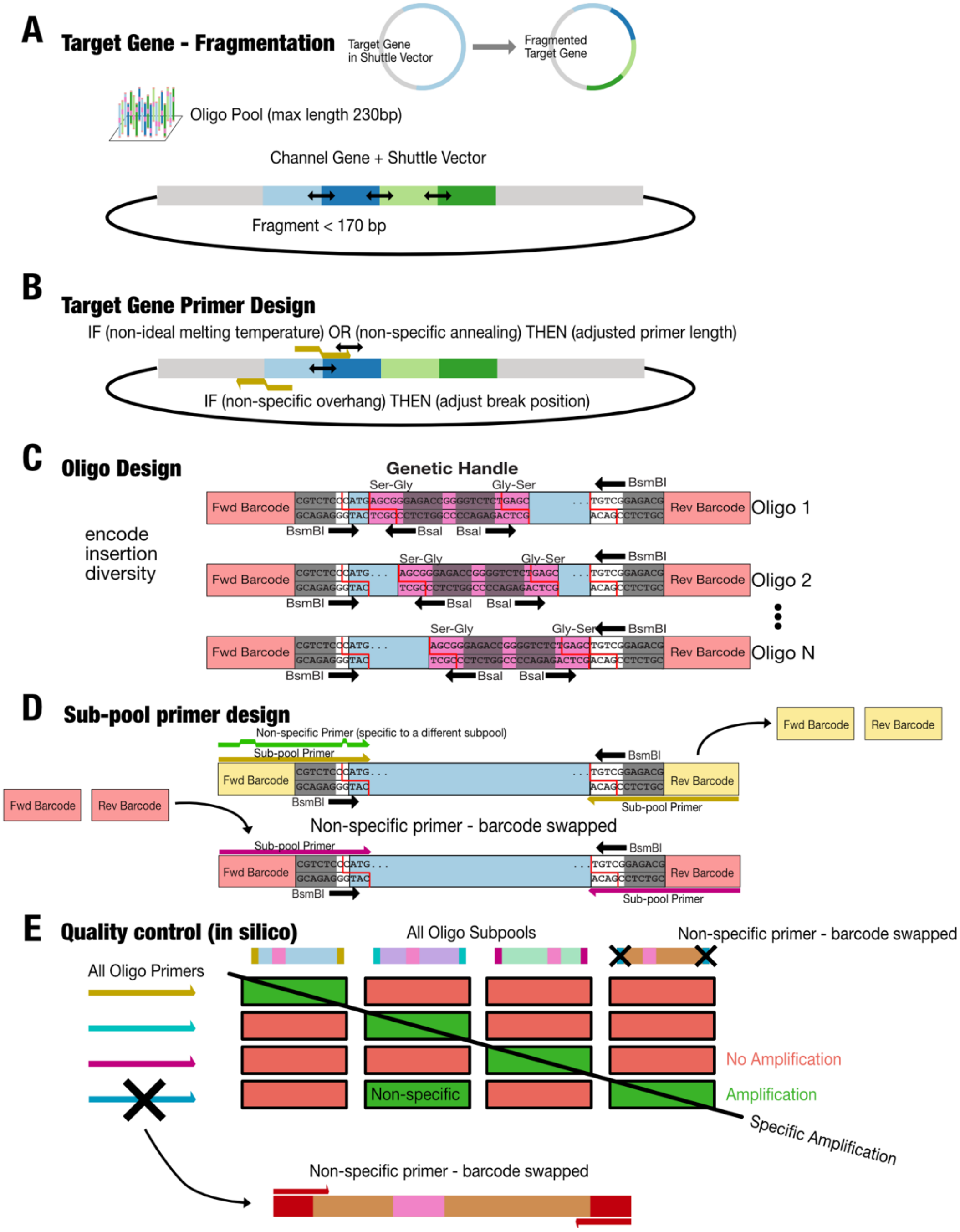

A
Target Gene - Fragmentation
Target Gene in Shuttle Vector
Fragmented Target Gene
Oligo Pool (max length 230bp)
Channel Gene + Shuttle Vector
Fragment < 170 bp
B
Target Gene Primer Design
IF (non-ideal melting temperature) OR (non-specific annealing) THEN (adjusted primer length)
IF (non-specific overhang) THEN (adjust break position)
C
Oligo Design
Genetic Handle
encode insertion diversity
Ser-Gly
Gly-Ser
BsmBI
BsaI
Fwd Barcode
Rev Barcode
CGTCTCCATGAGCGGGAGACCGGGTCTCTGAGC
GCAGAGGGTACTCGCCCTCTGGCCCAGAGACTCG
TGTCGGAGACG
ACAGCCTCTGC
Oligo 1
Oligo 2
Oligo N
D
Sub-pool primer design
Non-specific Primer (specific to a different subpool)
Sub-pool Primer
Fwd Barcode
Rev Barcode
Non-specific primer - barcode swapped
E
Quality control (in silico)
All Oligo Subpools
Non-specific primer - barcode swapped
All Oligo Primers
Non-specific
No Amplification
Amplification
Specific Amplification
Non-specific primer - barcode swapped

Figure 3-1: In silico design of oligos and primers. (**A**) A target gene (within its shuttle vector) is fragmented such that all fragments are less than the maximum oligo size minus the barcodes, restriction enzyme sequence, and genetic handle. Fragment break sites are adjusted for unique restriction enzyme cut overhangs. (**B**) A set of gene primers are designed for each fragment for inverse PCR. These primers will amplify everything except

the fragment and add an inward-facing BsmBI recognition site. (**C**) An oligo pool is designed for each fragment and within the pool an oligo is designed for each insertion position within that fragment. Each oligo consists of the fragment sequence it is replacing, sub-pool specific amplification barcodes, inward-facing BsmBI site that will match the cut site of the gene primers, and a genetic handle at every position in the gene. The genetic handle contains outward-facing BsaI recognition site for replacement with a domain of interest. (**D**) To retrieve a specific sub-pool of oligos, primers are designed based on bio-orthogonal barcodes. This amplification is made specific by swapping barcodes until unique amplification is found. (**E**) When combining the subpools from many genes, there is a chance of non-specific amplification. Quality control is performed on every oligo primer and oligo subpool for non-specific amplification. If found, the barcode is swapped for unique amplification.

<u>Design oligos that encode each insertion site (**Figure 3-1C**):</u> For each gene fragment, a loop is run to generate an oligo for each insertion position within that fragment, starting after the first codon and ending before the last codon to account for the cloning cut sites. Therefore, sequential fragments overlap by one codon. Oligos consist of a bio-orthogonal barcode for specific subpool amplification, BsmBI recognition sites, and the fragment sequence with a genetic handle insertion (**Figure 3-2B**). The genetic handle contains outward-facing BsaI restriction sites, which enable replacement of the handle with a domain of interest, and Ser-Gly and Gly-Ser flexible linkers at the beginning and end of the handle, respectively. Barcodes are courtesy of the Elledge lab (59). In detail, each oligo starts with a forward subpool specific barcode, appended with a forward-facing BsmBI recognition sequence plus one base to bring the cut site into frame. Next, the oligo is appended with the fragment sequence with the insertion handle inserted at the next amino acid position following the previous oligo. Finally, after the gene fragment section one base is added to bring the cut site into frame followed by a reverse facing BsmBI sequence, and a reverse subpool specific barcode.

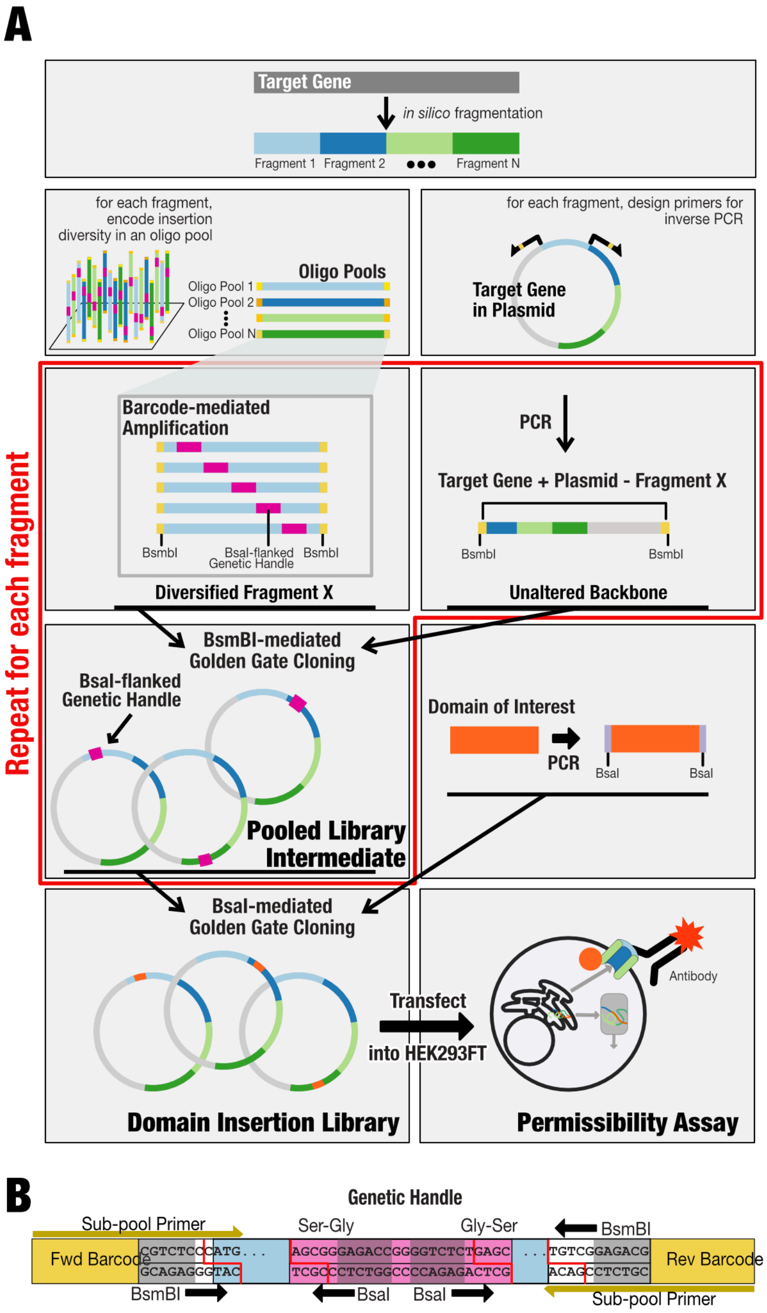


Figure 3-2: A) SPINE workflow. A target gene sequence is divided into shorter fragments. For each fragment, an oligo pool is generated with a genetic handle (purple) at each amino acid position. Flanking barcodes (different hues of yellow) mediate specific amplification of each subpool, which is then joined with the PCR-amplified target gene backbone in BsmBI-mediated Golden Gate cloning. This process is repeated for each fragment, and the resulting intermediate libraries are pooled. The genetic handle is replaced by a domain of interest (orange) through BsaI-mediated Golden Gate cloning, resulting in the final domain insertion library. **B) Barcode Design.** Each OLS subpool is designed with a bio-

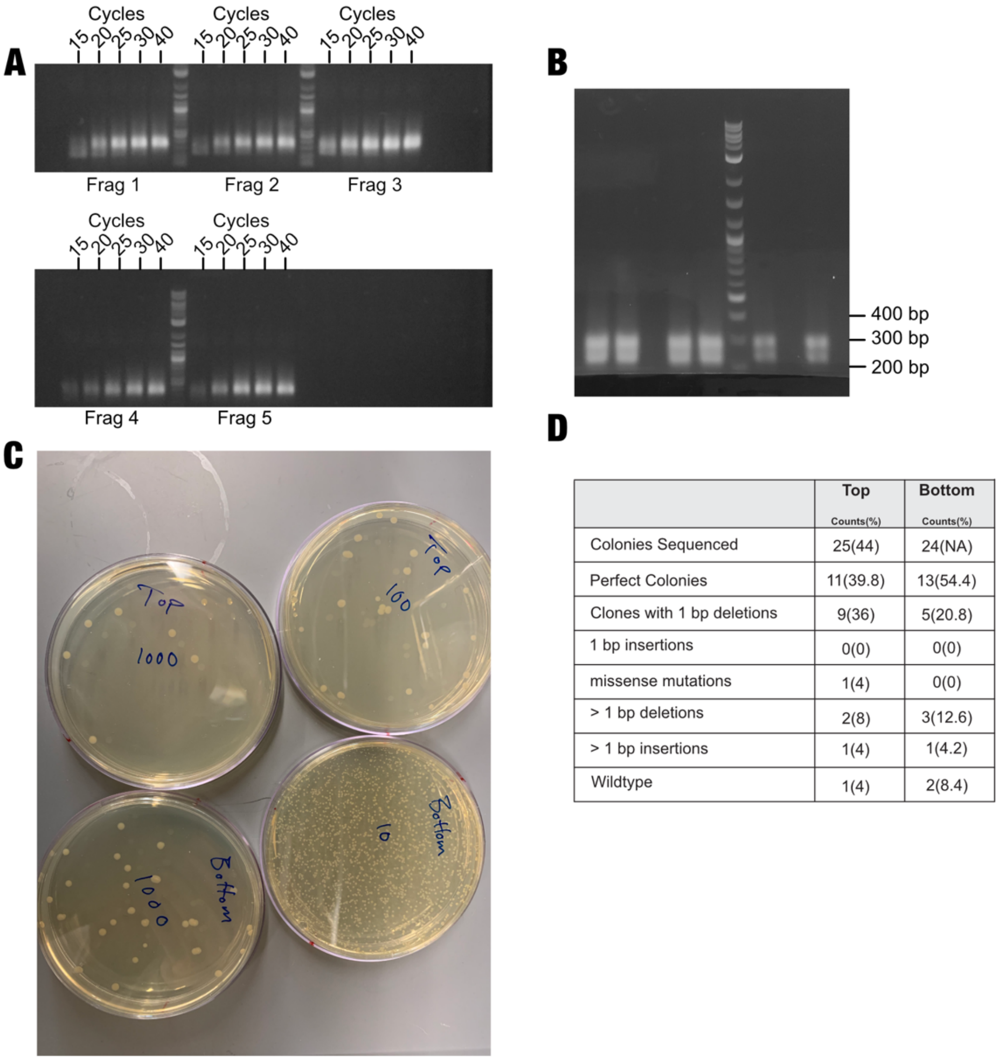
orthogonal barcode followed by a BsmBI recognition site that cuts within the sequence of a gene. Every barcode and BsmBI cut site are unique to a given subpool minimizing the chance for undesired assembly. The genetic handle is designed with outward-facing BsaI recognition sites that enable cutting within the beginning and ends of short flexible serine-glycine linkers. These linkers are the only scars that result from assembly and can be programmed to be any sequence at least 4 bp long.

Design of subpool amplifying oligos (**Figure 3-1D**): Forward and reverse subpool specific oligo primers are generated by testing annealing of a candidate primer sequence to the respective barcode, BsmBI recognition, and cut sequence. These primers are optimized for annealing temperature as described above, however, because the 3' end is limited to the cut site, melting temperatures are optimized by adjusting the 5' end or swapping the barcode sequence.

In Silico Quality Control (**Figure 3-1E**): A final *in silico* quality control is run to check for creation of new BsaI or BsmBI recognition sites and check for non-specific subpool primers across all oligos. If a BsaI or BsmBI recognition site is created, a codon within that recognition site will be changed to an alternative codon maintaining the amino acid sequence. Non-specific subpool primers are identified by an annealing temperature above 35˚C for any position in any oligo other than the designed position. If a primer is non-specific, that subpool amplification barcode is replaced with another barcode and quality control is repeated. All oligos and primers are exported as FASTA files for ordering.

Oligo Library Synthesis (OLS) Pool Amplification

A 7.5K oligo library synthesis (OLS) pool containing the 2,099 oligos for four target proteins (human Kir2.1 (Accession: NP_000882), Drosophila melanogaster Shaker (Accession: NP_728123), human α7 nAChR (Accession: NP_000735.1), and human ASIC1a (Accession: NP_001086.2)) was synthesized by Agilent and received as 10 pmoles of lyophilized DNA. This DNA was resuspended in 500 µl TE. OLS subpools corresponding to a given gene fragment were PCR amplified using Primestar GXL DNA polymerase (Takara Bio) according to the manufacturer's instructions in 50 µl reactions using 1 µl of the OLS pool as the template and 25 cycles of PCR. The entire PCR reaction was run on 1% agarose gels, visualized with Sybr safe (ThermoFisher), and gel purified (Zymo Research). See also **Figure 3-3**.



	Top Counts(%)	Bottom Counts(%)
Colonies Sequenced	25(44)	24(NA)
Perfect Colonies	11(39.8)	13(54.4)
Clones with 1 bp deletions	9(36)	5(20.8)
1 bp insertions	0(0)	0(0)
missense mutations	1(4)	0(0)
> 1 bp deletions	2(8)	3(12.6)
> 1 bp insertions	1(4)	1(4.2)
Wildtype	1(4)	2(8.4)

Figure 3-3: Optimization of OLS Amplification and purification. OLS subpool amplification is sensitive to the number of PCR cycles, and overamplification can result in library bias and unintended side products. Prior to amplifying all subpools, we therefore optimized the number of PCR cycles. (**A**) Using PrimeStar GXL (Takara Clontech), we tested 15, 20, 25, 30, 40 cycles on five OLS subpools. We decided on 25 cycles as there was sufficient PCR product in all reactions. **(B)** However, we noticed that in many of the reactions there were two PCR bands with the upper bands becoming dominant with more PCR cycles. This upper band is likely an artifact of PCR due to its dominance at higher cycles and it is larger than the expected ~230 bp product. **(C)** To test which band yields better transformation efficiencies, we gel-purified both PCR products for Golden Gate cloning into the recipient backbone, and then transformed the resulting product into chemically competent E.coli cells. Transformed E.coli were plated at different dilution factor (noted on each plate along with the band used). We found that the bottom band

yielded about 4x the transformation efficiency. **(D)** We also submitted 25 colonies from both transformations for Sanger Sequencing to test if the different products result in different insertion library fidelity. We found that mutation and wildtype rates were comparable. Based on these results, we suggest using 25 cycles and gel purifying the lower band after OLS amplification for the Deep Domain Insertion Profiling protocol to increase library diversity.

Combining OLS fragments and target gene backbone

To insert the OLS subpools into target gene backbones, complementary BsmBI sites to those on the OLS fragments of a respective subpool were added by PCR using Primerstar GXL DNA polymerase (Takara) and 100pg of wildtype channel as template DNA (**Figure 3-4A**). PCR products were run on 1% agarose gels, visualized with Sybr safe (ThermoFisher) and gel purified (Zymo Research) to remove any undesired PCR by-products.

Target gene backbone PCR product with added BsmBI sites and the corresponding OLS subpools were assembled using BsmBI-mediated Golden Gate cloning (56) (**Figure 3-4B**). Each 20ul Golden Gate reaction was composed of 100ng of backbone DNA, 20ng of OLS subpool DNA, 0.2ul BsmBI (New England Biolabs), 0.4ul T4 DNA ligase (New England Biolabs), 2ul T4 DNA ligase buffer, and 2ul 10mg/ml BSA (New England Biolabs). These reactions were placed in a thermocycler overnight with following program: (1) 5 minutes at 42degC, (2) 10 minutes at 16degC, (3) repeat (1) and (2) 40 times, (4) 42degC for 20 minutes, (5) 80degC for 10 minutes. Reactions were cleaned up using ZymoResearch Clean and Concentration kits, eluted in 10ul of elution buffer, transformed into E. cloni® 10G chemically competent cells (Lucigen) according to manufacturer's instructions. Cells were grown overnight at 30degC to avoid overgrowth in 50mL LB with 40 µg/mL kanamycin with shaking, and library DNA was isolated by miniprep (ZymoResearch). A small subset of the transformed cells was plated at varying cell density to assess transformation efficiency and validate successful insertions with colony PCR. All libraries at this step yielded >7,000 colonies corresponding to >45x coverage for perfect mutations assuming one-third of mutants are perfect. All libraries (corresponding to different subpools) of a given target gene were pooled together at an equimolar ratio, resulting in a mixture of insertions at every amino acid position (**Figure 3-4C**).

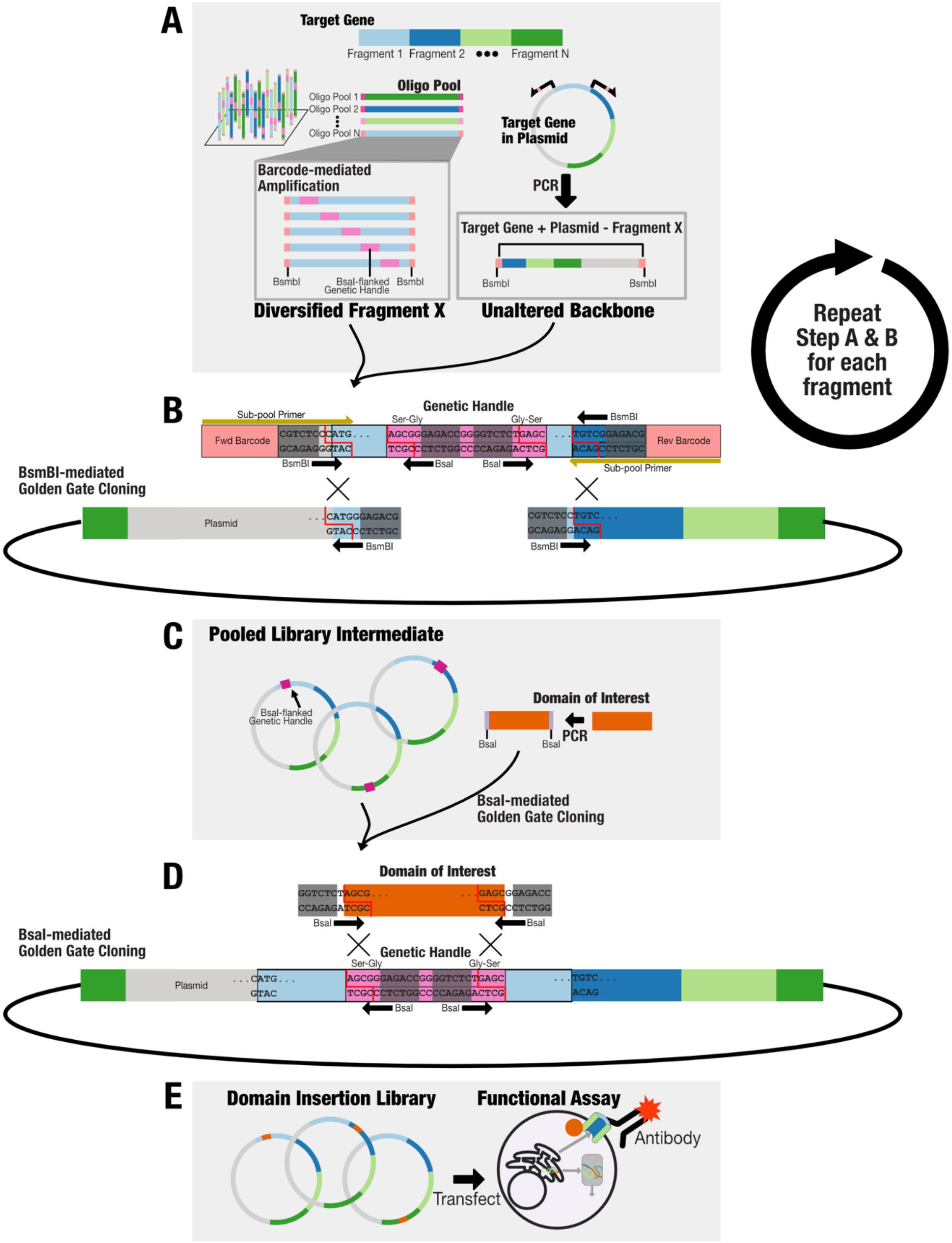


Figure 3-4: Detailed library assembly. (A) Amplification of oligo subpool and inverse PCR of the target gene and shuttle vector. **(B)** These two amplicons are combined by Golden Gate assembly using BsmBI restriction enzymes. Steps **A** and **B** are repeated for every fragment in the gene. **(C)** All assembled supools (result from step **A** and b for each

fragment) are mixed in equimolar ratio to yield an Intermediate Library. This mixture contains a genetic handle at every position in the gene and this can be replaced with any domain of interest by adding complementary BsaI recognition sites. **(D)** The domain of interest replaces the genetic handle in the Intermediate Library via Golden Gate cloning. **(E)** This yields the final SPINE library, which is then subjected to a functional assay.

Replacing the genetic handle with the domain of interest for ASIC1a, Shaker, and α7nAChR

Cib81 (60) was ordered as a gBlock (IDT DNA). BsaI sites complementary to those in the inserted genetic handle were added to Cib81 by PCR using Primestar Max (Takara Bio) according to the manufacturer's instructions (**Figure 3-4D**). The genetic handle in each target gene insertion library was replaced with Cib81 by BsaI-mediated Golden Gate cloning. Each 20ul Golden Gate reaction contained 100ng of backbone DNA, 15 ng of Cib81 DNA, 0.2 µl BsaI-HFv2 (New England Biolabs), 0.4 µl T4 DNA ligase (New England Biolabs), 2 µl T4 DNA ligase buffer, and 2 µl 10mg/ml BSA. These reactions were placed in a thermocycler overnight with following program: (1) 5 minutes at 37degC, (2) 10 minutes at 16degC, (3) repeat (1) and (2) 40 times, (4) 37degC for 20 minutes, (5) 80degC for 10 minutes. Reactions were cleaned up using ZymoResearch Clean and Concentration kits, eluted in 10ul of elution buffer, transformed into E. cloni® ELITE electrocompetent cells competent cells (Lucigen) in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ohms, 1.8 kV). Cells were grown overnight at 30degC to avoid overgrowth in 50mL LB with 40 µg/mL kanamycin with shaking, and library DNA was isolated by miniprep (ZymoResearch). A small subset of the transformed cells was plated at varying cell density to assess transformation efficiency and validate successful insertions with colony PCR. All libraries at this step yielded >7,000 colonies corresponding to >45x coverage for perfect mutations assuming one-third of mutants are perfect.

Replacing the genetic handle with the domain of interest for Kir2.1

We noticed that our libraries had contaminating WT DNA, which was likely due to trace amounts of template DNA left over from PCR amplification of target gene backbones, and which became enriched from multiple transformations. In preparation for the functional assay on Kir2.1-Cib81, we added an antibiotic selection step to remove WT DNA and enrich insertion variants. A chloramphenicol antibiotic cassette was amplified by PCR with primers to add BsaI sites complementary to the genetic handle, and outward-facing BsmBI sites, which enable replacement of the antibiotic cassette with a domain of interest, in this case, Cib81. BsaI-mediated Golden Gate followed the same scheme as replacing the genetic handle with the chloramphenicol antibiotic cassette. We transformed this Golden Gate reaction into E. cloni® 10G ELITE electrocompetent cells in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ohms, 1.8 kV). Cells were grown overnight at 30degC in 50mL LB with µg/mL kanamycin and 25 µg/mL Chloramphenicol LB with shaking to avoid overgrowth. Library DNA was isolated by midiprep (Zymo Research). A small subset of the transformed cells was plated at varying concentrations of cells to assess transformation efficiency and validate successful insertions with colony PCR. This library yielded >100,000 colonies corresponding to >300X coverage for perfect mutations assuming one-third of mutants are perfect.

We PCR-amplified Cib81 with BsmBI sites complementary to the antibiotic cassette. This antibiotic cassette was replaced with PCR amplified Cib81 using BsmBI-mediated Golden Gate as described above. Libraries were transformed into E. cloni® 10G ELITE electrocompetent cells in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ohms, 1.8 kV). Cells were grown overnight at 30degC in 50 mL LB with 40 µg/mL kanamycin with shaking to prevent overgrowth. Library DNA was isolated by midiprep (ZymoResearch). This library yielded >100,000 colonies corresponding to >300X coverage for perfect mutations assuming one-third of mutants are perfect.

MuA Transposon Mediated Domain Insertion

Transposition libraries were generated using 100ng MuA-BsaI engineered transposon and 1:2 molar ratio transposition target DNA in 20ul reactions with 4 µl 5x MuA reaction buffer and 1 µl 0.22 µg/µl MuA transposon (ThermoFisher). MuA-BsaI engineered transposon propagation plasmid or pUCKanR-Mu-BsaI was a gift from David Savage (Addgene plasmid # 79769). MuA-BsaI engineered transposon was digested with BglII and HindIII Fastdigest enzymes (ThermoFisher) and gel purified using gel purification kit (Zymo Research).

The transposition targets, human Kir2.1 (Accession: NP_000882), Drosophila melanogaster Shaker (Accession: NP_728123), human α7 nAChR (Accession: NP_000735.1), and human ASIC1a (Accession: NP_001086.2) including a porcine teschovirus ribosomal skipping sequence (P2A) (61), were codon-optimized for mouse, synthesized (Gen9) and subcloned with into pATT-Dest using NEB BamHI and HindIII. pATT-Dest was a gift from David Savage (Addgene plasmid # 79770). For Kir2.1, a FLAG tag was inserted after T115 using Q5 site-directed mutagenesis (New England Biolabs). MuA transposition reactions were incubated at 30degC for 18 hours for transposition, followed by 75degC for 10 minutes for heat inhibition. DNA from reactions was cleaned up (Zymo Research) and eluted in 10ul water. All 10ul were transformed into 30ul electrocompetent 10G ELITE E. coli (Lucigen) in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ohms, 1.8 kV). Cells were rescued and grown without antibiotics for 1 hour at 37degC. Aliquots were then serially diluted and plated on LB agar plates containing carbenicillin (100 µg/ml) and chloramphenicol (25 µg/ml) to assess library coverage. The remaining transformation mix was grown in 50 ml LB containing carbenicillin (100 µg/ml) and chloramphenicol (25 µg/ml). All transformed libraries yielded greater than 10^5 colonies, which for Kir2.1-P2A (1369bp) is >35x coverage. Plasmid DNA was purified by midi-prep kit (Zymo Research).

Transposition-inserted Kir2.1 variants were subcloned into an expression vector by amplifying channel variant genes adding on BsmBI sites, using 10 cycles of PCR using Primestar GXL (Takara Bio) and run on a 1% agarose gel. The larger band was cut out and gel purified (Zymo Research) to isolate channels with inserted transposons. A mammalian expression vector (pcDNA3.1) with EGFP was amplified to add on BsmBI sites complementary to those on Kir2.1-P2A. The Kir2.1-P2A (BsaI-transposon) variants were subcloned into this vector by BsmBI-mediated Golden Gate cloning (56). Reactions were cleaned (Zymo Research) and eluted with 10ul water. All 10ul were transformed into 30 µl Lucigen electrocompetent 10G ELITE E. coli and electroporated in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ohms, 1.8 kV). Cells were rescued and grown without antibiotics for 1 hour at 37degC then with an aliquot serially diluted plated on LB agar plates containing kanamycin (50 µg/ml) and chloramphenicol (25 µg/ml) to assess library coverage. The remaining transformation mix was grown in LB containing kanamycin (50 µg/ml) and chloramphenicol (25 µg/ml). All transformed libraries yielded greater than 10^5 colonies, which correspond to >35x coverage. Plasmid DNA was purified by midi-prep kit (Zymo Research).

Inserted Transposons were replaced with domains in individual reactions using BsaI-mediated Golden Gate cloning. Cib81 was PCR amplified to add on BsaI and linkers (Ala-

Ser and Ser-Ala-Gly), preceding and following the domain insertion) sites complementary to MuA-BsaI transposon sites for Golden Gate cloning. Domain amplicons were gel purified (Zymo Research). The product was further digested with AgeI-HF (NEB) and Plasmid-Safe ATP-dependent DNase (Epicentre) to remove any undigested transposon, then cleaned up (Zymo Research) and eluted with 10 µl water. All 10ul were transformed into 30 µl Lucigen electrocompetent 10G ELITE E. coli and electroporated in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ohms, 1.8 kV). Cells were rescued and grown without antibiotics for 1 hour at 37degC. An aliquot was serially diluted and plated LB agar plates containing kanamycin (50 µg/ml) to assess library coverage. The remaining transformation mix was grown in LB containing kanamycin (50 µg/ml). All transformed libraries yielded greater than 10^5 colonies meaning there is >35x coverage. Plasmid DNA was purified by midi-prep kit (Zymo Research).

Permissibility Assay

100 ng of MuA-generated and five concentrations of SPINE-generated Kir2.1 insertion library (50ng, 100ng, 200ng, 400ng, 600ng, 1.2 µg) were transfected with 36 µl of Turbofect (ThermoFisher) into 50% confluent HEK293FT (Invitrogen) with additional inert plasmid (pATT Dest) added to a total of 12 µg transfected DNA divided across a single 6 well dish (9.6 cm^2 / well). Multiple concentrations were used to artificially boost the noise level in the SPINE libraries to further challenge the assay. The 50ng (0.5%) data was not included in the downstream analysis as too few cells expressed Kir2.1 to yield complete permissibility data.

Cells from each well were detached using 1 ml Accutase (Stemcell Technologies) and twice spun down at 450xg and resuspended in FACS buffer (2% of FBS, 0.1% NaN3, 1xPBS). Cells were incubated with 1:200 anti-flag mouse antibody (Sigma) 1 hour rocking at 4 degC, washed twice with FACS buffer, covered with aluminum foil, and then incubated with 1:400 anti-mouse Alexa Fluorophore 568 (Thermo Fisher) for 30 minutes rocking at 4degC. We will refer to Alexa Fluorophore 568 as ‘Label’ from hereon. Cells were washed twice, resuspended in 3 ml FACS buffers, and filtered using cell strainer 5 ml tubes (Falcon). Cells were kept on ice and protected from light in the transfer to the flow cytometry core. Before cell sorting, a small aliquot of cells was saved as a control sample for sequencing.

Cells were sorted into EGFP high / Label low (transfected cells without surface expression) and EGFP high / Label high (transfected cells with surface expression) on a BD FACSAria II P69500132 flow cytometer. EGFP fluorescence was excited using a 488 nm laser, recorded with a 525/50 bandpass filter and a 505 long-pass filter. Alexa fluorophore 568 fluorescence was excited using a 561 nm laser and recorded with a 610/20 bandpass filter. Cells were gated on side scattering and forward scattering area to select whole HEK293FT cells, gated on forward scattering height and width to separate single cells, then gated on co-expressed EGFP to gate out cells that received a plasmid, then gated on cells that were labeled using the anti-flag antibody for surface-expressed channels. Gates were determined using single wildtype, EGFP only and unstained library samples. A representative example of this gating scheme is shown in **Figure 3-5**. EGFP high / Label low and EGFP high / Label high cells were collected into catch buffer (20% of FBS, 0.1% NaN3, 1xPBS). As many cells as possible (Between 2,000-100,000 cells) were collected for each sample/library pair which is ~4-250x coverage of all potentially productive (i.e., in-frame and forward) domain insertions.

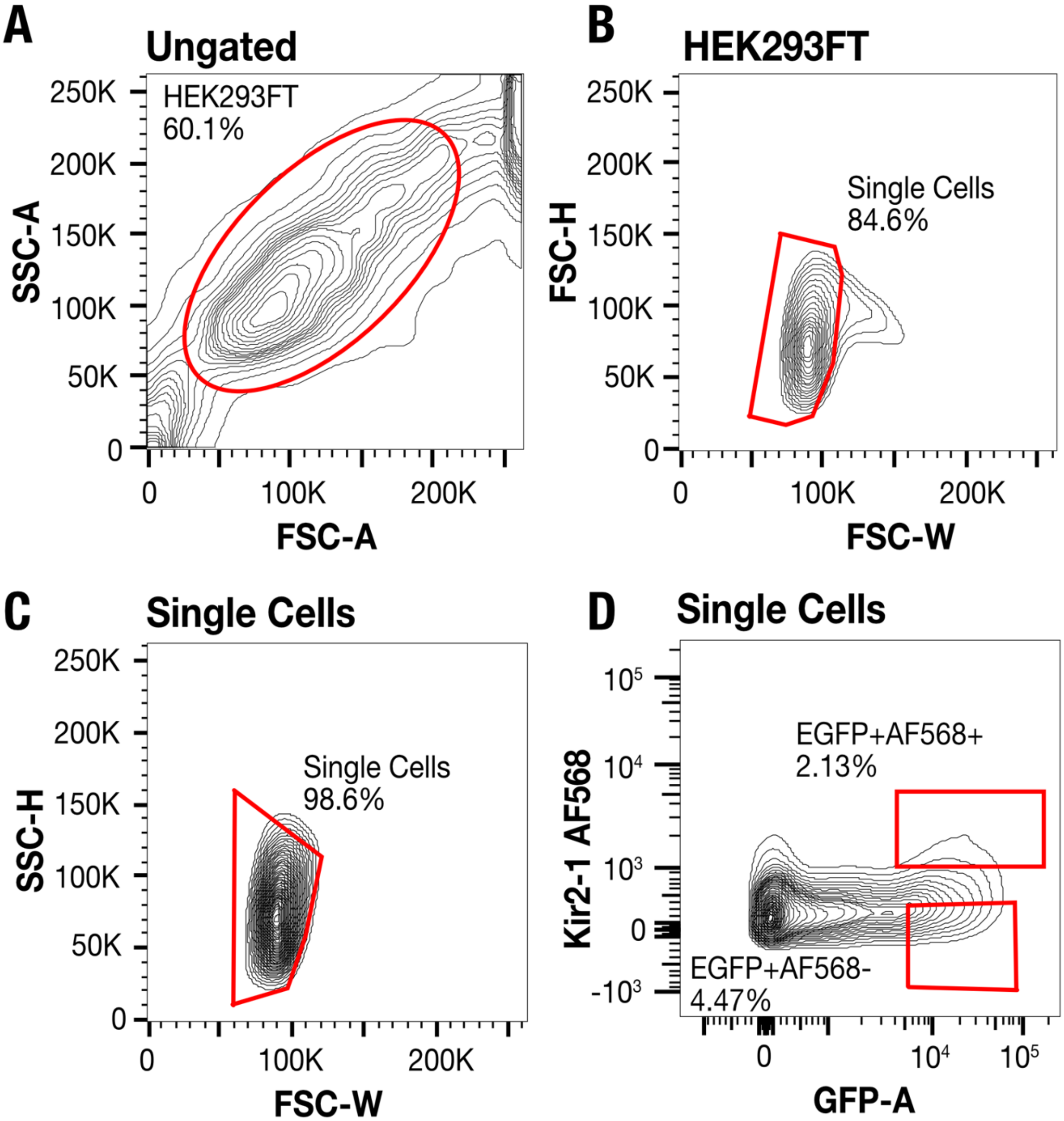
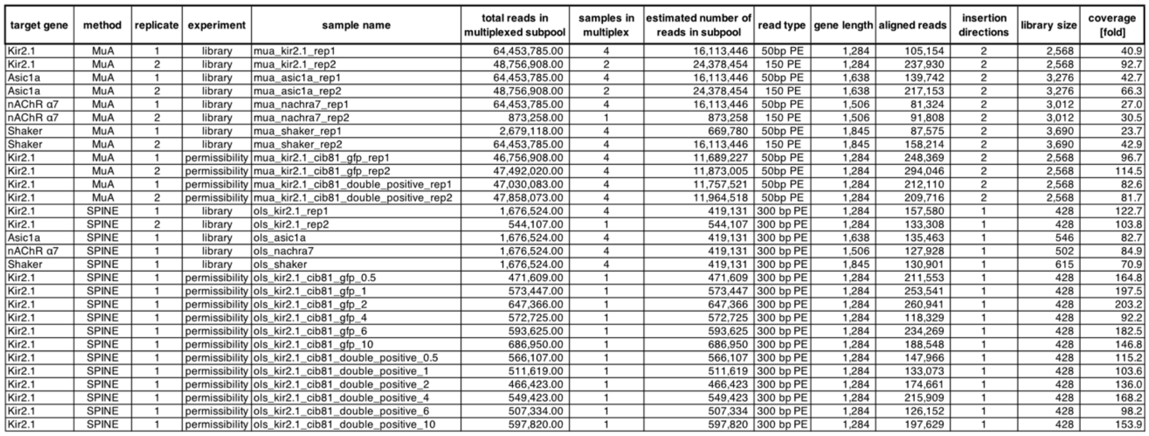


Figure 3-5: Permissibility assay gating scheme. (A) Whole HEK293 cells are gated on side (SSC-A) and forward scattering (FSC-A). **(B-C)** Side scattering height (SSC-H) and forward scattering width (FSC-W) are gated to select single cells. (**D**) EGFP high / Label low and EGFP high / Label high populations are gated based on EGFP (GFP-A) and Alexa fluorophore 568 fluorescence (secondary antibody, KIR2-1 AF568).

