Rapidly evolving proteins with conserved low-specificity

Lucas C. Wheeler[1,2] and Michael J. Harms[1,2*]

1. Institute of Molecular Biology, University of Oregon, Eugene OR 97403

2. Department of Chemistry and Biochemistry, University of Oregon, Eugene OR 97403

Abstract

sddasgf

Introduction

Changes in specificity are critical for protein evolution (1–8). Many proteins have low-specificity interfaces that interact with wide swaths of cellular targets. XX% of protein-protein interactions are thought to be mediated by interfaces that bind to XX or more targets. Key among these are protein-peptide interactions, where a protein recognizes short linear peptides of target proteins. Despite the importance and ubiquity of such interfaces, their evolution remains poorly understood.

A major barrier to answering questions such as these has been quantifying the specificity of low-specificity proteins. For a high-specificity interface, specificity can be readily summarized using a sequence logo or relatively simple structure-activity relationship. For a low-specificity interface, such a description can be difficult or impossible to define. Recent advances in high-throughput characterization and machine learning, however, can overcome this barrier. By combining these approaches, we can characterize changes in the specificity of low-specificity interfaces over evolutionary time.

Members of the S100 protein family are an excellent model system to study the evolution of low-specificity interfaces. These proteins bind to $\approx$ 12 amino acid linear peptide regions of target proteins to modulate their activity (Fig 1A) (9–17). Family members bind to extremely diverse peptide targets. The diversity is such that defin-

ing recognition rules is difficult. Fig 1XX shows a collection of peptides with known binding properties to human S100A6 (Fig 1B). All binding peptides—including several known biological targets—have similar affinity, despite their radically different sequences.

Surprisingly, despite the low-specificity of these proteins, there are strong evolutionary constraints on specificity. We previously tested the binding of a collection of diverse peptides to two closely-related S100 paralogs—S100A5 and S100A6—sampled from across amniotes. We found that specificity within each clade had been conserved for over 300 million years (Fig 1C). This specificity appears to be under selection, as it was maintained even as the proteins rapidly evolved. The average sequence identity across orthologs is XX% and XX% for S100A5 and S100A6, respectively. Additionally, the specificity does not appear to the result of chemical necessity, as we could alter the specificity of human S100A5 with a single point mutation.

Given the juxtaposition of a low-specificity interface with strong conservation of a subset of its partners, we set out to understand the evolution of these proteins in greater detail. In particular, we posed the following two questions. What are the recognition rules used by these proteins? How have they changed over time? We set out to define these rules using high-throughput phage display coupled to machine learning, while tracing how they evolved using ancestral sequence reconstruction. We find that both the modern and ancestral interfaces recognize shape complementarity, but that subtle changes in binding rules led to altered specificity over time.

Results

Binding peptides can be sampled by phage display

We set out to determine the peptide binding specificity of human S100A5 (hA5), human S100A6 (hA6), and their last common ancestor (ancA5/A6) (18). To account for phylogenetic uncertainty in the reconstruction, we also studied peptide binding to an alternate version of ancA5/A6 (altAll) (18).

We first assayed the binding of tens of thousands of peptides to each protein using phage display. We panned a commercial library of randomized 12-mer peptides expressed as fusions with the M13 phage coat protein. The S100 peptide-binding interface is only exposed upon Ca^{2+}-binding (Fig 1A); therefore, we performed phage panning experiments in the presence of Ca^{2+} and then eluted the bound phage using EDTA (Fig 2A). The population of enriched phage will be a mixture of phage that bind at the site of interest and phage that bind adventitiously (blue and purple phage, Fig 2A). Peptides in this latter category enrich in Ca^{2+}-dependent manner through avidity or binding at an alternate site (19, 20). To separate these populations, we repeated the panning experiment in the presence of a competitor peptide known to bind at the site of interest (Fig 2B) (18). This should lower enrichment of peptides that bind at the site of interest, while allowing any adventitious interactions to remain. By comparing the competitor and non-competitor pools, we can distinguish between actual and adventitious binders.

We performed this experiment with and without competitor, in biological duplicate, for all four proteins. We found that phage enriched strongly for all proteins

relative to a biotin-only control (Fig S1). Further, the addition of competitor binding knocked down enrichment in all samples (Fig S1). After panning, we sequenced the resulting phage pools, as well as the input library, using Illumina sequencing. We applied strict quality control, discarding any peptide that exhibited less than six counts (see methods, Fig S2). After quality control, we had a total of 265 million reads spread over 17 samples (Table S1).

We estimated changes in the frequencies of peptides between samples with and without competitor peptide. For each peptide i, we determined $E_i = -ln(\beta_i/\alpha_i)$, where β_i and α_i are the frequencies of the peptide in the non-competitor and competitor samples, respectively. Defined this way, a more negative value of E corresponds to a larger decrease in peptide frequency upon addition of competitor peptide. We used a clustering approach to estimate E for $\approx 40,000$ different peptides for each protein (see methods, Fig S3). We found that the distribution of E could be described distribution using two Gaussian distributions, apparently reflecting two underlying processes (Fig 3A, Fig S4). The dominant peak consists of "unresponsive" peptides whose frequencies change little in response to competitor peptide. A second, broader, distribution describes "responsive" peptides whose frequencies change with the addition of competitor.

There was no systematic difference between estimates of E between biological replicates. We used orthogonal distance regression to compare values of E for peptides seen in both biological replicates. The slopes of these lines ranged from 0.9 to 1.1, with intercepts between -0.09 and 0.11. hA5, for example, had a slope 1.06 and an intercept of -0.05 (Fig 3B). There are two distinct regions in these correlation

plots, corresponding to the unresponsive and responsive peptide distributions. The unresponsive distribution forms a large cloud about zero. In contrast, the responsive peptide distribution extends along the 1:1 line in a correlated fashion. If we focus on values of $E < -1$—peptides mostly in the "responsive" distribution—the 1:1 axis of variation explains 87.8% of the total variation in the data. The worst correlation, observed for altAll, was 75.8%.

Supervised machine learning reveals the rules for binding

As with previously studied peptides (Fig 1X), the peptides recovered in our phage display experiment are extremely diverse. This precludes using simple sequence-based rules to predict binders. We therefore set out to predict binding from chemical features for each amino acid sequence. We used supervised machine learning to train models against 57 chemical features that we could could readily calculate from an amino acid sequence. These included measures of hydrophobicity, hydrogen bonding, geometry, secondary structure propensity, and electrostatics. In addition to these specific features, we also defined 20 "meta" features by taking the principle components of the entire aaindex database (21), which reports 590 quantitative values for each of the 20 amino acids. For most chemical features, we simply added the values for each amino acid in a sequence. For example, we would sum up the number of hydrogen bond donors across the sequence and treat that as a chemical feature. We also used CIDER to calculate a few non-additive electrostatic features for each sequence (22), such as the isoelectric point. A full list of the features we calculated is given in Table S2.

We calculated these 57 features for the entire sequence and for all sliding windows ranging from 1 to 11 amino acids (Fig 4A). This introduces neighbor-neighbor correlation between features and improves model power. Overall, we calculated the features for 78 sliding windows on each peptide, giving us a total of $57 \times 78 = 4,446$ features per sequence (Fig 4A). We then trained a random forest regression model to predict E using the features of the $\approx 40,000$ we observed for each protein. A random forest model finds weights for a collection of random decision trees based on a set of input features (23). Prior to training, we withheld 10% of the peptides as a test set. We then optimized nuisance parameters such as the number of trees and choice of data weighting scheme using k-fold cross validation within the training set ($k = 10$). After training, the R^2 between our model and the training set was $\approx 97\%$ for all proteins (Table 1).

After our final optimization, we tested our models against their test sets. R^2 for test sets ranged from $80 - 85\%$ (Fig 4B, Table 1). For all models, the regression line reveals a slope slightly greater than one (e.g. 1.16 for hA5, Fig 4B). Further, the scatter is nonrandom, with the most negative values of E being overestimated and the most positive values underestimated. This makes intuitive sense, as the best-of-the-best and the worst-of-the-worst enriching sequences likely depend strongly on details not captured by our rather crude amino acid model.

All proteins recognize shape and hydrophobicity, not traditional high-specificity properties such as hydrogen polar contacts

Given that our model predicts ≈ 80% of the variation in enrichment, it can provide mechanistic insight into which chemical features and peptide positions are the most important for enrichment. We calculated the contribution of each feature. (by XX method) We then pooled features based on their similarity. For example, sidechain volume and beta-chain "knob" propensity were pooled into "geometry" (along with XX other terms). Predicted charge and number of hydrogen bonds, on the other hand, were pooled into "polar" interactions. A full list of the individual features and their bins is given in Table S3.

We then plotted the relative contribution of each property as a function of peptide position for hA5, hA6, ancA5/A6, and altAll (Fig 5). The different proteins led to highly similar models. For all four proteins, each site contributed almost equally to the predicted enrichment, ranging from XX% to XX%. In contrast, different molecular properties had radically different contribution levels. Sidechain geometry, relative hydrophobicity, and secondary structure propensity dominated the predictive power of the model. In contrast, the typical determinant of specificity—polar contacts—had almost no predictive power.

