Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evolutionary paths that link orthogonal pairs of binding proteins

Ziv Avizemer
Weizmann Institute of Science
Carlos Martí Gómez
Cold Spring Harbor Laboratory https://orcid.org/0000-0002-2042-843X
Shlomo Yakir Hoch
Weizmann Institute of Science
David M. McCandlish
Cold Spring Harbor Laboratory
Sarel J. Fleishman (sarel@weizmann.ac.il)
Weizmann Institute of Science

Research Article

Keywords: Epistasis, evolution, protein design, Rosetta, protein-protein interactions, orthogonality, fitness landscape

Posted Date: December 13th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2836905/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Evolutionary paths that link orthogonal pairs of binding proteins

Ziv Avizemer[1], Carlos Martí-Gómez[2], Shlomo Yakir Hoch[1], David M. McCandlish[2] and Sarel J. Fleishman[1]*

1 Department of Biomolecular Sciences, Weizmann Institute of Science, 7610001, Rehovot, Israel
2 Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
*Corresponding author: sarel@weizmann.ac.il

Summary

Some protein binding pairs exhibit extreme specificities that functionally insulate them from homologs. Such pairs evolve mostly by accumulating single-point mutations, and mutants are selected if their affinity exceeds the threshold required for function[1–4]. Thus, homologous and high-specificity binding pairs bring to light an evolutionary conundrum: how does a new specificity evolve while maintaining the required affinity in each intermediate[5,6]? Until now, a fully functional single-mutation path that connects two orthogonal pairs has only been described where the pairs were mutationally close thus enabling experimental enumeration of all intermediates[2]. We present an atomistic and graph-theoretical framework for discovering low molecular strain single-mutation paths that connect two extant pairs, enabling enumeration beyond experimental capability. We apply it to two orthogonal bacterial colicin endonuclease-immunity pairs separated by 17 interface mutations[7]. We were not able to find a strain-free and functional path in the sequence space defined by the two extant pairs. But including mutations that bridge amino acids that cannot be exchanged through single-nucleotide mutations led us to a strain-free 19-mutation trajectory that is completely viable *in vivo*. Our experiments show that the specificity switch is remarkably abrupt, resulting from only one radical mutation on each partner. Furthermore, each of the critical specificity-switch mutations increases fitness, demonstrating that functional divergence could be driven by positive Darwinian selection. These results reveal how even radical functional changes in an epistatic fitness landscape may evolve.

Main

Cellular signaling and metabolism rely on high interaction specificity[8]. Achieving high specificity may appear particularly challenging among pairs of proteins that exhibit sequence and structural similarity. Yet, such homologous pairs are, in fact, often functionally isolated, thereby establishing multiple orthogonal interactions[7,9,10]. The challenge to understanding how orthogonality is achieved among homologous pairs is exacerbated when considering that the evolution of a functionally orthogonal pair requires changes that destabilize interactions between the derived pair

and its ancestor. To avoid being purged by evolutionary selection, however, the changes that destabilize the interaction with the ancestral partners must accumulate in an order that maintains binding between the two members of the derived pair following every change[11]. How the proteins in a pair coevolve to exhibit high affinity in each intermediate while disabling their interactions with the ancestral proteins remains a key question in understanding diversification of gene families and protein interaction networks[12–14].

Some commonly used approaches to address questions on the evolution of binding pairs use ancestral-sequence reconstruction, laboratory evolution, and deep mutational scanning[2,3,12,15–21]. These strategies provided important insights on some of the constraints that shape evolutionary trajectories that lead to new or improved activities[12,22,23]. To date, however, reconstruction of a single-point mutation path among orthogonal pairs has been challenging and accomplished only when the number of mutations between the pairs was small enough to allow complete experimental enumeration of all intermediate mutants[2]. Thus, understanding how multiple incompatible interactions are exchanged between binding pairs is elusive, and fundamental questions in molecular evolution of orthogonal pairs have remained understudied[11,24,25]; among them: Do mutations on one side of the interface enable mutations on the other? Do evolutionary intermediates present dual-specificities or are the intermediates specific as observed in the extant pairs? Do specificity switches occur gradually through the accumulation of many mutations or abruptly through a handful? And how, despite the need to accumulate destabilizing mutations, are inactive intermediates avoided?