NGS Sequencing

DNA from pre-sort Control, EGFP high / Label low, and EGFP high / Label high cells for each library were extracted using a Microprep DNA kit (Zymo Research) and triple eluted with water. To remove chromosomal DNA, samples were digested with Plasmid-Safe

ATP-dependent DNase (Epicentre). The resulting plasmid DNA was further purified and concentrated using (Zymo Research). The product was used as a template for 12 cycles of PCR using Primestar GXL (Takara Clontech), run on a 1% agarose gel, and gel purified (Zymo Research) to remove primer dimers and non-amplicon DNA. Purified DNA was quantified using Picogreen DNA concentration and equal amounts of each domain insertion sample were pooled by cell sorting category (control, EGFP high / Label low, EGFP high / Label high). Pooled amplicons were prepared for sequencing using Nextera XT sample preparation workflows. Libraries were sequenced using Illumina MiSEQ in 300bp paired-end configuration. Read count statistics are provided in **Table 3-1**.



target gene	method	replicate	experiment	sample name	total reads in multiplexed subpool	samples in multiplex	estimated number of reads in subpool	read type	gene length	aligned reads	insertion directions	library size	coverage [fold]
Kir2.1	MuA	1	library	mua_kir2.1_rep1	64,453,785.00	4	16,113,446	50bp PE	1,284	105,154	2	2,568	40.9
Kir2.1	MuA	2	library	mua_kir2.1_rep2	48,756,908.00	2	24,378,454	150 PE	1,284	237,930	2	2,568	92.7
Asic1a	MuA	1	library	mua_asic1a_rep1	64,453,785.00	4	16,113,446	50bp PE	1,638	139,742	2	3,276	42.7
Asic1a	MuA	2	library	mua_asic1a_rep2	48,756,908.00	2	24,378,454	150 PE	1,638	217,153	2	3,276	66.3
nAChR α7	MuA	1	library	mua_nachra7_rep1	64,453,785.00	4	16,113,446	50bp PE	1,506	81,324	2	3,012	27.0
nAChR α7	MuA	2	library	mua_nachra7_rep2	873,258.00	1	873,258	150 PE	1,506	91,808	2	3,012	30.5
Shaker	MuA	1	library	mua_shaker_rep1	2,679,118.00	4	669,780	50bp PE	1,845	87,575	2	3,690	23.7
Shaker	MuA	2	library	mua_shaker_rep2	64,453,785.00	4	16,113,446	150 PE	1,845	158,214	2	3,690	42.9
Kir2.1	MuA	1	permissibility	mua_kir2.1_cib81_gfp_rep1	46,756,908.00	4	11,689,227	50bp PE	1,284	248,369	2	2,568	96.7
Kir2.1	MuA	2	permissibility	mua_kir2.1_cib81_gfp_rep2	47,492,020.00	4	11,873,005	50bp PE	1,284	294,046	2	2,568	114.5
Kir2.1	MuA	1	permissibility	mua_kir2.1_cib81_double_positive_rep1	47,030,083.00	4	11,757,521	50bp PE	1,284	212,110	2	2,568	82.6
Kir2.1	MuA	2	permissibility	mua_kir2.1_cib81_double_positive_rep2	47,858,073.00	4	11,964,518	50bp PE	1,284	209,716	2	2,568	81.7
Kir2.1	SPINE	1	library	ols_kir2.1_rep1	1,676,524.00	4	419,131	300 bp PE	1,284	157,580	1	428	122.7
Kir2.1	SPINE	2	library	ols_kir2.1_rep2	544,107.00	1	544,107	300 bp PE	1,284	133,308	1	428	103.8
Asic1a	SPINE	1	library	ols_asic1a	1,676,524.00	4	419,131	300 bp PE	1,638	135,463	1	546	82.7
nAChR α7	SPINE	1	library	ols_nachra7	1,676,524.00	4	419,131	300 bp PE	1,506	127,928	1	502	84.9
Shaker	SPINE	1	library	ols_shaker	1,676,524.00	4	419,131	300 bp PE	1,845	130,901	1	615	70.9
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_gfp_0.5	471,609.00	1	471,609	300 bp PE	1,284	211,553	1	428	164.8
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_gfp_1	573,447.00	1	573,447	300 bp PE	1,284	253,541	1	428	197.5
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_gfp_2	647,366.00	1	647,366	300 bp PE	1,284	260,941	1	428	203.2
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_gfp_4	572,725.00	1	572,725	300 bp PE	1,284	118,329	1	428	92.2
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_gfp_6	593,625.00	1	593,625	300 bp PE	1,284	234,269	1	428	182.5
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_gfp_10	686,950.00	1	686,950	300 bp PE	1,284	188,548	1	428	146.8
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_double_positive_0.5	566,107.00	1	566,107	300 bp PE	1,284	147,966	1	428	115.2
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_double_positive_1	511,619.00	1	511,619	300 bp PE	1,284	133,073	1	428	103.6
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_double_positive_2	466,423.00	1	466,423	300 bp PE	1,284	174,661	1	428	136.0
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_double_positive_4	549,423.00	1	549,423	300 bp PE	1,284	215,909	1	428	168.2
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_double_positive_6	507,334.00	1	507,334	300 bp PE	1,284	126,152	1	428	98.2
Kir2.1	SPINE	1	permissibility	ols_kir2.1_cib81_double_positive_10	597,820.00	1	597,820	300 bp PE	1,284	197,629	1	428	153.9

Table 3-1: Read sequencing statistics

Domain insertion permissibility enrichment

Alignments were done individually on both forward and reverse reads using a DIP-Seq pipeline (76,80), slightly modified for compatibility with updated python packages. In rare instances, both forward and reverse reads report domain insertion events, which results in duplicated domain insertion calls. In this event the duplicated domain insertion call is removed to avoid artificially boosting some events. This pipeline results in plaintext files indicating domain insertion positions and whether that insertion is in-frame and in the forward direction. Enrichment was calculated by comparing the change in EGFP high / Label low to EGFP high / Label high cells. Only positions with reads in both cell groups were used in enrichment calculations. All other positions are treated as 'NA' and not considered in downstream analysis and structure mappings, except for calculating correlations between datasets and correlations between insertion sites. In these correlation calculations, we treated 'NA's as '0's, because removing all the data will introduce more noise when comparing between datasets due to sampling limits.

Permissibility function for individual datasets comparing surface expressed (SE) and non-surface expressed (NSE) insertion variants:

$$F(i,j) = \frac{r^{i}_{j_{SE}}}{t_{j_{SE}}} - \frac{r^{i}_{j_{NSE}}}{t_{j_{NSE}}} \quad (1)$$

where *r* is the number of reads at amino acid position *i*, in the *j*th dataset divided by *t*, the total number of reads in the *j*th given sample.

Library Comparison

To compare read counts across multiple proteins, we normalized each gene by dividing each insertion site read count by the total number of reads for the respective gene. To account for variable gene length, we then multiplied the normalized read count by the number of amino acids for the respective gene to obtain normalized insertions per residue. Ideally, every insertion position would have a value of one, indicating an evenly distributed insertion library. To test how evenly distributed our libraries are, we compared the distribution using empirical cumulative probability density plot, which indicates both mean read count at 0.5 cumulative probability and the distribution of read counts by the slope. We also compared the library coverage (fraction of insertion positions) of each gene at increasing read depth thresholds (genes were normalized to 300 reads per position).

Domain insertion permissibility per position was z-scored:

$$z_i = \frac{x_i - \mu}{\sigma} \qquad (2)$$

where *x* is permissibility at amino acid position *i*, μ is the sample mean permissibility, and σ is the sample standard deviation.

Z-scored permissibility was mapped onto the structure of chicken Kir2.2 (PDB 3SPI) (87) using Chimera (155).

Sequence logos were generated using the *ggseqlogo* R package using the "twosamplelogo" sequence logo method which enables removal of any sequence background from the sequence logo resulting in a more accurate sequence logo (156).

Determining depletion of single base pair deletions and enrichment of in-frame domain insertions in the correct direction

To quantify single base pairs deletions (the predominant type of synthesis error with phosphoramidite chemistry (157–159)) in the SPINE permissibility sequencing data, we aligned paired-end reads from each dataset (Control, EGFP high / Label low, and EGFP high / Label high) to the sequence of Kir2.1 using the BBMap alignment package. We calculated the frequency of deletions in each dataset by dividing the number of 1 bp deletions detected in the aligned reads by the total number of aligned reads. To calculate enrichment of 1 bp deletion in EGFP high / Label low, and EGFP high / Label high datasets, we divided deletion frequency in these datasets by the deletion frequency of the corresponding control dataset.

To quantify incorrect reading frame insertion and directionality, we used the output from the DIPseq alignment pipeline for each dataset (Control, EGFP high / Label low, and EGFP high / Label high). The DIPseq alignment pipeline assigns a DNA insertion position and the direction of every insertion into your recipient gene. Using this data, we calculated frequencies for every reading frame (0, +1, +2) and insertion direction (plus, minus) as the number of reads in each of these six classes divided by the total number of reads. Enrichment for each class was then calculated by dividing each of these classes for EGFP high / Label low, and EGFP high / Label high dataset by the corresponding control dataset.

RESULTS

The SPINE workflow

SPINE is enabled by microarray-based massive oligonucleotide library synthesis (OLS) (150,151). OLS libraries are used for large-scale parallel gene synthesis (160,161) and generating saturated mutation libraries through oligo annealing (162) or recombination (11,163). Similarly, we combined OLS library synthesis with multi-step Golden Gate Cloning (152), to generate domain insertion libraries in a programmable fashion **(Figure 3-2A).**

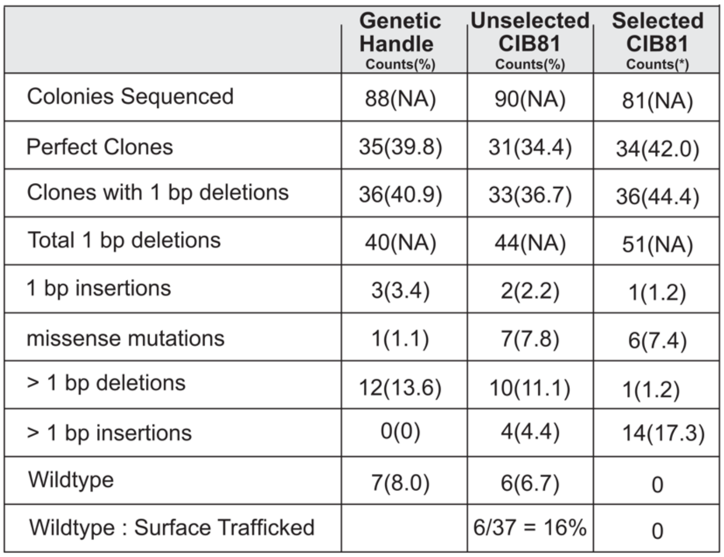
Current OLS can produce oligos with a maximum length of 230 base pairs (bp) (151). We broke up each target gene into fragments, whose insertional diversity is encoded by OLS library subpools. Each subpool contains about 170 bp of gene sequence flanked by biorthogonal barcodes for PCR amplification, and Golden Gate-compatible BsmBI sites for cloning the fragment into the target gene (**Figure 3-2B**). Varied between the oligos in each subpool is the genetic handle, which is inserted at every amino acid position of the target gene fragment corresponding to this subpool. Genetic handles are designed with Golden Gate-compatible BsaI sites at the beginnings and ends of linkers that allow replacement with any DNA sequence (in our case, a domain). The overhangs generated by these BsaI sites also encode the amino acids that serve as linkers between the target protein and the inserted domain. We here chose a short serine/glycine linker, which is widely used as a flexible linker (164), but any linker at least 2 amino acids long can be encoded in the BsaI overhangs.

For each fragment subpool, we generated corresponding target gene backbones plus complementary BsmBI cut sites by PCR amplifying, from a shuttle plasmid, all of the wildtype gene except for the region of the gene encoded by the fragment subpool. The OLS subpools were assembled with their corresponding backbone fragments in BsmBI-mediated Golden Gate reactions. This process was repeated for all fragment subpools and these libraries are combined in equimolar ratio to yield pooled intermediate libraries. The final domain insertion libraries were generated by replacing the genetic handles with a PCR-amplified domain of interest flanked by complementary BsaI cut sites and flexible linkers using BsaI-mediated Golden Gate cloning.

Domain Insertion Library Generation

Guided by our interest in probing the relationship between domain recombination and ion channel function, we generated domain insertion libraries with four ion channel genes, inward rectifier K^+ channel Kir2.1, voltage-dependent K^+ channel Shaker, α7 nicotinic acetylcholine receptor (α7nAChR), and the acid-sensing ion channel ASIC1a. In this proof-of-principle, we replaced the genetic handle with the 9 kDa plant protein domain Cib81 (92). Cib81 was chosen as a benchmark because we had used Cib81 in transposon-generated libraries (13).

To determine insertion library error rates and contamination with wildtype DNA (a leftover from the inverse PCR to generate the target gene backbone, which becomes enriched in the multistep cloning), we sequenced individual clones by Sanger sequencing from intermediate libraries (contain the genetic handle) and final domain-inserted libraries (contain Cib81). We found that ~40% of clones had the expected sequences without any errors (**Table 3-2**). Conversely, ~60% of clones had errors, with 1 bp deletions being the most frequent (41%) and 7% of clones were wildtype. In downstream functional assays, a wildtype channel would lead to significant false positives. We, therefore, replaced the genetic handle for Kir2.1 with a chloramphenicol antibiotic cassette to enrich for oligo incorporated plasmids before replacing with Cib81. This removed any contaminating wildtype DNA (Selected Cib81, **Table 3-2**). Overall, SPINE yielded similar percentages of perfect insertion libraries and wildtype to comparable targeted mutational approaches that use oligo library synthesis (162,163).



	Genetic Handle Counts(%)	Unselected CIB81 Counts(%)	Selected CIB81 Counts(*)
Colonies Sequenced	88(NA)	90(NA)	81(NA)
Perfect Clones	35(39.8)	31(34.4)	34(42.0)
Clones with 1 bp deletions	36(40.9)	33(36.7)	36(44.4)
Total 1 bp deletions	40(NA)	44(NA)	51(NA)
1 bp insertions	3(3.4)	2(2.2)	1(1.2)
missense mutations	1(1.1)	7(7.8)	6(7.4)
> 1 bp deletions	12(13.6)	10(11.1)	1(1.2)
> 1 bp insertions	0(0)	4(4.4)	14(17.3)
Wildtype	7(8.0)	6(6.7)	0
Wildtype : Surface Trafficked		6/37 = 16%	0

Table 3-2: Library selection error rates.

SPINE libraries have increased and more consistent saturation

The current state-of-the-art for generating domain insertion libraries relies on MuA transposase (76). However, MuA transposon-generated libraries have incomplete coverage (13)(80) and strong sequence bias (98,135,141–145).

To test whether SPINE libraries can overcome bias and low coverage problems that exist in MuA-based methods, we benchmarked them against transposon-generated libraries. The difference in coverage is easily apparent from visual inspection (**3-6A** log-transformed**, Figure 3-7** raw counts). We found that SPINE libraries had an average of 99.97% coverage compared to 49% for MuA transposase. In the most extreme case, α7nAChR, coverage went from less than 40% of positions having at least 5 reads with MuA transposase to a greater than 95% of positions having at least 55 reads per position using SPINE (**Figure 3-6B**). Furthermore, the probability of coverage stays flat for a considerable read depth range (1 – 80 reads), which suggests that coverage is less variable and more redundant.

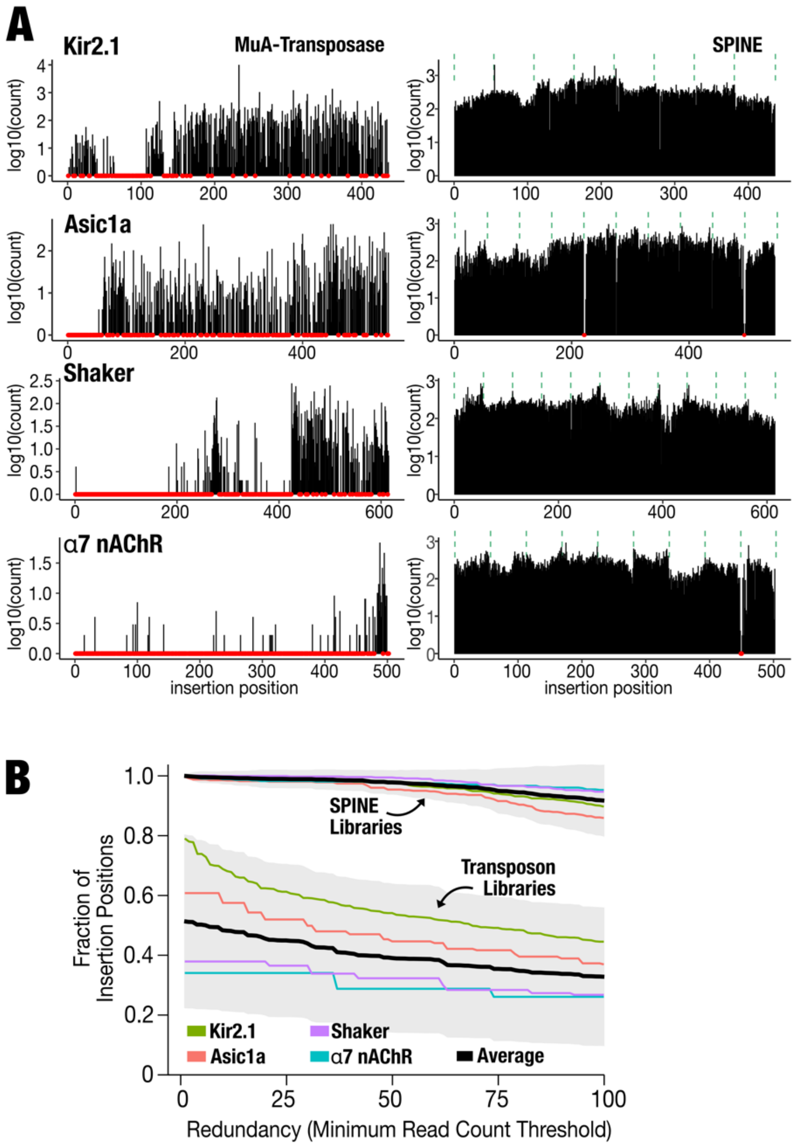


Figure 3-6: Library coverage. A) SPINE libraries are saturated. Comparison of MuA-tranposase generated insertion libraries and SPINE for four different ion channels. Red dots indicate missing positions. Green dashed lines indicate fragment boundaries for OLS-assembled libraries. **B) SPINE libraries have deep coverage.** Shown is the fraction of insertion positions for a given target gene that have the indicated coverage for each method. The average for each method is shown as a black line and the 95% confidence interval is shaded grey.

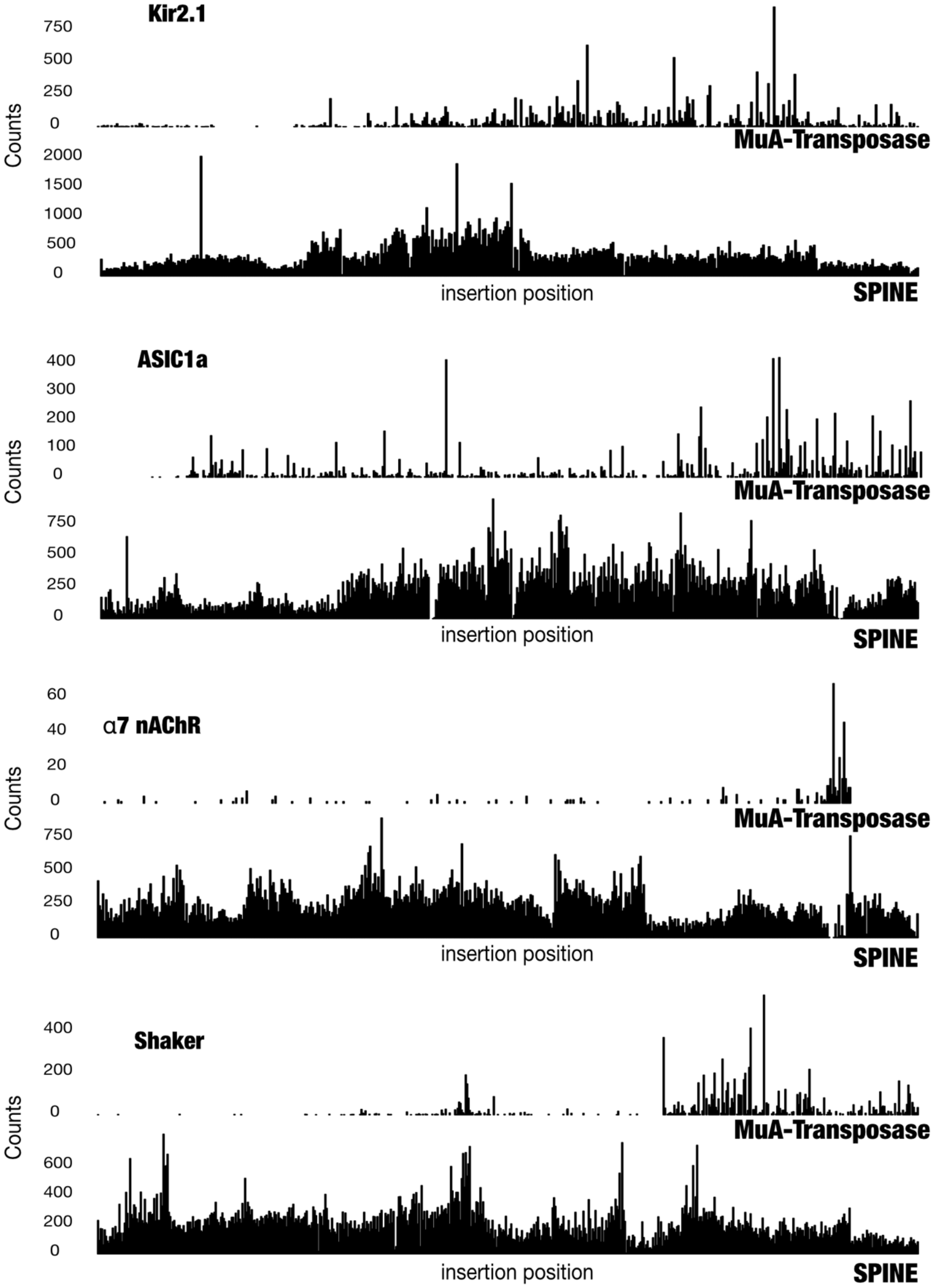


Figure 3-7: Raw Reads. Comparison of MuA-transposase and SPINE-generated insertion libraries for four different ion channels. Histograms show counts of Cib81 insertions for each amino acid position.

SPINE libraries have drastically reduced sequence bias

We compared replicates of generated libraries to test whether the uneven coverage we observed in MuA transposons was due to sampling or sequence bias. We found similar insertional maps from replicates in Kir2.1, ASIC1a, and Shaker transposon libraries, which reiterates previous reports on MuA transposase bias (98,143,144) (**Figure 3-8A**). That bias became apparent when we generated a sequence logo for MuA-mediated insertion positions in Kir2.1 (**Figure 3-8B**) and the other channels (**Figure 3-9B-C**). In agreement with known MuA bias, we found enrichment for insertions at trinucleotide CGG position (98,143–145). In contrast, SPINE library replication had lower insertional map similarity (**Figure 3-8A)**, and no strong and repeated sequence logo was apparent (Kir2.1 **Figure 3-8B**; Shaker, ASIC1a, α7nAChR **Figure 3-9B-C**), which shows that SPINE has drastically reduced bias. To compare the variability of random insertions with respect to targeted genes, we compared the empirical cumulative probability distribution functions (ECDF) for a simulated random distribution for each target gene with to those from MuA-generated libraries and SPINE (**Figure 3-8C**). While ECDFs for SPINE libraries are very similar to each other and similar to a random distribution (two-sample Kolmogorov-Smirnov test, D = 0.29084, p-value < 2.2e-16), for MuA libraries they are highly variable among each other and different from a random distribution (Two-sample Kolmogorov-Smirnov test, D = 0.74488, p-value < 2.2e-16).

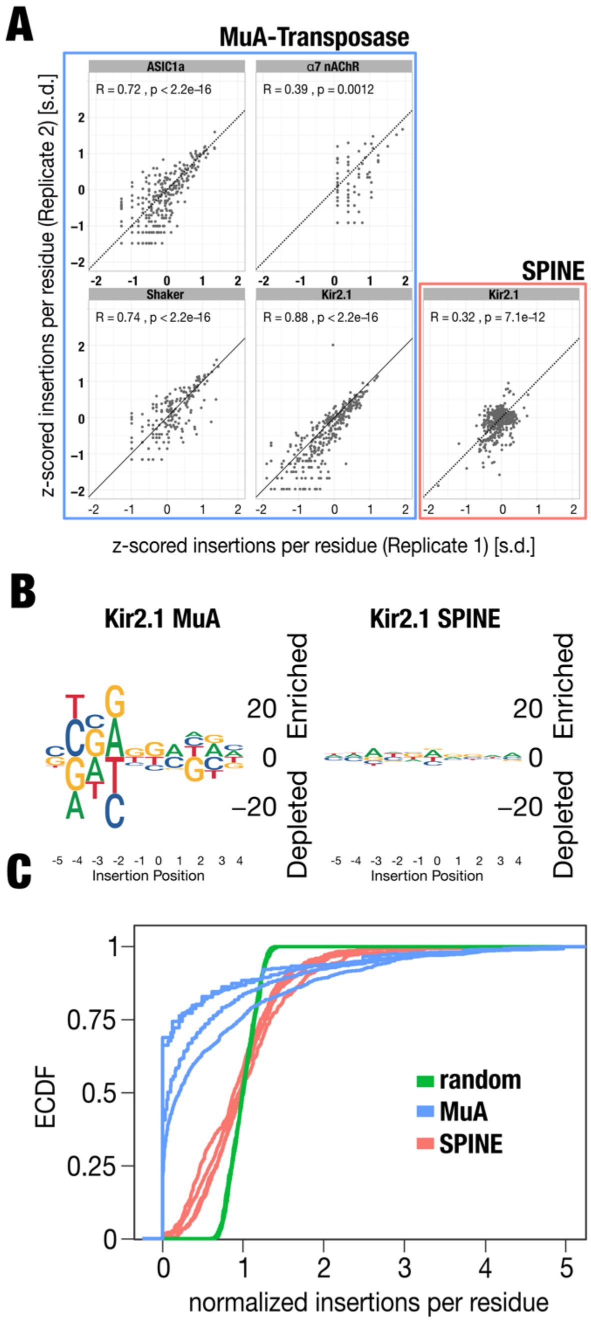


Figure 3-8: SPINE has drastically reduced bias. A) Scatterplots show z-scored insertions per residue for each biological replicate. Spearman correlation coefficients are inset. **B)** Sequence logos for insertion sites in Kir2.1 using MuA-transposition (left) and SPINE (right). **SPINE libraries are less sensitive to a targeted gene sequence. C)** Empirical cumulative density functions of four different target genes generate by MuA-transposition (blue lines) and SPINE (red lines). An idealized random library is shown (green lines). While ECDFs for SPINE libraries are very similar to each other and similar

to a random distribution (Two-sample Kolmogorov-Smirnov test, D = 0.29084, p-value < 2.2e-16), for MuA libraries they are highly variable among each other and different from a random distribution (Two-sample Kolmogorov-Smirnov test, D = 0.74488, p-value < 2.2e-16).

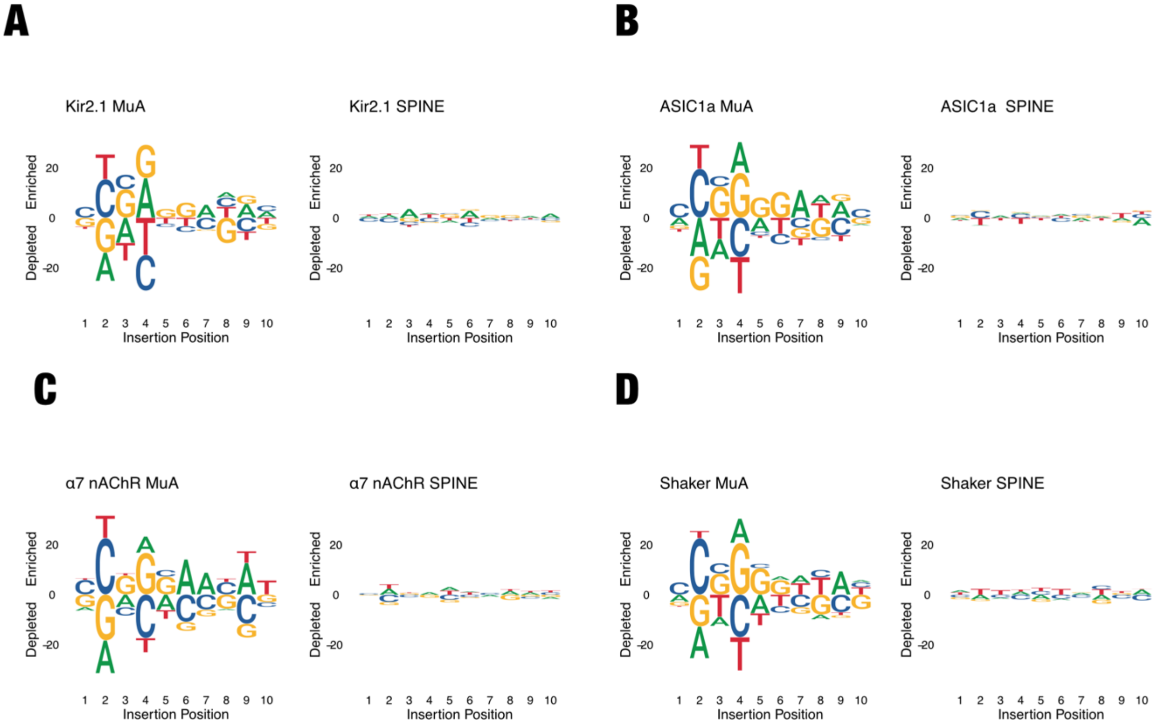


Figure 3-9: Sequence Logos for Kir2.1, ASIC1a, α7 nAChR, and Shaker libraries. Nucleotide bias of insertion position was calculated for MuA- and SPINE libraries using the *ggseqlogo* R package.

SPINE libraries only contain productive domain insertions

MuA transposition yields insertions in all 6 reading frames, which we confirmed for MuA libraries. Only 16% (1/6) of insertions are in the correct reading frame and direction (**Figure 3-10B**). In contrast, 99% of insertions were in-frame and forward in SPINE libraries due to this technique's programmed nature. Even if we account for SPINE's 58% error rate (**Table 3-2**) and make the best-case scenario assumption that MuA libraries have 0% errors apart from random insertion frame selection, SPINE results in more productive insertions than MuA transposons (44% in-frame and forward for SPINE vs. 16%).

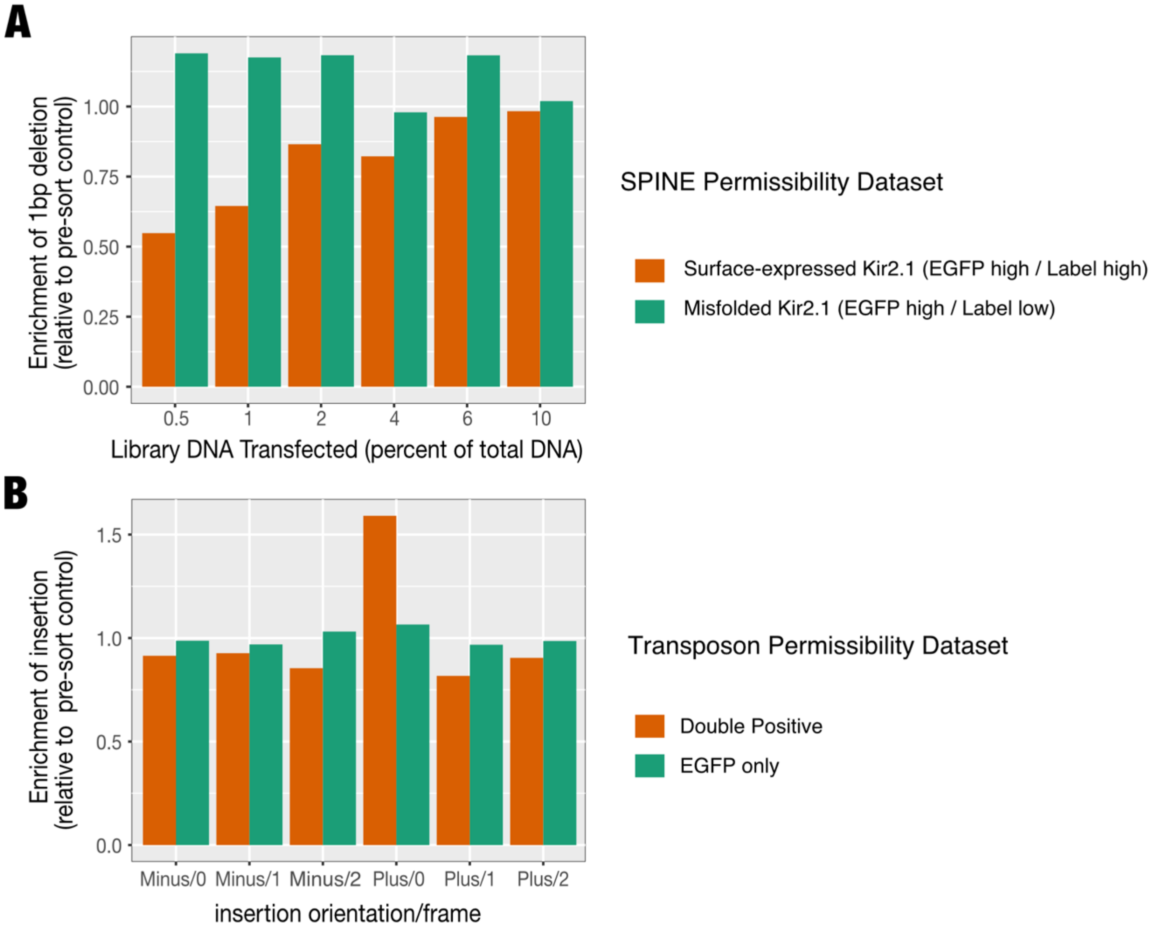


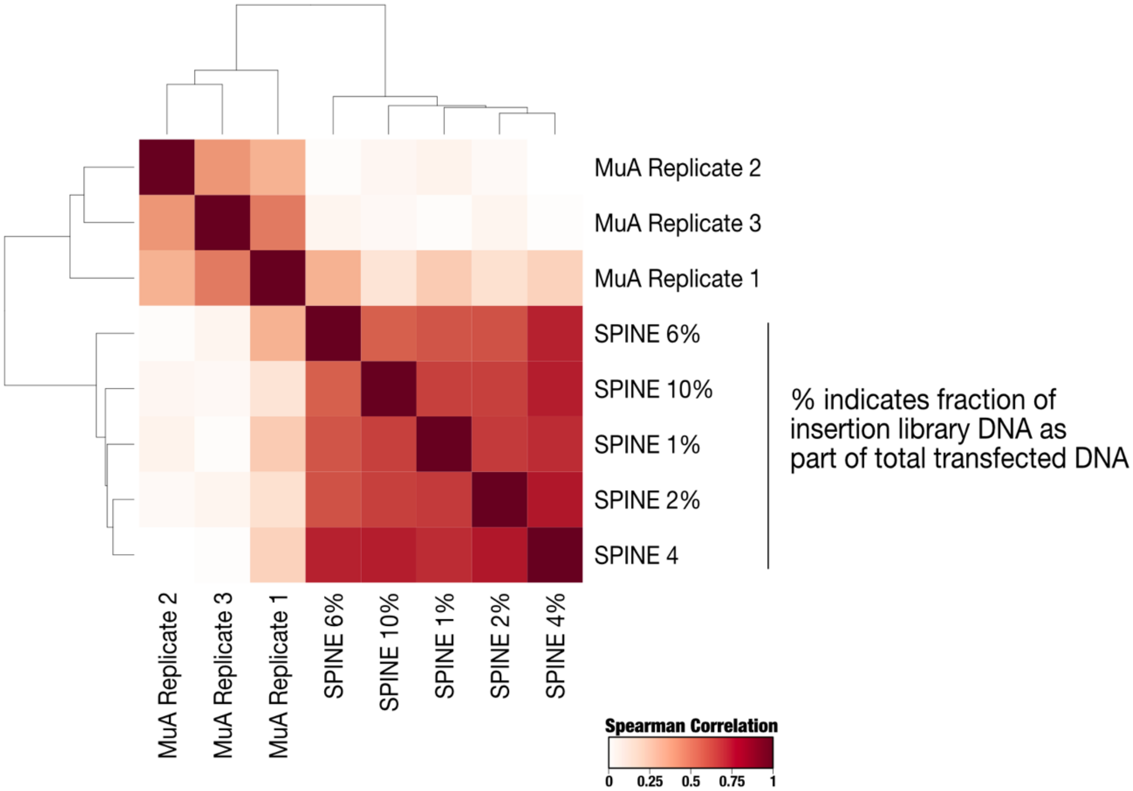
Figure 3-10: Signal-to-noise stemming from transient transfection in permissibility assays. **(A)** 1 basepair (bp) deletion frequency was quantified for SPINE-derived permissibility datasets from sorted cells (EGFP high / Label high, orange & EGFP high / Label low, green) and normalized to unsorted cells. Each dataset represents a different percentage of insertion library DNA as part of the total DNA amount used in the transient transfection of HEK293 cells. As expected, 1 bp deletions, which cause frameshift mutations, were depleted in cells expressing putative surface-trafficked insertion variants. **(B)** Insertion rates were quantified for all MuA transposase-generated permissibility datasets from from sorted cells (EGFP high / Label high, orange & EGFP high / Label low, green) and normalized to unsorted cells. For insertion variants recovered from cell expressing surface-trafficked Kir2.1, insertions are enriched only in the correct (Plus/0) reading frame.

