Evolution and the design of allosteric transcription factors toward novel ligand specificities

By

Kyle K. Nishikawa

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Biochemistry)

At the

UNIVERSITY OF WISCONSIN-MADISON

2022

Date of final oral exam: 4/12/2022

The dissertation is approved by the following members of the Final Oral Committee:
Srivatsan Raman, Assistant Professor, Biochemistry
Brian G. Fox, Professor, Biochemistry
James L. Keck, Professor, Biomolecular Chemistry
Eric V. Shusta, Professor, Chemical and Biological Engineering

Acknowledgments

This work was supported by is partially supported by the US Army Research Office (W911NF-17-1-0043). This research was performed using the compute resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. Use of the LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (Grant 085P1000817). I thank Spencer Anderson, Joseph Brunzelle, Elena Kondrashkina, and Zdzislaw Wawrzak (LS-CAT) and Craig Ogata, Michael Becker and Nagarajan Venugopalan (GM/CA@APS) for synchrotron beamline support. This research was performed using the instruments and assistance of the UW Biotech Center DNA Sequencing Core Facility. K.N is supported by NIH National Research Service Award T32 GM07215 and the Robert and Katherine Burris Biochemistry Fund. I thank my advisor Srivatsan Raman for his guidance in developing this research.

Table of Contents

Acknowledgments i
Table of Contents ii
Abstract iv
1.0.0 Transcription factor evolution and engineering 1
1.1.0 Introduction 2
1.2.0 Protein evolution, evolvability, and epistasis 4
1.2.1 Epistasis and fitness landscapes 5
1.2.2 Effect of epistasis on evolutionary trajectories 6
1.2.3 Types of epistasis 7
1.2.4 Directed evolution and ancestral reconstruction are tools to study epistasis 8
1.2.5 Epistasis in proteins with complex function 9
1.3.0 Engineering novel function into transcription factors 10
1.3.1 Challenges in engineering novel ligand affinity into aTFs 11
1.3.2 Computational design to engineer ligand binding 12
1.3.3 Fluorescence screens to find successful designs 13
1.3.4 RNA-Seq is an alternative to fluorescence screens 14
1.4.0 Probing epistasis and improving aTF redesign 15
1.5.0 References 16
2.0 Epistasis shapes the fitness landscape of an allosteric specificity switch 20
2.1 Abstract 21
2.2 Introduction 22
2.3 Computational design of ligand specificity switch 24
2.4 Epistasis shapes the fitness landscape of resveratrol response 27
2.5 Epistasis uniquely influences the fitness landscape of each ligand 32
2.6 Crystal structure reveals molecular basis of specificity of quadruple mutant 35
2.7 Relationship between biophysical affinity and biological response 37
2.8 Discussion 39
2.9 Methods 43
2.10 References 61
2.11 Figures 64
2.12 Supplementary Information 69
3.0 Radical redesign of allosteric transcription factors to bind to novel ligands 96
3.1 Abstract 97
3.2 Introduction 98
3.3 Validating RNA-Seq on a 16-member library 100
3.4 Identifying novel sensors in the agnostic library 102
3.5 Elucidating ligand-specific sequence preferences from RNA-Seq 104
3.6 DMS of TtgR reveals functional hotspots 108
3.7 Discussion 110
3.8 Methods 114
3.9 References 128
3.10 Figures 132

3.11 Supplementary Information 138
4.0.0 Conclusion and Future Directions 160
4.1.0 Conclusions 161
4.1.1 Epistasis across multiple fitness parameters 161
4.1.2 Engineering novel ligand affinity into TtgR 162
4.2.0 Future Directions 164
4.2.1 Epistasis in allosteric transcription factors 164
4.2.2 Computational design and RNA-Seq 166
4.3.0 References 168

Abstract

The emergence of novel functions in allosteric transcription factors is driven by mutations whose effects propagate through the allosteric network to create epistasis in gene expression. Elucidating the relationship between mutations and their global effects is essential to understanding the evolution of allosteric proteins. I integrate computational design, high-throughput screening, structural analysis, and biophysical characterization to create and probe the emergence of novel inducer affinity and specificity in an allosteric transcription factor. The functional parameters of gene expression create a multidimensional fitness surface that bypasses suboptimal regions of the fitness landscape by altering selective pressures. I develop a generalizable design and screening workflow that yields sensors for eight selected small molecules and can be used to probe the function of any protein that can control transcription using variant libraries through direct measurement of RNA abundance.

1.0.0 Transcription factor evolution and engineering

1.1.0 Introduction

Allosteric transcription factors (aTFs) have critical roles in cellular processes such as development, antibiotic resistance, and metabolism in both prokaryotes and eukaryotes[1-3]. Many aTFs control gene expression by interacting with small molecules inside the cell[4,5]. These proteins have evolved from ancient predecessors, establishing the need for gene expression control across evolutionary scales[6]. Gene duplication and mutagenesis created diverse functions in extant aTFs[7]. However, the intermolecular interactions that give rise to new function in aTFs remain largely unknown. The goal of this work is to characterize the molecular mechanisms underlying the emergence of novel function and to apply new techniques to engineer ligand affinity into aTFs.

Extant transcription factors have evolved from ancient precursors that carry the capacity for gene expression control and allosteric communication[8]. Transcription factors are a complex platform from which to probe evolution because gene expression control is the product of multiple functional parameters: affinity for DNA, affinity for the small molecule, and the capability to undergo allosteric changes that alter each. Mutations that yield novel function may affect each parameter differently. The effect of mutations is further complicated by epistasis, or the non-additive effect of combinations of mutations on protein function, but the prevalence of epistasis in complex functions has not been explored.

A deeper understanding of the emergence of novel transcription factor function creates the capacity to engineer new ligand affinity into existing aTFs. Changing ligand affinity and specificity in aTFs can also reveal amino acid identities and positions in the protein that are critical for these two functions. However, engineering novel ligand affinity challenging due to the intricacy of long-range interactions in allosteric proteins. Computational approaches can readily simulate protein-ligand interactions but lack the ability to account for allosteric function. In the absence of mechanistic understanding of allosteric interactions, high-throughput screening must be used to screen for designed aTFs that have preserved allosteric function.

Prokaryotic allosteric transcription factors are a simple, one component transcription regulation system that are simple models for understanding transcription control. The TetR family of transcriptional repressors control gene expression in response to antibiotics and other organic molecules[9]. In the absence of a small molecule inducer, these proteins bind to the operator sequence in the promoter of controlled genes to prevent RNA polymerase interaction. Once the small molecule is present, the proteins undergo an allosteric change that reduces affinity for the operator sequence and allows transcription of downstream genes[10]. For the TetR family, these downstream genes are typically antibiotic transporter proteins[9]. These proteins are a useful framework that can probe the role of epistasis in controlling gene expression and are a useful chassis to engineer novel function.

In chapter 2, I examine the molecular mechanisms underlying the evolution of ligand specificity in TtgR, a member of the TetR family. For an allosteric transcription factor (aTF), function is the combination of affinity for the inducer ligand, affinity for DNA, and allosteric changes that accompany binding to the ligand. I engineer a specificity switch into TtgR to model the acquisition of novel function and show that each parameter has unique patterns of epistasis even though the epistatic interactions are consistent across parameters. These interactions and the resulting altered specificity can be rationalized through the structural model of the engineered aTF bound to resveratrol. Finally, I compare biophysical affinity to sensitivity to emphasize that epistasis may also affect the allosteric response of the protein. The unique patterns of epistasis create functional trade-offs where optimizing one function comes at the cost of another. An evolving aTF simultaneously traverses these multiple fitness landscapes and can bypass regions of low fitness by switching selective pressures.

In chapter 3, I create a library of aTF variants and develop an RNA-Seq strategy to screen transcription factors for novel ligand affinity. I generate a ligand-agnostic library of variants using phylogeny and a set of computationally stable mutations[11]. The ligand-agnostic nature of this design workflow enables the same library to be screened across multiple ligands. Using RNA-Seq, I identified multiple transcription factors that could respond to at least one of nine different ligands. I also applied the RNA-Seq screening workflow to deep mutational scanning libraries of TtgR. I tested the performance of these libraries against two ligands and found functional hotspots in the DNA binding domain, at the interface between the DNA binding domain and the ligand binding domain, and in second-shell residues of the ligand binding pocket.

In chapter 4, I discuss the conclusions of this work and future directions for both the evolution of transcription factors and engineering aTF biosensors. This work is a preliminary glimpse into the complexity of evolution in multifunctional proteins and the role of epistasis in function switching. I explore the larger implications of the pervasiveness of epistasis on the evolution of novel function in aTFs. I elaborate on the next steps required to validate the engineering workflow and improve targeted computational design approaches.

1.2.0 Protein evolution, evolvability, and epistasis

In the following sections, I discuss the importance of epistasis on the evolution of new functions. Epistasis is primarily visualized through fitness landscapes and characterized through two lenses: specific and nonspecific epistasis. Epistasis has been characterized in different proteins through the use of ancestral protein reconstruction or directed evolution. I then examine the role of epistasis in the emergence of complex functions and consider the ability to bypass classical fitness barriers by deconstructing these functions into separate fitness parameters.

Proteins are a complex arrangement of amino acids in three-dimensional space. The creation of active sites, binding domains, and allosteric interactions are dependent on both the three-dimensional arrangement and chemical properties of groups of amino acids[12]. As proteins accumulate mutations during evolution, the chemical properties and spatial orientation of these amino acid groups change to create novel interactions and functions[13,14]. Understanding the relationship between these molecular changes and the resulting influence on protein function is essential to understanding protein evolution.

A central concept to the acquisition of novel functions is evolvability. Evolvability is the potential of an amino acid sequence to acquire mutations that confer a novel function. Functional promiscuity, the capacity to interact with multiple proteins or small molecules, has been argued as a key effector of the evolution of function[15-17]. Many monofunctional proteins reach an intermediate promiscuous stage before switching to be specific for the new function, indicating promiscuity plays an important role in the evolution of novel protein functions[15,16,18]. The amino acid interactions that underlie the changes in specificity and function are often epistatic in nature, highlighting the need to understand epistasis in the context of protein function[19].

1.2.1 Epistasis and fitness landscapes

The changes in protein function that arise from mutation are dependent on the chemical nature of the mutation and its interactions with nearby amino acids. The effect of a mutation on protein function can change depending on the amino acid identities at other positions; this phenomenon is called epistasis[20]. Epistasis is typically studied through a genetic lens but has been increasingly studied in the context of protein evolution.

A useful context used to imagine epistasis is a fitness landscape, typically visualized as a three-dimensional surface. The X and Y axes are used to describe sequence space, or the set of possible amino acid sequences. The fitness of each sequence can then be plotted on the z-axis to create a topographical map of protein function. As a protein evolves, it traverses sequence space through the accumulation and fixation of mutations that increase fitness under selective pressure[18]. If each mutation contributes additively to fitness, then the fitness landscape has a single maxima representing the sequence with optimal function[21]. The number of paths that are allowed between the starting sequence and the fitness maximum is equivalent to 2^n, where n is the number of mutations between the two points. The number of available paths to a peak in the fitness landscape is one method of characterizing epistasis.

1.2.2 Effect of epistasis on evolutionary trajectories

Epistasis results in certain combinations of mutations creating context-dependent changes in fitness and the shape of the fitness landscape. In an epistatic system, the fitness landscape becomes rugged, with multiple local maxima and minima corresponding to the epistatic interactions[21]. An inherent property of the ruggedness of the fitness landscape is the path-dependence of evolutionary trajectories. Combinations of mutations will be beneficial in one region of sequence space and be deleterious in another. Thus, the order of mutations constrains the available evolutionary pathways of a particular sequence to discrete regions of sequence space[20]. Similarly, the global maxima of the fitness landscape is not readily accessible from any point in sequence space because of multiple fitness minima, resulting in multiple “dead ends” of theoretical evolutionary trajectories[22].

The effect of epistatic interactions has been argued to be the key effector in mutation fixation during evolution[23-25]. In a blue fluorescent protein, epistatic interactions made up a minority of all interactions, but were sufficient to build a model that could accurately predict phenotypes of combinations of mutations with a correlation of 0.98[24]. Epistatic interactions are differentially distributed across protein domains, but contribute significantly to function[25]. Reshaping the fitness landscapes drastically alter the paths through sequence space that a protein can take to acquire novel function.

1.2.3 Types of epistasis

Epistatic interactions are characterized by both the effect on protein function and the synergy of the two contributing mutations. Specific epistasis refers to epistatic interactions in the context of biophysical properties. Nonspecific epistasis refers to nonlinear effects on biological properties like solubility, expression, or activity. In addition to these general categories based on functional effect, one can also describe epistatic interactions in terms of individual and combinatorial effects of mutations[26]. Magnitude epistasis occurs when the combinatorial effect of two mutations is amplified in comparison to the individual effect of each. Sign epistasis occurs when the effect of one mutation is dependent on the presence of the other. Reciprocal sign epistasis occurs when the effect of each mutation reverses in the presence of the other.

Specific epistasis encompasses function-switching mutations that can interact with DNA, small molecules, or other proteins, but are affected by the identity of physically interacting nearby residues. These nearby residues are permissive mutations, which can be required for the function switch to occur[27]. Permissive mutations must generally satisfy three categories: stabilization of protein structure, maintenance of free energy states in different conformations, and compatibility with parent and derived sequences[28]. These mutations, when combined, nonlinearly affect the biophysical properties contributing to novel function[8].

In contrast to specific epistasis, nonspecific epistasis affects the function at the biological level. Nonspecific epistasis is commonly observed in the context of protein stability across a large range of proteins[29]. The initial increase in stability of the thermostable variants were required for fixation of function-switching mutations[30,31]. Stability-mediated epistasis can be achieved through a variety of means; any stabilizing mutation will thus be epistatic with any destabilizing, function-conferring mutation. This nonspecific pairing is the root of nonspecific epistasis. The molecular mechanism of stabilization may differ, but the overall effect remains the same[32,33]. The key effect is that each mutation affects the same biophysical property, like stability, that then has a nonadditive effect on biological function. In these examples, any mutation that stabilizes the protein exhibits nonspecific epistasis for any destabilizing mutation that also changes function. Since the nonadditive effect manifests at the biological level, the mutations do not necessarily directly interact.

1.2.4 Directed evolution and ancestral reconstruction are tools to study epistasis

The influence of epistasis on protein evolution can only really be examined in depth when the sequence history of a protein is known. For example, stability-mediated epistasis in the evolution of influenza has benefitted from the availability of the genome sequence of previous historical strains that enabled closer examination of residue interactions[30]. Directed evolution is one laboratory technique that is used to engineer novel function and preserves the sequence history of the evolved protein[19,34]. However, directed evolution cannot be applied to examine protein evolution in existing aTFs as the process involves iterative mutagenesis.

Ancestral reconstruction is one approach that can enable the laboratory resurrection of ancient protein sequences using a combination of phylogenetically related sequences and maximum likelihood estimations[35]. Ancestral reconstructions of steroid hormone receptors have revealed numerous intricacies about the evolution of mammalian transcription factors[36]. Ancestral reconstruction has also been employed to examine the folding properties of RNAse H and ligand-specificity of methyl-parathion hydrolase. In both cases, subtle changes in structure lead to biological differences in either the folding pathway or the activity between ancestral and extant proteins[37,38]. Ancestral reconstruction provides the ability to test ancient protein sequences to investigate the role of epistasis in the evolution of natural proteins.

1.2.5 Epistasis in proteins with complex function

The influence and abundance of epistasis has been thoroughly examined in numerous proteins[28,37]. However, many studies focus on a singular aspect of protein function to measure the effects of epistasis. Stability, enzyme activity, DNA affinity, or small molecule affinity are all aspects of biological function; examining each individually will give an inaccurate representation of the influence of a mutation on protein fitness. For simple proteins with a singular biological function such as antibodies and binding affinity, mutations are fixed under a single biological context. However, complex proteins have multiple functions that can be optimized at different times during evolution. Often, these proteins evolve towards one specific function and become specialized after gene duplication events[39]. This process can occur through stepwise accumulation of mutations conferring additional specificity as subtle differences appear in the effector molecules[17].

There is a limited understanding of the effects of mutations in multifunctional proteins on multiple parameters and functions in an evolutionary context. In addition to selection pressures that may favor one biological function over another during evolution, each function can also be divided into multiple parameters that can individually affect fitness. Transcription factor gene expression control can be affected by changes in ligand binding affinity, DNA affinity, and allosteric communication. While evolution will drive biological function towards an optimal fitness under selection pressure, the biophysical mechanisms of improved fitness differ[32]. Like biological functions, each parameter may be selected independently during evolution based on the responsiveness of the transcription factor or the activity of the enzyme. Each function and parameter can be visualized with individual fitness landscapes. Elucidating the trajectories complex proteins undergo through sequence space during evolution requires understanding of the intersection of fitness parameter and function under selective pressure.

1.3.0 Engineering novel function into transcription factors

A fundamental property of many transcription factors is the ability to alter gene expression in response to a small molecule. These molecules may be metabolites, therapeutics, or solvents that the cell must respond to in order to survive in changing environments. The previous section details the complexity of interactions that create transcription factor function; a natural progression of this understanding is the engineering of new functions in characterized aTFs with rational approaches.

The capacity to engineer small molecule affinity into aTFs is important because an engineering workflow can be used to create novel biosensors. Biosensors are devices that use a biological component to sense analytes in the environment. Biosensors have also been used in environmental monitoring, food quality monitoring, and drug discovery[40]. Transcription factor biosensors are particularly useful because the biological sensing and the production of the transducer are incorporated into a single protein *in vivo*. Transcription factors have been used in a broad array of sensing applications such as detection of trace compounds, generation of complex gene circuits with natural and engineered transcription factors, and modulation of metabolic control[41-48]. The ability for cells to control the expression of key enzymes and proteins to optimize readout is critical to generating automated biosynthetic production pipelines.

1.3.1 Challenges in engineering novel ligand affinity into aTFs

Engineering novel ligand affinity into transcription factors poses two major challenges. First, redesigning the binding pocket of proteins to accommodate chemically distinct small molecules is a challenging task. These proteins are also allosteric[49,50]. As mutations are introduced into the ligand binding pocket to engineer affinity, the allosteric network of residues must be maintained so that the act of ligand binding can be communicated to the DNA binding domain.

Computational design and directed evolution are two approaches that can facilitate acquisition of new functions through mutagenesis. Directed evolution of an aTF is best suited for target molecules that weakly interact with the wildtype transcription factor because few initial substitutions must confer measurable improvement in function[51]. Computational design enables rapid testing of a large number of amino acid sequences *in silico*, producing a set of sequences that are most likely to interact with the ligand of interest[52]. This approach can be used to engineer affinity for molecules that are drastically different from the native ligand. While efficient at creating and optimizing close-range protein-ligand interactions, computational methods cannot account for the long-range interactions that arise in allosteric proteins[53]. This limitation creates the possibility of designed transcription factor sequences with high affinity for the target molecule, but without the ability to undergo allosteric changes.

Without prior knowledge of allosteric interactions, large numbers of designed aTFs must be tested. High-throughput screening techniques such as RNA-Seq and fluorescence-activated cell sorting (FACS) facilitate testing of many variants in a single experiment. The combination of computational design and high-throughput screening will create and isolate a transcription factor variant that has affinity for the target and maintains inherent allosteric properties.

1.3.2 Computational design to engineer ligand binding

Rosetta is a software suite developed to model the molecular interactions that comprise protein tertiary and quaternary structures but is insufficient to engineer transcription factor biosensors alone[54]. Rosetta has been extensively used to model novel protein interactions[55-59]. LacI was engineered to bind to novel sugar molecules like sucralose, gentiobiose, fucose, and lactitol using a combination of Rosetta design and high-throughput screening[60]. The success of the LacI redesign effort represents a small step in chemical space away from the natural ligand. Future endeavors must strive to push the boundaries of binding pocket design to create tools that allow more radical redesign of the transcription factor that expands the repertoire of biosensors.

Despite the wide range of successful protein designs using the Rosetta software suite, the computational design process has limitations that must be considered prior to its implementation. Any *in silico* model has errors in its energy functions used to model amino acid states that is propagated across all residues in the protein[61]. The Rosetta Energy Function is used to calculate the energy of an amino acid conformation and is a linear combination of different energy parameters[62]. These parameters model biophysical properties like electrostatics, repulsive forces between atoms, solvation energies, and hydrogen bonding energies. However, there is no guarantee that these computational models are accurate[63]. Over reliance on computational approaches of engineering novel affinity may be detrimental in the absence of thorough experimental validation and understanding of allosteric interactions.

1.3.3 Fluorescence screens to find successful designs

High throughput screens for computationally designed libraries can use the expression of a fluorescent protein as a marker for transcription factor function. Flow cytometry and fluorescence activated cell sorting (FACS) enables rapid screening of thousands of transcription factor variants using GFP fluorescence[60]. The transcription factor variants are assayed for the proper function based on GFP expression levels in the presence and absence of the target small molecule[57]. Each variant has its own unique fluorescence profile, and the fluorescence distribution of the resulting library is the summation of all the fluorescence distributions of the individual variants. Repeated sorting of different fluorescence populations in either the presence or the absence of the ligand can isolate functional designs.

The sorting process relies on the ability of the transcription factor variants to control gene expression in response to the small molecule. The function of the variant is then dependent on fraction of the population isolated for both sorts in the presence of the ligand and sorts in the absence of the ligand. Isolating a smaller fraction theoretically subsets variants with higher induced gene expression and lower basal expression but may also result in the loss of the variants with intermediate phenotypes. Once an observable shift in fluorescence between the uninduced and the induced libraries is obtained during the sorting process, clonal assays will identify the variants responsible.

1.3.4 RNA-Seq is an alternative to fluorescence screens

RNA-Seq is a high-throughput alternative to the fluorescence-based screens that overcomes several limitations of cell sorting approaches and generates functional scores for every variant in the library. Fluorescence assays can also be used to obtain the phenotypes of all library members by incorporating next-generation sequencing technology. These Sort-Seq approaches separate a fluorescence distribution into bins of discrete fluorescence ranges via fluorescence-activated cell sorting (FACS)[64-68]. Each bin is then sequenced with NGS to elucidate the abundance of each library member within the bin. The proportion of reads attributed to a single variant across all fluorescence bins is used to infer the fluorescence distribution of the variant. However, accurate reconstruction of the phenotypes of the library members requires careful consideration of the range of the bins and knowledge of the underlying individual distributions[64].

RNA-Seq based approaches map transcription factor function in without the need to partition the library. RNA-Seq relies on sequencing a short, random set of nucleotides called a "barcode". In a transcription factor library, the transcription factor variants control the expression of unique barcodes[69]. Each transcription factor variant will be linked to multiple barcodes. These barcodes can be mapped back to the transcription factor variant controlling its expression and then sequenced using NGS to elucidate the abundance of each barcode. The performance of the transcription factor variant is calculated by taking the ratio of these barcode abundances in the presence and absence of small molecule ligands. This approach enables the characterization of the entire transcription factor library in a single pooled assay.

One of the main challenges of the barcode-based RNA-Seq approach is mapping the randomized barcode to the variant responsible for its expression. One way to overcome this challenge was developed to characterize the architecture of σ^{70} promoters using a short barcode at the 3' of the *sfGFP* gene[70]. The barcodes were mapped to the promoter variant responsible for their expression prior to incorporation of the sfGFP gene. Another similar approach involved short barcodes mapping to GPCR activity in human cell lines[69], which established a generalizable method of screening eukaryotic transcription factors. PacBio long-read sequencing was used in another deep mutational scanning library of the SARS-CoV-2 receptor binding domain[71].

RNA-Seq yields counts of all barcodes extensively mapped to transcription factor variants, creating a fitness landscape of ligand responsiveness over the designed sequence space. This approach increases the number of variants that can be screened compared to Sort-Seq. Sort-Seq requires sequencing of multiple fluorescence partitions, which requires higher read volumes per library or fewer variants. Furthermore, RNA-Seq is a direct measure of aTF function as it measures the abundance of transcripts instead of GFP expression as a proxy. High-throughput screens of computationally designed aTF libraries would benefit from the improvements provided by RNA-Seq screens.

1.4.0 Probing epistasis and improving the aTF redesign

Despite their importance in biotechnology applications, relatively little is known about transcription factors and the molecular mechanisms used in these proteins to confer novel ligand affinity. These proteins often are simple, 1-component transcription regulators that control gene expression in response to small molecules. A key aspect of aTF function is allostery; these proteins undergo conformational or dynamic changes in response to ligand binding that alters their affinity for DNA. Controlling gene expression is a complex function that involves multiple functional parameters. Epistasis can affect each parameter to different extents, creating complex mutational interactions over the course of evolution. The goal of the first part of this work is to probe these intricate epistatic interactions across multiple functional parameters using changing ligand specificity.

Transcription factors represent a versatile chassis and untapped resource for sensing platforms. Expanding the use of aTF sensors requires expanding the repertoire of molecules that can be sensed with these proteins. However, previous methods that aimed to engineer novel ligand affinity had limited success. The latter part of this work is devoted to creating new computational and high-throughput screening workflows that can generate new aTF biosensors that increase the range of sensed molecules.

1.5.0 References

1 Li, Y., Luo, H., Liu, T., Zacksenhaus, E. & Ben-David, Y. The ets transcription factor Fli-1 in development, cancer and disease. *Oncogene* **34**, 2022-2031, doi:10.1038/onc.2014.162 (2015).

2 Gong, Z., Li, H., Cai, Y., Stojkoska, A. & Xie, J. Biology of MarR family transcription factors and implications for targets of antibiotics against tuberculosis. *J Cell Physiol* **234**, 19237-19248, doi:10.1002/jcp.28720 (2019).

3 Meinhardt, S. *et al.* Novel insights from hybrid LacI/GalR proteins: family-wide functional attributes and biologically significant variation in transcription repression. *Nucleic Acids Res* **40**, 11139-11154, doi:10.1093/nar/gks806 (2012).

4 Ogino, Y. *et al.* Functional distinctions associated with the diversity of sex steroid hormone receptors ESR and AR. *J Steroid Biochem Mol Biol* **184**, 38-46, doi:10.1016/j.jsbmb.2018.06.002 (2018).

5 Friedman, R. & Hughes, A. L. Molecular evolution of the NF-kappaB signaling system. *Immunogenetics* **53**, 964-974, doi:10.1007/s00251-001-0399-3 (2002).

6 Liu, Q. *et al.* Ancient mechanisms for the evolution of the bicoid homeodomain's function in fly development. *Elife* **7**, doi:10.7554/eLife.34594 (2018).

7 Bridgham, J. T., Carroll, S. M. & Thornton, J. W. Evolution of Hormone-Receptor Complexity by Molecular Exploitation. *Science* **312**, 5 (2006).

8 Anderson, D. W., McKeown, A. N. & Thornton, J. W. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *Elife* **4**, e07864, doi:10.7554/eLife.07864 (2015).

9 Cuthbertson, L. & Nodwell, J. R. The TetR family of regulators. *Microbiol Mol Biol Rev* **77**, 440-475, doi:10.1128/MMBR.00018-13 (2013).

10 Orth, P., Schnappinger, D., Hillen, W., Saenger, W. & Hinrichs, W. Structural basis of gene regulation by the tetracycline inducible Tet repressor–operator system. *Nat Struct Biol* **7**, 5 (2000).

11 Khersonsky, O. *et al.* Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol Cell* **72**, 178-186 e175, doi:10.1016/j.molcel.2018.08.033 (2018).

12 Wodak, S. J. *et al.* Allostery in Its Many Disguises: From Theory to Applications. *Structure* **27**, 566-578, doi:10.1016/j.str.2019.01.003 (2019).

13 Podgornaia, A. I. & Laub, M. T. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 5 (2015).

14 Tomatis, P. E. *et al.* Adaptive protein evolution grants organismal fitness by improving catalysis and flexibility. *PNAS* **105**, 6 (2008).

15 Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nat Genet* **37**, 73-76, doi:10.1038/ng1482 (2005).

16 Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A. & Sanchez-Ruiz, J. M. Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian beta-lactamases. *J Am Chem Soc* **135**, 2899-2902, doi:10.1021/ja311630a (2013).

17 Eick, G. N., Colucci, J. K., Harms, M. J., Ortlund, E. A. & Thornton, J. W. Evolution of minimal specificity and promiscuity in steroid hormone receptors. *PLoS Genet* **8**, e1003072, doi:10.1371/journal.pgen.1003072 (2012).

18 Smith, J. M. Natural Selection and the Concept of a Protein Space. *Nature* **225**, 2 (1970).

19 Kaltenbach, M., Jackson, C. J., Campbell, E. C., Hollfelder, F. & Tokuriki, N. Reverse evolution leads to genotypic incompatibility despite functional and active site convergence. *Elife* **4**, doi:10.7554/eLife.06492 (2015).

20 Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci* **25**, 1204-1218, doi:10.1002/pro.2897 (2016).

21 Kauffman, S. A. & Weinberger, E. D. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *J theor Biol* **141**, 35 (1989).

22 Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* **10**, 866-876, doi:10.1038/nrm2805 (2009).

23 Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535-538, doi:10.1038/nature11510 (2012).

24 Poelwijk, F. J., Socolich, M. & Ranganathan, R. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nat Commun* **10**, 4213, doi:10.1038/s41467-019-12130-8 (2019).

25 Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol* **24**, 2643-2651, doi:10.1016/j.cub.2014.09.072 (2014).

26 Miton, C. M., Buda, K. & Tokuriki, N. Epistasis and intramolecular networks in protein evolution. *Curr Opin Struct Biol* **69**, 160-168, doi:10.1016/j.sbi.2021.04.007 (2021).

27 McKeown, A. N. *et al.* Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell* **159**, 58-68, doi:10.1016/j.cell.2014.09.003 (2014).

28 Harms, M. J. & Thornton, J. W. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature* **512**, 203-207, doi:10.1038/nature13410 (2014).

29 Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *PNAS* **103**, 5 (2006).

30 Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* **2**, e00631, doi:10.7554/eLife.00631 (2013).

31 Patel, M. P. *et al.* Synergistic effects of functionally distinct substitutions in beta-lactamase variants shed light on the evolution of bacterial drug resistance. *J Biol Chem* **293**, 17971-17984, doi:10.1074/jbc.RA118.003792 (2018).

32 Hart, K. M. *et al.* Thermodynamic System Drift in Protein Evolution. *PLoS Biol* **12**, 8 (2014).

33 Tzul, F. O., Vasilchuk, D. & Makhatadze, G. I. Evidence for the principle of minimal frustration in the evolution of protein folding landscapes. *Proc Natl Acad Sci U S A* **114**, E1627-E1632, doi:10.1073/pnas.1613892114 (2017).

34 Dickinson, B. C., Leconte, A. M., Allen, B., Esvelt, K. M. & Liu, D. R. Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. *Proc Natl Acad Sci U S A* **110**, 9007-9012, doi:10.1073/pnas.1220670110 (2013).

35 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591, doi:10.1093/molbev/msm088 (2007).

36 Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409-413, doi:10.1038/nature23902 (2017).

37 Yang, G. *et al.* Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. *Nat Chem Biol* **15**, 1120-1128, doi:10.1038/s41589-019-0386-3 (2019).

38 Lim, S. A., Bolin, E. R. & Marqusee, S. Tracing a protein's folding pathway over evolutionary time using ancestral sequence reconstruction and hydrogen exchange. *Elife* **7**, doi:10.7554/eLife.38369 (2018).

39 Proulx, S. R. Multiple routes to subfunctionalization and gene duplicate specialization. *Genetics* **190**, 737-751, doi:10.1534/genetics.111.135590 (2012).

40 Altamirano, M. *et al.* A novel approach to improve specificity of algal biosensors using wild-type and resistant mutants: an application to detect TNT. *Biosens Bioelectron* **19**, 1319-1323, doi:10.1016/j.bios.2003.11.001 (2004).

41 Wang, B., Barahona, M. & Buck, M. A modular cell-based biosensor using engineered genetic logic circuits to detect and integrate multiple environmental signals. *Biosens Bioelectron* **40**, 368-376, doi:10.1016/j.bios.2012.08.011 (2013).

42 Wan, X. *et al.* Cascaded amplifying circuits enable ultrasensitive cellular sensors for toxic metals. *Nat Chem Biol* **15**, 540-548, doi:10.1038/s41589-019-0244-3 (2019).

43 Bashor, C. J. *et al.* Complex signal processing in synthetic gene circuits using cooperative regulatory assemblies. *Science* **364**, 5 (2019).

44 Garg, A., Lohmueller, J. J., Silver, P. A. & Armel, T. Z. Engineering synthetic TAL effectors with orthogonal target sites. *Nucleic Acids Res* **40**, 7584-7595, doi:10.1093/nar/gks404 (2012).

45 Koch, M., Pandi, A., Borkowski, O., Batista, A. C. & Faulon, J. L. Custom-made transcriptional biosensors for metabolic engineering. *Curr Opin Biotechnol* **59**, 78-84, doi:10.1016/j.copbio.2019.02.016 (2019).

46 Raman, S., Rogers, J. K., Taylor, N. D. & Church, G. M. Evolution-guided optimization of biosynthetic pathways. *Proc Natl Acad Sci U S A* **111**, 17803-17808, doi:10.1073/pnas.1409523111 (2014).

47 Snoek, T. *et al.* Evolution-guided engineering of small-molecule biosensors. *Nucleic Acids Res* **48**, e3, doi:10.1093/nar/gkz954 (2020).

48 Liu, Y., Landick, R. & Raman, S. A Regulatory NADH/NAD+ Redox Biosensor for Bacteria. *ACS Synth Biol* **8**, 264-273, doi:10.1021/acssynbio.8b00485 (2019).

49 Leander, M., Yuan, Y., Meger, A., Cui, Q. & Raman, S. Functional plasticity and evolutionary adaptation of allosteric regulation. *Proc Natl Acad Sci U S A* **117**, 25445-25454, doi:10.1073/pnas.2002613117 (2020).

50 Popovych, N., Sun, S., Ebright, R. H. & Kalodimos, C. G. Dynamically driven protein allostery. *Nat Struct Mol Biol* **13**, 831-838, doi:10.1038/nsmb1132 (2006).

51 Nannemann, D. P., Birmingham, W. R., Scism, R. A. & Bachmann, B. O. Assessing directed evolution methods for the generation of biosynthetic enzymes with potential in drug biosynthesis. *Future Med Chem* **3**, 809-819, doi:10.4155/fmc.11.48 (2011).

52 Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* **49**, 2987-2998, doi:10.1021/bi902153g (2010).

53 Suel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* **10**, 59-69, doi:10.1038/nsb881 (2003).

54 Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, 545-574, doi:10.1016/B978-0-12-381270-4.00019-6 (2011).

55 Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-1391 (2008).

56 Quijano-Rubio, A. *et al.* De novo design of modular and tunable protein biosensors. *Nature* **591**, 482-487, doi:10.1038/s41586-021-03258-z (2021).

57 Jha, R. K., Chakraborti, S., Kern, T. L., Fox, D. T. & Strauss, C. E. Rosetta comparative modeling for library design: Engineering alternative inducer specificity in a transcription factor. *Proteins* **83**, 1327-1340, doi:10.1002/prot.24828 (2015).

58 Soh, L. M. J. *et al.* Engineering a Thermostable Keto Acid Decarboxylase Using Directed Evolution and Computationally Directed Protein Design. *ACS Synth Biol* **6**, 610-618, doi:10.1021/acssynbio.6b00240 (2017).

59 Tinberg, C. E. *et al.* Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212-216, doi:10.1038/nature12443 (2013).

60 Taylor, N. D. *et al.* Engineering an allosteric transcription factor to respond to new ligands. *Nat Methods* **13**, 177-183, doi:10.1038/nmeth.3696 (2016).

61 Leaver-Fay, A. *et al.* Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol* **523**, 109-143, doi:10.1016/B978-0-12-394292-0.00006-0 (2013).

62 Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **13**, 3031-3048, doi:10.1021/acs.jctc.7b00125 (2017).

63 Schreier, B., Stumpp, C., Wiesner, S. & Hocker, B. Computational design of ligand binding is not a solved problem. *PNAS* **106**, 18491-18496 (2009).

64 Peterman, N. & Levine, E. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**, 206, doi:10.1186/s12864-016-2533-5 (2016).

65 Ding, N. *et al.* Programmable cross-ribosome-binding sites to fine-tune the dynamic range of transcription factor-based biosensor. *Nucleic Acids Res* **48**, 10602-10613, doi:10.1093/nar/gkaa786 (2020).

66 Rohlhill, J., Sandoval, N. R. & Papoutsakis, E. T. Sort-Seq Approach to Engineering a Formaldehyde-Inducible Promoter for Dynamically Regulated Escherichia coli Growth on Methanol. *ACS Synth Biol* **6**, 1584-1595, doi:10.1021/acssynbio.7b00114 (2017).

67 Townshend, B., Kennedy, A. B., Xiang, J. S. & Smolke, C. D. High-throughput cellular RNA device engineering. *Nat Methods* **12**, 989-994, doi:10.1038/nmeth.3486 (2015).

68 Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci U S A* **110**, 14024-14029, doi:10.1073/pnas.1301301110 (2013).

69 Jones, E. M. *et al.* Structural and functional characterization of G protein-coupled receptors with deep mutational scanning. *Elife* **9**, doi:10.7554/eLife.54895 (2020).

70 Urtecho, G., Tripp, A. D., Insigne, K. D., Kim, H. & Kosuri, S. Systematic Dissection of Sequence Elements Controlling sigma70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli. *Biochemistry* **58**, 1539-1551, doi:10.1021/acs.biochem.7b01069 (2019).

71 Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310 e1220, doi:10.1016/j.cell.2020.08.012 (2020).

2.0 Epistasis shapes the fitness landscape of an allosteric specificity switch

Kyle K. Nishikawa[1], Nicholas Hoppe[1], Robert Smith[1], Craig Bingman[1]

and Srivatsan Raman[1,2,3]

Published in Nature Communications

Author Contribution: I assisted in the computational design of TtgR to interact with novel ligands. I cloned the library and performed all fluorescence assays. I purified the proteins for crystallography and biophysical assays. I performed the ITC experiments.

1 Department of Biochemistry, University of Wisconsin-Madison, Madison, WI
2 Department of Bacteriology, University of Wisconsin-Madison, Madison, WI
3 Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, WI

2.1 Abstract

Epistasis is a major determinant in the emergence of novel protein function. In allosteric proteins, direct interactions between inducer-binding mutations propagate through the allosteric network, manifesting as epistasis at the level of biological function. Elucidating this relationship between local interactions and their global effects is essential to understanding evolution of allosteric proteins. We integrate computational design, structural and biophysical analysis to characterize the emergence of novel inducer specificity in an allosteric transcription factor. Adaptive landscapes of different inducers of the designed mutant show that a few strong epistatic interactions constrain the number of viable sequence pathways, revealing ridges in the fitness landscape leading to new specificity. The structure of the designed mutant shows a striking change in inducer orientation still retains allosteric function. Comparing biophysical and functional properties suggests a nonlinear relationship between local inducer affinity and allostery. Our results highlight the functional and evolutionary complexity of allosteric proteins.

2.2 Introduction

Interactions between mutations direct the evolution of protein function[1]. As proteins evolve, they follow paths through the fitness landscape to reach a fitness peak that represents a novel function[2]. For N mutations required to confer novel function, there are N! possible pathways connecting the start and end states. However, some pathways may not be evolutionarily favorable due to epistasis – a phenomenon that occurs when the sequence background into which a mutation is introduced changes the functional effect of that mutation. The non-additivity due to epistasis strongly influences the sequence trajectory a protein takes to gain new function[1,3-5]. Therefore, understanding the nature of epistatic interactions is the foundation for investigating the mechanisms leading to novel protein function[6].

Epistasis is generally categorized as specific or nonspecific based on cause-effect relationships between the interactions of mutations and their outcome. Specific epistasis occurs between a limited number of residues that typically physically interact, leading to nonadditive changes in thermodynamically-driven biophysical properties such as protein stability or affinity[7]. Specific epistasis has been extensively investigated in protein-protein, protein-ligand, protein-DNA interactions[5,8-16]. Nonspecific epistasis occurs when mutations are nonadditive with respect to protein traits when combined[17-20]. Such mutations can be spatially distant such as a global suppressor that can interact with many destabilizing mutations with low pairing specificity[4,21,22].

In this study, we examine the role of epistasis in the evolution of ligand specificity in an allosteric transcription factor. Allostery is a fundamental mechanism by which proteins recognize environmental cues (such as binding of an inducer or effector) within a localized region resulting in modulation of function at a distal site[23,24]. Mutations in the binding pocket that trigger the allosteric network have the potential to create new epistatic interactions at the level of protein function beyond the physical interactions commonly seen in specific epistasis and can create complex nonspecific interactions. As allosteric proteins evolve toward new function, such as orthologs in different organisms, their inducer specificity changes to adapt to the new environment[25]. Allosteric proteins may accrue mutations during evolution that would simultaneously affect specificities for old and new inducers. Further, these mutations may also impact function by affecting the capability of the protein to produce an allosteric change in response to an inducer[26,27]. For an allosteric transcription factor (aTF), function is the outcome of three parameters: affinity for the inducer ligand, affinity for DNA, and allosteric changes that accompany binding to the ligand. Each of these parameters will have its own fitness function mapped over the same sequence space, creating unique fitness landscapes. An aTF simultaneously traverses these multiple fitness landscapes which collectively govern the evolutionary trajectory of the aTF under selective pressure. Thus any one fitness landscape is not adequate as a global measure of transcription factor function. We need to examine multiple fitness landscapes and characterize epistasis in each individually to understand the evolutionary trajectory of an aTF.

Here, we integrate functional, structural, and biophysical analysis to characterize epistasis in the functional parameters of an allosteric transcription factor (aTF). Using computation-guided design, we changed the ligand specificity of TtgR, a promiscuous microbial aTF, to respond to one of its native ligands (resveratrol), but not to another (naringenin) by targeting mutations to positions that directly interact with the ligand[28,29]. By reconstructing all sequence pathways connecting the two states (promiscuous and specific), we found that epistatic interactions of two distinct sets of amino acids separately drive naringenin response while increasing resveratrol response (response is the reporter expression when induced by a ligand). We characterized the fitness landscapes of TtgR in terms of four functional parameters: fold change in gene expression, basal gene expression, maximum gene expression, and sensitivity to the ligand (EC_{50}) and showed that although ligand-induced allostery is a composite effect of all four parameters, each parameter shows unique patterns of epistasis, but also notable similarities. The crystal structure of the computationally designed mutant shows that one of the mutations reshapes the binding pocket to favor resveratrol over naringenin through a striking change in its binding orientation while maintaining allostery. We found that epistasis creates distinct biophysical and biological functional landscapes. Our results highlight the functional and evolutionary complexity of allosteric proteins because pathways can traverse through multiple adaptive landscapes under evolutionary pressure[29]. Our approach also provides a general conceptual and methodological framework to investigate epistasis in transcription factors.

2.3 Computational design of ligand specificity switch

We chose TtgR, a ligand-inducible aTF belonging to the diverse TetR-like protein family, as a target for computational engineering of ligand specificity[29]. TtgR is a 1-component transcriptional system and represents the simplest molecular mechanism for converting biophysical interaction between inducer and protein into a complex biological response like transcription[29]. In the uninduced state, TtgR physically obstructs the RNA polymerase by binding to DNA[29]. When induced, ligand-binding allosterically lowers affinity for DNA, thereby allowing transcription[29,30]. Since TtgR is found in a plant-associated microbe (*Pseudomonas putida*), it is induced by multiple plant molecules including resveratrol and naringenin[28]. Thus, TtgR provides a suitable functional backdrop to investigate the role of epistasis in emergence of novel function (ligand specificity) in an allosteric protein[28,31]. To emulate emergence of novel function, we engineered TtgR to respond to resveratrol and not to naringenin.

We used computational design (Rosetta software suite) to engineer TtgR specificity by generating function-switching mutations that directly interact with the ligand[32]. Less directed approaches may yield a specificity switch, but these can also include distal mutations whose effects on ligand affinity will confound our examination of epistasis[10,33]. Since our goal was to study how local interactions shape global function, computational design was the appropriate tool as *in silico* mutations are chosen based on interaction energies between protein and ligand[34,35].

To increase resveratrol specificity, we redesigned the ligand-contacting residues for greater affinity for resveratrol, assuming greater affinity may result in greater specificity. Since Rosetta is a structure-based design tool, the absence of a resveratrol-bound TtgR crystal structure made the design task challenging because the correct position of the ligand in the binding pocket was not known *a priori*. Therefore, we generated a set of diverse starting poses (16) by docking resveratrol conformers in different orientations within the binding pocket (Fig. 1). For each starting pose, we redesigned ligand-contacting residues while permitting constrained rigid-body flexibility of the ligand and torsional flexibility of the protein backbone. We computationally generated approximately 19,000 unique TtgR design variants with an average of 5 mutations per variant (Supplementary Fig. 1). After design, each output variant comes with a set of Rosetta-calculated scores that reflect physical properties such as stability, repulsion, hydrogen bonds, and protein-ligand affinity. The best variants for library construction can be selected from the distribution of all scores of output designs based on user-defined preferences. The variants were curated using parameter-specific median absolute deviation cutoffs on a select set of Rosetta scoring metrics of to yield a final list of approximately 3,500 unique sequences for experimental testing (Supplementary Fig. 2). The mutations generated in the 3,500 sequences are diverse, but designed sequences generally favor the wildtype amino acid at each mutable position (Supplementary Fig. 2). A few positions such as 96, 137, 168, and 175 have mutations that are more abundant than the wildtype amino acid. We synthesized oligonucleotides encoding approximately 3,500 designed variants as a pool of exact chip-DNA sequences (Twist Bioscience Inc).

To determine the activity of TtgR variants, we designed a pooled screen by sorting *E. coli* cells containing a GFP reporter system regulated by a TtgR operator adapted for *E. coli*. We quantified the activity of variants based on fold induction: the ratio of GFP expression with and without inducer. Fold induction is a simple measure of the transcriptional activity of an aTF that accounts for factors affected by epistasis including DNA affinity, ligand affinity and allostery[5,9]. The activity of the initial library was greater toward naringenin than resveratrol with a median fold induction of 21-fold and 2.4-fold with naringenin and resveratrol, respectively (Fig. 1). To enrich resveratrol-specific variants in the library, we devised a toggled screening scheme where we first sorted variants competent for binding to DNA (low GFP with no resveratrol) followed by sorting variants that can activate expression of the reporter (high GFP with resveratrol) (Supplementary Fig. 3). After three rounds of toggled screening, we observed much greater response to resveratrol than naringenin in the enriched population compared to the input population (Fig.1). From the enriched population, we isolated a resveratrol-specific TtgR variant with four mutations: C137I, I141W, M167L, and F168Y which we will henceforth refer to as the 'quadruple mutant'. All four mutations were in close proximity to the ligand and no mutations were found elsewhere on TtgR. The quadruple mutant gave 80- and 6-fold induction with 250μM resveratrol and 2mM naringenin, respectively, compared to 60- and 54-fold of wildtype TtgR (Fig. 1, Supplementary Fig. 4). These concentrations were selected based on maximum solubility in aqueous solution. The goal of Rosetta design was to narrow the potential designable sequence space to a subspace of sequences most likely to offer high resveratrol function. It is possible that other Rosetta designs were successful in generating ligand specificity, but were lost in the screening process that was engineered to identify only the most successful variants. We found that the while the quadruple mutant fell within the cutoffs imposed during the curation process, it was not the best in any scoring parameter. We chose the quadruple mutant as the functional endpoint for characterizing epistasis.