Despite minor quantitative differences, all four proteins exhibited nearly identical binding profiles. In all cases, the trained models indicate that specificity is determined entirely by shape complementary and hydrophobic surface area—not polar

contacts. Further, specificity is distributed across the peptide, rather than being concentrated onto one or two key sites. Thus, whatever differences in specificity exist between these proteins, it is achieved by slight modification of existing binding rules rather than through a radical change in recognition mechanism.

Trained models can be used to classify peptides as binders vs. non-binders

Although these proteins have similar recognition mechanisms, our previous low-throughput work indicated that each protein bound different binding targets. Specifically, we found that hA5 and hA6 bound to a subset of the peptides recognized by ancA5/A6 and altAll. To probe for this behavior, we next used our model to estimate the Venn diagram describing the binding sets for the modern and ancestral proteins by applying models for all four proteins to a common collection of 1,000,000 random 12-mer peptides.

Calculating this Venn diagram required classifying peptides as binders or non-binders. We therefore converted our quantitative model into a classifier. To facilitate this comparison, we normalized E for each protein such that the competitor peptide had an enrichment value of -1. We did this by $E_{norm} = E/|E_{comp}|$, where E_{comp} is the enrichment of the competitor peptide. We then classified peptides into the categories $E_{norm} < -1$ vs. $E_{norm} \geq -1$, corresponding to binding better or worse than the competitor peptide.

We swept along cutoffs in predicted values of E_{norm} and calculated our false positive and false negative rate using the measured values of E_{norm} for test-set peptides.

As expected, increasing the cutoff increased the false positive rate and decreased the false negative rate for each model. We quantified this behavior with Receiver Operator Characteristic (ROC) curves. A ROC curve is a plot of the true positive rate against the false positive rate as one changes the classifier cutoff. A perfect predictor will have a cutoff value where the false positive rate is 0 and the true positive rate is 1. As a consequence, the Area Under the Curve (AUC) will be 1.0. In contrast, a random predictor will follow the 1:1 line and will have an AUC of 0.5. All of our models had steep ROC curves that gave AUC values from 0.95 to 0.99 (Fig 4C). Given the amino acid sequence of a 12-mer peptide, we can therefore predict with high confidence whether a peptide enriches better or worse than the competitor peptide in a phage display experiment.

To compare our phage display results to our isolated peptide binding experiments, we next calibrated our phage enrichment values against binding of isolated peptides. We did this by calculating E_{norm} for 44 peptide/protein pairs and then measuring their binding using Isothermal Titration Calorimetry (Table S3). We used 17 synthetic peptides, some with known binding properties (10, 14, 16, 18), others that were in the freezer for other projects, and still others were extracted from the human proteome as possible S100 targets. We measured binding of 16 of these peptides to hA5, 13 to hA6, 8 to ancA5/A6, and 6 to altAll. We classified any peptide with a measurable binding constant by Isothermal Titration Calorimetry ($K_D \lesssim 100\ \mu M$) as "binding" and all others as "non-binding."

We then swept along E_{norm} and attempted to classify the 44 measured binders. The ROC curve came off the diagonal, with an AUC of 0.71 (Fig 4C). While this

is significantly worse than the prediction of enrichment in the phage display experiments, it is not unexpected. We trained our model on phage display data and are now attempting to use it to predict isolated peptide binding. To verify that this AUC curve indicated real binding signal, we simulated 44-observation ROC curves using a random predictor. We found that the probability of observing an AUC of 0.71 or greater by chance was 0.007—a strong indication of signal in our binding model (Fig S5).

To identify a cutoff for predicting binders, we plotted the false positive rate and false negative rate against E_{norm} for all 44 peptides. We then identified the value of E_{norm} that simultaneously minimized the false positive and false negative rates. To estimate the crossover point, we fit the modified Hill equation to each curve, which empirically captures the basic shape of these curves (Fig 4D). We found that these curves crossed for $E_{norm} = -1.19$, with false positive and false negative rates of ≈ 0.35. These rates are high and therefore preclude confidently predicting whether a given peptide binds. This is, however, sufficient to determine a Venn diagram for the binding specificity of these proteins.

Venn diagrams can be estimated from predicted binders

We next used our trained and calibrated models to estimate the Venn diagram describing the binding sets for the modern and ancestral proteins. We applied our models for all four proteins to a common collection of 1,000,000 random 12-mer peptides, classifying any peptide with $E_{norm} < -1.19$ as binding. We then calculated the overlap between these sets, placing the counts for each region of the Venn diagram

into the vector $\vec{V}_{obs}$.

Because we have high (and uncertain) false positive and false negative rates, the counts in $\vec{V}_{obs}$ may not be identical to the real populations of the Venn diagram ($\vec{V}$). We therefore sampled over counts in $\vec{V}$, as well as possible false positive and false negative rates, using Bayesian Markov Chain Monte Carlo (MCMC). We wrote a transition matrix $\mathbf{T}$ that maps $\vec{V}$ into $\vec{V}_{obs}$ ($\vec{V}_{obs} = \vec{V} \cdot \mathbf{T}$). $\mathbf{T}$ defines the probability of each class of miss-call given all false positive and false negative rates. As an example, one element in $\mathbf{T}$ encodes the probability that we mistakenly identify a hA5-specific peptide as a hA6-specific peptide (e.g. the false negative rate for hA5 times the false positive rate for hA6). The details of matrix construction are given in the supplemental text.

We allowed each protein to have its own false positive and false negative error rates. We set the prior probabilities for error rates by estimating the false positive and false negative rate for binding to each protein at the cutoff of $E_{norm} < -1.19$ (Fig S6, supplemental text). We then used MCMC to sample values of $\vec{V}$ and the error rates, comparing the resulting vector to $\vec{V}_{obs}$. We ran two samplers in parallel until convergence (≈ 2 million steps each). This strategy allowed us to estimate the Venn diagram while incorporating our uncertainty in its composition.

hA5 and hA6 exhibit opposite trends in specificity

We constructed a Venn diagram of the peptide targets bound by hA5, hA6, and ancA5/A6. We found that the total size of each binding set ranged from 1.3% [0.9,1.8] of peptides (for hA5) to 10.1% [9.2,11.7] of all peptides (for hA6) (Fig 6).

The values in the brackets denote the 95% credibility interval from the posterior distribution. The large sizes of these sets reflects the low-specificity, hydrophobic nature of the S100 binding interface (14, 18).

We found that hA5 exhibits increased specificity relative to ancA5/A6. The hA5 peptide set is a subset of the ancestral binding set (Fig 5). While 85% [xx,xx] of peptides are shared with the ancestor, only 9% [xx,xx] of peptides were specific to hA5. The hA5 peptide set was also largely a subset of the hA6 set: 80.6% [76.8,88.5] of hA5 peptides are also hA6 peptides. hA5 thus binds a subset of peptides that mostly overlap with both the ancestor and the hA6 paralog (Fig 6).

The hA6 binding set was much larger than hA5—consisting of 10.1% [9.2,11.7] of peptides. This is is expanded relative to the ancA5/A6 set. While there is a extensive overlap (37.4% [36.4,38.4]), most hA6 binding targets were acquired after gene duplication (Fig 5). Fully 62.0% [61.1,63.1] of peptides are unique to hA6. Relative to ancA5/A6, hA6 kept its ancestral partners, and then added a large collection of new partners. Thus, despite the apparent pattern of increased specificity for both proteins taken from the small peptide samples, hA5 and hA6 exhibit opposite changes in specificity relative to ancA5/A6.

The maximum-likelihood and altAll constructs give different results

We next compared the results for our two versions of the ancA5/A6 ancestor. In addition to the maximum likelihood ancestor characterized in Fig 5, we also characterized an "altAll" ancestor in which we substituted the amino acid state with the

next highest posterior probability at each ambiguous site (18, 24). This approach to characterizing uncertainty is relatively new and has, to this point, only shown results concordant with the ML ancestor (18, 24, 25) [xx Harms_eick_biophysics]. This is a very aggressive attempt at capturing uncertainty: 21 of 86 sites are different between ancA5/A6 and altAll.

We first directly compared the binding sets of ancA5/A6 and altAll. altAll had a much larger set of targets than ancA5/A6, binding to 25.4% [xx,xx] vs. 7% [xx,xx] of the random peptides (Fig 6A). Despite the large difference in set size, the two proteins shared many partners. Indeed, ancA5/A6 was essentially a subset of altAll, with 93% [xx,xx] of its binding set being within the altAll set (Fig 7A). We then compared altAll to hA5 and hA6. As with the ancA5/A6 construct, hA5 was essentially a subset of the ancestral state (Fig 7B). hA6, however, exhibited different behavior. Because the altAll set is so much larger than the ancA5/A6 set, the hA6 set is no longer larger than the ancestral state. It does have new partners relative to altAll, but 85% of its partners overlap with the altAll set. Thus, relative to altAll, both hA5 and hA6 gained specificity relative to the ancestor.