Taken together, SPINE enables the generation of saturated domain insertion libraries with drastically reduced insertion position bias, near-complete coverage, and redundant insertions at each position. SPINE libraries are furthermore enriched for productive in-frame insertions in a target gene.

SPINE enables saturated domain insertion profiling

We previously used transposon-mediated library generation to profile domain insertion permissibility in Kir2.1 (13). We transiently transfected insertion libraries into HEK293 cells and performed a functional assay to measure permissibility. Permissibility is the sensitivity of a channel to the insertion of a domain at a given position and is determined by measuring how well a channel variant folds, assembles, and traffics to the cell surface. All insertion variants express EGFP as a transfection marker, but only surface-expressed variants are fluorescently labeled via an extracellular FLAG tag. Using fluorescently activated cell sorting (**Figure 3-2A**, Functional Assay), we isolate cells that express insertion variants that fold, assemble, and traffic well (EGFP high / Label high), from insertion variants that do not (EGFP high / Label low). We connect genotype (insertion variant) to phenotype (permissibility) by recovering and sequencing plasmids in sorted populations.

A potential problem with transient transfection is that each cell expresses a mix of insertion variants. When we sort a cell that contains a well-expressing insertion variant, sequencing will recover the coding sequence for a folding variant (the signal) and sequences that are unrelated to the phenotype (noise). Second, K^+ channels form tetramers which might be composed of monomers with different insertion variants; also increasing the noise. While the signal-to-noise was sufficient to conduct our work in Kir2.1 with transposon-generated insertion libraries, we wanted to established that a surface expression assay coupled to transient transfection with diluted DNA still yields sufficient signal-to-noise in the background of SPINE libraries. As determined earlier, 60% of clones in a SPINE library have errors that stem from inefficiencies in oligo synthesis (160). The predominant error is 1 bp deletions (157–159). Deletions will lead to frameshift mutations and premature stop codons, which should disrupt ion channel folding, assembly, and surface trafficking. When we determined enrichment/depletion of 1 bp deletions relative to the pre-sort control, we found slight enrichment in cells expressing non-surface trafficked insertion variants (**Figure 3-10A**). Importantly, in cells with surface-trafficked insertion variants, they were depleted. Degree of enrichment and depletion appears dependent on how much library DNA was used in the transfection. Specifically, when library DNA made up only 0.5% of the total amount of transfected DNA, 1 bp deletions where enriched ~12% in cells expressing predominantly misfolded Kir2.1, while they were depleted by ~50% in cells expressing predominantly surface-expressed Kir2.1. With an increasing amount of library DNA, that difference grew smaller until no depletion or enrichment (in comparison to pre-sort control) was observed. We also found that despite the increased noise from increasing amounts of library DNA, permissibility assay with SPINE-generated libraries where more repeatable (0.56 SPINE vs. 0.38 MuA transposase mean Spearman correlation coefficients, **Figure 3-11-12**).



% indicates fraction of insertion library DNA as part of total transfected DNA

Figure 3-11: Hierarchical Clustering by Spearman correlations between permissibility datasets. Spearman correlation coefficients were calculated for all Kir2.1 permissibility datasets. The three MuA permissibility datasets are derived from transient transfection with the same concentration of library DNA (1% of total transfected DNA). SPINE-derived insertion library permissibility assays are measured at 5 library DNA concentrations (1%, 2%, 4%, 6%, 10% of total DNA). Despite this increased experimental variability, OLS permissibility datasets have higher correlation between replicates. This suggest high reproducibility. Hierarchical clustering of OLS datasets shows that the low concentration replicates (1, 2, 4) and high concentration replicates (6, 10) cluster together.

This may reflect the higher probability of different insertion variants assembling into mosaic tetramers at higher library transfection amount, which would decrease the signal to noise in high vs. low concentration experiments.

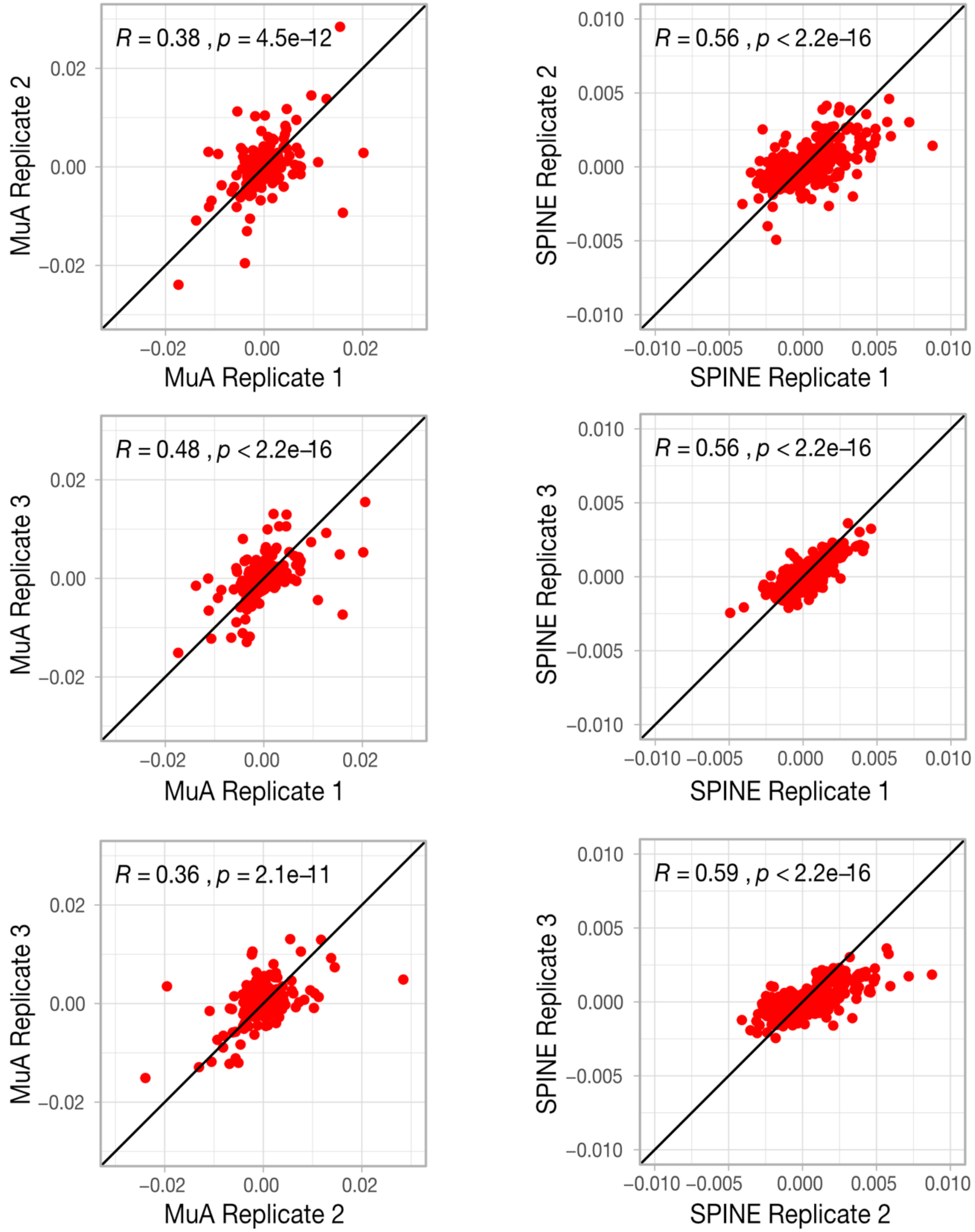


Figure 3-12: Comparison of permissibility assay replicates. Scatterplots show z-scored permissibility (red points) for the indicated Kir2.1 replicates derived from MuA-generated insertion libraries (left panels) or SPINE-generated libraries (right panels). Spearman correlation coefficients (R) for each replicate pair are inset.

In aggregate, these data agree with the expectation that deletions cause frameshift mutations or premature stop codons, which would cause ion channels to incorrectly fold, assemble, or traffic. Given that deletions in a cell with predominantly permissive insertion variants are depleted, suggests that even with transient transfection, our permissibility assay has sufficient signal-to-noise. The data also suggests this phenotype (surface-expression) is unlikely to be influenced by the higher mutation rate in SPINE libraries.

Having tested the sensitivity of our permissibility assay, we explored whether SPINE could improve permissibility map resolution in Kir2.1 compared to MuA transposition. With Cib81 as the inserted domain, we found SPINE improved permissibility maps. A visual inspection of permissibility data (averaged across three independent replicates) mapped onto the crystal structure of human Kir2.2 (87), visualizes the striking difference in saturation and dynamic range (**Figure 3-13A**). While MuA library data is sparse with 71 sites missing and noisy, SPINE library data is almost complete (1 site is missing) and has a high dynamic range between highest and lowest permissibility. Plotting permissibility along sequence position shows that formerly missing regions are now filled in (**Figure 3-13B**). For example, for a large region at the beginning of the gene (amino acid positions 1-150) little permissibility information is available from MuA libraries (which had poor insertion coverage and depth in this region) while permissibility is measured for the entire region with SPINE libraries. In this region, there are the unstructured N terminus, several regulatory sites, and M1 transmembrane domain (**Figure 3-13B,** protein topology cartoon) that are functionally important (87). The interface between the M1 and M2 helix is now well resolved, while most positions were missing in MuA libraries (**Figure 3-13C**). For all other regions, SPINE conforms to previous permissibility patterns while providing a more complete and dynamic data set. There now appear to be four levels of permissibility in Kir2.1: high permissibility for the unstructured C-terminus, moderate permissibility in the N-terminus, low permissibility in the structured cytosolic regions and no permissibility in transmembrane regions. In addition, within regions with expected high (flexible N/C termini) or no permissibility (transmembrane domain), there are fewer probable false positives or negatives (15/119 MuA vs. 6/157 SPINE; two-sided z score, p-value: <0.0064). Further emphasizing the improved quality of permissibility maps is that insertion into a known Golgi export signal (82) have clearer negative permissibility in the SPINE data.

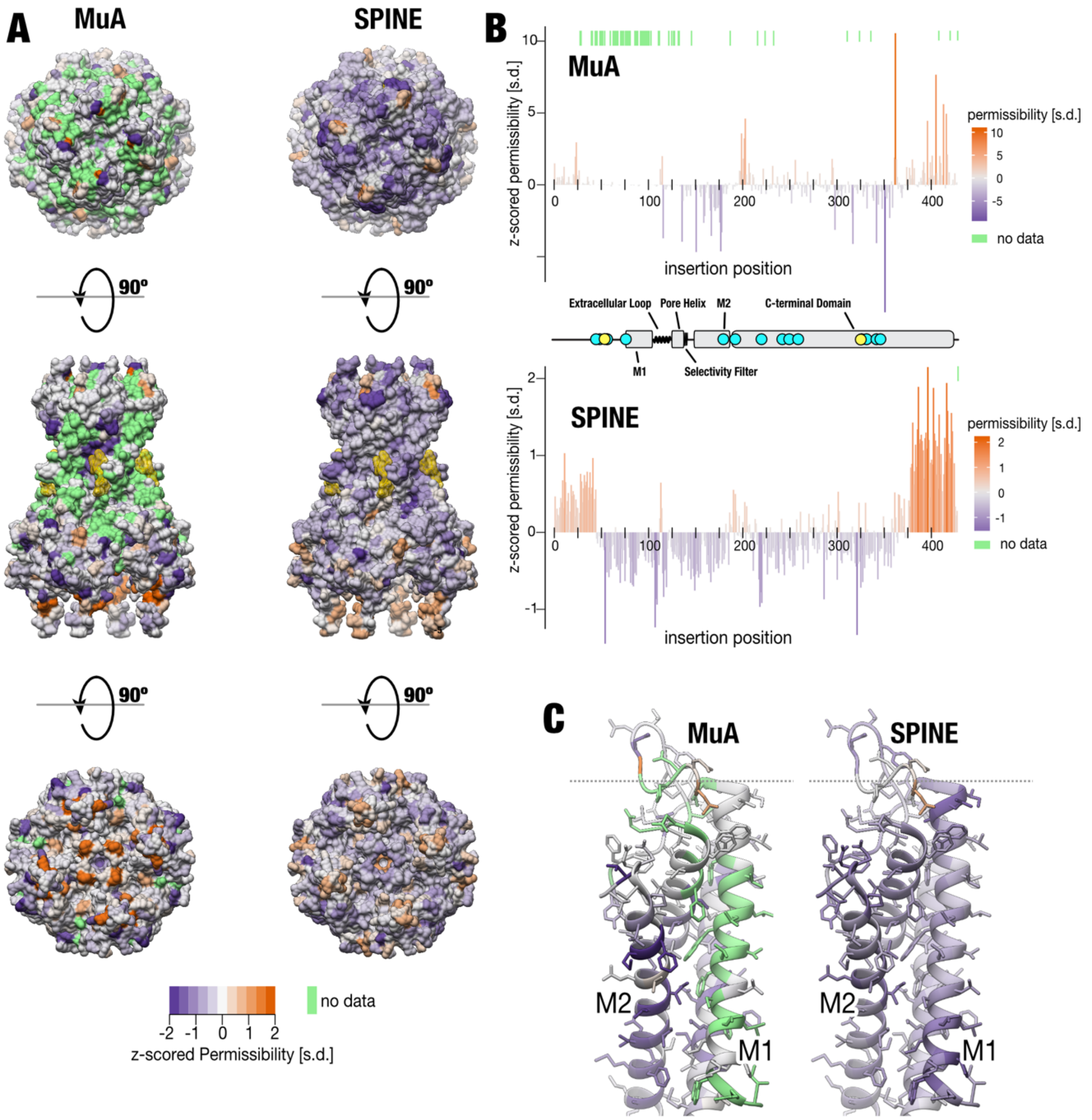


Figure 3-13: Domain Insertion Permissibility. A) Permissibility data for Cib81 insertion libraries derived from MuA-transposition and SPINE is mapped on the crystal structure of chicken Kir2.2 (PDB 3SPI (90)) Green indicates missing data. PIP_2, an allosteric modulator of Kir, is rendered yellow. **B)** Secondary structure elements (center) are shown along with z-scored permissibility for Cib81 insertion for MuA- and SPINE- generated libraries. Cyan

dots indicate functionally important sites. Yellow dots indicate trafficking signals important for surface expression. Green lines above each dataset indicate missing data. **C)** Comparison of permissibility coverage for transmembrane domains (M1 and M2).

DISCUSSION

Transposase-mediated domain insertion is widely used to address both basic science and biomedical engineering questions (72,131–133). We developed SPINE as an alternative approach that uses oligo library synthesis and multi-step Golden Gate cloning to assemble domain insertion libraries in a programmable fashion. Which approach investigators choose depends on what best meets the experimental requirements; SPINE compares favorably in several aspects.

The sequence bias and variable efficiency of transposases is well established (98,135,141–145). We and others showed, in different protein families, that this can result in domain insertion libraries that have bias, incomplete coverage, and variable coverage redundancy (13,76,140). For all tested genes SPINE has reduced bias, near-complete coverage, and superior coverage redundancy. The success of using transposon-based domain insertion to construct, for example, biosensors (76) may suggest that transposon-based approaches work well enough. And in light of the same general trends observed in domain insertion permissibility maps for Kir2.1 in this study, one could argue that bias, lack of coverage, and depth do not matter. However, for some target genes –such as nAChR in this study– transposon-generated domain insertion libraries have such severe bias and marginal saturation that they are effectively unusable for applications that derive insight from comprehensive mapping of all possible domain insertions (13). It is hard to intuit for which target genes transposon-mediated domain insertion will perform poorly and there may be limited recourse. For example, changing codon usage, did not improve nAChR libraries (unpublished observation). In other cases, functionally important regions have no domain insertion events, such as the S4-S5 linker in Shaker (important in mediating channel opening in response to change in voltage (165)) and a Na+ binding pocket in Kir2.1 (88,97). There is value in a domain insertion method that is predictable and dependable. Furthermore, lack of bias, near-complete, and redundant coverage result in richer functional data. In this study, this manifests as improved dynamic range of the Kir2.1 domain insertion permissibility signal. For other engineered proteins the case remains to be made, but we predict that SPINE will produce more complete domain insert maps, which will increase the likelihood of finding, for example, a functional biosensor.

In the DIP-seq approach, type IIS restriction sites are embedded in the MuA transposase recognition sites to mediate the exchange of the transposon inserted into the target gene with a domain of interest with compatible flanking overhangs (76). The simultaneous requirements of maintaining transposition efficiency (146,166) and restriction efficiency puts sequence constraints on restriction enzyme recognition sequences and overhangs. Because overhang sequences are added to the original target at the insertion site and encode linkers, the amino acid composition of these linkers is constrained. Linker optimization is a critical aspect of fusion protein engineering (74, 80). SPINE offers a significant advantage over MuA-transposon approaches because it puts no constraint on linker composition. We used a serine/glycine linker, which is used as a flexible linker (164), but any linker sequence of 2 amino acids at either side of the inserted domain can be used. This enables full exploration of how linker length and composition impact target protein function independent of the inserted domain. If the first and last two amino acids of the inserted domain are included as overhangs in the genetic handle, it is possible to insert a domain without any linkers. However, this would require a new OLS library for each inserted domain.

MuA can insert a transposon into any of the 6 reading frames, while the programmable nature of SPINE results in enriched in-frame insertions. When phenotyping assays are coupled to sequencing, as is the case for DIP-seq (76) or CPP-seq (139), SPINE allows for more efficient use of the specified sequencing output because fewer reads are spent on unproductive insertions. Furthermore, SPINE insertion libraries can be targeted to single or multiple regions of the target gene and, thus, avoid undesired insertions. Achieving the same with MuA transposases requires multiple intermediate staging libraries that contain the targeted regions, which then are subcloned with the remainder of the target gene. This feature of SPINE not only simplifies domain insertion workflows, but provide easier access to complex domain insertion library designs. In Kir2.1 for example, targeting domain insertions to known allosteric sites in Kir2.1 while avoiding transmembrane region or trafficking signals could be a promising strategy to efficiently construct light- and drug-switchable versions of this ion channel.

SPINE relies on microchip-synthesized oligonucleotides which have an overall error rate of ~0.2% (1 in 500 bp) (160). This means that only ~50% of the oligos in a 230bp OLS pool are expected to have the correct sequence. Since we do not (but could in the future) use enzymatic error-correction (160,167), the number of assembled domain insertion variants carrying mutations is high (~60% in this study). Owing to inefficiencies in the phosphoramidite chemistry used in oligo library synthesis, the predominant error is single-base deletions (157–159) (36% in this study). Single-base deletions result in frameshift mutations that introduce premature stop codons and therefore non-functional proteins. Our data supports this by showing that single-base deletions are strongly depleted in cells expressing surface-expressed protein. Missense mutations are rare (1-8% in this study) and considering the large number if possible combinations of missense mutations and domain insert sites (>2 million for Kir2.1) it is unlikely that the same missense mutation occurs frequently enough with the same domain insertion to influence the observed phenotype. Overall, the majority of mutations introduced by SPINE do not substantially impact downstream assay fidelity. Lastly, new chemistries and processes continuously improve oligo synthesis sequence fidelity, which can benefit SPINE in the future.

Transposon-based approaches and SPINE both operate at the nucleic acid level and can be applied to arbitrary protein coding and non-coding sequences. Some domestication is required with SPINE in the form of removing certain type IIS restriction sites (here, BsaI and BsmBI), however, in the age of relatively cheap DNA synthesis this a low barrier. While the same OLS pool can be reused to insert different domains into the same target (protein) sequence, each additional target requires a new pool. In light of these requirements, transposon-based domain insertion library construction holds a measurable cost and ease-of-use advantage, in particular, if the number of targeted proteins in large and the number of inserted domains is small. For applications that require drastically reduced bias, complete coverage, and more redundancy these advantages may be less relevant. Under such circumstances, SPINE offers distinct cost and time advantages as an approach that will likely work on the first try.

SPINE is to our knowledge the first method to enable saturated domain insertion profiling. This puts domain insertion profiling on the same level as deep mutagenesis as a method that enables experimental evolution. Like mutations, domain insertion is a major source of genetic variation that underlies natural evolution. By virtue of the programmable nature of

OLS, other types of genetic variation can conceivably be combined with domain insertion, including any combination of single amino acid mutations, insertions, or deletions. This opens up the possibility to study how the effects of domain insertion depend on sequence context, i.e. epistasis (168,169). Saturated domain insertion profiling, made possible by SPINE, can be a window into the relationship between domain insertion and the emergence of new protein function and how this relationship is shaped by other evolutionary forces.

From a practical perspective, SPINE could also prove instrumental in protein engineering. Rational approaches explicitly leverage structural and functional information (85), however in the absence of such information, they reach their limits. Computational approaches (e.g., coevolution analysis (79,170) work best in large protein families with wide-spread and homogenously distributed similarity. Rule-based *de novo* protein design (171) (88) is rapidly advancing, but does not capture protein dynamics that underlie allosteric transitions (28,172). Domain insertion profiling is a scalable method that can provide a window into protein evolution, dynamics and allostery. For example, we used this approach to identify sites with engineerable allostery in the Inward Rectifier K^+ channel Kir2.1, and inserting a light-switchable domain into these sites rendered Kir2.1 activity sensitive to light (13). Perhaps other channel-based opto- and chemogenetic reagents can be constructed in a similar manner. The SPINE-generated insertion library can be used with different downstream genotype-phenotype assays other than measuring surface expression, including measuring abundance as a proxy for protein stability (30) or enzyme activity coupled to cell survival (139). This makes SPINE a broadly useful insertional mutagenesis technique that offers the opportunity to generate large-scale domain insertion datasets to exhaustively explore the critical parameters that contribute to the construction of synthetic fusion proteins, such as, the location of the insertion, linker length, and linker composition. Empirical rules for protein engineering derived from SPINE-generated datasets may be useful to improve algorithms used in rationale, computational, and rule-based approaches.

Chapter 4

Note this was originally published here:

Coyote-Maestas W, Nedrud D, Antonio Suma, Matreyek KA, Fowler DM, Carnevale Vi, et al. The biophysical basis of protein domain compatibility in ion channels. Biorxiv. :69.

Author Contributions

W.C.-M., D.S., and D.N. conceived the study. W.C.-M. and D.N. generated libraries and performed insertional scans. D.N. coded alignment and enrichment pipelines for data analysis. W.C.-M. carried out machine learning, correlation analysis, and data mining. D.S. conducted clustering analysis and structural mapping. A.S. and V.C. conducted molecular dynamics simulations. K.A.M. and D.M.F. provided reagents and technical advice to construct mammalian cell lines from libraries. C.L.M. provided expertise for random forest model building and data mining. W.C.-M. and D.S. co-wrote the manuscript with input from all co-authors. This paper is referenced in the theses of both W.C.M. and D.N.

The biophysical basis of protein domain compatibility in ion channels

INTRODUCTION

Proteins domains are basic evolutionary units that allows the rapid emergence of new proteins by domain recombination (173). Accordingly, domain recombination-based engineered is often used to generate synthetic proteins in biomedical engineering (174). However, synthetically recombined proteins that fold and function well are typically the result of trial-and-error design and iterative optimization. Furthermore, deriving practical rules that accelerate domain recombination-based protein design is challenging because structure/function relationships of isolated and recombined domain differ (28).

To derive rules for productive domain recombination, we generated 760 polypeptide motifs (donors) insertions at all 436 amino acids of the inward rectifier K$^+$ channel Kir2.1 (recipient) and measured cell surface expression of channel-insertion variants. Previously, we found surprising variability between three insertional motif profiles that imply donor-recipient compatibility is complex (13). We chose 760 donor motifs as a representative sample to exhaustively study compatibility (**Table 4-1**). The massive scale of these experiments (over 300,000 variants) is possible due to insertional libraries with little bias (12) (**Figure 4-1**) and recombining libraries into stable cell lines (175).

Motif	Number of motifs	Number of motifs pass QC (%)	Ref.
common domains in extant prot.	20	20 (100%)	(34)
disordered protein fragments	105	89 (85%)	(35)
disordered proteins	54	27 (50%)	(35)
manually curated motifs	15	15 (100%)	n/a
polypeptide linkers	5	5 (100%)	n/a
ancestral motifs	40	38 (97%)	(36)
small non-domain proteins	6	5 (83%)	n/a
smotifs	39	38 (97%)	(37)
natural proteins < 50 AA	467	391 (84%)	(17)
peptide toxins	9	9 (100%)	n/a
Total	760	637 (84%)	

Table 4-1: Motif group statistics for Kir2.1 760 motif dataset.

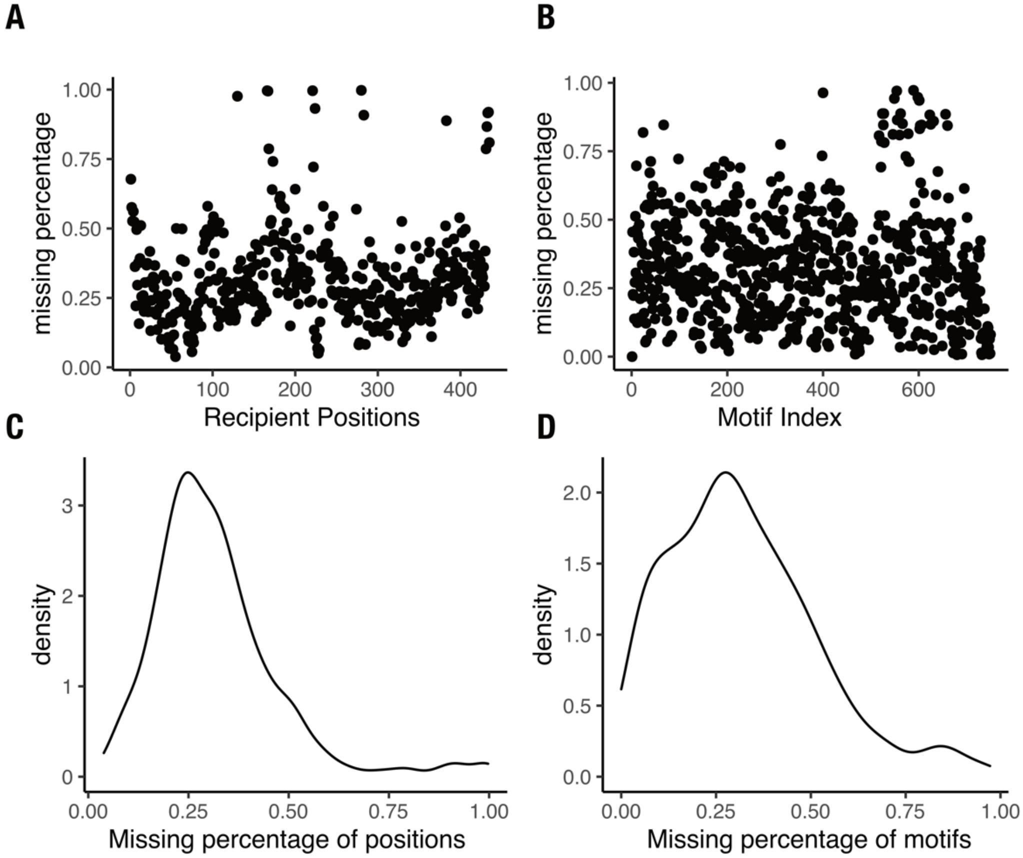


Figure 4-1: Insertional fitness coverage. (A-B) Scatter plots with the percent missing of Kir2.1 insertion fitness data after alignment by **(A)** position and **(B)** motif. **(C-D)** Density plots of Kir2.1 insertion fitness data percent missing by **(C)** position and **(D)** motif.

RESULTS

Systematic motif insertions reveal strong fitness pattern consistent with known ion channel biochemistry

Kir2.1 helps maintain the excitability of cells (83). For Kir2.1 to function it must fold, tetramerize, and traffic to the plasma membrane (82,95,176–178). We measure the impact of insertions on Kir2.1 with fluorescent antibody labeling and fluorescently activated cell sorting coupled to sequencing (**Figure 4-2A**). To quantify how motif insertions affect Kir2.1 surface expression, we calculate surface expression fitness of insertion variants as enrichment or depletion of surface expressed vs. non-surface expressed variants. This data is consistent with expected biochemistry (**Figure 4-2B-C**). Insertions into the extracellular FLAG tag used to label surface-expressed Kir2.1 decrease trafficking fitness because they disrupt antibody binding. Transmembrane region insertions (M1, M2, Pore, Filter) strongly decrease fitness (Wilcoxon rank sum test p-value < 2.2e-16) by impairing membrane insertion (177,179). Similarly, insertions in folding-critical core beta sheets of the C-terminal domain (CTD) (117) decrease fitness. Conversely, most insertions in the unstructured N- or C-termini are tolerated. As expected, insertions into Golgi export signals decrease surface expression. This is particularly strong for a N-terminal signal with tertiary structure (**Figure 4-2B**, positions 46-50, (82)). Insertion phenotypes in an ER export signal (the unstructured FCYENE signal (178), **Figure 4-2B**, positions 382-387) are more varied with some insertions not affecting surface trafficking. Perhaps the specific residue orientations that is required for function in structured export signal renders them more sensitive to motif insertion, while linear unstructured signals that rely on localized charge or hydrophobicity are more robust. Although insertional profiling reveals fitness patterns consistent with known biochemistry overall, the variability of insertion fitness across donor motifs and recipient insertion implies more complex mechanisms.

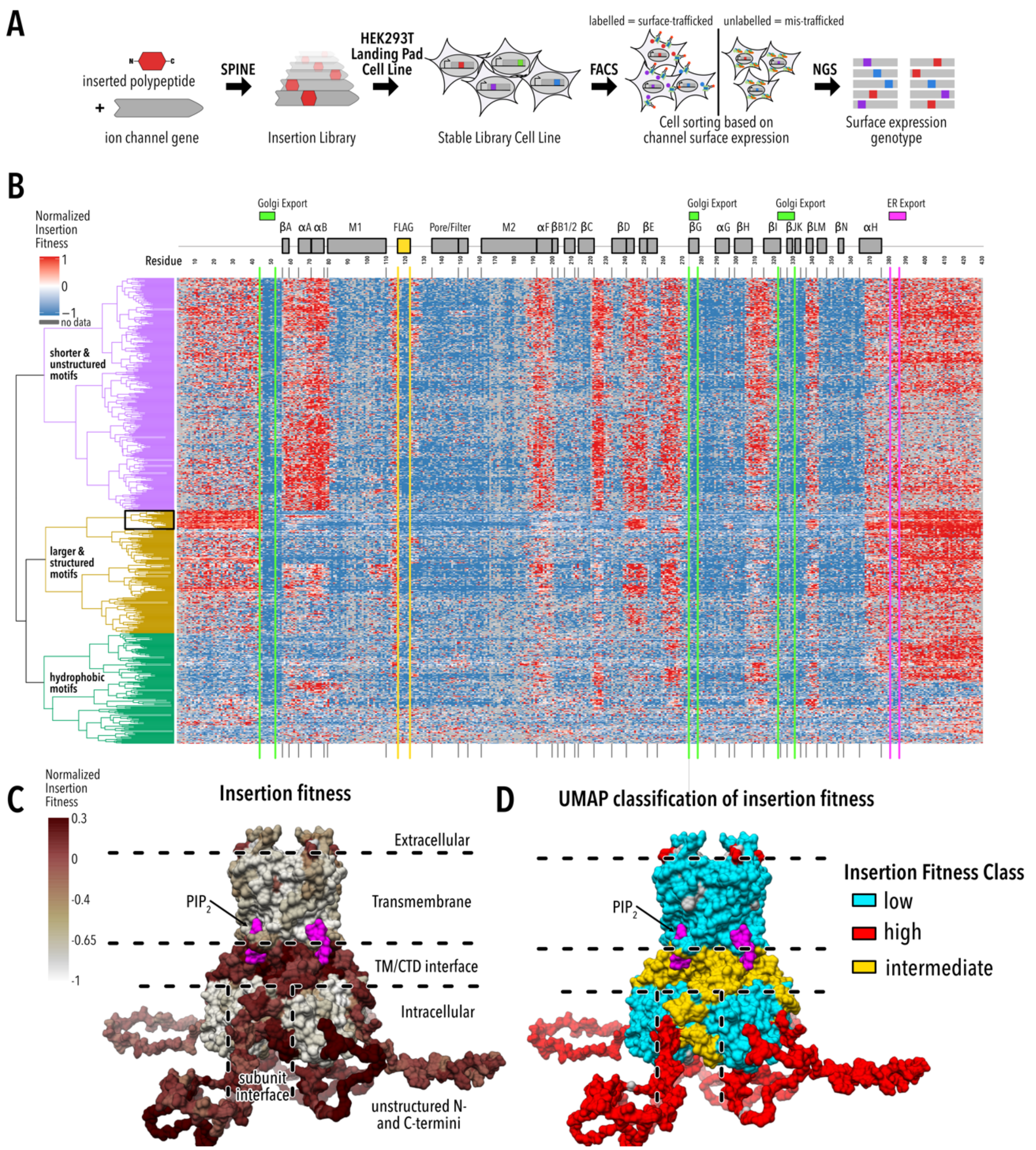


Figure 4-2: Large-scale insertional fitness profiling. (A) Motifs are inserted into all positions of a recipient protein using SPINE (12). A stable single-copy insertion library is generated by BxBI-mediated recombination in HEK293T (*6*). Cells are sorted based on channel surface expression determined by antibody labelling directed to an extracellular FLAG tag. Genotypes of each sorted cell populations are recovered by NGS. **(B)** Insertion fitness heatmap of 760 motifs inserted into all positions of Kir2.1. Secondary structural elements (grey boxes) are Kir2.1 are shown above, along known Golgi and ER export signals (green and magenta boxes, respectively). Motifs are hierarchically clustered by on a cosine distance metric. Dendrograms are colored by major motifs classes. The black

box indicates a subset of ‘well-structured motifs’ (see Figure 2F-H). **(C-D)** Mean normalized insertion fitness (C) or UMAP classification of Kir2.1 insertion fitness mapped onto the structure of Kir2.2 (PDB: 3SPI (87); 70% identity with Kir2.1). Fitness classes describe highly flexible and unstructured N/C termini (red), conformationally rigid and structured pore domain and CTD beta sheet core (cyan), and structured yet dynamic interface between TM and CTD, or between subunit in the CTD (yellow).

Recipient and donor properties interact to determine insertion fitness

To learn if donor properties affect fitness, we hierarchically clustered insertion fitness by motif. This revealed three groups: short unstructured motifs, larger folded motifs, and hydrophobic motifs (**Figure 4-2B**). Unstructured motifs are allowed in many parts of Kir2.1. Structured motifs, which contain nearly all motifs longer than 90 amino acids, are most allowed at the termini and spuriously in structured Kir2.1 regions. Hydrophobic motifs are distinct from other motifs clusters. They decrease fitness in regions (e.g., N terminus) that are universally compatible with the other two motif groups. Some hydrophobic motifs can be inserted where no other motifs can (e.g. beginning of M1 and end of M2 transmembrane helices). Taken together, this suggests that insertion fitness is influenced by the inserted motif’s properties.

To learn if recipient protein properties affect fitness, we used Uniform Manifold Approximation and Projection (UMAP (180)) clustering by insertion position. Three distinct clusters emerge (**Figure 4-3A**) corresponding to contiguous regions of Kir2.1 (**Figure 4-3D**). These regions represent the (1) pore domain and CTD core beta sheets, (2) unstructured N- and C-termini, and (3) PIP_2 (Kir2.1’s activator) binding sites, interfaces between the pore domain / CTD, and monomer interfaces within CTD. The emergence of discrete contiguous Kir2.1 regions from unbiased clustering suggests that local Kir2.1 properties influence insertional fitness, as well.