2.4 Epistasis shapes the fitness landscape of resveratrol response

We constructed multiple fitness landscapes derived from dose-response curves to examine epistatic constraints in the transition from wildtype TtgR to the resveratrol-specific quadruple mutant. We made all single, double, and triple mutation combinations of the four mutations that provide resveratrol specificity as individual clones, resulting in a total of 16 variants (including endpoints). Experimental fitness landscapes are a useful framework for characterizing epistasis by revealing fitness pathways through mutational intermediates that connect two functional states. Fitness landscapes are commonly illustrated as a series of nodes and edges. Each node is designated by a binary string in which each number corresponds to a mutable position. A zero indicates the wildtype amino acid identity and a one indicates the substituted amino acid. The positions in order from left to right are: 137, 141, 167, and 168 (0000 is wildtype TtgR, 1111 is quadruple mutant, and 0100 represents the I141W mutant).

The ability of a transcription factor to control gene expression in response to a small molecule is broadly described by four parameters – (1) fold change in gene expression upon induction (fold induction), (2) basal gene expression without the inducer, (3) maximum gene expression upon induction, and (4) sensitivity to ligand concentration or EC_{50}. These parameters capture the mechanistic properties of binding to inducer, binding to DNA, and allosteric communication of ligand binding. To investigate how the same set of binding pocket mutations might uniquely affect each parameter, we constructed the fitness landscape of each parameter individually. We quantified the number of viable pathways in the resveratrol landscape by requiring that each additional mutation must increase parameter fitness if the quadruple mutant performs better than wildtype or decrease parameter fitness if the quadruple mutant performs worse than wildtype. There are 24 possible pathways from wildtype to quadruple mutant (Fig. 2a). Each functional parameter shows distinctive patterns of epistasis, although some are closely related. In the fold induction landscape, viable pathways must go through 0010 as all other single mutants have lower resveratrol response relative to wildtype TtgR (Fig. 2a). This restricts the number of available pathways from 24 to a maximum of 6. From 0010, there are three possible double mutants: 0011, 0110 and 1010. Both 0110 and 0011 are not viable as their activity substantially decreases compared to 0010 (Fig. 2a). However, 1010 is viable as it gives modestly higher resveratrol response (Fig. 2a). Both C137I and M167L manifest as key permissive intermediates in the fitness landscape that allows I141W (1110) or F168Y (1011) to be added. Since 1010 is the only viable double mutant, the number of available pathways reduces to two (Fig. 2a, bold red lines). Both triple mutants (1011 and 1110) have higher resveratrol response than 1010 which allows two viable pathways to reach the quadruple mutant, which is the global maxima of this fitness landscape (Fig. 2a).

The fitness landscape of basal gene expression resembles the fold induction landscape, with identical viable pathways, as the nodes with lower basal gene expression also show higher fold induction. All the nodes along viable pathways have lower basal gene expression than wildtype TtgR (0000) with the quadruple mutant ranking among mutants with lowest basal gene expression (Fig. 2b). The adaptive landscapes of maximum gene expression and EC_{50} show similar features to each other including a general trend of increasing magnitude from 0000 to 1111 (Fig. 2c,d). Since the global maxima for maximum gene expression is 0111 (not 1111), all pathways on the maximum gene expression landscape terminate at 0111 (Fig. 2c). Six pathways are allowed in the EC_{50} landscape because of the general tendency of mutations to increase EC_{50} regardless of mutational background (Fig. 2d). There is an interesting dependence between maximum gene expression and EC_{50} where nodes with high expression tended to also have high EC_{50} (low ligand sensitivity), indicating a likely trade off where high gene expression comes at the expense of ligand sensitivity. In other words, it may be difficult to achieve an ultrasensitive response concomitantly with a large change in gene expression.

Next, we delved deeper into the key epistatic interactions that shape the fitness landscapes. Epistatic interactions are classified as magnitude, sign, or reciprocal sign based on the combined effect of a pair of mutations relative to the effect of each mutant individually. Magnitude epistasis occurs when both mutations individually are beneficial or detrimental and their combined effect is greater in magnitude than sum of their individual effects (Supplementary Fig. 5). Sign epistasis occurs when the effect of one mutation switches from beneficial to deleterious or vice versa depending on if the other mutation is present (Supplementary Fig. 5). Reciprocal sign epistasis occurs when both mutations switch effects when paired (Supplementary Fig. 5).

Two epistatic interactions, C137I-I141W and M167L-F168Y, play important roles in modulating basal gene expression and fold induction. C137I mutation makes epistatic interactions with all the other three mutations (1100, 1010, or 1001) which are critical to control basal gene expression through sign or reciprocal sign epistasis (Fig. 2b). This is best exemplified by the interaction between C137I (1000) and I141W (0100) in the basal gene expression landscape. Both 1000 and 0100 have high basal gene expression while the double mutant 1100 has low basal gene expression leading to reciprocal sign epistasis. This interaction shows mutations in the binding pocket trigger the allosteric network to create new epistatic interactions at a distal site (in this case, the DNA-binding interface). The other double mutants that contain C137I (1010 and 1001) also have decreased basal gene expression, which is maintained through the quadruple mutant by non-epistatic (1100-1111, 1010-1111, and 1001-1111) interactions (Fig. 2b). The I141W mutation is also key modulator of fold induction that manifests through controlling basal gene expression. Although this mutation by itself causes high basal gene expression (low fold induction) when paired with either M167L (0110) or F168Y (0101) in any combination, in the 1100 background both M167L (1110) and F168Y (1101) have low basal gene expression (high fold induction) and form a magnitude epistasis interaction to generate the phenotype of the quadruple mutant (Fig. 2b).

The M167L mutation makes a strong epistatic pair with the F168Y mutation, creating a reciprocal sign epistasis interaction in the EC_{50} landscape and sign epistasis in the basal gene expression, maximum gene expression, and fold induction landscapes. In the EC_{50} landscape, M167L is the only node that decreases EC_{50} that does not contain C137I (Fig. 2d). However, this effect is masked by the addition of either C137I or I141W. The two mutations show sign epistasis in the maximum gene expression landscape in the C137I background (1000-1011) and magnitude epistasis in the I141W or C137I-I141W background, indicating that the pair behavior is dependent on the background mutations (Fig. 2c).

While a qualitative description of epistasis is easy to visualize, we wanted to also quantify the extent of and characterize the type of epistasis within all individual subnetworks and the entire 16-variant system. We used Bahadur expansion to describe all pairwise and higher order interactions (see methods)[36]. The Bahadur expansion models the activity of the landscape using a linear sum of interaction terms and coefficients. Orders of interactions (first [solo], second [pairwise], third [three way], or fourth [four way]) can be included in this sum to understand their contribution to modeling the behavior of all variants. For each subnetwork, we computed the correlation coefficient between a linear sum of first order interaction terms and actual experimental data. In the simplest case of no epistasis, the correlation coefficient of this comparison (R^2) is close to 1, but any deviation ($R^2<1.0$) indicates prevalence of epistasis. The patterns of epistasis for fold induction, basal gene expression, maximum gene expression, and EC_{50} are all different. Of the 24 possible subnetworks, 11 subnetworks are epistatic in the fold induction landscape which includes seven, four, and two instances of sign, reciprocal sign and magnitude epistasis, respectively (Fig. 2e). In the basal gene expression landscape, 12 subnetworks are epistatic with ten sign and two reciprocal sign epistasis subnetworks (Fig. 2f). The maximum gene expression landscape has 24 epistatic subnetworks: 12 magnitude, 10 sign, and 2 reciprocal sign (Fig. 2g). The EC_{50} landscape has 19 epistatic subnetworks: 7 magnitude, 9 sign, and 3 reciprocal sign (Fig. 2h). The magnitude and location of the epistatic interactions in the fitness landscapes are unique to their respective fitness property.

Since small deviations in activity may be permitted during evolution, we relaxed the requirement that each subsequent step through sequence space change fitness to be more like the quadruple mutant. We allowed small losses in function of 25% between nodes and found that additional pathways are tolerated in the basal gene expression, maximum gene expression, and EC_{50} landscapes. No additional pathways exist in the resveratrol fold induction landscape (Supplementary Fig. 6).

Epistasis thus has a large role in shaping the fold induction landscape between the promiscuous wildtype and resveratrol-specific quadruple mutant through key interactions. These same interactions create unique epistatic interactions in the fitness landscapes of basal gene expression, maximum gene expression, and EC_{50}. Although the global expansion first-order terms explain the majority of the variance in the fold induction landscape, higher-order epistatic interactions influence resveratrol fold induction by modulating interactions in secondary and tertiary subnetworks to improve the resveratrol response (Supplementary Fig. 7).

2.5 Epistasis uniquely influences the fitness landscape of each ligand

As inducer specificity changes, the fitness landscape of the same mutational intermediates will differ for each inducer. These differences may reveal alternative adaptive pathways in the fitness landscape of one inducer that circumvent functional "dead ends" in the fitness landscape of another inducer. Therefore, we examined the fitness landscape of naringenin-induced response by evaluating the same four parameters: fold induction, basal gene expression, maximum gene expression (at 2000μM), and EC_{50} of all 16 variants for comparison with the fitness landscapes of resveratrol. We determined the number of viable pathways by requiring that each additional mutation must have a change in fitness that bridges wildtype and the quadruple mutant to emulate the progressive change in function during evolution.

In the fold induction landscape, none of the 24 possible pathways viably connect wildtype to quadruple mutant because the global minima (variant with lowest naringenin response) in the landscape is the double mutant 0110, not the quadruple mutant (1111) (Fig. 3a). In the basal gene expression landscape, three pathways connect wildtype to the quadruple mutant through the C137I (1000) mutation (Fig. 3b). Pathways emerging from 1000 pass through two double mutants, 1001 and 1100, with lower basal gene expression. The basal gene expression of 1001 is higher than 1100, allowing 1001 to link to both triple mutants (1011 and 1101) compared to the single triple mutant from 1100 (1110). The maximum gene expression landscape contains two pathways connecting wildtype to quadruple mutant (Fig. 3c). Although many nodes have lower maximum gene expression compared to the preceding node, most are not part of pathways that bridge wildtype and the quadruple mutant. Two single mutants (1000 and 0100) have lower maximum gene expression than wildtype, but only one is connected to a viable double mutant (0110). Both triple mutants (0111 and 1110) accessible from 0110 connect to the quadruple mutant. Like the EC_{50} landscape of resveratrol, the EC_{50} landscape of naringenin is characterized by a general increase in EC_{50} as mutations accumulate (Fig. 3d). There are 8 possible pathways that link wildtype to the quadruple mutant. Three of the four single mutants increase EC_{50} (0100, 0010, and 0001). Four of the double mutants and all the triple mutants are accessible by at least one of the preceding nodes, but not every double or triple mutant is accessible from all preceding nodes due to minor deviations in the general trend of increasing EC_{50}. No additional mutational pathways are tolerated even when increases of up to 25% naringenin response are allowed between nodes for the naringenin fold induction landscape (Supplementary Fig. 8). Similarly to the resveratrol landscapes, the basal gene expression, maximum gene expression, and EC_{50} landscapes show additional pathways at this tolerance.

Closer examination of the role of individual mutations shows that C137I and I141W have strong effects on multiple landscapes. C137I (1000) is the only mutation that decreases EC_{50} relative to wildtype (Fig. 3d). Two additional double mutants 1010 and 1001 further decrease EC_{50} but pairing C137I with I141W (1100) or C137I with both M167L and F168Y (1011) increases in EC_{50}, suggesting that these mutational combinations may mask the effect of C137I. As with the resveratrol landscapes, the I141W mutation has an important role in modulating basal gene expression and fold induction (Fig. 3a,b). Any mutant containing I141W, but not C137I has higher basal gene expression (low folder induction) than wildtype. Combining I141W and C137I results in a large decrease in basal gene expression, which further decreases upon addition of either M167L (1110) or F168Y (1101). M167L and F168Y individually result in incremental changes in basal gene expression, maximum gene expression, and EC_{50} (Fig. 3b,c,d). However, the M167L-F168Y double mutant shows interesting context-dependent effects due to epistasis. For example, in the fold induction landscape, the combination of M167L and F168Y is beneficial in 1000 background, but detrimental in the 1100 background (Fig. 3a). This dependent behavior extends to all the other fitness landscapes even though the mutational background and types of epistasis change.

Next, we quantified epistasis both within all individual subnetworks and the entire 16-variant system. Epistasis was much more prevalent in the subnetworks of the fitness landscapes of naringenin than those in the fitness landscapes of resveratrol. In the fold induction landscape, 19 of the 24 subnetworks show epistasis (Fig. 3e). Nine were sign, six magnitude, and four reciprocal sign. In the basal gene expression landscape, 16 subnetworks show epistasis with 3 examples of magnitude epistasis and 13 examples of sign epistasis (Fig. 3f). The maximum gene expression landscape has 24 epistatic subnetworks with 11 magnitude, 6 sign, and 7 reciprocal sign subnetworks (Fig. 3g). The EC_{50} landscape has 17 epistatic subnetworks: 8 magnitude epistasis, 8 sign epistasis, and 1 reciprocal sign epistasis (Fig. 3h). The same set of mutations that create epistatic interactions giving rise to high resveratrol response forge ligand-specific epistatic patterns in the fold induction, basal gene expression, maximum gene expression, and EC50 landscapes (Supplementary Fig. 9).

Epistasis shapes the fitness landscape of each function (naringenin and resveratrol) in distinct ways. Furthermore, each functional parameter (basal gene expression, maximum gene expression, or EC_{50}) is affected uniquely by the addition of multiple combinations of mutations. I141W controls high basal gene expression and strongly modulates fold induction regardless of ligand. In contrast, C137I is more context-dependent; it is responsible for low EC_{50} values solo or in combination with either M167L or F168Y in the naringenin landscape, but is strongly influenced by M167L in the resveratrol EC_{50} landscape. Some epistatic pairs are consistent between resveratrol and naringenin. The C137I+I141W pair strongly affects basal gene expression and fold induction for both ligands. The M167L+F168Y pair has unique behavior in all fitness landscapes that is dependent on the mutation background into which they are introduced. However, the pair's effect on the wildtype background is stronger in resveratrol compared to naringenin for all parameters.

2.6 Crystal structure reveals molecular basis of specificity of quadruple mutant

To understand the structural basis of TtgR-ligand interactions, we solved high-resolution crystal structures of quadruple mutant (resveratrol-bound and apo) and wildtype TtgR (resveratrol-bound) at a resolution of 1.9Å or better (Table S2). TtgR is a compact, dimeric, all-helical transcription factor with a large cavity between five angled helices forming the ligand binding pocket (Supplementary Fig. 10a,b). The quadruple mutant bound to resveratrol (PDB: 7KD8) is structurally very similar to the wildtype with an all-atom RMSD of 1.2Å over the entire structure. The DNA binding domains of the resveratrol-bound quadruple mutant and the resveratrol-bound wildtype are extremely similar with an all-atom RMSD of 1.0Å (Supplementary Fig. 11). The four mutations do not substantially change the volume of the pocket (215Å^3 in wildtype compared to 234Å^3 in the quadruple mutant) or the surface area of the pocket (184Å in wildtype compared to 186Å in the quadruple mutant) (Supplementary Fig. 12). The position and orientation of resveratrol in the wildtype TtgR structure (PDB: 7K1C) resembles the position and orientation of naringenin in a previously solved co-crystal structure of TtgR (PDB: 2UXU)[28]. In both structures, the ligands bind in a vertical mode such that the plane of the molecule is roughly perpendicular to DNA (Supplementary Fig. 10c). In wildtype TtgR, the four mutated positions (C137, I141, M167 and F168) are located approximately in the center of the binding pocket and make nonspecific van der Waals interactions with resveratrol (Fig. 4a, upper panel). Other neighboring residues N110, D172 and H114 make specific hydrogen bonds that stabilize resveratrol in the vertical orientation (Fig. 4a, lower panel). Although both naringenin and resveratrol bind in the vertical orientation, only N110 is able to make a hydrogen bond with both naringenin and resveratrol[28]. The ability of wildtype TtgR to bind multiple ligands likely arises from the nonspecific interactions made by the nonpolar amino acids in the binding pocket.

Structure of the quadruple mutant reveals the role of individual residues in ligand specificity. I141W, a mutation critical for resveratrol specificity, creates a large steric barrier that alters the shape of the pocket and obstructs the vertical binding orientation of ligands (Fig. 4b, upper panel). Resveratrol is accommodated in the binding pocket in a horizontal binding orientation almost parallel to the plane of the tryptophan. Unlike I141W which plays a clear steric role, the other three mutations (C137I, M167L and F168Y) have a more subtle effect in reshaping the binding pocket through nonpolar interactions. C137I mutation creates a protrusion in the binding pocket that increases shape complementarity to resveratrol (Supplementary Fig. 13a). M167L is buried between the residues in the binding pocket and the dimerization helix and may play a role in positioning the I141W tryptophan to stabilize its horizontal orientation through van der Waals interactions (Supplementary Fig. 13b). F168Y allows the formation of multiple hydrogen bonds with nearby water molecules and may serve to stabilize the structure (Supplementary Fig. 13b). A different hydrogen bonding network consisting of D71, R75, and E78 make hydrogen bonds with the resveratrol molecules in chain A (Supplementary Fig. 13c) and D71, E78, D172, and a nearby water molecule make a hydrogen bond with the single resveratrol molecule in chain B (Fig. 4b, lower panel).

Although resveratrol and naringenin share similar chemical backbones, naringenin is bulkier than resveratrol due to the fused carbon rings of the chromanone. This reduces shape complementarity of naringenin to the redesigned binding pocket despite the similarity in volume of the quadruple mutant and wildtype binding pockets (Supplementary Fig. 12, 14).. The 4-hydroxyphenyl moiety and the carbonyl group of the 4-chromanone backbone of naringenin could create steric clashes with residues lining the wall of the pocket and cause the ligand to sample less space in the pocket compared to resveratrol, which provides a reasonable structural basis for ligand specificity.

The new binding mode of the quadruple mutant was not predicted in the original design scheme. We seeded the input structures for Rosetta design with resveratrol docked in the vertical orientation to mimic the binding mode of the wildtype structure. The design process is only able to make minor alterations to the position and angle of the ligand in the binding pocket (Supplementary Fig. 15).

The structural basis of ligand specificity relies on the I141W substitution to create a steric barrier to prevent binding in the vertical orientation, which is observed in wildtype TtgR for multiple ligands. In the novel horizontal mode, other ligands may be occluded from the pocket through steric clashes with wildtype residues in the pocket. The epistatic interactions observed in the fitness landscapes for naringenin and resveratrol can be rationalized through examination of the structure. The C137I-I141W pair increases shape complementarity to resveratrol while M167L-F186Y contact the dimerization helix and potentially affect the positioning of nearby residues that interact with the ligand. The altered binding mode establishes that allostery is robust to major changes in binding mode in TtgR.

2.7 Relationship between biophysical affinity and biological response

Ligand response of an aTF is a complex combination of both biophysical interactions and allostery. Mutations that affect aTF fold induction can do so by altering ligand affinity, DNA affinity, or the allosteric signal upon ligand binding. Since all four mutations are localized to the binding pocket, the observed changes in fold induction of TtgR are likely due to altered binding affinity to ligand, transmission of allosteric signal, or both. To understand the relationship between biophysical affinity and biological response, we compared changes in ligand affinity (K_d) to changes in ligand sensitivity (EC_{50}) for both naringenin and resveratrol. We chose mutants in the 0000-1000-0100-1100 subnetwork because it is important for the high resveratrol response in the quadruple mutant. Further, this network shows a strong manifestation of epistasis through reciprocal sign change and is therefore a good model to understand the relationship between biophysical affinity and biological response. We estimated ligand affinity using isothermal titration calorimetry (ITC) of purified proteins and ligand sensitivity from dose-response curves. Ligand sensitivity is derived from reporter expression and is thus a combination of both allostery and affinity.

Affinity and sensitivity of resveratrol for different variants are generally concordant for resveratrol, with the exception of 1100 (Fig. 5a). We note though that the ITC and dose response curves for some variants did not plateau due to poor ligand solubility at high concentrations resulting in imprecise estimates of K_d and EC_{50}. Nonetheless, qualitative comparisons can be made to gain useful insight. For instance, comparison of ITC profiles of 0000 and 1111 for resveratrol shows weaker binding for 1111 even though the precise K_d may be difficult to measure. Similarly, dose response curves show weaker EC_{50} for 1111 than 0000 even though it is not fully saturated. The C137I mutation appears to be largely responsible for the affinity in 1100, but the I141W mutation causes the increase in sensitivity. In general, as mutations accumulate from wildtype, the affinity and sensitivity generally decrease, suggesting a decreased ability to undergo allosteric changes is likely due to weaker binding (Fig. 5a). The discordance between affinity and sensitivity is much greater for naringenin than resveratrol. In the case of naringenin, no relationship was evident between affinity and sensitivity across the subnetwork (Fig. 5b). Although the quadruple mutant has higher resveratrol fold induction than wildtype, its affinity and sensitivity for resveratrol is lower than that of wildtype (Fig. 1, 5a). In essence, these examples illustrate the complex relationship between local interactions and their global effects in allosteric proteins.

The 0000-1000-0100-1100 subnetwork displays a unique, ligand-specific pattern of epistasis for biophysical and biological parameters. The mutations we introduced into TtgR suggest an effect on allostery changes in EC_{50} as the complexities of function may not be simply explained by changes in biophysical affinity. These measurements also suggest that by optimizing a particular protein function (fold induction), other parameters (sensitivity or affinity) may not necessarily stay at fitness maxima as the 1111 mutant shows poor sensitivity to both ligands.

2.8 Discussion

In this study, we describe the pervasive effects of epistasis on ligand specificity in a simple allosteric transcription factor by the examining fold induction, basal gene expression, maximum gene expression, and EC_{50} of two ligands across multiple mutants. By leveraging computational protein design, we engineered four mutations into TtgR, a promiscuous transcription factor that can normally bind to both resveratrol and naringenin, to only bind to resveratrol. By characterizing the functional response to both resveratrol and naringenin across all combinations of mutations, we show that the extent of epistasis between mutations affecting multiple protein functions is specific for each ligand. For instance, 50% of subnetworks meet the criteria for epistasis for resveratrol fold induction while 83% of subnetworks are epistatic for naringenin fold induction. However, the fitness landscapes of both ligands are shaped by common critical pairs of epistatic interactions (C137I and I141W or M167L and F168Y), though their behavior may be different depending on the functional parameter. Biological effects of these mutations are further validated by the crystal structures. The four mutations localize to one face of the binding pocket, making nonpolar interactions with the ligand. C137I and I141W increase shape complementarity of the pocket for resveratrol, but only in an alternative horizontal binding pose. The four mutations that confer ligand specificity decrease both affinity and sensitivity suggesting that the changes in sensitivity could be a consequence of lower affinity and not necessarily a purely allosteric effect.

Our study used a constrained set of mutations chosen through *in silico* selection as opposed to natural selection of random mutations found in *bona fide* evolutionary pathways. An evolutionary process may have selected a different set of mutations to confer the same functional outcome, leading to the presence of a different pattern of epistasis for either naringenin or resveratrol response. Often in natural evolution, mutations that are distal to the site of interest have a profound effect on protein function[8,21]. These background mutations complicate any examination of key mutations within the targeted area of the protein and their influence on protein function. By utilizing a combination of computational design and high-throughput screening, we targeted mutations to a discrete set of ligand-interacting positions within the binding pocket. Our approach enabled us to examine the propensity of epistasis in a constrained setting where mutations are limited to those that interact directly with the ligand, enabling the examination of the intersection of mutation, biophysical epistasis, and biological epistasis.

Our results highlight the dependence of epistasis on protein function and the prevalence of distinctive adaptive landscapes for multiple functions within the same set of mutations. This process highlights the functional tradeoffs that occur during an evolutionary process and raises the implication that proteins with multiple functions may readily traverse nonoptimal sequence space through varying selective pressures. These landscapes can thus become interconnected by changing selection pressures between different protein functions. On an evolutionary scale, simultaneously changing protein sequence and selection pressure may enable improbable trajectories by bypassing epistatic barriers to reach previously inaccessible mutational states. In our case, higher order epistasis which prevents access to the quadruple mutant in the naringenin fold induction landscape, could be bypassed by toggling between naringenin and resveratrol selection pressures. The evolution of allosteric proteins is inherently dependent on epistasis and the interactions arising between mutations in these proteins uniquely affects multiple adaptive landscapes.

Acknowledgements

We thank Dr. Tina Wang, Dr. Phil Romero and Phil Huss for critical review of the manuscript. This work was supported by is partially supported by the US Army Research Office (W911NF-17-1-0043). This research was performed using the compute resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. Use of the LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (Grant 085P1000817). We thank Spencer Anderson, Joseph Brunzelle, Elena Kondrashkina, and Zdzislaw Wawrzak (LS-CAT) and Craig Ogata, Michael Becker and Nagarajan Venugopalan (GM/CA@APS) for synchrotron beamline support. K.N is supported by NIH National Research Service Award T32 GM07215 and the Robert and Katherine Burris Biochemistry Fund.

Author contributions

K.N and S.R designed the study, analyzed the data, and wrote the manuscript. K.N performed all experiments. N.H carried out computational design. R.S and C.B purified and crystallized the proteins.

Competing Interests

We declare no competing interests.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

2.9 Methods

Computational Design:

Protein modeling and design was performed with Rosetta version 3.5 (2015.19.57819)[35,37]. Python and shell scripts for generating input from Rosetta and analyzing from Rosetta are available at: https://github.com/raman-lab/biosensor_design

Structure and ligand preparation:

The high-resolution TtgR structure co-crystalized with tetracycline was selected as the starting point for computational design (PDB ID: 2UXH)[28]. The structure was prepared for use in Rosetta by performing an all-atom, coordinate-constrained relaxation[38].

Commands:

```
Rosetta/main/source/bin/idealize_jd2.linuxgccrelease -database Rosetta/main/database/ -
in::file::fullatom -s 2UXH.pdb -extra_res_fa LG.params -no_optH false -flip_HNQ
```

```
Rosetta/main/source/bin/relax.linuxgccrelease -database Rosetta/main/database/ -
relax::sequence_file always_constrained_relax_script -constrain_relax_to_native_coords -
relax::coord_cst_width 0.25 -relax::coord_cst_stdev 0.25 -s 2UXH_idealized.pdb -in::file::native
2UXH_idealized.pdb -extra_res_fa LG.params -in::file::fullatom -no_optH false -flip_HNQ
```

```
Rosetta/main/source/scripts/python/public/molfile_to_params.py -n resveratrol.params -p
resveratrol.pdb
```

Protein design simulations:

The RosettaScripts protocol used to design the ligand binding pocket of each starting TtgR-resveratrol complex was based on enzyme design protocols[32,39].

Command:

```
Rosetta/main/source/bin/rosetta_scripts.linuxgccrelease -database Rosetta/main/database/ -
parser::protocol enzdes.xml -in::file::s 2UXH_resvertrol.pdb -extra_res_fa resv.params -
use_input_sc -packing:linmem_ig 10 -ex1-ex2 -run:preserve_header -enzdes_out -
enzdes:bb_min_allowed_dev 0.2 -enzdes:loop_bb_min_allowed_dev 0.5 -
enzdes:minimize_ligand_torsions 15 -parser::script_vars ligchain=X resfile=TtgR.resfile -
out::pdb -nstruct 10
```

The TtgR.resfile is a plain text file containing the amino acid position numbers that were able to be mutated during design, and these were positions 137, 141, 167, 168, 171, 172, 175, and 176. We used UW-Madison's Center for High Throughput Computing computer cluster to perform 320,000 different design simulations. The resulting designed structures were curated to yield the set of sequences that we synthesized to isolate resveratrol-specific TtgR variants.

Selection of designs for synthesis:

We selected computational designs for synthesis by first removing designs that were repetitive and then removing designs that were energetically unfavorable. The criteria for unfavorable energies were selected empirically based on the distribution of energies for all designs to yield approximately 10^4 sequences for synthesis. Specifically, on each unique design, ΔΔG stability calculations were performed on designed residues to ensure the number of destabilizing changes was limited. If the mutation destabilized the TtgR-resveratrol complex by 0.5 Rosetta Energy Units (REU), the residue was reverted to its wild-type identity. After this, non-unique designs were again removed. The unique designs were filtered using distance from the median absolute deviation of several salient Rosetta scoring metrics including total ligand binding energy, hydrogen bond energy, Leonard-Jones repulsive energy, solvation energy, and total score, which is a weighted, linear combination of all score terms in the energy function[34]. Designs that passed this filter were synthesized for library screening.

Commands:

```
./biosensor_design/fas_from_pdb_stdout.py *.pdb > TtgR_resveratrol_all_designs.fasta


./biosensor_design/uniquify_fas.py TtgR_resveratrol_all_designs.fasta >
TtgR_resveratrol_unique_designs.fasta


./ddg_monomer.static.linuxgccrelease -database ./database @ddg_flags -in:file:s
design_pdb.pdb -ddg::mut_file list_of_positions_to_calc_ddg.mutfile -ddg::iterations 50


./gen_enzdes_cutoffs.py concatentated_design_score_file.sc -c
median_abolute_deviation_cutoffs.txt -o designs_passing_filter.sc
```

The median absolute deviation cutoffs used were:

total_score < +1 MAD

fa_rep < +3 MAD

hbond_sc < +3 MAD

tot_burunsat_pm < +3 MAD

%(LIG)s_fa_rep < +3 SD

%(LIG)s_hbond_sc < +3 MAD

%(LIG)s_burunsat_pm < 2.5 ABS

%(LIG)s_total_score < -1 MAD

Library synthesis:

Creating sfGFP reporter plasmid:

The sfGFP reporter plasmid was constructed using a backbone containing the ColE1 origin and a kanamycin resistance gene. The TtgR operator sequence was modified to contain canonical -10 (5'-TATAAT-3') and -35 (5'-TTGACA-3') elements in the promoter. A strong RBS (g10) was chosen for high sfGFP expression[40]. The TtgR operator-RBS sequence was constructed via sequential PCR reactions with overlapping primers containing homology to the pColE1 backbone 5' of sfGFP. The plasmid was annealed using isothermal assembly using 0.16pmol of backbone and 0.43pmol of promoter[41]. DH10B cells (NEB) were transformed with the pColE1 reporter plasmid and plated on LB-kanamycin agar (50μg/mL). A colony was selected and grown in LB-kanamycin media (50μg/mL) shaking for 16 hours at 37°C. An aliquot of the culture was stored at -80°C in 25% glycerol. Plasmids were isolated using a DNA miniprep kit (Omega BioTek) according to the manufacturer's protocol. The insertion of TtgR operator sequence was confirmed via Sanger sequencing.

Creating TtgR expression plasmid:

The TtgR expression plasmid used the SC101 origin and a spectinomycin resistance gene. The constitutive promoter-RBS combination apFAB61-BBa_J61132 and the TtgR gene were amplified via KAPA HiFi PCR mix (Roche) using primers with homology to the pSC101 backbone[42]. The TtgR-pSC101 construct was generated using isothermal assembly (0.046pmol backbone and 0.24pmol TtgR) and DH10B cells were transformed with the TtgR-pSC101 construct. A colony was selected and grown in LB-spectinomycin media (50μg/mL) shaking for 16 hours at 37°C. An aliquot was stored at -80°C and plasmids were isolated and verified as described previously.

Library cloning:

Rosetta-designed sequences were synthesized as exact oligos (Twist Biosciences). Oligos were converted to double-strand DNA using qPCR and purified on a spin column (EZNA Cycle Pure kit from Omega BioTek). The pSC101 backbone was amplified with two separate primer pairs encoding BsaI cut sites that matched the insertion location of the oligos on the TtgR gene. The amplified backbone was treated with Dpn1 for 16 hours at 37°C (NEB) followed by a purification using a spin column. The backbone was treated with BsaI (NEB) for 2.5 hours at 37°C followed by purification using a spin column. The digested backbone was treated with Antarctic phosphatase (NEB) for 1 hour at 37°C followed by purification using a spin column. A golden gate reaction (NEB) was performed using 0.12pmol backbone and 0.89pmol library oligo in roughly a 1:7 molar ratio and incubating for 30 cycles of 37°C for 5min and 16°C for 5 min followed by 60°C for 5min. A control reaction was made using just the pSC101 backbone with no Rosetta oligos added. The golden gate reactions were dialyzed using semi-permeable membranes (Millipore) for 1 hour at 25°C against dH_2O. 25μL of C3020 cells (NEB) were transformed with 2μL of the dialyzed golden gate mixture via electroporation. Cells recovered for 1 hour in SOC media shaking at 37°C and were diluted 5X with LB. Dilutions of 100X, 500X, and 1,000X were plated to calculate transformation efficiency relative to the control. A transformation was considered successful when CFU/mL of the Golden Gate reactions exceeded CFU/mL of control reactions by a factor of 10 or more. Cells grew for 6 hours post transformation before the culture was diluted 50X and grown overnight shaking at 37°C for 16 hours. Plasmids of the library were harvested using a DNA miniprep kit and stored at -20°C.

Preparing electrocompetent cells with reporter plasmid:

An aliquot of the pColE1 frozen stock was streaked on a LB-kanamycin agar plate and grown for 16 hours at 37°C. A single colony was selected and grown in LB-kanamycin media shaking for 16 hours at 37°C. The culture was diluted 50X and grown at 37°C to an OD_{600} of 0.6. Cells were placed on ice and 5mL aliquots were centrifuged at 5,500*g* for 5 minutes at 4°C. Pellets were resuspended, washed with ice cold dH2O, and spun at 5,500*g* twice. The cells were resuspended in 20μL of water to create electrocompetent DH10B containing the pColE1 plasmid. DH10B *E.coli* containing the pColE1 reporter plasmid were transformed with the initial Rosetta library in pSC101 via electroporation. The transformed cells were recovered for 1 hour shaking at 37°C before dilutions were plated on LB-kanamycin/spectinomycin agar plates (50μg/mL each) to calculate transformation efficiency. The remaining cells were diluted 5X with LB- kanamycin/spectinomycin media and grown shaking at 37°C for 16 hours. A frozen stock was made with 25% glycerol.

Sorting the resveratrol library:

50μL aliquots of the co-transformed Rosetta libraries were thawed on ice and inoculated into 5mL of LB-kanamycin/spectinomycin and grown shaking at 37°C to an OD_{600} of 0.2. Wildtype co-transformed TtgR sensor+reporter was also inoculated as a reference. These were then split into 4 1mL aliquots and inoculated with either 500μM naringenin (DMSO), 95μM resveratrol (ethanol), DMSO, ethanol and grown for 14 hours at 37°C shaking. Cells were diluted 50X in ice cold PBS (137mM NaCl, 2.7mM KCl, 10mM Na_2HPO_4, 1.8mM KH_2PO_4) and stored on ice prior to sorting. Sorting was conducted using a Sony SH800 cell sorter. Cells were excited by a 488nm laser and GFP fluorescence was captured through a 525/50 filter. Gain settings were adjusted such that all cells fell between 10^2 and 10^6 RFU. 100,000 event measurements of all libraries, induced and repressed, were taken to draw gates according to population percentage.

Sorting followed an induced-repressed schema; the first library sort consists of taking 500,000 cells of median 50% of fluorescence from the nontreated distribution. This sort isolates cells that contain TtgR variants capable of repressing GFP expression. Cells were sorted into 2mL of LB. LB as added to a final volume of 5mL and incubated for 1 hour at 37°C shaking. Kanamycin and spectinomycin were added after 1 hour to a final concentration of 50μg/mL each from 1mg/mL stocks. These grew to an OD600 of 0.2 before frozen stocks were made in 25% glycerol. A small aliquot was stored as a frozen stock at -80°C in 25% glycerol. The remaining culture was induced with naringenin, resveratrol, DMSO, or ethanol at an OD600 of 0.2.

The next sort consisted of isolating 100,000 cells in the top 5% of fluorescence from the resveratrol-induced library. This subpopulation was grown as described previously and induced with 95μM resveratrol at an OD600 of 0.2. The final sort consisted of isolating 500,000 cells the bottom 60% of the nontreated fluorescence distribution. The sorted cells were incubated at 37°C until the culture reached an OD600 of 0.2. A frozen stock was stored at -80°C in 25% glycerol.

Clonal Testing:

Aliquots of the sorted library, wildtype TtgR, and a GFP-positive control were thawed on ice. 50μL of the library was plated on LB-kanamycin/spectinomycin and incubated at 37°C for 16 hours. The GFP control aliquot was streaked on LB-kanamycin and the wildtype TtgR aliquot was streaked on LB-kanamycin/spectinomycin and incubated in the same fashion. Colonies were selected from each plate and inoculated into 150μL of LB in a 96 well plate. The colonies were incubated at 37C shaking in a SBT1500-H microplate shaker (Southwest Science) and grew to saturation (approximately 8 hours). The cultures were diluted 15X into fresh LB with either 500μM naringenin or 95μM resveratrol and incubated in a Synergy HTX plate reader (BioTek) for 16 hours at 37°C. The performance of each colony was measured using the ratio of fluorescence to optical density (RFU/OD_{600}). The ratio of this measurement in the presence and absence of ligand defined the response to each ligand. Successful colonies had higher response for resveratrol than for naringenin. These colonies were sequenced using Sanger sequencing.

Testing of combinatorial mutants:

Generation of combinatorial mutants:

The 14 mutational intermediates were generated using eight primers specifically encoding combinations of either 137+141 or 167+168. The resulting oligos were inserted into the TtgR-pSC101 plasmid using isothermal assembly using .042pmol of backbone and 0.8pmol TtgR. DH10B *E.coli* cells (NEB) were transformed with the resulting reaction via electroporation. Colonies were selected and sequenced to verify the correct mutations were present. The correct colonies were inoculated into LB-spectinomycin and incubated at 37°C for 16 hours. An aliquot was stored at -80°C in 25% glycerol and plasmids were harvested from the remaining culture. DH10B cells were cotransformed with the 14 TtgR-pSC101 plasmids and the pColE1 reporter plasmid. These were grown for 16 hours shaking at 37°C in LB-kanamycin/spectinomycin media and frozen in 25% glycerol at -80C.

Dose response curves:

A 250mM stock of naringenin was made in DMSO and a 100mM stock of resveratrol was made in ethanol. The TtgR-pSC101/pColE1 frozen stocks were struck out onto LB-kanamycin/spectinomycin plates. Colonies were selected and inoculated into 150uL LB in a 96-well plate. These grew in a microplate shaker to saturation (approximately 8 hours) at 37°C. The cultures were diluted 15X into fresh LB-kanamycin/spectinomycin in a 96-well plate with varying concentrations of either naringenin (0μM, 10μM, 25μM, 50μM, 75μM, 100μM, 250μM, 500μM, 750μM, 1000μM, 1500μM, 2000μM) or resveratrol (0μM, 2.5μM, 5μM, 7.5μM, 10μM, 25μM, 50μM, 75μM, 100μM, 150μM, 200μM, 250μM). The concentration series for each ligand differ due to solubility limits in aqueous solutions. A series of naringenin and resveratrol stock concentrations were made such that a 50X or a 100X dilution, respectively, would yield the desired concentrations in the assay. Most variants were assayed with three biological replicates. Variants with more biological noise (1010, 1001, 1110, and 1101 for naringenin and 1001, 1000, 0001, and 0011 for resveratrol) were assayed with six replicates. The assay was incubated in the microplate shaker for 14 hours at 37°C shaking. Cells containing wildtype TtgR pSC101 with the pColE1 reporter and cells containing pColE1 reporter alone served as controls and were included on every plate. A set of 6 biological replicates of a sfGFP positive control were induced with both sets of ligands and concentrations.

Cells were diluted 50X in ice cold PBS. Fluorescence measurements were conducted on a LSR-Fortessa system (BD Biosciences) using a 488nm laser for excitation and a 530/30 filter for fluorescence emission. Using gates on FSC-H vs FSC-A, 100,000 events were gathered per well. To account for changes in fluorescence that are independent of TtgR function, raw fluorescence values were normalized by fold changes in sfGFP fluorescence in the positive control (N=6). The median values of the fluorescence distributions were used as the basis for fold induction calculations. Fold induction as calculated by obtaining the ratio of induced average median fluorescence to baseline average median fluorescence.

Quantifying epistasis:

Analyzing fluorescence data:

The mean and standard deviation of each concentration of ligand for each combinatorial mutant were used to calculate a fit using the Hill equation as a function of ligand concentration (x)[43].

$$f(x, n, EC_{50}) = F_{baseline} + ((F_{max} - F_{baseline}) * \left(\frac{x^n}{{EC_{50}}^n + x^n}\right)) \text{ (1)}$$

TtgR function was defined as the maximum fold induction of the system, which is the ratio of the median fluorescence at the highest ligand concentration and the median fluorescence at 0μM ligand:

$$fold\ induction = \frac{F_{max}}{F_{baseline}} \text{ (2)}$$

The Python 2.7 function curve_fit() from the Scipy module was used to fit the dose response curves to the Hill equation (Supplementary Fig. 16, Supplementary Fig. 17)[44]. This function provides both fit parameters and error as a covariance matrix as output. Basal gene expression was the fluorescence at 0μM ligand. Maximum gene expression was the fluorescence at the highest ligand concentration. EC_{50} was estimated using the Hill equation.

Bahadur expansion:

The Bahadur expansion was used to analyze the data[36]. Fitness for the bahadur expansion was defined as:

$$fitness_{variant} = log_{10}(\frac{fold\ induction_{variant}}{fold\ induction_{wildtype}}) \text{ (3)}$$

Fold induction in Eq.2 was changed to "basal gene expression", "maximum gene expression", or "EC_{50}" for each functional parameter. Each mutant can be represented as a numerical string (z string), where each mutable position is one number (z_i) in the string. A wild type residue at a position is designated by a -1 while the mutated residue is designated by a 1. The mutant M167L+F168Y thus becomes [-1, -1, 1, 1]. The interaction terms can be modeled as follows:

$$\varphi_0 = 1$$

$$\varphi_1, \varphi_2, \dots, \varphi_n = z_1, z_2, \dots, z_n$$

$$\varphi_{n+1}, \varphi_{n+2}, \dots, \varphi_{n+C_2^n} = z_1 z_2, z_1 z_3, \dots, z_{n-1} z_n$$

$$\dots$$

$$\varphi_{2^n-1} = z_1 z_2 \dots z_n$$

An orthonormal matrix of psi-values is created based on the combinations of mutations within the set (Supplementary Table 3). The Bahadur coefficients can be calculated using this orthonormal matrix and a fluorescence values f(x) for a particular mutant x in the set of all mutants X.

$$w_i = \frac{1}{2^n} \sum_{x \in X} f(x) \varphi_i(x) \text{ (4)}$$

The fluorescence of each combinatorial mutant can be calculated based on the Bahadur coefficients and z string.

$$f(x) = \sum_{i=0}^{2n-1} w_i \varphi_i(x) \text{ (5)}$$

The R^2 between the modeled fluorescence values and the experimental data is 1.0 when all interaction terms are included in the expansion. By truncating Eq. (5) to contain only low-order interactions, the effect of these contributions to the model can be determined. The expansion was applied to the full set of mutations (4 positions) and modeled using first order terms; first and second order terms; first, second, and third order terms; and all terms (Supplementary Fig. 18). An identical approach was applied to all 24 subnetworks and utilized only first order terms in the reconstruction (Supplementary Fig 19).

Errors in the R^2 statistics were estimated using a Monte Carlo simulation. 500 sets of fluorescence values for all mutants were sampled based on experimental fluorescence means and standard deviations following a Gaussian distribution using the NumPy module in Python 2.7[45,46]. Eq. (4) and (5) were applied to reconstruct the fluorescence values and calculate R^2 values between the sampled model and the sampled data to give a distribution of R^2 values. Bias-corrected adjusted 95% confidence intervals were calculated by obtaining the average R^2 of 10,000 bootstrap iterations of the Monte Carlo simulation R^2. The bahadur expansion was applied to each functional parameter.

A control set of additive data was used to calculate the R^2 of data showing no epistasis (Supplementary Table 1). This subnetwork was analyzed using the same approach as the subnetwork workflow.

Protein characterization:

Purifying proteins for isothermal titration calorimetry:

The TtgR gene for variants 0000, 1000, 0100, 1100, and 1111 were cloned into a pET31B vector downstream of the T7 promoter for lac-inducible transcription control using isothermal assembly with 0.18pmol backbone and 0.392pmol TtgR. MBP was amplified with primers to add a C-terminal His-tag and TEV site and inserted into the TtgR-pET31B vector upstream of TtgR to create a MBP-His-TtgR fusion with a TEV cleavage site between the His-tag and the TtgR protein. BL21 chemically competent cells (NEB) were transformed with 20ng of pET31B vector. Dilutions of transformants were plated on LB-ampicillin agar. A colony was selected and grown in 5mL LB-ampicillin media shaking at 37°C for 16 hours. This culture was added to 500mL autoinduction media (Terrific Broth, 0.8% glycerol, 2mM $MgSO_4$, 0.375% (w/v) aspartic acid, 0.015% (w/v) glucose, 0.5% (w/v) lactose) and grown for 8 hours at 37°C shaking. The culture was grown for an additional 16 hours at 25°C shaking.

The cells were spun down at 5,500*g* for 15 minutes at 4°C. The supernatant was removed and the cells were resuspended in a lysis buffer (300mM NaCl, 50mM HEPES, 1mM PMSF, 1mg/mL Lysozyme, 5mM BME, 10% glycerol, pH 7.5). A Q500 sonicator (Qsonica) was used to lyse cells using a 5 second on, 15 second off sonication protocol for 4 minutes total sonication time. The lysate was centrifuged at 14,000*g* for 45 minutes at 4°C. The supernatant was isolated and filtered through a 0.22μm filter. The filtered supernatant was purified on an Akta Start using 2 5mL HisTrap HP columns. The column was washed with 5 column volumes (CV) IMAC-A (500mM

NaCl, 20mM Imidazole, 20mM MOPS, 0.3mM TCEP, pH 7). MBP-6His-TtgR was eluted with a gradient of 100% IMAC-A to 100% IMAC-B (500mM NaCl, 500mM Imidazole, 20mM MOPS, 0.3mM TCEP, pH7) over 5CV and collected in 2mL fractions. Fractions with the highest absorbance at 280nm (A280) were combined and dialyzed in 8L of dialysis buffer A (100mM NaCl, 20mM MOPS, 0.3mM TCEP, pH 7.5). TEV was added to the proteins prior to dialysis at a ratio of 1:50 w/w TEV:TtgR. Dialysis occurred over a 16 hour interval at 4°C while stirring at low speed. Dialyzed protein was centrifuged at 14,000*g* for 10 minutes at 4°C. The supernatant was passed through a 0.22μm filter and loaded onto the HisTrap columns at 5mL/min. The column was washed with 5CV of IMAC-A and 2mL fractions were collected. 5CV of IMAC-B was used to remove the MBP-6His from the column. The column was washed with an additional 10CV IMAC-A. Wash fractions with high A280 were combined and reapplied to the column. The column was washed with 5CV of IMAC-A and 2mL fractions were collected. 5CV of IMAC-B was used to strip the MBP-6His from the column. Fractions with high A280 were combined and dialyzed in 4L of dialysis buffer C (100mM NaCl, 20mM MOPS, 10mM MgCl2, 0.3mM TCEP, pH 7.8). The protein was centrifuged at 14,000*g* for 10 minutes at 4°C. The supernatant was passed through a 0.22μm filter. The protein was concentrated to approximately 9mg/mL and frozen in 60μL aliquots in liquid nitrogen before storing at -80°C. Dialysis buffer C was passed through a 0.22μm filter and stored at 4°C for ITC experiments.

Determining binding affinity of TtgR variants:

Stocks of 250mM naringenin and 100mM resveratrol were diluted to 500μM and 250μM, respectively, in dialysis buffer C. Aliquots of TtgR were thawed on ice and diluted to a final concentration of 7.5μM. DMSO or ethanol was added to the TtgR solution to match the solution composition of the naringenin or resveratrol dilutions. An aliquot of dialysis buffer C was also prepared with DMSO or ethanol for a control injection and to wash the sample cell between ITC injections.