We focused on bacterial colicin endonuclease/immunity pairs[7]. The colicins are SOS-induced nonspecific DNA endonuclease toxins (ColE) that coexpress with cognate high-affinity immunity proteins (Im) that protect the producing cell from degradation of its genomic DNA. Following secretion of this complex from the producing cell, the ColE receptor-binding domain attaches to the outer-membrane target receptor of a neighboring cell. Next, the cytotoxic endonuclease domain (E) is unfolded, dissociating from the protective Im protein, and is then translocated to the host cytoplasm. In the host cytoplasm, the E refolds and degrades DNA, killing the cell unless the host harbors a cognate and protective Im[7,26]. Cognate E/Im pairs exhibit ultrahigh binding affinity ($K_D \leq 10^{-14}$ M) and Im proteins that bind E at $K_D > 10^{-10}$ M are not protective[27]. During evolution, the colicins have diversified to four homologous pairs that exhibit ultrahigh specificities of 5-10 orders of magnitude with respect to one another[6,27,28]. Thus, the four Im proteins protect the producing cell against the toxicity of their cognate E proteins but provide no protection against the homologous noncognate E proteins. This system thereby presents a clear example of functional orthogonality under strong evolutionary selection. Due to their remarkably high affinities and specificities, the colicin E/Im pairs have served as model systems to study evolutionary dynamics[3,29] and biomolecular recognition in protein engineering and design[6,30–35]. Insights from these studies and the straightforward approach to test the evolutionary fitness of pairs *in vivo* within their native host attracted us to use this system to study the coevolution of orthogonal binding pairs. Because only four E/Im pairs are known[7,36], however, ancestral sequences cannot be reliably

reconstructed as in other systems[4,13,21,22]. Here, we develop an atomistic strategy that models the stability of each possible intermediate in the complete combinatorial binding landscape and searches for a single-mutation path between the extant pairs while minimizing molecular strain that may arise from protein instability or incompatible interactions between the partners. This framework allows testing different mechanistic hypotheses on the sequence and physical constraints that shape the fitness landscape.

Results

Orthogonality and structural incompatibilities

We used the two most closely related E/Im pairs, known as pairs 2 and 9, as the subjects of our study. In these pairs, the interaction hotspot region, comprising E:Phe86 and Im:Tyr54-Tyr55 (**Fig. 1A**, spheres), is responsible for a large portion of the binding energy and is conserved in sequence and conformation (amino acid numbering according to E2/Im2[34]). Conversely, other interfacial positions contribute less to affinity but determine binding specificity[27,35] and are therefore the focus of our analysis. Some of these positions exhibit dramatic changes in their physicochemical properties. For instance, in E9/Im9, interfacial positions surrounding the hotspot include Val98, Tyr83 (E9) and Leu33 and Val34 (Im9) compared respectively to Arg, Lys, Asp, and Asn in E2/Im2 (**Fig. 1A**). The overall effect of these four radical mutations is a striking change from a polar and charged interface in E2/Im2 that is mediated by a buried interfacial electrostatic interaction between E:Arg98 and Im:Asp33 to a largely hydrophobic interaction in E9/Im9 (**Fig. 1A & B**). Additionally, the Im protein backbone differs in two unstructured regions connecting helices 1 and 2 and helices 3 and 4 (**Fig. 1C**), both containing critical specificity-determining positions[27,34]. Thus, despite the high sequence and structure similarity between the two pairs, their binding interfaces feature mutually incompatible networks of molecular interactions (hydrophobic versus polar/charged) that underlie the ultrahigh specificity barrier between them.