Discussion

These high-throughput experiments, coupled to machine learning, provide useful insight into the evolution of these low-specificity proteins. First, they reveal that the basic rule-set has remained unchanged between S100A5 and S100A6 paralogs: recognition is mediated through shape recognition. This is borne out structurally

for the one crystal structure of rabbit S100A6 bound to a peptide target. The interaction is mediated by a long spine of hydrophobic interactions, with strikingly few polar contacts. We would predict that other S100/peptide interactions in this S100 subfamily would exhibit similar recognition rules. Despite sharing a similar overall rule set, however, these proteins do indeed bind to very different sets of peptides. S100A5 binds to a much smaller set of peptides than S100A6.

On the sizes of sets

The large sets for each protein likely reflect the hydrophobic nature of the hA5 and hA6 binding interfaces (14, 18). The binding set of hA6 may be larger than that of hA5 due to its extended binding surface relative to other S100 proteins (10). This larger extended surface may allow it to accommodate peptides that wrap around the protein and bind into an extended groove. This may explain both its broader specificity and the acquisition of targets not observed in the ancestral protein.

It remains unknown whether the hA5 and hA6 binding sets are shared among modern orthologs, or whether these sets have fluctuated relative to one another. We previously found strong evidence for conservation of specificity—for a small set of peptides—in orthologs across amniote species (18). This suggested an overall conservation of biochemical specificity in the S100s. However, as noted above there is insufficient sampling in the low throughput experiments to distinguish differences in the overall specificity of the proteins. Thus, the high-throughput approach used in this study would need to be applied to sets of orthologs to determine the degree to which specificity is conserved across orthologs.

Biological targets

Interestingly, the scope of these binding set sizes mirrors the tissue distributions of the two proteins. In mammals, S100A5 has an extremely narrow tissue distribution, being found primarily in the olfactory bulb and olfactory sensory neurons (26–28). In contrast, S100A6 is expressed ubiquitously. This is counterintuitive if one starts with the "parsing environment" perspective, as S100A6 has broader specificity even while experiencing more diverse environments (29).

There are, at least, two evolutionarily relevant ways to view protein specificity. The first is specificity between biological targets. Can a protein discriminate between targets A and B, both seen in its cellular context? The second is specificity between *all* possible targets, whether seen biologically or not. Understanding the latter class of specificity is particularly important for understanding the evolution of new functions. Possible, but unrealized, molecular partners—sometimes called "promiscuous" partners—are a rich source of raw material for future evolution. When a partner in the promiscuous arises in the cell, a new interaction already exists that can then be optimized by natural selection.

Previous studies of the evolution of specificity have focused largely on the evolution of specificity for biological targets. Such studies have provided deep insight into the evolution and mechanism of biological systems. They have also revealed a common theme: after gene duplication, ancestral partners are often partitioned among the descendant paralogs (1, 7, 8, 18, 29–36).

Phylogenetic uncertainty and specificity

We observed a large difference in the binding sets for our two ancestral reconstructions, ancA5/A6 and altAll. The altAll protein is a very aggressive attempt to incorporate phylogenetic uncertainty, simultaneously introducing alternate amino acid states at 21 sites. The overall behavior of the protein may thus be compromised by the combination of a large number of unlikely and potentially unfavorable residues. We therefore believe the ancA5/A6 protein gives the best estimate of the evolutionary process: the overall binding set shrank along the hA5 lineage and grew along the hA6 lineage. This said, we cannot exclude the possibility that the actual evolutionary transition more resembles that of altAll or is somewhere between the two sets.

Although these two proteins are only two data points, they are consistent with a relationship between sequence error and specificity. It is possible that, as more errors accumulate in a reconstruction, the less specific the protein is. This could be tested systematically by measuring the overall specificity of the same ancestor as a function of introducing less and less likely states. If such a relationship holds, one might even imagine that the overall trend of low-specificity ancestors is an artifact of errors in the reconstruction. Maybe historical proteins had identical levels of specificity as modern proteins, but our method for studying them has led apparent low specificity. This would also point to a different starting point for protein engineers. Maybe, rather than starting from the best estimate of the ancestral protein, they should start with something like an "altAll." This introduces noise that may lead to a less optimal protein that can act as a starting point for future evolution.

On the evolution of increased specificity

One intriguing suggestion is that, on average, proteins become more specific over evolutionary time (37–39). If true, this would be a directional "arrow" for protein evolution (32, 39–41). Such features are controversial (39, 42), but could ultimately provide fundamental insights into the evolutionary process. For example, increasing specificity might indicate that proteins become less evolvable over time, as they have fewer promiscuous interactions that can be exploited to acquire new functions (2, 38). From a practical standpoint, it has also been suggested that less-specific reconstructed ancestors would be powerful starting points for engineering new protein functions (30).

Much of the empirical support for the increasing-specificity hypothesis comes from ancestral reconstruction studies, for example, that shown in Fig 1C. Our work, however, demonstrates that existing data are insufficient to answer these questions one way or the other. Observations about specificity made on small numbers of biological targets are insufficient to reveal changes in the overall binding set of the protein. We then saw this empirically: while a low-throughput analysis showed that hA5 and hA6 both increased their specificity relative to ancA5/A6 (Fig 1A), a high-throughput analysis of the same three proteins showed a different pattern: hA5 increased its overall specificity, while hA6 decreased its overall specificity and gained entirely new targets (Fig 5).

The ready availability of high-throughput experiments and powerful statistical approaches such as machine learning now make it possible to characterize the overall specificity of proteins—at least, for individual classes of partners such as peptides.

By applying these to reconstructed proteins over deep evolutionary time, we may be able to detect any trends in overall specificity.

We speculate, however, that there is ultimately no trend towards increased specificity in proteins. Protein specificity is likely in an evolutionary regime of mutation-selection balance. If we start with the assumption that many more protein sequences encode low-specificity than high-specificity, random mutations will tend to decrease specificity. Selection will maintain specificity for biological targets, but not promiscuous interactions that are not realized biologically. As a result, proteins will only as specific as they need to be, but will be largely nonspecific with regard to non-biological targets. Even if selection for new biological specificity causes a brief increase in overall specificity, drift will cause specificity between non-biological targets to relax back to the lowest specificity compatible with the protein's biological function. This process does not mean a protein will never increase its overall specificity—as we observed, for example, for hA5 (Fig 5)—but it does imply that absent selection for higher specificity, proteins will tend to lose what specificity they have.

Materials and Methods

Molecular cloning, expression and purification in of S100 proteins

Proteins were expressed in a pET28/30 vector containing an N-terminal His tag with a TEV protease cleavage site (Millipore). For each protein, expression was carried out in Rosetta *E.coli* (DE3) pLysS cells. 1.5 L cultures were inoculated at a 1:100

ratio with saturated overnight culture. *E.coli* were grown to high log-phase (OD_{600} ≈0.8–1.0) with 250 rpm shaking at 37 $°C$. Cultures were induced by addition of 1 mM IPTG along with 0.2% glucose overnight at 16 $°C$. Cultures were centrifuged and the cell pellets were frozen at 20 $°C$ and stored for up to 2 months. Lysis of the cells was carried out via sonication on ice in 25 mM Tris, 100 mM NaCl, 25 mM imidazole, pH 7.4. The initial purification step was performed at 4 $°C$ using a 5 mL HiTrap Ni-affinity column (GE Health Science) on an Äkta PrimePlus FPLC (GE Health Science). Proteins were eluted using a 25 mL gradient from 25-500 mM imidazole in a background buffer of 25 mM Tris, 100mM NaCl, pH 7.4. Peak fractions were pooled and incubated overnight at 4 $°C$ with ≈1:5 TEV protease (produced in the lab). TEV protease removes the N-terminal His-tag from the protein and leaves a small Ser-Asn sequence N-terminal to the wildtype starting methionine. Next hydrophobic interaction chromatography (HIC) was used to purify the S100s from remaining bacterial proteins and the added TEV protease. Proteins were passed over a 5 mL HiTrap phenyl-sepharose column (GE Health Science). Due to the Ca^{2+}-dependent exposure of a hydrophobic binding, the S100 proteins proteins adhere to the column only in the presence of Ca^{2+}. Proteins were pre-saturated with 2mM Ca^{2+} before loading on the column and eluted with a 30mL gradient from 0 mM to 5 mM EDTA in 25 mM Tris, 100 mM NaCl, pH 7.4.