The ITC experiments were conducted on a VP-ITC (MicroCal). An initial control injection scheme consisted of loading the sample cell with dialysis buffer C and performing a series of 10 10μL ligand injections with 10 minute intervals at 25°C. The sample cell was washed 5 times with dialysis buffer C before the 7.5μM protein solution was loaded. 25 10μL naringenin injections or 28 10μL resveratrol injections occurred in 10 minute intervals at 25°C.

Data analysis was primarily conducted using Origin 7.0 (MicroCal). The heats of injection from the control sample were averaged. The protein-ligand injection profile was subtracted by this average heat prior to curve fitting. Due to low affinity for both naringenin and resveratrol, the stoichiometry of binding was fixed to 1 to reduce the degrees of freedom prior to fitting. The curves were fit with the single binding site model (Supplementary Fig. 20).

X-ray crystallography:

Purifying Proteins for X-ray crystallography:

TtgR-pET31B vector was electroporated into BL21 cells (NEB) and recovered in 1mL SOC. The cells were incubated for 1 hour at 37°C before serial dilutions were plated on LB-ampicillin (100μg/mL) plates. A single colony was selected and incubated in 5mL LB-ampicillin (100μg/mL) at 37°C shaking for 3 hours. The 5mL culture was added to 500mL LB-ampicillin media and incubated at 37°C shaking at 250rpm for approximately 3 hours until the OD_{600} reached 0.6. The culture was induced with 100μM IPTG followed by an incubation at 16°C for 16 hours shaking at 250rpm.

The cells were spun down at 5,500*g* for 15 minutes at 4C. The supernatant was removed and the cells were resuspended in a lysis buffer (300mM NaCl, 50mM HEPES, 1mM PMSF, 1mg/mL

Lysozyme, 5mM BME, 10% glycerol, pH 7.5). A Q500 sonicator (Qsonica) was used to lyse cells using a 25 second on, 50 second off sonication protocol for 3 minutes and 45 seconds total sonication time. The lysate was centrifuged at 14,000*g* for 45 minutes at 4°C. The supernatant was isolated and filtered through a 0.22μm filter. The filtered supernatant was purified on an Akta Start (Cytiva) using a 5mL HisTrap HP columns (Cytiva). The supernatant was loaded onto the column at a flow rate of 5mL/min. The column was washed with 5 column volumes (CV) IMAC-A. MBP-6His-TtgR was eluted with a gradient of 100% IMAC-A to 100% IMAC-B over 10CV and collected in 2mL fractions. Fractions with the highest absorbance at 280nm (A280) were combined and dialyzed in 8L of dialysis buffer A. TEV was added to the proteins prior to dialysis at a ratio of 1:50 w/w TEV:TtgR. Dialysis occurred over a 16 hour interval at 4°C while stirring at low speed.

TtgR was isolated from MBP-6His through a subtractive IMAC protocol using the Akta Start and 5mL HisTrap HP column. The dialyzed protein was centrifuged at 4,000*g* for 10 minutes at 4C. Supernatant was passed through a 0.22μm filter and applied to the HisTrap column at 5mL/min. 5CV IMAC-A was used to wash the column while 2mL fractions were collected. 2.5CV IMAC-B was used to remove the MBP from the column and 5mL fractions were collected. Wash fractions with high A280 were combined and dialyzed in 4L of dialysis buffer B (50mM NaCl, 5mM MOPS, 0.3mM TCEP, pH 7.5). EDTA was added to the protein wash fractions to a final concentration of 10mM prior to dialysis. Dialysis occurred over a 16 hour interval at 4C while stirring at low speed. TtgR was concentrated to 10mg/mL using spin concentrators. Samples were spun at intervals of 3,500*g* for 5 minutes and mixed via pipette between spins. Concentrated TtgR was separated into 60μL aliquots and frozen in liquid nitrogen prior to storage at -80°C.

Size exclusion chromatography:

Samples of TtgR wildtype and mutant proteins were received frozen in 5 mM MOPS, pH 7.4, 50 mM NaCl, 0.3 mM TCEP. Samples were thawed and centrifuged for 5 minutes at 21,130*g*. Sample supernatants were filtered with a 0.22 micron MillexGV syringe filter unit (Millipore) before applying to an equilibrated 10 mm x 300 mm Superdex 200 column (GE Healthcare). Chromatography was performed on a GE AKTA FPLC system. Column buffer was 20 mM HEPES, pH 7.5, 350 mM NaCl, 0.3 mM TCEP. Two primary peaks were obtained from each sample with major peak at approximately 45kD MW and a minor peak at approximately 79kD. The fractions corresponding to the major peak were pooled and concentrated with an Amicon Ultracel-10 centrifugal filter device (Millipore) and dialyzed vs. 5 mM HEPES, pH 7.5, 50 mM NaCl, 0.3 mM TCEP. Samples collected after dialysis were divided into small aliquots and flash frozen in PCR tubes with liquid nitrogen.

Crystallization screening and optimization:

Crystallization screening and optimization was conducted in the Collaborative Crystallography Core in the Department of Biochemistry and the University of Wisconsin-Madison. Crystallization experiments were set up using a SPT Labtech mosquito® crystallization robot in MRC SD-2 crystallization plates at 4°C and 20°C (277 and 293 K.) Crystals progressing to diffraction experiments were all obtained at 20°C. Two general screens, Hampton Research IndexHT and Molecular Dimensions JCSG+ were used in this study[47]. Crystals were detected using brightfield and UV fluorescence imaging with a JANSi UVEX-P crystallization plate imaging system supplementing visual inspection with stereomicroscopes. Initial rounds of crystallization optimization were performed in SD2 plates using the mosquito to expand 24 solution conditions by setting columns of experiments in four different sample to reservoir volume ratios.

Cryoprotected crystals were harvested in Mitegen micro mounts and flash cooled by immersion in liquid nitrogen.

Crystallography:

Crystals were screened and X-ray diffraction data was collected at Advanced Photon Source (APS) beamlines LS-CAT and GM/CA@APS, universally on crystals cooled to 100K. Diffraction data was reduced using XDS and scaled with XSCALE[48,49]. Structures were solved by molecular replacement with Phaser within the Phenix suite of programs, automatically rebuilt with phenix.autobuild, iteratively improved with alternating rounds of rebuilding in Coot and refinement using phenix.refine, and validated using MOLPROBITY[50-54].

7K1A crystals providing diffraction data were grown by mixing 200 nL of protein at 9.7 mg/mL in sample buffer (5mM HEPES pH 7.5, 50 mM NaCl, 0.3 mM TCEP) with 150 nL of reservoir solution, was equilibrated against 150 nL 20% MEPEG, 0.2M MgCl2, 0.1M bistris HCL pH 6.5 equilibrated against 50 microliters of reservoir solution in a SD2 plate. Samples were cryoprotected with reservoir solution supplemented to 35% MEPEG 2000. A 360° sweep of data (720 frames) was collected on a MAR 300 CCD detector at LS-CAT beamline 21ID-G on 2018-12-16 using 0.97856 Å X-rays. The phase problem was solved using 2UXU(A) as a molecular replacement model[28].

7K1C crystals of wild-type TtgR with resveratrol were prepared by incubating 0.41 mM protein (9.8 mg/mL) and 0.5 mM resveratrol dissolved in sample buffer for 30 minutes at room temperature prior to setting up crystallization experiments. The crystal yielding the best diffraction data was grown by mixing 200 nL of the protein-ligand sample with 250 nL reservoir (18%

PEG4000, 0.2M MgCl2, 0.1M bistris HCl pH 6.5) equilibrated against 50 microliters of reservoir a SD2 plate. Samples were cryoprotected with reservoir solution supplemented with 35% PEG4000. A 360° sweep of data (720 frames) was collected on a MAR 300 CCD detector at LS-CAT beamline 21ID-G on 2018-12-16 using 0.97856 Å X-rays. The phase problem was solved using 2XDN as a molecular replacement model.

7KD8 crystals were prepared by incubating 0.43 mM (10.4 mg/mL) quadruple mutant protein with 1 mM resveratrol in sample buffer for 30 minutes prior to setting up crystallization experiments. Crystals providing the reported diffraction data set grew from 2 microliters of sample mixed with 2 microliters of reservoir solution (12% MEPEG 2000, 5% 2-methyl-2,4-pentanediol, 0.3 M MgCl2, 0.1 M bistris buffer at pH 6.5 equilibrated in a hanging drop experiment using a siliconized glass cover slip. Samples were cryoprotected with reservoir solution supplemented to 30% MEPEG 2000. A 360°(3600 frames) shutterless data set was collected at LS-CAT 21ID-D on 2019-05-30 with an Eiger 9M direct detector and 1.07812 Å X-rays. The phase problem was solved using 7K1A as a molecular replacement model.

Figures and scripts:

All figures were generated using the Matplotlib module in Python 2.7[55]. Scripts used in data analysis and figure generation can be found at: https://github.com/raman-lab/epistasis. POVME 3.0 was used to calculate pocket volumes based on the location of resveratrol[56].

References

1 Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535-538, doi:10.1038/nature11510 (2012).

2 Smith, J. M. Natural Selection and the Concept of a Protein Space. *Nature* **225**, 2 (1970).

3 Lunzer, M., Golding, G. B. & Dean, A. M. Pervasive Cryptic Epistasis in Molecular Evolution. *PLOS Genetics* **6**, e1001162, doi:10.1371/journal.pgen.1001162 (2010).

4 Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* **2**, e00631, doi:10.7554/eLife.00631 (2013).

5 Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife* **5**, doi:10.7554/eLife.16965 (2016).

6 Miton, C. M., Buda, K. & Tokuriki, N. Epistasis and intramolecular networks in protein evolution. *Curr Opin Struct Biol* **69**, 160-168, doi:10.1016/j.sbi.2021.04.007 (2021).

7 Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci* **25**, 1204-1218, doi:10.1002/pro.2897 (2016).

8 McKeown, A. N. *et al.* Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell* **159**, 58-68, doi:10.1016/j.cell.2014.09.003 (2014).

9 Anderson, D. W., McKeown, A. N. & Thornton, J. W. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *Elife* **4**, e07864, doi:10.7554/eLife.07864 (2015).

10 Dickinson, B. C., Leconte, A. M., Allen, B., Esvelt, K. M. & Liu, D. R. Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. *Proc Natl Acad Sci U S A* **110**, 9007-9012, doi:10.1073/pnas.1220670110 (2013).

11 Podgornaia, A. I. & Laub, M. T. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 5 (2015).

12 Kaltenbach, M., Jackson, C. J., Campbell, E. C., Hollfelder, F. & Tokuriki, N. Reverse evolution leads to genotypic incompatibility despite functional and active site convergence. *Elife* **4**, doi:10.7554/eLife.06492 (2015).

13 Lunzer, M., Miller, S. P., Felsheim, R. & Dean, A. M. The Biochemical Architecture of an Ancient Adaptive Landscape. *Science* **310**, 3 (2005).

14 Wilson, C. *et al.* Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* **347**, 5 (2015).

15 Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409-413, doi:10.1038/nature23902 (2017).

16 Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol* **24**, 2643-2651, doi:10.1016/j.cub.2014.09.072 (2014).

17 Hart, K. M. *et al.* Thermodynamic System Drift in Protein Evolution. *PLoS Biol* **12**, 8 (2014).

18 Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A. & Sanchez-Ruiz, J. M. Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian beta-lactamases. *J Am Chem Soc* **135**, 2899-2902, doi:10.1021/ja311630a (2013).

19 Patel, M. P. *et al.* Synergistic effects of functionally distinct substitutions in beta-lactamase variants shed light on the evolution of bacterial drug resistance. *J Biol Chem* **293**, 17971-17984, doi:10.1074/jbc.RA118.003792 (2018).

20 Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929-932, doi:10.1038/nature05385 (2006).

21 Tomatis, P. E. *et al.* Adaptive protein evolution grants organismal fitness by improving catalysis and flexibility. *PNAS* **105**, 6 (2008).

22 Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *PNAS* **103**, 5 (2006).

23 Wodak, S. J. *et al.* Allostery in Its Many Disguises: From Theory to Applications. *Structure* **27**, 566-578, doi:10.1016/j.str.2019.01.003 (2019).

24 Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* **508**, 331-339, doi:10.1038/nature13001 (2014).

25 Eick, G. N., Colucci, J. K., Harms, M. J., Ortlund, E. A. & Thornton, J. W. Evolution of minimal specificity and promiscuity in steroid hormone receptors. *PLoS Genet* **8**, e1003072, doi:10.1371/journal.pgen.1003072 (2012).

26 Lisi, G. P., East, K. W., Batista, V. S. & Loria, J. P. Altering the allosteric pathway in IGPS suppresses millisecond motions and catalytic activity. *Proc Natl Acad Sci U S A* **114**, E3414-E3423, doi:10.1073/pnas.1700448114 (2017).

27 Skerker, J. M. *et al.* Rewiring the specificity of two-component signal transduction systems. *Cell* **133**, 1043-1054, doi:10.1016/j.cell.2008.04.040 (2008).

28 Alguel, Y. *et al.* Crystal structures of multidrug binding protein TtgR in complex with antibiotics and plant antimicrobials. *J Mol Biol* **369**, 829-840, doi:10.1016/j.jmb.2007.03.062 (2007).

29 Teran, W. *et al.* Antibiotic-Dependent Induction of Pseudomonas putida DOT-T1E TtgABC Efflux Pump Is Mediated by the Drug Binding Repressor TtgR. *Antimicrobial Agents and Chemotherapy* **47**, 3067-3072, doi:10.1128/aac.47.10.3067-3072.2003 (2003).

30 Daniels, C., Daddaoua, A., Lu, D., Zhang, X. & Ramos, J. L. Domain cross-talk during effector binding to the multidrug binding TTGR regulator. *J Biol Chem* **285**, 21372-21381, doi:10.1074/jbc.M110.113282 (2010).

31 Teran, W., Krell, T., Ramos, J. L. & Gallegos, M. T. Effector-repressor interactions, binding of a single effector molecule to the operator-bound TtgR homodimer mediates derepression. *J Biol Chem* **281**, 7102-7109, doi:10.1074/jbc.M511095200 (2006).

32 Fleishman, S. J. *et al.* RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One* **6**, e20161, doi:10.1371/journal.pone.0020161 (2011).

33 Xiong, D. *et al.* Improving key enzyme activity in phenylpropanoid pathway with a designed biosensor. *Metab Eng* **40**, 115-123, doi:10.1016/j.ymben.2017.01.006 (2017).

34 Leaver-Fay, A. *et al.* Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol* **523**, 109-143, doi:10.1016/B978-0-12-394292-0.00006-0 (2013).

35 Taylor, N. D. *et al.* Engineering an allosteric transcription factor to respond to new ligands. *Nat Methods* **13**, 177-183, doi:10.1038/nmeth.3696 (2016).

36 Matsuura, T., Kazuta, Y., Aita, T., Adachi, J. & Yomo, T. Quantifying epistatic interactions among the components constituting the protein translation system. *Mol Syst Biol* **5**, 297, doi:10.1038/msb.2009.50 (2009).

37 Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, 545-574, doi:10.1016/B978-0-12-381270-4.00019-6 (2011).

38 Nivon, L. G., Moretti, R. & Baker, D. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLoS One* **8**, e59004, doi:10.1371/

10.1371/journal.pone.0059004.g001 (2013).

39 Siegel, J. B. *et al.* Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **329**, 309-313 (2010).

40 Olins, P. O., Devine, C. S., Rangwala, S. H. & Kavka, K. S. The T7 phage gene 10 leader RNA, a ribosome-binding site that dramatically enhances the expression of foreign genes in Escherichia coli *Gene* **73**, 9 (1988).

41 Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6**, 343-345, doi:10.1038/nmeth.1318 (2009).

42 Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci U S A* **110**, 14024-14029, doi:10.1073/pnas.1301301110 (2013).

43 Hill, A. V. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J Physiol (Lond.)* **40**, iv-vii (1910).

44 Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261-272, doi:10.1038/s41592-019-0686-2 (2020).

45 Oliphant, T. E. *A guide to NumPy*. (Trelgol Publishing, 2006).

46 Walt, S. v. d., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* **13**, 22-30, doi:10.1109/MCSE.2011.37 (2011).

47 Page, R. *et al.* Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the Thermotoga maritima proteome. *Acta Crystallographica Section D* **59**, 1028-1037, doi:doi:10.1107/S0907444903007790 (2003).

48 Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125-132 (2010).

49 Diederichs, K., McSweeney, S. & Ravelli, R. B. G. Zero-dose extrapolation as part of macromolecular synchrotron data reduction. *Acta Crystallographica Section D* **59**, 903-909, doi:doi:10.1107/S0907444903006516 (2003).

50 Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallographica Section D: Structural Biology* **75**, 861-877 (2019).

51 Terwilliger, T. SOLVE and RESOLVE: automated structure solution, density modification and model building. *Journal of synchrotron radiation* **11**, 49-52 (2003).

52 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography* **66**, 486-501 (2010).

53 Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix. refine. *Acta Crystallographica Section D: Biological Crystallography* **68**, 352-367 (2012).

54 Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **66**, 12-21 (2010).

55 Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90-95, doi:10.1109/MCSE.2007.55 (2007).

56 Wagner, J. R. *et al.* POVME 3.0: Software for Mapping Binding Pocket Flexibility. *Journal of Chemical Theory and Computation* **13**, 4584-4592, doi:10.1021/acs.jctc.7b00500 (2017).

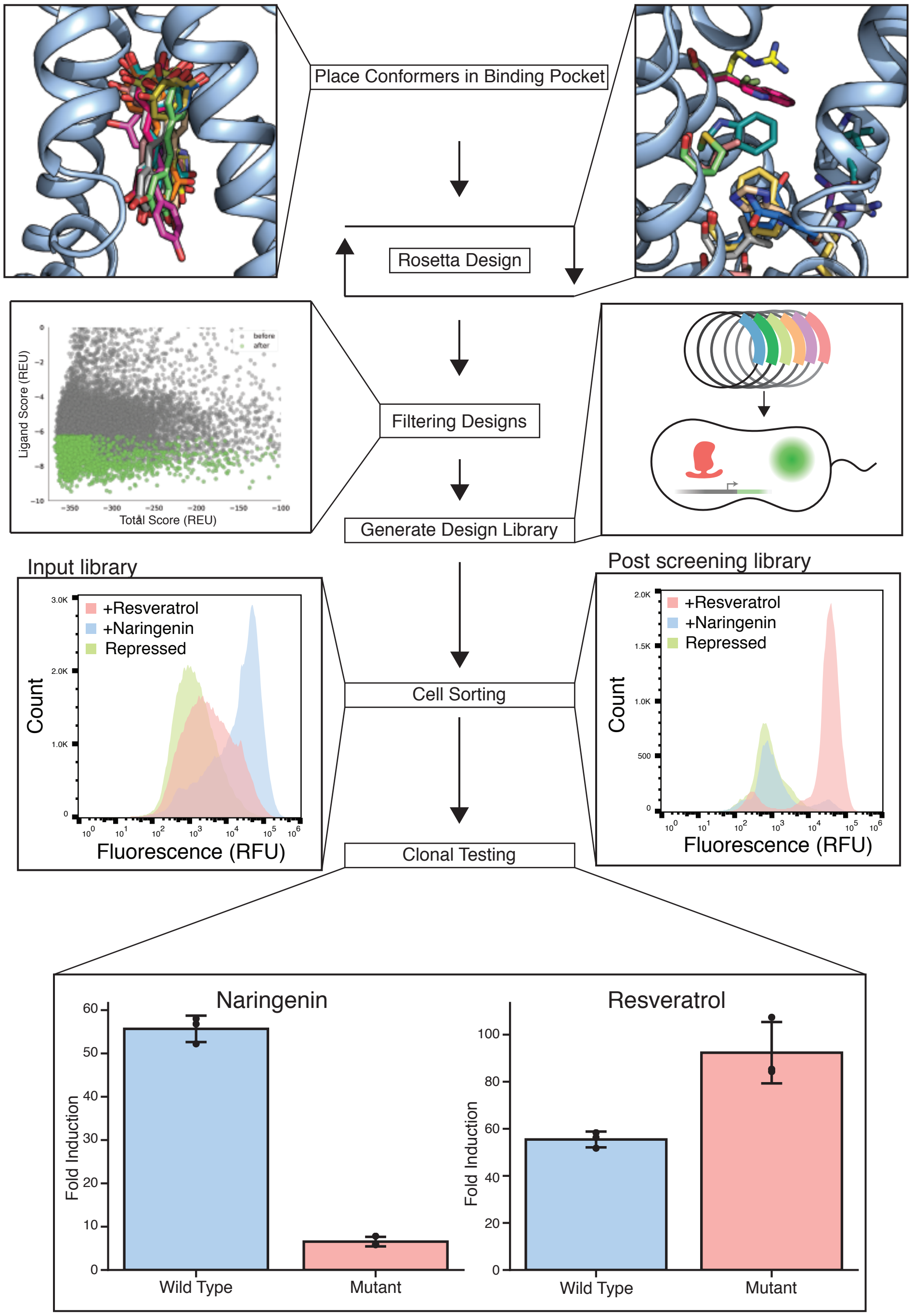


Figure 1: Design of resveratrol-specific TtgR variant. Resveratrol conformers are docked into TtgR followed by Rosetta-based computational design of the binding pocket. Candidates with favorable Rosetta score metrics (green points) are synthesized and cloned into an expression vector. Distribution of fluorescence in cells containing uninduced TtgR variant library (light green), induced with naringenin (light blue) and resveratrol (red) before sorting (Pre-Sort) and after three rounds of sorting (Post-Sort) are shown. Colony screening identified a quadruple mutant showing resveratrol specificity: C137I/I141W/M167L/F168Y.

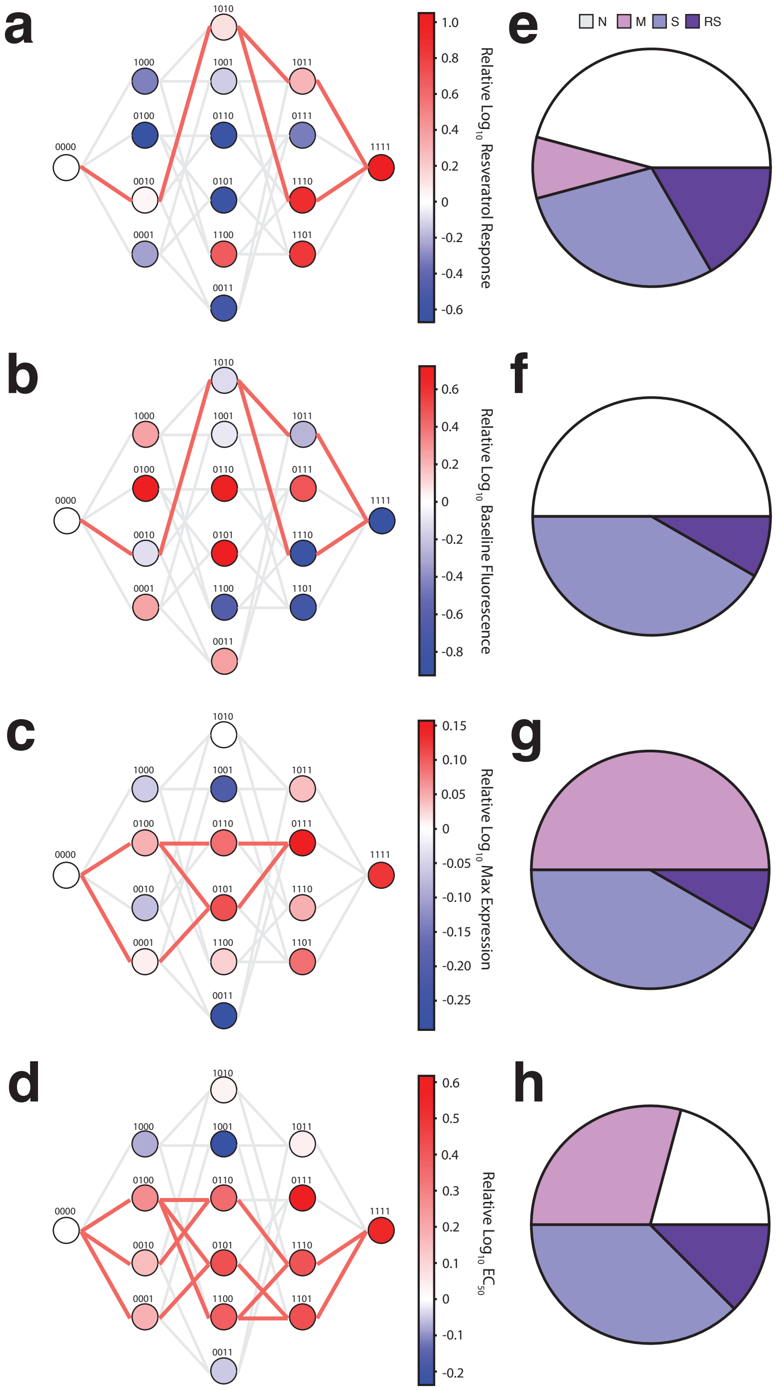


Figure 2: Fitness landscapes for multiple functional parameters in response to induction with resveratrol. Fitness landscapes of ***(a)*** fold induction, ***(b)*** basal gene expression, ***(c)*** maximum gene expression, and ***(d)*** EC$_{50}$ parameters for all 16 TtgR variants in response to resveratrol with each variant shown as a node in the graph. Each variant is labeled with a binary string corresponding to the presence (1) or absence (0) of a mutation at position 137, 141, 167, or 168 in order. Nodes separated by a single mutation are connected by edges showing viable (bold red) and unviable paths (light gray) through sequence space. Nodes are shaded by log$_{10}$ of the fold induction ratio at 250µM resveratrol normalized to the fold induction ratio of wildtype TtgR. Number of epistatic subnetworks in the resveratrol ***(d)*** fold induction, ***(e)*** basal gene expression, ***(f)*** maximum gene expression, and ***(g)*** EC50 landscape determined by Bahadur expansion.

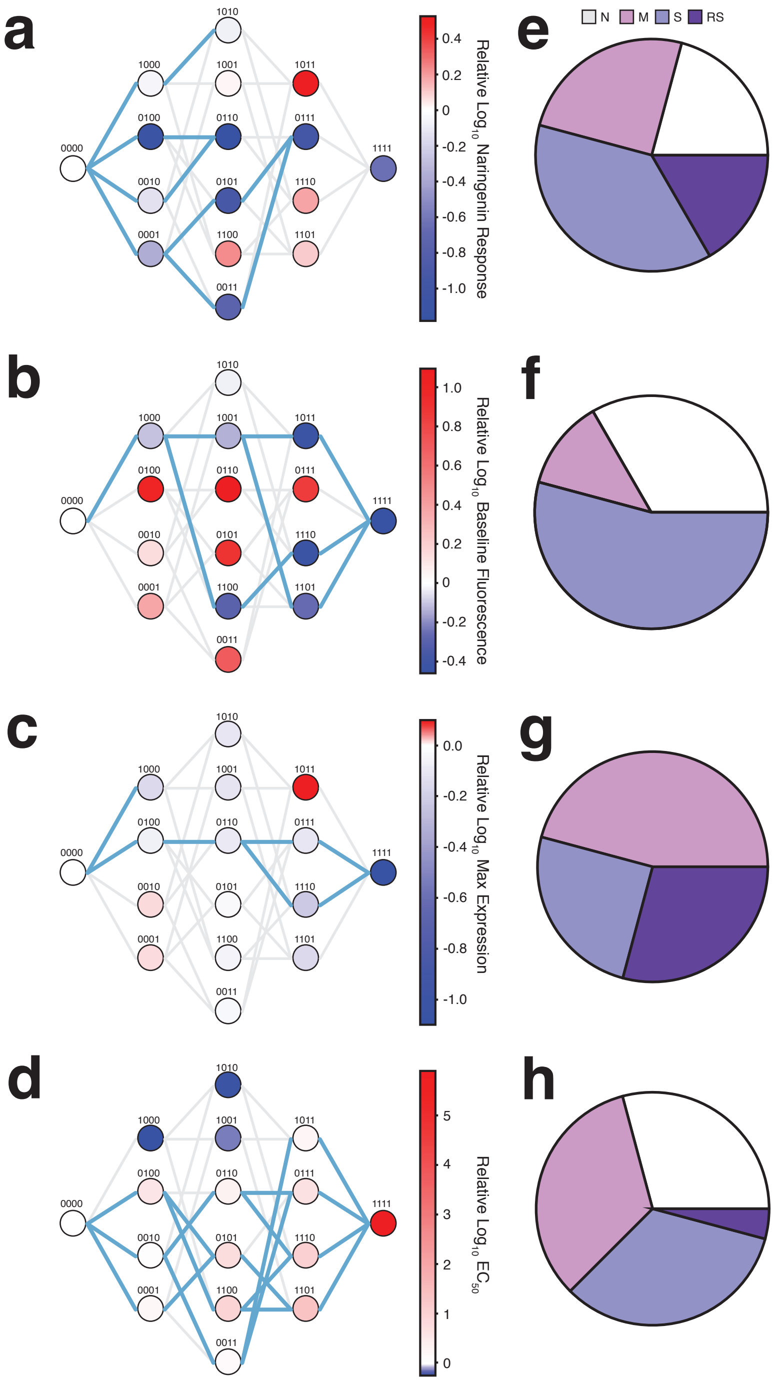


Figure 3: Fitness landscapes for multiple functional parameters in response to induction with naringenin. Fitness landscapes of ***(a)*** fold induction, ***(b)*** basal gene expression, ***(c)*** maximum gene expression, and ***(d)*** EC$_{50}$ parameters for all 16 TtgR variants in response to naringenin with each variant shown as a node in the graph. Each variant is labeled with a binary string corresponding to the presence (1) or absence (0) of a mutation at position 137, 141, 167, or 168 in order. Nodes separated by a single mutation are connected by edges showing viable (bold blue) and unviable paths (light gray) through sequence space. Nodes are shaded by log$_{10}$ of the fold induction ratio at 2000μM resveratrol normalized to the fold induction ratio of wildtype TtgR. Number of epistatic subnetworks in the resveratrol ***(e)*** fold induction, ***(f)*** basal gene expression, ***(g)*** maximum gene expression, and ***(h)*** EC$_{50}$ landscape determined by Bahadur expansion.

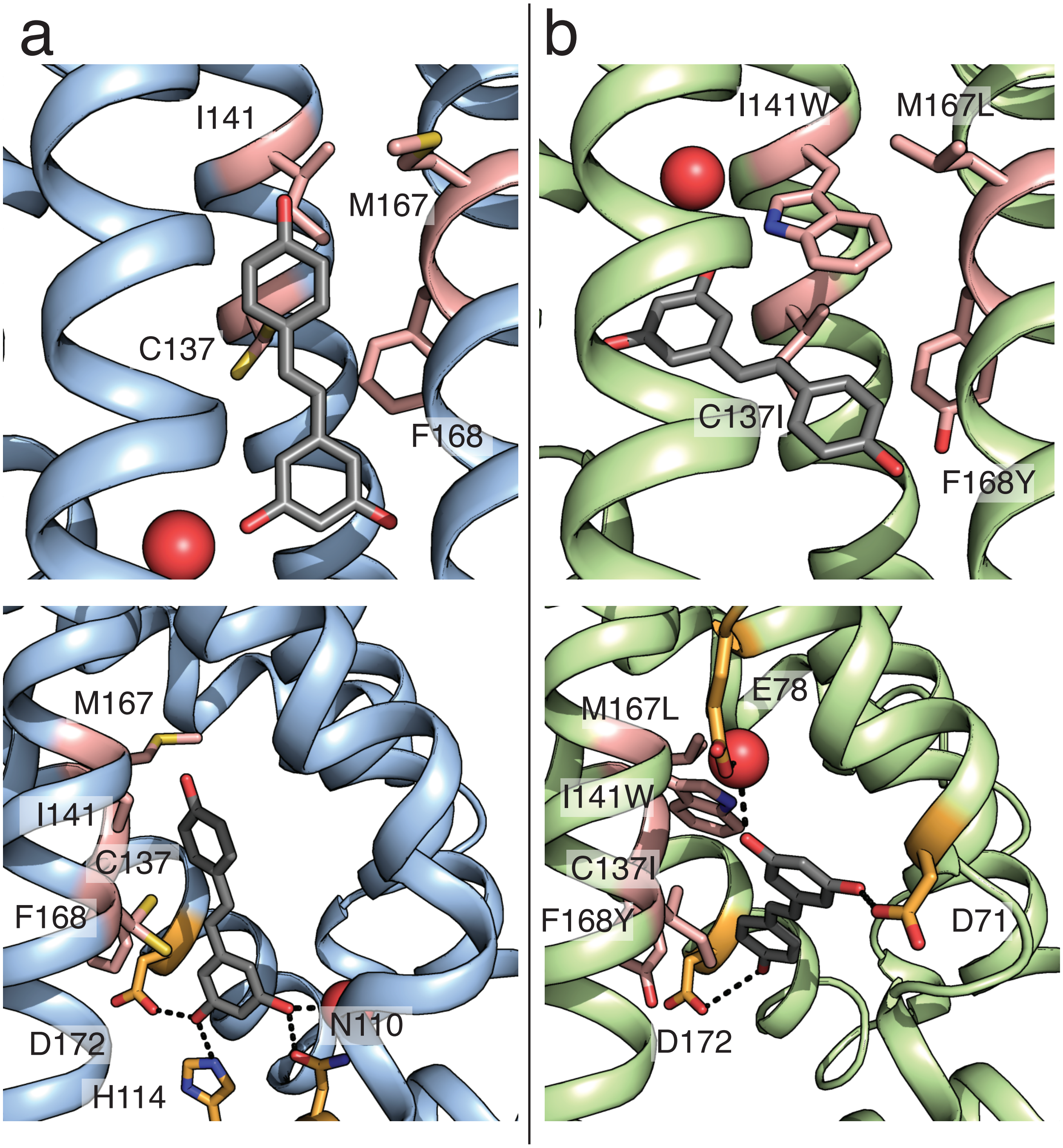


Figure 4: Structural basis for ligand specificity. Wildtype TtgR and quadruple mutant are shown in blue and green ribbons, respectively. Positions 137, 141, 167, and 168 are colored in pink. Resveratrol is shown as gray sticks. Water molecules are shown as red spheres. ***(a)*** Binding pocket of resveratrol-bound wildtype TtgR (PDB ID: 7K1C) (upper panel) with residues making hydrogen bonds to resveratrol highlighted in orange (lower panel). ***(b)*** Binding pocket of resveratrol-bound quadruple mutant TtgR (PDB ID: 7KD8) (upper panel) with residues making hydrogen bonds to resveratrol highlighted in orange (lower panel).

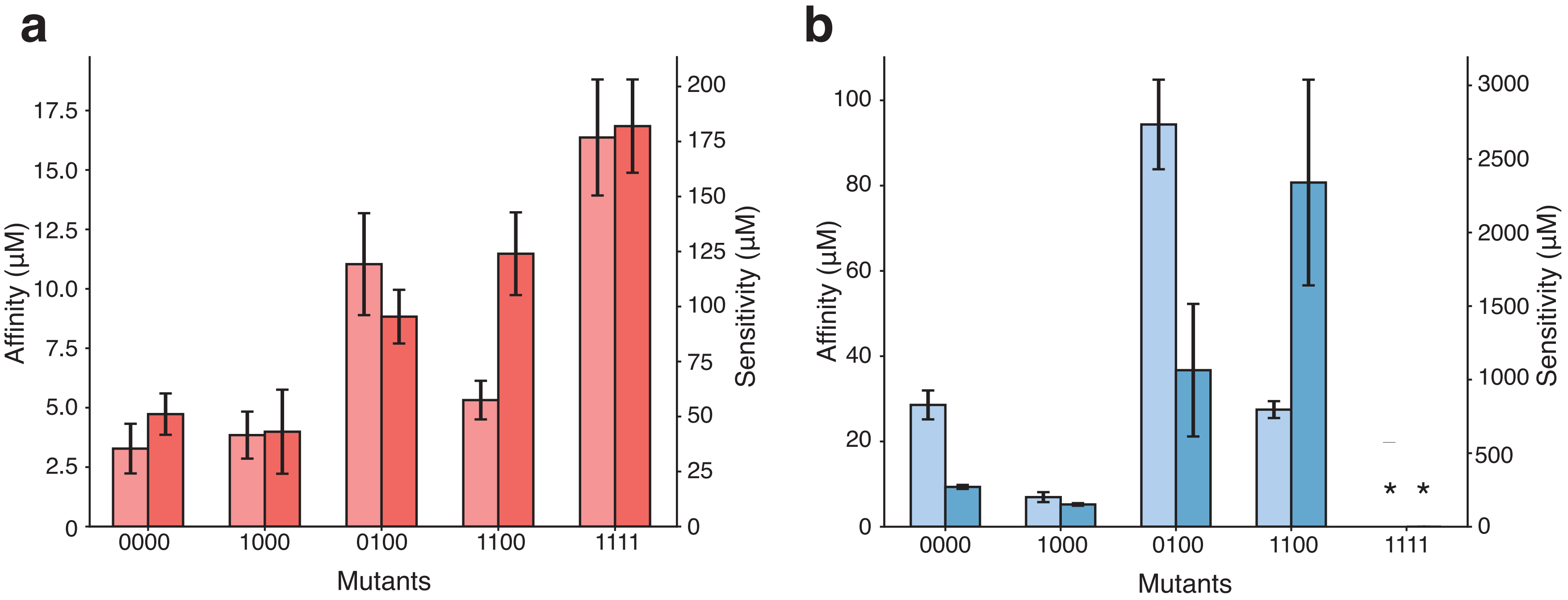
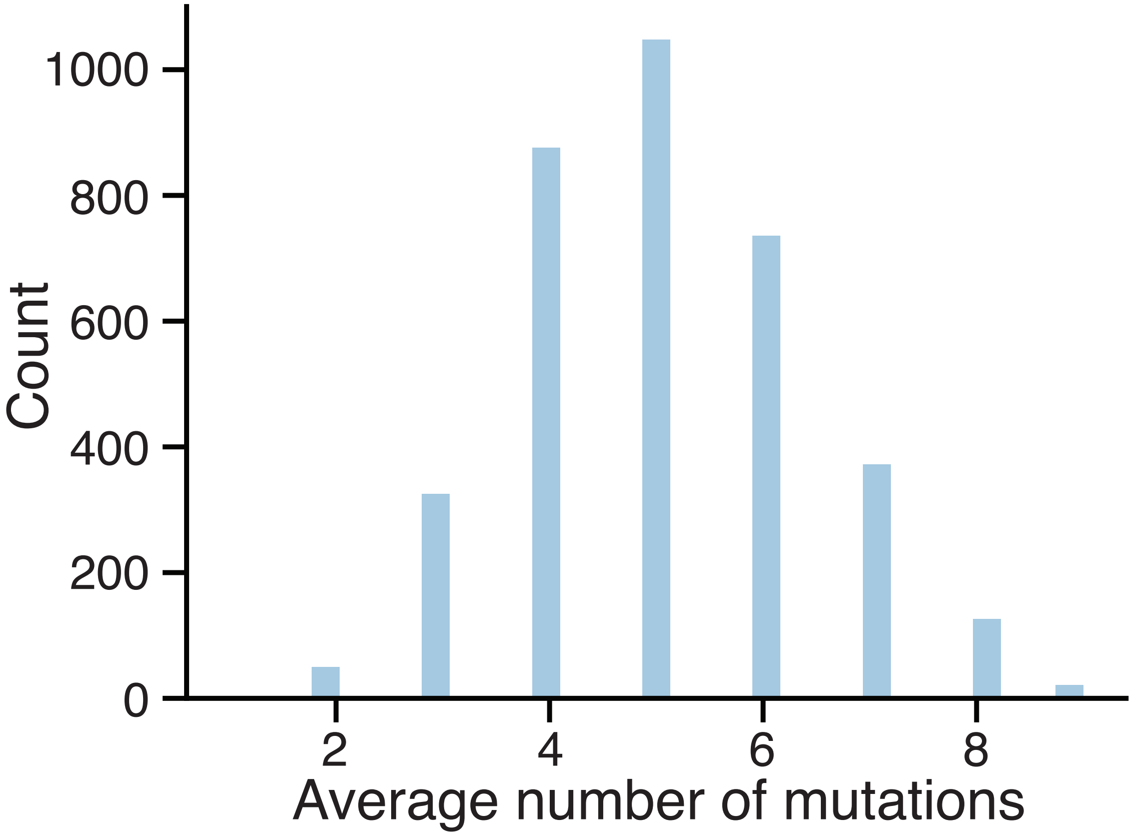
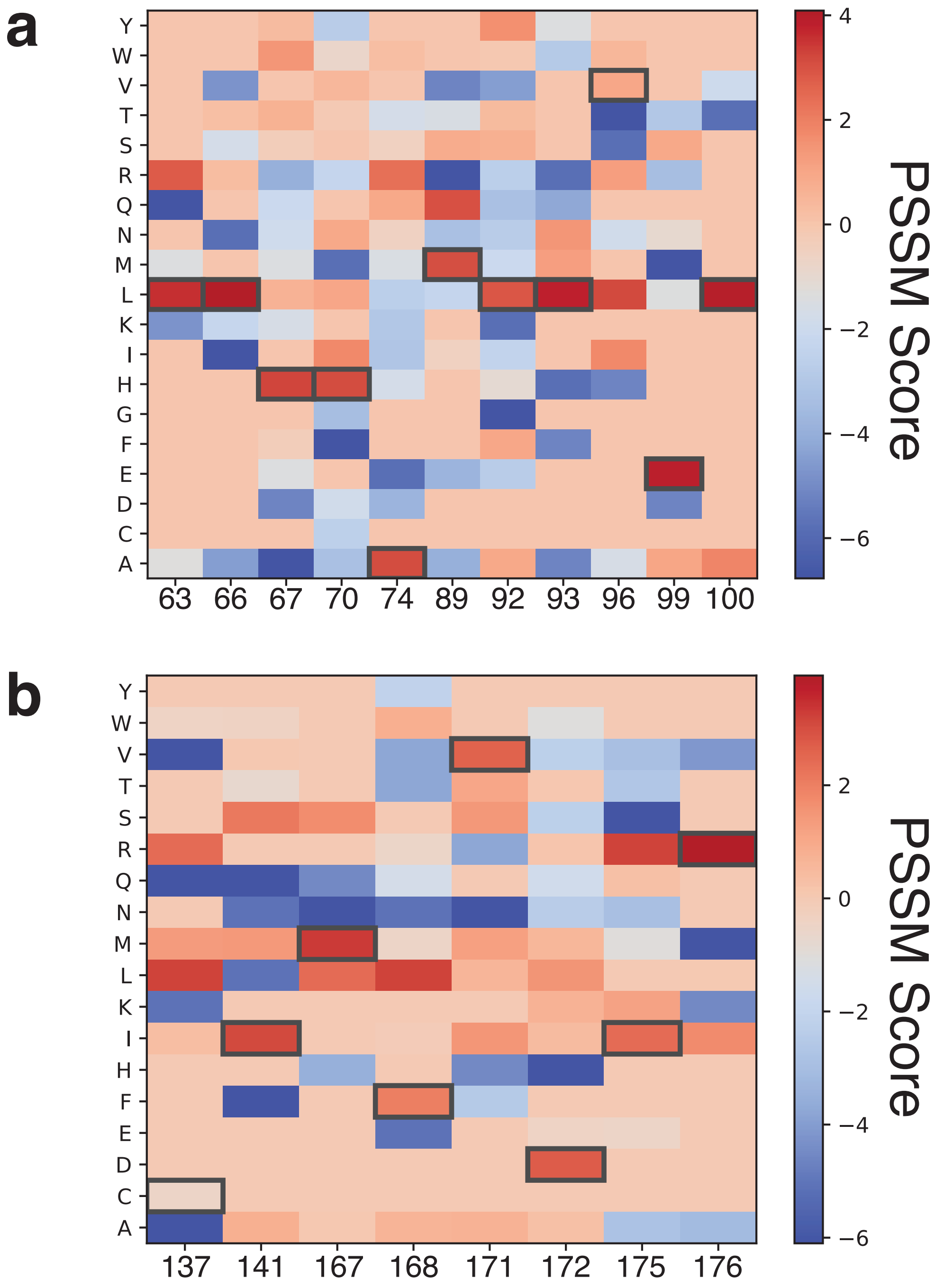


Figure 5: Comparison of biophysical and biological properties of TtgR variants. Ligand affinity (light bar) and EC_{50} sensitivity (dark bar) for resveratrol ***(a)*** and naringenin ***(b)*** are shown for TtgR variants 0000, 1000, 0100, 1100, and 1111. Ligand affinity was determined by isothermal calorimetry and EC_{50} sensitivity from fitting dose response curves to the Hill equation. EC_{50} values and error are calculated based on fitting to triplicate dose response curves. ITC values and error are generated from a one-site binding model (see methods).