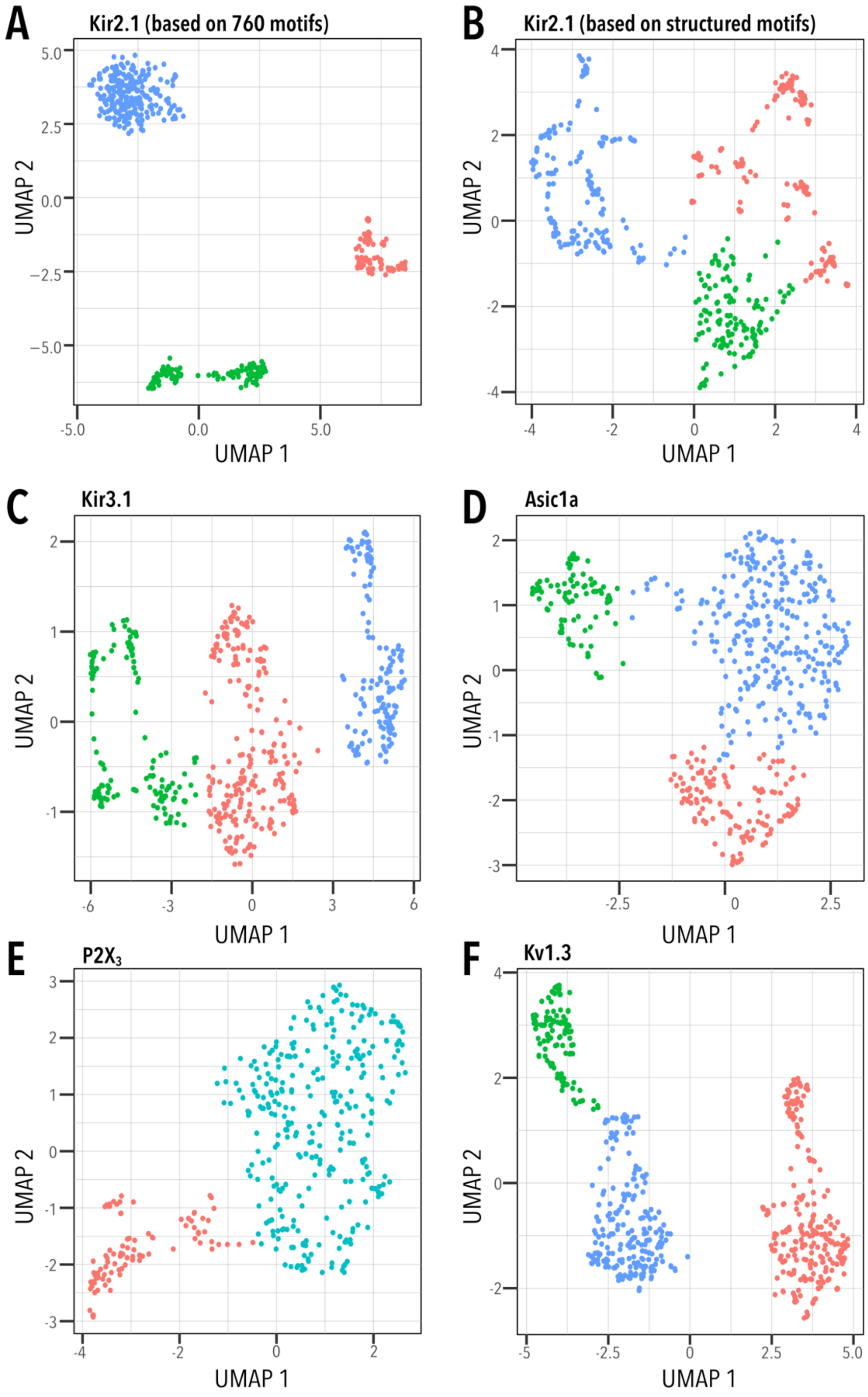
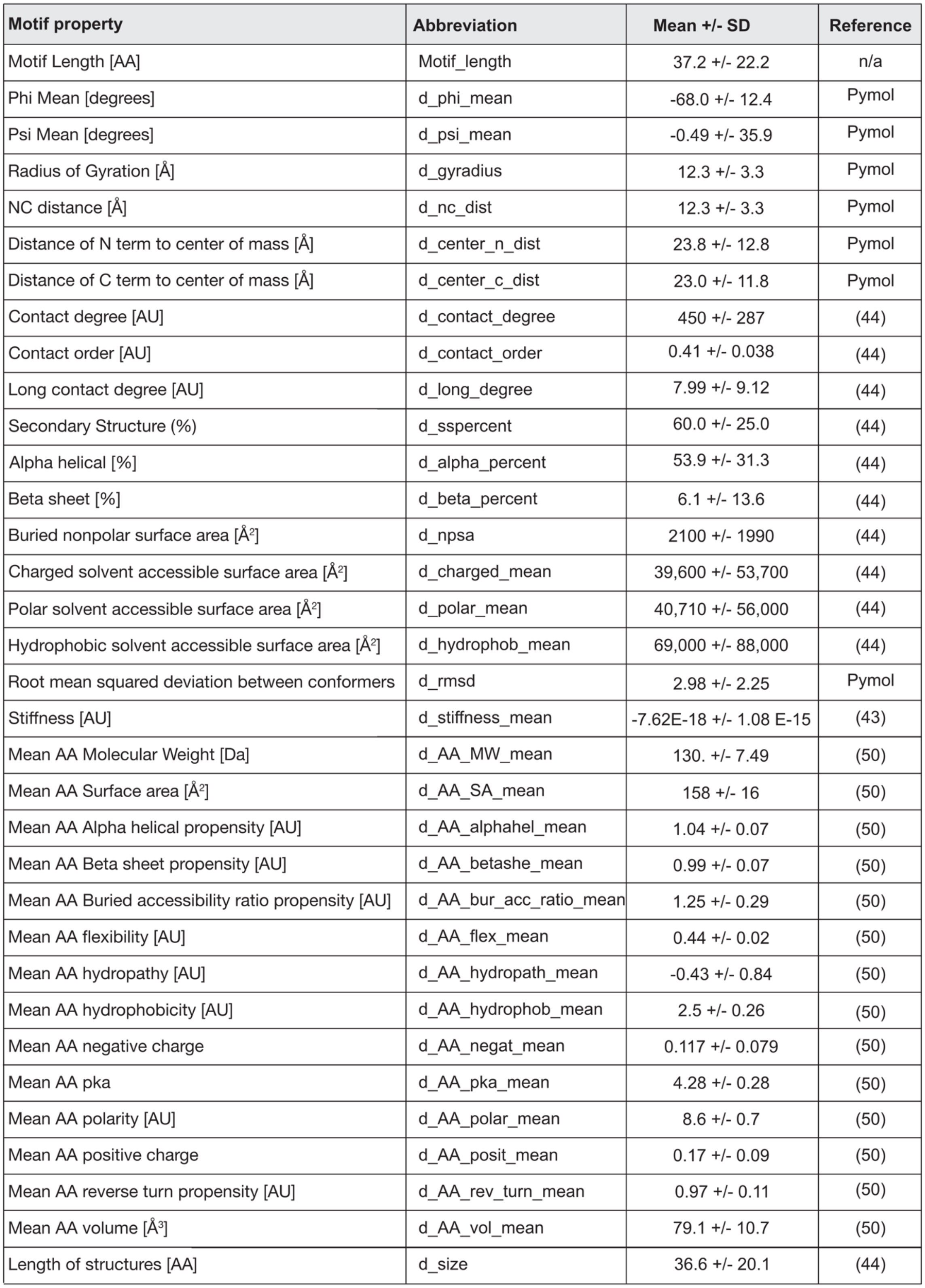


Figure 4-3: Unbiased clustering of insertion fitness. Uniform Manifold Approximation Projection (UMAP) was used to cluster insertion fitness of each channel. Cluster

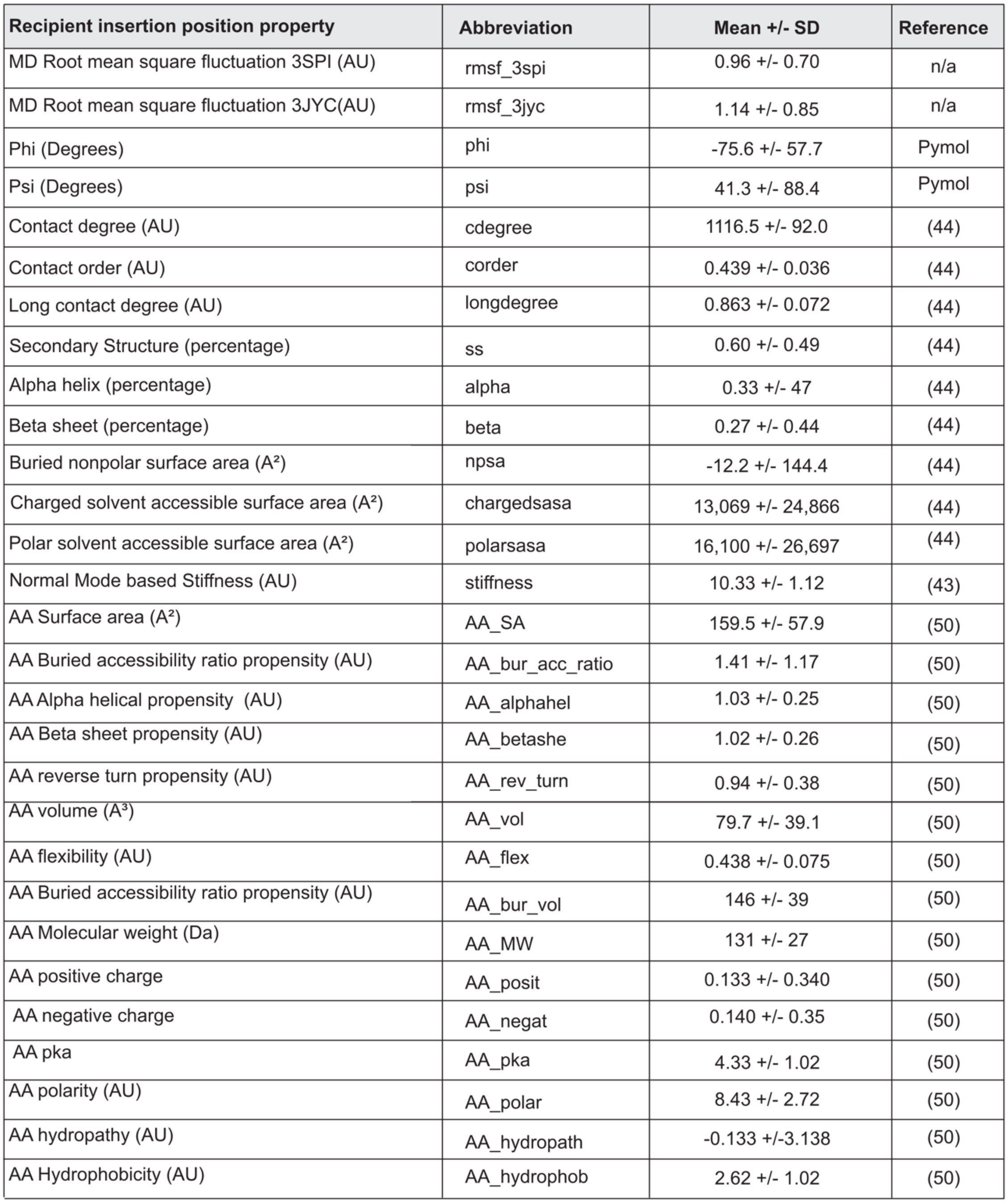
membership of each residue is indicated by color. Optimal cluster number was determined using Nbclust using the majority rule.

To identify the underlying biophysical properties that influence insertion fitness, we calculated sequence-, structure-, and dynamics-based properties of inserted motifs (**Table 4-2**) and recipient Kir2.1 (**Table 4-3**). We find that insertion fitness has moderately positive correlation with Kir2.1 backbone flexibility (molecular dynamics-derived root mean square fluctuation and anisotropic network model-derived stiffness; Pearson correlation coefficient 0.48 and -0.41, respectively, **Figure 4-4A**) implying that Kir2.1 rearranges structurally after motif insertion. Available space at insertion sites (e.g., contact degree) has a non-monotonic relationship (**Figure 4-4B**). Inserted motifs clusters have discrete property distributions, implying that motif biophysical properties make specific contributions to how insertions influence surface expression (**Figure 4-4C-H**). This is illustrated by a subcluster comprised of longer motifs containing hydrophobic and negatively charged residues (black box in **Figure 4-2B, Figure 4-4F-H**). While motif properties are clearly important, they behave non-linearly. For example, correlation of insertion fitness with motif length is negative for motifs under 25 amino acids but becomes positive for longer motifs (-0.33 and 0.22 Pearson coefficients, respectively, **Figure 4-4I**). Remarkably, all motif properties correlate positively and negatively with fitness dependent on insertion position. Motif lengths, for example, is positively correlated in flexible termini and loops but negatively correlated in the G-loop (**Figure 4-4M**). Our data provide highly-resolved information about both donor motifs and the recipient channel that captures the specific rules that govern insertional compatibility (**Figure 4-4M, Figure 4-5**). Hierarchical clustering correlations between fitness and motif properties at each residue separates Kir2.1 into three distinct classes (**Figure 4-4L, Figure 4-6**). These classes are similar to UMAP clustering of fitness alone (compare **Figure 4-2D** and **Figure 4-4L**, Pearson's $\chi 2$ test p-value < 2.2e-16, Cramer's V 0.42), suggesting motif and recipient properties can explain insertion fitness. Within each class, correlation sign (positive or negative) between fitness with inserted donor properties is identical. For example, all residues in the pore domain and beta sheet core of the CTD class positively correlate with motif hydrophobicity and negatively with polarity (**Figure 4-6**). In aggregate, this suggests that biophysical properties underlie insertional compatibility and properties of Kir2.1 (recipient) and inserted motif (donor) interact to determine fitness.



Motif property	Abbreviation	Mean +/- SD	Reference
Motif Length [AA]	Motif_length	37.2 +/- 22.2	n/a
Phi Mean [degrees]	d_phi_mean	-68.0 +/- 12.4	Pymol
Psi Mean [degrees]	d_psi_mean	-0.49 +/- 35.9	Pymol
Radius of Gyration [Å]	d_gyradius	12.3 +/- 3.3	Pymol
NC distance [Å]	d_nc_dist	12.3 +/- 3.3	Pymol
Distance of N term to center of mass [Å]	d_center_n_dist	23.8 +/- 12.8	Pymol
Distance of C term to center of mass [Å]	d_center_c_dist	23.0 +/- 11.8	Pymol
Contact degree [AU]	d_contact_degree	450 +/- 287	(44)
Contact order [AU]	d_contact_order	0.41 +/- 0.038	(44)
Long contact degree [AU]	d_long_degree	7.99 +/- 9.12	(44)
Secondary Structure (%)	d_sspercent	60.0 +/- 25.0	(44)
Alpha helical [%]	d_alpha_percent	53.9 +/- 31.3	(44)
Beta sheet [%]	d_beta_percent	6.1 +/- 13.6	(44)
Buried nonpolar surface area [Å^2]	d_npsa	2100 +/- 1990	(44)
Charged solvent accessible surface area [Å^2]	d_charged_mean	39,600 +/- 53,700	(44)
Polar solvent accessible surface area [Å^2]	d_polar_mean	40,710 +/- 56,000	(44)
Hydrophobic solvent accessible surface area [Å^2]	d_hydrophob_mean	69,000 +/- 88,000	(44)
Root mean squared deviation between conformers	d_rmsd	2.98 +/- 2.25	Pymol
Stiffness [AU]	d_stiffness_mean	-7.62E-18 +/- 1.08 E-15	(43)
Mean AA Molecular Weight [Da]	d_AA_MW_mean	130. +/- 7.49	(50)
Mean AA Surface area [Å^2]	d_AA_SA_mean	158 +/- 16	(50)
Mean AA Alpha helical propensity [AU]	d_AA_alphahel_mean	1.04 +/- 0.07	(50)
Mean AA Beta sheet propensity [AU]	d_AA_betashe_mean	0.99 +/- 0.07	(50)
Mean AA Buried accessibility ratio propensity [AU]	d_AA_bur_acc_ratio_mean	1.25 +/- 0.29	(50)
Mean AA flexibility [AU]	d_AA_flex_mean	0.44 +/- 0.02	(50)
Mean AA hydropathy [AU]	d_AA_hydropath_mean	-0.43 +/- 0.84	(50)
Mean AA hydrophobicity [AU]	d_AA_hydrophob_mean	2.5 +/- 0.26	(50)
Mean AA negative charge	d_AA_negat_mean	0.117 +/- 0.079	(50)
Mean AA pka	d_AA_pka_mean	4.28 +/- 0.28	(50)
Mean AA polarity [AU]	d_AA_polar_mean	8.6 +/- 0.7	(50)
Mean AA positive charge	d_AA_posit_mean	0.17 +/- 0.09	(50)
Mean AA reverse turn propensity [AU]	d_AA_rev_turn_mean	0.97 +/- 0.11	(50)
Mean AA volume [Å^3]	d_AA_vol_mean	79.1 +/- 10.7	(50)
Length of structures [AA]	d_size	36.6 +/- 20.1	(44)

Table 4-2: Inserted motif properties. This table only contains means and standard deviations of the insertion position properties. ≈ refers to Angstroms, AA refers to amino acids, Da refers to Daltons, and AU to arbitrary units.



Recipient insertion position property	Abbreviation	Mean +/- SD	Reference
MD Root mean square fluctuation 3SPI (AU)	rmsf_3spi	0.96 +/- 0.70	n/a
MD Root mean square fluctuation 3JYC(AU)	rmsf_3jyc	1.14 +/- 0.85	n/a
Phi (Degrees)	phi	-75.6 +/- 57.7	Pymol
Psi (Degrees)	psi	41.3 +/- 88.4	Pymol
Contact degree (AU)	cdegree	1116.5 +/- 92.0	(44)
Contact order (AU)	corder	0.439 +/- 0.036	(44)
Long contact degree (AU)	longdegree	0.863 +/- 0.072	(44)
Secondary Structure (percentage)	ss	0.60 +/- 0.49	(44)
Alpha helix (percentage)	alpha	0.33 +/- 47	(44)
Beta sheet (percentage)	beta	0.27 +/- 0.44	(44)
Buried nonpolar surface area (A²)	npsa	-12.2 +/- 144.4	(44)
Charged solvent accessible surface area (A²)	chargedsasa	13,069 +/- 24,866	(44)
Polar solvent accessible surface area (A²)	polarsasa	16,100 +/- 26,697	(44)
Normal Mode based Stiffness (AU)	stiffness	10.33 +/- 1.12	(43)
AA Surface area (A²)	AA_SA	159.5 +/- 57.9	(50)
AA Buried accessibility ratio propensity (AU)	AA_bur_acc_ratio	1.41 +/- 1.17	(50)
AA Alpha helical propensity (AU)	AA_alphahel	1.03 +/- 0.25	(50)
AA Beta sheet propensity (AU)	AA_betashe	1.02 +/- 0.26	(50)
AA reverse turn propensity (AU)	AA_rev_turn	0.94 +/- 0.38	(50)
AA volume (A³)	AA_vol	79.7 +/- 39.1	(50)
AA flexibility (AU)	AA_flex	0.438 +/- 0.075	(50)
AA Buried accessibility ratio propensity (AU)	AA_bur_vol	146 +/- 39	(50)
AA Molecular weight (Da)	AA_MW	131 +/- 27	(50)
AA positive charge	AA_posit	0.133 +/- 0.340	(50)
AA negative charge	AA_negat	0.140 +/- 0.35	(50)
AA pka	AA_pka	4.33 +/- 1.02	(50)
AA polarity (AU)	AA_polar	8.43 +/- 2.72	(50)
AA hydropathy (AU)	AA_hydropath	-0.133 +/-3.138	(50)
AA Hydrophobicity (AU)	AA_hydrophob	2.62 +/- 1.02	(50)

Table 4-3: Recipient insertion position properties. This table only contains means and standard deviations of the insertion position properties. ≈ refers to Angstroms, AA refers to amino acids, Da refers to daltons, and AU to arbitrary units.

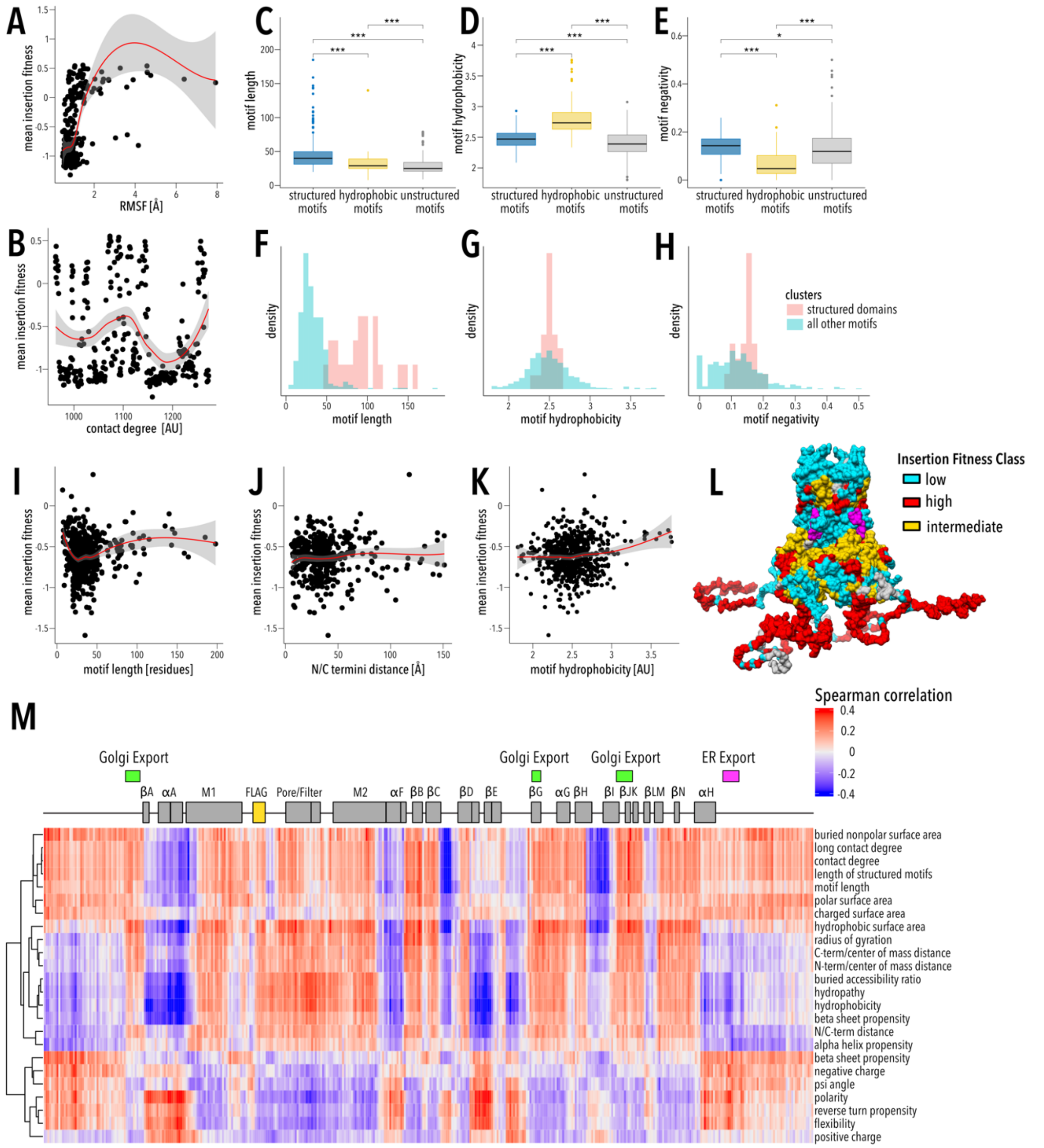


Figure 4-4: Relationships between fitness data and computed properties. Pairwise scatterplots between recipient properties (**A** – RMSF, **B** – contact degree) and insertion fitness. **(C-E)** Boxplots of motif **(C)** length, **(D)** hydrophobicity, and **(E)** negativity across the three motif clusters from **Figure 4-2B**. Median is marked with a block line, boxes represent the interquartile range, outlier points are shown, and P-values from a pairwise Wilcoxon tests are shown. **(F-H)** Density plots of motif **(F)** length, **(G)** hydrophobicity, and **(H)** negativity of the domain cluster and all other motifs colored. Density is weighted group size to allow direct comparison between different sized groups. **(I-K)** Pairwise scatterplots

between motif properties (**I** – motif length and **J** – NC termini distance, **K** – motif hydrophobicity) and insertion fitness. **L**. Hierarchical clusters of motif properties corelations with Kir2.1 position (**Figure 4-5**) is mapped onto the structure of Kir2.2 (PDB: 3SPI (87);70% identity with Kir2.1). The regulator PIP_2 is shown in magenta. **M.** Spearman correlation plot between motif properties and the fitness of that motif at each position. Properties are hierarchically clustered. A LOESS regression curve is fitted to each scatterplot, with the red line represents the fit and the gray area represents the 95% confidence interval. Boxplot significance levels are *** p<0.001, ** p<0.01, and * p<0.05, respectively.

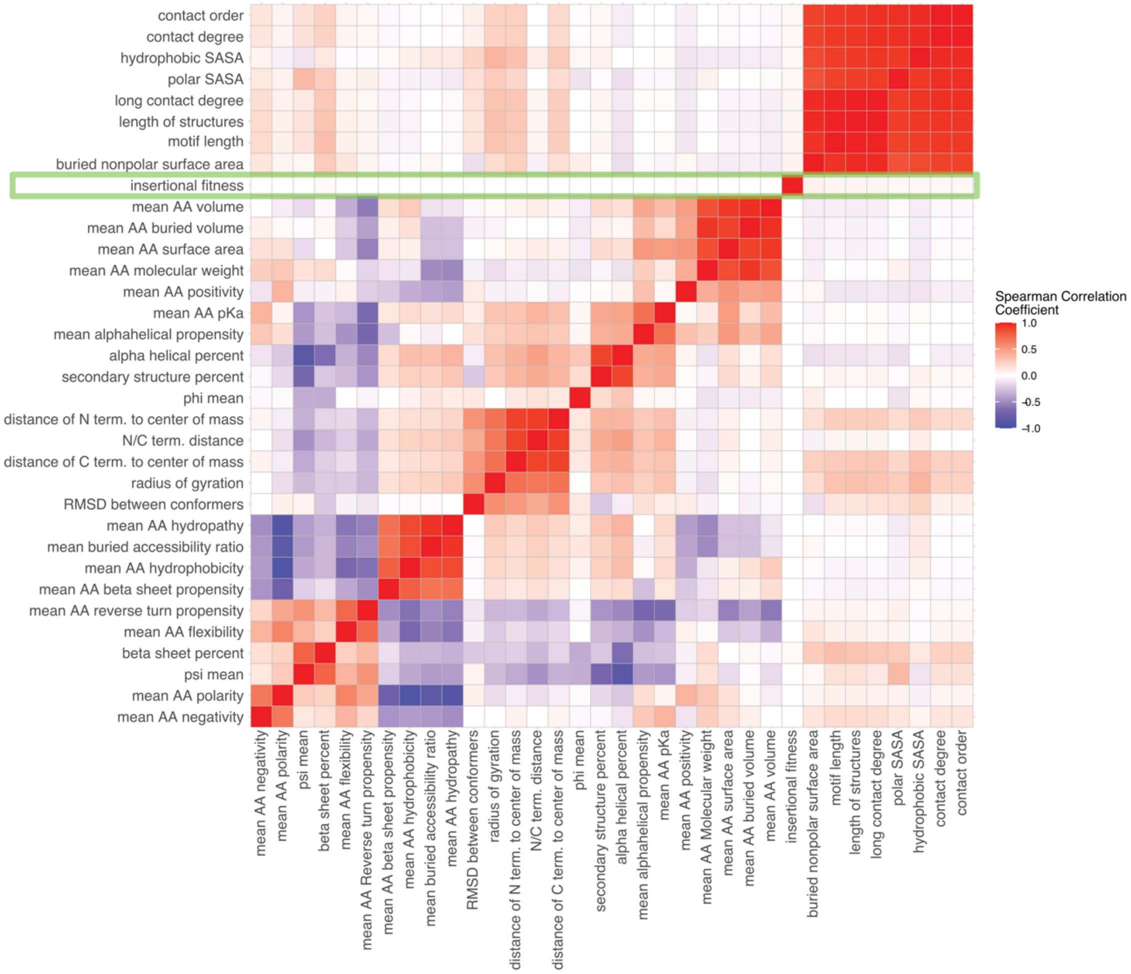


Figure 4-5: Motif properties and insertional fitness correlations. Correlation plot between motif property and the fitness across all positions. Insertional fitness is not correlated with any motif property. The motif properties and positions are hierarchically clustered (dendrograms not shown) and the plot is colored with spearman correlations

increasing from blue-to-red. AA refers to amino acids and SASA refers to solvent accessible surface area.

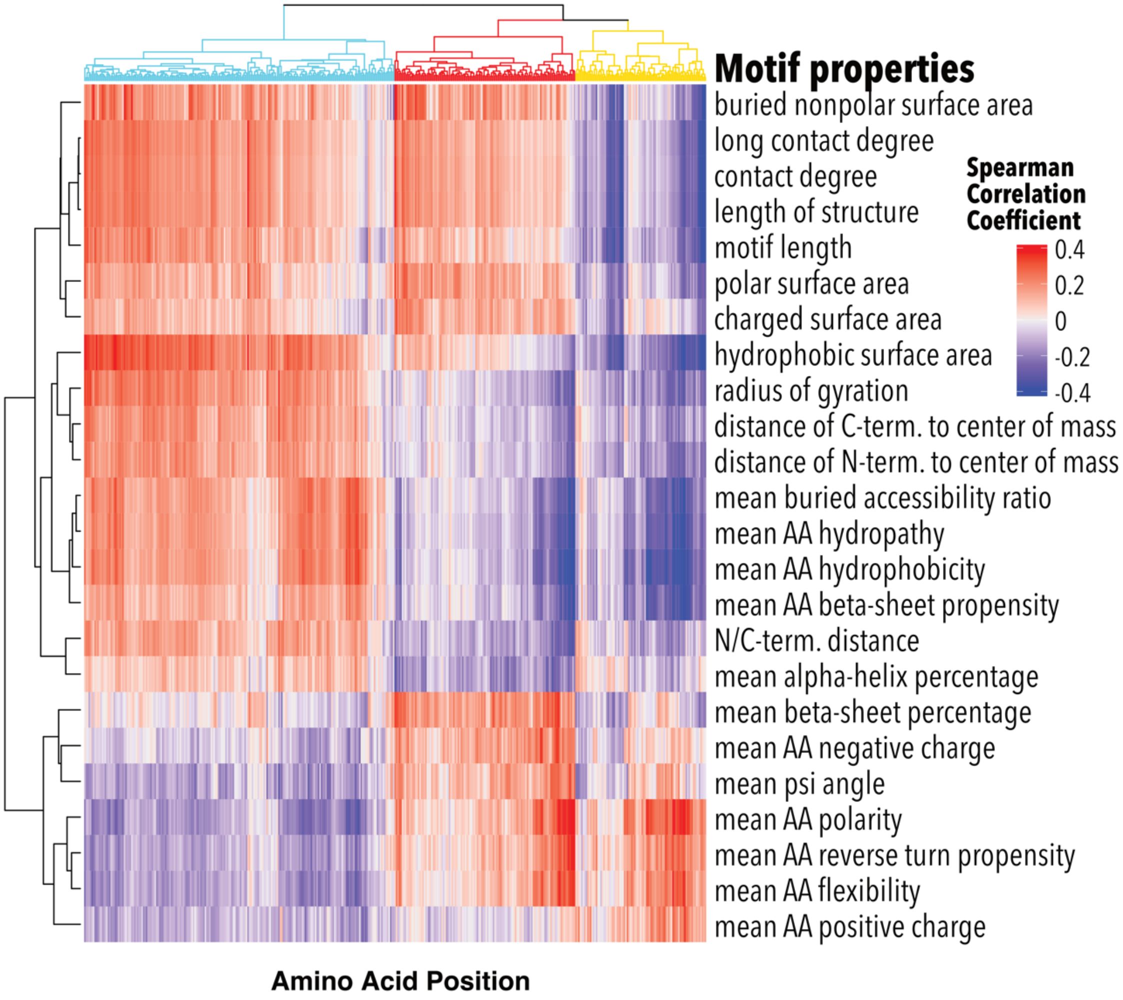


Figure 4-6: Clustered positions and properties correlation plot. Correlation plot between motif properties and the fitness of that motif at each position. The motif properties and positions are hierarchically clustered. Position clusters dendrogram branches are colored (cyan, red, yellow) as in **Figure 4-4L**.

Machine learning reveals the basis for donor/recipient compatibility

To identify which donor and recipient properties are important and how they interact in compatible insertions, we used Machine Learning (ML). While ML methods are sometimes treated as black boxes, they are useful for exploring rich genotype/phenotype datasets with non-linear interactions (181). We trained and tested regression random forests to predict insertional fitness at every amino acid position based on recipient and motif properties. To identify the most important properties and aid interpretation, we reduced properties from over 900 to 10 based on redundancy and feature importance with little impact on performance (**Figure 4-7, Table 4-4**). The final model successfully predicts insertional fitness for all positions and motifs of data withheld from model training (**Figure 4-8**).

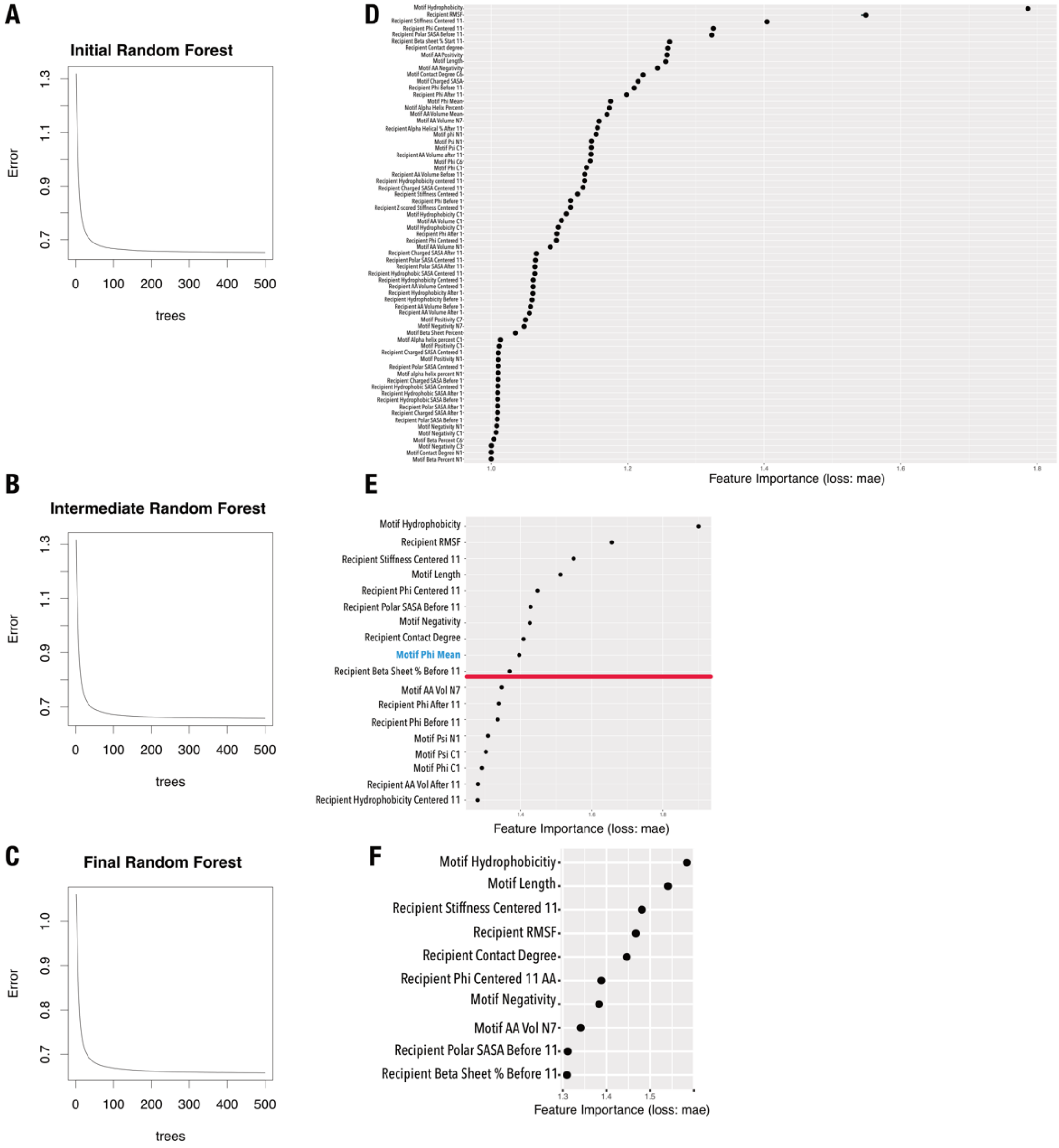
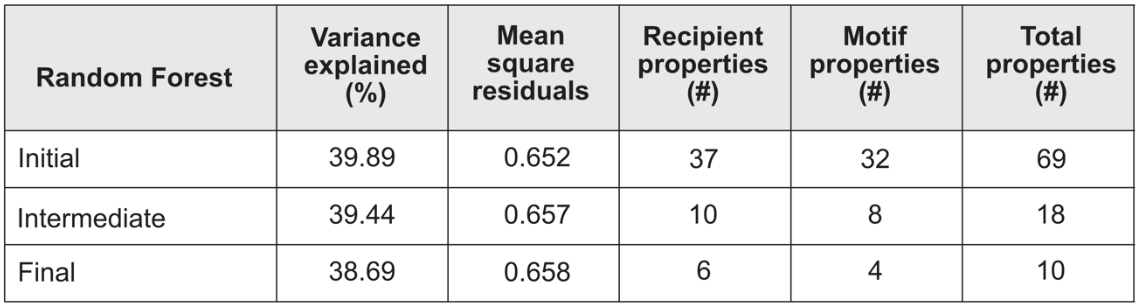


Figure 4-7: Random forest model iteration training and property importance. (A-C) Error curves with mean squared error plotted against number of trees in the **(A)** initial, **(B)** intermediate, and **(C)** final Random forest models. As more trees are added there is less

error. **(D-F)** Bar plots of the importance of features in predicting insertional fitness in the **(D)** initial, **(E)** intermediate, and **(F)** final Random forest models. In (**E)** the threshold that was used to trim features is marked with a red line. In addition, mean motif phi angle (blue) was removed because it required motifs to have solved structures, which substantially limited the number of motifs we could include. Property importance is based on the mean absolute error (mae) of removing properties from the predictive model. Further details can be found in the *Materials and Methods*.



Random Forest	Variance explained (%)	Mean square residuals	Recipient properties (#)	Motif properties (#)	Total properties (#)
Initial	39.89	0.652	37	32	69
Intermediate	39.44	0.657	10	8	18
Final	38.69	0.658	6	4	10

Table 4-4: Random forest parameters. Despite substantially reducing the number of properties, model performance based on variance explained and mean squared residuals are not significantly impacted.