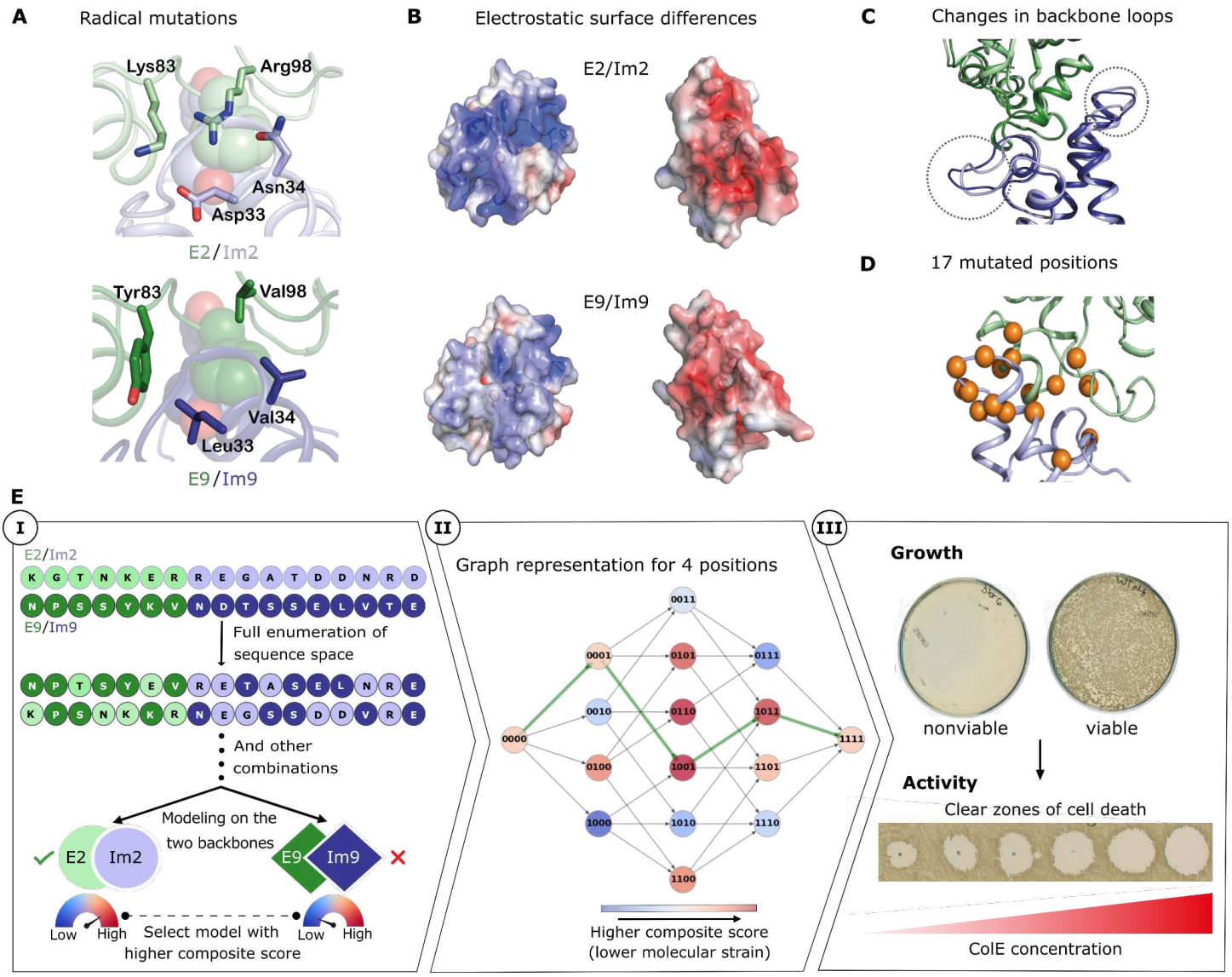


Figure 1. The structural basis for orthogonal binding specificities in colicin E2/Im2 and E9/Im9 and an energy-based strategy to discover evolutionarily plausible paths. (A) Four radical mutations between E2/Im2 and E9/Im9 (PDB entries: 3u43 and 1emv, respectively). The conserved hotspot residues, E:Phe86 and Im:Tyr54-Tyr55, are shown in spheres in all molecular representations. (B) Open-book representation of the two complexes with solvent-accessible surfaces colored according to electrostatic representation. (C) Aligned structures of E2/Im2 and E9/Im9. Two loops that exhibit backbone conformation differences are marked in dashed circles. (D) 17 interfacial positions that differ between E2/Im2 and E9/Im9 are represented as orange spheres on the structure of E2/Im2. (E) (**Step I**) Starting from the two wild type pairs, all possible sequences are modeled on both structures. For each sequence, the lowest-strain (best score) backbone is selected. (**II**) An exemplary subgraph for four positions. Mutants (represented by nodes) are connected if they differ by one mutation. '0' and '1' represent the start and end amino acid identities. The minimum-strain path between the terminal nodes is marked in green. (**III**) Two-step experimental validation of the calculated minimum-strain path. First, we test the growth of bacteria carrying a plasmid encoding a mutant pair (**Fig. S1**). Second, we purify the ColE/Im complex and apply it in tenfold serial dilutions to naive JM83 *E. coli* cells. Clear zones of cell death indicate ColE toxicity.

In vivo validation of a single-mutation path

We develop an atomistic modeling framework to discover coevolutionarily plausible (single-amino acid change) paths that connect two extant binding pairs through functional intermediates. The interfaces of E2/Im2 and E9/Im9 differ by 17 interfacial positions (7 and 10 positions on E and Im, respectively) (**Fig. 1D** and **Table S1**). The sequence space that encompasses both wild type pairs comprises all combinations of mutations at these interfacial positions, for a total of 2^{17} (approximately 131,000) possible mutants. In addition to the 17 interfacial mutations, the two colicin pairs differ by 33 mutations outside the interface. To establish a common genetic background for assessing the evolutionary intermediates, we tested a variant of E9/Im9 that