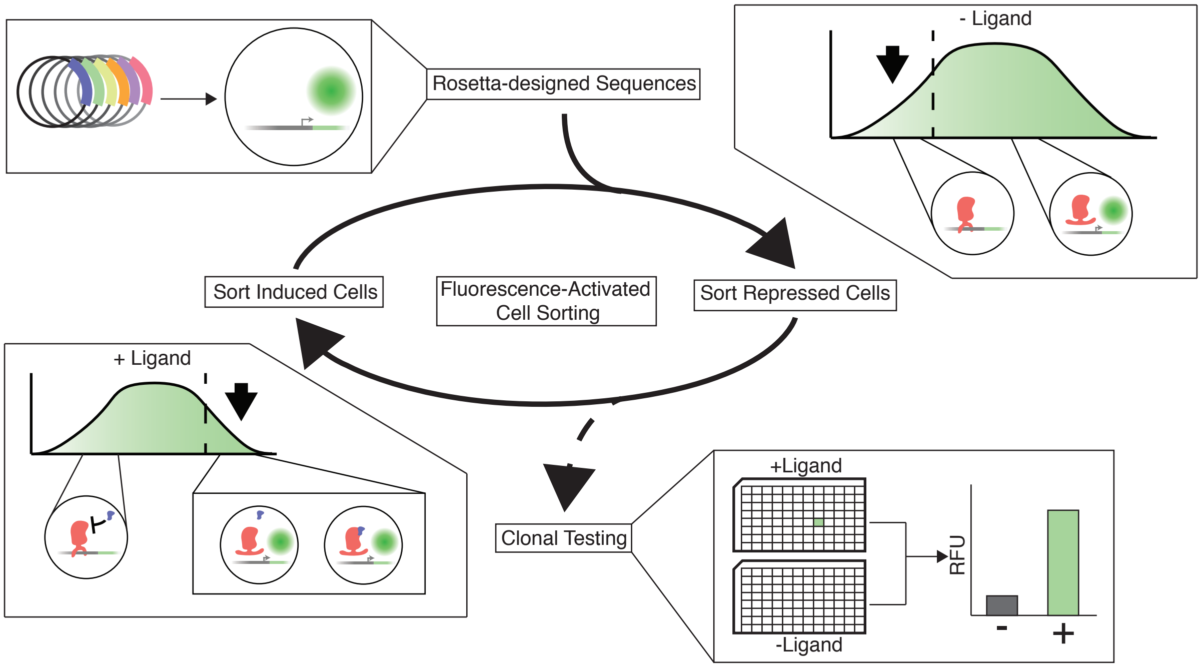


Supplementary Figure 1: Mutation distribution for synthesized resveratrol designs

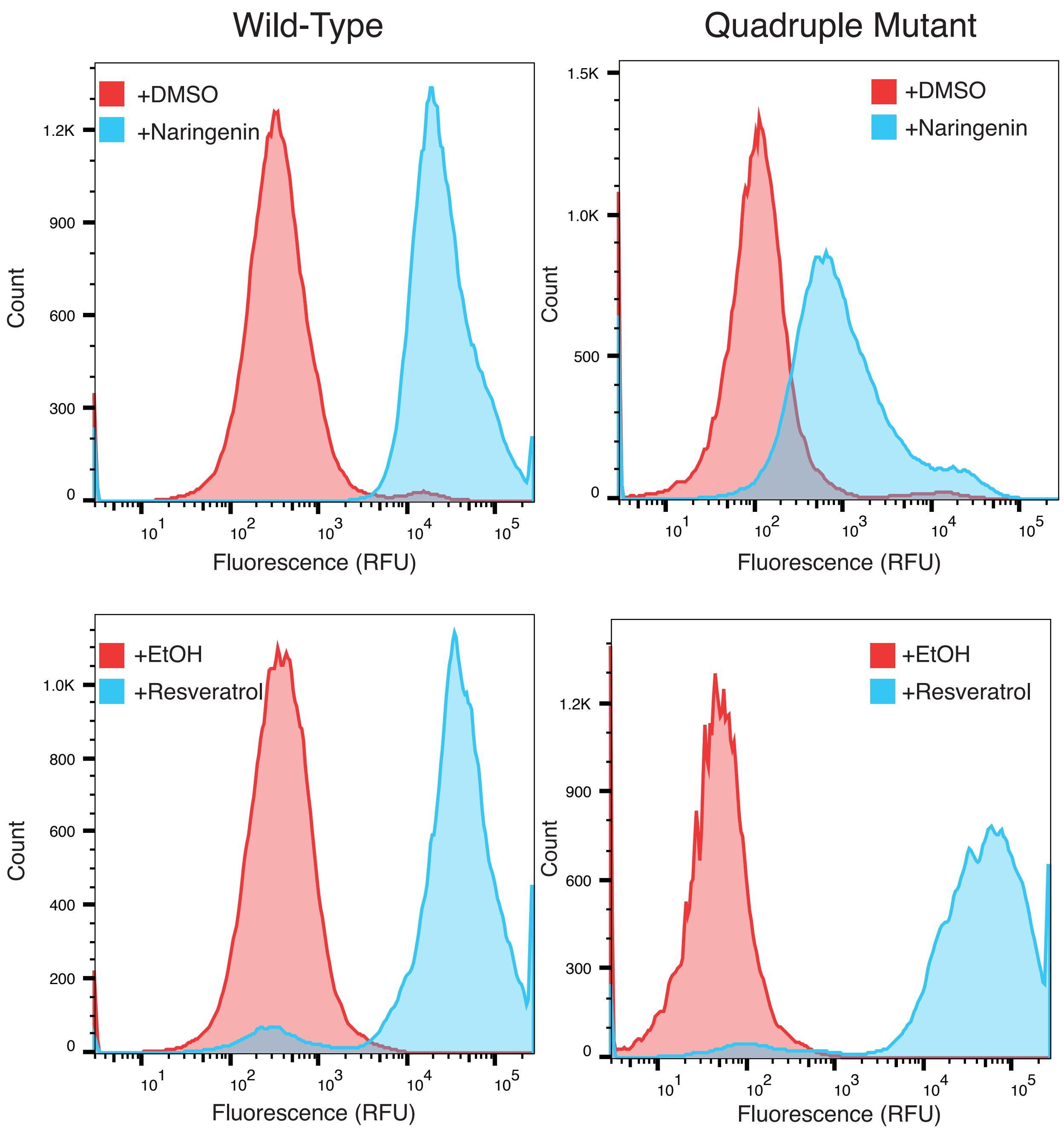
Histogram of the number of mutations in the library of experimentally screened, Rosetta-generated designs. The average number of mutations was 5.1 with a variance of 1.8.



Supplementary Figure 2: Mutation heatmap for synthesized resveratrol designs Heatmaps are colored by PSSM score calculated from the set of curated Rosetta designs. A black box is drawn around the wildtype amino acid identity at each position. ***(a)*** Heatmap of the first of two designed regions of TtgR. ***(b)*** Heatmap of the second designed region in TtgR.

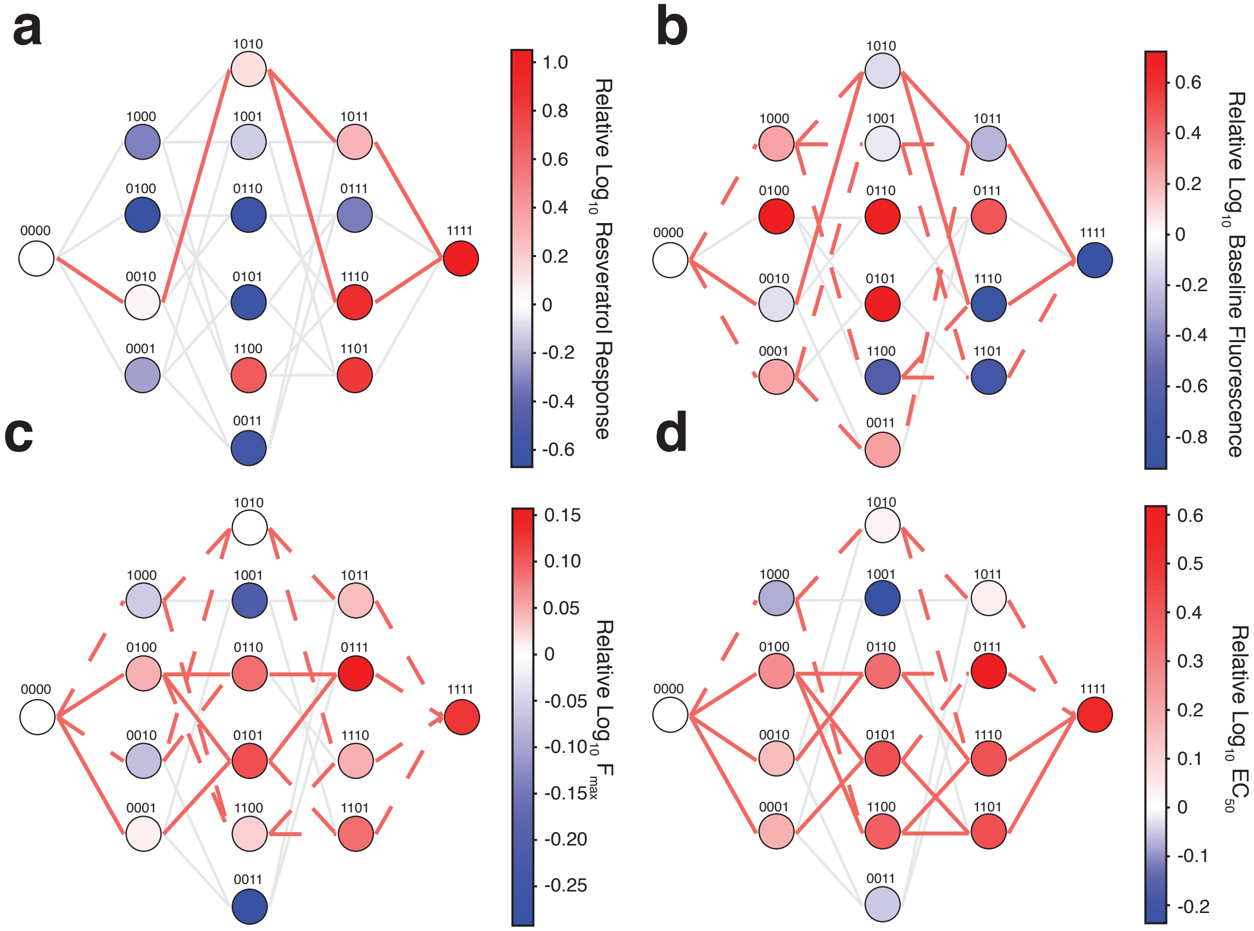


Supplementary Figure 3: Workflow for screening ligand-specific TtgR variants by fluorescence activated cell sorting
Rosetta-designed TtgR variants are transformed into *E. coli* cells carrying the reporter plasmid. TtgR variants are sorted by toggling between repressed and induced states (solid arrow). The lower 50% of fluorescent cells are sorted in the absence of inducer to isolate variants that are able to repress transcription. Subsequently, the sorted population are grown and induced with resveratrol. The top 5% of fluorescent cells are sorted to isolate variants capable of binding to the ligand and inducing GFP expression. After toggling multiple times, the repressed sort is repeated a final time before the subpopulation is clonally tested with both naringenin and resveratrol (dashed arrow).

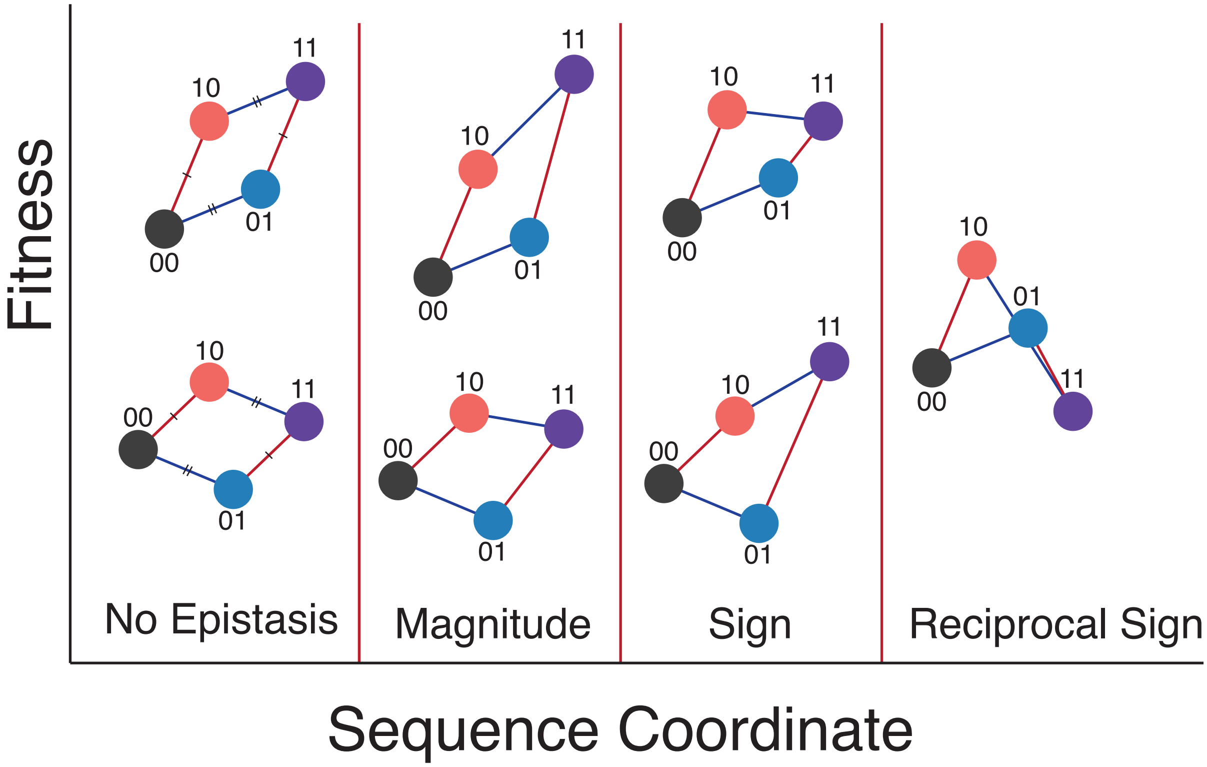


Supplementary Figure 4: Fluorescence distributions of wildtype TtgR and quadruple mutant

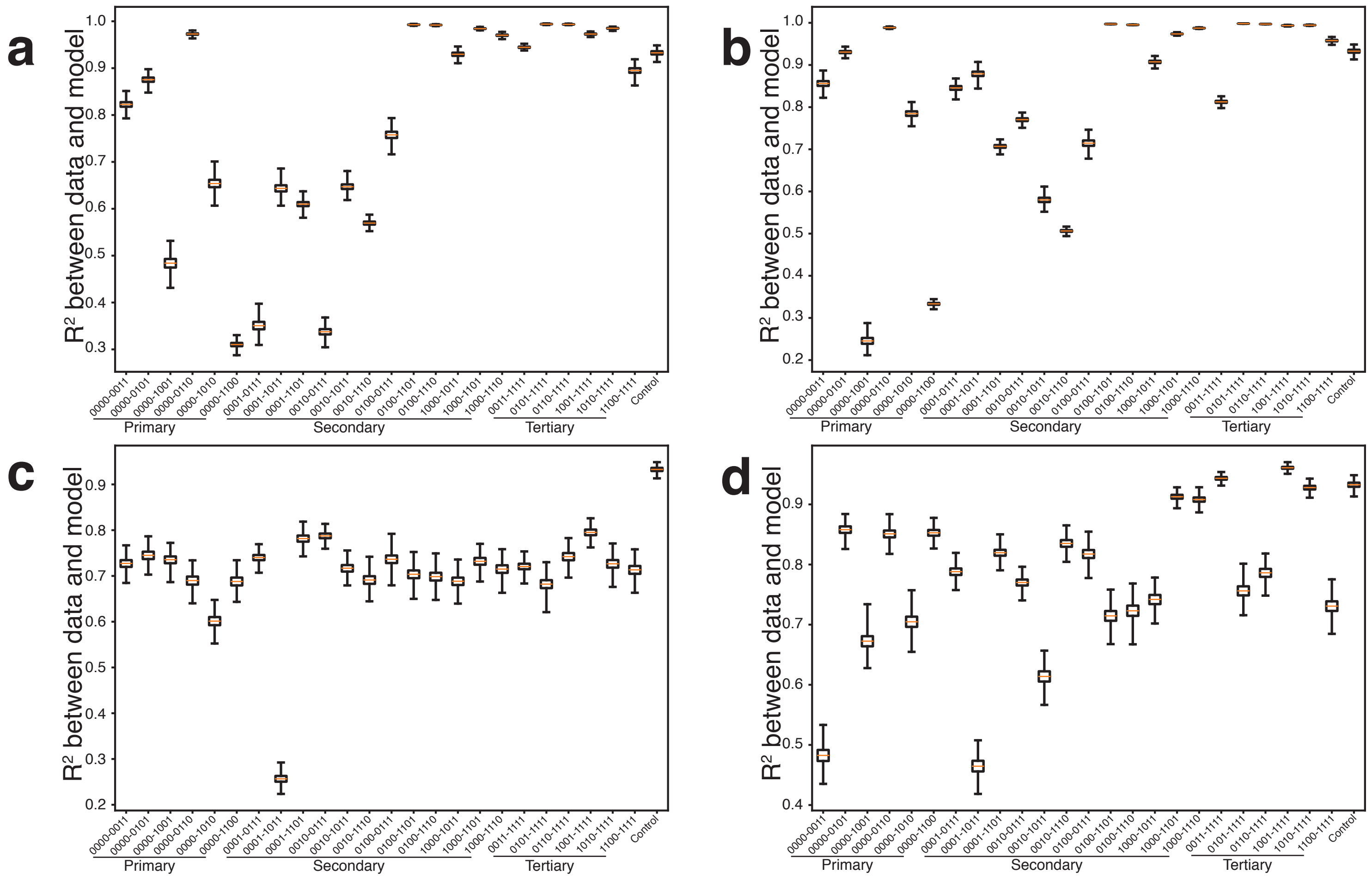
Flow cytometry histograms of wildtype TtgR and quadruple mutant TtgR with and without inducers. Naringenin 2000µM (blue) dissolved in DMSO and DMSO-only control (red), Resveratrol 250µM (blue) dissolved in ethanol and ethanol-only control (red) are shown.



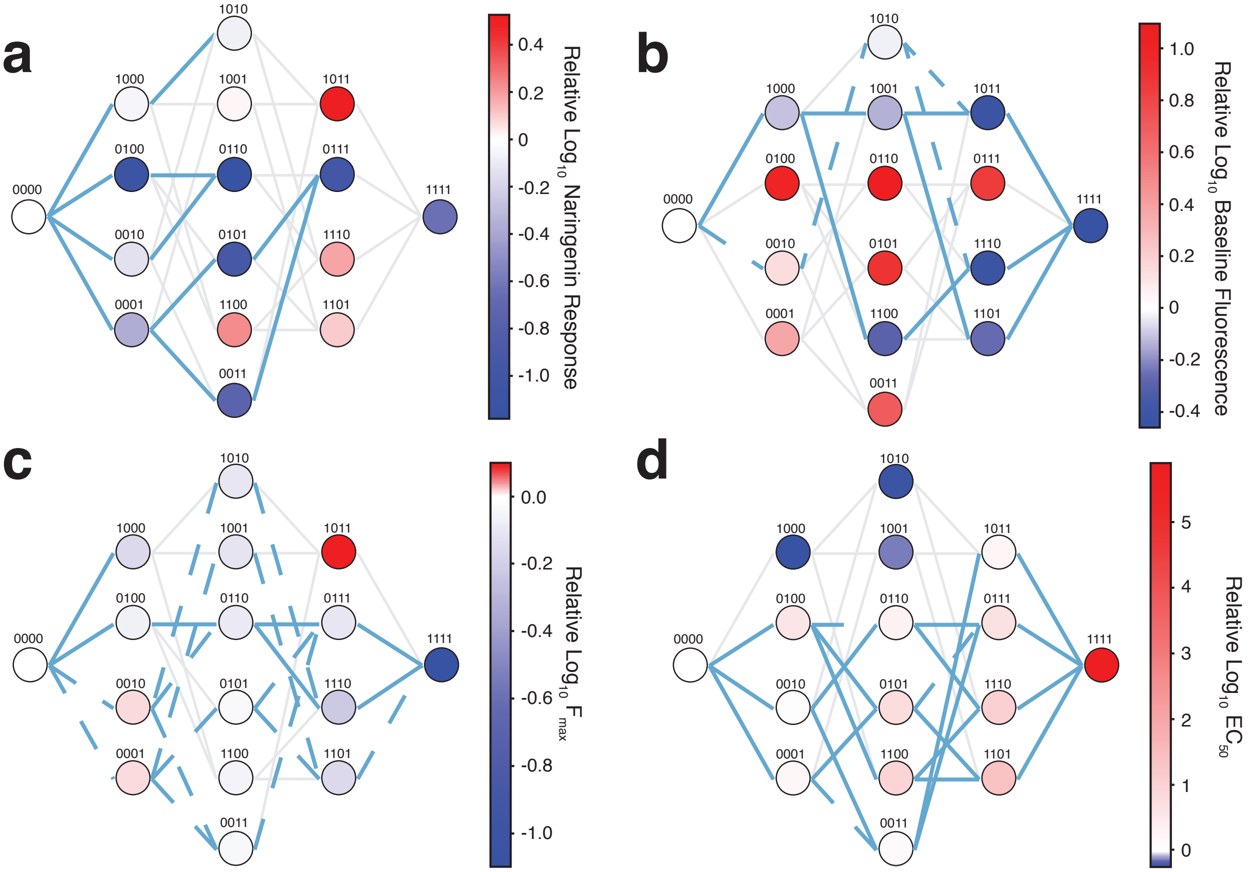
Supplementary Figure 5: Additional mutational pathways permitted by a 25% tolerance window for resveratrol functional parameters
The tolerance window describes the acceptance of a mutation that performs worse than the background variant when describing allowed pathways through sequence space. Each variant is labeled with a binary string corresponding to the presence (1) or absence (0) of a mutation at position 137, 141, 167, or 168 in order. Nodes separated by a single mutation are connected by edges showing viable (bold red) and unviable paths (light gray) through sequence space. Nodes are shaded by log$_{10}$ of the fitness parameter at 250μM resveratrol normalized to the fitness of wildtype TtgR. All new tolerated pathways are shown as red dashed lines. Additional pathways have been calculated for resveratrol ***(a)*** fold induction, ***(b)*** basal expression, ***(c)*** maximum expression, and ***(d)*** EC$_{50}$ landscapes. The fold induction landscape shows no additional pathways while the basal expression, maximum fluorescence, and EC$_{50}$ landscapes show 11, 8, and 8 additional paths, respectively.



Supplementary Figure 6: Visual definition of different types of epistasis
Visual representation of different types of epistasis. This graphical representation separates example subnetworks based on the type of epistasis. An arbitrary fitness metric is plotted against a sequence coordinate where each mutation is represented by a binary string. A system is non-epistatic when the combined effect of mutations is the sum of their individual effects. Magnitude epistasis occurs when the combined effect of mutations is greater than the sum of their individual effects (no change in direction). Sign epistasis occurs when one mutation switches direction from beneficial to detrimental (or vice versa) depending on the background in which it is introduced. Reciprocal sign epistasis occurs when both mutations switch direction depending on the background.

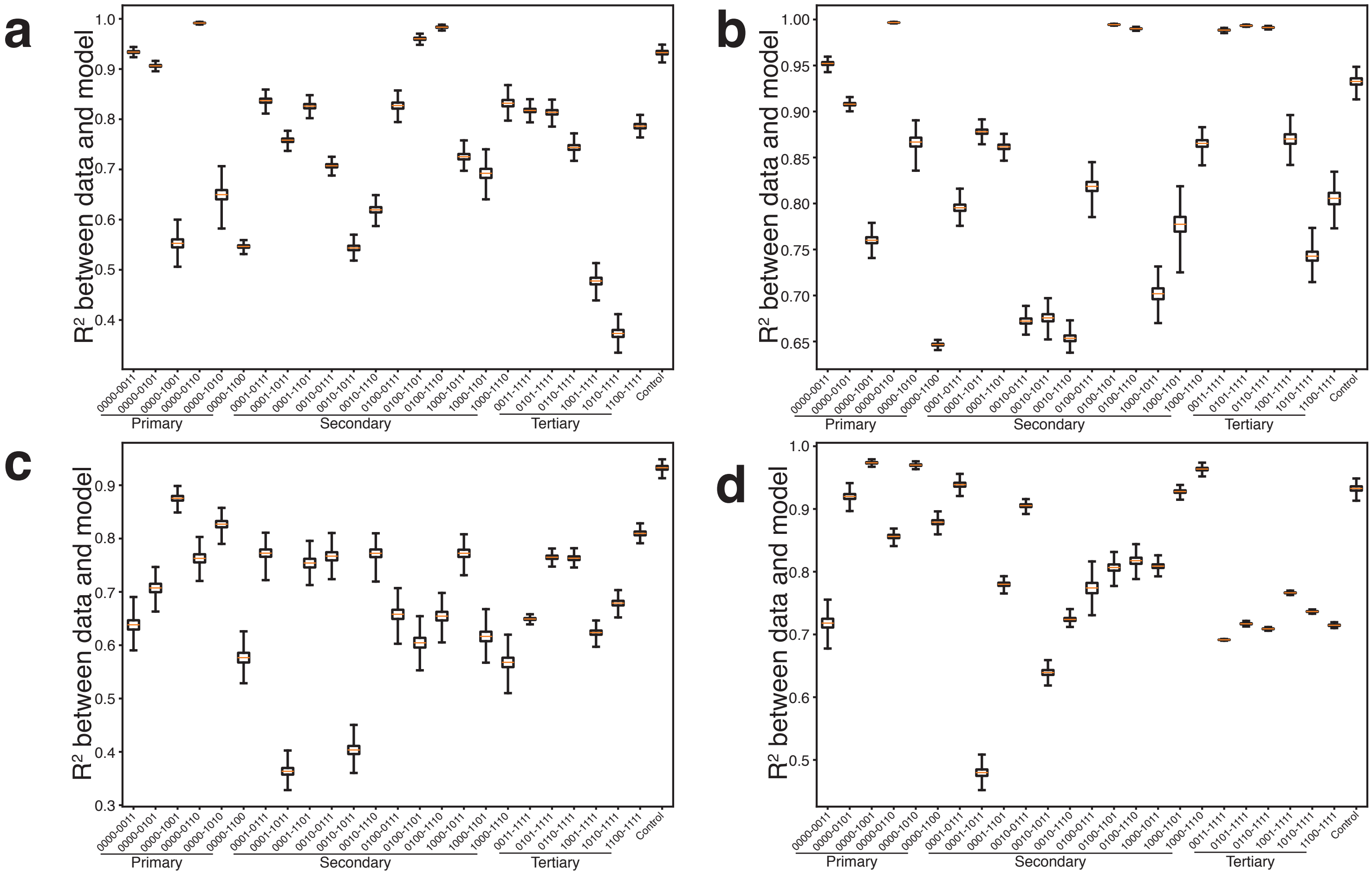


Supplementary Figure 7: Bahadur expansion of subnetworks in resveratrol functional parameters
Bahadur expansion was applied to the 24 subnetworks of the ***(a)*** fold induction, ***(b)*** basal expression, ***(c)*** maximum expression, and ***(d)*** EC_{50} landscapes. The box plots show the bootstrap averages (N=10,000 bootstrap replicates). Epistatic subnetworks were defined as those with an R^2 value of less than 0.9, based on simulated additive data (“Control”). The box denotes the interquartile range and the orange line denotes the median R^2 value for the bootstrap averages. The whiskers extend to the maximum and minimum R^2 values. The fold induction landscape shows that the majority of subnetworks in a wildtype or single-mutant background show epistasis while those in double-mutant backgrounds are less likely to show epistasis. The basal expression landscape shows similar patterns of R^2 values as the fold induction landscape. All of the subnetworks in the maximum expression landscape are epistatic. The EC_{50} landscape shows more epistasis than the fold induction landscape in subnetworks with a double mutant background.

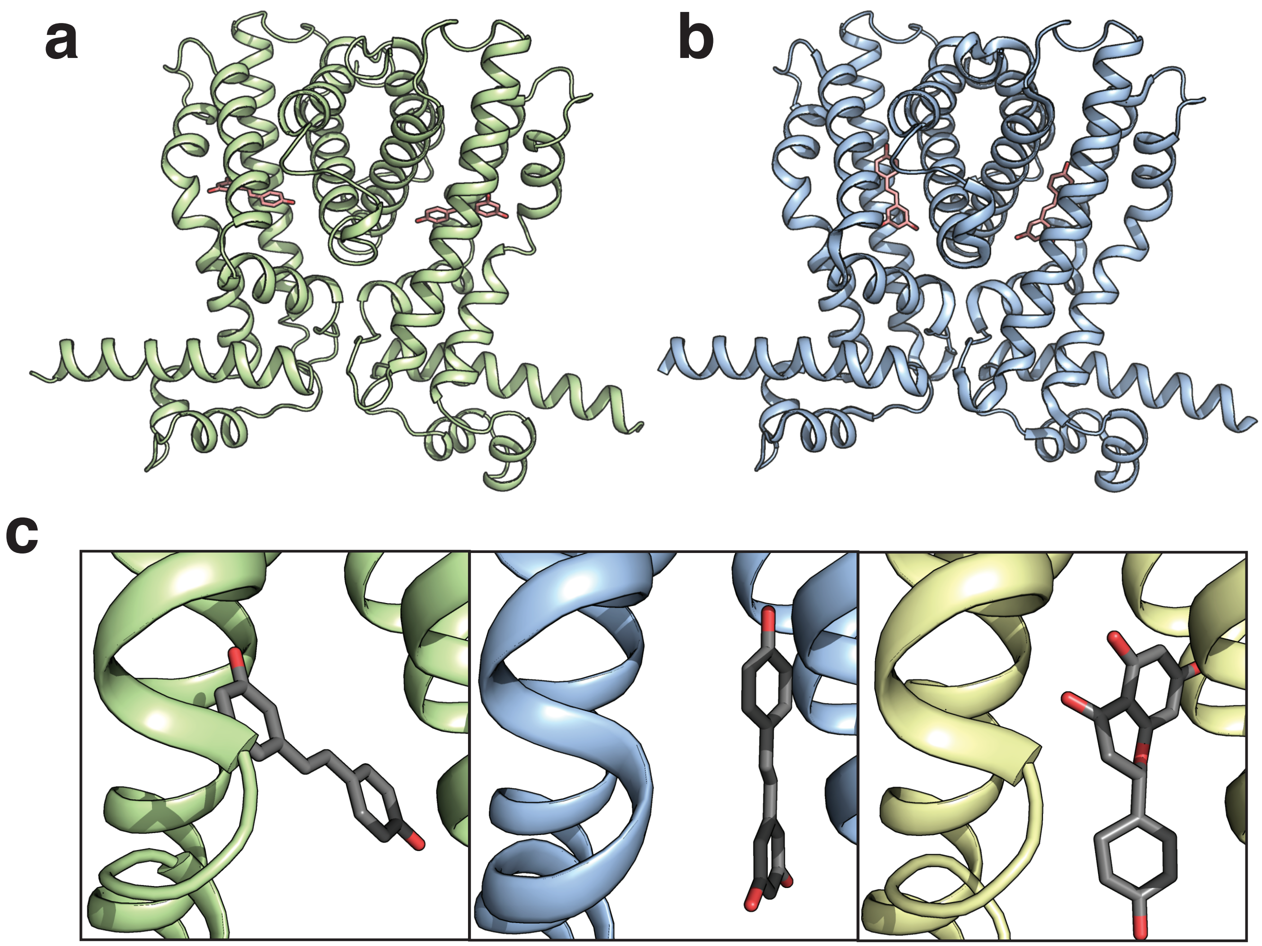


Supplementary Figure 8: Additional mutational pathways permitted by a 25% tolerance window for naringenin functional parameters

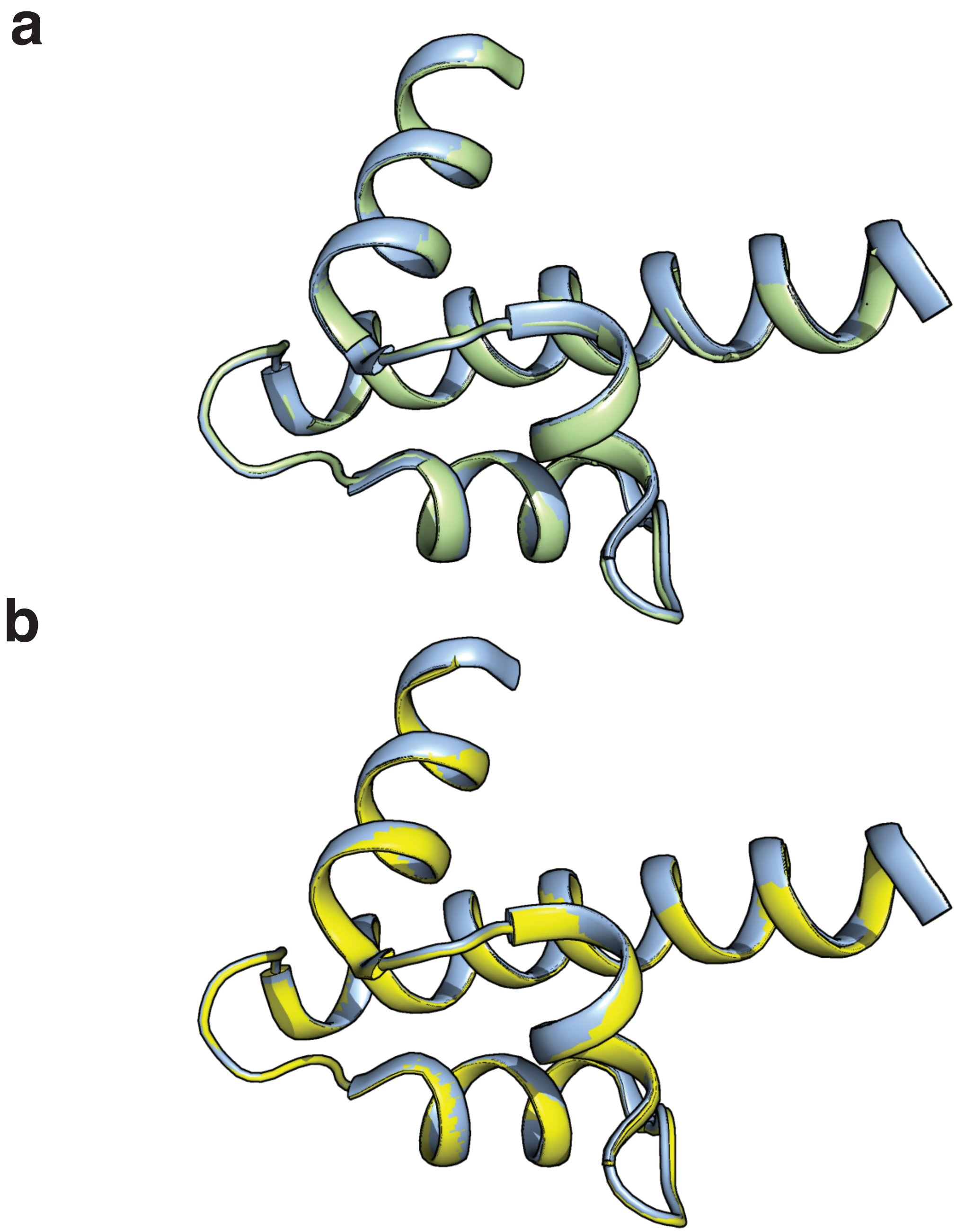
The tolerance window describes the acceptance of a mutation that performs worse than the background variant when describing allowed pathways through sequence space. Each variant is labeled with a binary string corresponding to the presence (1) or absence (0) of a mutation at position 137, 141, 167, or 168 in order. Nodes separated by a single mutation are connected by edges showing viable (bold red) and unviable paths (light gray) through sequence space. Nodes are shaded by log$_{10}$ of the fitness parameter at 2000μM naringenin normalized to the fitness of wildtype TtgR. All new tolerated pathways are shown as blue dashed lines. Additional pathways have been calculated for resveratrol ***(a)*** fold induction, ***(b)*** basal expression, ***(c)*** maximum expression, and ***(d)*** EC$_{50}$ landscapes. The fold induction landscape shows no additional pathways while the basal expression, maximum fluorescence, and EC$_{50}$ landscapes show 2, 8, and 6 additional paths, respectively.



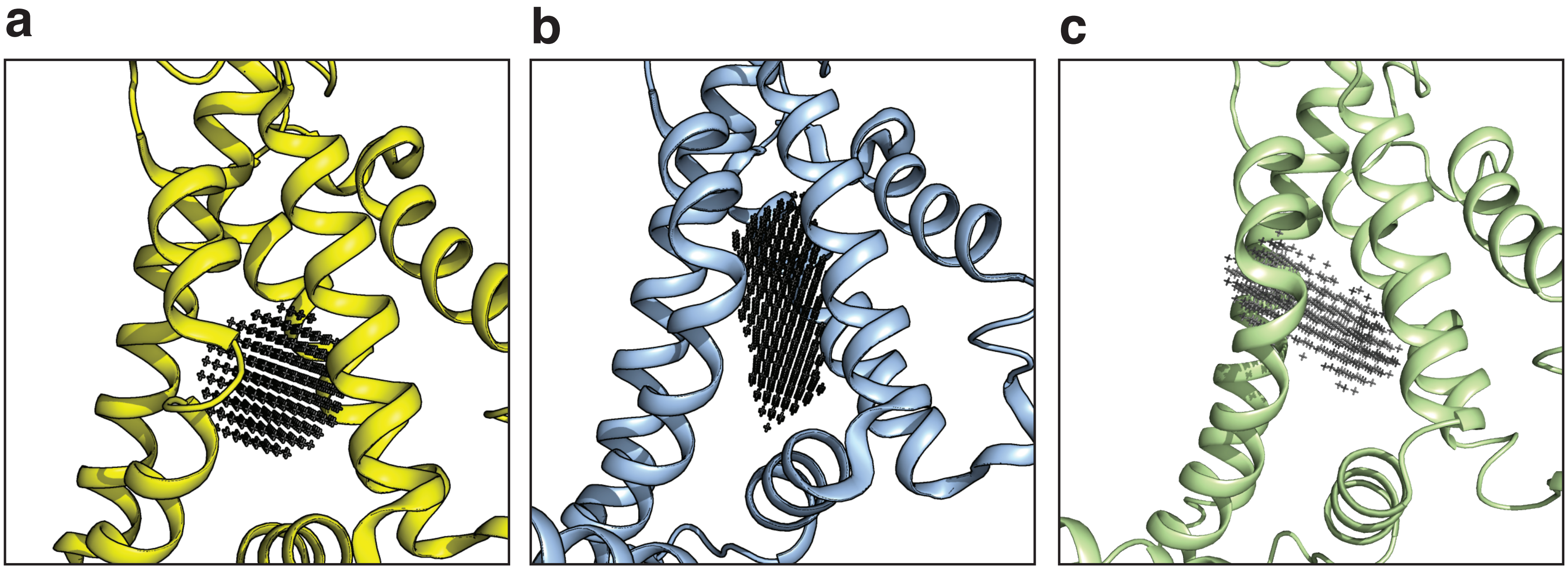
Supplementary Figure 9: Bahadur expansion of subnetworks in naringenin functional parameters
Bahadur expansion was applied to the 24 subnetworks of the ***(a)*** fold induction, ***(b)*** basal expression, ***(c)*** maximum expression, and ***(d)*** EC_{50} landscapes. Epistatic subnetworks are defined in the same fashion as Supplementary Fig. 7. The box plots show the bootstrap averages (N=10,000 bootstrap replicates). The box denotes the interquartile range and the orange line denotes the median R^2 value for the subnetwork. The whiskers extend to the maximum and minimum R^2 values. The fold induction landscape shows that only a small number of subnetworks in the wildtype or single mutant background are not epistatic. In contrast, the basal expression landscape has nonepistatic subnetworks in the wildtype, single, and double mutant backgrounds. Like the subnetworks of the resveratrol maximum fluorescence landscape, the majority of subnetworks in the naringenin maximum fluorescence landscape show epistasis. The EC_{50} landscape shows similarity to the fold induction landscape with nonepistatic subnetworks in the wildtype and single mutant backgrounds.



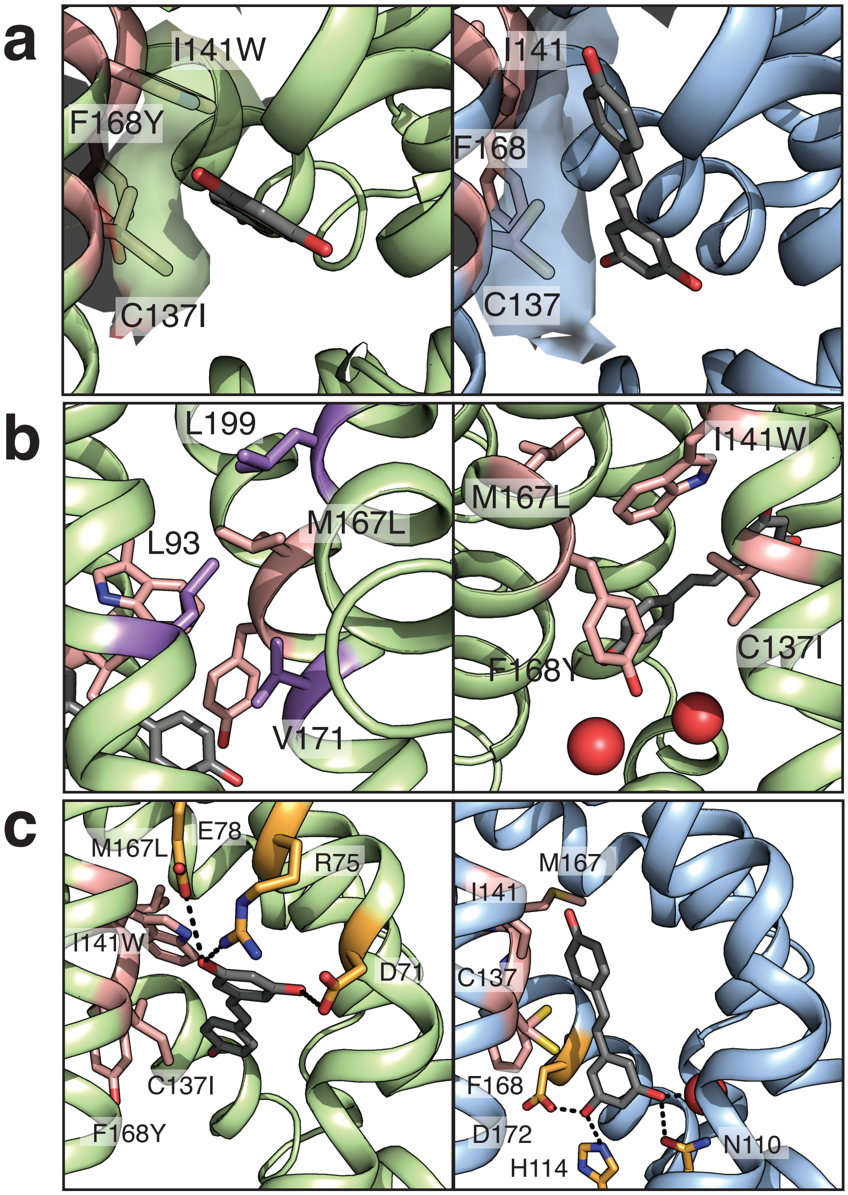
Supplementary Figure 10: Structure of ligand-bound wildtype and quvadruple mutant
TtgR is as an all-helical dimer. The helix-turn-helix domain at the N-terminal end binds to DNA. The ligand binding pocket is enclosed by five angled helices. An additional helix at the C-terminal end forms the dimerization interface. The quadruple mutant (PDB: 7KD8) ***(a)*** is structurally identical to the wildtype (PDB: 7K1C) ***(b)***. Resveratrol is shown as pink sticks in both. ***(c)*** A close-up view of the binding orientation of resveratrol in the pocket. The quadruple mutant (left) binds to resveratrol in the horizontal orientation. Wildtype TtgR binds to resveratrol (middle) or to naringenin (right, PDB ID: 2UXU) in the vertical orientation. Resveratrol is shown as grey sticks.



Supplementary Figure 11: Alignment of DNA binding domains for resveratrol-bound quadruple mutant, resveratrol-bound wildtype, and apo quadruple mutant TtgR
(a) The DNA binding domains of the resveratrol-bound quadruple mutant (green) and the resveratrol-bound wildtype (blue) TtgR. The RMSD of these two domains is 1.03Å. ***(b)*** The DNA binding domains of resveratrol-bound wildtype (blue) and apo wildtype (yellow) TtgR. The RMSD of these two domains is 1.35Å.

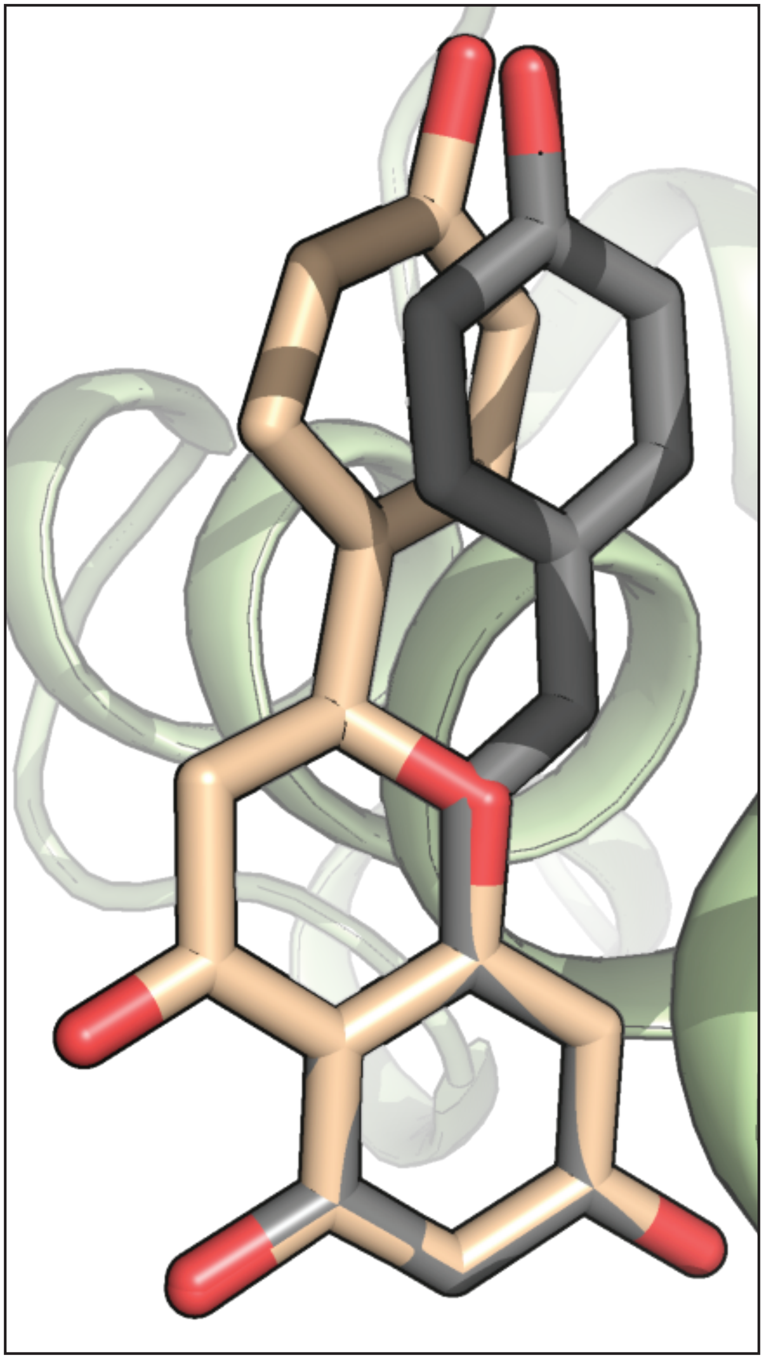


Supplementary Figure 12: TtgR binding pocket volume visualization
Binding pocket volumes were calculated using POVME3.0 and visualized in Pymol. ***(a)*** The pocket volume of apo wildtype TtgR, represented by the Xs, is 170Å^3. The pocket was not predefined in this calculation. ***(b)*** The pocket volume of resveratrol-bound wildtype TtgR is 215Å^3. ***(c)*** The pocket volume of resveratrol-bound quadruple mutant TtgR is 234Å^3. The pockets for (b) and (c) were defined using the resveratrol molecule.