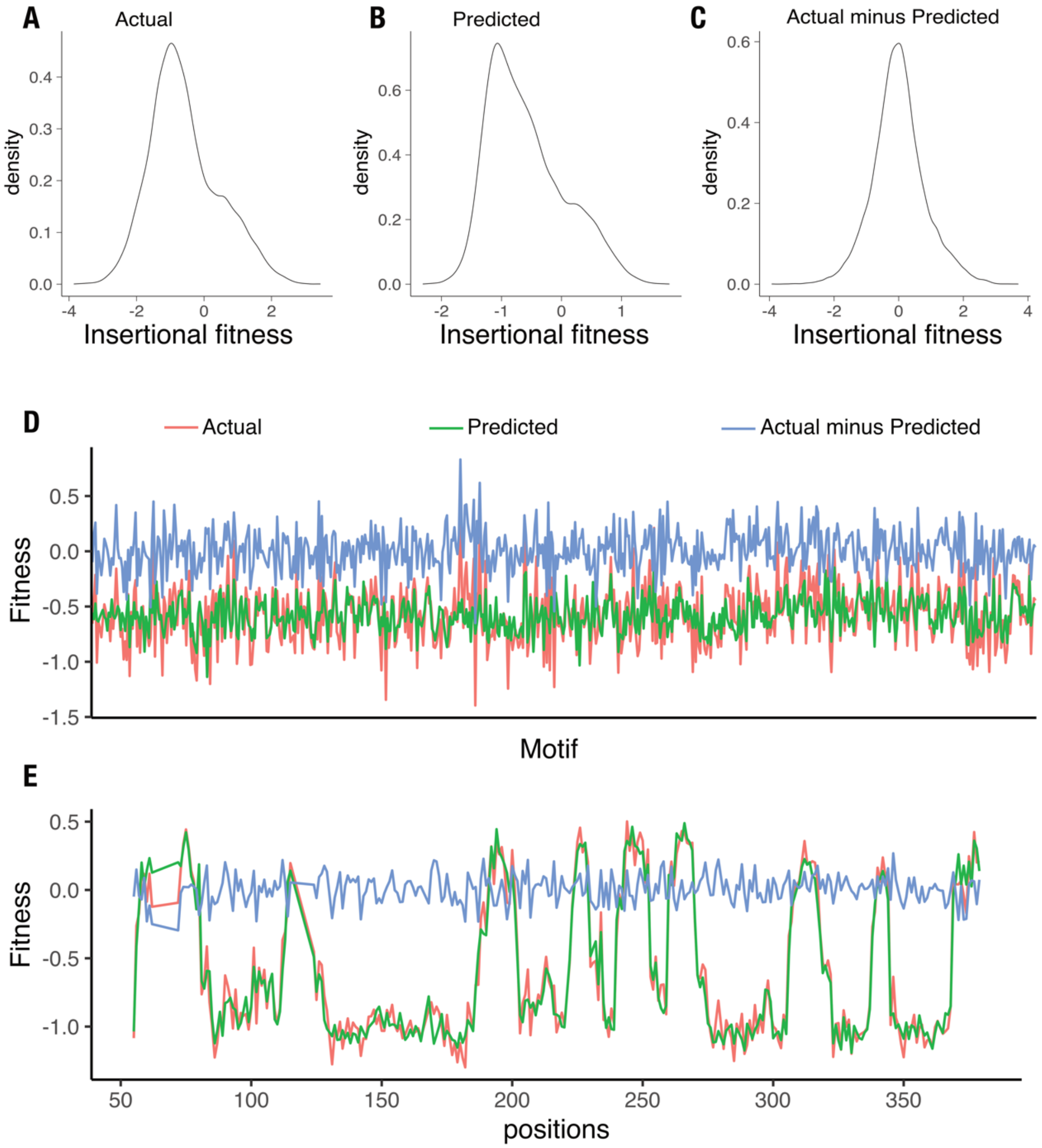


Figure 4-8: Model performance plots. (**A-C**) Density plots for (**A**) actual, (**B**) predicted, and (**C**) difference between actual and predicted insertional fitness. (**D**) Insertional fitness actual, predicted, and the difference per domain. (**E**) Insertional fitness actual, predicted, and difference per recipient insertion position. All model performance is reported based on data withheld from all random forest training.

Local Kir2.1 flexibility (RMSF and stiffness) is important for model performance and is positively associated with insertion fitness (**Figure 4-9A,E, Figure 4-10C,G**). Insertion position space (contact degree) plays a major non-linear role (**Figure 4-9A,E**). Apart from contact degree, all recipient properties have simple monotonic relationships with insertional fitness meaning recipient properties determine whether an insertion is viable (**Figure 4-10**).

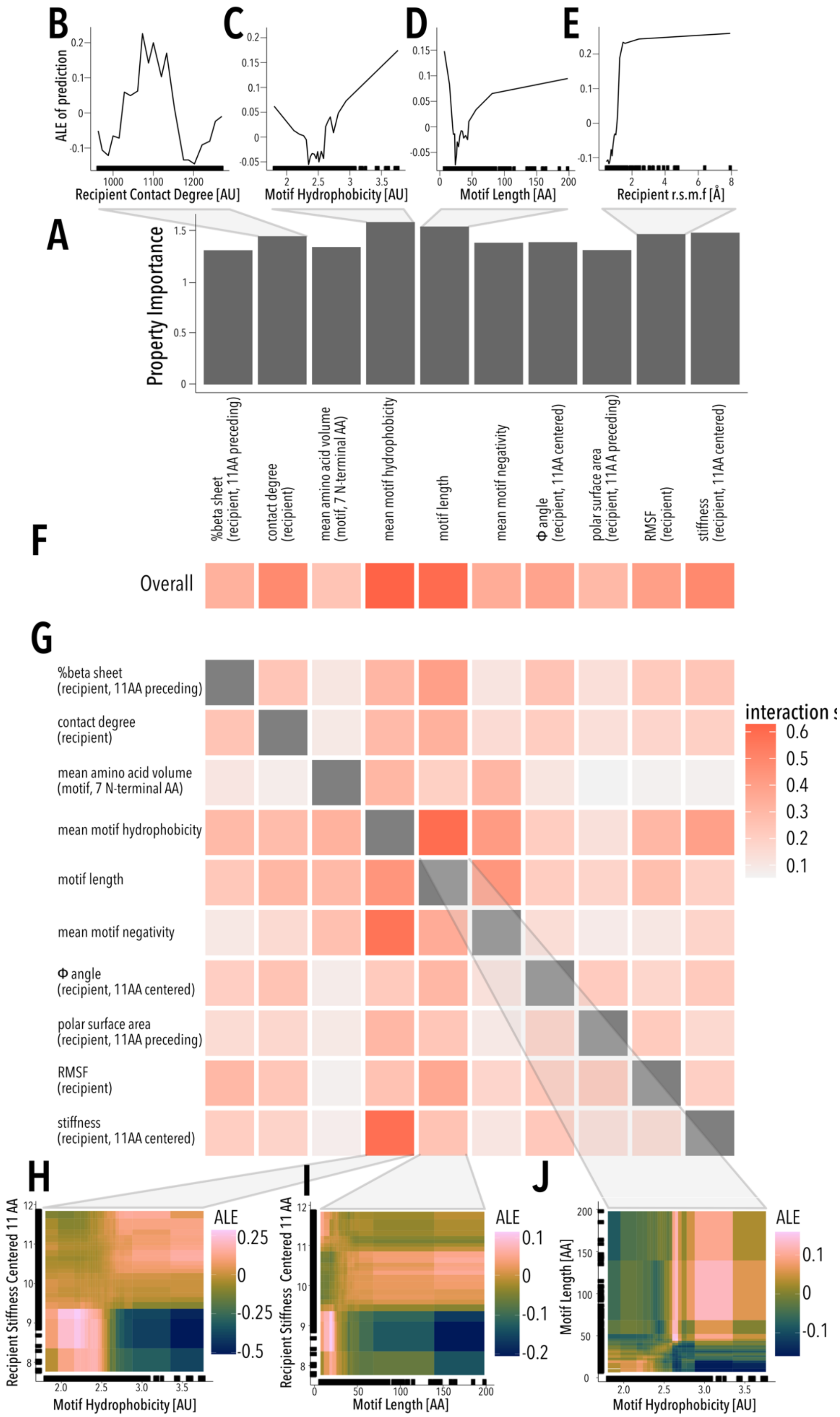

B
C
D
E
ALE of prediction
0.2
0.1
0
-0.1
1000
1100
1200
Recipient Contact Degree [AU]
0.2
0.15
0.1
0.05
0
-0.05
2
2.5
3
3.5
Motif Hydrophobicity [AU]
0.15
0.1
0.05
0
-0.05
0
50
100
150
200
Motif Length [AA]
0.2
0.1
0
-0.1
2
4
6
8
Recipient r.s.m.f [Å]
A
Property Importance
1.5
1
0.5
0
%beta sheet (recipient, 11AA preceding)
contact degree (recipient)
mean amino acid volume (motif, 7 N-terminal AA)
mean motif hydrophobicity
motif length
mean motif negativity
Φ angle (recipient, 11AA centered)
polar surface area (recipient, 11AA preceding)
RMSF (recipient)
stiffness (recipient, 11AA centered)
F
Overall
G
%beta sheet (recipient, 11AA preceding)
contact degree (recipient)
mean amino acid volume (motif, 7 N-terminal AA)
mean motif hydrophobicity
motif length
mean motif negativity
Φ angle (recipient, 11AA centered)
polar surface area (recipient, 11AA preceding)
RMSF (recipient)
stiffness (recipient, 11AA centered)
interaction s
0.6
0.5
0.4
0.3
0.2
0.1
H
I
J
Recipient Stiffness Centered 11 AA
12
11
10
9
8
2.0
2.5
3.0
3.5
Motif Hydrophobicity [AU]
ALE
0.25
0
-0.25
-0.5
Recipient Stiffness Centered 11 AA
12
11
10
9
8
0
50
100
150
200
Motif Length [AA]
ALE
0.1
0
-0.1
-0.2
Motif Length [AA]
200
150
100
50
0
2.0
2.5
3.0
3.5
Motif Hydrophobicity [AU]
ALE
0.1
0
-0.1

Figure 4-9: Machine learning model. (A) Bar plots of recipient or donor property importance in predicting insertion fitness. Importance is based on the mean absolute error of removing features from the predictive model. (**B-E)** Plots of the Accumulated local effects (ALE) of properties on prediction insertion fitness for **(B)** recipient contact degree, **(C)** motif hydrophobicity, **(D)** motif length, and **(E)** recipient RMSF. **(F, G)** Heatmap of each property's interaction strength **(F)** overall and **(G)** pairwise with every other property. **(H-J)** Pairwise ALE plots investigate how pairwise interactions contribute to prediction of **(H)** recipient stiffness-motif hydrophobicity, **(I)** recipient stiffness-motif length, and **(J)** motif hydrophobicity-motif length. Pairwise ALE plots are colored from dark blue to pink with increasing ALE scores. Marginal ticks **(B-E, H-J)** indicate values that are covered used in the property data.

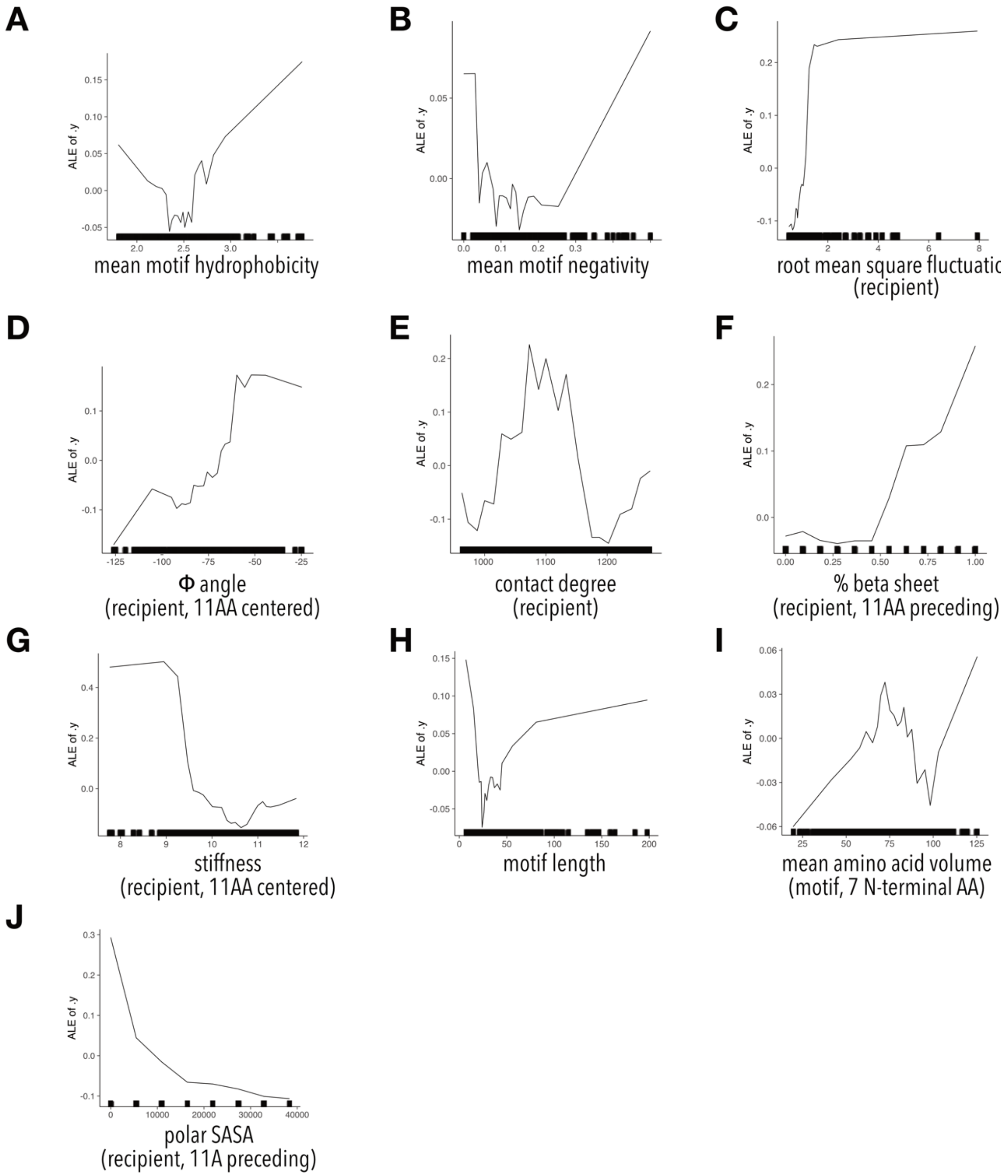


Figure 4-10: ALE plots for final model properties. Plots of the Accumulated Local Effects (ALE) of properties on the prediction of insertional fitness for **(A)** mean motif hydrophobicity, **(B)** mean motif negativity, **(C)** recipient root mean square fluctuation (based on MD simulation, PDB code 3JYC), **(D)** mean recipient phi angles of 11 AA centered around insertion site, **(E)** recipient contact density, **(F)** mean beta sheet content 11 AA before insertion site **(G)** mean recipient stiffness of 11 AA centered around insertion

site, **(H)** motif length, **(I)** mean amino acid volume of the motif's 7 N terminal AA, and **(J)** polar surface accessible surface area of 11 AA before insertion site.

The most important motif properties are length and hydrophobicity, which are both bimodal (**Figure 4-9C-D**). To understand why length and hydrophobicity are bimodal and how properties interact, we explored all property interactions **(Figure 4-9F-G)**. While recipient properties do not interact with each other, we find motif properties interact amongst themselves and with recipient properties **(Figure 4-9G)**. This suggests motif property interactions determine insertion fitness and not all insertions are equally compatible with each insertion position.

By exploring property interactions, we learn why different motifs behave distinctly (**Supp. Note 1, 4-11 4-13**). For example, low contact degree is strongly beneficial for large motifs (**Figure 4-11H**). Highly hydrophobic short donor motifs are deleterious within flexible regions (small flexible loops) likely because their solvent-exposed hydrophobic residues will be destabilizing and promote aggregation (**Figure 4-9H-I**) (182). The small motif cluster contains motifs that are shorter and less hydrophobic, making them less disruptive (**Figure 4-4C-D, Figure 4-12B-C**). In contrast, highly hydrophobic motifs are best allowed in buried regions with high stiffness and contact degree because these insertion positions minimize solvent exposure (**Figure 4-9H-I, Figure 4-13G**). Longer motifs benefit from strong positive interactions between motif length and moderate hydrophobicity likely allowing the formation of a hydrophobic core that can promote folding (**Figure 4-12J, Figure 4-11D**) (183). Well-folded domains can be stabilizing and promote insertion fitness when there is sufficient space otherwise large insertions disrupt the recipient protein’s folding (**Figure 4-11H**). Formation of a stable hydrophobic core as a desirable property of engineered domains corroborates conclusions from high-throughput protein design experiments (11).

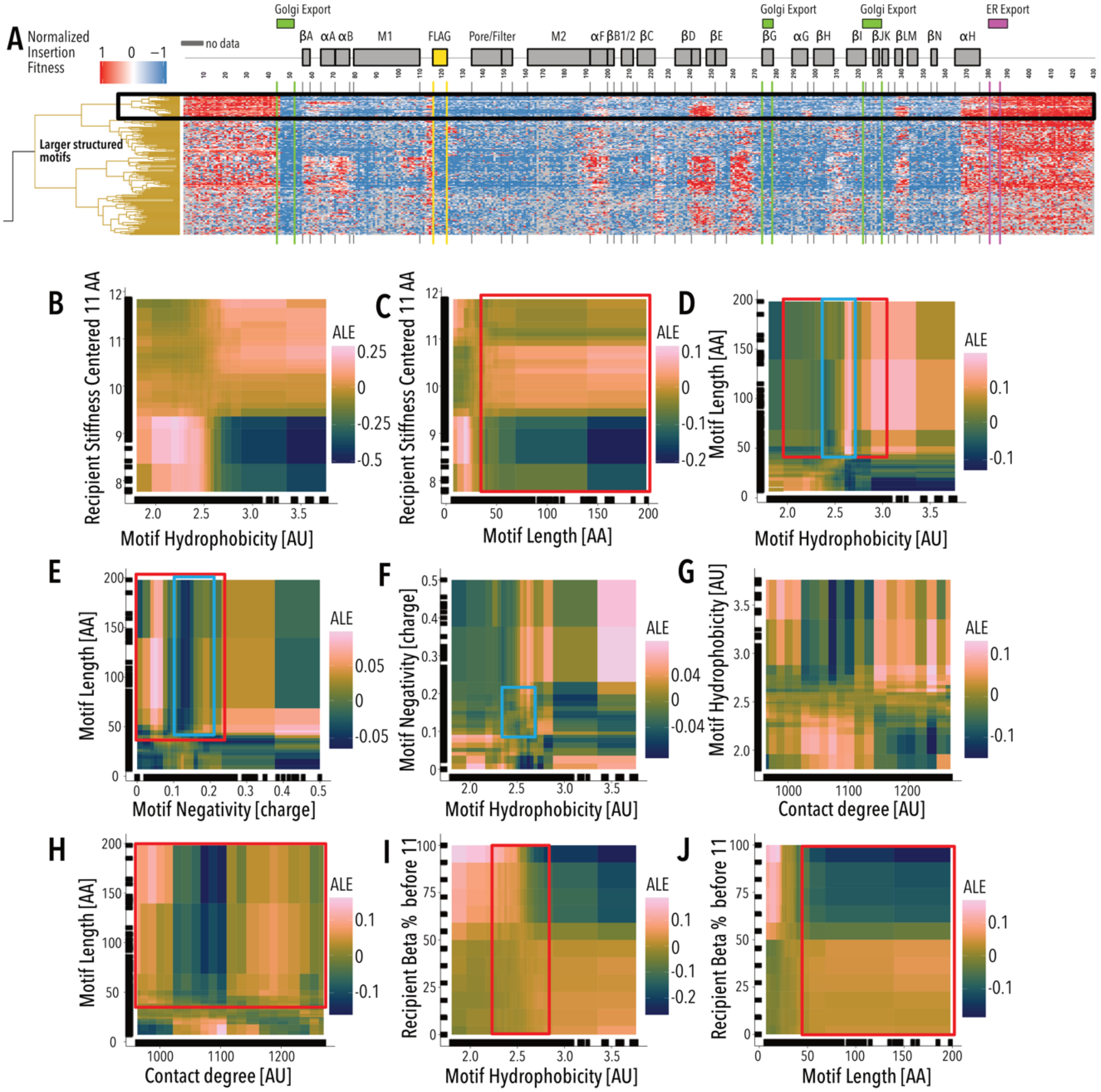


Figure 4-11: Larger structured motif cluster pairwise ALE Exploration. (A) Insertion fitness heatmap of structured motifs inserted into all positions of Kir2.1. Secondary structural elements (grey boxes) for Kir2.1 are shown above, along known Golgi and ER export signals (green and magenta boxes, respectively). Motifs are hierarchically clustered by on a cosine distance metric. The black box indicates a subset of 'well-structured motifs' (see **Figure 2F-H**). **(B-J)** Pairwise ALE plots investigate how pairwise interactions contribute to prediction of **(B)** recipient stiffness - motif hydrophobicity, **(C)** recipient stiffness - motif length, **(D)** motif hydrophobicity - motif length, **(E)** motif length - motif hydrophobicity, **(F)** motif negativity - motif hydrophobicity, **(G)** motif hydrophobicity - recipient contact degree, **(H)** motif length - recipient contact degree, **(I)** recipient Beta % - motif hydrophobicity, and **(J)** recipient beta % - motif length. Pairwise ALE plots are

colored from dark blue to pink with increasing ALE scores. The distribution of larger motifs cluster is boxed in red and the distribution of the well-structured is boxed in blue. Marginal ticks **(B–J)** indicate data point used in model building.

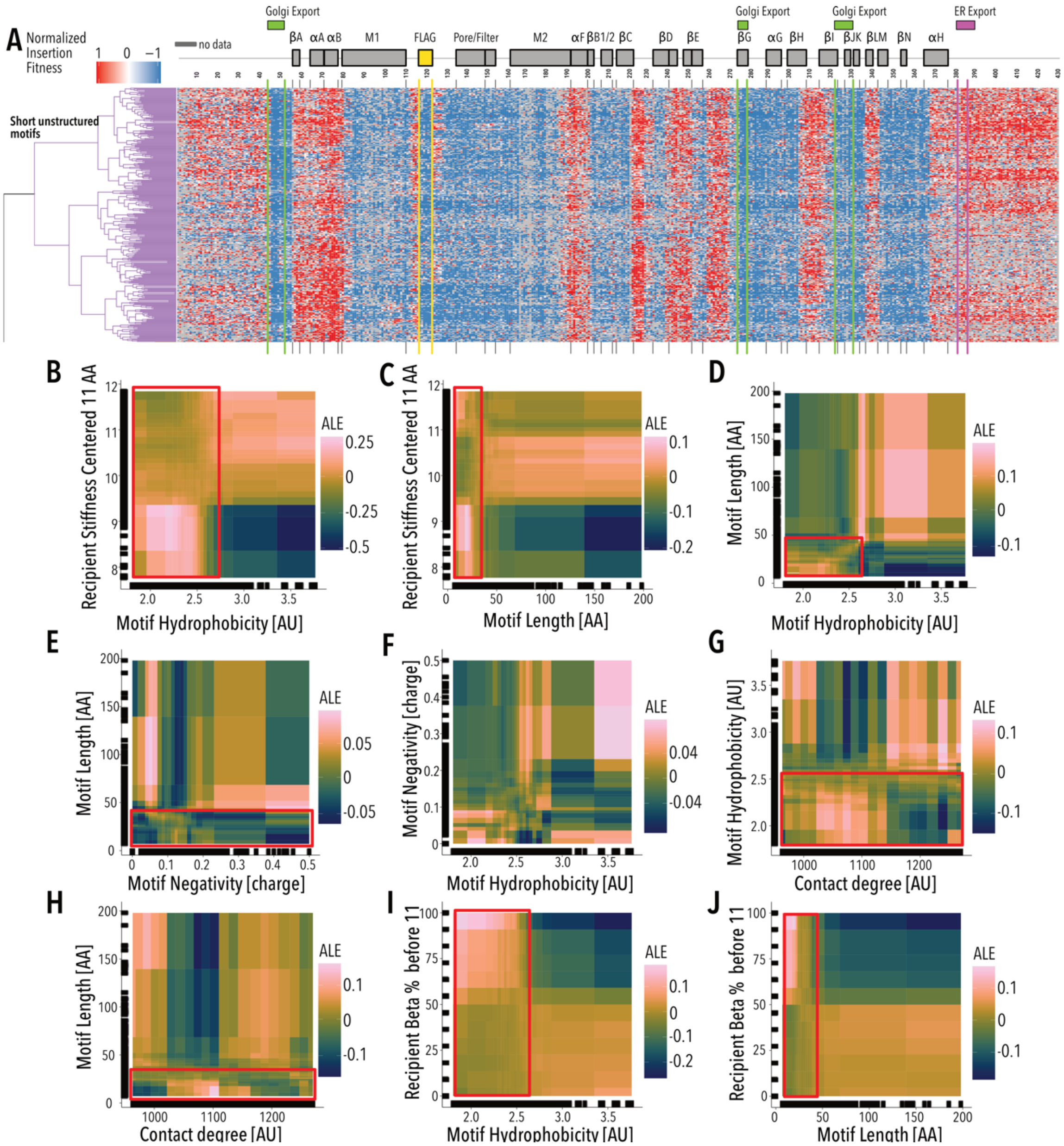


Figure 4-12: Short unstructured motif cluster pairwise ALE Exploration. (A) Insertion fitness heatmap of short unstructured motifs inserted into all positions of Kir2.1. Secondary structural elements (grey boxes) are Kir2.1 are shown above, along known Golgi and ER export signals (green and magenta boxes, respectively). Motifs are hierarchically clustered by on a cosine distance metric.**(B-J)** Pairwise ALE plots investigate how pairwise

interactions contribute to prediction of **(B)** recipient stiffness - motif hydrophobicity, **(C)** recipient stiffness - motif length, **(D)** motif hydrophobicity - motif length, **(E)** motif length - motif hydrophobicity, **(F)** motif negativity - motif hydrophobicity, **(G)** motif hydrophobicity - recipient contact degree, **(H)** motif length - recipient contact degree, **(I)** recipient Beta % - motif hydrophobicity, and **(J)** recipient beta % - motif length. Pairwise ALE plots are colored from dark blue to pink with increasing ALE scores. The distributions of hydrophobic motifs cluster are boxed in red. Marginal ticks **(B–J)** indicate data point used in model building.

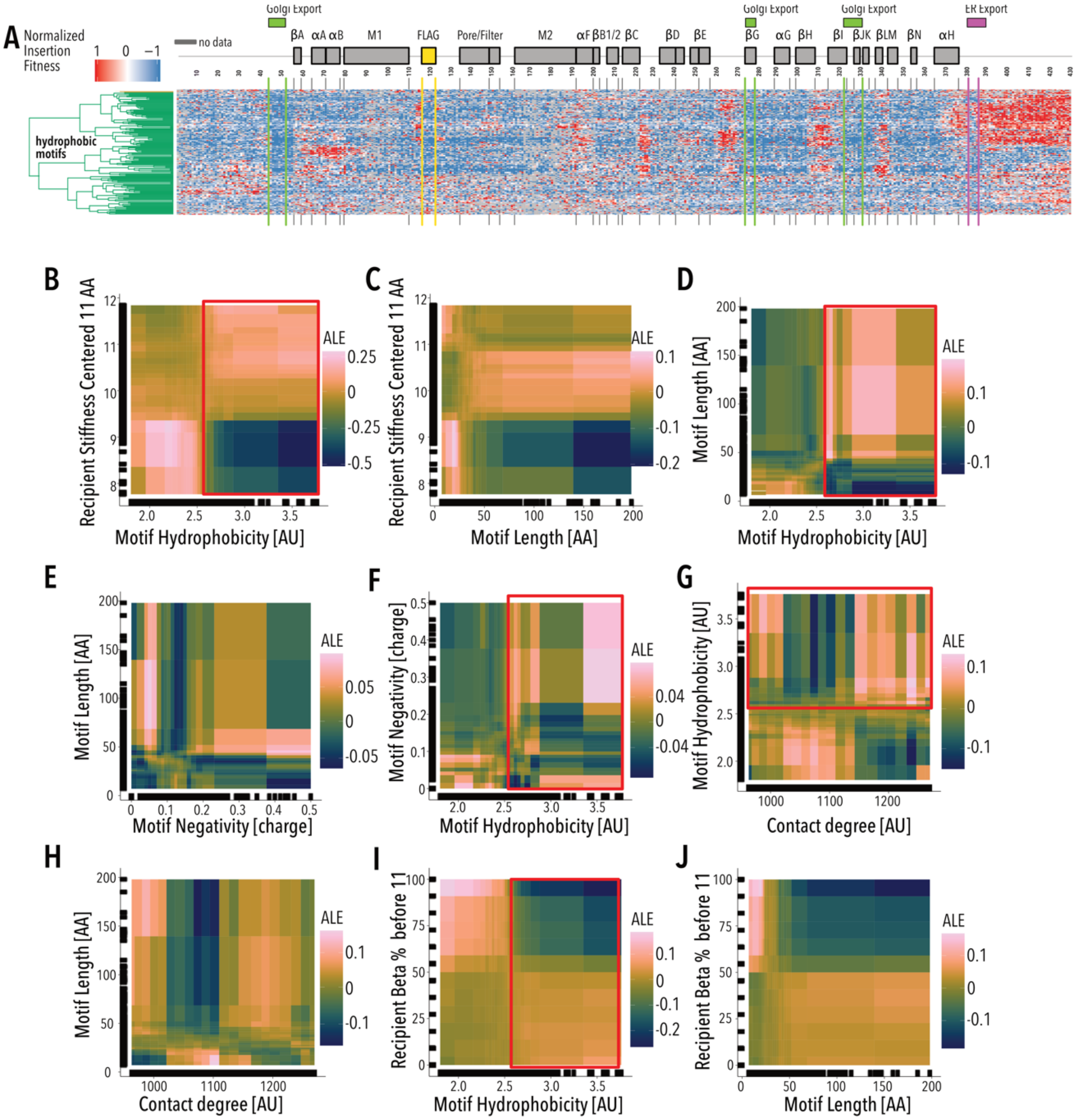
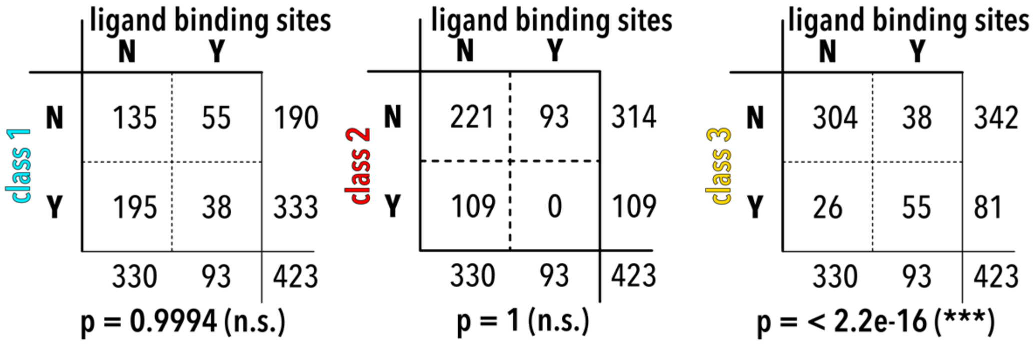


Figure 4-13: Hydrophobic motif cluster pairwise ALE Exploration. (A) Insertion fitness heatmap of hydrophobic motifs inserted into all positions of Kir2.1. Secondary structural elements (grey boxes) for Kir2.1 are shown above, along known Golgi and ER export signals (green and magenta boxes, respectively). Motifs are hierarchically clustered by on a cosine distance metric. **(B-J)** Pairwise ALE plots investigate how pairwise interactions contribute to prediction of **(B)** recipient stiffness - motif hydrophobicity, **(C)** recipient stiffness - motif length, **(D)** motif hydrophobicity - motif length, **(E)** motif length - motif hydrophobicity, **(F)** motif negativity - motif hydrophobicity, **(G)** motif hydrophobicity - recipient contact degree, **(H)** motif length - recipient contact degree, **(I)** recipient Beta % - motif hydrophobicity, and **(J)** recipient beta % - motif length. Pairwise ALE plots are colored from dark blue to pink with increasing ALE scores. The distributions of hydrophobic motifs cluster are boxed in red. Marginal ticks **(B–J)** indicate data point used in model building.

The ML model allows us to propose practical rules for successfully inserting donor motifs into recipient proteins. Insertion positions are ideally located in flexible protein regions with sufficient space. To form a well-folded domain, motifs need sufficient length and hydrophobic amino acid content to form a well-ordered hydrophobic core. If a desired insertion position is located within a buried and rigid region an inserted motif should be hydrophobic. More flexible regions prefer small non-hydrophobic insertions, and larger more structured domains will only be allowed if there is sufficient space and flexibility. Most significantly, the interactions between motifs and recipient properties determine the outcome of protein recombination. This stands in contrast to other domain recombination approaches that implicitly treat donor motifs as interchangeable (184).

Motif and recipient property interactions produce the discrete classes of motifs and regions (**Figure 4-2B,D, Figure 4-4L**). The rigid class with TM and CTD core beta sheets requires specific conformations to achieve a stable fold and allows few insertions. The flexible class with the N/C termini can adopt many conformations and allows most insertions. The class representing interfaces is an intermediate that is structured and dynamic. It contains many Kir2.1 regions (PIP_2 binding site, TM/CTD and subunit interfaces) that conformationally change upon PIP_2 binding and during closed to open state transitions (87,185). Since gating mechanisms are conserved across the inward rectifier family (186), the interface class may also be enriched for other inward rectifier regulator binding sites, such as $G\beta\gamma$

(GIRK), and ATP (Kir6.2). This is indeed the case (p-value < 2e-16, two-sided Fisher's Exact test, **Figure 4-14**). Taken together, classes suggest a hierarchical organization of inward rectifiers that balance the stability needed for folding with the conformational dynamics required for function.



		ligand binding sites		
		N	Y	
class 1	N	135	55	190
	Y	195	38	333
		330	93	423

p = 0.9994 (n.s.)

		ligand binding sites		
		N	Y	
class 2	N	221	93	314
	Y	109	0	109
		330	93	423

p = 1 (n.s.)

		ligand binding sites		
		N	Y	
class 3	N	304	38	342
	Y	26	55	81
		330	93	423

p = < 2.2e-16 (***)

Figure 4-14: Class / ligand binding sites contingency tables. Independence of inward rectifier ligand binding sites (PIP_2 – Kir2.1, Kir3.1, Kir6.2, Gβγ – Kir3.1 only, ATP – Kir6.2 only) with respect to different residue classes identified by unbiased clustering of insertion fitness was tested using two-sided Fisher's Exact tests. Only class 3 (colored yellow in **Figure 1D**) is enriched for ligand binding sites.

A hierarchical organization of ion channels that balances stability and flexibility for folding and function

To test if our compatibility framework and the hierarchical organization generalizes, we profiled surface expression fitness in the inward rectifier Kir3.1 (GIRK), the voltage-dependent K^+ channels Kv1.3, the purinoreceptor $P2X_3$, and the acid-sensing channel Asic1a by inserting a smaller set of 15 motifs (**Figure 4-15A**, **Table 4-5, Figure 4-16**). Kir3.1 is a G-protein regulated paralog of Kir2.1 with very similar structure (88) but requires co-expression of Kir3.2 for effective trafficking (176). Kv1.3, $P2X_3$, and Asic1a have different folds, gating, and regulation (165,186,187).

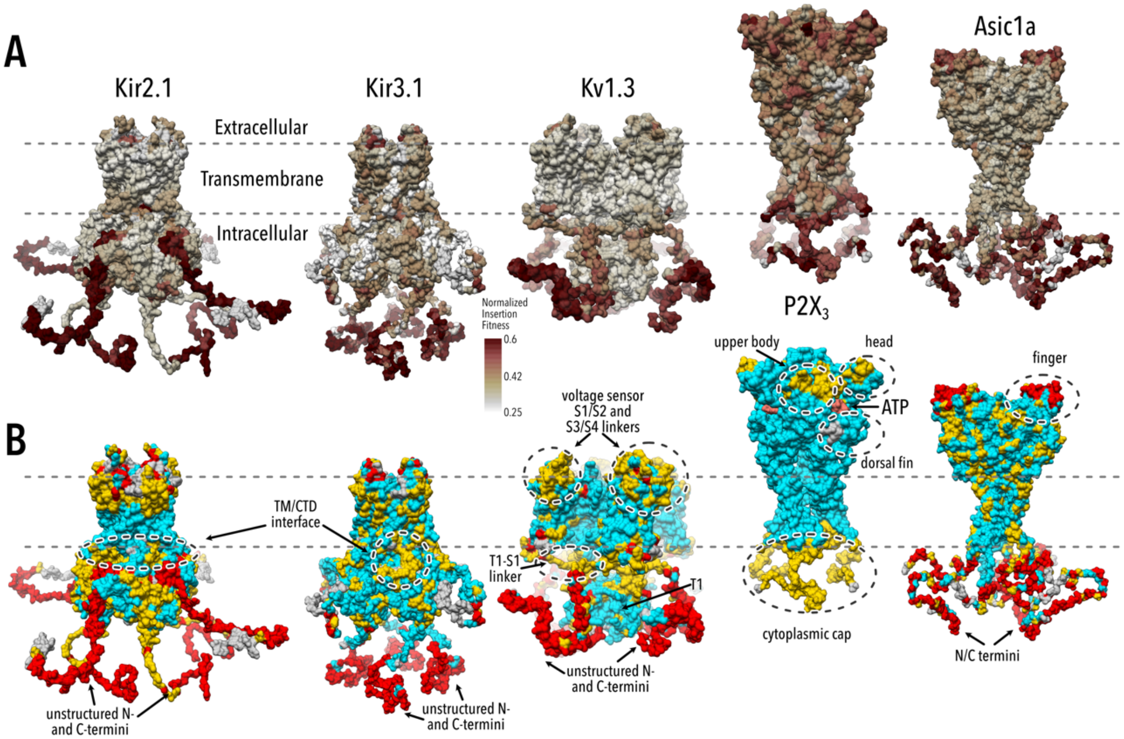
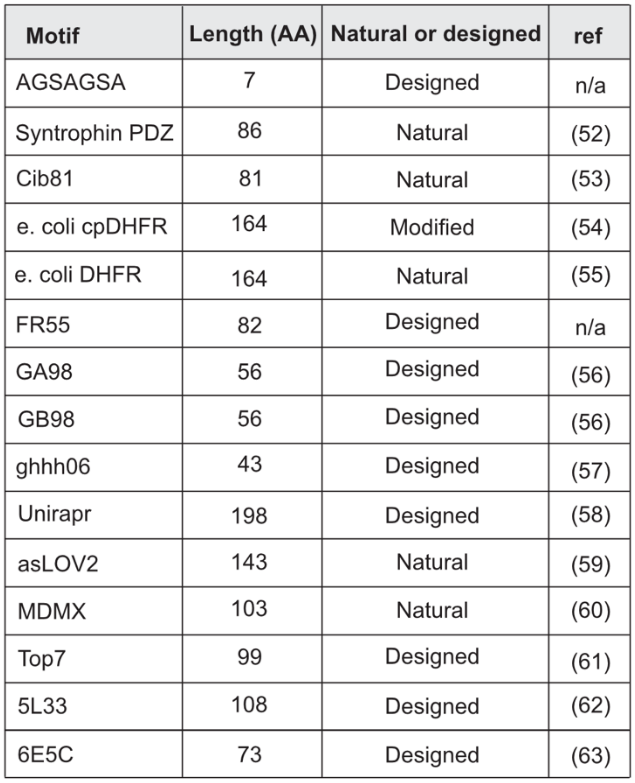


Figure 4-15: Generalization to other ion channels. Mean insertion fitness (**A**) and UMAP insertion fitness classification (**B**) mapped onto the crystal structures of Kir2.2 (PDB 3SPI (87); 70% identity with Kir2.1), Kir3.2 (PDB 4KFM (88); 45% identity with Kir3.1), Kv1.2/Kv2.1 paddle chimera (PDB 2R9R (165), 62% identical with Kv1.3), $P2X_3$ (PDB 5SVK (186)), and Asic1a (PDB 6AVE (187)). For all channels apart from P2X3, fitness classes describe highly flexible and unstructured regions (red), conformationally rigid and structured regions (cyan), and structured yet dynamic regions (yellow) that are enriched for ion channel structural elements important for gating transition (dashed circles). In P2X3, there are two regions with the unstructured and structured yet dynamic regions and unstructured regions combined (yellow) and the rigid region still discrete (cyan).