Peak fractions were pooled and dialyzed against 4 L of 25 mM Tris, 100 mM NaCl, pH 7.4 buffer overnight at 4 $°C$ to remove excess EDTA. The proteins were then passed once more over the 5 mL HiTrap Ni-affinity column (GE Health Science) to remove any uncleaved His-tagged protein. The cleaved protein was collected in

the flow-through. Finally, protein purity was examined by SDS-PAGE. If any trace contaminants appeared to be present we performed anion chromatography with a 5 mL HiTrap DEAE column (GE). Proteins were eluted with a 50 mL gradient from 0-500 mM NaCl in 25 mM Tris, pH 7.4 buffer. Pure proteins were dialyzed overnight against 2L of 25 mM TES (or Tris), 100 mM NaCl, pH 7.4, containing 2 g Chelex-100 resin (BioRad) to remove divalent metals. After the final purification step, the purity of proteins products was assessed by SDS PAGE and MALDI-TOF mass spectrometry to be > 95. Final protein products were flash frozen, dropwise, in liquid nitrogen and stored at $-80\ ^{\circ}C$. Protein yields were typically on the order of 25 mg/1.5 L of culture.

Isothermal Titration Calorimetry

For all peptides, we attempted to measure binding at $25\ ^{\circ}C$. ITC experiments were performed in 25 mM TES, 100mM NaCl, 2 mM $CaCl_2$, 1mM TCEP, pH 7.4. Samples were equilibrated and degassed by centrifugation at $18,000 \times g$ at the experimental temperature for 35 minutes. Synthetic peptides (purchased from GenScript) were dissolved directly into the experimental buffer prior to each experiment. All experiments were performed on a MicroCal ITC-200. Gain settings were determined on a case-by-case basis to ensure quality data. A 750 rpm syringe stir speed was used for all experiments. Spacing between injections ranged from 300s-900s depending on gain settings and relaxation time of the binding process. These setting were optimized for each binding interaction that was measured. A single-site binding model was fit to the titration data using the Bayesian MCMC fitter in pytc

(https://github.com/harmslab/pytc). The ML estimate was used as a starting guess and the likelihood surface was then explored with 100 walkers, each taking 5,000 steps. The first 10% of steps were discarded as burn in. For each protein/peptide combination, one clean ITC trace was used to fit the binding model. Negative results were double-checked to ensure accuracy.

Preparation of biotinylated proteins for phage display

A mutant version of hA5 with a single N-terminal Cys residues were generated via site-directed mutagenesis using the QuikChange lightning system (Agilent). The Cys was introduced in the Ser-Asn tag leftover from TEV protease cleavage as Ser-Asn-Cys. The proteins were expressed and purified as described in the previous section. A small amount of the purified proteins were biotinylated using the EZ-link BMCC-biotin system (ThermoFisher Scientific). ≈1 mg BMCC-biotin was dissolved directly in 100% DMSO to a concentration of 8 mM for labeling. Proteins were exchanged into 25mM phosphate, 100mM NaCl, pH 7.4 using a Nap-25 desalting column (GE Health Science) and degassed for 30 min at 25 $^{\circ}C$ using a vacuum pump (Malvern Instruments). While stirring at room temperature, 8mM BMCC-biotin was added dropwise to a final 10X molar excess. Reaction tubes were sealed with PARAFILM (Bemis) and the maleimide-thiol reactions were allowed to proceed for 1 hour at room temperature with stirring. The reactions were then transferred to 4$^{\circ}C$ and incubated with stirring overnight to allow completion of the reaction. Excess BMCC-biotin was removed from the labeled proteins by exchanging again over a Nap-25 column (GE Health Science), and subsequently a series of 3 concentration-wash steps on

a NanoSep 3K spin column (Pall corporation), into the Ca-TeBST loading loading buffer. Complete labeling was confirmed by MALDI-TOF mass spectrometry by observing the ≈540Da shift in the protein peak. Final stocks of labeled proteins were prepared at 10 μM by dilution into the loading buffer.

Phage display

Phage display experiments were performed using the PhD-12 peptide phage display kit (NEB). All steps involving the pipetting of phage-containing samples was done using filter tips (Rainin). We prepared 100 μL samples containing phage (5.5×10^{11} PFU) and 0.01 μM biotin-protein (or biotin alone in the negative control) at room temperature in a background of Ca^{2+}-TeBST loading buffer (50mM TES, 100mM NaCl, 2mM $CaCl_2$, 0.01% Tween-20, pH 7.4) to ensure Ca^{2+}-saturation of the S100 proteins. For the experiments using a peptide competitor, we included the peptide RSHSGFDWRWAMEALTGGSAE at 20 μM in the loading buffer. This peptide (named A6cons in the original report), binds all four proteins at the canonical binding site with K_D between 1 and 8 μM (18). Samples were incubated at room temperature for 2hr. Each sample was then applied to one well of a 96-well high-capacity streptavidin plate (previously blocked using PhD-12 kit blocking buffer and washed 6X with 150 μL loading buffer). Samples were incubated on the plate with gentle shaking for 20min. 1 μL of 10 mM biotin (NEB) was then added to each sample on the plate and incubated for an additional five minutes to compete away purely biotin-dependent interactions. Samples were then pulled from the plate carefully by pipetting and discarded. Each well was washed 5X with 200 µL of loading buffer

by applying the solution to the well and then immediately pulling off by pipetting. Finally, 100 µL of EDTA-TeBST elution buffer (50mM TES, 100mM NaCl, 5mM EDTA, 0.01% Tween-20, pH 7.4) was applied to each well and the plate was incubated with gentle shaking for 1hr at room temperature to elute. Eluates were pulled from the plate carefully by pipetting and stored at $4°C$. Eluates were titered to quantify eluted phage as follows. Serial dilutions of the eluates from $1:10-1:10^5$ were prepared in LB medium. These were used to inoculate 200 µL aliquots of mid-log-phase ER2738 *E. coli* (NEB) by adding 10 µL to each. Each 200 µL aliquot was then mixed with 3mL of pre-melted top agar, applied to a LB agar XGAL/IPTG (Rx Biosciences) plate, and allowed to cool. The plates were incubated overnight at $37°C$ to allow formation of plaques. The next morning, blue plaques were counted and used to calculate PFU/mL phage concentration. Enrichment was calculated as a ratio of experimental samples to the biotin-only negative control.

To generate the pre-conditioned phage library the naïve library was first screened in duplicate against each of the four proteins as described above. Each of these lineages was subsequently amplified in ER2738 *E. coli* (NEB) as follows. 20mL 1:100 dilutions of an ER2738 overnight culture were prepared. Each 20mL culture was inoculated with one entire sample of remaining phage eluate. The cultures were incubated at $37°C$ with shaking for 4.5 hours to allow phage growth. Bacteria were then removed by centrifugation and the top 80% of the culture was removed carefully with a filtered serological pipette and transferred to a fresh tube containing 1/6 volume of PEG/NaCl (20% w/v PEG-8000, 2.5M NaCl). Samples were incubated overnight at $4°C$ to precipitate phage. Precipitated phage were isolated by centrifu-

gation and subsequently purified by an additional PEG/NaCl precipitation on ice for 1hr. These individually amplified pools were then resuspended in 200 µL each of terile loading buffer and mixed together to form a pre-conditioned library in order to minimize the impact of sampling on the subsequent panning experiment. The pool was diluted 1:1 with 100% glycerol and stored at $-20°C$ for use in the final panning experiments.

Preparation of deep sequencing libraries

Phage genomic ssDNA was isolated from leftover amplified eluates from each round of panning using the M13 spin kit (Qiagen). Products were stored in low TE buffer. These ssDNA were used as the template for 2 replicate PCRs with the Cs1 forward (5'—ACACTGACGACATGGTTCTACAGTGGTACCTTTCTATTCTCACTCT—3') and PhD96seq-Cs2 reverse (5'—TACGGTAGCAGAGACTTGGTCTCCCTCATAGT-TAGCGTAACG—3') primers. Products were isolated from these PCR products using the GeneJet gel extraction kit (Thermo Scientific) and pooled. The pooled products were then used as templates for a secondary reaction with the barcoded primers. Products were isolated from these final PCRs using the GeneJet gel extraction kit. Concentration of barcoded samples was measured by A_{260}/A_{280} using a 1mm cuvette on an Eppendorf biospectrometer. Multiplexing was done by mixing samples according to mass. The concentration of the multiplexed library was corrected using qPCR with the P5 and P7 Illumina flow-cell primers. The library was then diluted to a final concentration of 10nM and Illumina sequenced on two lanes of a HiSeq 4000 instrument, using the Cs1 F' as the R1 sequencing primer. The lanes

were spiked with 20% PhiX control DNA due to the relatively low diversity of the library.

Phage display analysis pipeline

We performed quality control on three read features. First, we verified that the sequence had exactly the anticipated length from the start of the phage sequence through the stop codon. Second, we only took sequences in which the invariant phage sequence differed by at most one base from the anticipated sequence. This allows for a single point mutation and or sequencing errors, but not wholesale changes in the sequence. Finally, we took only reads with an average phred score better than 15. The vast majority of the reads that failed our quality control did not have the variable region, representing reversion to phage with a wildtype-like coat protein. This analysis is encoded in the *hops_count.py* script (https://github.com/harmslab/hops), which takes a gzipped fastq file as input and returns the counts for every peptide in the file.