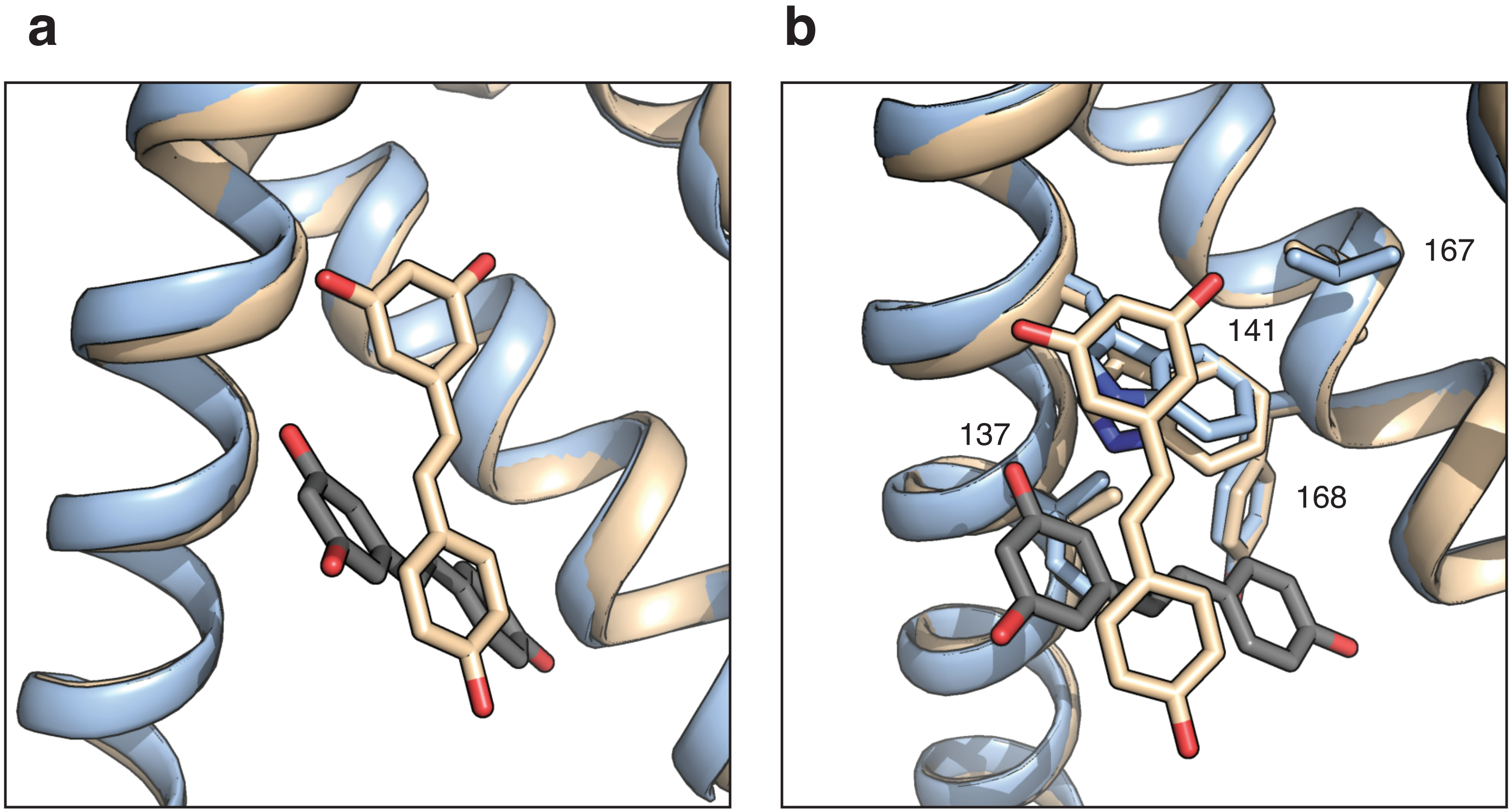


Supplementary Figure 13: Interactions of mutated positions and alternate hydrogen bonding networks

(a) The C137I mutation creates a small cleft in the binding pocket that can enhance shape complementarity to resveratrol in the horizontal binding mode. The quadruple mutant (left) is shown in comparison to wildtype (right). The van der Waals surface of residue 137 and 141 is shown for both structures. Positions 137, 141, 167, and 168 are shown as pink sticks. ***(b)*** (Left) M167L creates nonpolar interactions with residues in helices composing the binding pocket and dimerization interface (purple sticks). Mutated positions 137, 141, 167, and 168 are shown in pink. 167 also plays a role in positioning the I141W side chain. (Right) The F168Y substitution enables the formation of additional hydrogen bonds to solvent that can create a hydrogen bond network with D172. Water molecules are shown as red spheres. ***(c)*** The hydrogen bond network differs for the quadruple mutant between chain A and chain B due to the slightly different position of the resveratrol molecules in each. (Left) In the quadruple mutant, D71, R75 and E78 (orange) make hydrogen bonds with the resveratrol molecules. (Right) The hydrogen bonding network of wildtype chain A is identical to chain B. Water molecules are shown as red spheres.

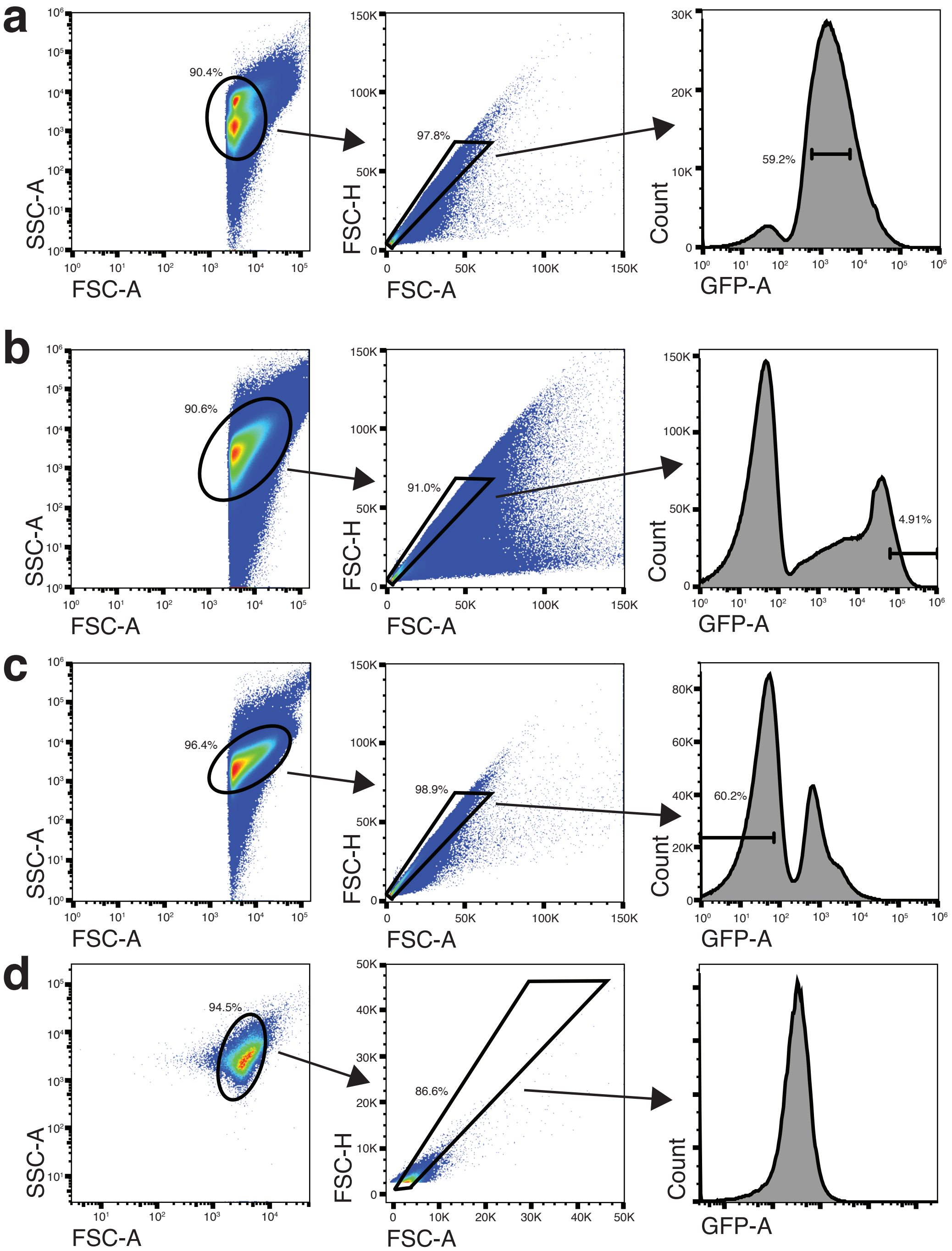


Supplementary Figure 14: Naringenin and resveratrol overlap
Naringenin (in brown) derived from a previous structure (PDB: 2UXU) is overlapped with resveratrol (in grey) via the pair_fit function in Pymol. The structures of each ligand are similar with respect to the location of hydroxyl groups, but differ by the addition of a carbonyl in the 4-chromanone backbone of naringenin.



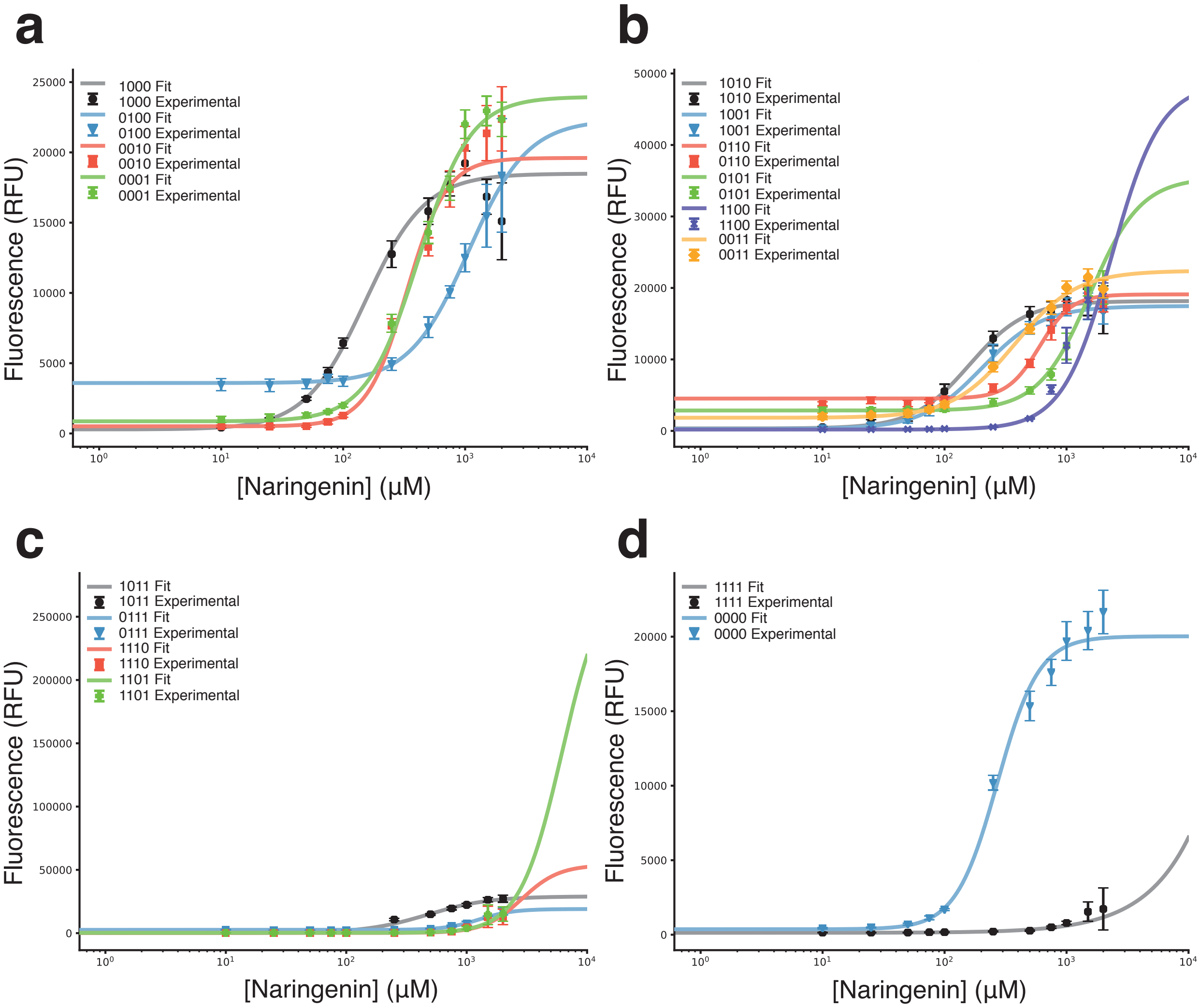
Supplementary Figure 15: Designed TtgR quadruple mutant binding pocket compared to crystal structure of TtgR quadruple mutant

The designed TtgR quadruple mutant (brown) is aligned to the crystal structure of the TtgR quadruple mutant (blue). ***(a)*** Resveratrol positions of the designed TtgR (brown) and the crystal structure (grey). ***(b)*** Mutated positions are shown as sticks. The design was unable to model the resveratrol in the horizontal orientation even though the residue rotamer states are similar to those in the crystal structure.

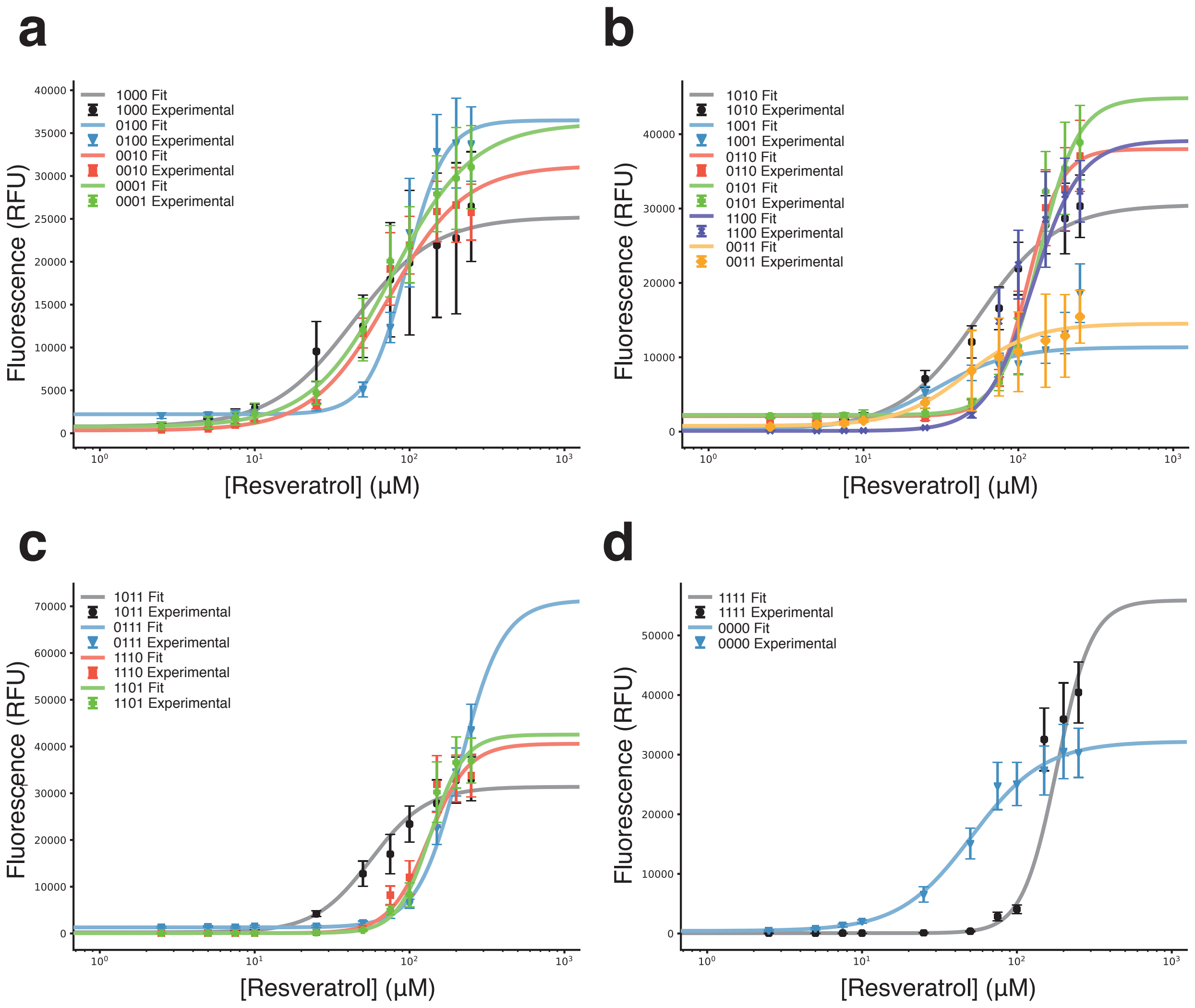


Supplementary Figure 16: Gating scheme for fluorescence-activated cell sorting and flow cytometry

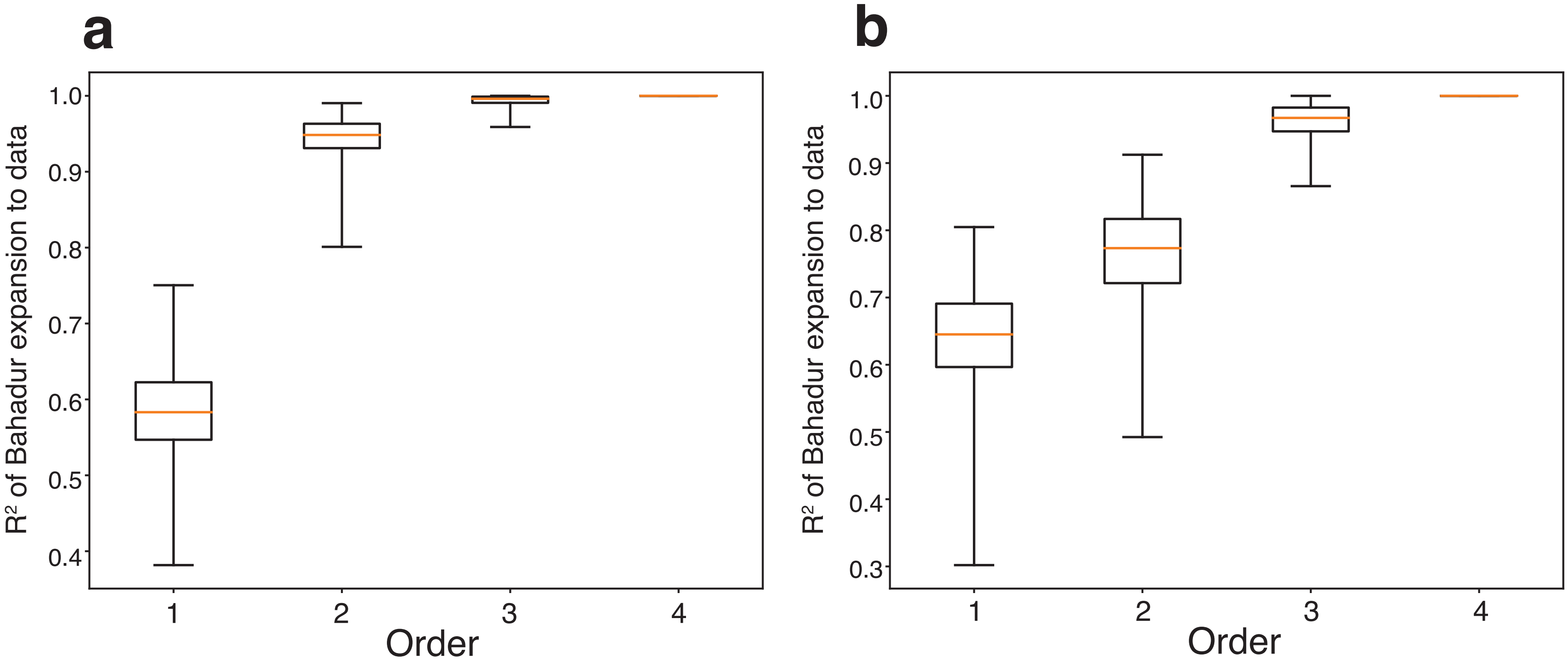
(a) Gating strategy to sort cells containing a TtgR variant that is able to repress *sfGFP* expression. ***(b)*** Gating strategy to sort cells containing a TtgR variant that is able to induce *sfGFP* expression when exposed to 95.5µM resveratrol. ***(c)*** Gating strategy to isolate cells containing a TtgR variant that is able to both repress *sfGFP* expression in the absence of resveratrol and induce *sfGFP* expression in the presence of resveratrol. The sorting gating schemes are presented in Figure 1 and Supplementary Fig. 3. ***(d)*** Gating strategy to calculate fluorescence for flow cytometry experiments presented in Fig. 2, Fig. 3, Supplementary Fig. 4, Supplementary Fig. 6, Supplementary Fig. 7, Supplementary Fig. 8, Supplementary Fig. 9, Supplementary Fig. 16, and Supplementary Fig. 17.



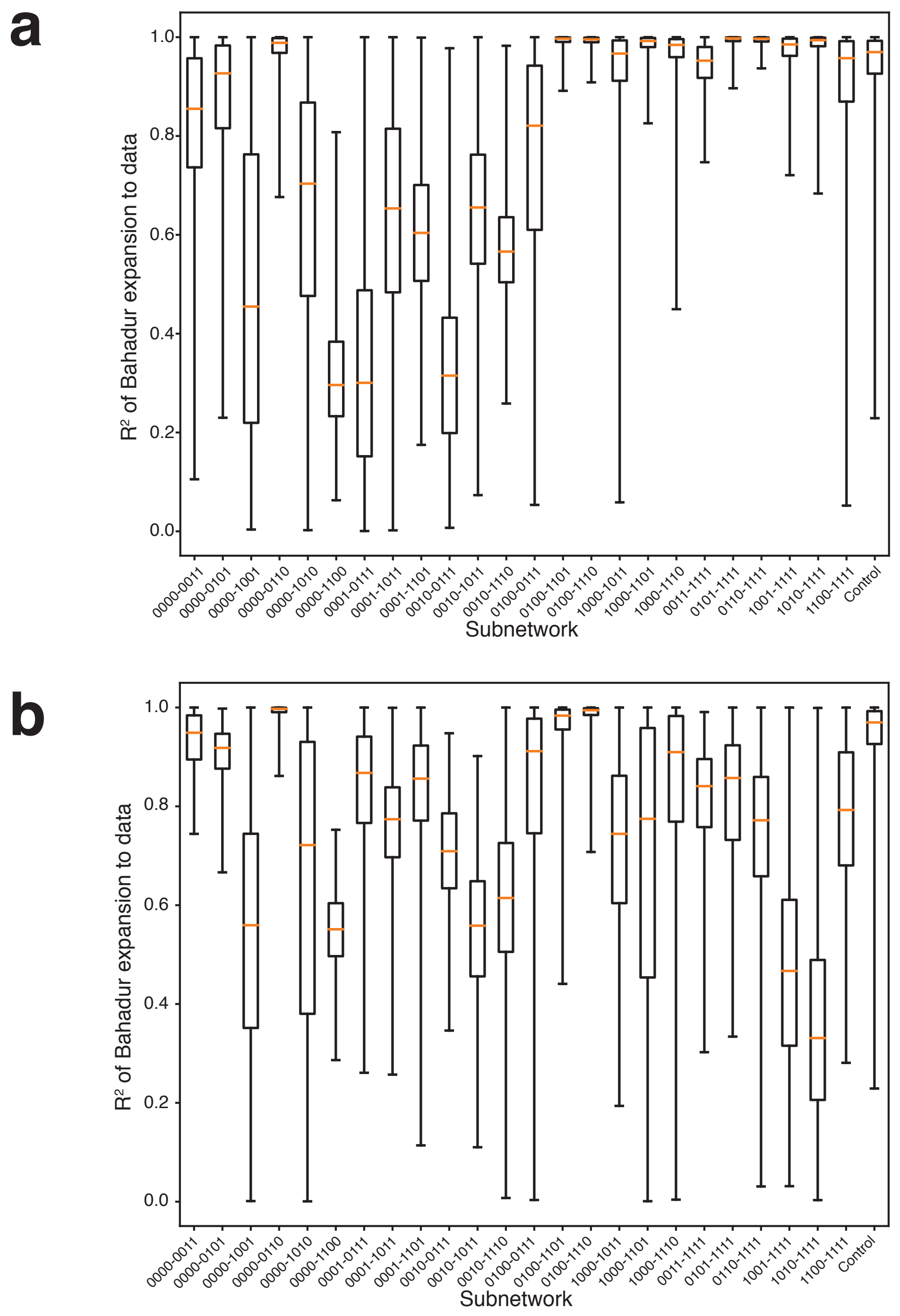
Supplementary Figure 17: Naringenin dose response curves
Dose response curves to naringenin for all 16 mutational combinations. Naringenin concentration varied between 0μM and 2000μM naringenin. Fit is shown as a solid line and experimental data is shown as markers with error bars. The marker denotes the averages and the error bars show the standard deviations of biological triplicate measurements (n=3) unless otherwise specified (see Methods). ***(a)*** Single mutant fits (1000, 0100, 0010, and 0001). ***(b)*** Double mutant fits (1010, 1001, 0110, 1100, and 0011). ***(c)*** Triple mutant fits (1011, 0111, 1110, and 1101). ***(d)*** Wildtype (0000) and quadruple mutant (1111) dose response curves.



Supplementary Figure 18: Resveratrol dose response curves
Dose response curves for all 16 mutational combinations to resveratrol. Resveratrol concentration varied between 0μM and 250μM resveratrol. Fit is shown as a solid line and experimental data is shown as markers with error bars. The marker denotes the averages and the error bars show the standard deviations of biological triplicate measurements (n=3) unless otherwise specified (see Methods). ***(a)*** Single mutant fits (1000, 0100, 0010, and 0001). ***(b)*** Double mutant fits (1010, 1001, 0110, 1100, and 0011). ***(c)*** Triple mutant fits (1011, 0111, 1110, and 1101). ***(d)*** Wildtype (0000) and quadruple mutant (1111) dose response curves.

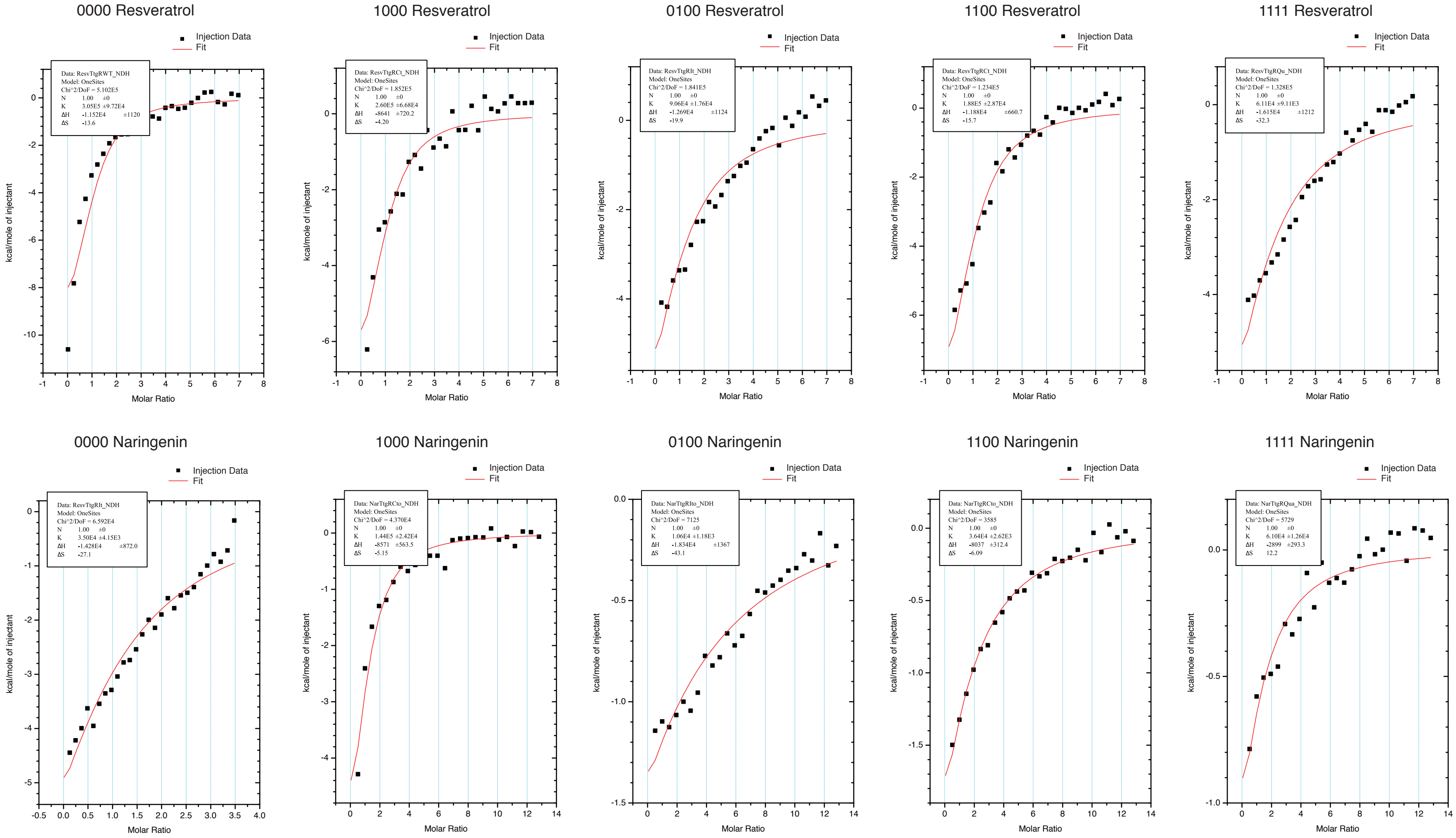


Supplementary Figure 19: Distribution of R^2 for first and higher order interactions for the full network
Distribution of R^2 values from the Bahadur expansion model applied to the full 16-member network after stochastic sampling (N=500) of fold induction values based on experimental averages and standard deviations. Boxplots of first, second, third, and fourth order interactions are shown for ***(a)*** resveratrol and ***(b)*** naringenin. The orange line is the median R^2 value for the distribution and the box encloses the interquartile range (IQR). The whiskers extend to the maximum and minimum R^2 values. The raw R^2 values for the fold induction, baseline fluorescence, EC_{50}, and can be found online at https://github.com/raman-lab/epistasis.

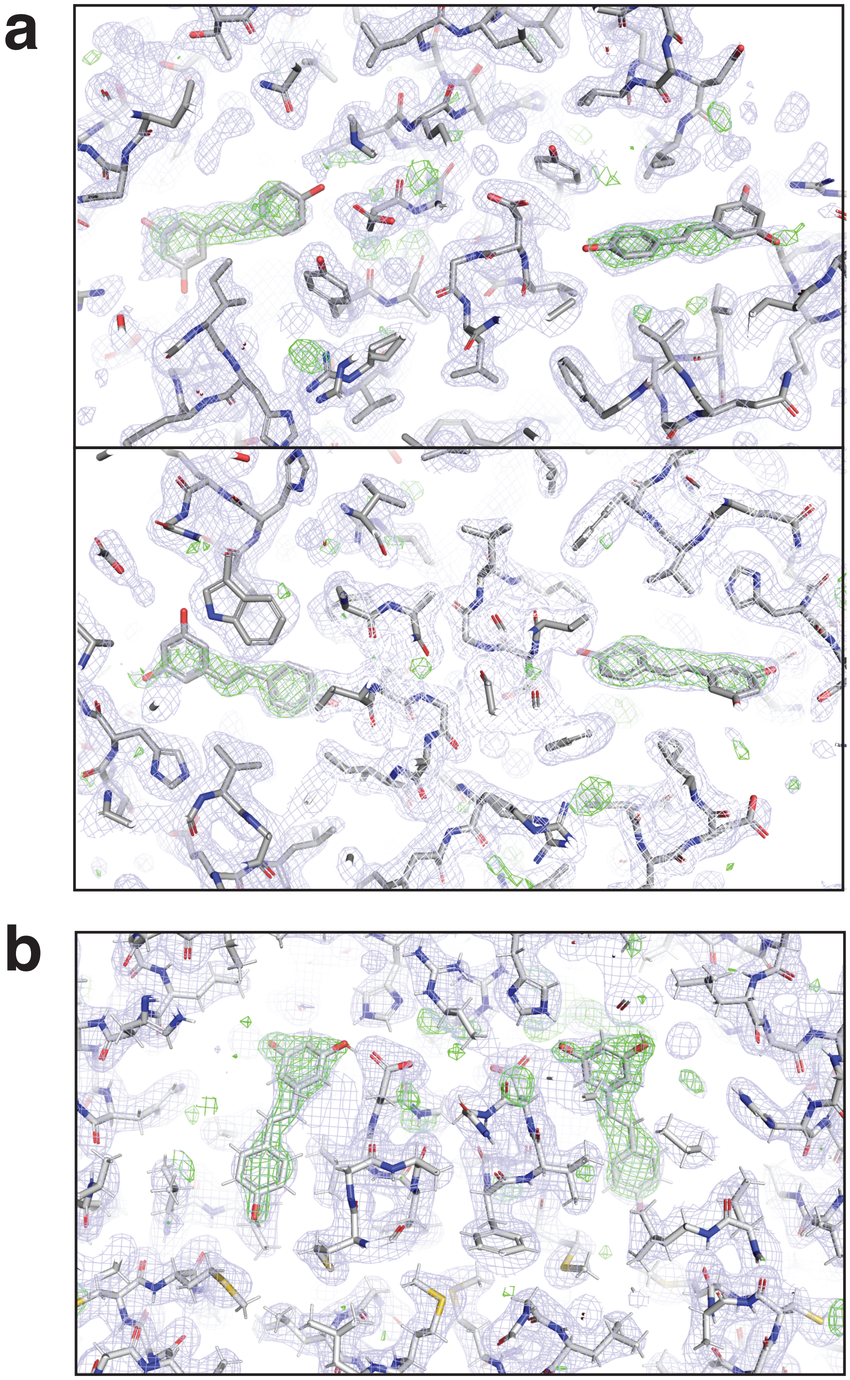


Supplementary Figure 20: Distribution of R^2 for first order interactions for individual subnetworks

Distribution of R^2 values from the Bahadur expansion model applied to each subnetwork after stochastic sampling of experimental fold induction values for ***(a)*** resveratrol or ***(b)*** naringenin based on experimental averages and standard deviations. Each subnetwork network was modeled 500 times by stochastic sampling. The orange line is the median R^2 value for the distribution and the box encloses the interquartile range (IQR). The whiskers extend to the maximum and minimum R^2 values.The raw R^2 values for the fold induction, baseline fluorescence, EC$_{50}$, and can be found online at https://github.com/raman-lab/epistasis.



Supplementary Figure 21: Estimating binding parameters from isothermal calorimetry of wildtype TtgR and variants
Isothermal titration calorimetry experimental data for affinity of TtgR mutants to either naringenin or resveratrol. Heat per mole of ligand injected (kCal/mol) is plotted as a function of the molar ratio of ligand:protein. Binding parameters are estimated from single site binding model fits using Origin 7.0 software (MicroCal). Due to low affinities for both naringenin and resveratrol, stoichiometry was fixed to 1 for both naringenin and resveratrol (see methods).



Supplementary Figure 22: mF$_o$-DF$_c$ and 2mF$_o$-DF$_c$ omit maps for resveratrol-bound quadruple mutant TtgR and resveratrol-bound wildtype TtgR

Maps for the protein density were calculated in phenix from deposited models and structure factor amplitudes. The mF$_o$-Df$_c$ map is contoured at 3σ and the 2mF$_o$-DF$_c$ maps were contoured at 2σ. The mF$_o$-DF$_c$ omit map is shown as green wires while the 2mF$_o$-DF$_c$ omit map is shown in grey. ***(a)*** mF$_o$-DF$_c$ and 2mF$_o$-DF$_c$ omit maps for resveratrol-bound quadruple mutant TtgR. Chain A is shown in the top panel on the left and chain B is on the right. The lower panel depicts chain C (left) and chain D (right). ***(b)*** mF$_o$-DF$_c$ and 2mF$_o$-DF$_c$ omit maps for resveratrol-bound wildtype TtgR. Chain A is shown on the left and chain B is shown on the right.

Tables

Name	Sequence
KN_E1	TATCACGAGGCCCTTTCGTCTTCACCACCCAGCAGTATTGACAAACAAC
KN_E2	TTCATGGTTGTTTGTCAATACTGCTGGGTGggcgcgccatgactaagcttttcattgtct
KN_E3	aaagttaaatgTTGCTAAGGATTATACTTACATTCATGGTTGTTTGTCAATACTGCTGGG
KN_E4	atgtatatctccttcttaaagttaaatgTTGCTAAGGATTATACTTA
KN_E5	cagctcttcgcctttacgcatatgtatatctccttcttaaagttaaatgTT
KN_E6	GTGAAGACGAAAGGGCCTCG
KN_E7	atgcgtaaaggcgaagagctg
KN_E8	catgctgcttcatGtggtcc
KN_E9	GCTGGCAATTCCGACGTC
KN_E10	TTGACAATTAATCATCCGGC
KN_E11	CGAGCCGGATGATTAATTGTCAA
KN_E12	TGAattagcagaaagtcaaaagcctccga
KN_E13	tcggaggcttttgactttctgctaatTCATTATTTGCGCAGCGCCGG
KN_E14	gCGATCGTGCCCACCT
KN_E15	GTGCGGGCTCCAACT
KN_E16	ggCTGGTGCGTCGTCT
KN_E17	cGGGAAGTGTTCGCCG
KN_E18	GGTCTCGGTTCTGGATGCACGTACCCGTCGC
KN_E19	GGTCTCGCAGTGCCTGAACCAGTTCGGC
KN_E20	GGTCTCGCGTTGGCTGCTGCTGCCGGATAG
KN_E21	GGTCTCGATCCAGCACGGCGCTCTGGCGC
KN_E22	GAAATTCGTCAGCAGCGCCAGAGCGCCGTGCTGGATattCATAAAGGTATCACC
KN_E23	GAAATTCGTCAGCAGCGCCAGAGCGCCGTGCTGGATTGTCATAAAGGTtggACC
KN_E24	GAAATTCGTCAGCAGCGCCAGAGCGCCGTGCTGGATattCATAAAGGTtggACC
KN_E25	GAAATTCGTCAGCAGCGCCAGAGCGCCGTGCTGGATTGTCATAAAGGTATCACC
KN_E26	CAGCAGCAGCCAACGGCGAATCAGGCCATCCACATAGGCAAAcagCGCAACCGC
KN_E27	CAGCAGCAGCCAACGGCGAATCAGGCCATCCACATAGGCataCATCGCAACCGC
KN_E28	CAGCAGCAGCCAACGGCGAATCAGGCCATCCACATAGGCatacagCGCAACCGC
KN_E29	CAGCAGCAGCCAACGGCGAATCAGGCCATCCACATAGGCAAACATCGCAACCGC
KN_E30	TTTTGTTTAACTTTAAGAAGGAGATATACATATGaaaatcgaagaaggtaaactggtaat
KN_E31	CATATGTATATCTCCTTCTTAAAGTTAAACAAAA
KN_E32	attgaaaatataaattttcGTGGTGGTGGTGGTGGTGgtcgccgttaattaaagtctgcg
KN_E33	CCACCACCACgaaaatttatattttcaatctATGGTGCGTCGCACCAAAGAAGAAG
KN_E34	CTTTGTTAGCAGCCGGATCTCATTATTTGCGCAGCGCCGGGCTCAG
KN_E35	TGAGATCCGGCTGCTAACAAAGCCCGAAAGGA

Supplementary Table 1: Primers

Names and sequences of all primers used in this study.

Binary	Average	St. Dev
00	3.1	0.93
01	24.5	7.35
10	56.1	16.83
11	74.4	22.32

Supplementary Table 2: Control additive data set
A set of random values that are additive with respect to the mean. The standard deviation of each datapoint is 30% of the mean. This control set was used to calculate the R^2 for comparison to the subnetwork Bahadur expansions.

	Quadruple Mutant (Apo) 7K1A	TtgR QM STL 7KD8	wtTtgR STL 7K1C
Wavelength	0.9786	1.078	0.9786
Resolution range	32.14 - 1.75 (1.813 - 1.75)	25.65 - 1.71 (1.771 - 1.71)	28.63 - 1.9 (1.968 - 1.9)
Space group	C 2 2 21	P 1	C 2 2 21
Unit cell	57.92 64.28 223.87 90 90 90	43.497 43.587 115.942 97.969 98.648 96.761	57.73 64.5 223.22 90 90 90
Total reflections	578957 (41577)	311123 (31645)	481016 (49244)
Unique reflections	42585 (4137)	82029 (8133)	33367 (3316)
Multiplicity	13.6 (10.1)	3.8 (3.9)	14.4 (14.9)
Completeness (%)	99.71 (97.73)	91.78 (91.33)	99.88 (99.97)
Mean I/sigma(I)	16.20 (1.29)	13.88 (1.92)	22.92 (2.06)
Wilson B-factor	35.18	31.18	36.4
R-merge	0.09432 (1.303)	0.04573 (0.5776)	0.07704 (1.294)
R-meas	0.09805 (1.373)	0.05349 (0.6697)	0.07997 (1.339)
R-pim	0.02647 (0.4257)	0.02758 (0.3382)	0.02123 (0.3446)
CC1/2	0.987 (0.578)	0.998 (0.901)	0.999 (0.739)
CC*	0.997 (0.856)	1 (0.974)	1 (0.922)
Reflections used in refinement	42566 (4136)	81923 (8104)	33349 (3315)
Reflections used for R-free	1979 (197)	1975 (197)	2020 (202)
R-work	0.1966 (0.3962)	0.1927 (0.2891)	0.1833 (0.3081)
R-free	0.2400 (0.3854)	0.2406 (0.3556)	0.2246 (0.3837)
CC(work)	0.955 (0.728)	0.959 (0.912)	0.963 (0.810)
CC(free)	0.920 (0.647)	0.930 (0.871)	0.948 (0.650)
Number of non-hydrogen atoms	3609	7247	3552
macromolecules	3321	6841	3322
ligands	2	72	36
solvent	286	334	194
Protein residues	413	830	415
RMS(bonds)	0.003	0.006	0.016
RMS(angles)	0.48	0.72	1.23
Ramachandran favored (%)	99.27	99.27	99.51
Ramachandran allowed (%)	0.73	0.61	0.49
Ramachandran outliers (%)	0	0.12	0
Rotamer outliers (%)	0.58	0.84	0.86
Clashscore	2.11	2.17	2.53
Average B-factor	39.41	40.02	42.83
macromolecules	39.27	39.83	42.66
ligands	35.92	53	52.81
solvent	41.05	41.07	43.92

Number of TLS groups	14	22	1

Supplementary Table 3: Crystallography refinement statistics

Refinement statistics for three structures: apo quadruple mutant TtgR (7K1A), wildtype TtgR bound to resveratrol (7K1C), and quadruple mutant TtgR bound to resveratrol (7KD8). Statistics for the highest resolution shell are shown in parentheses.

	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111	Residue Interactions
Ψ_0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
Ψ_1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	1
Ψ_2	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	1	1	2
Ψ_3	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	3
Ψ_4	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	4
Ψ_5	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1-2
Ψ_6	1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1-3
Ψ_7	1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	1-4
Ψ_8	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	2-3
Ψ_9	1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	2-4
Ψ_{10}	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	3-4
Ψ_{11}	-1	-1	1	1	1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	1-2-3
Ψ_{12}	-1	1	-1	1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	1-2-4
Ψ_{13}	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1	1	-1	-1	1	1-3-4
Ψ_{14}	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	2-3-4
Ψ_{15}	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1	1-2-3-4

Supplementary Table S4: Psi values for Bahadur Expansion

Psi values for all orders of interactions (right column) for each mutant.

3.0 Radical redesign of allosteric transcription factors to bind to novel ligands

Kyle K. Nishikawa[1], Srivatsan Raman[1,2,3]

Manuscript in preparation

Author Contribution: I designed and created the plasmid construct used for the RNA-Seq experiments. I cloned the agnostic and DMS libraries into the RNA-Seq construct. I cloned all of the libraries into the mapping construct. I prepared all samples for next-generation sequencing. I cloned the top variants into the fluorescent reporter construct. I performed all sorting experiments. I did all of the data analysis.

1 Department of Biochemistry, University of Wisconsin-Madison, Madison, WI
2 Department of Bacteriology, University of Wisconsin-Madison, Madison, WI
3 Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, WI

3.1 Abstract

Engineering novel ligand affinity into allosteric transcription factors has enormous importance in biotechnology, where these proteins can serve as natural biosensors. Efforts to engineer these proteins has been limited due to extensive long-range interactions that create the allosteric response. In the absence of prior knowledge of these interactions, novel function must be found through many empirical measurements of function. Here, we create a novel computational design and high-throughput screening workflow that incorporates ligand-agnostic variants with RNA-Seq to engineer new biosensors. This approach generated variants with affinity to eight nonnative ligands. Sequence analysis of high performing variants revealed distinct sequence profiles for different ligand specificities. We also apply the screening workflow to characterize functional hotspots in a DMS library, revealing important locations at the interface between the ligand binding domain and the DNA binding domain. This workflow can be applied to screen function in any protein whose function is a measurable change in transcription.

3.2 Introduction

The production of value-added chemicals such as proteins, pharmaceuticals, polymer precursors, and biofuels has exploded in part due to laboratory development of novel biosynthetic processes[1-3]. Creating and optimizing novel biosynthetic pathways to the point of commercialization often involves iterative design-build-test-learn cycles[4]. Rational design and random mutagenesis approaches can create thousands of pathway variants. The bottleneck of this approach is testing, where direct measurement of the value-added compound may rely on direct quantification through inherently low-throughput techniques[3]. High-throughput screens using fluorescence reporters and flow cytometry workflows create the potential for testing a vast library for optimization in a single experiment[5]. Transcription factor biosensors are increasingly valuable in this process as these proteins have natural small-molecule sensing capabilities. Furthermore, the utility of transcription factors in biotechnology extends far beyond metabolic engineering as these proteins are also critical in the detection of trace compounds in the environment, as a diagnostic tool, and in the generation of complex gene circuits[6-8].

Allosteric transcription factors (aTFs) control gene expression in response to changes in the environment[9]. Prokayrotic transcription factors such as LacI or TetR have the capacity to detect small molecules and have a simple mechanism of gene expression control, making these proteins logical candidates for biosensing applications[10]. In the absence of the small molecule inducer, the transcription factor remains bound to the operator sequence in the promoter of controlled genes, physically preventing RNA polymerase from interacting with the promoter[11]. aTF binding to the inducer causes an allosteric change that decreases affinity for the operator sequence, allowing downstream gene expression[12].

A major limitation of implementing transcription factor biosensors is the narrow range of molecules that can be bound by characterized transcription factors. While genome mining has been able to identify novel transcription factors for select target molecules, this approach cannot be realistically applied to any arbitrary molecule of interest (Fig. 1a)[13,14]. The capacity to engineer novel ligand specificity into existing, well-characterized aTF scaffolds will greatly expand the uses of these biosensors in metabolic engineering, both as control systems in circuits and sensing platforms for pathway improvement[15,16]. Currently, design approaches have been used to engineer novel affinity for molecules that are structurally like the wildtype ligand[17-19]. Widespread adoption of transcription factor biosensors in biotechnology requires the ability to design known aTFs to bind to a diverse array of small molecules (Fig 1a)[20]. However, designing aTFs for novel ligand affinity has two major challenges: mutating the ligand binding pocket for affinity for the target molecule and maintaining allosteric function.

Computational design is one solution to the challenge of engineering affinity for a target molecule[21-23]. Unlike random mutagenesis, computational approaches are not reliant on iterative experimental workflows as these algorithms search sequence space for optimal interactions *in silico*[24]. This process is able to analyze millions of sequences and generate a subset of mutations that are most likely to increase affinity for the target molecule. However, computational tools cannot account for long-range interactions required for the allosteric changes in response to ligand binding and has had limited success in aTF scaffolds[25].

High-throughput screening approaches are required for searching massive computationally designed libraries to isolate variants that can both interact with the target ligand and maintain allosteric function. While fluorescence-based Sort-Seq approaches can recreate the fluorescence profile of all variants in the library, this approach is limited by the number of gates and the individual fluorescence distribution of each library member[5,26,27]. Furthermore, sorting cannot be scaled across multiple ligands. RNA-Seq is a pooled alternative to fluorescence-based screens to gain quantitative measurement of transcription factor function and has been used to quantify promoter activity and GPCR variant libraries[28,29].

Here, we describe a ligand-agnostic computational design approach coupled with an RNA-Seq workflow for quantitative analysis of transcription factor function. We leverage the evolutionary history of TtgR to create a library of phylogenetically derived, computationally stable amino acid substitutions at key locations in the binding pocket[30]. We screen this library of TtgR variants against nine different ligands for functional aTFs and show that RNA-Seq is also applicable in mutational scanning libraries by screening a TtgR deep mutational scanning (DMS) library against endoxifen and tamoxifen. We find groups of variants change gene expression in response to each ligand and validate top performing variants in a fluorescence-based assay. Furthermore, we show that the variant library contains unique patterns of ligand specificity across all tested ligands, which are reflected by amino acid preferences at important positions. Finally, the DMS screen is able to identify allosterically important regions connecting the DNA binding domain and ligand binding pocket. This work establishes a novel approach to create new transcription factor biosensors that is also pertinent to basic science applications and can be applied to any protein whose functional readout can be quantified via transcription.