Motif	Length (AA)	Natural or designed	ref
AGSAGSA	7	Designed	n/a
Syntrophin PDZ	86	Natural	(52)
Cib81	81	Natural	(53)
e. coli cpDHFR	164	Modified	(54)
e. coli DHFR	164	Natural	(55)
FR55	82	Designed	n/a
GA98	56	Designed	(56)
GB98	56	Designed	(56)
ghhh06	43	Designed	(57)
Unirapr	198	Designed	(58)
asLOV2	143	Natural	(59)
MDMX	103	Natural	(60)
Top7	99	Designed	(61)
5L33	108	Designed	(62)
6E5C	73	Designed	(63)

Table 4-5: Smaller set of 15 motifs.

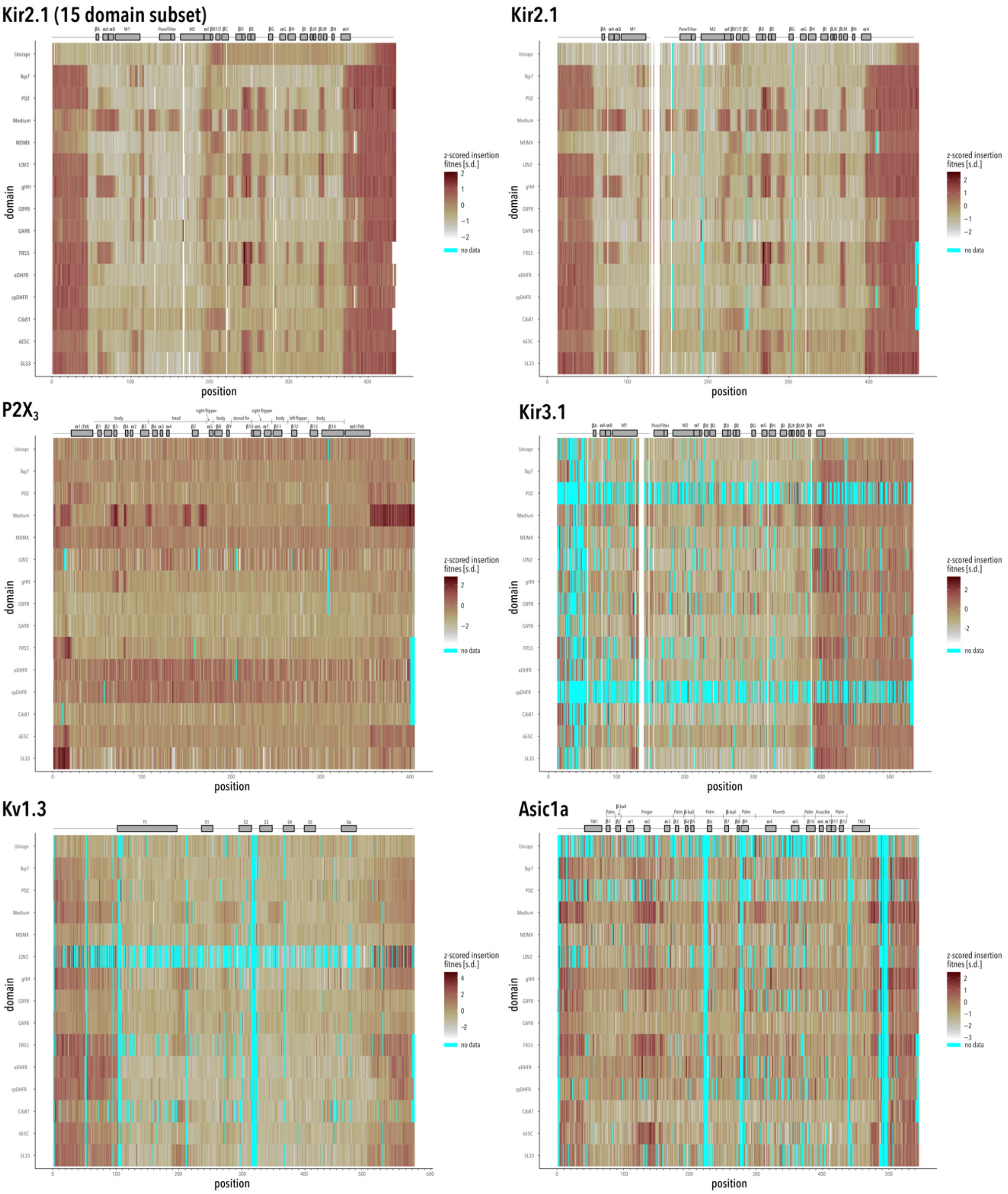


Figure 4-16: All datasets are based on at least two biological replicates. Two datasets are shown for Kir2.1 that were collected with different sequencing chemistry. Secondary structure elements (and topological organization; P2X3 and Asic only) are shown as cartoons.

The general patterns of surface expression in inward rectifiers also apply to Kv1.3, P2X$_3$, and Asic1a. There is weak to moderate correlation between the relative impact of each domain (**Figures 4-15 to 4-16**) in different channels, suggesting that while inserted motifs have similar effects across channels, the recipient channel's properties dominate. For related channels –Kir2.1 and Kir3.1– insertion profiles are fairly correlated (Pearson correlation coefficient 0.56). Insertions in membrane-embedded regions are deleterious, insertions into termini are allowed, and different inserted motifs give rise to distinct fitness profiles (**Figure 4-15**). This suggests that properties that dictate fitness in Kir2.1 are generalizable to other ion channels.

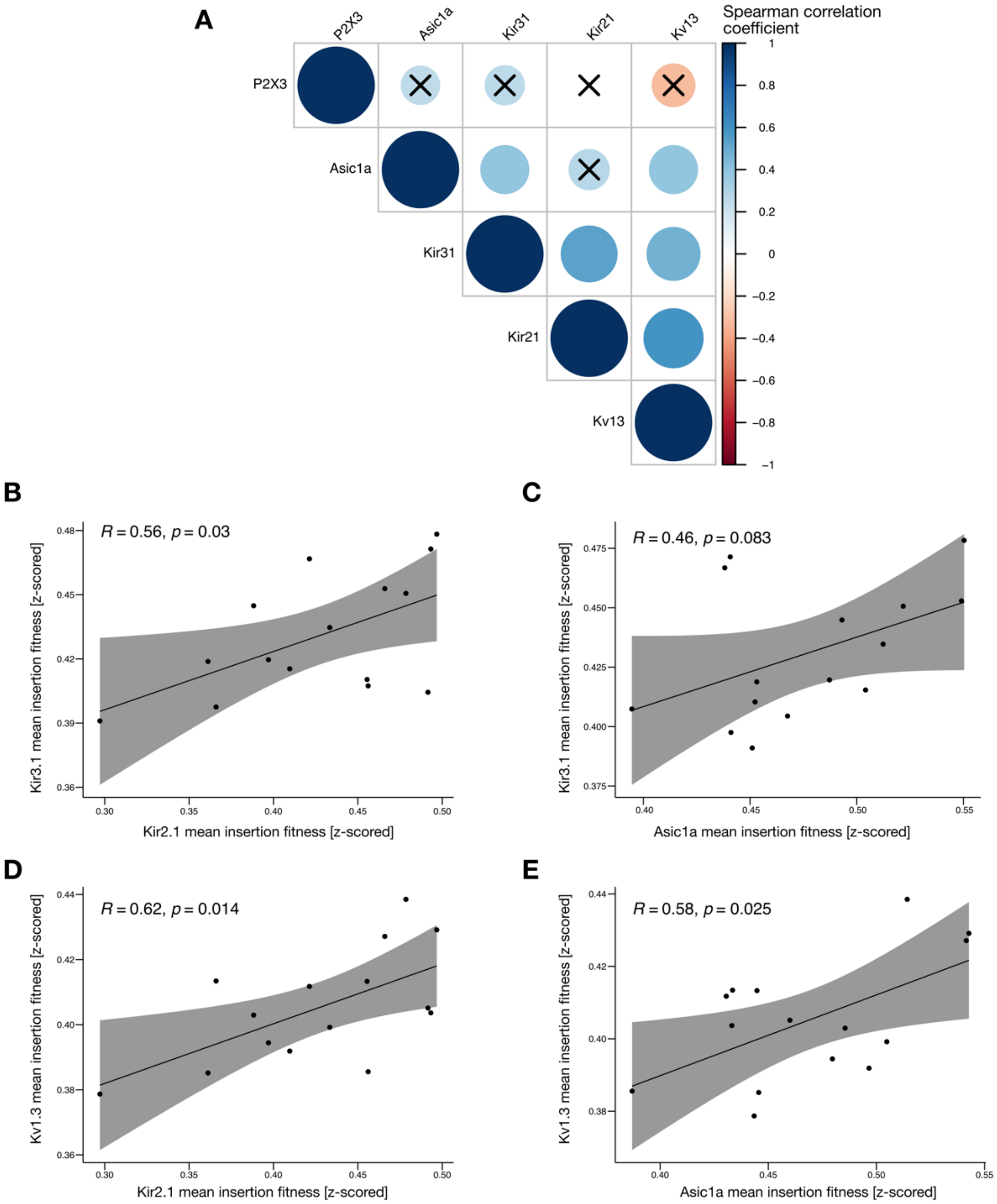


Figure 4-17: Correlation of domain insertion fitness in different ion channels. (A) Spearman correlation of mean insertion fitness (across all channel positions and motifs) between different channel pairs. Crosses indicate coefficient p-values > 0.05 (i.e., not significant). (**B-E**) Scatterplots of mean insertion fitness (across all channel position) for each inserted motifs. The solid black line indicates a linear regression and the grey shaded area indicates a 90% confidence interval. Spearman correlation coefficient and p-value are shown for each channel combination. Overall, correlation of motif effects on insertion fitness is moderate, suggesting a minor role relative to recipient channel properties.

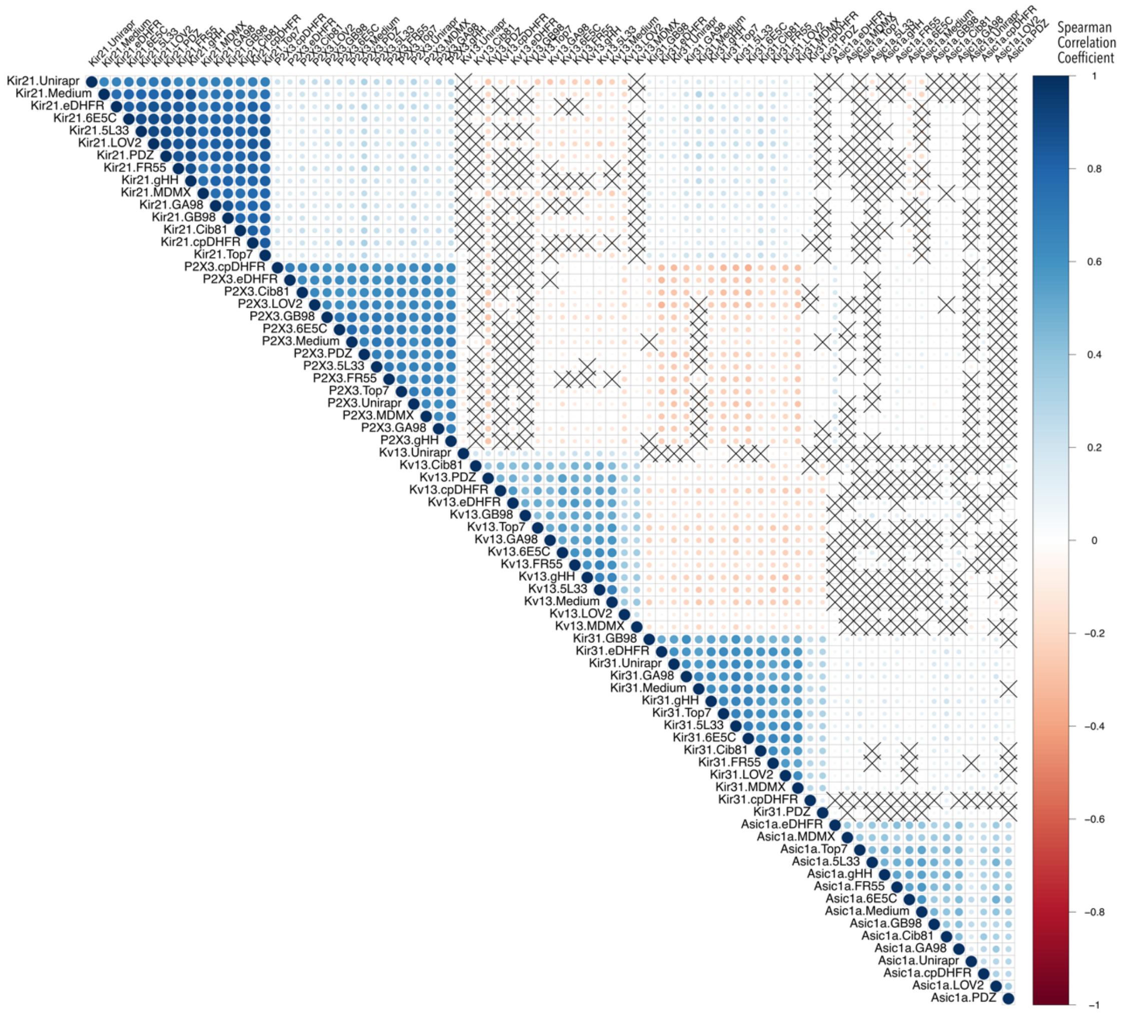


Figure 4-18: Correlations of insertion fitness for motifs in different channels. Spearman correlation of mean insertion fitness (across all channel position) of a specific motif in a specific channels with all other combinations. Strong correlations of different motifs in the same channel background dominate, suggesting that the recipient's

properties influence on fitness is strong. Crosses indicate coefficient p-values > 0.05 (i.e., not significant).

Since properties manifested as distinct classes in Kir2.1, we wondered if this concept would also apply to Kir3.1, Kv1.3 and P2X3. Applying the same UMAP-based clustering approach we used for Kir2.1, we find discrete insertion fitness classes in all channels (**Figure 4-15B**). As expected from shared fold architecture, Kir3.1's classes resemble Kir2.1's (Pearson's $\chi2$ test p-value <2.2e-16, Cramer's V 0.36) with three classes encompassing the TM and CTD core, regulator binding sites and interfaces, and termini. Using established structure/function data, we can infer that classes in each channel have distinct roles in folding stability and conformational dynamics. In each channel, there is a class that allows few insertions and corresponds to structural element required for tetramerization (Kv1.3 T1 tetramerization domain), folding (inward rectifier CTD Ig-like fold (117), P2X3 disulfide-stabilized ecto-domain(186), and ASIC1a beta sheets), or membrane insertion (transmembrane helices). Most channels have a class that allows nearly all insertions, and which coincides with flexible protein termini. The final class is intermediate, allowing only certain insertions. The intermediate class is enriched for residues that conformationally change during gating or regulation (e.g., Kir TM/CTD interface (87), Kv1.3 S1-T1 linker (188), $P2X_3$ cytoplasmic cap (186)).

DISCUSSION

We propose class organization is a universal feature of ion channels that results from constraints on channel structure to satisfy folding, assembly, and interaction with trafficking partners while providing flexibility for allosteric regulation and conformational changes during channel opening and closing. Other studies proposed a similar protein 'sector' concept, based on analyzing coevolution of residue pairs in large alignments across homologues (189). In contrast, our classes emerge from direct experimental data that are not constrained by statistical modeling's limitations and reflect underlying biophysical properties. Insertional profiling could be useful as a high-throughput coarse-grain structural biology method to study protein folding and dynamics from steady-state biochemical experiments. Further experiments are required to establish whether the hierarchical organization of insertion fitness extends to all protein classes.

Our dataset provides an unprecedented depth of information across hundreds of inserted donor motifs and several recipient ion channels. Using this dataset, we build a quantitative biophysical model of domain recombination in ion channels. Our discovery of specific interactions between donor and recipient properties is a crucial step towards universal domain recombination 'grammar' (190) for rational engineering of fusion proteins. Unbiased clustering of insertion fitness reveals a hierarchical organization of ion channels into regions with different material properties (rigid, semi-flexible, flexible) that play distinct roles to balance the stability needed for trafficking and the dynamics required for gating. As a universal organizing framework this may explain how contradictory requirements for stability and flexibility can be balanced to allow for well-folded and functional proteins.

Acknowledgements

We thank James Fraser, Gabriella Estevam, Margaret Titus, the Schmidt Lab, and Fraser lab for helpful feedback and discussion. We acknowledge support from the University of Minnesota Flow Cytometry Resource, in particular Rashi Arora, Therese Martin, and Jason Motl for providing flow cytometry technical support. Andrei Lupas and Vikram Alva kindly provided motif sequences from previous studies. Hellen Farrants and Kai Johnsson kindly provided DHFR and cpDHFR DNA for experiments.

Funding

This work was supported by the National Institutes of Health [MH109038 to D.S.] and University of Minnesota Genome Center Illumina S2 grant. W.C.-M. is supported by a National Science Foundation Graduate Research Fellowship and a Howard Hughes Medical Institute Gilliam Fellowship for Advanced Study.

Material & Methods

Choice of domains: We curated 760 motifs a representative sample of biophysical properties that drive donor/recipient compatibility (**Table 4-1**). Common domains in extant proteins are selected from SMART domain groups, focusing on those with available structural information, and varying range of frequencies within the human genome (188). The disordered protein fragments and proteins are from a curated disordered protein database, DISPROT (191). The protein fragments are derived from proteins with disordered regions, and the proteins are entire proteins that are disordered. The manually curated motifs include natural, synthetic proteins, several switchable proteins, and a flexible GSAG linker (**Table 4-5**). The polypeptide linkers are manually selected hydrophobic and hydrophilic subsections from Kir2.1. Ancestral motifs have been proposed by Alva et al. (192). The small non-domain proteins are manually selected monomeric small proteins which are not commonly recombined. The smotifs are super-secondary structural motifs that are common across proteins (193). The natural proteins <50 AA acid motifs are a set of proteins under 50 amino acids that do not contain cysteines that were used in a massive protein stability assay (11). Peptide toxins are a set of genetically encodable disulfide-rich neurotoxin peptides.

Molecular Biology: Genes encoding human Kir2.1 (Uniprot P63252), human Kir3.1 (Uniprot P48549), human Kir3.2 (Uniprot P48051), human Asic1a (Uniprot P78348), human $P2X_3$ (Uniprot P56373), and human Kv1.3 (Uniprot P22001) were produced by DNA synthesis (Twist Bioscience). A Kozak sequence (GCCACC) and P2A-EGFP were added prior and after each open reading frame, respectively. FLAG tag epitopes were added into previous described extracellular loops of Kir2.1 (between S116 and K117 (95)), Kir3.1 (between K114 and A115 (176)), Asic1a (between F147 and K148 (194)), and $P2X_3$ (between N72 and R73 based on insertion into paralog $P2X_2$ (195)). Golden Gate compatible 5' and 3' sites where added to each gene by inverse PCR.

Library generation: We generated motif insertion libraries using Saturated Programmed Insertional Engineering (SPINE) (12). Briefly, we use multi-step Golden Gate cloning to insert a series of motifs in between all consecutive residue pairs of a gene. We break up a gene into fragments (~169 bps or 53 amino acids) with a genetic handle cassette inserted at every amino acid position. The genetic handle has outward-facing BsaI type IIS restriction sites, which are replaced with any DNA fragment with short N-terminal Ser-Gly and C-terminal Gly-Ser of the inserted motif. We include an antibiotic cassette, Chloramphenicol, to remove background wildtype DNA and select for inserted library members. As a quality control step, we sequence all our libraries for baseline coverage prior to screens (**Figure 4-19)**.

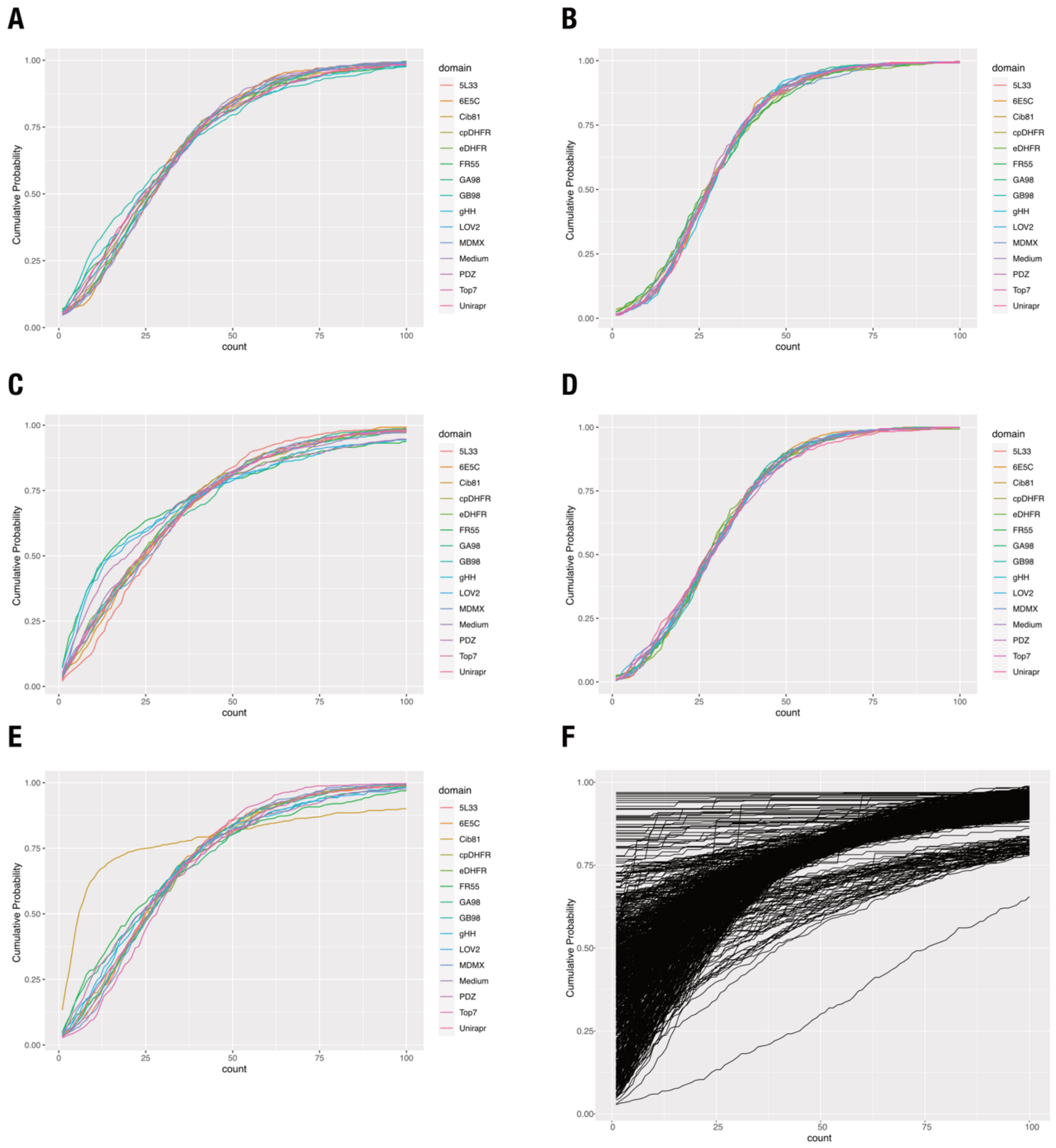


Figure 4-19: Baseline profiles for each domain and gene combination. (A-F) Empirical cumulative distribution plots for **(A)** ASIC1a, **(B)** Kir2.1, **(C)** Kir3.1, **(D)** P2X3, **(E)** Kv1.3, **(F)** Large domain set for Kir2.1. Each domain was normalized to have 30x coverage before calculating empirical cumulative distribution function. Plots show cumulative probability for each count threshold from 1 to 100. This indicates distribution of insertions in a given gene with distributions shifted to the right being more evenly distributed.

Cloning domains: The common domains, hand-curated motifs, and non-domain proteins were ordered as gene fragments (Twist Bioscience). The disordered, gene fragments, ancestral, structural, and motifs PDBs <50 amino acids were ordered in the form of an

OLS pool (Agilent). All motifs were mammalian codon optimized and designed with amplifiable barcodes and BsaI type IIs restriction sites complementary to those in the inserted genetic handle. Golden gate cloning is conducted with BsaI-v2 HF (NEB), T4 Ligase (NEB) following manufacturer’s instructions. Completed Golden Gate reactions were cleaned with Zymo Clean Concentrate kits and transformed into Lucigen E. cloni™ electrocompetent cells. Diversity was maintained at every step such that there are at least 30x successfully transformed colony forming units as determined by serial dilutions and plating an aliquot of liquid cultures.

Library cell line construction: To generate cell lines, we used a rapid single-copy mammalian cell line generation pipeline (*6*). Briefly, insertion libraries are cloned into a staging plasmid with BxBI-compatible *attB* recombination sites using BsaI Golden Gate cloning. We amplify the backbone using inverse PCR and the library of interest with primers that add complementary BsaI cut sites. Golden Gate cloning is conducted with BsaI-v2 HF (NEB), T4 Ligase (NEB) following manufacturer’s instructions. Completed Golden Gate reactions were cleaned with Zymo Clean Concentrate kits and transformed into Lucigen E. cloni™ electrocompetent cells. Diversity was maintained at every step such that there are at least 30x successfully transformed colony forming units as determined by serial dilutions and plating an aliquot of liquid cultures. Completed library landing pad constructs are co-transfected with a BxBI expression construct (pCAG-NLS-Bxb1) into (TetBxB1BFP-iCasp-Blast Clone 12 HEK293T cells). This cell line has a genetically integrated tetracycline induction cassette, followed by a BxBI recombination site, and split rapalog inducible dimerizable Casp-9. Cell are maintained in D10 (DMEM, 10% w/v fetal bovine serum (FBS), 1% w/v sodium pyruvate, and 1% w/v penicillin/streptomycin). Two days after transfection, doxycycline (2 ug/ml, Sigma-Aldrich) is added to induce expression of our genes of interest (successful recombination) or the icasp9 selection system (no recombination). Successful recombination shifts the iCasp-9 out of frame, thus only cells that have undergone recombination survive, while those that haven’t will die from iCasp-9-induced apoptosis. One day after doxycycline induction, AP1903 (10 nM, MedChemExpress) is added to cause dimerization of Casp9 and selectively kill cells without successful recombination. One day after AP1903-Casp9 selection, media is changed back to D10 + Doxycyline (2 ug/ml, Sigma-Aldrich) for recovery. Two days after cells have recovered, cells are reseeded to enable normal cell growth. Once cells reach confluency, library cells are frozen in glycerol stocks in aliquots for assays.

Sequencing-based surface expression assay: To measure how inserted motifs disrupt channel expression, we measured surface expression of all variants. We thawed glycerol stocks of library cell lines into wells of a 6 well dish, swapped media the following day to D10, grew cells to confluency, split once to ensure maximum cell health, and swapped media for D10 + doxycycline (2 ug/ml, Sigma-Aldrich). Kir3.1 cannot homo-tetramerize and therefore requires a co-expressed Kir3.2 or Kir3.4 inward rectifier to surface express (*21*). For this reason, 48 hours prior to sorting Kir3.1 libraries, we transiently transfected the stable Kir3.1 insertion library cell line with 2 ug Kir3.2-P2A-miRFP670 and 6ul Turbofect per well of a 6 well plate. For all libraries except for Kv1.3, we detached cells with 1 ml Accutase (Sigma-Aldrich), spun down and washed three times with FACS buffer (2% FBS, 0.1% NaN_3, 1X PBS), incubated for 1-hour rocking at 4degC with a BV421 anti-flag antibody (BD Bioscience), washed twice with FACS buffers, filtered with cell strainer 5 ml tubes (Falcon), covered with aluminum foil, and kept on ice for transfer to the flow cytometry core. For Kv1.3, cells were detached and washed the same except after initial washing cells were brought up in FACS buffer with Agitoxin-2-Cys-TAMRA (5nM, Alomone), filtered with cell strainer 5 ml tubes, and brought to cell sorting facility on ice. Before sorting, 5% of cells were saved as a control sample for sequencing prior to sorting. All cells except for Kir3.1 were sorted into unlabeled and labeled (either BV421 or Agitoxin-Cys-TAMRA) populations based on $EGFP^{high}/label^{low}$ and $EGFP^{high}/label^{high}$, respectively. On a BD FACSAria II P69500132 cell sorter, EGFP fluorescence was excited with a 488 nm laser and recorded with a 525/50 nm bandpass filter and 505 nm long-pass filter. BV421 fluorescence was excited using a 405 nm laser and recorded with a 450/50 nm bandpass filter, TAMRA fluorescence was excited using a 561 nm laser and recorded with a 586/15 nm bandpass filter, and miRFP670 was excited with a 640 nm laser and recorded with 670/30 nm bandpass filter.

All cells (expect those expressing Kir3.1) were gated on forward scattering area and side scattering area to find whole cells, forward scattering width, and height to separate single cells, EGFP for cells that expressed variants without errors (our library generation results in single base pair deletions that will not have EGFP expression because deletions will shift EGFP out of frame (12)), and label for surface expressed cells. Kir3.1 library cells were gated on forward scattering area and side scattering area to find whole cells, forward scattering width and height to separate single cells, miRFP670 5 times to get varying levels of Kir3.2 co-expression, GFP for cells that expressed variants without errors, and label for surface expressed cells. For simplicity, we only report Kir3.1 enrichment for one level of Kir3.2 (Kir3.2 #4). The surface expression label gate boundaries were determined based on unlabeled cells from the same population because controls tend to have non-representative distributions. Examples of the gating strategy for each channel is depicted in **Figures 4-20 to 4-24**.

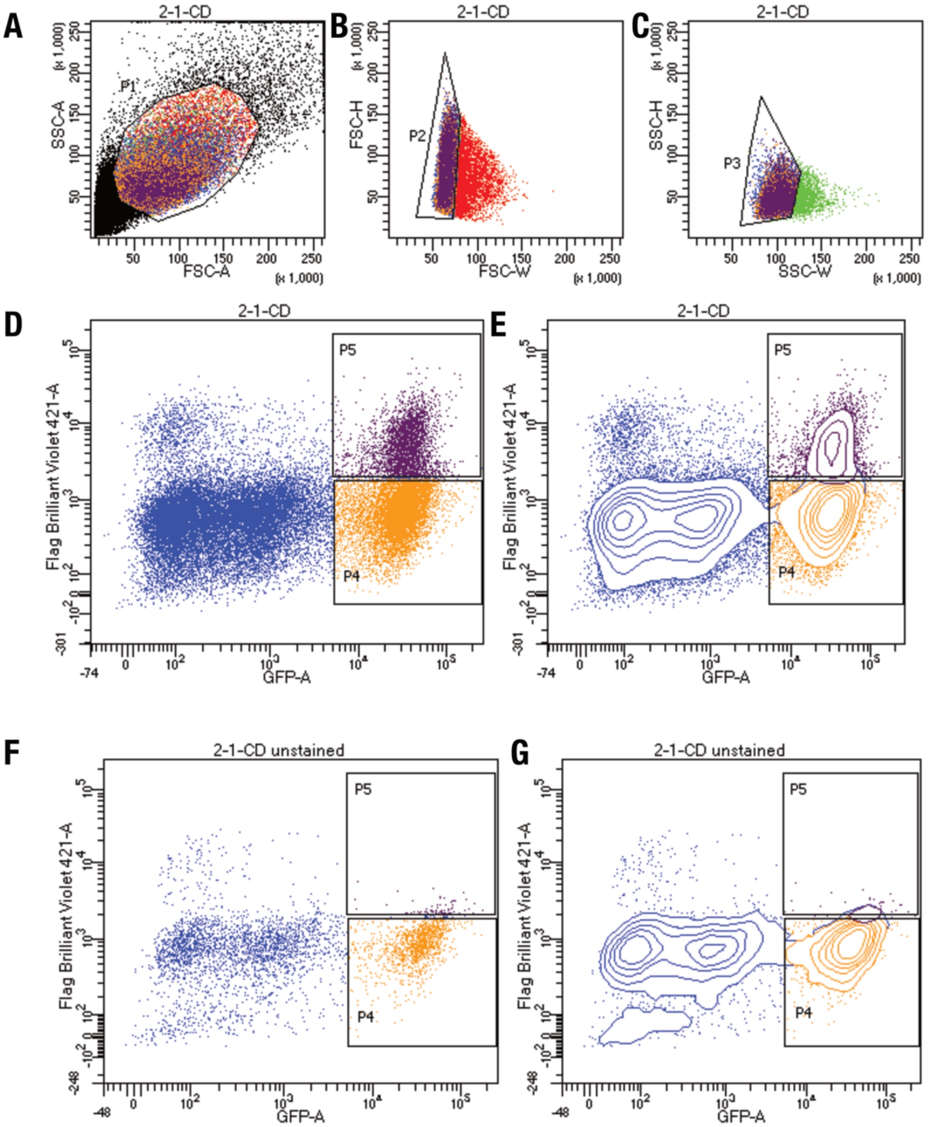


Figure 4-20: Kir2.1 surface expression assay gating scheme. (A) Whole HEK293 cells are gated on side (SSC-A) and forward scattering (FSC-A). **(B-C)** Forward scattering height (SSC-H), forward scattering width (FSC-W), and Side scattering width (SSC-W) are used to gate single cells. (**D-G**) EGFP^{high}/Label^{low} and EGFP^{high}/Label^{high} populations are

gated based **(D-E)** stained and (**F-G)** unstained on EGFP (GFP-A) of Anti-Flag Brilliant Violet-421 fluorescence with **(D,F)** scatterplot and **(E,G)** contour plots shown. Contour plots represent 95% confidence intervals with outliers shown as dots.

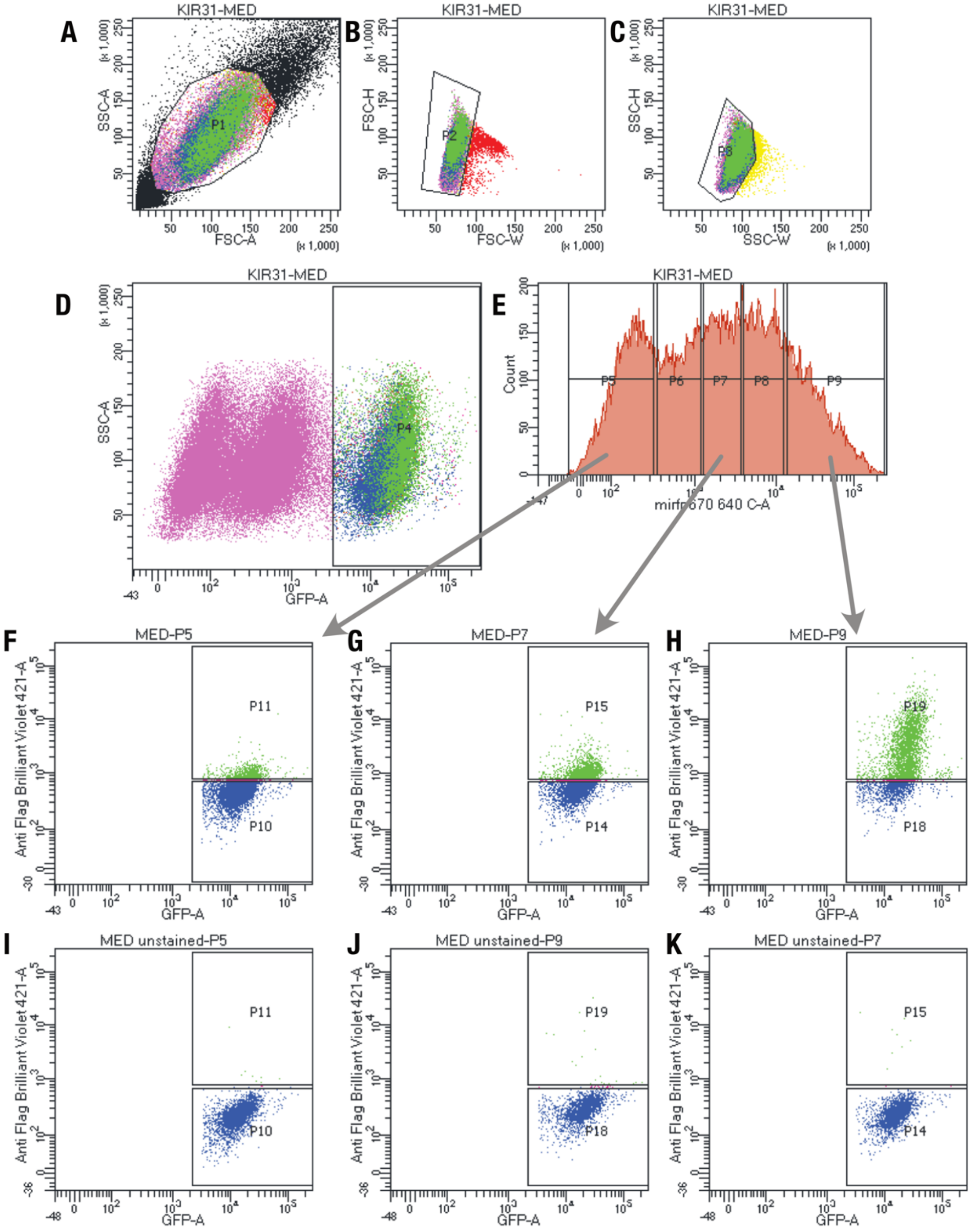


Figure 4-21: Kir3.1 surface expression assay gating scheme. (A) Whole HEK293 cells are gated on side (SSC-A) and forward scattering (FSC-A). **(B-C)** Forward scattering height (SSC-H), forward scattering width (FSC-W), and Side scattering width (SSC-W) are

used to gate single cells. (**D**) Cells are gated on EGFP positive cells to isolate successfully recombined libraries. **(E)** Cells are further split into 5 populations to separate out different populations of Kir3.2 co-expressed miRFP670. **(F-K)** EGFP^{high}/ Label^{low} and EGFP^{high}/Label^{high} populations are gated based **(F-H)** stained and (**I-K)** unstained on EGFP (GFP-A) of Anti-Flag Brilliant Violet-421 fluorescence. The data from 3 highest levels of miRFP670 were combined and reported as fitness.