Identifying the read count cutoff

One critical question is at what point the number of reads correlates with the frequency of a peptide. If we set the cutoff too low, we incorporate noise into downstream analyses. If we set the cutoff too high, we remove valuable observations from our dataset. To identify an appropriate cutoff, we studied the mapping between c_i (the number of reads arising from peptide i) and f_i (the actual frequency of peptide i in the experiments). Our goal was to find $P(f_i|c_i, N)$: the probability peptide i is

at f_i given we observe it c_i times in N counts. Using Bayes theorem, we can write

$$P(f_i|c_i, N) = \frac{P(c_i|f_i, N)P(f_i)}{P(c_i)},$$

where N is the total number of reads. We calculated $P(c_i|f_i, N)$ assuming a binomial sampling process: what is the probability of observing exactly c counts given N independent samples when a population with a peptide frequency f_i? This gives the curve seen in Fig S2A. We then estimated $P(\hat{f}_i)$ from the distribution of frequencies in the input library, constructing a histogram of apparent peptide frequencies (Fig S2B). Empirically, we found that frequencies followed an exponential distribution over the measurable range of frequencies. Finally, we assumed that all counts have equal prior probabilities, turning $P(c_i)$ into a scalar that normalizes the integral of $P(f_i|c_i, N)$ so it sums to 1.

Using the information from Fig S2A and B, we could then calculate $P(f_i|c_i, N)$ for any number of reads in an experiment N. Fig S2C shows this calculation for $N = 2.0 \times 10^7$ reads—a typical number of reads from our experimental replicates. This curve is linear above 6 reads. Below this, counts no longer correlates linearly with frequency, as it is possible to obtain 5 reads random sampling from low frequency library members. We therefore used a cutoff of 6 counts for all downstream analyses. In total, 74.0% of reads passed our quality control and read cutoff (Table S1).

Measuring enrichment values

We next set out to measure changes in the frequency of peptides between the competitor and non-competitor samples. The simplest way to do this would be to iden-

tify peptides seen in both experiments, and then measure how their frequencies change between conditions. Unfortunately, these proteins all bind a wide swath of peptide targets and relatively few peptides were shared between conditions. This approach would thus exclude the majority of sequences. For example, only 8,672 of the 112,681 unique peptides observed for hA5 were present in both the competitor and non-competitor, even after pooling biological replicates. Worse, because we are interested in peptides that are lost when competitor peptide is added, ignoring peptides with no counts in the competitor sample means ignoring some of the most informative peptides.

To solve this problem, we clustered similar peptides and measured enrichment for peptide clusters rather than individual peptides. We extracted all peptides that were observed across the competitor and non-competitor samples for a given protein. We then used DBSCAN to cluster those peptides according to sequence similarity, as measured by their their Damerau-Levenshtein distance (43, 44). This revealed extensive structure in our data. For example, hA5 yielded 8,645 clusters with more than one peptide, incorporating more than half of the unique peptides (Fig 3A, Fig S3A). We chose clustering parameters that led to highly similar peptides within each cluster, as can be seen by the representative sequence logos for three clusters of hA5 (Fig S3B). Sequences that were not placed in clusters were treated as clusters with a size of one.

We then used the enrichment of each cluster to estimate the enrichment of indi-

vidual peptides. We defined enrichment as:

$$E_{cluster} = -ln\left(\frac{\sum_{i=1}^{i\leq N}\beta_i}{\sum_{i=1}^{i\leq N}\alpha_i}\right), \quad (1)$$

where N is the total number of peptides in the cluster, β_i is the frequency of peptide i in the competitor sample, and α_i is the frequency of peptide i in the non-competitor sample. We then made the approximation that all members of the cluster have the same enrichment:

$$E_i \approx E_{cluster}, \quad (2)$$

allowing us to estimate the enrichment of all i peptides in the cluster (Fig S3C). Peptides lost because of competition for the interface will add zeros to the numerator of Eq. 1, leading to an overall decrease in enrichment. Peptides missed because of finite sampling will add zeros evenly to the competitor and non-competitor samples, leading to no net enrichment.

We tested this cluster-based approximation using the 8,672 peptides of hA5 for which we could directly calculate enrichment (that is, those peptides seen in both the competitor and non-competitor experiments). We calculated the enrichment of each peptide individually and compared these values to those obtained by the cluster method. There is no systematic difference in the values estimated using the two methods, and the linear model explains 98.4% of the variation between the two methods.

We clustered peptides using our own implementation of the DBSCAN algorithm (44) using the Damerau-Levensthein distance (43). The main parameter for DB-

SCAN clustering is ε—the neighborhood cutoff. Clusters are defined as sequences that can be reached through a series of ε-step moves. We found that $\varepsilon = 1$ gave the best results for our downstream machine learning analysis. Our whole enrichment pipeline—including clustering—can be run given a peptide count file for the non-competitor experiment and a peptide-count file for the competitor experiment using the *hops_ enrich.py* script (https://github.com/harmslab/hops).

Principle Component Analysis

We implemented our machine learning model in Python 3 extended with numpy (45), scipy (46), and matplotlib (47). We used sklearn for our random forest regression (23, 48, 49). A full list of the calculated features is shown in Table S2. As noted, some features were calculated using CIDER (22). Our full implementation, including all data files, is available at https://github.com/harmslab/hops.

To generate the aaindex meta features, we performed a principle component analysis on all 590 features from the aaindex database. Any missing value was assigned the mean value of that feature. Prior to performing the PCA, we standardized all values to a mean of zero and a standard deviation of 1. This yielded 20 principle components.

Incorporating uncertainty into an estimate of a Venn diagram

We used a Bayesian approach to estimate the overlaps between the binding sets of proteins, despite high false positive and false negative rates. Consider a set of peptides binding to the proteins A and B. The binding of these peptides can be

described by a Venn diagram with four regions: $[A \cup B]^c$ (peptides that bind neither A nor B), $A \backslash B$ (peptides that bind A alone), $B \backslash A$ (peptides that bind B alone), and $A \cap B$ (peptides that bind both A and B). The number of peptides in each region is given by $\vec{V}$, while the number of peptides observed in each region is given by $\vec{V}_{obs}$. $\vec{V}$ and $\vec{V}_{obs}$ can differ as there may be both false positives (at rates m_A and m_B) and false negatives (at rates n_A and n_B). We can write a row-stochastic matrix that describes the probability of observing a peptide in a region given its actual region as:

$$\mathbf{T} = \begin{bmatrix} P([A \cup B]^c | [A \cup B]^c) & P(A \backslash B | [A \cup B]^c) & P(B \backslash A | [A \cup B]^c) & P(A \cap B | [A \cup B]^c) \\ P([A \cup B]^c | A \backslash B) & P(A \backslash B | A \backslash B) & P(B \backslash A | A \backslash B) & P(A \cap B | A \backslash B) \\ P([A \cup B]^c | B \backslash A) & P(A \backslash B | B \backslash A) & P(B \backslash A | B \backslash A) & P(A \cap B | B \backslash A) \\ P([A \cup B]^c | A \cap B) & P(A \backslash B | A \cap B) & P(B \backslash A | A \cap B) & P(A \cap B | A \cap B) \end{bmatrix}$$

where each conditional probability $P(X|Y)$ describes the probability of observing the peptide in region X given it is actually in region Y. If we know this matrix and we know the real population in each region, we can calculate $\vec{V}_{obs}$ by:

$$\vec{V}_{obs} = \vec{V} \cdot \mathbf{T}.$$

We can construct $\mathbf{T}$ using the false positive and false negative rates for binding to protein A or B. For example, the probability of seeing a peptide that binds to A

alone when it actually does not bind to either A or B would be

$$P(A \backslash B | [A \cup B]^c) = m_A - m_A m_B :$$

the probability of a false positive for A less the probability of a false positive for both A and B. Using appropriate combinations of false positive and false negative rates, we can calculate every value in $\mathbf{T}$:

$$\mathbf{T} = \begin{bmatrix} 1-(m_A+m_B-m_Am_B) & m_A-m_Am_B & m_B-m_Am_B & m_Am_B \\ n_A-n_Am_B & 1-(n_A+m_B-n_Am_B) & n_Am_B & m_B-n_Am_B \\ n_B-m_An_B & m_An_B & 1-(m_A+n_B-m_An_B) & m_A-m_An_B \\ n_An_B & n_B-n_An_B & n_A-n_An_B & 1-(n_A+n_B-n_An_B) \end{bmatrix}.$$

This can be readily extended to any number of proteins with any number of possible overlaps.