3.3 Validating RNA-Seq on a 16-member library

To assay the function of a library of transcription factor variants, we elected to assay transcript quantity directly via RNA-Seq. One of the major challenges was linking the expression of a reporter gene to the transcription factor variant responsible for controlling its expression. Transcription factors can be uniquely identified using the expression of a short, randomized barcode as the reporter gene. To link the transcription factor to the barcode, we created a plasmid that contained both the transcription factor variant and the barcode on the same piece of DNA (Fig. 1b). A second construct was used to map the aTF variant to the barcode with next-generation sequencing (Fig. 1c). Once the transcription factor and barcode pairings are known, transcription factor function is a measure of the abundance of the barcodes during RNA-Seq (Fig. 1c). E. coli containing the plasmid library are dosed with either the target ligand or a vehicle control and harvested in log phase to obtain both total RNA and the library plasmids (Fig. 1d). The RNA provides a measure of function while the plasmids facilitate normalization to prevent library skew from affecting results.

We used a small test library of 16 TtgR variants that have differential response to naringenin[31]. Gene fragments encoding the variants were inserted with random barcodes into our expression vector. Barcodes were mapped to variants in a separate next-generation sequencing run. For barcodes that are mapped to multiple variants, the majority variant was selected if the read counts for each other variant amounted to less than 10% of the read count of the most abundant variant. Each variant in the test library had approximately 8,000 barcodes (Fig. 1e).

To analyze the performance of RNA-Seq on a transcription factor library, we compared RNA-Seq fold enrichment to qRT-PCR fold enrichment. 8 of the 16 variants were isolated from random colony screening of the test library for individual quantification. The test library and clonal variants were dosed with either 1mM naringenin or DMSO as a vehicle control. The RNA-Seq data was subset to barcodes that appeared in all four conditions (naringenin RNA, naringenin DNA, DMSO RNA, and DMSO DNA). The performance of a variant was calculated as the sum of the barcode counts for each variant (Fig. 1f). Comparison of the qRT-PCR fold enrichment and the RNA-Seq data showed high correlation (R^2=0.83) (Fig. 1g). The RNA-Seq approach readily replicates differences observed via qRT-PCR across a range of functions.

A bigger library with 8,000 barcodes per variant will be too large to sequence thoroughly. We hypothesized that down-sampling fewer barcodes per variant will affect the accuracy of the fold enrichment calculation. We used a Monte Carlo sampling approach to randomly select 10, 25, 50, and 100 barcodes per variant for 500 trials each and scored each sample by its correlation to the qRT-PCR dataset (Fig. 1h). Each bootstrap group shows, on average, similar correlation to the qRT-PCR assay compared to the 8,000 barcodes per variant. Thus, larger libraries can be accommodated with a smaller barcode to variant ratio to reduce the sequencing volume requirements.

3.4 Identifying novel sensors in the agnostic library

We tested an agnostic library against a range of small molecules to find new biosensors (Supplementary Fig. 1). We selected four derivatives of tamoxifen (Tam), a breast cancer therapeutic, to create specific and multi-specific sensors. Tamoxifen is a selective estrogen receptor (ER) modulator that is metabolized by cytochrome P450 into 4-hydroxy-tamoxifen (4Hy) and N-desmethyltamoxifen (Ndes). These two metabolites are then catabolized to endoxifen (End). Endoxifen and 4Hy are the most abundant metabolites and show high activity[32,33]. By using sensing platforms for active metabolites of Tam like End and 4Hy, physicians can ensure maximum efficacy during treatment.

We also selected quinine (Quin), naltrexone (Nal), and ellagic acid (EllA) as targets for the agnostic library. Quinine is a small molecule therapeutic used to treat malaria. It is isolated from the bark of the cinchona tree; a sensor for quinine will be useful for creating a biosynthetic pathway in prokaryotes for scalable production. Furthermore, a quinine sensor may be useful for monitoring quinine resistance and variations in pharmacokinetics during treatment[34]. Naltrexone is used to treat addiction as an opioid receptor antagonist[35]. Chemically, it shares many similarities to other compounds that interact with the opioid receptors like morphine and heroin. Naltrexone is chemically distinct from TtgR's native ligands and thus poses a challenging target for affinity engineering. By obtaining an opioid sensor, we can develop portable devices for quick detection of this class of compounds. Ellagic acid is a plant polyphenol with a highly conjugated chemical structure and shares chemical features with native ligands of TtgR (Supplementary Fig. 1).

We obtained a set of computationally stable substitutions at key positions in the ligand binding domain of TtgR using the FuncLib tool and constructed a 17,737-member library comprising of 1-4 mutations[30]. Mapping barcode-variant pairs identified 17,533 variants (98.8%) with an average of 20 barcodes per variant (Supplementary Fig. 2). RNA-Seq of the 16N barcodes corresponding to TtgR variants contained barcodes associated with 17,365 TtgR variants (97.9%). Each ligand had a wide range of functional responses; top performing variants are candidates for novel biosensors (Fig. 2a).

The variants in the top standard deviation for any one ligand were selected for further analysis (red points) (Fig. 2a). These variants performed the best against naringenin and phloretin, the native ligands of TtgR, and had the weakest response to ellagic acid. The top 40 sequences for each ligand were selected for validation in a fluorescence-based assay. The best performers of the naringenin and phloretin ligands shared many variants (Supplementary Fig. 3). Similarly, the tamoxifen, endoxifen, 4-hydroxytamoxifen, and N-desmethyltamoxifen sets shared variants (Supplementary Fig. 3) The quinine, naltrexone, and ellagic acid top performing variants were unique to each ligand (Supplementary Fig. 3).

The 251 unique top variants were cloned into an expression vector containing *sfGFP* under control of the TtgR operator sequence (Fig. 2b). Variants capable of repressing transcription in the absence of any small molecule were first isolated via cell sorting (Supplementary Fig 4). These variants were then exposed to each ligand and the high fluorescence cells were sorted (Supplementary Fig. 5). The performance of these variants is the fold change in percent population of the high fluorescence and repressed sorts.

The fold change of variant abundance in the sorted library indicated that each ligand had a subset of functional transcription factors. However, the library showed no change in fluorescence when inoculated with ellagic acid, suggesting that a fluorescence assay is insufficient for screening this ligand. Variants showed response to all other ligands (Fig. 2c). Although the naltrexone top variants only showed high activity for naltrexone, these variants showed generalized response across most ligands (Fig. 2c). In contrast, quinine top performers are generally specific for quinine (Fig. 2c). Surprisingly, naringenin and phloretin top variants do not show strong response to these ligands, but the low fold change may be due to the large number of variants that were isolated in the high fluorescence sort during the sorting process (Supplementary Fig. 5). The percent abundance in the induced state will thus differ less for each variant compared to the abundance in the repressed state. The cross reactivity of the top performing variants to eight of the nine ligands highlights the broad applicability of this approach. The majority of sequences with high function in the RNA-Seq dataset also show response in a fluorescence-based assay.

3.5 Elucidating ligand-specific sequence preferences from RNA-Seq

The agnostic library was designed to provide a set of stable substitutions without optimizing affinity for any ligand. This approach enables the same library to be screened across multiple ligands, as each aTF variant has the potential to interact with a new subset of small molecules due to the redesigned ligand binding pocket. Given these design constraints, we expect that many variants will have similar ligand specificity profiles. We can leverage the RNA-Seq to gain information about mutations that create and affect function across all aTF variants and ligands. We selected all 16,190 variants (91.2% of all variants) with data for at least 6 of the 9 ligands and imputed the missing data using KNN imputation (see methods). Variants that performed at least 1.5 fold better than wildtype (3,135 variants) on at least one of the nine ligands were selected for hierarchical clustering via the UPGMA algorithm with a correlation distance metric and a target of 12 clusters (Fig. 3a, Supplementary Fig. 6)[36]. The ligand clusters (top dendrogram) are grouped appropriately based on the chemical structure (Supplementary Fig. 1). The tamoxifen ligands are most closely related by their performance. Similarly, naringenin and phloretin, the two native ligands of TtgR, also cluster together. In contrast, the ligands with the most structural diversity (ellagic acid, naltrexone, and quinine) are the most distant.

The variant clusters (left dendrogram) display unique sequence specificity profiles based on the normalized fold enrichment across the 9 ligands. Cluster 1 (blue, top) is characterized by higher 4-hydroxytamoxifen and endoxifen normalized fold enrichment. The third cluster (green) contains variants with high naltrexone response. The fourth cluster (red) is primarily composed of variants with high quinine normalized fold enrichment. Cluster 7 (pink) is characterized by variants with high N-desmethyltamoxifen and tamoxifen normalized fold enrichment.

We wanted to characterize the sequence profiles of the cluster members to understand the important substitutions that contribute to the unique specificity profile of each cluster. The agnostic library is a combination of selected mutations across a limited number of positions in the binding pocket with potential to directly interact with small molecules (Fig. 3b). We calculated the relative positional entropy of all mutable positions for each cluster in comparison to the entire 16,190 sequences (see methods). Relative positional entropy quantifies the change in amino acid distribution after clustering. In this context, high relative entropy indicates that certain amino acids are preferred at a particular position once clustered. Each cluster except cluster 2, cluster 10, and cluster 11 has one or more positions with high selective pressure (Fig. 3c). L113 and H114 show low relative positional entropy across all clusters, suggesting that these two positions do not contribute to unique protein-ligand interactions (Fig. 3c). These two residues are located at the bottom of the binding pocket and have the potential to make hydrogen bond and nonpolar interactions with small molecules that are oriented in the native "vertical" binding pose for many ligands in wildtype TtgR (Fig. 3b). Surprisingly, N110 is adjacent to L113 and H114 and has the potential to make similar interactions with ligands at the bottom of the binding pocket as L113 and H114, but has high entropy in Cluster 7 (Fig. 3b, 3c). N-desmethyltamoxifen and tamoxifen thus interact with this position in preference over L113 and H114. All other positions make up one face of the ligand binding pocket and have at least one cluster that has high relative positional entropy.

Comparing amino acid sequence preferences at high selectivity positions across clustered sequences and top performing variants for each ligand can identify distinct amino acid trends for each ligand. Some clusters are very similar to one another across the nine ligands. Clusters 7, 11, and 12 all share high naringenin and phloretin fold response but differing tamoxifen response profiles (Fig. 3a). Cluster 7 has higher N-desmethyltamoxifen and tamoxifen response while cluster 12 has higher 4-hydroxytamoxifen and endoxifen response (Fig. 3a). Cluster 11 shows high naringenin and phloretin response but does not respond to any of the tamoxifen ligands. We examined the positions in clusters 7, 11, and 12 in the top 80th percentile of relative positional entropy to identify large changes in amino acid distribution before and after clustering. Cluster 7 has two positions with large changes: 92 and 110. Cluster 11 has no positions with high relative entropy. Cluster 12 shows high relative positional entropy at 67, 78, and 92. In Cluster 7, position 92 favors the wildtype leucine substitution over all possible mutations. Only the best tamoxifen variants also show this behavior (Fig. 3d). N-desmethyltamoxifen, tamoxifen, and phloretin favor the valine substitution while naringenin does not. In contrast, naringenin favors alanine, isoleucine, or methionine at 92 (Fig. 3d). These differences highlight the generalizable approach of the agnostic library.

Selecting amino acid substitutions from a limited set at a limited number of positions can drastically change ligand response. Cluster 11, which contains variants with high response naringenin, phloretin, and quinine, contains no positions with high relative entropy. Since naringenin and phloretin are two of the native ligands of TtgR and every position has a high percentage of wildtype residues due to the limited number of mutations in the library, many variants show little change in the amino acid distributions in response to clustering (Fig. 3d). The small change in amino acid distribution before and after clustering across the positions implies that wildtype response is largely maintained regardless of amino acid substitution. Often, the change in amino acid frequency in the clusters are matched by the changes observed in the top performers of associated ligands. For example, cluster 7, which has high N-desmethyltamoxifen and tamoxifen shows similar amino acid preferences as the top variants of N-desmethyltamoxifen. The amino acid preferences of each cluster and the top performing variants imply that the sequence-based analysis can highlight key functional substitutions that are associated with high ligand response.

The agnostic design scheme of the aTF library creates a set of variants with the potential to interact with a wide variety of ligands while having a constrained set of mutations that are selected for stability. By clustering the variants based on their performance across ligands, we can gain insight into the amino acid preferences at each position and an initial understanding of the importance of each in conferring novel ligand affinity. Comparing the sequence profiles of clusters with shared naringenin and phloretin performance highlights the unique solutions that can give rise to novel patterns of ligand specificity. The library thus contains variants with unique sets of ligand specificity, creating a wide variety of potential sensors with tunable response profiles.

3.6 DMS of TtgR against endoxifen and tamoxifen highlights functional hotspots

One benefit of an RNA-Seq based screening workflow is that many libraries can be screened in parallel to create a functional landscape of variants. We tested a deep-mutational scan library of TtgR consisting of single point mutations of every position to all 19 other amino acids against tamoxifen and endoxifen[37]. The library was split into 6 different pools spanning amino acids 1-39, 40-77, 78-115, 116-153, 154-191, and 192-210. The majority of single point mutations have little effect on function, with 3,897 variants in the endoxifen dataset and 3,996 variants in the tamoxifen dataset between 1.2 and 0.8 of wildtype fold enrichment (Fig 4A, 4B).

The DNA binding domain has regions of high and low function across multiple substitutions. The helix of the HTH motif that directly interacts with the major groove of the operator sequence contains many positions where the majority (61%) of mutations decrease aTF function (Fig. 4a, 4b). Substitutions at positions between 81 and 101 are often detrimental (approximately 53%) to protein function than the wildtype residue. These positions correspond to a solvent-facing region of a helix in the ligand binding pocket, suggesting an allosteric role in gene expression control. In contrast, 65% of mutations to positions between 116 and 153 confer increased function (Fig. 4a, 4b). These positions compose a single helix of the ligand binding pocket, but high-performing positions are agnostic of orientation on the helix. Surprisingly, the helix associated with dimerization (186-210) also contains many positions where the majority of substitutions (approximately 64%) are beneficial to protein function.

The variants that show high performance or low performance may be indicative of hotspots that are critical for protein design. Hotspots were selected based on the number of mutations whose performance fell outside the interquartile range of fold enrichment normalized to wildtype performance (see methods). The majority of identified hotspots are near the DNA binding domain and the interface between the DNA binding domain and the ligand binding domain.

Positions within the DNA binding domain can be classified as solvent interactions, DNA interactions, or potential allosteric interactions (Fig. 4c). T5, K6, A9, and R27 are located in the first helix of the DNA binding domain and do not have any significant interactions with the operator sequence, the ligand, or the other TtgR monomer. The high mutability of positions 5, 6, 9, and 27 may be because these positions have nonspecific interactions with the solvent. A38 is located in the recognition helix that interacts with the major groove of the operator sequence and has a direct effect on the capacity of TtgR to repress gene expression. Six of the seven positions at A38 drastically reduce function. A19, A23, A30, and R31 have possible structural or allosteric functions based on their location. A19 and A23 make van der Waals interactions with I37 and L40 in the recognition helix and possibly stabilize the position of the recognition helix conformation. These positions may also be important for the allosteric changes that occur in response to ligand binding that decrease affinity for the operator sequence. The backbone amide of A30 makes a hydrogen bond with D118. Mutating A30 to polar residues increases function by creating additional interactions with A30 and R31 on the opposite monomer. R31 makes hydrogen bond interactions with T120 and D122 on the opposite monomer. The hydrogen bond interactions are substituted for van der Waals interactions in the majority of mutations at this position.

The positions in the ligand binding domain either directly interact with the ligand or interface with the DNA binding domain. L93 and N110 directly interact with the ligand and are mutable positions in the agnostic library. Mutating position 93 results in a loss of function for the majority of mutations. However, tamoxifen and endoxifen yield different sets of mutations that improve function. In contrast, mutations at position 110 largely improve function with the exception of the cysteine mutation, which is consistent across both ligands. I112 is located at the interface of the ligand binding domain and the DNA binding domain. The side group of 112 is located between F24 and Y25 in the DNA binding domain. The size and polarity of this position is important as only the leucine substitution increases function. Any substitution to residues with different shape or polarity results in a loss of function. D84 is a solvent-accessible position located in a loop in the ligand binding pocket. Selecting positions that have multiple substitutions that confer either high or low function via the RNA-Seq approach identifies functional hotspots in TtgR in the DNA binding domain and the ligand binding domain. The importance of these hotspots can be rationalized based on their location in the structure of TtgR.

3.7 Discussion

Transcription factor biosensors have an important role in developing novel metabolic engineering pathways[4]. However, creating new biosensors with affinity for any desired target molecule is challenging because allosteric properties of the aTF must also be maintained. To solve this challenge, we created an agnostic library of TtgR variants and used RNA-Seq to screen for affinity to tamoxifen, endoxifen, N-desmethyltamoxifen, 4-hydroxytamoxifen, naringenin, phloretin, naltrexone, ellagic acid, and quinine.

Screening the agnostic library with RNA-Seq to quantify barcode expression revealed multiple variants with high performance to each ligand. These top performers were validated using a fluorescence-based cell sorting scheme. The fold change in variant abundance resulting from the sorting workflow showed that these variants had a range of both specificities and activity against eight of the nine ligands. Some variants, like those that performed best on quinine, were largely specific for that one molecule. Others, like naltrexone top performers, often responded to multiple ligands.

Although the agnostic library is not designed for affinity to any one ligand, our results suggest that unique ligand specificities arise from screening the same ligand-agnostic library against multiple ligands. Even a small subset of mutations in the allowed set enables drastic function-switching phenotypes across dissimilar ligands. Variants with similar ligand specificity profiles can be found using hierarchical clustering and we show that each cluster has unique amino acid compositions at critical positions that may enable function switching.

We have also shown that the RNA-Seq workflow can be applied to functional screens the application of a DMS library to endoxifen and tamoxifen. Hotspots were identified in the DNA binding domain and the ligand binding domain. Some of these positions, like T5, K6, A9, and R27, are solvent-exposed and likely have a wide range of tolerable mutations. Others are in direct contact with either the DNA or the ligand and thus have a large number of mutations that decrease function. The last group lies at the interface of the DNA binding domain and the ligand binding domain, indicating a potential allosteric or a structural role in function.

These results validate an RNA-Seq based approach for assaying transcription factor function. However, a similar workflow can be adapted to any protein whose function results in altered gene expression. This work establishes a base for expanded aTF usage in biotechnology by increasing the design potential of well-characterized transcription factors. In the future, improvements of this technology will enable designer aTFs to be created for any desired small molecule.

Despite the initial potential of this method, we acknowledge that additional data is required to completely validate this design and screening system. Although initial RNA-Seq and fluorescence screens have both shown changes in gene expression across top performing variants for multiple ligands, clonal assays of each variant via transcript measurement (qRT-PCR) or GFP fluorescence (flow cytometry) will prove that the responses observed in the RNA-Seq and fluorescence screens are real. Crystal structures of the top hits will reveal the molecular interactions responsible for the altered binding specificities. Finally, the RNA-Seq data can be applied to improve the accuracy of targeted Rosetta design methods by using machine learning to update the scoring process used to evaluate designed sequences. Improving the computational design process will decrease the number of variants tested enabling additional ligands to be assayed for the same cost.

New sensors have been found for the majority of ligands tested, indicating that TtgR is an extraordinary scaffold that is amenable to the acquisition of many novel functions. This functional plasticity may be a byproduct of the evolutionary history of TtgR. TtgR naturally controls the expression of TtgABCD in *pseudomonas putida*[38]. These proteins form subunits of a multidrug exporter; TtgR must also be able to sense multiple ligands in order to control exporter expression[39]. The ability to interact with multiple ligands has been posited as a key characteristic of more evolvable proteins[40]. Thus, future biosensor design efforts may warrant additional focus on scaffolds that already have multifunctionality. QacR is another aTF that can bind to a wide variety of ligands and is a potential candidate for redesign towards novel ligand specificity[41].

Acknowledgments

This research was performed using the compute resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. This research was performed using the sequencing resources of the UW-Madison Biotechnology Center DNA Sequencing facility. K.K.N. is supported by NIH National Research Service Award T32 GM07215 and the Robert and Katherine Burris Biochemistry Fund.

Competing Interests

We declare no competing interests.

Author Contributions

K. K. N. and S. R. designed the study, analyzed the data, and wrote the manuscript. K. K. N. performed all experiments.

Code availability

All figures were generated using the Matplotlib and the Seaborn module in Python 3.9[49]. Initial NGS analysis was performed with PEAR, NGmerge, fastp, UMI-Tools, and BBTools[42-44]. Scripts used in data analysis of NGS data and figure generation can be found at: https://github.com/raman-lab/rna-seq. SciKit-Learn was used for KNN imputation. SciPy was used for hierarchical clustering[47]. Clustering was visualized using Seaborn. Flow cytometry data were analyzed in FlowJo V10.

3.8 Methods

General cloning methods

Plasmid creation:

Amplicons are generated using Kapa HiFi (Roche) PCR kits following the manufacturer protocol (Supplementary Table 1). Amplicons are treated with 15U of Dpn1 (NEB) for 2.5 hours at 37°C followed by 20 minutes at 80°C. PCR amplicons are then purified using EZNA Cycle Pure kits (Omega BioTek). Isothermal assembly followed Gibson Assembly protocols (NEB), but contained 100 mM Tris-HCl pH 7.5, 20 mM $MgCl_2$, 0.2 mM dATP, 0.2 mM dCTP, 0.2 mM dGTP, 10 mM dTT, 5% PEG-8000, 1 mM NAD+, 4 U/ml T5 exonuclease, 4 U/µl Taq DNA ligase, and 25 U/ml Phusion polymerase. Isothermal assembly reactions are diluted 10X in dH_2O prior to transformation. DH10B (NEB) electrocompetent cells are transformed with 2µL of diluted isothermal assembly reaction. Transformants are recovered in 700µL SOC for 1 hour at 37°C. Dilutions are plated on LB-kanamycin (50µg/mL) plates and incubated at 37°C overnight. Colony PCR is performed using Kapa Robust (Roche) using a single colony diluted in 100µL of dH_2O. Plasmid purifications are performed using the ZR Plasmid Miniprep Classic kit (Zymo).

Library creation:

Plasmid libraries are generated using Golden Gate Assembly Kits (NEB, BsaI-HFv2). The reactions undergo a cycling protocol of 30 alternating 5-minute 37°C and 16°C cycles followed by a final 60°C 5-minute hold. The reactions are dialyzed against dH_2O on semi-permeable membranes (Millipore) for 1 hour at room temperature. DH10B (NEB) cells were transformed with 3μL of dialyzed reaction via electroporation. Transformants were recovered in 1mL of SOC and then diluted 2X, 5X, and 10X with fresh SOC. Each dilution recovered for 1 hour shaking at 37°C. 4mL of LB-kanamycin (50μg/mL) was added to each dilution after recovery and 50X and 500X dilutions were plated of each recovered dilution to calculate transformation efficiency. The remaining transformants were grown for 6 hours shaking at 225rpm. A frozen stock was made in 25% glycerol and stored at -80 for each dilution. Fresh cultures were created by diluting each 6-hour growth 50X into fresh LB-kanamycin. These were grown overnight and plasmids were harvested via ZR Plasmid Miniprep Classic kit (Zymo).

RNA purification:

Cells were struck out on an LB-Kan plate and grown overnight at 37°C. Three colonies were inoculated into LB-kanamycin for overnight growth. The overnight cultures were diluted 50X into fresh LB-kanamycin containing either ligand or solvent. The cultures were grown at 37°C shaking at 250rpm in an Innova 4230 (New Brunswick Scientific). At the targeted OD_{600}, cultures were placed on ice for 10 minutes. $5*10^8$ cells were harvested by centrifugation at 5,500*g* based on the OD_{600} and the assumption that 1.0 OD_{600} cultures have $8*10^8$ cells/mL. The pelleted cells were decanted and stored at -80°C. This process was repeated in biological triplicate for each target OD_{600} with new colonies.

RNA was purified from cell pellets via Trizol reagent (Invitrogen). 1mL of Trizol reagent (Invitrogen) was added to each cell pellet and vortexed briefly. The samples incubated at room temperature for 5 minutes. 200μL of chloroform (Sigma Aldrich) was added to each sample. The samples incubated at room temperature for 2 minutes and were centrifuged at 12,000*g* for 15 minutes at 4°C. 300μL of the aqueous phase was transferred to a clean 2mL centrifuge tube and placed on ice. RNA was purified from the aqueous phase using the RNA Clean and Concentrator 5 kit (Zymo) and eluted in 15μL Ultrapure RNase-free dH2O (Invitrogen). The purified RNA was digested using 4U DnaseI (NEB) in a 50μL reaction incubated at 37°C for 30 minutes. The digestion reactions were purified using the RNA Clean and Concentrator 5 kit (Zymo) and eluted in 15μL Ultrapure RNase-free dH_2O (Invitrogen). Concentrations were measured using a Nanodrop instrument (Thermo Fisher).

qRT-PCR quantification of transcript abundance:

The abundance of the *sfGFP* and *rrsA* transcripts were measured via qRT-PCR. Each biological triplicate RNA was run in technical triplicate in a MicroAmp Fast Optical 96-well plate (Life Technologies). 1ng of RNA was added to Luna Universal One-Step qRT-PCR mix (NEB) containing 4μmol of each primer on ice. The standard cycling protocol was used according to the manufacturer's suggestion. Each sample consisted of a set of reactions containing *sfGFP*-specific primers and another set containing *rrsA*-specific primers. The reactions were run on a CFX Connect Real Time PCR Detection System (BioRad).

Fold enrichment was calculated using equations (1) and (2). The error was propagated from the technical replicates and biological replicates using (3).

$$fold\ enrichment = 2^{-\Delta\Delta C_t}\ (1)$$

$$\Delta\Delta C_t = (C_{t\,GFP} - C_{t\,rrsA})_{+Ligand} - (C_{t\,GFP} - C_{t\,rrsA})_{-Ligand}\ (2)$$

$$error = \sqrt{\sum(\sigma_i)}\ (3)$$

Short barcode oligo synthesis:

Pre-defined or random barcodes were synthesized as a short primer (IDT). These barcode primers were combined separately with another constant primer to create short double-stranded fragments containing the barcode flanked by BsaI cut sites in a single cycle of PCR using Kapa HiFi (Roche). 1μL of this reaction was added into a second Kapa HiFi (Roche) reaction with additional primers to increase the length of the amplicon over 18 cycles. The resulting amplicon was purified using the DNA Clean and Concentrator-5 kit (Zymo).

Barcode-variant mapping via next-generation sequencing:

Two primer groups were used to add Illumina sequencing regions to the barcode-spacer-variant region of the mapping plasmid libraries. Each primer group consisted of three primers with different numbers of Ns (0N, 3N, or 6N) to increase positional base diversity during runs. The adapter primers had complementarity to the plasmid and contained Illumina sequencing primer binding regions. Stem primers had the i7 and i5 indices and the adapter sequence to anneal to the sequencing flow cell. The adapter regions were added using 1ng of template, 0.6μL of 10μM primers, and Kapa HiFi mix (Roche) for 14 cycles. These reactions were purified using the DNA Clean and Concentrator 5 kit (Zymo). The stem primers were used in a second PCR reaction using 4μL of the first reaction for 10 cycles.

Sample preparation for sequencing:

For MiSeq-based sequencing, the proper band was isolated using gel extraction on a 0.5% agarose gel followed by purification with the EZNA gel extraction kit (Omega BioTek). The concentration of the DNA was measured using AccuClear (Biotium) following manufacturer protocols. The flow cell was loaded with 15pM DNA with 5% PhiX. For NovaSeq-based sequencing, samples were purified using PippinHT (Sage Science) and the concentration was measured via 4200 TapeStation (Agilent). The size selection, concentration measurement, and NovaSeq runs were performed by the University of Wisconsin Madison Biotechnology Center (UWBC).

Mapping Data Analysis:

The FastQ output was merged using PEAR. A C++ script was used to filter poor-scoring reads based on Q-scores. Reads that passed the quality filter were then filtered on constant regions surrounding the barcode and TtgR variants. Barcodes that had read counts greater than 10 and were unique for a single TtgR variant were mapped to a that variant. If a barcode mapped to more than one TtgR variant, then the TtgR variant that had the most reads was selected if each other variant was less than 10% of the reads of the most abundant variant.

RNA-Seq preparation:

RNA is harvested according to the RNA purification protocol. cDNA synthesis uses approximately 3μg total RNA, a primer encoding a 16nt unique molecular identifier (UMI), and the Maxima H Minus Double-Stranded cDNA Synthesis Kit. The cDNA is purified using the DNA Clean and Concentrator 5 kit (Zymo). The Illumina sequencing regions are added in 2 PCR reactions in the same manner as the MiSeq barcode-variant mapping reactions. Three sets of primers containing the Illumina sequencing primer and a predefined barcode (ATCG, CGAT, and GTCA) were used in the first PCR reaction to add the Illumina sequencing regions (11 cycles). One set of primers was used for each biological replicate. The first reaction is purified using the DNA Clean and Concentrator 5 kit (Zymo). The second reaction uses 4μL of the first reaction and primers that add i5 and i7 indices in 8 cycles. The final amplicons are purified again. All replicates were combined in an equal molar ratio after purification.

Plasmids are harvested from the remaining culture of the RNA preparation step. The UMI is added to the plasmid-derived samples in a 2-cycle PCR reaction using 100ng of template. The amplification of all DNA libraries followed an identical protocol to the RNA preparation.

The cDNA and DNA samples are sequenced using either a NovaSeq SP chip (test library) or a NovaSeq S4 chip (DMS and agnostic libraries) by the UWBC.

RNA-Seq Data Analysis:

Fastq files were merged using NGmerge and filtered using Fastp based on average Q-score > Q30 for reads[42,43]. Reads containing the 5' and 3' constant regions were isolated using UMI-Tools and counted using Tally[44,45]. Reads containing the central constant region were isolated and UMI sequences were removed with UMI-Tools. The barcodes were then counted with Tally. RNA-Seq barcodes were matched to mapped barcode-variant pairs with a Hamming distance tolerance of 1 using Seal (sourceforge.net/projects/bbmap/).

RNA-Seq barcodes that were successfully mapped to known barcode-variant pairs were analyzed across the induced RNA, induced DNA, control RNA, and control DNA samples. A barcode both had to be found in all four datasets to be included in downstream analysis. The read counts for a variant were then a sum of the barcode counts for all barcodes mapped and found in all four datasets. No read count threshold was imposed during analysis. The fold enrichment calculation uses equation (4).

$$Fold\ Enrichment = \frac{\frac{RNA_{+Lig}}{DNA_{+Lig}}}{\frac{RNA_{-Lig}}{DNA_{-Lig}}} \text{ (4)}$$

If biological replicates were available for each condition, the fold enrichment per variant was curated based on the coefficient of variation (CV). Percent deviation is calculated with equation (5).

$$CV = \frac{\sigma}{\bar{x}} (5)$$

In this equation, σ is the standard deviation of the fold enrichment and $\bar{x}$ is the mean fold enrichment across replicates. A 30% CV cutoff was imposed for the agnostic dataset and a 20% deviation cutoff was imposed on the DMS dataset (Supplementary Fig. 7, Supplementary Fig. 8, Supplementary Fig. 9).

All variants were normalized to wildtype fold enrichment for each replicate. Heatmaps were constructed using the average performance of each variant after normalization.

Cell Sorting:

An overnight culture is diluted 50X in phosphate buffered saline (137mM NaCl, 2.7mM KCl, 10mM Na_2HPO_4, 1.8mM KH_2PO_4) and placed on ice for 10 minutes prior to sorting. Sorting was performed on an SH800 (Sony) using the 488nm laser and a 525±25 filter. Sorted cells were grown for 1 hour shaking at 37°C in 5mL LB. Kanamycin was added to a final concentration of 50μg/mL and the culture was grown overnight. An aliquot of the sorted culture was stored at -80°C in 25% glycerol. Plasmids were isolated from the remaining culture using the ZR Plasmid Miniprep – Classic kit (Zymo).

Creating TtgR_pBBR1_SPS_V2:

The plasmid containing the *TtgR* gene and the *sfGFP* gene under control of the TtgR operator sequence was created using two Gibson Assembly reactions. The *sfGFP* gene was under control of a modified TtgR operator sequence with canonical -10 (5'-TATAAT-3') and -35 (5'-TTGACA-3') elements in the promoter. The backbone contains the *TtgR* gene under an apFAB61-BBaJ61132 constitutive operator sequence, a kanamycin resistance marker, and the pBBR1 origin (TtgR_pBBR1). The *sfGFP* gene was inserted into the pBBR1 backbone following the standard methodology. Next, a terminator was placed at the 3' end of the *sfGFP* gene according to protocol. This plasmid was labeled as TtgR_pBBR1_SPS_V2.

Creating TtgR_ColE1_SPS_V5:

The pBBR1 origin was exchanged for a ColE1 origin. The *sfGFP* fragment was amplified from TtgR_pBBR1_V2 using primers specific for the *sfGFP* region with 5' ends complementary to the destination ColE1 backbone and to the *sfGFP* amplicon. The *TtgR* gene was amplified from the TtgR_SC101BBa plasmid with primers containing complementary regions to the backbone and GFP amplicon. The *sfColE1* backbone amplicon contains a kanamycin marker and a ColE1 origin. Plasmids were labeled as TtgR_ColE1_SPS.

The *sfGFP* promoter was modified to have the wildtype TtgR operator sequence. *sfGFP* with the wildtype operator sequence was amplified from a separate plasmid using primers with overlap to the TtgR_ColE1_SPS plasmid. The backbone amplicon was amplified from TtgR_ColE1_SPS and consisted of the TtgR gene, the Kanamycin resistance marker, and the ColE1 origin. The plasmid was labeled as TtgR_ColE1_SPS_V2.

A third Gibson assembly reaction was required to insert stop codons and BsaI cut sites into the middle of the GFP gene to create the barcode insertion site. The stop codons and BsaI sites were encoded on overlapping primers and added to the TtgR_ColE1_SPS_V2 plasmid. The backbone as annealed to itself in a 1-part isothermal assembly. This construct was labeled TtgR_ColE1_SPS_V5.

Creating GFP control:

To create a GFP positive control, the *TtgR* gene was removed from the TtgR_ColE1_SPS_V2 plasmid. The backbone was amplified with primers that had complementary overlap with the *sfGFP* gene. The *sfGFP* gene was amplified with primers complementary to the backbone. The BsaI cut sites and early stop codons were inserted into *sfGFP* in the same fashion as the creation of TtgR_ColE1_SPS_V5. The plasmid was labeled TtgR_ColE1_SPS_V3_GFPControl. Three pre-defined 20nt barcodes (AAACCCTGTGCCAGAGGGTG, GAGTGACCTTAAGTCAGGGA, and GCTTCTGTCCAAGCAGGTTA) were generated according to standard protocols. The barcodes were inserted into the TtgR_ColE1_SPS_V3_GFPControl using Golden Gate assembly.

OD_{600} optimization:

mRNA levels were assayed at three different OD_{600} values: 0.6, 1.2, and ~2.8 (overnight growth). Testing was performed with the TtgR_ColE1_SPS_V2 plasmid with primers specific for the 5' region of *sfGFP*. *rrsA*, a ribosomal subunit and constitutively expressed gene, was used as a reference. 2mL cultures were grown and RNA harvested according to standard protocols. The abundance of the *sfGFP* and *rrsA* transcripts were measured via qRT-PCR following the standard protocol (Supplementary Fig. 10a).

Length Optimization:

The OD_{600} 0.6 induction samples were used to test the effect of different amplicon lengths on qRT-PCR fold enrichment. One of the TtgR_ColE1_SPS_V2 samples assayed with DMSO was used to calculate the primer efficiency of three different primer pairs. Each pair shared the same forward primer but had differing reverse primers that yielded amplicon lengths of 75bp, 150bp, and 300bp. 0.001ng, 0.01ng, 0.1ng, or 1ng of RNA was added to Luna Universal One-Step qRT-PCR mix (NEB) containing 4μmol of each primer on ice. These RNA amounts were also assayed with the *rrsA* primers in the same manner. The abundance of the *sfGFP* and *rrsA* transcripts were measured via qRT-PCR following the standard protocol (Supplementary Fig. 10b).

Creating TtgR Test Library:

The TtgR gene variants were isolated from a set of 16 pre-existing plasmids each containing a single TtgR variant. 100ng of each amplicon was combined into a single aliquot and stored at -20°C. Barcodes for the RNA-Seq were 16nt in length and were encoded on a ssDNA primer (IDT). The TtgR_ColE1_SPS_V5 backbone was amplified using primers that encompassed the *sfGFP* gene, the ColE1 origin, and the kanamycin resistance marker. The barcodes, *TtgR* gene variants, and backbone were assembled in a single Golden Gate reaction (NEB) according to standard protocols.

Mapping test library barcode-variant pairs:

A 60nt spacer was created to bring the random barcode and TtgR variants physically adjacent on the same plasmid to enable short read next generation sequencing mapping of barcode-variant pairs. The test library plasmids were amplified with primers encoding BsaI cut sites that would place the spacer between the barcode and the TtgR variant region. The spacer was inserted into the backbone using Golden Gate (NEB) following standard protocols. The resulting library was sequenced on a 15M 2x250 MiSeq chip (Illumina). Data analysis followed standard protocols.

RNA-Seq of test library:

The induction of the test library used either 1mM naringenin or DMSO as a control. DH10B containing each barcoded GFP Control plasmid were struck out on LB-kanamycin plates. One colony was selected from each barcoded DH10B and grown in 3mL LB-kanamycin overnight. These barcoded control cultures were combined in equal ratio and added to the test library culture to a final composition of 0.25% control. The induced cultures were grown and prepared following standard protocols.

Validating RNA-Seq test library results:

The test library frozen stock was struck out on LB-kanamycin and grown overnight. 16 colonies were selected and the TtgR variants were identified using colony PCR per standard protocols. 8 of the 16 total variants were verified via sequencing and were stored at -80°C. RNA was harvested from each variant in biological triplicate under 1mM naringenin and DMSO conditions. qRT-PCR was used to assay barcode transcript enrichment.

Creating agnostic libraries:

FuncLib-tolerated mutations were encoded into short oligos (Agilent) consisting of the TtgR gene region flanked by BsaI cut sites for Golden Gate assembly. Four pools of approximately 4,400 variants were created by randomly combining between 1 and 5 tolerated mutations. Each pool had unique priming sequences to isolate from a pooled sample. The pooled library was diluted to 0.005μM in Tris-HCl (pH 7.5). Each pool was amplified using Kapa HiFi and 1μL of the diluted library in 15 cycles in triplicate. The amplified reactions were pooled together and purified using the DNA Clean and Concentrator 5 kit (Zymo). The pooled oligos were cloned into the TtgR_ColE1_SPS_V5 backbone using Golden Gate assembly (NEB). The libraries with approximately 15 barcodes per variant, calculated by CFU/mL, were selected for RNA-Seq.

Mapping agnostic Libraries:

The mapping process was performed as described in the general cloning methods. The spacer library was sequenced using an 2x250 NovaSeq SP chip (Illumina) by the UWBC.

Agnostic RNA-Seq:

The agnostic libraries were induced with the ligand (Supplementary Table 2). DMSO, dH_2O, and EtOH were included as solvent controls. The four pools were grown individually in 5mL LB-kanamycin overnight in triplicate. The four pools were combined prior to inoculation in 25mL LB-kanamycin for the RNA harvest. GFP Control barcoded cells were spiked into the combined agnostic replicates at a final concentration of 0.25%. The same pooled replicates were used for all ligand inductions. Read volumes were calculated by targeting 500 reads per barcode with the assumption that 50% of the reads will be lost due to filtering criteria.

RNA-Seq data analysis:

Data analysis followed the RNA-Seq pipeline described above. Variants with data passing CV thresholds for more than 5 ligands and performed at least 1.5 times better than wildtype were selected for clustering. Missing data was imputed using KNN methods in SciKit Learn[46]. The UPGMA algorithm with a correlation distance metric and a target of 12 clusters was used to cluster in SciPy[36,47]. The number of clusters was selected by plotting the silhouette score against the number of clusters (Supplementary Fig. 6). The relative positional entropy was calculated for each cluster compared to the total set of variants using equation (6)[48]:

$$RE = \sum_a f_{cluster,a} \left(\frac{f_{cluster,a}}{f_{all,a}} \right) \text{(6)}$$

In this equation, a is the set of all amino acids observed at a single position and f is the frequency with which that amino acid is observed. This equation compares clustered sets of sequences compared to all possible agnostic sequences. This equation was only applied to clusters with more than 20 sequences.

Testing Top Hits:

Top performing variants were selected based on the mean rank of each variant across the three biological replicates. These variants were encoded in gene fragments (Twist) and synthesized in a 96-well plate format. The fragments were resuspended to a final concentration of 10ng/μL, pooled together, and cloned into the TtgR_ColE1_SPS_V2 backbone using Golden Gate Assembly. The resulting library was sorted based on fluorescence.

LB media is inoculated with 50μL of the frozen stock of the library and grown overnight shaking at 37°C. Sorting was performed according to the Cell Sorting protocol. 500,000 cells were isolated of the lower 70% of the population based on fluorescence. Plasmids were isolated from the remaining culture using the ZR Plasmid Miniprep – Classic kit (Zymo). DH10B (NEB) were transformed with the purified plasmid library according to the Library Creation protocol.

LB media is inoculated with 50μL of the frozen stock of the repressed library and grown overnight shaking at 37°C. The culture was diluted 50X into fresh LB and grown overnight at 37°C shaking with the ligands (Supplementary Table 2). Sorting was performed according to the Cell Sorting protocol. 400,000 cells were isolated using a gate that encompassed the top 0.5% of the population based on the fluorescence distribution in the absence of any ligand. Plasmids were isolated from the remaining culture using the ZR Plasmid Miniprep – Classic kit (Zymo).

The abundance of variants was determined using next-generation sequencing. Sequencing amplicons were generated using primers that had complementarity to the *TtgR* gene around the gene fragment insertion site. The amplification process followed the "Barcode-variant mapping via next-generation sequencing" protocol. The concentration of the DNA was measured using Qubit Fluorometric Quantification (Thermo Fisher) following manufacturer protocols. The flow cell was loaded with 15pM DNA with 5% PhiX. Sequencing was performed on a MiSeq instrument (Illumina).

Fastq files were merged using NGmerge and filtered using Fastp based on average Q-score > Q30 for reads[42,43].

DMS Library synthesis:

The TtgR DMS libraries were created from pre-existing TtgR DMS libraries. This DMS library was split into 6 different segments that encompassed the length of the TtgR gene. Each segment was a separate plasmid library. These segments were amplified separately and cloned into the TtgR_ColE1_SPS_V5 backbone using Golden Gate assembly (NEB). Libraries with approximately 10 barcodes per variant were selected for RNA-Seq.

Mapping DMS Libraries:

The mapping process was performed as described in the general cloning methods. The spacer library was sequenced using an 2x250 NovaSeq SP chip (Illumina) by the UWBC (Supplementary Fig. 11).

DMS RNA-Seq:

The DMS libraries were induced with the either 50μM tamoxifen, 50μM endoxifen, or EtOH. The six pools were grown individually in 5mL LB-kanamycin overnight in triplicate. Each segment was induced separately. Read volumes were calculated by targeting 500 reads per barcode with the assumption that 50% of the reads will be lost due to filtering criteria.

Data analysis followed the RNA-Seq pipeline described above. The 90th percentile of positions by number of mutations outside the interquartile range of all variants were defined as functional hotspots.

3.9 References

1 Rogers, J. K. & Church, G. M. Genetically encoded sensors enable real-time observation of metabolite production. *Proc Natl Acad Sci U S A* **113**, 2388-2393, doi:10.1073/pnas.1600375113 (2016).

2 Dellomonaco, C., Clomburg, J. M., Miller, E. N. & Gonzalez, R. Engineered reversal of the beta-oxidation cycle for the synthesis of fuels and chemicals. *Nature* **476**, 355-359, doi:10.1038/nature10333 (2011).

3 Lin, Y., Shen, X., Yuan, Q. & Yan, Y. Microbial biosynthesis of the anticoagulant precursor 4-hydroxycoumarin. *Nat Commun* **4**, 2603, doi:10.1038/ncomms3603 (2013).

4 Rogers, J. K., Taylor, N. D. & Church, G. M. Biosensor-based engineering of biosynthetic pathways. *Curr Opin Biotechnol* **42**, 84-91, doi:10.1016/j.copbio.2016.03.005 (2016).

5 Rohlhill, J., Sandoval, N. R. & Papoutsakis, E. T. Sort-Seq Approach to Engineering a Formaldehyde-Inducible Promoter for Dynamically Regulated Escherichia coli Growth on Methanol. *ACS Synth Biol* **6**, 1584-1595, doi:10.1021/acssynbio.7b00114 (2017).

6 Wang, B., Barahona, M. & Buck, M. A modular cell-based biosensor using engineered genetic logic circuits to detect and integrate multiple environmental signals. *Biosens Bioelectron* **40**, 368-376, doi:10.1016/j.bios.2012.08.011 (2013).

7 Wan, X. *et al.* Cascaded amplifying circuits enable ultrasensitive cellular sensors for toxic metals. *Nat Chem Biol* **15**, 540-548, doi:10.1038/s41589-019-0244-3 (2019).

8 Bashor, C. J. *et al.* Complex signal processing in synthetic gene circuits using cooperative regulatory assemblies. *Science* **364**, 5 (2019).

9 Kochanowski, K. *et al.* Few regulatory metabolites coordinate expression of central metabolic genes in Escherichia coli. *Mol Syst Biol* **13**, 903, doi:10.15252/msb.20167402 (2017).

10 Cuthbertson, L. & Nodwell, J. R. The TetR family of regulators. *Microbiol Mol Biol Rev* **77**, 440-475, doi:10.1128/MMBR.00018-13 (2013).

11 Orth, P., Schnappinger, D., Hillen, W., Saenger, W. & Hinrichs, W. Structural basis of gene regulation by the tetracycline inducible Tet repressor–operator system. *Nat Struct Biol* **7**, 5 (2000).

12 Sevvana, M. *et al.* An exclusive alpha/beta code directs allostery in TetR-peptide complexes. *J Mol Biol* **416**, 46-56, doi:10.1016/j.jmb.2011.12.008 (2012).

13 Hanko, E. K. R., Minton, N. P. & Malys, N. A Transcription Factor-Based Biosensor for Detection of Itaconic Acid. *ACS Synth Biol* **7**, 1436-1446, doi:10.1021/acssynbio.8b00057 (2018).

14 Peters, G. *et al.* Development of N-acetylneuraminic acid responsive biosensors based on the transcriptional regulator NanR. *Biotechnol Bioeng* **115**, 1855-1865, doi:10.1002/bit.26586 (2018).

15 Monteiro, F. *et al.* Measuring glycolytic flux in single yeast cells with an orthogonal synthetic biosensor. *Mol Syst Biol* **15**, e9071, doi:10.15252/msb.20199071 (2019).

16 Tsuruno, K., Honjo, H. & Hanai, T. Enhancement of 3-hydroxypropionic acid production from glycerol by using a metabolic toggle switch. *Microb Cell Fact* **14**, 155, doi:10.1186/s12934-015-0342-1 (2015).

17 Taylor, N. D. *et al.* Engineering an allosteric transcription factor to respond to new ligands. *Nat Methods* **13**, 177-183, doi:10.1038/nmeth.3696 (2016).