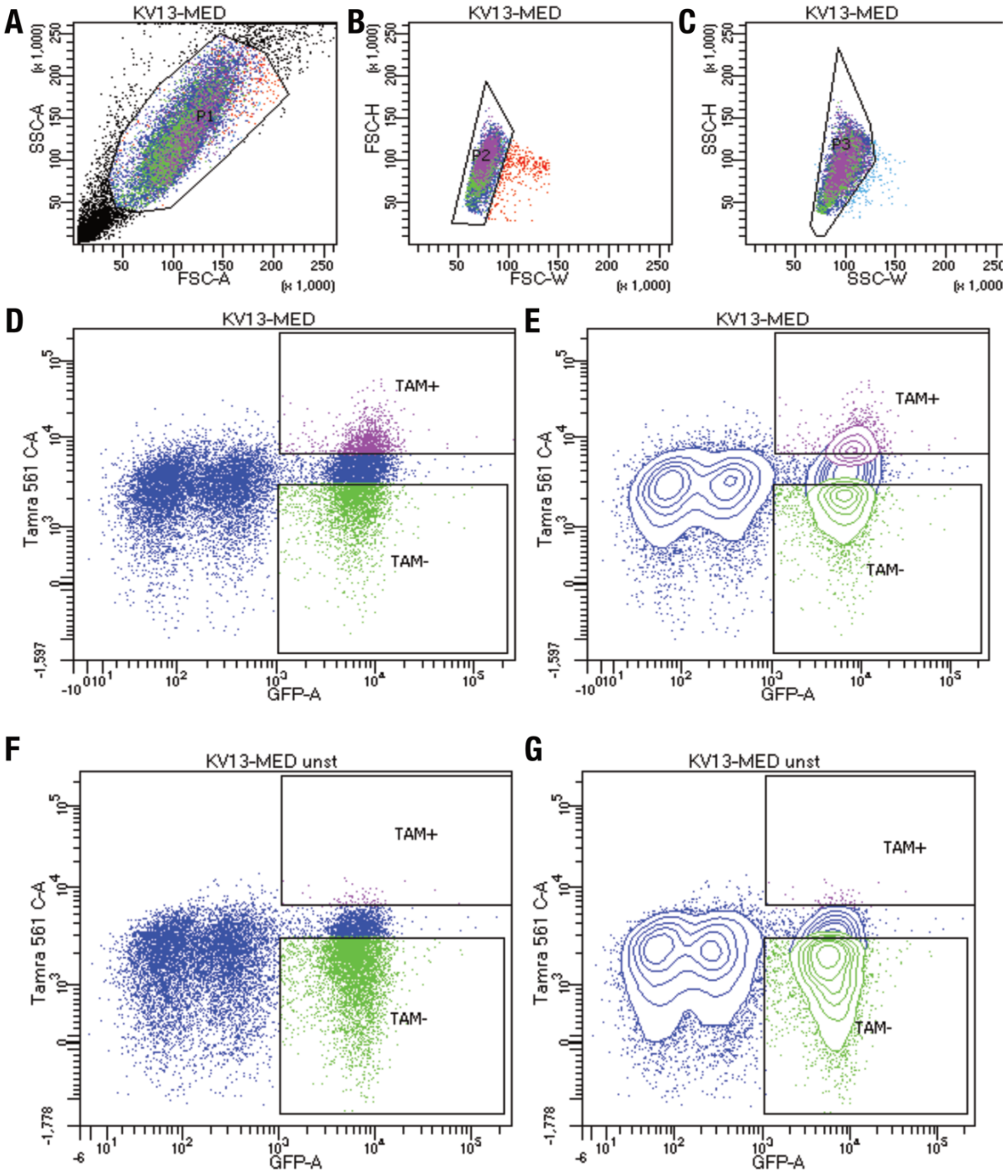


Figure 4-22: Kv1.3 Surface expression assay gating scheme. (A) Whole HEK293 cells are gated on side (SSC-A) and forward scattering (FSC-A). **(B-C)** Forward scattering

height (SSC-H), forward scattering width (FSC-W), and Side scattering width (SSC-W). (**D-G**) EGFP^{high}/ Label^{low} and EGFP^{high}/Label^{high} populations are gated based **(D-E)** stained and (**F-G)** unstained on EGFP (GFP-A) of Kv1.3 specific Agitoxin-Tamra fluorescence with **(D,F)** scatterplot and **(E,G)** contour plots shown. Contour plots represent 95% confidence intervals with outliers shown as dots.

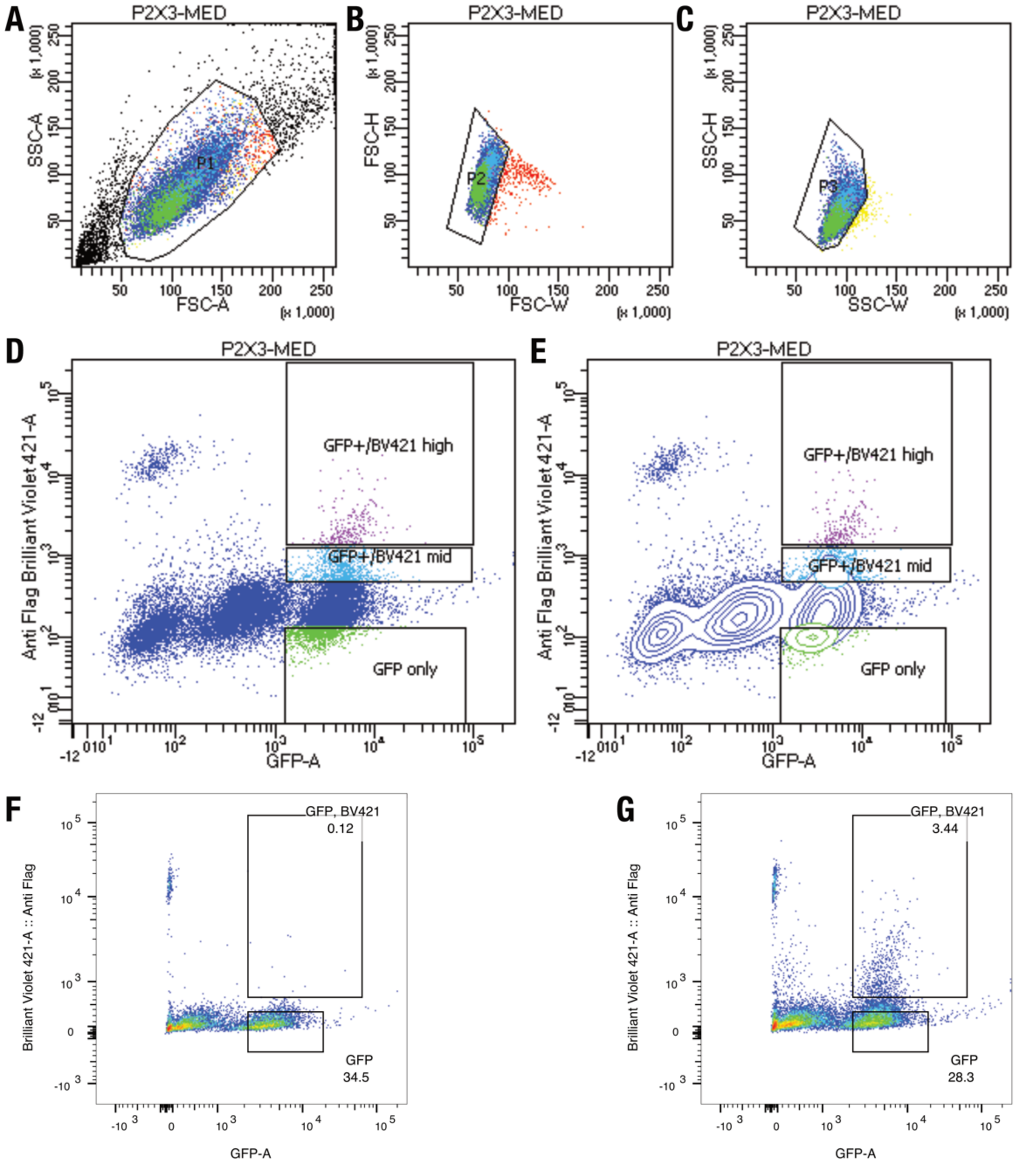


Figure 4-23: P2X3 Surface expression assay gating scheme. (A) Whole HEK293 cells are gated on side (SSC-A) and forward scattering (FSC-A). **(B-C)** Forward scattering height (SSC-H), forward scattering width (FSC-W), and Side scattering width (SSC-W) are used to gate single cells. (**D-E**) EGFP^{high}/ Label^{low}, EGFP ^{high}/Label^{low}, EGFP^{high}/Label^{med}

and EGFP^{high}/Label^{high} populations are gated based **(F)** stained and **(G)** unstained on EGFP (GFP-A) of Anti-Flag Brilliant Violet-421 fluorescence with **(D)** scatterplot, **(E)** Contour plot, and **(F-G)** pseudo color plots. In post sample collection *Mid* and *High* label populations were combined ratiometrically based on percent populations in corresponding gates. Contour plots represent 95% confidence intervals with outliers shown as dots. Pseudocolor plots represent density of points with a blue-to-red color scale with increasing density.

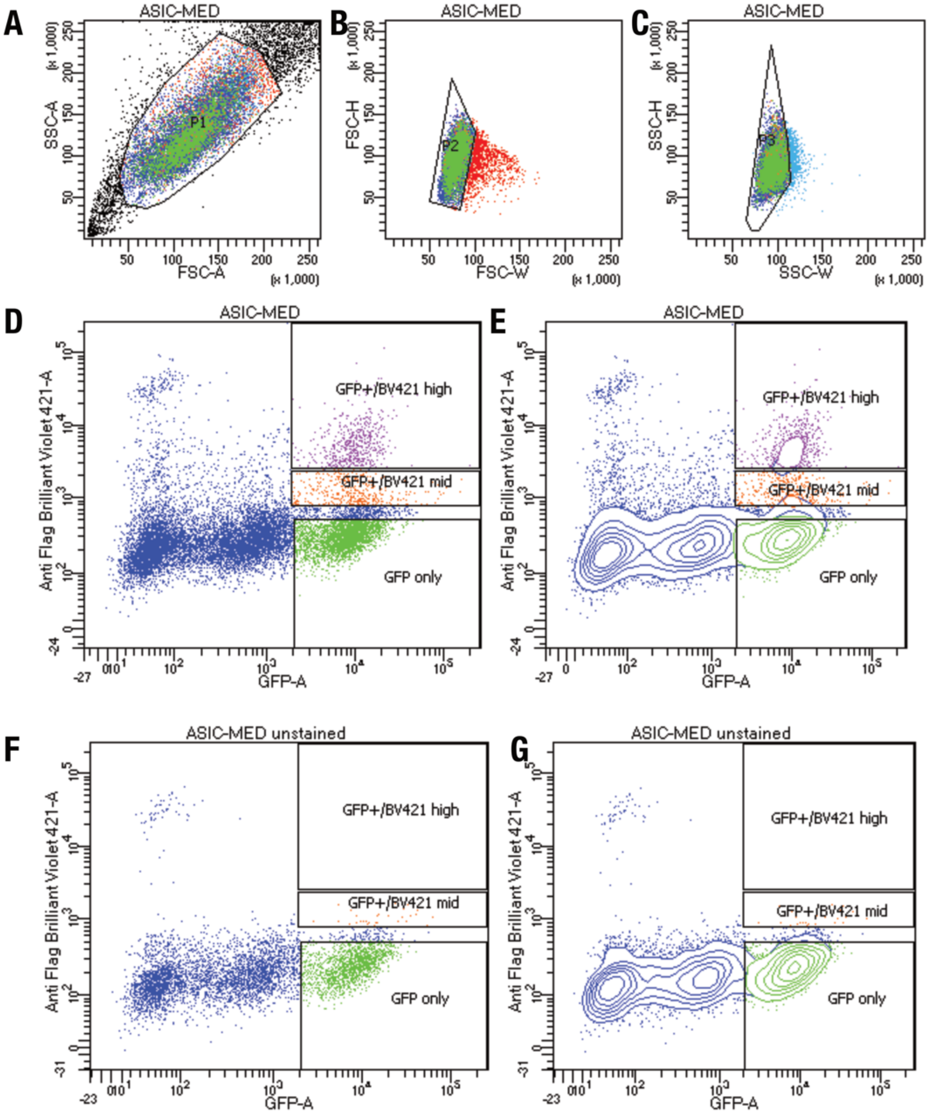


Figure 4-24: ASIC1a Surface expression assay gating scheme. (A) Whole HEK293 cells are gated on side (SSC-A) and forward scattering (FSC-A). **(B-C)** Forward scattering height (SSC-H), forward scattering width (FSC-W), and Side scattering width (SSC-W) are

used to gate single cells. (**D-G**) EGFP^{high}/ Label^{low} and EGFP^{high}/Label^{high} populations are gated based **(D-E)** stained and (**F-G)** unstained on EGFP (GFP-A) of Anti-Flag Brilliant Violet-421 fluorescence with **(D,F)** scatterplot and **(E,G)** contour plots shown. Contour plots represent 95% confidence intervals with outliers as shown as dots.

EGFP^{high}/label^{low} and EGFP^{high}/label^{high} cells were collected into catch buffer (20% FBS, 0.1% NaN_3, 1x PBS. For larger pooled sublibrary samples, we collected between at least 100,000 to 500,00 cells per gate which is 8-35x coverage. 15,000 cells in both gates of a Kir2.1 library with a small flexible ASGASGA linker was collected each day to normalize all the pooled libraries. For smaller 15 motifs samples, we collected between 4,000-50,000 of each sample/library pair which is ~10-120x coverage for all libraries. We find the more disruptive an insertion the more difficult it is to collect sufficient surface-labeled cells to reach 30x coverage. This means that our lower coverage is assuming all positions are represented in surface expressed cells.

Sequencing: DNA from pre-sort control and sorted cells were extracted with Microprep DNA kits (Zymo Research) and triple eluted with water. The elute was diluted such that no more than 1.5ug of DNA is used per PCR reaction and amplified for 20 cycles of PCR using Primestar GXL (Takara Clonetech), run on a 1% agarose gel, and gel purified. Primers that bind outside the recombination site ensure leftover plasmid DNA from the original cell line construction step is not amplified. Purified DNA was quantified using Picogreen DNA quantification. Equal amounts by mass of each domain insertion sample were pooled by cell sorting category and split into two domain sets per channel library set to segregate highly similar motifs sequences. Final amplicon pools were as follows: control, surface expression low 1, surface expression high 1, function low1, function high 1, surface expression low 2, surface expression high 2, function low 2, and function high 2. Pooled amplicons were prepared for sequencing using the Nextera XT sample preparation workflow, and sequenced using Illumina Novaseq in 2x150bp mode.

Enrichment Calculations: Forward and reverse reads were aligned individually using a DIP-seq pipeline (76), slightly modified for SPINE compatibility and for updated python packages. If both forward and reverse reads report an insertion, duplicated domain insertion calls are removed to avoid artificially boosting counts. This pipeline results in .csv spreadsheets indicating insertion position, direction, and whether it is in frame.

Surface expression enrichment was calculated by comparing the change in EGFP^{high}/label^{low} to EGFP^{high}/label^{high}. Enrichment calculation was based on Enrich2 software (*39*) and written in R. Only positions with reads in both label^{low} and label^{high} groups were used in enrichment calculations. For each cell group, the percentage of reads at each position was calculated after adding 0.5 to assist positions with very small counts. Enrichment was calculated by taking the natural logarithm of EGFP^{high}/label^{high} percentage divided by the EGFP^{high}/label^{low} percentage for each position (i).

$$Enrichement_i = \log \frac{0.5 + Count\_High_i}{\sum_i^n 0.5 + Count\_High_i} \Big/ \frac{0.5 + Count\_Low_i}{\sum_i^n 0.5 + Count\_Low_i}$$

All datasets were z-scored to an internal control flexible linker motif (AGSAGSA) enrichment (separate for each sequencing subpool) by subtracting the average medium enrichment and dividing by the standard deviation of the medium enrichment. Replicates (r) were combined by a weighted average, which was calculated by a restricted maximum likelihood estimate (M) and standard error (SE) using 50 Fisher scoring iterations.

$$Enrichment_i = \sum_r^n Enrichment_{i,r} * \frac{\sqrt{M_r + SE_r{}^2}}{\sum_r^n \sqrt{M_r + SE_r{}^2}}$$

Standard error was calculated assuming a Poisson distribution.

$$SE_i = \sqrt{\frac{1}{Count\_High_i + 0.5} + \frac{1}{Count\_Low_i + 0.5} + \frac{1}{\sum_i^n 0.5 + Count\_High_i} + \frac{1}{\sum_i^n 0.5 + Count\_Low_i}}$$

All other positions are treated as NA and are not considered in further analysis (exclusion criteria), except for correlations between datasets as removing data adds more noise than treating NAs as 0s due to sampling.

Data quality: Inserting 760 motifs into 432 Kir2.1 positions yields a total theoretical library diversity of 328,320 variants. Each sub-pooled library we generated and screened encompassed 12,500 variants. Due to random variance, some datasets were incomplete

(**Figure 4-1**). To make downstream analysis more robust, we only included motifs with data (after exclusion criteria outlined in *Enrichment Calculations)* in >80% of positions. This left us with 637 out of 760 motifs (further details in **Table 4-1**).

Clustering: All motif insertional profiling data was clustered by calculating a cosine distance matrix and clustering it with Ward's hierarchical clustering method using the hclust function in R with the 'ward.D2' method. Uniform Manifold Approximation Projection (UMAP)-based clustering was done using the uwot R package using cosine or Euclidean distance metrics, and a local neighborhood size of 10 sample points. Neighborhood size influences how UMAP balances local versus global structure in the data. Within a range of neighborhood sizes tested (2-50), our choice best conveys the broader structure of the data.

Ensemble Network Model: To calculate dynamics of the recipient and motifs with available PDBs, we used the Prody Python package (196). For this we used code from from Golinski et al. (197) as a starting point kindly provided by Alexander Golinski and Benjamin Hackel (University of Minnesota). We calculated mean stiffness of each backbone based on weighted sums of normal modes from an Anisotropic Network Model of vibration. We calculated summed recipient stiffness for varying lengths (1, 3, 5, 7, 9, 11 amino acids) before, centered on, and after an insertion position. Motif stiffness was summed for the entire motif and for varying lengths of the N- and C-termini (1, 2, 3, 4, 5, and 6 amino acids).

Molecular dynamics simulations: All-atom force-field based molecular dynamics simulations were carried out to sample multi-μs trajectories. Our structural models (agonist-bound PDB 3SPI and apo state PDB 3JYC) are constituted by the channel embedded in a bilayer of ~1300 POPC lipids hydrated by two slabs containing ~170,000 waters and ~600 KCl ion pairs, for a total of ~700,000 atoms. We first generated the coordinates of the missing amino acids in the experimental structures (mostly located in unstructured regions) using ROSETTA (for this purpose we generated 10,000 models and kept the representative structure of the most populated cluster). We then used charmm-gui (198) to model the bilayer and the aqueous compartment. Simulations are being performed with the charmm36 force field (199) at a temperature of T=303.15K, using the highly parallel computational code NAMD2.12 (200) on 280 processors cores from Temple

University's Owlsnest. Per residue root mean squared fluctuations (RMSF) were calculated using the R bio3D package (201) as root mean squared position fluctuation of each residue's C-alpha atom across each simulation.

Amino acid scoring: We calculated bioinformatic scores for amino acids using the Quantiprot python package (202). For scores we used: molecular weight, surface area, alpha helical propensity, beta sheet propensity, buried accessibility ratio propensity, flexibility, hydropathy, hydrophobicity, negative charge, pKa, polarity, positive charge, reverse turn propensity, and volume. These scores were calculated for both recipient and donors. We calculated summed recipient scores for varying lengths before, centered on, and after an insertion position (1, 3, 5, 7, 9, 11 amino acids). Motif sequence scores were summed for the entire motif and for varying lengths of the N and C termini (1, 2, 3, 4, 5, and 6 amino acids). Motif length was also included.

Protein Structural Properties: A series of properties were calculated with heavily modified code previously used to calculate properties of protein domains kindly provided by Alexander Golinski and Benjamin Hackel (197) that uses Pymol called from python scripts. Recipient protein PDBs were trimmed of any ions, water, and other none protein of interest molecules. Recipient protein phi, psi, contact degree, contact order, long contact degree, secondary structure percentage, alpha helical percentage, beta sheet percentage, nonpolar solvent accessible surface area (SASA), charged SASA, and hydrophobic SASA. For each of these properties, we summed recipient structural scores for varying lengths (1, 3, 5, 7, 9, 11 amino acids) before, centered on, and after an insertion position. For motifs with structures, the mean phi angle, mean psi angle, radius of gyration, distance between n and c termini, distance of N and C termini to center of mass, motif size in Daltons, mean contact degree, mean contact order, mean long contact degree, mean secondary structure percentage, mean alpha helical percentage, mean beta sheet percentage, mean nonpolar SASA, mean charged SASA, mean hydrophobic SASA, and RMSD if there were multiple conformers were calculated. In addition to mean motif structural properties, N- and C-terminal varying lengths (1, 2, 3, 4, 5, and 6 amino acids) sums were calculated for the phi angle, psi angle, contact degree, contact order, long contact degree, secondary structure percentage, alpha helical percentage, beta sheet percentage, nonpolar SASA, charged SASA, hydrophobic SASA, and RMSD.

Choosing features to train Random Forest: To allow for greater interpretability of our Random Forest-based models, we filtered the input features for redundancy. Our approach to reduce property redundancy was as follows: For motifs, we took the shortest and longest N- and C-terminal features as well as the mean motif features. We identified redundant motif properties by setting a +/- 0.8 correlation cutoff calculated between the motif property and permissibility across all motifs for a given site. We chose the most explanatory of highly correlated motif properties based on summed absolute correlative value across all positions. For recipient properties, we took the longest and shortest of each mean property before, centered and after the insertion position. We identified redundant recipient properties by setting a +/- 0.8 correlation cutoff calculated between the recipient property and permissibility across all positions for a given motif. We chose the most explanatory of highly correlated recipient properties based on summed absolute correlative value across all motifs. These steps reduced our recipient properties from 908 (520 recipient and 388 motif) properties down to 64 (32 recipient and 32 motif) properties.

Random Forests: Once we had a non-redundant set of 64 properties, we trained a preliminary random forest model with 500 trees (**Figure 4-7**). Based on this preliminary model, we further trimmed the properties down to the most explanatory 20 (12 recipient and 8 motif properties). We retrained the model without a significant drop in model performance (39.98% variance explained for 69 properties and 39.44% for 20 properties, **Table 4-4**). However, at this point we were including motif structural properties. This meant that we were not able to include any motifs without structural data. As only 1 of the top 10 most predictive properties ('Motif Phi Mean' as the 9th most predictive) were from the structured domain set, we decided to exclude structure-based motif features altogether. This allowed us to include more motifs and reduce our non-redundant properties set down further (39.44% variance explained for 20 properties and 38.69% for 10 properties, **Table 4-4**). We ended up choosing the top 10 most predictive features which included 6 recipient features (stiffness, phi angle of 11 AA centered around insertion site, MD simulation RMSF, contact degree at insertion site, polar surface area of 11 AA preceding insertion site, beta sheet content in 11 AA preceding insertion site) and 4 motif features (mean hydrophobicity, motif length, mean negative charge, mean amino acid volume of 7 N-terminal residues). This final model was trained using 85% of the data, with the other 15% withheld for testing, and performed well on the test dataset (**Figure 4-8**). All random forests were trained using the Randomforest package in R with 500 trees and localimp = ‘TRUE’ with all model parameters set to default values.

Supplemental Note 1: Detailed rules for protein recombination from machine learning.

Properties that guide recombination: Random forest models allow us to study how a set of properties interact non-linearly to give rise to a phenotype. We trained a random forest model on a set of recipient and motif properties to learn what determines productive protein motif insertions into our recipient protein Kir2.1. We calculate feature importance for every property by looking at how model performance is impacted when a given property is not included in the model. We find the most important property overall is motif hydrophobicity, with recipient flexibility (stiffness and RMSF), motif length, and recipient space around an insertion site (contacts) close behind. The most important motif properties are the motifs length and hydrophobicity, and the most important recipient properties are contact degree and stiffness. However, based on feature importance alone, we do not know how properties relate to insertions.

We can further investigate how properties give rise to productive insertions through accumulated local effects (ALE) plots (**Figure 4-9B-E, Figure 4-10 7**). These plots summarize the local effects of a property on the model’s prediction. For example, flexibility appears to have switch-like interactions whereby, below a threshold rigidity, it is quite deleterious (**Figure 4-9E, Figure 4-10C** Positive relationship in RMSF and **Figure 4-10G** negative relationship in stiffness). Other recipient properties also have straightforward positive or negative relationships such as polar solvent accessible surface area (SASA) (negative, **Figure 4-10J**), beta sheet % (positive, **Figure 4-10F**), and Phi angle (positive, **Figure 4-10D**). Contact degree on the other hand has a nonlinear and non-monotonic interaction suggesting this recipient property is more complex (**Figure 4-9B**, **Figure 4-10E**). Overall, recipient features appear to determine insertional fitness in relatively simple ways, such as flexibility and beta sheets 11 amino acids prior to an insertion position are positive, which likely means flexible loops are desirable insertion positions. This result is in line with previous insertion strategies (*20*).

In contrast, all motif properties have more complex relationships to insertional fitness. For example, lower motifs hydrophobicity appears to be deleterious (1.8-2.5) then becomes beneficial at higher values. Similarly, motif length is negative until it becomes beneficial in the model at about 25 amino acids. This is true for the other motif features as well: motif negativity (**Figure 4-10B**) is initially negative (albeit noisy) then becomes positive. N-terminal 7 amino acid volume (**Figure 4-10I**) that is initially positive, becomes negative, and returns to be positive. Overall, this suggests motif properties have more complex relationships to insertional fitness. Motif properties are beneficial in some contexts and deleterious in others.

Taken together, recipient properties behave as expected in which flexible loops appear to be beneficial. In contrast to existing approaches to engineer synthetic fusion protein (e.g., (*20*)) that consider inserts to be interchangeable and solely focus on the properties of insertion positions, we propose that inclusion of motifs properties and their interactions is crucial to understand whether an insertions is viable at a given insertion position.

Interactions between properties: Random forests are comprised of many decision trees built from random subsets of features that in aggregate predict a desired outcome from properties. Decision trees make predictions by splitting a dataset at property thresholds set on each input feature. Thresholds on multiple input features enable decision trees, and by extension forests, to capture non-linear interactions between properties if they are predictive of the class being modeled. These non-linear interactions are why a property such as motif length can be positive and negative in different contexts.

To interpret why motif properties and contact density behave non-linearly, we explored their interactions (**Figure 4-9F-J, Figure 4-11 to 4-13**). We find that motif hydrophobicity and length interact substantially more than all other properties with recipient contact degree and stiffness the next highest interactions (**Figure 4-9F**). Motif properties having more interactions than recipient properties makes sense in light of our earlier observation that motif properties are more likely to non-linearly impact insertional fitness. However, just looking at overall interactions' strength does not tell us which features interact.

To identify which properties are interacting with which, we calculated pairwise interaction strengths between all properties (**Figure 4-9G**). The strongest interactions overall in-order of strength are pairwise interactions between motif hydrophobicity with motif length, negativity, and stiffness. Overall, there are many pairwise interactions between all the motif features and limited interactions between motif and recipient properties. For recipient properties there are no strong interactions between recipient properties and very few interactions with motif features. That said, recipient stiffness interacts with motif hydrophobicity and length. There are also moderate pairwise interactions between recipient contact degree with hydrophobicity and motif length. Overall, this means that motif properties interact with each other to determine how a motif behaves when inserted into a position and secondarily with recipient properties to determine whether a motif feature set is beneficial.

To learn which interactions are driving insertional fitness, we calculated and plotted pairwise ALE. It is important to note that pairwise ALE only represents the interaction that contributes to insertional fitness and does not consider how either property contributes alone.

When looking at the strongest interaction overall, motif hydrophobicity and recipient stiffness it is apparent that very high hydrophobicity is extremely deleterious within very flexible regions, low hydrophobicity is very beneficial in flexible regions, and high hydrophobicity is moderately beneficial in stiff (likely buried) regions (**Figure 4-9H**). Observing non-linear interactions help us build hypotheses of underlying biophysical mechanisms, such as hydrophobic residues when exposed and inserted into flexible surface exposed regions are extremely deleterious, whereas when these same motifs are inserted into buried likely more hydrophobic regions these become beneficial. In addition, interactions between motif length (> ~25 AA) and stiffness demonstrate a different trend, where long insertions into very flexible regions are deleterious (these are regions at the termini of the structure likely needed for folding and small flexible loops) and very rigid regions are also deleterious for long motifs (**Figure 4-9I**). Whereas longer motifs are beneficial in intermediate flexibility regions which are regions within the structured C-terminal domains that move (e.g., flexible loops and the PIP_2 binding sites). By comparing these two pairwise ALEs (**Figure 4-9H** Stiffness-Hydrophobicity and **Figure 4-9I** Stiffness-Length), we can see that short non-hydrophobics are most preferred within very flexible regions, short hydrophobics are most preferred within very flexible regions, and longer partially hydrophobic motifs are preferred in semi-flexible regions. Furthermore, hydrophobicity is deleterious for short motifs, beneficial for longer motifs, and extremely deleterious for short motifs (**Figure 4-9J**). Perhaps in longer motifs, hydrophobic residues provide stabilization by virtue of well-formed hydrophobic cores, whereas shorter motifs lack well-formed hydrophobic cores and instead expose hydrophobic residues thus becoming very disruptive by promoting aggregation. Overall, this analysis points to motif hydrophobicity and length interacting to determine how a motif behaves within the context of a recipient property. These interactions give rise to the classes of motifs and regions, we observe in clustering **(Figure 4-2B, D).**

To further investigate what drove specific motif cluster behavior, we investigated calculated and annotated ALE plots based on where a motif class properties are located (**Figure 4-4C-E**).

Unstructured short cluster behavior: For the short unstructured motifs, non-hydrophobicity and length are important within unstructured regions because these regions prefer polar hydrophilic motifs as these will be solvent exposed (**Figure 4-12B-D**). These motifs however are not allowed well in buried regions based on high contacts being deleterious for small motifs (**Figure 4-12B, G**). In general negativity appears to play a weak negative role (**Figure 4-12E**). Finally, there is a strong beneficial interaction in regions with beta sheets in the 11 amino acids preceding – perhaps implying flexible loops (**Figure 4-12I, J**). Flexible motifs are overwhelmingly inserted within flexible loops or at the termini of beta sheets (**Figure 4-12A**). This class is primarily best allowed within flexible and non-buried regions. Motifs fall into this class if they are non-hydrophobic and small meaning they will be non-disruptive from the perspective of space (contact degree), flexibility (stiffness), and surface exposure (beta sheet %).

Hydrophobic motifs: For the hydrophobic motifs, it is quite clear that hydrophobicity drives the behavior of this class. The motif length is not as important because hydrophobic motifs range in size. Hydrophobic motifs mostly benefit from little negativity, which makes sense as many hydrophobic motifs are best allowed with small segments of the transmembrane M1 and negativity would be disruptive when interacting with lipids (**Figure 4-13F**). Hydrophobic motifs are very deleterious when inserted within very flexible regions and beneficial within rigid regions (**Figure 4-13B**). This combined with highly hydrophobic motifs being beneficial within high contact regions (**Figure 4-13G**) means hydrophobics are beneficial when inserted within buried regions. Hydrophobics are highly deleterious in and around beta sheets (**Figure 4-13I**). Overall, this means hydrophobics behave inversely to the unstructured short cluster. Hydrophobics are mostly deleterious but can be inserted in some buried and transmembrane regions where they will not be disruptive. That said, several recipient flexible loops can accept either motif class (βC-βD, βE-βG, βH-βI, βL-βM). Interestingly, the βD-βE loop and unstructured termini that strongly allows and prefers longer more structured motifs does not allow for most hydrophobic inserts, perhaps because hydrophobics would interact with the solvent to cause misfolding and aggregation.

Larger structured motifs: Larger more structured motifs contain nearly all folded proteins and are most interesting from an engineering perspective. This class is overwhelmingly determined by length, with hydrophobicity being intermediate and negativity only slightly higher than other groups. While the overall class does appear to be driven by length, length interacts strongly with hydrophobicity and weakly with negativity (**Figure 4-11D-E**). Hydrophobicity is positive for long motifs likely representing the ability to form a hydrophobic core and fold. This interaction becomes even more clear when focusing on a subset of motifs within this class that are commonly recombined domains and other well folded larger proteins (**Figure 4-11A,D**). There is a clear demarcation above which hydrophobicity is highly beneficial (**Figure 4-11D**), which is likely why folded proteins has such a tight band of hydrophobicity (**Figure 4-4G**). There is a similarly tight distribution of negativity and may be an impact, but it is not nearly as strong (**Figure 4-11E-F**). Large motifs in very flexible (and generally small loops) large insertions are deleterious but intermediate stiff regions are more amenable to larger insertions (**Figure 4-11C**). That space is a fundamental determinant for larger motifs is best illustrated by the interactions with contact degree, where low contact degree is beneficial for the largest motifs (**Figure 4-11H**). Insertions of long motifs appear very deleterious in beta sheet rich regions, which likely disrupt formation of the immunoglobulin-like C-terminal domain of Kir2.1 (**Figure 4-11J**). Overall, motif length and hydrophobicity strongly interact positively to give rise to increased insertional fitness likely through improving folding. Whether this is beneficial is dependent on where an insertion occurs. Regions with some flexibility and sufficient space are deleterious. However, if there is sufficient space (N -and C-termini and βD-βE loop) insertions are actually quite beneficial. To better design domains for recombination, it would be ideal to have stable domains that have sufficient size and hydrophobicity to be able to maintain their fold after recombination, otherwise their folding thermodynamics will likely be overruled by the recipient protein.

Chapter 4:

Summary

Ion channels as with many other proteins evolved through domain recombination by assembling transmembrane and sensory domains, for example cyclic nucleotide binding or voltage sensing domains are combined with channel pores to generate cyclic nucleotide or voltage sensitive channels (53). Similarly, protein engineers recombine domains to generate useful tool and therapeutics, such as Car-T cell cancer therapies (21) and genetically encoded calcium sensors (22). However, these approaches are challenging meaning many new tools either require years of optimization or never find common use (74). These challenges are because we lack mechanistic rules for assembling protein domains that both limit our understanding of a fundamental principle of protein evolution and engineering of new proteins.

We developed high throughput domain insertion pipelines to engineer light-switchable ion channels, study fundamentals of ion channel structure-function, and identify mechanistic rules for domain-recipient compatibility. To start, we used a simple potassium channel, Kir2.1, as a model protein with a limited set of 3 inserted domains (13). However, there was a tremendous amount of variability between inserted domains that couldn't be easily explained. To develop a mechanistic model, we needed to scale up to a larger set of motifs. Unfortunately, the molecular biology pipelines the field used at the time were too biased and incomplete for insertional scanning to be generalizable. We developed a new library generation pipeline, SPINE, that solved all the problems with the existing transposon approach (12).

With the scalability of SPINE, we scaled up massively to over 700 different inserted motifs and used machine learning to develop a mechanistic and interpretable model of protein compatibility (14). We discover that interactions between the recipient protein and inserted domains drive whether a given insertion is productive with space and flexibility at an insertion position being critical and the folding of the inserted motifs. We then further validated our findings by scanning 4 additional recipient ion channels. In every channel we find that data from insertional scanning experiments reveals distinct roles for regions of the protein in folding and function. This finding implies that channels (and perhaps all proteins) balance the stability needed for folding with the dynamics needed for function. From a technology perspective, this means insertional scanning can be used as a coarse-grained experimental method for structural biology to identify regions of a protein that may be involved in function.

Impact and future directions

Most domain-based protein engineering efforts are trial and error because we lack mechanistic frameworks to explain how to assemble protein domains. Prior to our work, no one had compared how different domains are differentially compatible with a recipient protein. Most engineers have treated domains as interchangeable, which has led to many failed tools and wasted years of optimization (184). Our work is by far the largest study of it's kind with previous studies only including at most 2 different inserts. By exhaustively testing many different insertions, we demonstrate the importance of the properties of both the recipient protein and inserted motif. The guidelines from our machine learning models can be immediately useful to other protein engineers in their efforts. We've laid a foundation from which a universal protein computability framework can be built. By expanding outwards to additional backbones including diverse cytosolic proteins such as enzymes and kinases in insertional scans we could train and test universal machine learning models to allow anyone to engineer their favorite protein more easily.

There are many unanswered questions in this field that could and should be addressed through similar approaches. For example, many protein engineers include linkers during optimization, but this is largely a random and iterative practice. Perhaps generating massive domain-linker libraries space could provide rules to speed up engineering. Other challenging problems include identifying positions of a protein that best allow for engineering allosteric control or developing reporters. There are numerous other areas of protein engineering that can and will benefit from similar exhaustive approaches. Our approach can be broadly used as a approach for improving protein engineering by generating massive libraries and developing mechanistic rulesets. We provide a high-level example of a truly massive library-based approach (over 300k) that can be readily done in mammalian cell-based experiments.