We can then estimate $\vec{V}$ using Bayesian Markov Chain Monte Carlo (MCCE). We first write a likelihood function:

$$ln\left[P(\vec{V}_{obs}|\vec{V},\{m\},\{n\})\right] = -\frac{1}{2}\sum_i \left[(\vec{V}_{obs,i}-\vec{V}_i\mathbf{T})^2/\sigma_i^2 + ln(\sigma_i^2)\right]$$

where i indexes regions in the Venn diagram, σ_i^2 is the uncertainty of the counts in region i, $\{m\}$ is the set of false positive rates and $\{n\}$ is the set of false negative rates. We can then sample values in $\vec{V}$, $\{m\}$ and $\{n\}$ by MCCE. For $\vec{V}$, we used the prior:

$$ln\left[P(\vec{V})\right] = \begin{cases} -\infty & \vec{V} < 0 \\ 0 & \vec{V} \geq 0 \end{cases},$$

thus requiring all regions to have positive counts. We also constrained the number of counts in $\vec{V}$ be within 5% of the number of counts in $\vec{V}_{obs}$ (N):

$$ln[P(\vec{V})] = \begin{cases} -\infty & \sum \vec{V} < 0.95N \\ 0 & 0.95N \leq \sum \vec{V} \leq 1.05N \\ -\infty & \sum \vec{V} > 1.05N \end{cases}$$

For every false positive or false negative rate (denoted as r_j), we used the prior:

$$ln\left[P(r_j)\right] = \begin{cases} -\infty & r_j < 0 \\ -\frac{(r_j - \hat{\mu}_j)^2}{2\sigma_j^2} + \sqrt{2\pi\sigma_j^2} & 0 \leq r_j \leq 1, \\ -\infty & r_j > 1 \end{cases}$$

where $\hat{\mu}_j$ is the estimate of the value of r_j from our binding experiments and σ_j was set to 0.2. For values outside of 0 and 1, the log prior is $-\infty$, enforcing bounds on these parameters.

References

[1] Carroll SM, Bridgham JT, Thornton JW. Evolution of Hormone Signaling in Elasmobranchs by Exploitation of Promiscuous Receptors. Molecular Biology and Evolution. 2008 Dec;25(12):2643–2652. Available from: https://academic.oup.com/mbe/article/25/12/2643/1110104/Evolution-of-Hormone-Signaling-in-Elasmobranchs-by.

[2] Khersonsky O, Tawfik DS. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. Annual Review of Biochemistry. 2010;79(1):471–505. Available from: http://dx.doi.org/10.1146/annurev-biochem-030409-143718.

[3] Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. Nature Reviews Genetics. 2010 Aug;11(8):572–582. Available from: http://www.nature.com/nrg/journal/v11/n8/full/nrg2808.html.

[4] Kanzaki H, Yoshida K, Saitoh H, Fujisaki K, Hirabuchi A, Alaux L, et al. Arms race co-evolution of Magnaporthe oryzae AVR-Pik and rice Pik genes driven by their physical interactions. The Plant Journal. 2012 Dec;72(6):894–907. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2012.05110.x/abstract.

[5] Reinke AW, Baek J, Ashenberg O, Keating AE. Networks of bZIP Protein-Protein Interactions Diversified Over a Billion Years of Evolution. Science. 2013 May;340(6133):730–734. Available from: http://science.sciencemag.org/content/340/6133/730.

[6] Kaltenbach M, Tokuriki N. Dynamics and constraints of enzyme evolution. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 2014 Nov;322(7):468–487. Available from: http://onlinelibrary.wiley.com/doi/10.1002/jez.b.22562/abstract.

[7] Clifton B, Jackson C. Ancestral Protein Reconstruction Yields Insights into Adaptive Evolution of Binding Specificity in Solute-Binding Proteins. Cell Chemical Biology. 2016 Feb;23(2):236–245. Available from: http://www.sciencedirect.com/science/article/pii/S2451945616000313.

[8] Alhindi T, Zhang Z, Ruelens P, Coenen H, Degroote H, Iraci N, et al. Protein interaction evolution from promiscuity to specificity with reduced flexibility in an increasingly complex network. Scientific Reports. 2017 Mar;7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5364480/.

[9] Santamaria-Kisiel L, Rintala-Dempsey AC, Shaw GS. Calcium-dependent and -independent interactions of the S100 protein family. Biochemical Journal. 2006 Jun;396(2):201–214. Available from: http://www.biochemj.org/content/396/2/201.

[10] Lee YT, Dimitrova YN, Schneider G, Ridenour WB, Bhattacharya S, Soss SE, et al. Structure of the S100A6 Complex with a Fragment from the C-Terminal Domain of Siah-1 Interacting Protein: A Novel Mode for S100 Protein Target Recognition. Biochemistry. 2008 Oct;47(41):10921–10932.

[11] Bertini I, Gupta SD, Hu X, Karavelas T, Luchinat C, Parigi G, et al. Solution

Structure and Dynamics of S100A5 in the Apo and Ca2+-Bound States. JBIC Journal of Biological Inorganic Chemistry. 2009 Sep;14(7):1097–1107.

[12] Leclerc E, Fritz G, Vetter SW, Heizmann CW. Binding of S100 Proteins to RAGE: An Update. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2009 Jun;1793(6):993–1007.

[13] Leśniak W, S\lomnicki \P, Filipek A. S100A6 – New Facts and Features. Biochemical and Biophysical Research Communications. 2009 Dec;390(4):1087–1092.

[14] Streicher WW, Lopez MM, Makhatadze GI. Annexin I and Annexin II N-Terminal Peptides Binding to S100 Protein Family Members: Specificity and Thermodynamic Characterization. Biochemistry. 2009 Mar;48(12):2788–2798.

[15] S\lomnicki \P, Nawrot B, Leśniak W. S100A6 Binds P53 and Affects Its Activity. The International Journal of Biochemistry & Cell Biology. 2009 Apr;41(4):784–790.

[16] Liriano MA. Structure, Dynamics and Function of S100B and S100A5 Complexes [Ph.D.]. University of Maryland, Baltimore. United States – Maryland; 2012.

[17] Donato R, R Cannon B, Sorci G, Riuzzi F, Hsu K, J Weber D, et al. Functions of S100 Proteins. Current molecular medicine. 2013;13(1):24–57.

[18] Wheeler LC, Anderson JA, Morrison AJ, Wong CE, Harms MJ. Conservation of

specificity in two low-specificity proteins. bioRxiv. 2017 Oct;p. 207324. Available from: https://www.biorxiv.org/content/early/2017/10/25/207324.

[19] Sidhu SS, Lowman HB, Cunningham BC, Wells JA. Phage Display for Selection of Novel Binding Peptides. Methods in Enzymology. 2000 Jan;328:333–IN5.

[20] Willats WGT. Phage Display: Practicalities and Prospects. Plant Molecular Biology. 2002 Dec;50(6):837–854.

[21] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: Amino Acid Index Database, Progress Report 2008. Nucleic Acids Research. 2008;36(Database issue):D202–205.

[22] Holehouse AS, Ahad J, Das RK, Pappu RV. CIDER: Classification of Intrinsically Disordered Ensemble Regions. Biophysical Journal. 2015 Jan;108(2):228a.

[23] Breiman L. Random Forests. Machine learning. 2001;45(1):5–32.

[24] Eick GN, Bridgham JT, Anderson DP, Harms MJ, Thornton JW. Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty. Molecular Biology and Evolution. 2017 Feb;34(2):247–261.

[25] McKeown A, Bridgham J, Anderson D, Murphy M, Ortlund E, Thornton J. Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. Cell. 2014 Sep;159(1):58–68. Available from: http://www.sciencedirect.com/science/article/pii/S0092867414011143.

[26] Knott TK, Madany PA, Faden AA, Xu M, Strotmann J, Henion TR, et al. Olfactory Discrimination Largely Persists in Mice with Defects in Odorant Receptor Expression and Axon Guidance. Neural development. 2012;7(1):17.

[27] McIntyre JC, Davis EE, Joiner A, Williams CL, Tsai IC, Jenkins PM, et al. Gene Therapy Rescues Cilia Defects and Restores Olfactory Function in a Mammalian Ciliopathy Model. Nature medicine. 2012;18(9):1423–1428.

[28] Olender T, Keydar I, Pinto JM, Tatarskyy P, Alkelai A, Chien MS, et al. The Human Olfactory Transcriptome. BMC genomics. 2016 Aug;17(1):619.

[29] Eick GN, Colucci JK, Harms MJ, Ortlund EA, Thornton JW. Evolution of Minimal Specificity and Promiscuity in Steroid Hormone Receptors. PLoS Genetics. 2012 Nov;8(11). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3499368/.

[30] Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM. Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian beta-Lactamases. Journal of the American Chemical Society. 2013 Feb;135(8):2899–2902. Available from: http://dx.doi.org/10.1021/ja311630a.

[31] Pougach K, Voet A, Kondrashov FA, Voordeckers K, Christiaens JF, Baying B, et al. Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. Nature Communications. 2014 Sep;5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4172970/.

[32] Risso VA, Gavira JA, Sanchez-Ruiz JM. Thermostable and promiscuous Precambrian proteins. Environmental Microbiology. 2014 Jun;16(6):1485–1489. Available from: http://onlinelibrary.wiley.com/doi/10.1111/1462-2920.12319/abstract.

[33] Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. Molecular Biology and Evolution. 2015 Jan;32(1):132–143. Available from: https://academic.oup.com/mbe/article/32/1/132/2925568/Evolution-of-Conformational-Dynamics-Determines.