18 Wise, A. A. & Kuske, C. R. Generation of Novel Bacterial Regulatory Proteins That Detect Priority Pollutant Phenols. *Appl. Environ. Microbiol.* **66**, 7 (2000).

19 Galvao, T. C., Mencia, M. & de Lorenzo, V. Emergence of novel functions in transcriptional regulators by regression to stem protein types. *Mol Microbiol* **65**, 907-919, doi:10.1111/j.1365-2958.2007.05832.x (2007).

20 Mitchler, M. M., Garcia, J. M., Montero, N. E. & Williams, G. J. Transcription factor-based biosensors: a molecular-guided approach for natural product engineering. *Curr Opin Biotechnol* **69**, 172-181, doi:10.1016/j.copbio.2021.01.008 (2021).

21 Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-1391 (2008).

22 Jha, R. K., Chakraborti, S., Kern, T. L., Fox, D. T. & Strauss, C. E. Rosetta comparative modeling for library design: Engineering alternative inducer specificity in a transcription factor. *Proteins* **83**, 1327-1340, doi:10.1002/prot.24828 (2015).

23 Quijano-Rubio, A. *et al.* De novo design of modular and tunable protein biosensors. *Nature* **591**, 482-487, doi:10.1038/s41586-021-03258-z (2021).

24 Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, 545-574, doi:10.1016/B978-0-12-381270-4.00019-6 (2011).

25 Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **13**, 3031-3048, doi:10.1021/acs.jctc.7b00125 (2017).

26 Peterman, N. & Levine, E. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**, 206, doi:10.1186/s12864-016-2533-5 (2016).

27 Ding, N. *et al.* Programmable cross-ribosome-binding sites to fine-tune the dynamic range of transcription factor-based biosensor. *Nucleic Acids Res* **48**, 10602-10613, doi:10.1093/nar/gkaa786 (2020).

28 Jones, E. M. *et al.* Structural and functional characterization of G protein-coupled receptors with deep mutational scanning. *Elife* **9**, doi:10.7554/eLife.54895 (2020).

29 Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci U S A* **110**, 14024-14029, doi:10.1073/pnas.1301301110 (2013).

30 Khersonsky, O. *et al.* Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol Cell* **72**, 178-186 e175, doi:10.1016/j.molcel.2018.08.033 (2018).

31 Nishikawa, K. K., Hoppe, N., Smith, R., Bingman, C. & Raman, S. Epistasis shapes the fitness landscape of an allosteric specificity switch. *Nat Commun* **12**, 5562, doi:10.1038/s41467-021-25826-7 (2021).

32 Johnson, M. D. *et al.* Pharmacological characterization of 4-hydroxy-N-desmethyl tamoxifen, a novel active metabolite of tamoxifen. *Breast Cancer Res. Treat.* **85**, 9 (2004).

33 Lien, E. A., Solheim, E. & Ueland, P. M. Distribution of Tamoxifen and Its Metabolites in Rat and Human Tissues during Steady-State Treatment. *Cancer Res.* **51**, 8 (1991).

34 Achan, J. *et al.* Quinine, an old-antimalarial drug in a modern world: role in the treatment of malaria. *Malaria Journal* **10**, 12 (2011).

35 Niciu, M. J. & Arias, A. J. Targeted opioid receptor antagonists in the treatment of alcohol use disorders. *CNS Drugs* **27**, 777-787, doi:10.1007/s40263-013-0096-4 (2013).

36 Sokal, R. R., Michener, C. D. & Kansas, U. o. *A Statistical Method for Evaluating Systematic Relationships*. (University of Kansas, 1958).

37 Leander, M., Yuan, Y., Meger, A., Cui, Q. & Raman, S. Functional plasticity and evolutionary adaptation of allosteric regulation. *Proc Natl Acad Sci U S A* **117**, 25445-25454, doi:10.1073/pnas.2002613117 (2020).

38 Teran, W. *et al.* Antibiotic-Dependent Induction of Pseudomonas putida DOT-T1E TtgABC Efflux Pump Is Mediated by the Drug Binding Repressor TtgR. *Antimicrobial Agents and Chemotherapy* **47**, 3067-3072, doi:10.1128/aac.47.10.3067-3072.2003 (2003).

39 Alguel, Y. *et al.* Crystal structures of multidrug binding protein TtgR in complex with antibiotics and plant antimicrobials. *J Mol Biol* **369**, 829-840, doi:10.1016/j.jmb.2007.03.062 (2007).

40 Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nat Genet* **37**, 73-76, doi:10.1038/ng1482 (2005).

41 Schumacher, M. A. *et al.* Structural Mechanisms of QacR Induction and Multidrug Recognition. *Science* **294**, 6 (2001).

42 Gaspar, J. M. NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics* **19**, 536, doi:10.1186/s12859-018-2579-2 (2018).

43 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890, doi:10.1093/bioinformatics/bty560 (2018).

44 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499, doi:10.1101/gr.209601.116 (2017).

45 Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. & Enright, A. J. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* **63**, 41-49, doi:10.1016/j.ymeth.2013.06.027 (2013).

46 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *JMLR* **12**, 6 (2011).

47 Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261-272, doi:10.1038/s41592-019-0686-2 (2020).

48 Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc Natl Acad Sci U S A* **112**, 7159-7164, doi:10.1073/pnas.1422285112 (2015).

49 Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90-95, doi:10.1109/MCSE.2007.55 (2007).

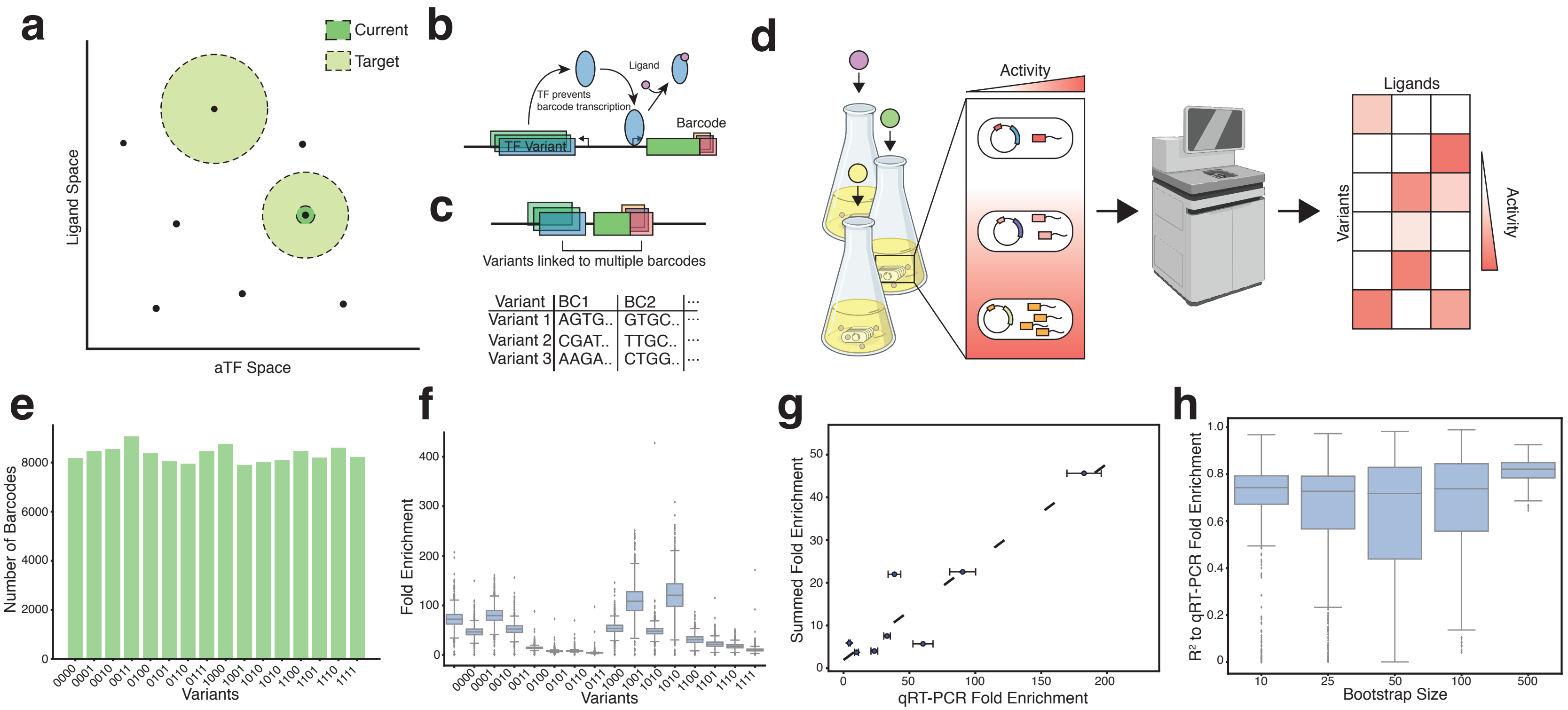

a
Current
Target
Ligand Space
aTF Space
b
Ligand
TF prevents
barcode transcription
TF Variant
Barcode
c
Variants linked to multiple barcodes
Variant
BC1
BC2
...
Variant 1
AGTG..
GTGC..
Variant 2
CGAT..
TTGC..
Variant 3
AAGA..
CTGG..
d
Activity
Ligands
Variants
Activity
e
Number of Barcodes
0
2000
4000
6000
8000
0000
0001
0010
0011
0100
0101
0110
0111
1000
1001
1010
1100
1101
1110
1111
Variants
f
Fold Enrichment
0
100
200
300
400
0000
0001
0010
0011
0100
0101
0110
0111
1000
1001
1010
1100
1101
1110
1111
Variants
g
Summed Fold Enrichment
0
10
20
30
40
50
0
50
100
150
200
qRT-PCR Fold Enrichment
h
R² to qRT-PCR Fold Enrichment
0
0.2
0.4
0.6
0.8
1.0
10
25
50
100
500
Bootstrap Size

Figure 1: Validating RNA-Seq on a 16-member library. ***(a)*** The space of transcription factors and small molecules that can be sensed with different approaches. Black points represent specific ligand:aTF pairs. The dark green circle represents the extent to which methods can currently expand aTF:ligand affinity. The light green circle represents the extent to which new methodologies must increase aTF:ligand pairs. ***(b)*** Construct design pairs aTF variants to randomized barcodes. ***(c)*** A separate construct is used to pair barcodes to aTF variants such that RNA-Seq of barcodes can be translated to aTF function. ***(d)*** Methodology to harvest plasmids and RNA from E. coli transformed with aTF libraries. This approach can easily be scaled across multiple ligands and multiple libraries. ***(e)*** The number of barcodes per variant identified using next-generation sequencing of the construct in (C). Each variant is identified by a separate binary string. ***(f)*** Box plots of fold enrichment for each variant via RNA-Seq. The box represents the interquartile range. Whiskers extend to 1.5 times the IQR. Fliers denote points that lie outside the whiskers. Variants are represented by binary strings. ***(g)*** Correlation of qRT-PCR data and fold enrichment from RNA-Seq. qRT-PCR fold enrichment was measured via biological replicates of clonal strains of 8 of the 16 variants (see methods). The RNA-Seq fold enrichment value was calculated by summing the counts of all barcodes associated with a particular variant (see methods). The R2 for this dataset is 0.83. ***(h)*** Bootstrap correlation of qRT-PCR fold enrichment to RNA-Seq data for 8 of the 16 variants. Groups of 10, 25, 50, 100, or 500 barcodes were sampled for each variant across 500 cycles. The resulting correlation for each cycle is plotted.

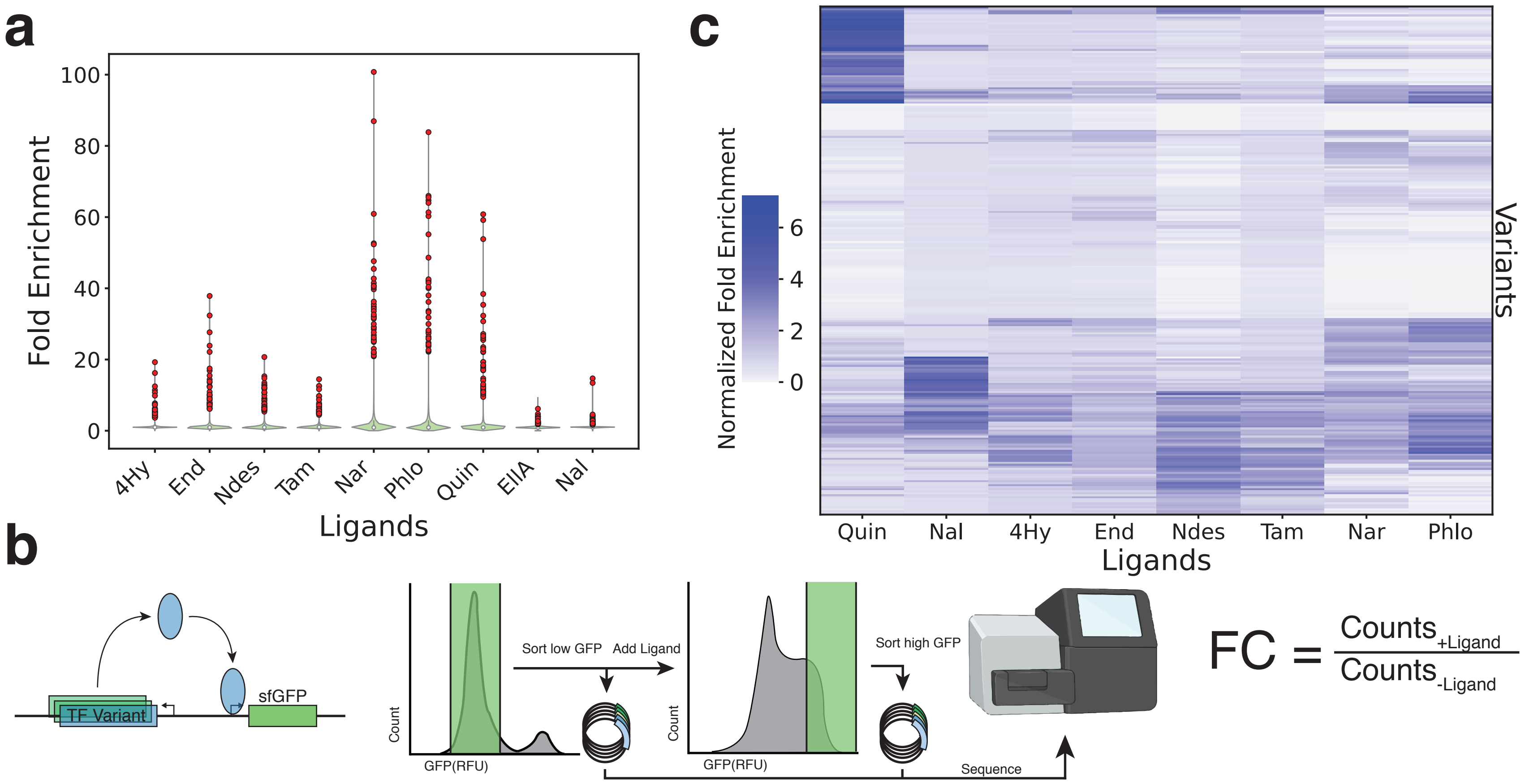


Figure 2: Identifying novel sensors in the agnostic library. ***(a)*** The top 40 best variants from each ligand were selected for a fluorescence-based screen (red points). The violin plot shows the fold enrichment calculated by RNA-Seq for all 17,365 variants for each ligand. The circle inside the violin plot denotes the median of fold enrichment for each ligand. The thick grey line inside the violin plot represents the IQR and the thin grey line extends to 1.5 times the IQR. ***(b)*** Fluorescence screening workflow that incorporates a construct with *sfGFP*. A single repressed sort and an induced sort were sequenced (see methods). Fold change (FC) was calculated as the ratio of percent change in the population with and without ligand. ***(c)*** Fold change for each ligand across the 251 best performing variants.

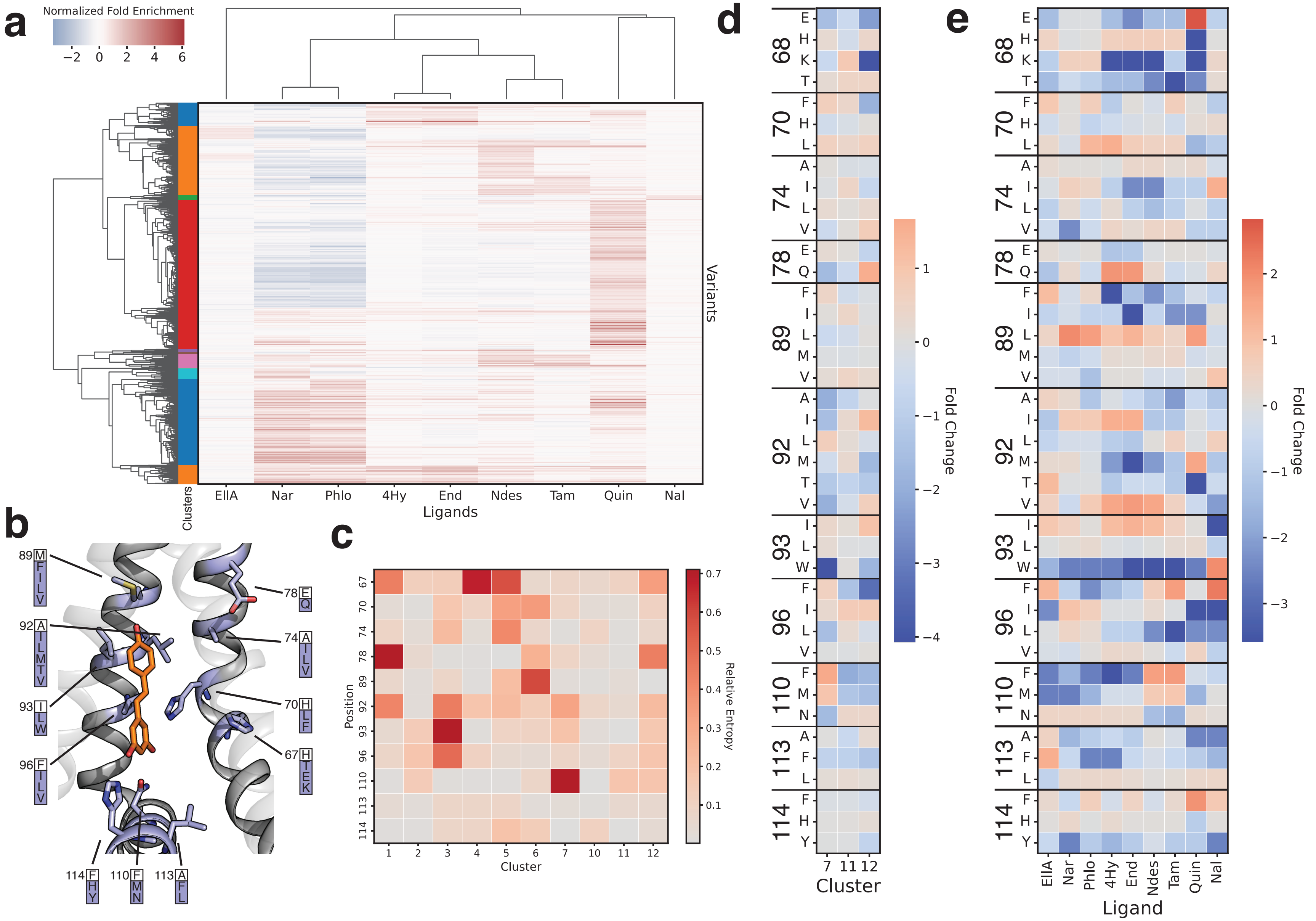

a
Normalized Fold Enrichment
−2 0 2 4 6
Variants
Clusters
EIIA Nar Phlo 4Hy End Ndes Tam Quin Nal
Ligands
b
89 M F I L V
92 A I L M T V
93 I L W
96 F I L V
78 E Q
74 A I L V
70 H L F
67 H T E K
114 F H Y
110 F M N
113 A F L
c
Position
67 70 74 78 89 92 93 96 110 113 114
1 2 3 4 5 6 7 10 11 12
Cluster
Relative Entropy
0.1 0.2 0.3 0.4 0.5 0.6 0.7
d
68 E H K T
70 F H L
74 A I L V
78 E Q
89 F I L M V
92 A I L M T V
93 I L W
96 F I L V
110 F M N
113 A F L
114 F H Y
7 11 12
Cluster
Fold Change
1 0 −1 −2 −3 −4
e
68 E H K T
70 F H L
74 A I L V
78 E Q
89 F I L M V
92 A I L M T V
93 I L W
96 F I L V
110 F M N
113 A F L
114 F H Y
EIIA Nar Phlo 4Hy End Ndes Tam Quin Nal
Ligand
Fold Change
2 1 0 −1 −2 −3

Figure 3: Elucidating ligand-specific sequence preferences from RNA-Seq. ***(a)*** RNA-Seq fold enrichment data for 3,3135 variants across nine ligands. Ligands and variants have been clustered via the UPGMA algorithm with a correlation distance metric and a target of 12 clusters (see methods). The different clusters are denoted by the colored bars on the right of the heatmap. aTF function is shown as the log2(fold enrichment) normalized to wildtype. ***(b)*** Structure of TtgR with tolerated mutations at each position (PDB ID: 7K1C). The wildtype residue is highlighted at each position as purple sticks. Resveratrol, a natural ligand of TtgR, is shown as orange sticks. The tolerated mutations at each position are shown with the violet background while the wildtype identity is shown in white. ***(c)*** Sequence relative positional entropy of the clustered data for all tolerated positions. In this plot, a higher relative entropy indicates a changed amino acid distribution after clustering. ***(d)*** Heatmap of the fold change in amino acid abundance across allowed positions for clusters 7, 11, and 12. The fold change of frequency is the log2 of the ratio of amino acid frequency after clustering to the frequency before clustering. ***(e)*** Heatmap of the fold change in amino acid abundance across allowed positions for the top 40 variants for each ligand. Fold change is calculated identical to (D).

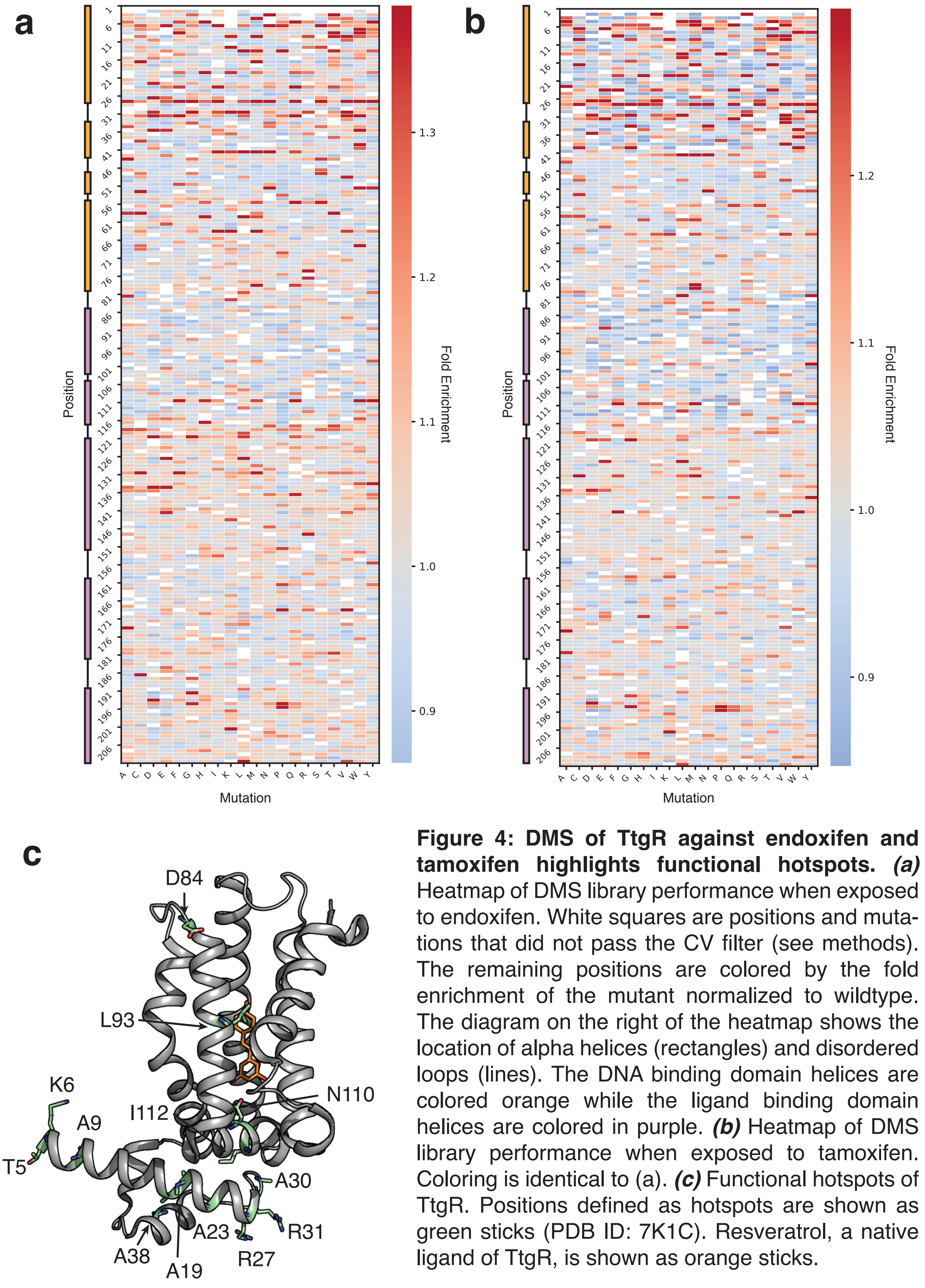
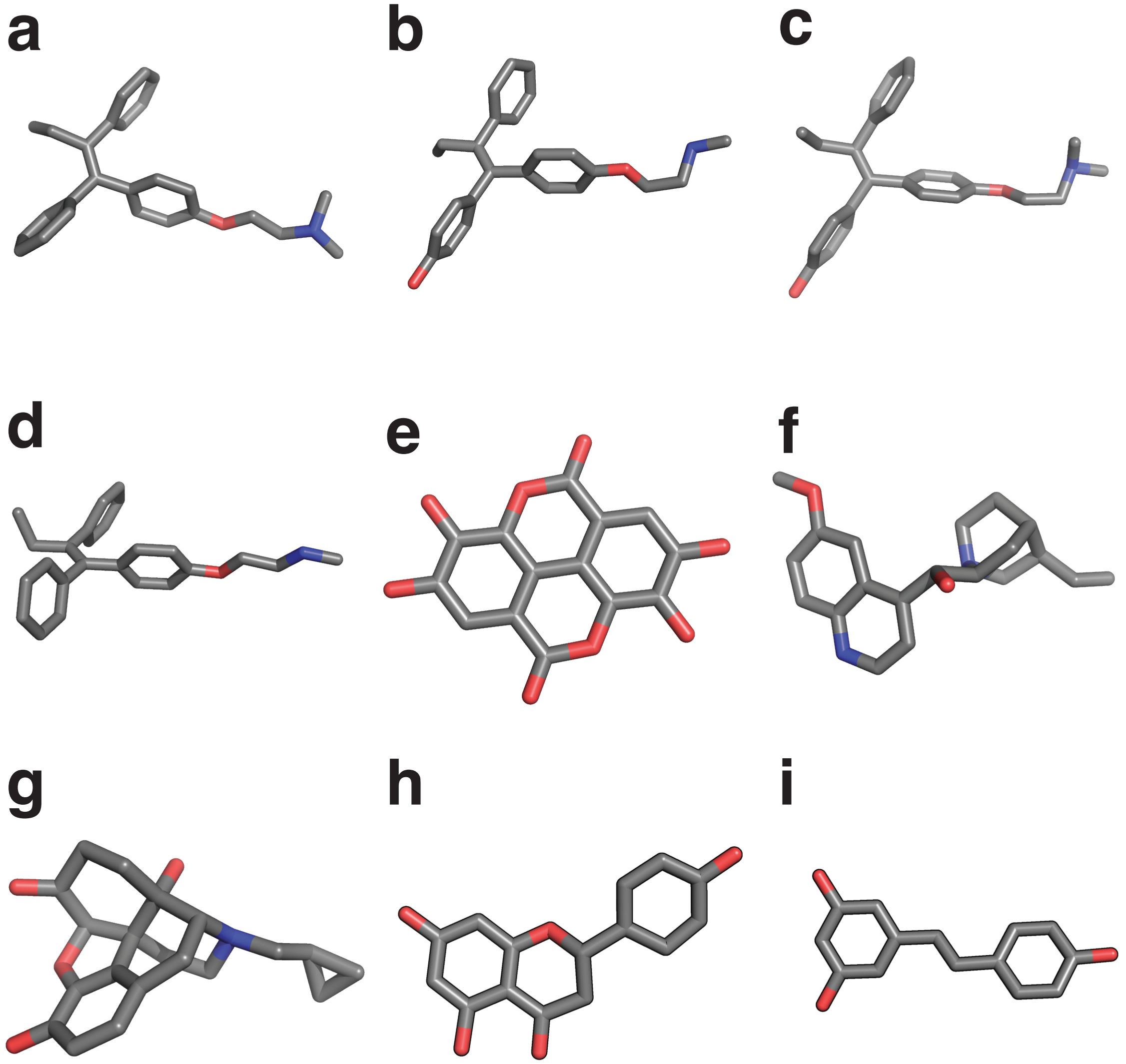
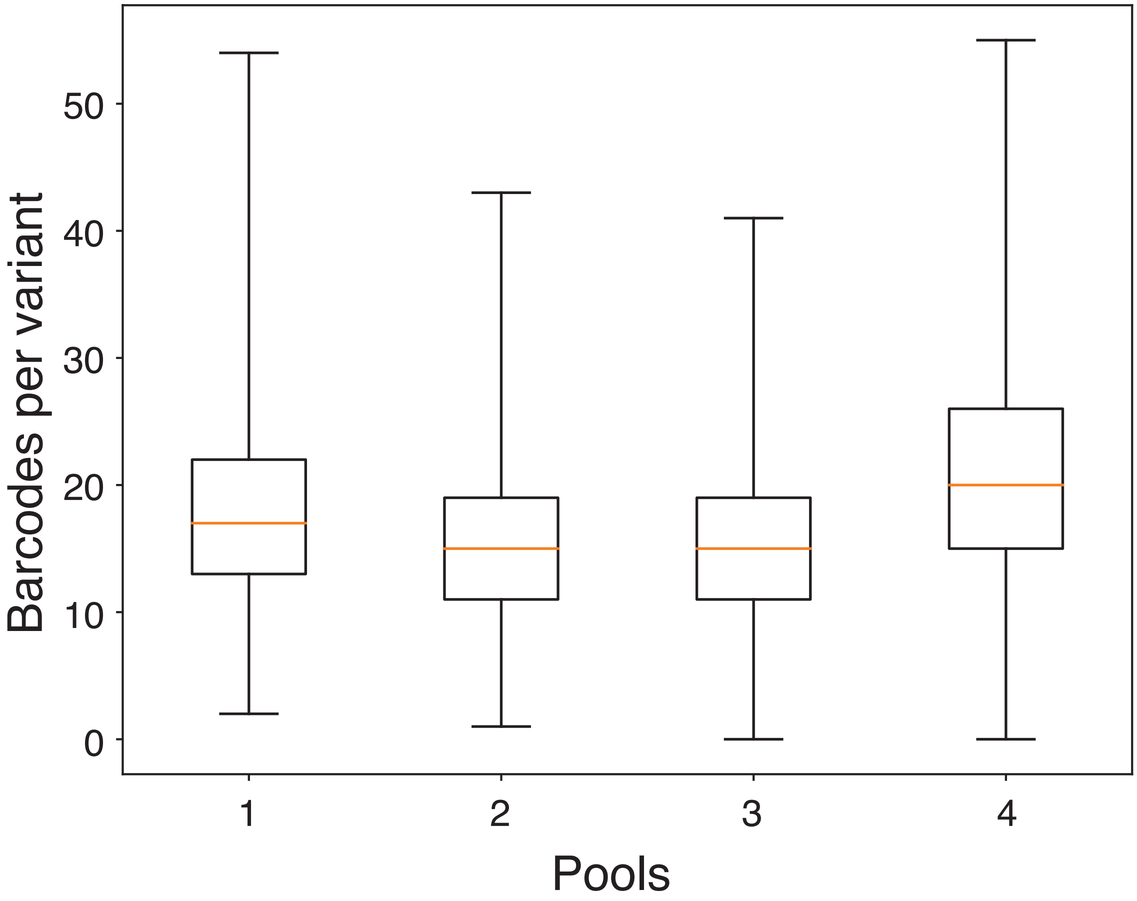


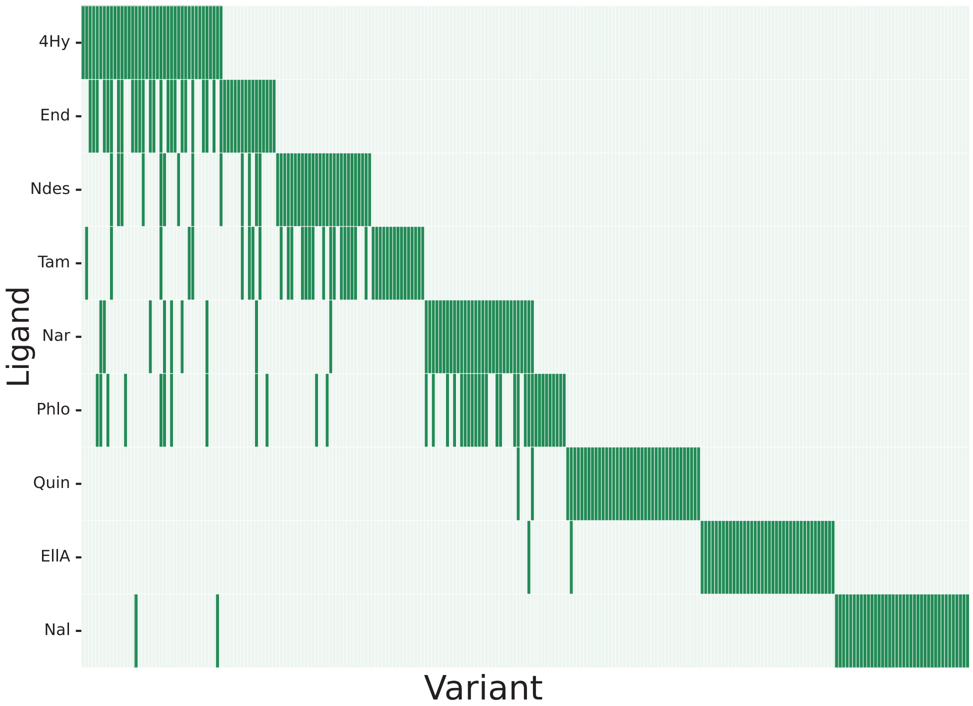
Figure 4: DMS of TtgR against endoxifen and tamoxifen highlights functional hotspots. ***(a)*** Heatmap of DMS library performance when exposed to endoxifen. White squares are positions and mutations that did not pass the CV filter (see methods). The remaining positions are colored by the fold enrichment of the mutant normalized to wildtype. The diagram on the right of the heatmap shows the location of alpha helices (rectangles) and disordered loops (lines). The DNA binding domain helices are colored orange while the ligand binding domain helices are colored in purple. ***(b)*** Heatmap of DMS library performance when exposed to tamoxifen. Coloring is identical to (a). ***(c)*** Functional hotspots of TtgR. Positions defined as hotspots are shown as green sticks (PDB ID: 7K1C). Resveratrol, a native ligand of TtgR, is shown as orange sticks.



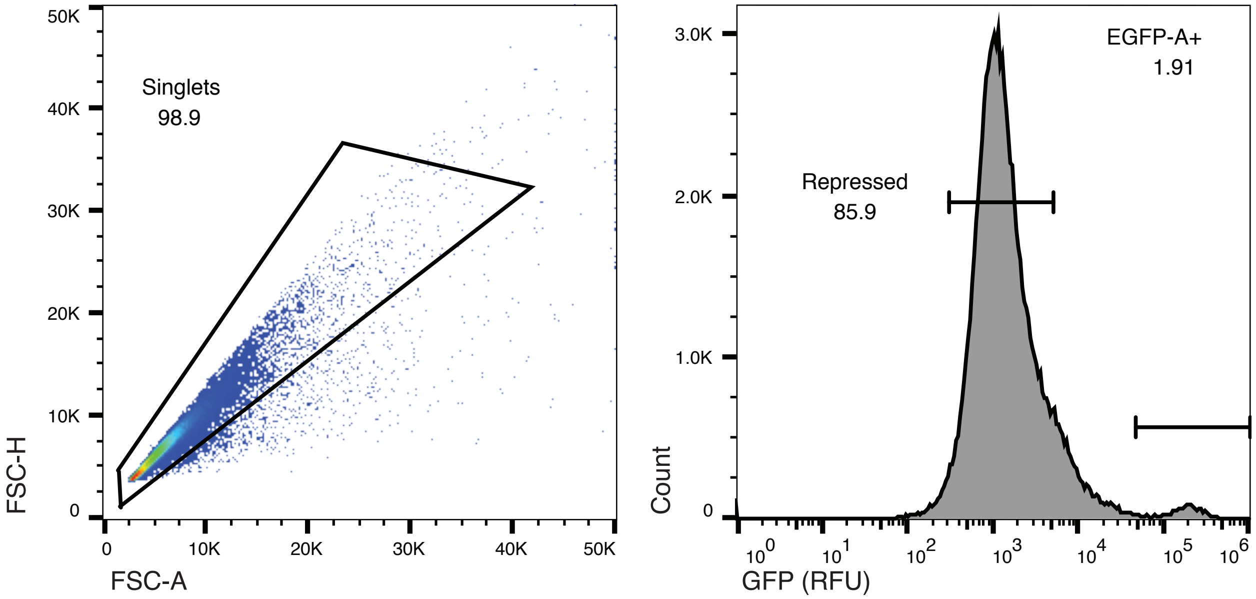
Supplementary Figure 1: Ligands used in this study
Structures of ***(a)*** tamoxifen, ***(b)*** endoxifen, ***(c)*** 4-hydroxytamoxifen, ***(d)*** N-desmethyltamoxifen, ***(e)*** ellagic acid, ***(f)*** quinine, ***(g)*** naltrexone, ***(h)*** naringenin, and ***(i)*** resveratrol. Carbon atoms are shown in grey, oxygen atoms in red, and nitrogen atoms in blue. Hydrogens are not shown.



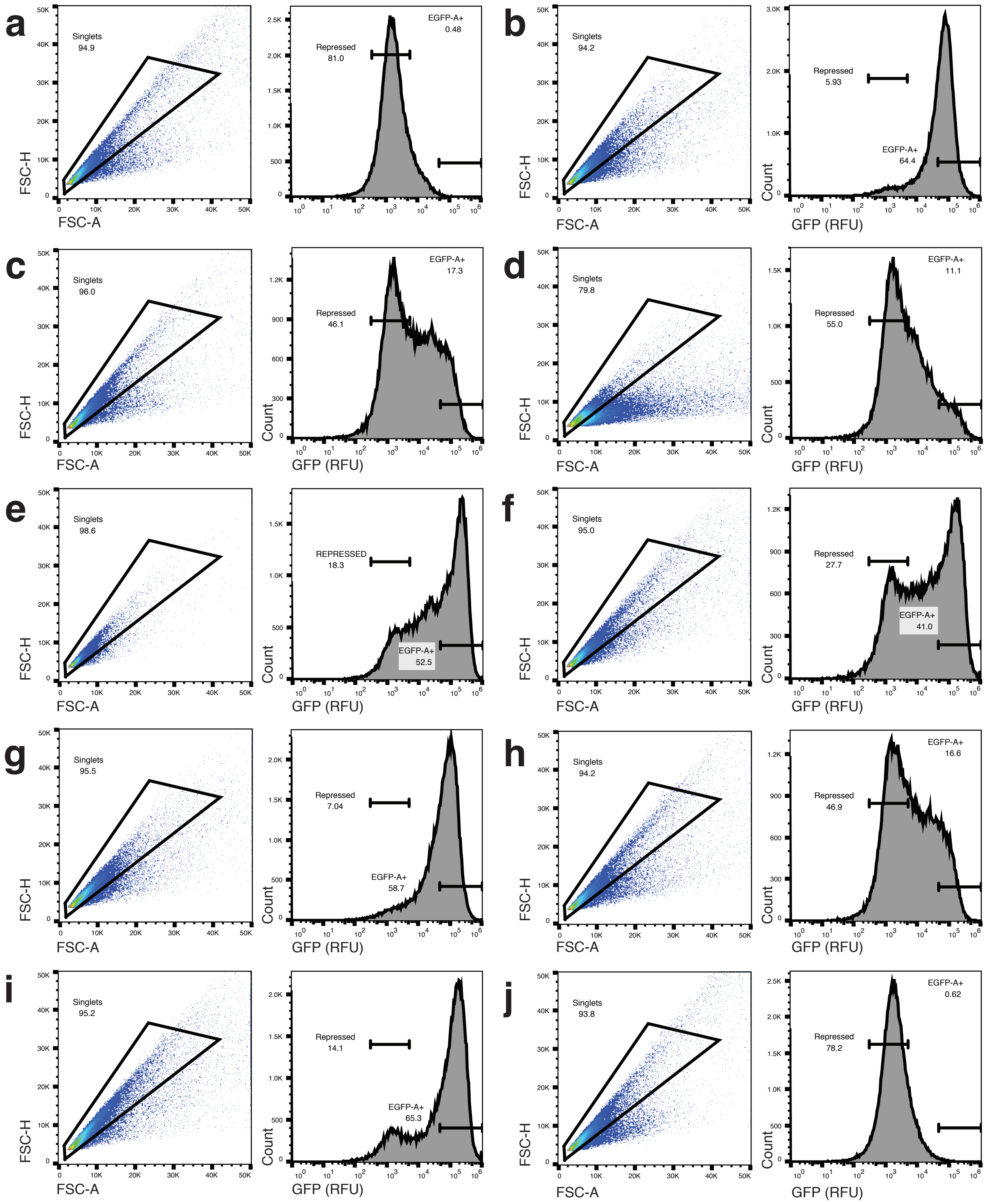
Supplementary Figure 2: Barcodes per variant of agnostic library from mapping data
The agnostic library was split into 4 pools containing approximately 4,400 variants (see methods). Barcodes and variants were mapped as separate pools. The box plot represents the number of barcodes mapped to each variant in the pools.



Supplementary Figure 3: Shared sequences in top performing variants
The top 40 variants for each ligand were selected and listed (x-axis). Variants are marked in green if they are within the top 40 for a particular ligand. There are 251 unique variants in the set of top performers across all ligands.

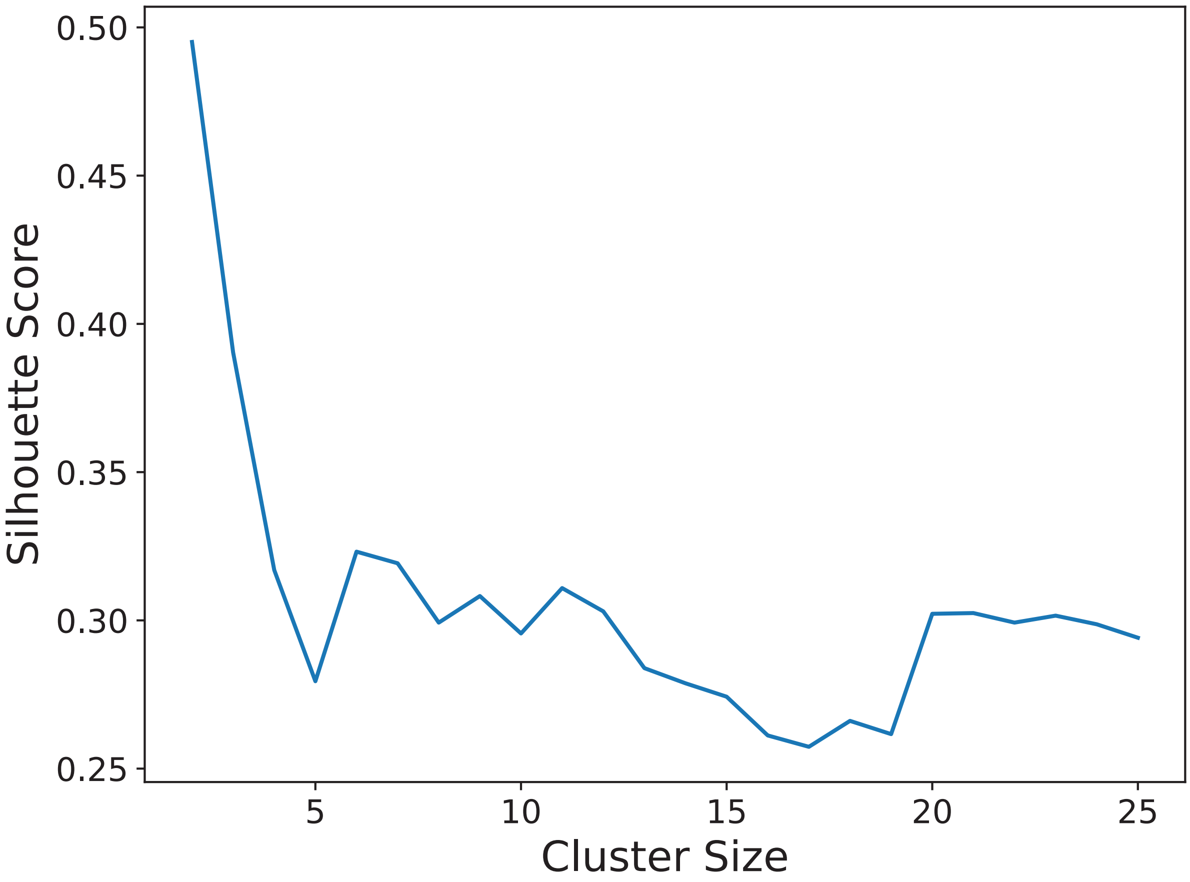


Supplementary Figure 4: Cell sorting gates for repressed sort
Flow cytometry scatterplot (left) and fluorescence histogram (right) for the library of top variants with no ligand. The scatterplot shows forward scatter area (FSC-A) and forward scatter height (FSC-H). The singlet gate was used to subset the population. Cells falling into the repressed gate in the fluorescence histogram were sorted.