In our initial study, we engineered a light-switchable potassium channel (13). This channel through further optimization could allow user-programmable control of the resting membrane potential of cells. This could allow others to explore how resting membrane potential controls various physiological and disease states in a living animal. Furthermore, by finding small molecule-switchable new chemical-genetic therapeutics could be developed that could be used to treat seizures and other excitability associated diseases. In addition to protein engineering, we find insertional scanning can be used as a coarse-grained structural biology method for studying the intrinsic biophysics of ion channel architectures. We find in each ion channel that insertional scanning identifies regions that are sensitive but allow some insertion which are involved in regulation and function whereas other regions are rigid and must be maintained for folding. This implies a hierarchical organization of channels that balances contradictory needs of stable regions that can fold and dynamic regions that can function. Perhaps this is universal to all proteins. With insertional scanning we could potential explore fundamental principles of protein evolution, organization, and function. Furthermore, this insertional scanning could be a first pass method for identifying regions of a protein that are coupled to function. Perhaps from these approaches we can discover cryptic druggable pockets and hidden conformational states. Insertional scanning and other high throughput perturbation assays hold the potential to open up whole new types of biological experiments. Where we can understand how each region of a protein contributes to function and perturb conformational and energetic landscapes of a protein to study fundamentals of protein dynamics.

From our screens, we identified regions of inward rectifying potassium channels that are sensitive to insertion and are the binding site for allosteric regulators in close paralogs. We followed this up by inserting light-switchable domains at several of these sites and controlling activity. This means that there is regulatory potential at these sites and perhaps these latent regulatory potentials was harnessed in regulation to connect different components of a signaling network. Inward rectifying channel's DE loop is a Gprotein binding site in G-protein protein coupled inwardly rectifying potassium channels (GIRK). This DE loop could have intrinsic regulatory potential that was harnessed in evolution to connect GPCRs to GIRKs to allow neurotransmitter control of a potassium channel. Through many such evolutionary events signaling networks could have evolved. Further studies should be done further exploring how gprotein regulation evolved in inward rectifiers. These findings could illuminate insights in a fundamental questions in biology, how does a regulatory site evolve to allow the complex cell signaling networks that underlie multi-cellular life.

REFERENCES

1. Houdusse A, Sweeney HL. How Myosin Generates Force on Actin Filaments. Trends in Biochemical Sciences. 2016 Dec;41(12):989–97.

2. Palczewski K. G Protein–Coupled Receptor Rhodopsin. Annu Rev Biochem. 2006 Jun;75(1):743–67.

3. Isacoff EY, Jan LY, Minor DL. Conduits of Life's Spark: A Perspective on Ion Channel Research since the Birth of Neuron. Neuron. 2013 Oct;80(3):658–74.

4. Chen K, Arnold FH. Engineering new catalytic activities in enzymes. Nat Catal. 2020 Mar;3(3):203–13.

5. Vallée-Bélisle A, Plaxco KW. Structure-switching biosensors: inspired by Nature. Current Opinion in Structural Biology. 2010 Aug;20(4):518–26.

6. Tobin P, Richards D, Callender R, Wilson C. Protein Engineering: A New Frontier for Biological Therapeutics. CDM. 2015 Jan 26;15(7):743–56.

7. Elowitz M, Lim WA. Build life to understand it. Nature. 2010 Dec;468(7326):889–90.

8. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol. 2007 Dec;8(12):995–1005.

9. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016 Jun;17(6):333–51.

10. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014 Aug;11(8):801–7.

11. Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. Science. 2017 Jul 14;357(6347):168–75.

12. Coyote-Maestas W, Nedrud D, Okorafor S, He Y, Schmidt D. Targeted insertional mutagenesis libraries for deep domain insertion profiling. Nucleic Acids Research. 2020 Jan 24;48(2):e11–e11.

13. Coyote-Maestas W, He Y, Myers CL, Schmidt D. Domain insertion permissibility-guided engineering of allostery in ion channels. Nat Commun. 2019 Dec;10(1):290.

14. Coyote-Maestas W, Nedrud D, Antonio Suma, Matreyek KA, Fowler DM, Carnevale Vi, et al. The biophysical basis of protein domain compatibility in ion channels. Biorxiv. :69.

15. Nedrud D, Coyote-Maestas W, Schmidt D. A survey of pairwise epistasis supports an outside-in hierarchy of clade-specifying and function-defining residues in PSD95 PDZ3 [Internet]. Biochemistry; 2020 Jun [cited 2020 Aug 30]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.06.26.174375

16. Andersson DI, Jerlström-Hultqvist J, Näsvall J. Evolution of New Functions De Novo and from Preexisting Genes. Cold Spring Harb Perspect Biol. 2015 Jun;7(6):a017996.

17. Keefe AD, Szostak JW. Functional proteins from a random-sequence library. Nature. 2001 Apr 5;410(6829):715–8.

18. Taylor JS, Raes J. Duplication and Divergence: The Evolution of New Genes and Old Ideas. Annu Rev Genet. 2004 Dec;38(1):615–43.

19. Albery WJ, Knowles JR. Evolution of enzyme function and the development of catalytic efficiency. Biochemistry. 1976 Dec;15(25):5631–40.

20. Bhattacharyya RP, Reményi A, Yeh BJ, Lim WA. Domains, Motifs, and Scaffolds: The Role of Modular Interactions in the Evolution and Wiring of Cell Signaling Circuits. Annu Rev Biochem. 2006 Jun;75(1):655–80.

21. Almåsbak H, Aarvak T, Vemuri MC. CAR T Cell Therapy: A Game Changer in Cancer Treatment. Journal of Immunology Research. 2016;2016:1–10.

22. Dana H, Sun Y, Mohar B, Hulse BK, Kerlin AM, Hasseman JP, et al. High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. Nat Methods. 2019 Jul;16(7):649–57.

23. Ponting CP, Russell RR. The Natural History of Protein Domains. Annu Rev Biophys Biomol Struct. 2002 Jun;31(1):45–71.

24. Pan X, Thompson MC, Zhang Y, Liu L, Fraser JS, Kelly MJS, et al. Expanding the space of protein geometries by computational design of de novo fold families. 2020;6.

25. Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, et al. Computational design of ligand-binding proteins with high affinity and selectivity. Nature. 2013 Sep;501(7466):212–6.

26. Foight GW, Wang Z, Wei CT, Jr Greisen P, Warner KM, Cunningham-Bryant D, et al. Multi-input chemical control of protein dimerization for programming graded cellular responses. Nat Biotechnol. 2019 Oct;37(10):1209–16.

27. Lajoie MJ, Boyken SE, Salter AI, Bruffey J, Rajan A, Langan RA, et al. Designed protein logic to target cells with precise combinations of surface antigens. Science. 2020 Aug 20;eaba6527.

28. Vishwanath S, de Brevern AG, Srinivasan N. Same but not alike: Structure, flexibility and energetics of domains in multi-domain proteins are influenced by the presence of other domains. Jacobs D, editor. PLoS Comput Biol. 2018 Feb 12;14(2):e1006008.

29. Matreyek KA, Stephany JJ, Fowler DM. A platform for functional assessment of large variant libraries in mammalian cells. Nucleic Acids Research. 2017 Jun 20;45(11):e102–e102.

30. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet. 2018 Jun;50(6):874–82.

31. Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, Sander C, et al. Inferring protein 3D structure from deep mutation scans. Nat Genet. 2019 Jul;51(7):1170–6.

32. Braberg H, Echeverria I, Bohn S, Cimermancic P, Shiver A, Alexander R, et al. Genetic interaction mapping informs integrative structure determination of protein complexes. Science. 2020 Dec 11;370(6522):eaaz4910.

33. Leander M, Yuan Y, Meger A, Cui Q, Raman S. Functional plasticity and evolutionary adaptation of allosteric regulation. Proc Natl Acad Sci USA. 2020 Oct 13;117(41):25445–54.

34. Tack DS (Fed). The genotype-phenotype landscape of an allosteric protein. :28.

35. Kozek KA, Glazer AM, Ng C-A, Blackwell D, Egly CL, Vanags LR, et al. High-throughput discovery of trafficking-deficient variants in the cardiac potassium channel KV11.1. Heart Rhythm. 2020 Dec;17(12):2180–9.

36. Chao JT, Hollman R, Meyers WM, Meili F, Matreyek KA, Dean P, et al. A Premalignant Cell-Based Model for Functionalization and Classification of *PTEN* Variants. Cancer Res. 2020 Jul 1;80(13):2775–89.

37. Reeb J, Wirth T, Rost B. Variant effect predictions capture some aspects of deep mutational scanning experiments. BMC Bioinformatics. 2020 Dec;21(1):107.

38. Dunham A, Beltrao P. Exploring amino acid functions in a deep mutational landscape [Internet]. Systems Biology; 2020 May [cited 2020 Dec 15]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.05.26.116756

39. Shah NH, Kuriyan J. Understanding molecular mechanisms in cell signaling through natural and artificial sequence variation. Nat Struct Mol Biol. 2019 Jan;26(1):25–34.

40. Azeloglu EU, Iyengar R. Signaling Networks: Information Flow, Computation, and Decision Making. Cold Spring Harb Perspect Biol. 2015 Apr;7(4):a005934.

41. Stein V, Alexandrov K. Synthetic protein switches: design principles and applications. Trends in Biotechnology. 2015 Feb;33(2):101–10.

42. Kennis JTM, van Stokkum IHM, Crosson S, Gauden M, Moffat K, van Grondelle R. The LOV2 Domain of Phototropin: A Reversible Photochromic Switch. J Am Chem Soc. 2004 Apr;126(14):4512–3.

43. Farrants H, Tarnawski M, Müller TG, Otsuka S, Hiblot J, Koch B, et al. Chemogenetic Control of Nanobodies. Nat Methods. 2020 Mar;17(3):279–82.

44. Dagliyan O, Shirvanyants D, Karginov AV, Ding F, Fee L, Chandrasekaran SN, et al. Rational design of a ligand-controlled protein conformational switch. Proceedings of the National Academy of Sciences. 2013 Apr 23;110(17):6800–4.

45. Kullmann DM. The neuronal channelopathies. Brain. 2002 Jun;125(6):1177–95.

46. Bernard G, Shevell MI. Channelopathies: A Review. Pediatric Neurology. 2008 Feb;38(2):73–85.

47. Fernández-Falgueras A, Sarquella-Brugada G, Brugada J, Brugada R, Campuzano O. Cardiac Channelopathies and Sudden Death: Recent Clinical and Genetic Advances. Biology. 2017 Jan 29;6(4):7.

48. Litan A, Langhans SA. Cancer as a channelopathy: ion channels and pumps in tumor development and progression. Front Cell Neurosci [Internet]. 2015 Mar 17 [cited 2020 Dec 15];9. Available from: http://journal.frontiersin.org/Article/10.3389/fncel.2015.00086/abstract

49. Boyden ES, Zhang F, Bamberg E, Nagel G, Deisseroth K. Millisecond-timescale, genetically targeted optical control of neural activity. Nat Neurosci. 2005 Sep;8(9):1263–8.

50. Fenno L, Yizhar O, Deisseroth K. The Development and Application of Optogenetics. Annu Rev Neurosci. 2011 Jul 21;34(1):389–412.

51. Govorunova EG, Sineshchekov OA, Li H, Spudich JL. Microbial Rhodopsins: Diversity, Mechanisms, and Optogenetic Applications. Annu Rev Biochem. 2017 Jun 20;86(1):845–72.

52. Nedrud D, He Y, Schmidt D. Efficient mesoscale phenotypic screening in cultured primary neuron culture [Internet]. Neuroscience; 2019 Feb [cited 2020 Dec 15]. Available from: http://biorxiv.org/lookup/doi/10.1101/562298

53. Nayak S, Batalov S, Jegla T, Zmasek C. Evolution of the Human Ion Channel Set. CCHTS. 2009 Jan 1;12(1):2–23.

54. Möglich A, Yang X, Ayers RA, Moffat K. Structure and Function of Plant Photoreceptors. Annu Rev Plant Biol. 2010 Jun 2;61(1):21–47.

55. Darwin C. On the Origin of Species by Means of Natural Selection. 1869.

56. Harding SD, Sharman JL, Faccenda E, Southan C, Pawson AJ, Ireland S, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. Nucleic Acids Research. 2018 Jan 4;46(D1):D1091–106.

57. Miesenbock G. The Optogenetic Catechism. Science. 2009 Oct 16;326(5951):395–9.

58. Urban DJ, Roth BL. DREADDs (Designer Receptors Exclusively Activated by Designer Drugs): Chemogenetic Tools with Therapeutic Utility. Annu Rev Pharmacol Toxicol. 2015 Jan 6;55(1):399–417.

59. Perutz M. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. Nature; 1960.

60. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 2005 Mar;6(3):197–208.

61. Cooper A, Dryden DTF. Allostery without conformational change: A plausible model. Eur Biophys J. 1984 Oct;11(2):103–9.

62. Popovych N, Sun S, Ebright RH, Kalodimos CG. Dynamically driven protein allostery. Nat Struct Mol Biol. 2006 Sep;13(9):831–8.

63. Motlagh HN, Hilser VJ. Agonism/antagonism switching in allosteric ensembles. Proceedings of the National Academy of Sciences. 2012 Mar 13;109(11):4134–9.

64. Hilser VJ, Wrabl JO, Motlagh HN. Structural and Energetic Basis of Allostery. Annu Rev Biophys. 2012 Jun 9;41(1):585–609.

65. Hilser VJ, Thompson EB. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. Proceedings of the National Academy of Sciences. 2007 May 15;104(20):8311–5.

66. Choi JH, Laurent AH, Hilser VJ, Ostermeier M. Design of protein switches based on an ensemble model of allostery. Nat Commun. 2015 Nov;6(1):6968.

67. Coyle SM, Flores J, Lim WA. Exploitation of Latent Allostery Enables the Evolution of New Modes of MAP Kinase Regulation. Cell. 2013 Aug;154(4):875–87.

68. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. J R Soc Interface. 2014 Nov 6;11(100):20140419.

69. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc R Soc Lond B. 1979 Sep 21;205(1161):581–98.

70. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nat Biotechnol. 2009 Feb;27(2):157–67.

71. Guntas G, Mansell TJ, Kim JR, Ostermeier M. Directed evolution of protein switches and their application to the creation of ligand-binding proteins. Proceedings of the National Academy of Sciences. 2005 Aug 9;102(32):11224–9.

72. Edwards WR, Busse K, Allemann RK, Jones DD. Linking the functions of unrelated proteins using a novel directed evolution domain insertion method. Nucleic Acids Research. 2008 Aug;36(13):e78–e78.

73. Siegel MS, Isacoff EY. A Genetically Encoded Optical Probe of Membrane Voltage. Neuron. 1997 Oct 1;19(4):735–41.

74. Cosentino C, Alberio L, Gazzarrini S, Aquila M, Romano E, Cermenati S, et al. Engineering of a light-gated potassium channel. Science. 2015 May 8;348(6235):707–10.

75. Chen T-W, Wardill TJ, Sun Y, Pulver SR, Renninger SL, Baohan A, et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature. 2013 Jul;499(7458):295–300.

76. Nadler DC, Morgan S-A, Flamholz A, Kortright KE, Savage DF. Rapid construction of metabolite biosensors using domain-insertion profiling. Nat Commun. 2016 Nov;7(1):12266.

77. Tucker CL, Fields S. A yeast sensor of ligand binding. Nat Biotechnol. 2001 Nov;19(11):1042–6.

78. Lockless SW. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. Science. 1999 Oct 8;286(5438):295–9.

79. Lee J, Natarajan M, Nashine VC, Socolich M, Vo T, Russ WP, et al. Surface Sites for Engineering Allosteric Control in Proteins. Science. 2008 Oct 17;322(5900):438–42.

80. Oakes BL, Nadler DC, Flamholz A, Fellmann C, Staahl BT, Doudna JA, et al. Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. Nat Biotechnol. 2016 Jun;34(6):646–51.

81. Bendahhou S, Donaldson MR, Plaster NM, Tristani-Firouzi M, Fu Y-H, Ptácek LJ. Defective Potassium Channel Kir2.1 Trafficking Underlies Andersen-Tawil Syndrome. J Biol Chem. 2003 Dec 19;278(51):51779–85.

82. Ma D, Taneja TK, Hagen BM, Kim B-Y, Ortega B, Lederer WJ, et al. Golgi Export of the Kir2.1 Channel Is Driven by a Trafficking Signal Located within Its Tertiary Structure. Cell. 2011 Jun;145(7):1102–15.

83. Hibino H, Inanobe A, Furutani K, Murakami S, Findlay I, Kurachi Y. Inwardly Rectifying Potassium Channels: Their Structure, Function, and Physiological Roles. Physiological Reviews. 2010 Jan;90(1):291–366.

84. Guo D, Ramu Y, Klem AM, Lu Z. Mechanism of Rectification in Inward-rectifier K+ Channels. Journal of General Physiology. 2003 Mar 17;121(4):261–76.

85. Kubo Y, Murata Y. Control of rectification and permeation by two distinct sites after the second transmembrane region in Kir2.1 K $^{+}$ channel. The Journal of Physiology. 2001 Mar;531(3):645–60.

86. Nishida M, Cadene M, Chait BT, MacKinnon R. Crystal structure of a Kir3.1-prokaryotic Kir channel chimera. The EMBO Journal. 2007 Sep 5;26(17):4005–15.

87. Hansen SB, Tao X, MacKinnon R. Structural basis of PIP2 activation of the classical inward rectifier K+ channel Kir2.2. Nature. 2011 Sep;477(7365):495–8.

88. Whorton MR, MacKinnon R. X-ray structure of the mammalian GIRK2–βγ G-protein complex. Nature. 2013 Jun;498(7453):190–7.

89. Martin GM, Yoshioka C, Rex EA, Fay JF, Xie Q, Whorton MR, et al. Cryo-EM structure of the ATP-sensitive potassium channel illuminates mechanisms of assembly and gating. eLife. 2017 Jan 16;6:e24149.

90. Savilahti H, Rice PA, Mizuuchi K. The phage Mu transpososome core: DNA requirements for assembly and function. The EMBO Journal. 1995 Oct;14(19):4893–903.

91. Harris BZ, Lim WA, Harris BZ, Lim WA. Mechanism and role of PDZ domains in signaling complex assembly. :13.

92. Taslimi A, Vrana JD, Chen D, Borinskaya S, Mayer BJ, Kennedy MJ, et al. An optimized optogenetic clustering tool for probing protein interaction and function. Nat Commun. 2014 Dec;5(1):4925.

93. Pegan S, Arrabit C, Slesinger PA, Choe S. Andersen's Syndrome Mutation Effects on the Structure and Assembly of the Cytoplasmic Domains of Kir2.1 †,‡. Biochemistry. 2006 Jul;45(28):8599–606.

94. Tinker A, Jan YN, Jan LY. Regions Responsible for the Assembly of Inwardly Rectifying Potassium Channels. Cell. 1996 Nov;87(5):857–68.

95. Stockklausner C, Ludwig J, Ruppersberg JP, Klöcker N. A sequence motif responsible for ER export and surface expression of Kir2.0 inward rectifier K $^{+}$ channels. FEBS Letters. 2001 Mar 30;493(2–3):129–33.

96. Collins A, Chuang H -h., Jan YN, Jan LY. Scanning mutagenesis of the putative transmembrane segments of Kir2.1, an inward rectifier potassium channel. Proceedings of the National Academy of Sciences. 1997 May 13;94(10):5456–60.

97. Lee S-J, Wang S, Borschel W, Heyman S, Gyore J, Nichols CG. Secondary anionic phospholipid binding site and gating mechanism in Kir2.1 inward rectifier channels. Nat Commun. 2013 Dec;4(1):2786.

98. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. Insertion site preference of Mu, Tn5, and Tn7 transposons. Mobile DNA. 2012;3(1):3.

99. Nehring RB, Wischmeyer E, Döring F, Veh RW, Sheng M, Karschin A. Neuronal Inwardly Rectifying K $^{+}$ Channels Differentially Couple to PDZ Proteins of the PSD-95/SAP90 Family. J Neurosci. 2000 Jan 1;20(1):156–62.

100. Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C. MISTIC: mutual information server to infer coevolution. Nucleic Acids Research. 2013 Jul 1;41(W1):W8–14.

101. Harrison SC, Durbin R. Is there a single pathway for the folding of a polypeptide chain? Proceedings of the National Academy of Sciences. 1985 Jun 1;82(12):4028–30.

102. Adams DS, Levin M. Measuring Resting Membrane Potential Using the Fluorescent Voltage Reporters DiBAC4(3) and CC2-DMPE. Cold Spring Harbor Protocols. 2012 Apr 1;2012(4):pdb.prot067702-pdb.prot067702.

103. Mase Y, Yokogawa M, Osawa M, Shimada I. Structural Basis for Modulation of Gating Property of G Protein-gated Inwardly Rectifying Potassium Ion Channel (GIRK) by i/o-family G Protein α Subunit (Gα $_{i/o}$). J Biol Chem. 2012 Jun 1;287(23):19537–49.

104. Bodhinathan K, Slesinger PA. Alcohol modulation of G-protein-gated inwardly rectifying potassium channels: from binding to therapeutics. Front Physiol [Internet]. 2014 [cited 2020 Dec 15];5. Available from: http://journal.frontiersin.org/article/10.3389/fphys.2014.00076/abstract

105. Rodriguez GJ, Yao R, Lichtarge O, Wensel TG. Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. Proc Natl Acad Sci USA. 2010 Apr 27;107(17):7787–92.

106. Kim JH, Lee S-R, Li L-H, Park H-J, Park J-H, Lee KY, et al. High Cleavage Efficiency of a 2A Peptide Derived from Porcine Teschovirus-1 in Human Cell Lines, Zebrafish and Mice. Thiel V, editor. PLoS ONE. 2011 Apr 29;6(4):e18556.

107. Engler C, Gruetzner R, Kandzia R, Marillonnet S. Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes. Peccoud J, editor. PLoS ONE. 2009 May 14;4(5):e5553.

108. Vriend G. WHAT IF: A molecular modeling and drug design program. Journal of Molecular Graphics. 1990 Mar;8(1):52–6.

109. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017 Feb;35(2):128–35.

110. Li H, Chang Y-Y, Yang L-W, Bahar I. *i* GNM 2.0: the Gaussian network model database for biomolecular structural dynamics. Nucleic Acids Res. 2016 Jan 4;44(D1):D415–22.

111. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes11Edited by G. von Heijne. Journal of Molecular Biology. 2001 Jul 6;310(2):311–25.

112. Batey S, Nickson AA, Clarke J. Studying the folding of multidomain proteins. HFSP J. 2008 Dec;2(6):365–77.

113. Richardson JS. The Anatomy and Taxonomy of Protein Structure. In: Anfinsen CB, Edsall JT, Richards FM, editors. Advances in Protein Chemistry [Internet].

Academic Press; 1981 [cited 2020 Dec 15]. p. 167–339. Available from: http://www.sciencedirect.com/science/article/pii/S0065323308605203

114. Scaiewicz A, Levitt M. The Language of the Protein Universe. Curr Opin Genet Dev. 2015 Dec;35:50–6.

115. Chothia C, Gough J. Genomic and structural aspects of protein evolution. Biochemical Journal. 2009 Apr 1;419(1):15–28.

116. Doolittle RF. The multiplicity of domains in proteins. Annu Rev Biochem. 1995 Jun 1;64(1):287–314.

117. Fallen K, Banerjee S, Sheehan J, Addison D, Lewis LM, Meiler J, et al. The Kir channel immunoglobulin domain is essential for Kir1.1 (ROMK) thermodynamic stability, trafficking and gating. Channels. 2009 Jan;3(1):57–68.

118. Lin MZ, Schnitzer MJ. Genetically encoded indicators of neuronal activity. Nature Neuroscience. 2016 Sep;19(9):1142–53.

119. VanEngelenburg SB, Palmer AE. Fluorescent biosensors of protein function. Current Opinion in Chemical Biology. 2008 Feb 1;12(1):60–5.

120. Gao XJ, Chong LS, Kim MS, Elowitz MB. Programmable protein circuits in living cells. Science. 2018 Sep 21;361(6408):1252–8.

121. Peisajovich SG, Garbarino JE, Wei P, Lim WA. Rapid Diversification of Cell Signaling Phenotypes by Modular Domain Recombination. Science. 2010 Apr 16;328(5976):368–72.

122. Schmiedel JM, Lehner B. Determining protein structures using deep mutagenesis. Nature Genetics. 2019 Jul;51(7):1177–86.

123. Bandaru P, Shah NH, Bhattacharyya M, Barton JP, Kondo Y, Cofsky JC, et al. Deconstruction of the Ras switching cycle through saturation mutagenesis. eLife. 2017 Jul 7;6:e27810.

124. Wright CM, Wright RC, Eshleman JR, Ostermeier M. A protein therapeutic modality founded on molecular regulation. Proceedings of the National Academy of Sciences. 2011 Sep 27;108(39):16206–11.

125. Judd J, Wei F, Nguyen PQ, Tartaglia LJ, Agbandje-McKenna M, Silberg JJ, et al. Random Insertion of mCherry Into VP3 Domain of Adeno-associated Virus Yields Fluorescent Capsids With no Loss of Infectivity. Mol Ther Nucleic Acids. 2012 Nov;1(11):e54.

126. Kolkman JA, Stemmer WPC. Directed evolution of proteins by exon shuffling. Nature Biotechnology. 2001 May;19(5):423–8.

127. Cherry JR, Lamsa MH, Schneider P, Vind J, Svendsen A, Jones A, et al. Directed evolution of a fungal peroxidase. Nature Biotechnology. 1999 Apr;17(4):379–84.

128. Luckow B, Renkawitz R, Schütz G. A new method for constructing linker scanning mutants. Nucleic Acids Research. 1987 Jan 26;15(2):417–29.

129. Guntas G, Mitchell SF, Ostermeier M. A Molecular Switch Created by In Vitro Recombination of Nonhomologous Genes. Chemistry & Biology. 2004 Nov 1;11(11):1483–7.

130. Ostermeier M. Designing switchable enzymes. Curr Opin Struct Biol. 2009 Aug;19(4):442–8.

131. Sheridan DL, Berlot CH, Robert A, Inglis FM, Jakobsdottir KB, Howe JR, et al. A new way to rapidly create functional, fluorescent fusion proteins: random insertion of GFP with an in vitro transposition reaction. BMC Neuroscience. 2002;11.

132. Mealer R, Butler H, Hughes T. Functional Fusion Proteins by Random Transposon-Based GFP Insertion. In: Methods in Cell Biology [Internet]. Academic Press; 2008 [cited 2020 Dec 15]. p. 23–44. (Fluorescent Proteins; vol. 85). Available from: http://www.sciencedirect.com/science/article/pii/S0091679X08850029

133. Shah V, Pierre B, Kim JR. Facile construction of a random protein domain insertion library using an engineered transposon. Analytical Biochemistry. 2013 Jan 15;432(2):97–102.

134. Osawa M, Erickson HP. Probing the domain structure of FtsZ by random truncation and insertion of GFP. Microbiology. 2005 Dec 1;151(12):4033–43.

135. Ason B, Reznikoff WS. DNA Sequence Bias During Tn5 Transposition. Journal of Molecular Biology. 2004 Jan 30;335(5):1213–25.

136. Giraldez T, Hughes TE, Sigworth FJ. Generation of Functional Fluorescent BK Channels by Random Insertion of GFP Variants. J Gen Physiol. 2005 Nov;126(5):429–38.

137. Gregory JA, Becker EC, Jung J, Tuwatananurak I, Pogliano K. Transposon Assisted Gene Insertion Technology (TAGIT): A Tool for Generating Fluorescent Fusion Proteins. Sandler SJ, editor. PLoS ONE. 2010 Jan 15;5(1):e8731.

138. Mehta MM, Liu S, Silberg JJ. A transposase strategy for creating libraries of circularly permuted proteins. Nucleic Acids Research. 2012 May 1;40(9):e71–e71.

139. Atkinson JT, Jones AM, Zhou Q, Silberg JJ. Circular permutation profiling by deep sequencing libraries created using transposon mutagenesis. Nucleic Acids Research. 2018 Jul 27;46(13):e76–e76.

140. Younger AKD, Su PY, Shepard AJ, Udani SV, Cybulski TR, Tyo KEJ, et al. Development of novel metabolite-responsive transcription factors via transposon-mediated protein fusion. Protein Engineering, Design and Selection. 2018 Feb 1;31(2):55–63.

141. Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF. Transposable Elements and Genome Organization: A Comprehensive Survey of Retrotransposons Revealed by the Complete *Saccharomyces cerevisiae* Genome Sequence. Genome Res. 1998 May 1;8(5):464–78.

142. Peters JE, Craig NL. Tn7 recognizes transposition target structures associated with DNA replication using the DNA-binding protein TnsE. Genes Dev. 2001 Mar 15;15(6):737–47.

143. Haapa-Paananen S, Rita H, Savilahti H. DNA Transposition of Bacteriophage Mu: A QUANTITATIVE ANALYSIS OF TARGET SITE SELECTION *IN VITRO*. J Biol Chem. 2002 Jan 25;277(4):2843–51.

144. Manna D, Deng S, Breier AM, Higgins NP. Bacteriophage Mu Targets the Trinucleotide Sequence CGG. JB. 2005 May 15;187(10):3586–8.

145. Mizuuchi M, Mizuuchi K. Target site selection in transposition of phage Mu. Cold Spring Harb Symp Quant Biol. 1993;58:515–23.

146. Allet B. Mu insertion duplicates a 5 base pair sequence at the host inserted site. Cell. 1979 Jan 1;16(1):123–9.

147. Sampson J, Jacobs K, Yeager M, Chanock S, Chatterjee N. Efficient study design for next generation sequencing. Genetic Epidemiology. 2011;35(4):269–77.

148. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics. 2014 Feb;15(2):121–32.

149. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016 Nov 1;5(4):221–32.

150. Kosuri S, Church GM. Large-scale de novo DNA synthesis: technologies and applications. Nature Methods. 2014 May;11(5):499–507.

151. LeProust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, et al. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic Acids Research. 2010 May;38(8):2522–40.

152. Engler C, Kandzia R, Marillonnet S. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. El-Shemy HA, editor. PLoS ONE. 2008 Nov 5;3(11):e3647.

153. Sugimoto N, Nakano S, Yoneyama M, Honda K. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. Nucleic Acids Res. 1996 Nov 15;24(22):4501–5.

154. SantaLucia J, Hicks D. The Thermodynamics of DNA Structural Motifs. Annual Review of Biophysics and Biomolecular Structure. 2004;33(1):415–40.

155. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera?A visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct;25(13):1605–12.

156. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics. 2017 Nov 15;33(22):3645–7.

157. Carr PA, Park JS, Lee Y-J, Yu T, Zhang S, Jacobson JM. Protein-mediated error correction for de novo DNA synthesis. Nucleic Acids Research. 2004 Oct 15;32(20):e162–e162.

158. Ellington A, Pollard JD. Introduction to the Synthesis and Purification of Oligonucleotides. Current Protocols in Nucleic Acid Chemistry. 2000;00(1):A.3C.1-A.3C.22.

159. Hecker KH, Rill RL. Error Analysis of Chemically Synthesized Polynucleotides. BioTechniques. 1998 Feb;24(2):256–60.

160. Kosuri S, Eroshenko N, LeProust EM, Super M, Way J, Li JB, et al. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. Nature Biotechnology. 2010 Dec;28(12):1295–9.

161. Plesa C, Sidore AM, Lubock NB, Zhang D, Kosuri S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. Science. 2018 Jan 19;359(6373):343–7.

162. Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. Massively parallel single-amino-acid mutagenesis. Nature Methods. 2015 Mar;12(3):203–6.

163. Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Research. 2014 Aug 18;42(14):e112–e112.

164. Chen X, Zaro JL, Shen W-C. Fusion protein linkers: Property, design and functionality. Advanced Drug Delivery Reviews. 2013 Oct 15;65(10):1357–69.

165. Long SB. Voltage Sensor of Kv1.2: Structural Basis of Electromechanical Coupling. Science. 2005 Aug 5;309(5736):903–8.

166. Goldhaber-Gordon I, Early MH, Baker TA. MuA Transposase Separates DNA Sequence Recognition from Catalysis †. Biochemistry. 2003 Dec;42(49):14633–42.

167. Quan J, Saaem I, Tang N, Ma S, Negre N, Gong H, et al. Parallel on-chip gene synthesis and application to optimization of protein expression. Nature Biotechnology. 2011 May;29(5):449–52.

168. McCandlish DM, Shah P, Plotkin JB. Epistasis and the Dynamics of Reversion in Molecular Evolution. Genetics. 2016 Jul;203(3):1335–51.

169. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. Epistasis as the primary factor in molecular evolution. Nature. 2012 Oct;490(7421):535–8.

170. Reynolds KA, McLaughlin RN, Ranganathan R. Hot Spots for Allosteric Regulation on Protein Surfaces. Cell. 2011 Dec 23;147(7):1564–75.

171. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, et al. Principles for designing ideal protein structures. Nature. 2012 Nov;491(7423):222–7.

172. Motlagh HN, Wrabl JO, Li J, Hilser VJ. The ensemble nature of allostery. Nature. 2014 Apr;508(7496):331–9.

173. Chothia C. Evolution of the Protein Repertoire. Science. 2003 Jun 13;300(5626):1701–3.

174. Lin C-Y, Liu JC. Modular protein domains: an engineering approach toward functional biomaterials. Current Opinion in Biotechnology. 2016 Aug;40:56–63.

175. Matreyek KA, Stephany JJ, Chiasson MA, Hasle N, Fowler DM. An improved platform for functional assessment of large protein libraries in mammalian cells. Nucleic Acids Research. 2019 Oct 15;gkz910.

176. Ma D, Zerangue N, Raab-Graham K, Fried SR, Jan YN, Jan LY. Diverse Trafficking Patterns Due to Multiple Traffic Motifs in G Protein-Activated Inwardly Rectifying Potassium Channels from Brain and Heart. Neuron. 2002 Feb 28;33(5):715–29.

177. Papazian DM. Potassium Channels: Some Assembly Required. Neuron. 1999 May 1;23(1):7–10.

178. Ma D. Role of ER Export Signals in Controlling Surface Potassium Channel Numbers. Science. 2001 Jan 12;291(5502):316–9.

179. Popot JL, Engelman DM. Membrane protein folding and oligomerization: the two-stage model. Biochemistry. 1990 May;29(17):4031–7.

180. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:180203426 [cs, stat] [Internet]. 2020 Sep 17 [cited 2020 Dec 15]; Available from: http://arxiv.org/abs/1802.03426

181. Xu C, Jackson SA. Machine learning and complex biological data. Genome Biol. 2019 Dec;20(1):76, s13059-019-1689–0.

182. Campioni S, Mannini B, Zampagni M, Pensalfini A, Parrini C, Evangelisti E, et al. A causative link between the structure of aberrant protein oligomers and their toxicity. Nature Chemical Biology. 2010 Feb;6(2):140–7.

183. Munson M, Balasubramanian S, Fleming KG, Nagi AD, O'Brien R, Sturtevant JM, et al. What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. Protein Sci. 1996 Aug;5(8):1584–93.

184. Dagliyan O, Tarnawski M, Chu P-H, Shirvanyants D, Schlichting I, Dokholyan NV, et al. Engineering extrinsic disorder to control protein activity in living cells. Science. 2016 Dec 16;354(6318):1441–4.

185. Zangerl-Plessl E-M, Lee S-J, Maksaev G, Bernsteiner H, Ren F, Yuan P, et al. Atomistic basis of opening and conduction in mammalian inward rectifier potassium (Kir2.2) channels. J Gen Physiol. 2019 Nov 19;jgp.201912422.

186. Mansoor SE, Lü W, Oosterheert W, Shekhar M, Tajkhorshid E, Gouaux E. X-ray structures define human P2X 3 receptor gating cycle and antagonist action. Nature. 2016 Oct;538(7623):66–71.

187. Yoder N, Yoshioka C, Gouaux E. Gating mechanisms of acid-sensing ion channels. Nature. 2018 Mar;555(7696):397–401.

188. Minor DL, Lin Y-F, Mobley BC, Avelar A, Jan YN, Jan LY, et al. The Polar T1 Interface Is Linked to Conformational Changes that Open the Voltage-Gated Potassium Channel. Cell. 2000 Sep 1;102(5):657–70.

189. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. Cell. 2009 Aug;138(4):774–86.

190. Gimona M. Protein linguistics — a grammar for modular protein assembly? Nat Rev Mol Cell Biol. 2006 Jan;7(1):68–73.

191. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, et al. DisProt: the Database of Disordered Proteins. Nucleic Acids Research. 2007 Jan 3;35(Database):D786–93.

192. Alva V, Söding J, Lupas AN. A vocabulary of ancient peptides at the origin of folded proteins. eLife. 2015 Dec 14;4:e09410.

193. Pugalenthi G, Suganthan PN, Sowdhamini R, Chakrabarti S. MegaMotifBase: a database of structural motifs in protein families and superfamilies. Nucleic Acids Research. 2007 Dec 23;36(Database):D218–21.

194. Chen X, Gründer S. Permeating protons contribute to tachyphylaxis of the acid-sensing ion channel (ASIC) 1a. J Physiol. 2007 Mar 15;579(Pt 3):657–70.

195. Richler E, Shigetomi E, Khakh BS. Neuronal P2X2 Receptors Are Mobile ATP Sensors That Explore the Plasma Membrane When Activated. Journal of Neuroscience. 2011 Nov 16;31(46):16716–30.

196. Bakan A, Meireles LM, Bahar I. ProDy: Protein Dynamics Inferred from Theory and Experiments. Bioinformatics. 2011 Jun 1;27(11):1575–7.

197. Golinski AW, Holec PV, Mischler KM, Hackel BJ. Biophysical Characterization Platform Informs Protein Scaffold Evolvability. ACS Comb Sci. 2019 Apr 8;21(4):323–35.

198. Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: A web-based graphical user interface for CHARMM. Journal of Computational Chemistry. 2008;29(11):1859–65.

199. Huang J, MacKerell AD. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. Journal of Computational Chemistry. 2013;34(25):2135–45.

200. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. Journal of Computational Chemistry. 2005;26(16):1781–802.

201. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 2006 Nov 1;22(21):2695–6.

202. Konopka BM, Marciniak M, Dyrka W. Quantiprot - a Python package for quantitative analysis of protein sequences. BMC Bioinformatics. 2017 Dec;18(1):339.