[34] Devamani T, Rauwerdink AM, Lun-zer M, Jones BJ, Mooney JL, Tan MAO, et al. Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. Journal of the American Chemical Society. 2016 Jan;138(3):1046–1056. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5466365/.

[35] Ma S, Martin-Laffon J, Mininno M, Gigarel O, Brugière S, Bastien O, et al. Molecular Evolution of the Substrate Specificity of Chloroplastic Aldolases/Rubisco Lysine Methyltransferases in Plants. Molecular Plant. 2016 Apr;9(4):569–581.

[36] Rauwerdink A, Lunzer M, Devamani T, Jones B, Mooney J, Zhang ZJ, et al. Evolution of a Catalytic Mechanism. Molecular Biology and Evolution. 2016 Apr;33(4):971–979.

[37] Jensen RA. Enzyme Recruitment in Evolution of New Function. Annual Review

of Microbiology. 1976;30(1):409–425. Available from: http://dx.doi.org/10.1146/annurev.mi.30.100176.002205.

[38] Copley SD. Toward a Systems Biology Perspective on Enzyme Evolution. The Journal of Biological Chemistry. 2012 Jan;287(1):3–10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3249082/.

[39] Wheeler LC, Lim SA, Marqusee S, Harms MJ. The thermostability and specificity of ancient proteins. Current Opinion in Structural Biology. 2016 Jun;38:37–43. Available from: http://www.sciencedirect.com/science/article/pii/S0959440X16300501.

[40] Gaucher EA, Govindarajan S, Ganesh OK. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature. 2008 Feb;451(7179):704–707. Available from: http://www.nature.com/nature/journal/v451/n7179/abs/nature06510.html?foxtrotcallback=true.

[41] Mannige RV, Brooks CL, Shakhnovich EI. A Universal Trend among Proteomes Indicates an Oily Last Common Ancestor. PLOS Computational Biology. 2012 Dec;8(12):e1002839. Available from: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002839.

[42] Williams PD, Pollock DD, Blackburne BP, Goldstein RA. Assessing the Accuracy of Ancestral Protein Reconstruction Methods. PLOS Computational Biology. 2006 Jun;2(6):e69. Available from: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0020069.

[43] Damerau FJ. A Technique for Computer Detection and Correction of Spelling Errors. Commun ACM. 1964 Mar;7(3):171–176. Available from: http://doi.acm.org/10.1145/363958.363994.

[44] Ester M, Kriegel HP, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. AAAI Press; 1996. p. 226–231.

[45] Walt Svd, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science Engineering. 2011 Mar;13(2):22–30.

[46] Jones E, Oliphant T, Peterson P, others. SciPy: Open source scientific tools for Python; 2001. Available from: http://www.scipy.org/.

[47] Hunter JD. Matplotlib: A 2D graphics environment. Computing In Science & Engineering. 2007;9(3):90–95.

[48] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. CRC press; 1984.

[49] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

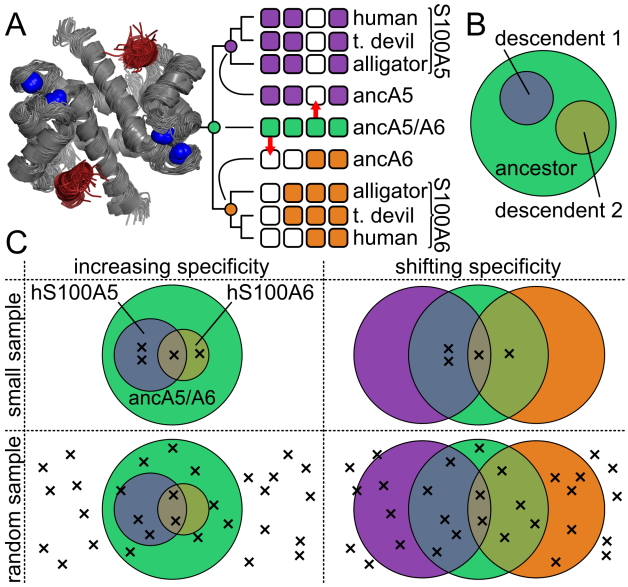

Fig 1. Testing the increased specificity hypothesis requires extensive sampling of targets. A) Venn diagram of the increasing-specificity hypothesis. The large circle is set of targets recognized by the ancestor; the smaller circles are sets of targets represented its descendants. There is no strict requirement that descendants be subsets of the ancestor. B) Experimentally measured changes in peptide binding specificity for S100A5 and S100A6 (taken from (18)). Structure: location of peptide (red) binding to a model of S100A5 (gray, PDB: 2KAY). Bound Ca^{2+} are shown as blue spheres. Phylogeny: Boxes represent binding of four different peptides (arranged left to right) to nine different proteins (arranged top to bottom). A white box indicates the peptide does not bind that protein; a colored box indicates the peptide binds. Colors denote ancA5/A6 (green), S100A5 (purple), and S100A6 (orange). Red arrows highlight ancestral peptides lost in the modern proteins. C) Venn diagrams show overlap in peptide binding sets between ancA5/A6, S100A5, and S100A6. Crosses denote experimental observations. Columns show two evolutionary

scenarios: increasing specificity (left) versus shifting specificity (right). Rows show to different sampling methods: small sample (top) versus random sampling (bottom). Colors are as in panel B.

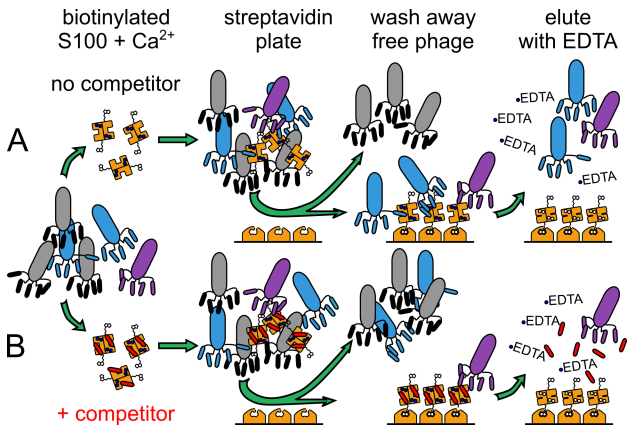

Fig 2. Set of binding peptides can be estimated using phage display. Rows show two different experiments, done in parallel, for each protein. Biotinylated, Ca^{2+}-loaded, S100 is added to a population of phage either alone (row A) or with saturating competitor peptide added in trans (row B). Phage that bind to the protein (blue or purple) are pulled down using a streptavidin plate. Bound phage are then eluted using EDTA, which disrupts the peptide binding interface. In the absence of competitor (row A), phage bind adventitiously (purple) as well as at the interface of interest (blue). In the presence of competitor (row B), only adventitious binders are present.

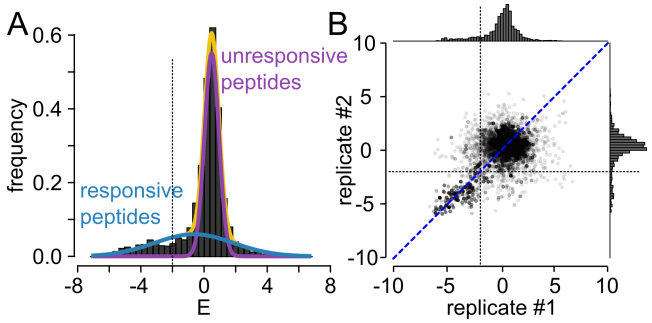

Fig 3. A subpopulation of the phage respond to the addition of competitor peptide. A) Distribution of enrichment values for peptides taken from pooled biological replicates of hA5. The measured distribution (gray) can be fit by the sum of two Gaussian distributions: responsive (blue) and unresponsive (purple), which sum to the total (yellow). B) Enrichment values from biological replicates are strongly correlated. Axes are enrichment for replicate #1 or replicate #2. Points are individual peptides. Distributions for each replicate are shown on the top and right, respectively. The red dashed line is the best fit line (orthoganol distance regression), explaining ∼81% of the variation in the data.