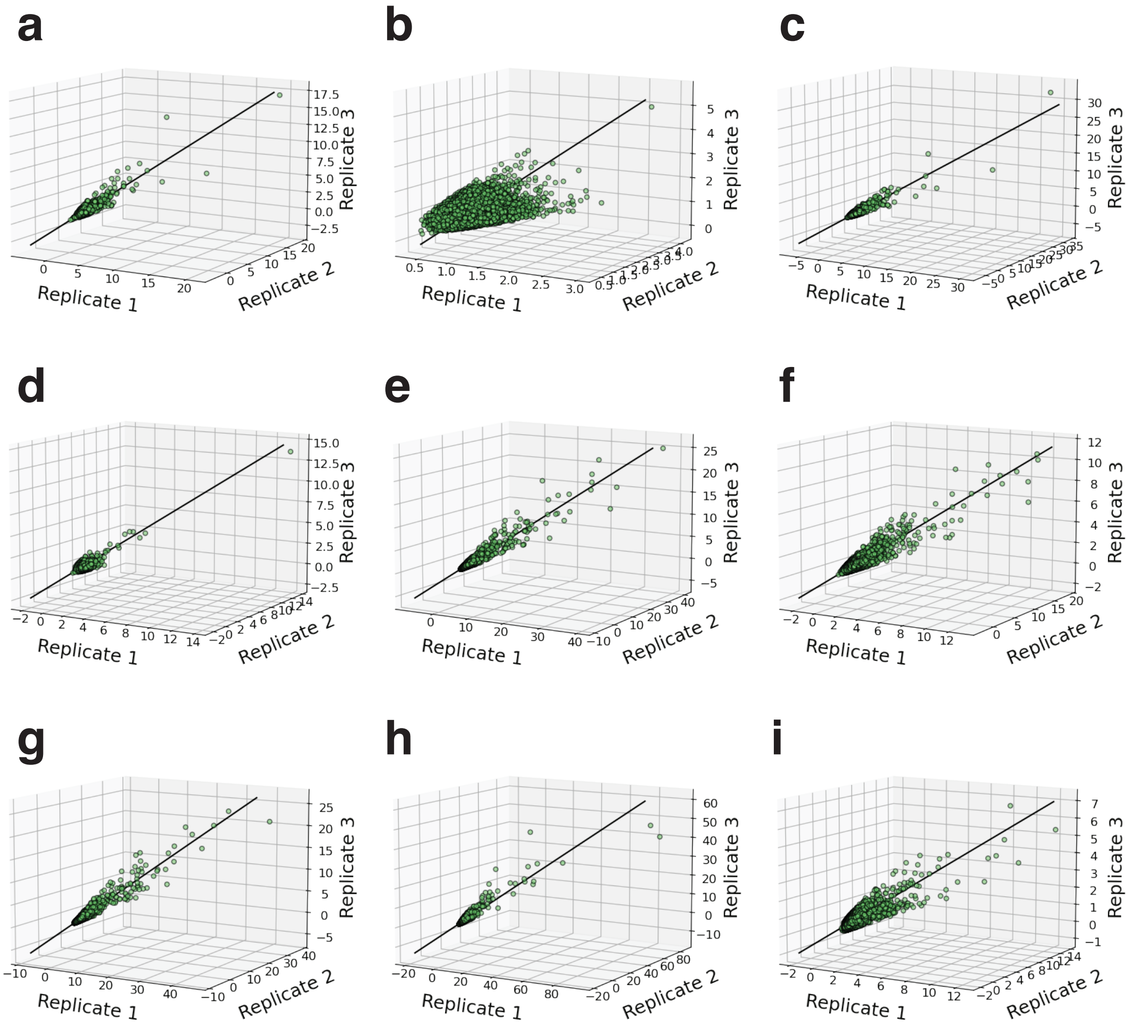


Supplementary Figure 5: Cell sorting gates for ligand-induced cultures

Flow cytometry scatterplots and histograms for libraries induced with ***(a)*** no ligand, ***(b)*** naringenin, ***(c)*** tamoxifen, ***(d)*** naltrexone, ***(e)*** quinine, ***(f)*** endoxifen, ***(g)*** phloretin, ***(h)*** 4-hydroxytamoxifen, ***(i)*** N-desmethyltamoxifen, ***(j)*** ellagic acid. The scatterplot shows forward scatter height versus area; cells falling into the singlet gates were sorted based on the fluorescence histogram (EGFP-A+ gate). The repressed gate indicates the major peak of the no ligand population.

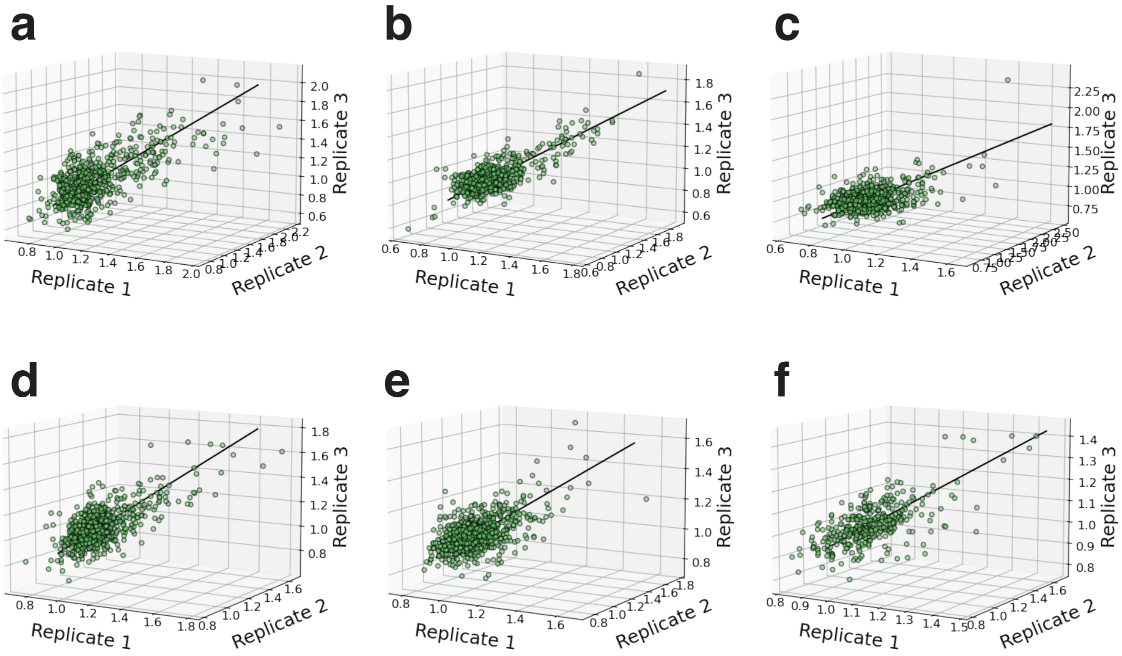


Supplementary Figure 6: Silhouette score of cluster sizes
The silhouette score is plotted against cluster size for hierarchical clustering with the UPGMA algorithm with a correlation distance metric (see methods).



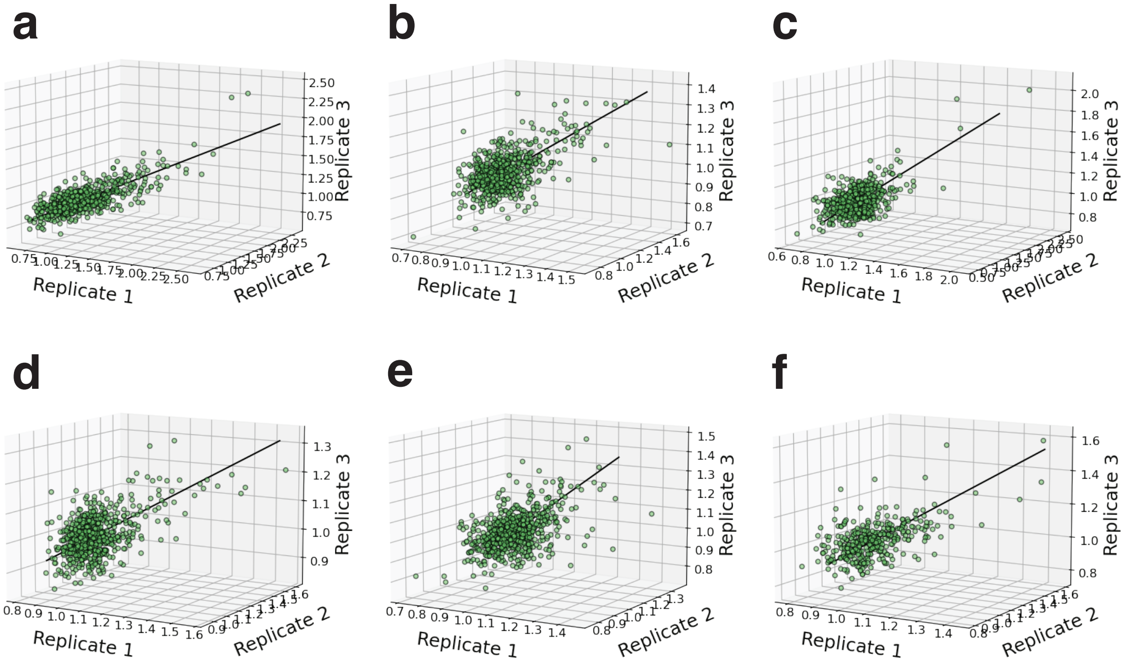
Supplementary Figure 7: CV filter of RNA-Seq fold enrichment for the agnostic library

Scatter plots of agnostic variants after applying a 30% CV filter for ***(a)*** 4-hydroxytamoxifen, ***(b)*** ellagic acid, ***(c)*** endoxifen, ***(d)*** naltrexone, ***(e)*** naringenin, ***(f)*** N-desmethyltamoxifen, ***(g)*** phloretin, ***(h)*** quinine, and ***(i)*** tamoxifen. The fold enrichment of each variant (green circle) is plotted across the three biological replicates on the X, Y, and Z axes. The best fit line is shown in black and is calculated using a least squares approach.



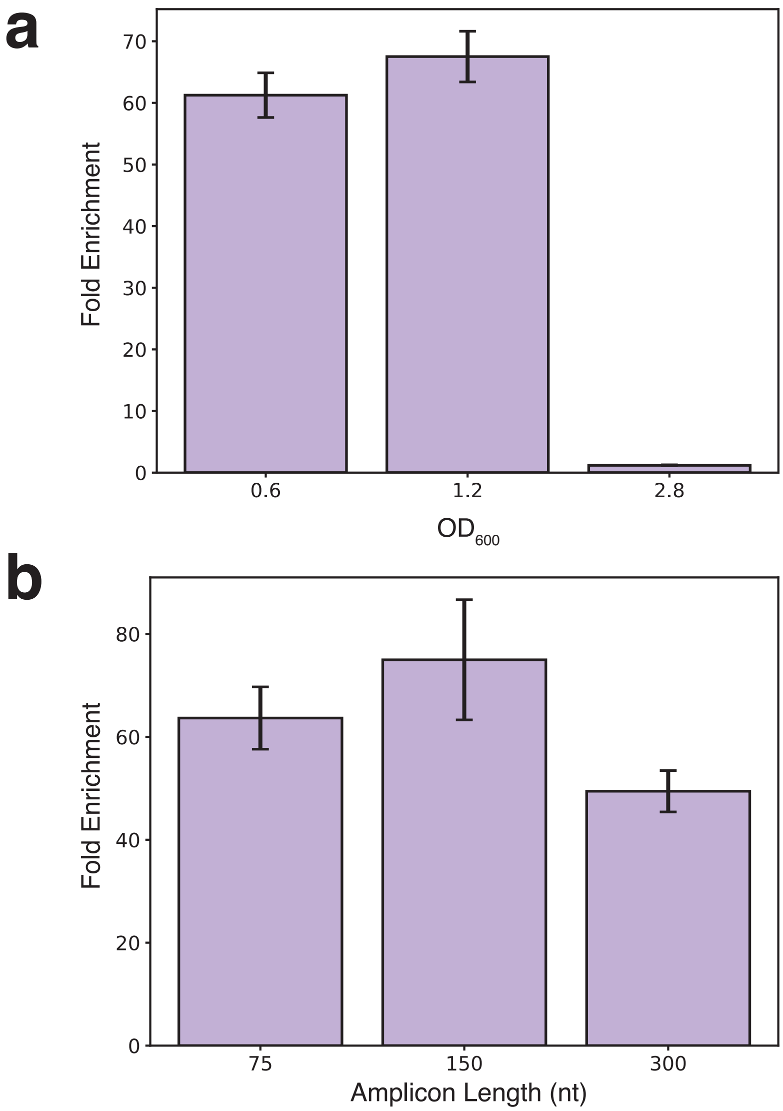
Supplementary Figure 8: CV filter of RNA-Seq fold enrichment for the DMS library with endoxifen

Scatter plots of DMS variants after applying a 30% CV filter for DMS segments ***(a)*** 1, ***(b)*** 2, ***(c)*** 3, ***(d)*** 4, ***(e)*** 5, and ***(f)*** 6 after dosing with endoxifen. Plot layout is identical to Supplementary Fig. 7.

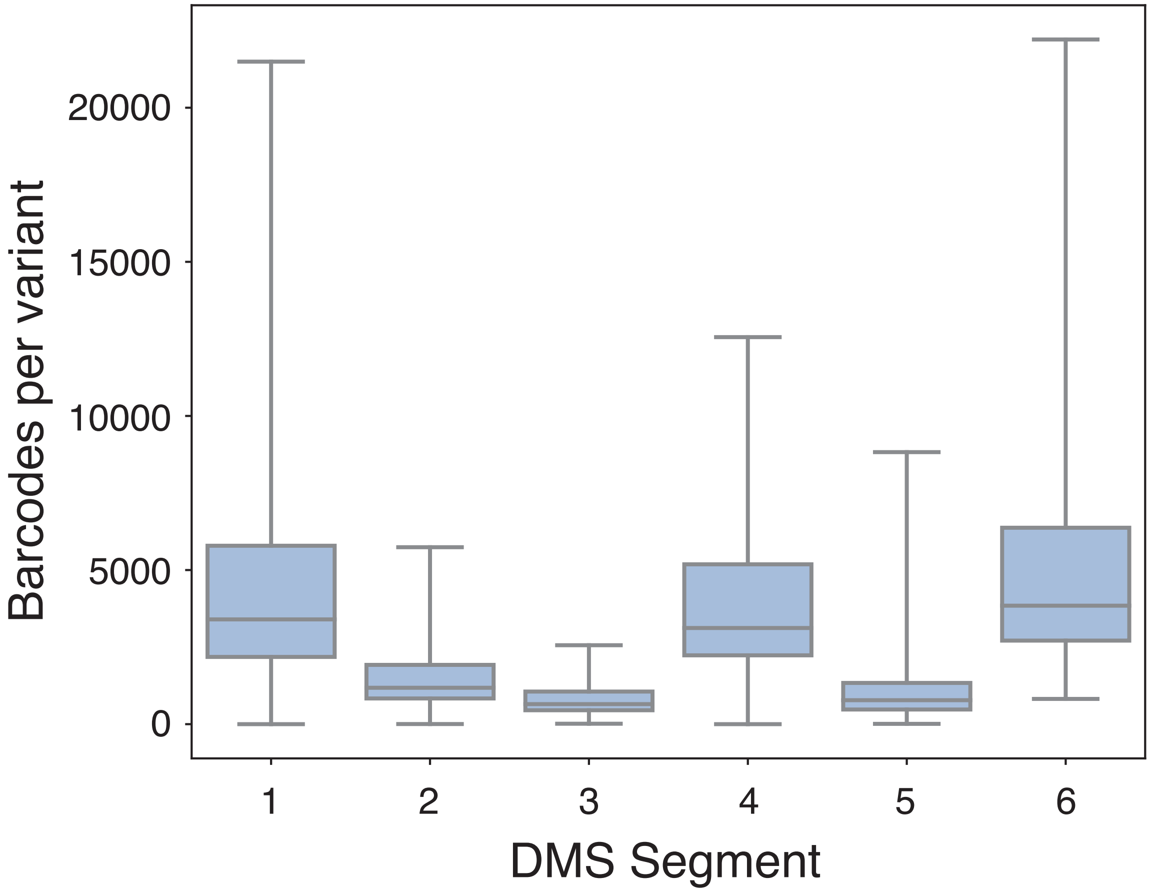


Supplementary Figure 9: CV filter of RNA-Seq fold enrichment for the DMS library with tamoxifen

Scatter plots of DMS variants after applying a 30% CV filter for DMS segments ***(a)*** 1, ***(b)*** 2, ***(c)*** 3, ***(d)*** 4, ***(e)*** 5, and ***(f)*** 6 after dosing with tamoxifen. Plot layout is identical to Supplementary Fig 7.



Supplementary Figure 10: Optimizing RNA expression and amplification
(a) qRT-PCR fold enrichment of a 150nt amplicon in the sfGFP gene at different OD$_{600}$ values at the time of RNA harvesting. RNA was harvested at OD$_{600}$ 0.6, 1.2, and 2.8 (x-axis). Each OD$_{600}$ value was measured in biological triplicate at 1mM naringenin. Fold enrichment of the amplicon was calculated as the $2^{-\Delta\Delta Ct}$ value of the *sfGFP* gene in comparison to a constitutively expressed control (see methods). Error bars are the standard error propagated from the technical replicates. ***(b)*** qRT-PCR fold enrichment of different amplicon sizes. Each amplicon used the same 5' primer, but differing 3' primers to yield a 75nt, 150nt, and 300nt amplicon. The same samples in biological triplicate were used as template for the qRT-PCR experiment. Fold enrichment was calculated using the same methods as (a).



Supplementary Figure 11: Barcodes per variant of DMS library from mapping data
The DMS library was split into 6 pools containing approximately 700 variants (see methods). Barcodes and variants were mapped as separate segments. The box plot represents the number of barcodes mapped to each variant in the segment.

Name	Sequence
KN_124	gtcagtgtcgtgccatagatccacgaggcccttttttcgtc
KN_125	gggatctcgacgctctcccttatgactgattaccgcctttgagtgag
KN_126	tcataagggagagcgtcgagatccc
KN_127	ctcactcaaaggcggtaatcagggccgccaccgc
KN_128	gcggtggcggccctgattaccgcctttgagtgag
KN_129	gtcgagatcccgggcgcgccAAAAAATTTATTTGCTTTCAGGAAAA
KN_130	CTTCTTCTTTGGTGCGACGCACCATAAAGGTTCCACTGCTAGATT
KN_131	ATGGTGCGTCGCACCAAAGAAGAAG
KN_132	ggcgcgcccgggatctcgac
KN_133	gtcgagatcccgggcgcgccgcttgatatcgaattcctgcagcccg
KN_134	cgggctgcaggaattcgatatcaagcggcgcgcccgggatctcgac
KN_135	tcataagggagagcgtcgagatcccGGCGCGCCTTGACAATTAATCATC
KN_136	CTTCTTCTTTGGTGCGACGCACCATCATATGAAAAGATCCCGGGCTAGATTAAG
KN_137	gggggatcccatggtacgc
KN_138	aagacgaaaaaagggcctcgtg
KN_139	cacgaggcccttttttcgtcttttatttgtacagttcatccatacc
KN_140	gcgtaccatgggatccccccacctcgagatgctagc
KN_141	ccgacgtctaagaaaccattattatcacgaggcccttttttcgtctt
KN_142	gtcagatagcaccacatagcaggatctatggcacgacactgac
KN_143	TCGCCAGCAGGCCTTTTTATTTG
KN_144	gtcagatagcaccacatagcagTAATAATCATCGCGAAGACTTGATCG
KN_145	gggagagcgtcgagatcccTTGACAATTAATCATCCGGCTCGTATAATAG
KN_146	CAAATAAAAAGGCCTGCTGGCGATTATTTGCGCAGCGCCGG
KN_147	cacctcgagatgctagcaaaaaaagagtaCACCCAGCAGTATTTACAAACAACC
KN_148	CAAATAAAAAGGCCTGCTGGCGAgggatctcgacgctctcc
KN_149	CCCAGATACGCTGTTTCAATTCCTTTATTATTATTTGCGCAGCGCCG
KN_150	ACCGCACAGGTTGCCCACTTGACAATTAATCATCCGGCTCG
KN_156	gtcggccaaggtaccgg
KN_157	tggtTtcgtcActattctggtgg
KN_160	tgaagagtttgatcatggctcag
KN_161	tttcccagacattactcacccg
KN_162	tcaccctcgccacgca
KN_163	cgcgttttgtacgtgccg
KN_164	ggtctcCACTGCTGGATTCTCTGCACG
KN_165	ggtctcCAACGACGAATCAGGCCATCC
KN_166	agtgagttgattgctacgtaaggcttcggactgGGTCTCcaattCATCgACgtctGctgc

KN_167	acGGTCTCtcgtgtataatggNNNNNNNNNNNNNNNNgcagCagacGTcGATGaattg
KN_168	ccccgaaaagtgccacctggcggcgttgtgacaatttaagtgagttgattgctacgtaag
KN_169	gcgaCTGaaTATtgcggcacgAtacCATTTgCTCataagacGGTCTCtcgtgtataatgg
KN_170	ctGGTCTCGcacgaggcccttttttcg
KN_171	ctGGTCTCGaattgttacgtagcaatcaactc
KN_172	AGGCGTCTTTCTTAGCCGGCGGTCTCcaattgttacgtagcaatcaactca
KN_173	GCCGGCTAAGAAAGACGCCTGGTCTCgcacgaggcccttttttcgt
KN_174	gagttgattgctacgtaacaatt
KN_175	ctgcAAACCCTGTGCCAGAGGGTGccattatacacgaggcccttttttcg
KN_176	GGCACAGGGTTTgcagCagacGTcGATGaattgttacgtagcaatcaactc
KN_177	tGctgcGAGTGACCTTAAGTCAGGGAccattatacacgaggcccttttttcg
KN_178	AAGGTCACTCgcagCagacGTcGATGaattgttacgtagcaatcaactc
KN_179	tgcGCTTCTGTCCAAGCAGGTTAccattatacacgaggcccttttttcg
KN_180	TTGGACAGAAGCgcagCagacGTcGATGaattgttacgtagcaatcaactc
KN_193	CGTCTGGTCTCAGCGTGCCGAACTGGTTCAGGC
KN_194	GAGTCGGTCTCCTCTCggatgaactgtacaaataaaagacg
KN_195	CGTCTGGTCTCAGCGTTGCGAATTTACGGATGATATGTGTG
KN_196	ggcccttttttcgtcttTTAagtcggccaaggtaccggca
KN_197	GCCGGCTAAGAAAGACGCCTGGTCTCgcacgTTATTAcagttccaccagaatagTgacga
KN_198	CAGGCGTCTTTCTTAGCCGGCGGTCTCcGaACgatggtgatgtcaacggtcat
KN_199	GAGTCGGTCTCCTCTCcgtcActattctggtggaactgTAA
KN_200	TCCACGCGATGGCCCNNNNNNNNNNNNNNNNNNNNNttgacatcaccatcGTtCCATC
KN_201	tctggtggaactgTAATAAcgtgta
KN_202	AGCCTCCTGGGCGGGTCATGNNNNNNNNNNNNNNNNtcaccctcgccacgca
KN_204	GAGTCGGTCTCgatcccTTGACAATTAATCATCCGGCTCGTAT
KN_205	ACCGTGGTCTCCGCGAgaggcttttgactttctgctaatttat
KN_206	CCGACGGTCTCGggatctcgacgctctcccttatgac
KN_207	CGAAGGGTCTCGTCGCCAGCAGGCCTTTTTATTTGGGGG
KN_208	CGTCTGGTCTCAGCGTTCATTAGAGTCTAGAGAAAGACAGGATT
KN_209	CGTCTGGTCTCAGCGTCGTACCACGCTGGCAG
KN_210	CGTCTGGTCTCAGCGTACGAAACGCATGATCACCTGG
KN_211	CGTCTGGTCTCAGCGTACGTACCCGTCGCATTAATGAAATC
KN_212	CGTCTGGTCTCAGCGTGCACTGGCAAACGCAGTTC
KN_213	CGTCTGGTCTCAGCGTGCCGGATAGTGTTGATCTGCTG

SPS_7b_MS_SPR_R1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCAGCAGCCAACGACGA
SPS_7b_MS_SPR_R2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNGCAGCAGCCAACGACGA
SPS_7b_MS_SPR_R3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNGCAGCAGCCAACGACGA
SPS_NovaSeq_Barcode_F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcaattCATCgACgtctGctgc
SPS_NovaSeq_Barcode_F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNcaattCATCgACgtctGctgc
SPS_NovaSeq_Barcode_F3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNcaattCATCgACgtctGctgc
SPS_S1_spacer_miseq_R1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAGTGCCAGGGTGATACC
SPS_S1_spacer_miseq_R2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNCAGTGCCAGGGTGATACC
SPS_S1_spacer_miseq_R3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNCAGTGCCAGGGTGATACC
SPS_V5_UMI2_BC_F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCCTCCTGGGCGGGTCATG
SPS_V5_UMI2_BC_F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNAGCCTCCTGGGCGGGTCATG
SPS_V5_UMI2_BC_F3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNAGCCTCCTGGGCGGGTCATG
ML_Seg1_SPR1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGTAGATTGCACCGCGGG
ML_Seg1_SPR2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNAGTAGATTGCACCGCGGG
ML_Seg1_SPR3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNAGTAGATTGCACCGCGGG
ML_Seg2_SPR1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGCAGCCCAGCGGG
ML_Seg2_SPR2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNATGCAGCCCAGCGGG
ML_Seg2_SPR3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNATGCAGCCCAGCGGG
ML_Seg3_SPR1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGCTGCTGACGAATTTCAC
ML_Seg3_SPR2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNCGCTGCTGACGAATTTCAC
ML_Seg3_SPR3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNCGCTGCTGACGAATTTCAC
ML_Seg4_SPR1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACCGCTGCGCGTTC
ML_Seg4_SPR2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNACCGCTGCGCGTTC
ML_Seg4_SPR3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNACCGCTGCGCGTTC

ML_Seg5_SPR1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGCAGCATATCCAGACCG
ML_Seg5_SPR2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNCGCAGCATATCCAGACCG
ML_Seg5_SPR3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNCGCAGCATATCCAGACCG
ML_Seg6_SPR1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCCTGCTGGCGAgagg
ML_Seg6_SPR2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNGCCTGCTGGCGAgagg
ML_Seg6_SPR3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNGCCTGCTGGCGAgagg
SPS_UMI2_BC_rep1_F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGAGCCTCCTGGGCGGGTCATG
SPS_UMI2_BC_rep1_F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNACTGAGCCTCCTGGGCGGGTCATG
SPS_UMI2_BC_rep1_F3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNACTGAGCCTCCTGGGCGGGTCATG
SPS_UMI2_BC_rep2_F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATAGCCTCCTGGGCGGGTCATG
SPS_UMI2_BC_rep2_F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNCGATAGCCTCCTGGGCGGGTCATG
SPS_UMI2_BC_rep2_F3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNCGATAGCCTCCTGGGCGGGTCATG
SPS_UMI2_BC_rep3_F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCAAGCCTCCTGGGCGGGTCATG
SPS_UMI2_BC_rep3_F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNGTCAAGCCTCCTGGGCGGGTCATG
SPS_UMI2_BC_rep3_F3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNGTCAAGCCTCCTGGGCGGGTCATG
Adap_TtgR_S2F_ext	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNCGCTGGCAGATATTGCAGAA
Adap_TtgR_S4R_ext	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNCCAGTTCACCCGGCAG

Supplementary Table 1: Primers

Names and sequences of primers used in this study.

Ligand	Concentration (mol/L)
Tamoxifen	0.00005
Endoxifen	0.00005
4-hydroxy tamoxifen	0.00005
N-desmethyl tamoxifen	0.00005
Naltrexone	0.001
Quinine	0.0005
Ellagic Acid	0.00015
Naringenin	0.001
Phloretin	0.0003

Supplementary Table 2: Ligand concentrations used in the RNA-Seq experiment. Concentrations were based on solubility in aqueous solution. No more than 2% v/v (DMSO) or 1% v/v (EtOH and H_2O) of solvent were tolerated.

Condition	Replicate	Has Constants	Reads Mapped
4Hy	1	3.44E+07	2.12E+07
	2	3.38E+07	2.09E+07
	3	3.37E+07	2.09E+07
DMSO	1	7.50E+07	4.60E+07
	2	7.32E+07	4.49E+07
	3	7.25E+07	4.46E+07
EIIA	1	1.80E+06	1.11E+06
	2	1.87E+06	1.16E+06
	3	8.82E+05	5.42E+05
End	1	1.93E+07	1.19E+07
	2	1.91E+07	1.18E+07
	3	1.97E+07	1.22E+07
EtOH	1	5.30E+07	3.27E+07
	2	5.69E+07	3.52E+07
	3	5.53E+07	3.42E+07
H2O	1	5.79E+07	3.57E+07
	2	5.94E+07	3.66E+07
	3	5.48E+07	3.39E+07
NaI	1	5.06E+07	3.16E+07
	2	6.29E+07	3.91E+07
	3	5.87E+07	3.66E+07
Nar	1	4.05E+07	2.45E+07
	2	4.87E+07	2.92E+07
	3	4.65E+07	2.79E+07
Ndes	1	4.12E+07	2.35E+07
	2	4.04E+07	2.35E+07
	3	3.88E+07	2.22E+07
Phlo	1	4.10E+07	2.47E+07
	2	5.38E+07	3.21E+07
	3	4.99E+07	2.98E+07
Quin	1	3.87E+07	2.38E+07
	2	4.21E+07	2.58E+07
	3	4.13E+07	2.54E+07
Tam	1	3.01E+07	1.84E+07
	2	3.14E+07	1.93E+07
	3	3.20E+07	1.97E+07

Supplementary Table 3: Read counts for DNA from the agnostic library for different conditions.

Read counts after merging and quality filter based on the presence of the constant regions surrounding the 16nt barcode and UMI. Reads mapped are the number of reads that correspond to a barcode identified in the mapping sequencing run.

Condition	Replicate	Has Constants	Reads Mapped
4Hy	1	2.10E+07	9.66E+06
	2	2.22E+07	1.02E+07
	3	2.21E+07	1.01E+07
DMSO	1	2.38E+07	1.04E+07
	2	2.55E+07	1.12E+07
	3	2.41E+07	1.05E+07
EllA	1	1.55E+07	6.70E+06
	2	2.32E+07	1.00E+07
	3	2.32E+07	1.00E+07
End	1	2.43E+07	1.14E+07
	2	2.33E+07	1.09E+07
	3	2.54E+07	1.19E+07
EtOH	1	1.96E+07	8.70E+06
	2	2.28E+07	1.01E+07
	3	2.33E+07	1.03E+07
H2O	1	2.09E+07	8.98E+06
	2	2.19E+07	9.36E+06
	3	2.26E+07	9.77E+06
NaI	1	1.90E+07	8.31E+06
	2	2.18E+07	9.41E+06
	3	2.15E+07	9.35E+06
Nar	1	2.47E+07	1.34E+07
	2	2.77E+07	1.50E+07
	3	2.68E+07	1.45E+07
Ndes	1	2.28E+07	1.09E+07
	2	2.31E+07	1.10E+07
	3	2.44E+07	1.17E+07
Phlo	1	2.52E+07	1.37E+07
	2	2.71E+07	1.47E+07
	3	2.72E+07	1.47E+07
Quin	1	2.13E+07	9.92E+06
	2	2.39E+07	1.11E+07
	3	2.40E+07	1.11E+07
Tam	1	2.60E+07	1.23E+07
	2	2.49E+07	1.17E+07
	3	2.44E+07	1.13E+07

Supplementary Table 4: Read counts for DNA from the agnostic library for different conditions.

Read counts after merging and quality filter based on the presence of the constant regions surrounding the 16nt barcode and UMI. Reads mapped are the number of reads that correspond to a barcode identified in the mapping sequencing run.

		EtOH		End		Tam	
Segment	Replicate	Has Constants	Reads Mapped	Has Constants	Reads Mapped	Has Constants	Reads Mapped
1	1	5.09E+06	1.77E+06	5.22E+06	1.82E+06	4.42E+06	1.54E+06
	2	4.90E+06	1.71E+06	6.02E+06	2.09E+06	4.95E+06	1.72E+06
	3	4.90E+06	1.71E+06	5.88E+06	2.04E+06	4.96E+06	1.72E+06
2	1	3.44E+06	8.17E+05	4.26E+06	1.00E+06	4.63E+06	1.09E+06
	2	4.18E+06	9.93E+05	4.38E+06	1.03E+06	5.09E+06	1.20E+06
	3	4.66E+06	1.10E+06	4.68E+06	1.10E+06	5.30E+06	1.25E+06
3	1	3.95E+06	6.99E+05	3.87E+06	6.77E+05	3.67E+06	6.48E+05
	2	3.89E+06	6.94E+05	4.78E+06	8.43E+05	3.75E+06	6.67E+05
	3	5.34E+06	9.48E+05	4.29E+06	7.56E+05	3.68E+06	6.49E+05
4	1	7.53E+06	2.29E+06	7.46E+06	2.26E+06	6.93E+06	2.10E+06
	2	8.86E+06	2.69E+06	8.78E+06	2.65E+06	8.89E+06	2.69E+06
	3	9.02E+06	2.74E+06	8.06E+06	2.44E+06	8.80E+06	2.67E+06
5	1	4.67E+06	1.26E+06	3.62E+06	9.71E+05	5.50E+06	1.47E+06
	2	5.90E+06	1.58E+06	3.53E+06	9.47E+05	5.97E+06	1.59E+06
	3	6.23E+06	1.66E+06	3.56E+06	9.53E+05	6.90E+06	1.83E+06
6	1	2.76E+06	8.28E+05	2.28E+06	6.85E+05	2.25E+06	6.72E+05
	2	3.13E+06	9.39E+05	2.32E+06	6.97E+05	2.56E+06	7.68E+05
	3	2.93E+06	8.83E+05	2.58E+06	7.77E+05	2.57E+06	7.76E+05

Supplementary Table 5: Read counts for DNA from the DMS library under different conditions.
Read counts after merging and quality filter based on the presence of the constant regions surrrounding the 16nt barcode and UMI. Reads mapped are the number of reads that correspond to a barcode identified in the mapping sequencing run.

		EtOH		End		Tam	
Segment	Replicate	Has Constants	Reads Mapped	Has Constants	Reads Mapped	Has Constants	Reads Mapped
1	1	2.18E+06	6.77E+05	4.63E+06	1.48E+06	5.82E+06	1.85E+06
	2	2.18E+06	6.74E+05	4.83E+06	1.55E+06	6.23E+06	1.96E+06
	3	1.97E+06	6.10E+05	4.66E+06	1.48E+06	6.18E+06	1.94E+06
2	1	3.45E+06	6.77E+05	8.34E+06	1.68E+06	7.87E+06	1.56E+06
	2	2.96E+06	5.82E+05	5.06E+06	1.03E+06	7.21E+06	1.45E+06
	3	2.71E+06	5.32E+05	7.27E+06	1.47E+06	7.51E+06	1.49E+06
3	1	2.32E+06	3.03E+05	5.30E+06	7.38E+05	4.50E+06	6.04E+05
	2	2.88E+06	1.68E+05	4.79E+06	6.69E+05	4.58E+06	6.17E+05
	3	2.00E+06	2.62E+05	5.94E+06	8.21E+05	4.38E+06	5.91E+05
4	1	2.53E+06	6.81E+05	6.13E+06	1.70E+06	7.05E+06	1.92E+06
	2	2.61E+06	6.97E+05	6.42E+06	1.77E+06	7.29E+06	1.98E+06
	3	2.30E+06	6.17E+05	6.78E+06	1.87E+06	5.69E+06	1.55E+06
5	1	1.50E+06	3.28E+05	3.78E+06	8.75E+05	4.24E+06	9.54E+05
	2	1.48E+06	3.24E+05	3.72E+06	8.58E+05	4.28E+06	9.58E+05
	3	1.39E+06	3.03E+05	3.62E+06	8.37E+05	3.88E+06	8.79E+05
6	1	2.43E+06	6.48E+05	5.37E+06	1.50E+06	5.27E+06	1.46E+06
	2	2.28E+06	6.07E+05	5.27E+06	1.47E+06	5.33E+06	1.47E+06
	3	2.06E+06	5.50E+05	5.25E+06	1.45E+06	4.95E+06	1.35E+06

Supplementary Table 6: Read counts for RNA from the DMS library under different conditions.
Read counts after merging and quality filter based on the presence of the constant regions surrrounding the 16nt barcode and UMI. Reads mapped are the number of reads that correspond to a barcode identified in the mapping sequencing run.

4.0.0 Conclusions and Future Directions

4.1.0 Conclusions

Allosteric transcription factors (aTFs) have fundamental roles in a wide range of cellular processes due to their ability to alter gene expression at the transcription level. aTFs have the capacity to interact with small molecules or proteins, bind to DNA, and undergo allosteric changes. However, little is known about the molecular mechanisms that give rise to these functions. These proteins have evolved from ancient predecessors; the molecular interactions that confer the ability to control gene expression have been preserved under selective pressures. Mutations fixed during evolution can affect each fitness parameter independently and have a nonadditive effect on gene expression control. This work took the initial steps in understanding the epistatic effect of amino acid interactions across multiple fitness parameters and the sequence-function relationship underlying ligand specificity. I developed a novel design-screening workflow that can be applied to create novel aTF biosensors.

4.1.1 Epistasis across multiple fitness parameters

To understand the complexity of the evolution of novel function in ATFs, I used an engineered TtgR variant with resveratrol specificity through the addition of four mutations: C137I, I141W, M167L, and F168Y. By assaying all combinations of these four mutations against naringenin, the native ligand, and resveratrol, the target ligand, I show that epistasis affects both protein functions but to different extents. I used dose response curves to characterize the epistatic interactions in basal gene expression, maximum gene expression, and EC50 between both ligands. Epistasis is pervasive, but unique through across all parameters of transcription factor function. However, pairs of epistatic residues, such as C137I+I141W and M167L+F168Y were consistent across multiple fitness parameters.

This study utilized computational design to engineer a specificity switch into TtgR, which constrained mutations to a select set of residues in the ligand binding pocket. In the natural evolution of an aTF, mutations that increase fitness can occur throughout the primary sequence and become fixed through natural selection. The mutations presented in this work are a single solution to confer resveratrol specificity, but evolution could have selected a different set of mutations at different positions. By selecting mutations at positions that directly interact with the ligand, I focused on the effect of epistasis on both biophysical interactions and biological function.

Epistasis in the development of novel function in aTFs is intricately linked to all facets of gene expression control. Classically, epistasis has been visualized as the ruggedness of a fitness landscape[1,2]. In these examples, the height of the landscape is the measure of a single functional parameter like binding affinity, stability, or catalytic activity. An evolutionary process will navigate a combination of all parameters in complex functions, which can be envisioned as a multidimensional fitness surface. I show that a single sequence may not be optimal in all parameters. For example, combinations of the four mutations in this study showed that as basal fluorescence decreases, sensitivity also decreases. Furthermore, this work implies that varying selection pressures on an evolutionary scale may enable multifunctional proteins to bypass fitness barriers in different landscapes. In our case, higher order epistasis prevents access to the quadruple mutant in the naringenin fold induction landscape but could be bypassed by selection pressure in the resveratrol response fitness landscape. The evolution of allosteric proteins is inherently dependent on epistasis and the interactions arising between mutations in these proteins uniquely affects multiple adaptive landscapes.

4.1.2 Engineering novel ligand affinity into TtgR

Allosteric transcription factors have important applications as biosensors in biotechnology as these proteins naturally control gene expression in response to small molecules in the environment[3]. However, these proteins remain difficult engineering candidates despite high demand to bring designer transcription factors into biotechnology as specific, sensitive, *in vivo* biosensors[4,5]. A natural transcription factor that interacts with the desired molecules with high sensitivity may not exist, creating a need to engineer new affinities into existing aTFs. This redesign requires the manipulation of ligand binding interactions and the preservation of existing allosteric interactions.

In chapter 3, I use a combination of a phylogenetically derived library and an RNA-Seq screening workflow that enables screening of 16,000 transcription factor variants across multiple ligands. By selecting a ligand-agnostic library design workflow, the resulting variants may have altered affinity for any small molecule[6].

The RNA-Seq workflow was validated with a small library consisting of the combinatorial mutants assayed in chapter 2. I used the RNA-Seq workflow to probe the agnostic library for response on tamoxifen, 4-hydroxytamoxifen, endoxifen, N-desmethyltamoxifen, naltrexone, ellagic acid, quinine, naringenin, and phloretin. I selected the top performing variants from each ligand based on RNA-Seq data and inserted them into a *sfGFP* reporter system to screen via fluorescence-activated cell sorting. The fluorescence screen indicated that the variants responded to the ligands, validating the existence of novel sensors in the agnostic library.

Sequence trends between variants suggest that the RNA-Seq screening approach reveals amino acid preferences that are consistent in the top performers. By analyzing the positional entropy and amino acid distribution for each mutable position, I found that similar ligands had similar preferences for amino acids at positions with high selectivity. These preferences were also shared on the subset of best-performing variants for each ligand, suggesting that the RNA-Seq can be used to inform future design workflows by incorporating fold enrichment data.

In addition to screening the agnostic library, I applied the RNA-Seq approach to a deep mutational scan library of TtgR against endoxifen and tamoxifen. I found that this approach can characterize thousands of variants across multiple biological replicates. Mutational hotspots were across the protein. A small subset of these positions was located between the DNA binding domain and the ligand binding domain, suggesting structural or functional importance. Thus, the RNA-Seq workflow can also be applied to understand the underlying biology of gene expression.

4.2.0 Future Directions

4.2.1 Epistasis in allosteric transcription factors

The concept of multidimensional epistasis in aTF evolution has not been examined exhaustively. With the increase in next-generation sequencing read capacity, assaying both gene expression and its functional parameters is now possible. Assuming that read volumes are no longer a limitation in the future, assaying large combinations of mutations to probe the depth and strength of epistasis can be done more deeply than before[7]. This work is the first step in observing the effects of mutations across multiple fitness parameters. Next steps include expanding the transcription factor variant library and increasing the number of ligands screened.

The set of four mutations that provided resveratrol specificity is only one solution to changing ligand specificity in TtgR. The next step is to ascertain if these four mutations are the only solution at both the four positions and across the entire protein sequence. Searching sequence space at the four mutable positions (137, 141, 167, and 168) would give greater insight into the epistatic relationship between each position. C137I and I141W created an epistatic interaction in this work, but the pairing may be specific for these two substitutions. This experiment would elucidate the underlying intersection of fitness parameters across multiple substitutions and positions. I expect that a global maximum of fitness across all parameters will be inaccessible as fitness tradeoffs are an innate property of evolution[8,9]. This result can be visualized as a Pareto front across each functional parameter.

Mutagenesis across the protein domain in a combinatorial fashion has been performed to a limited extent[10]. Future experiments need to increase the number of combinatorial mutations sampled to better characterize the extent of epistasis in the fitness landscape. For TtgR, a protein with approximately 200 amino acids, mutating any four positions to the non-wildtype amino acids will yield a library of approximately 10^{19} variants. Assuming that technology has advanced to the point that sequencing this number of variants is feasible, this theoretical library will reveal both functional hotspots and epistatic trends at a greater depth than previous experiments.

Understanding epistasis in promiscuous aTFs requires assaying both an extensive transcription factor library across multiple ligands. Additional ligands, like phloretin or tetracycline, can give insight into the entire functional landscape of TtgR. Response to these ligands may be individually affected by the mutations across the three different fitness parameters. I expect that each ligand will have a unique set of optimal sequences and that the current sequence of TtgR is not optimal for each ligand based on other works examining the evolution of functional specificity[11,12].

By increasing the number of mutations, positions, and ligands in transcription factor variant libraries, the next studies can probe the fitness surface of aTF and examine higher-order epistatic interactions. In the future, I believe that libraries of all combinations of mutations can be created and assayed using next-generation sequencing. These studies can determine the exact prevalence and importance of epistasis across transcription factor function.

4.2.2 Computational design and RNA-Seq

I have shown that the computational design workflow and RNA-Seq screening approach have successfully generated variants with function on eight novel ligands. This workflow requires final validation of variant function, but additional ligands, mutations, and scaffolds create the potential for developing an extensive set of designer biosensors for biotechnology.

Each functional sequence identified via fluorescence screen must be validated in a clonal qRT-PCR and fluorescence assay to verify function. I expect that top performing variants have some response to at least one of the ligands but may be promiscuous. Clonal validation of function will provide a small list of variants that can then be crystallized with one of the eight ligands. At this point, specificity or fold enrichment can be improved using directed evolution to create highly sensitive, highly active sensors.

Using an agnostic approach to computational design had multiple advantages over targeted design. First, mutations selected through the computational workflow were derived from a phylogenetic alignment. The evolutionary history of these mutations theoretically creates more stable alterations in the protein structure compared to random mutagenesis. Second, the same library can be screened across multiple ligands, streamlining the screening process. I believe that as the capacity for chip oligo synthesis and next generation sequencing increases, the next agnostic libraries can incorporate additional positions and mutations. Combinatorial mutations across all allowed positions in TtgR creates a library of 151,165,440 variants. Future screens can incorporate these variants, which have mutations throughout the entire binding pocket, to create new biosensors for additional small molecules. I believe that mutating additional positions within the pocket will yield a larger number of functional variants. The set of positions in this study are largely localized to one face of the binding pocket; increasing the potential interactions throughout the binding pocket will enable higher affinities for different molecules.

One goal of the RNA-Seq workflow is to obtain data that can be used to improve targeted computational design. The data from the agnostic library can be used to improve the scoring function weights across the nine ligands to decrease the number of designs that are selected for screening. However, this screening approach should also be applied to the combinatorial mutagenesis libraries described above to create a functional landscape across all positions and multiple ligands. I believe that the nuanced data from this experiment would be invaluable for machine learning approaches to iteratively improve the Rosetta design process beyond the limited set of mutations tested here.

This work is the first step towards creating a toolkit for biosensor design. Now that a workflow to create new sensors is established, I think an immediate follow to this work would be using computational and RNA-Seq screening process on a new system. Although I observed numerous functional variants to the eight selected ligands, TtgR may not be a suitable scaffold to engineer for any random target ligand affinity. However, TtgR is a promiscuous scaffold, which may make it more amenable to adopting novel functions than more specific aTFs like TetR[13]. I believe that other promiscuous scaffolds will have the greatest potential for adopting novel ligand affinity. QacR is another TetR family member that can bind to multiple organic compounds[14]. These proteins could create a central set of scaffolds from which many additional biosensors can be created at will. Future researchers should be able to pick a ligand of interest and have a method to search for the most suitable candidate scaffold from the set. Applying either the targeted design or the agnostic approach with RNA-Seq screening would then generate the novel sensor for metabolic engineering, environmental monitoring, or gene circuits.

4.3.0 References

1 Smith, J. M. Natural Selection and the Concept of a Protein Space. *Nature* **225**, 2 (1970).

2 Kauffman, S. A. & Weinberger, E. D. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *J theor Biol* **141**, 35 (1989).

3 Cuthbertson, L. & Nodwell, J. R. The TetR family of regulators. *Microbiol Mol Biol Rev* **77**, 440-475, doi:10.1128/MMBR.00018-13 (2013).

4 Taylor, N. D. *et al.* Engineering an allosteric transcription factor to respond to new ligands. *Nat Methods* **13**, 177-183, doi:10.1038/nmeth.3696 (2016).

5 Jha, R. K., Chakraborti, S., Kern, T. L., Fox, D. T. & Strauss, C. E. Rosetta comparative modeling for library design: Engineering alternative inducer specificity in a transcription factor. *Proteins* **83**, 1327-1340, doi:10.1002/prot.24828 (2015).

6 Khersonsky, O. *et al.* Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol Cell* **72**, 178-186 e175, doi:10.1016/j.molcel.2018.08.033 (2018).

7 Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409-413, doi:10.1038/nature23902 (2017).

8 McKeown, A. N. *et al.* Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell* **159**, 58-68, doi:10.1016/j.cell.2014.09.003 (2014).

9 Tomatis, P. E. *et al.* Adaptive protein evolution grants organismal fitness by improving catalysis and flexibility. *PNAS* **105**, 6 (2008).

10 Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol* **24**, 2643-2651, doi:10.1016/j.cub.2014.09.072 (2014).

11 Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nat Genet* **37**, 73-76, doi:10.1038/ng1482 (2005).

12 Eick, G. N., Colucci, J. K., Harms, M. J., Ortlund, E. A. & Thornton, J. W. Evolution of minimal specificity and promiscuity in steroid hormone receptors. *PLoS Genet* **8**, e1003072, doi:10.1371/journal.pgen.1003072 (2012).

13 Alguel, Y. *et al.* Crystal structures of multidrug binding protein TtgR in complex with antibiotics and plant antimicrobials. *J Mol Biol* **369**, 829-840, doi:10.1016/j.jmb.2007.03.062 (2007).

14 Schumacher, M. A. *et al.* Structural Mechanisms of QacR Induction and Multidrug Recognition. *Science* **294**, 6 (2001).