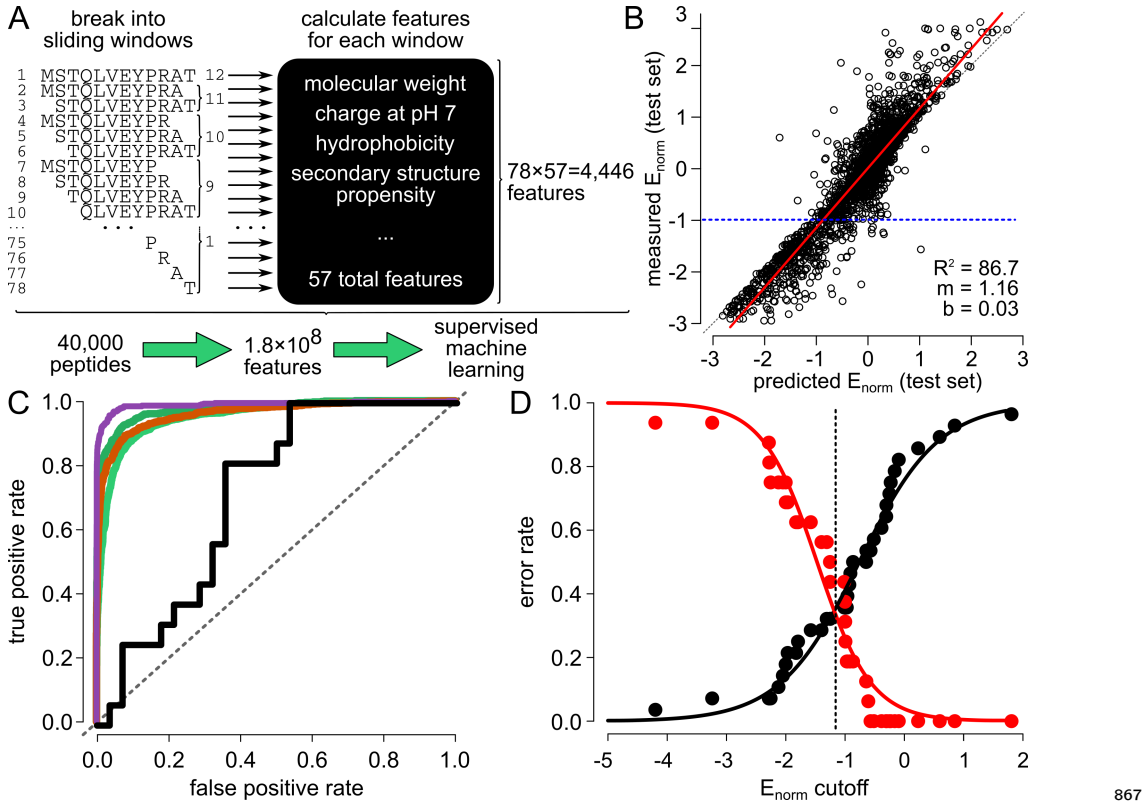

Fig 4. Peptide binding can be predicted from amino acid sequence. A) Schematic showing our strategy for training a binding model. We break the 12-mer peptide into 78 different sliding windows. For each peptide, we calculate 57 features (black box), giving a total of 4,446 features per peptide. We then use 40,000 peptides to train a model predicting E (green arrows). B) Correlation between predicted E_{norm} and measured E_{norm} for $\sim 4,000$ peptides in test set for hA5. Each point is a peptide. Red line is least squares regression line. Blue dashed line is our classification line (see panel C). C) Receiver Operator Characteristic (ROC) curves for binding models. Colored series show ability of models to classify measured E_{norm} as ≤ -1 (the blue dashed line form panel B). Curves are hA5 (purple), hA6 (orange),

ancA5/A6 (dark green), and altAll (light green). Black line is the ROC curve for predicting the binding of 44 isolated peptides. D) Error rates for predicting isolated peptides that bind as function of E_{norm} cutoff for the classifier. False negative rate (red) and false positive rate (red) cross at $E_{norm} = -1.19$ (dashed line) with a value of ≈ 0.35. Solid lines are fits of the modified Hill equation to the to error rates.

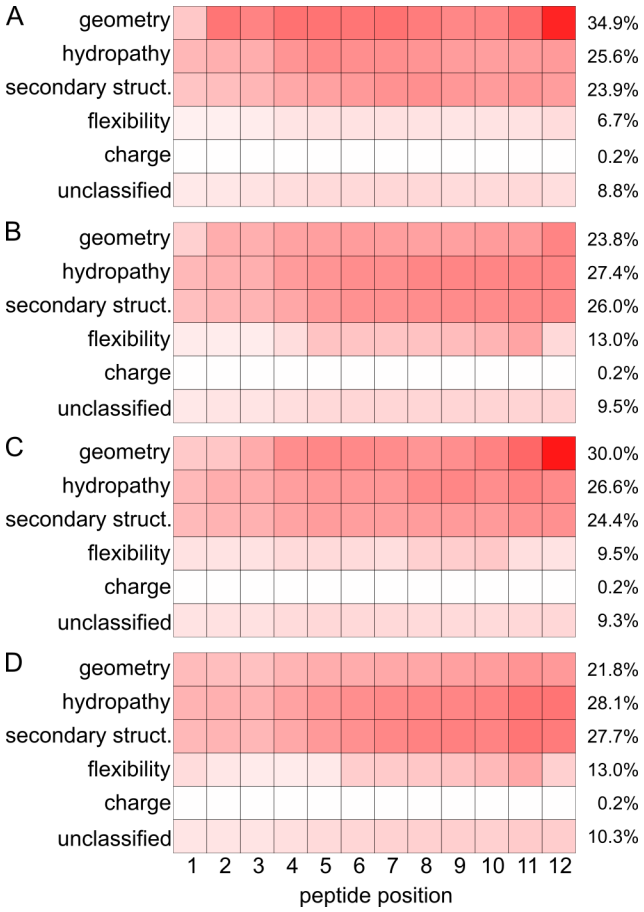

Fig 5. Machine learning model reveals consistent determinants of specificity. Squares denote the contribution of each peptide position (left-to-right) and different chemical features (top-to-bottom). Color indicates relative contribution from red (strong) to white (no contribution). Marginal contribution of each chemical feature is shown to the right of each plot. Panels correspond to hA5 (A), hA6 (B), ancA5/A6 (C), and altAll (D).

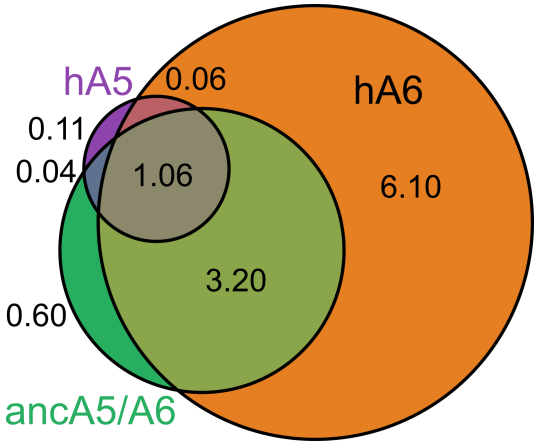

Fig 6. hA5 and hA6 exhibit divergent changes in specificity. Circles denote estimated binding sets for hA5 (purple), hA6 (orange), and ancA5/A6 (green). Areas and numbers in each region indicate the percent of all peptides that are within that region of the Venn diagram. 88.8% of peptides are not predicted to bind to any of the proteins.

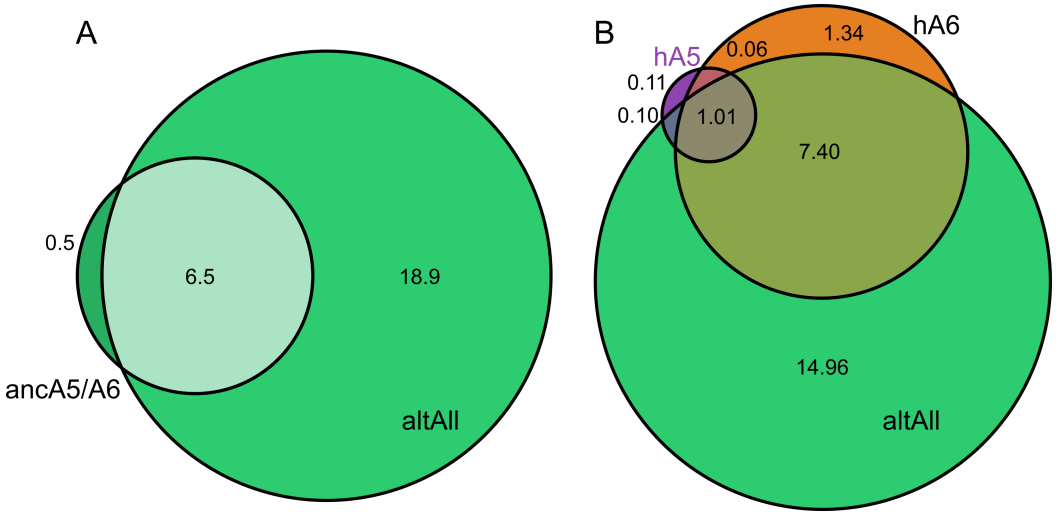

Fig 7. The altAll construct binds more peptides than the ML reconstruction. A) Circles denote binding sets for ancA5/A6 (small circle) and altAll (larger green). Areas and numbers in each region indicate the percent of random peptides in that region of the Venn diagram. B) Overlap between altAll, hA5 and hA6.

Table 1: Protein binding model statistics

protein	num. training observations	R^2_{train}	R^2_{test}	AUC	FPR	FNR
hA5	40,887	97.6	85.1	98.9	0.35	0.35
hA6	42,156	97.4	82.9	96.1	0.41	0.41
ancA5/A6	43,938	97.7	84.2	97.4	0.35	0.35
altAll	51,903	96.6	80.0	95.1	0.45	0.15