encodes all of the E2/Im2 mutations outside the binding interface (termed E9/Im9$_{chimera}$) (**Table S2**). We confirmed that E9/Im9$_{chimera}$ was as viable as E9/Im9 (**Fig. S4**) and that its endonuclease toxicity remained equivalent to that of the two wild type pairs (**Fig. S5**). Additionally, Im9$_{chimera}$ maintained and even exhibited somewhat improved neutralization of E9 and E9$_{chimera}$ relative to Im9 (**Fig. S4**). We thus used wild type E2/Im2 and E9/Im9$_{chimera}$ as the genetic backgrounds for testing intermediates, allowing us to focus only on the interfacial mutations. For design calculations, however, we maintained the wild type identities in all positions outside the interface to minimize the degree of extrapolation from the experimentally determined structures.

Because simultaneous multinucleotide mutations are less likely to occur in evolution compared to single-nucleotide mutations[37], we checked whether some of the 17 interface mutations required several nucleotide exchanges. Indeed, at seven interfacial positions the amino acids cannot be exchanged by single-nucleotide mutations, suggesting that unseen amino acid intermediates may have been necessary to bridge the two extant identities (codon bridges). We analyzed homologs of colicin pairs 2 and 9 to find indications for possible bridging identities that may have been used by evolution but found almost no interface mutations in the natural diversity (**Table S3 and S4**). Because the sequence space that includes all possible codon bridges is too large for atomistic modeling (>10^8 intermediates), we selected at each position one bridging mutation that was physicochemically intermediate between the two extant identities (**Table S1**); for example, in position E:98, we enlarged the sequence space by adding Met to the two extant identities, Arg and Val, and in E:73, we added Ala to the extant identities, Gly and Pro (**Fig. 2A**).

In an evolutionarily plausible path, each intermediate must exhibit low molecular strain to be functional and confer viability on the producing cell. Therefore, we score each mutant according to its system and binding energies using the Rosetta all-atom energy function[38]. The energies are then transformed using a logistic function into values between 0 and 1 that approximate the likelihood that the proteins are stable and exhibit high affinity, and these values are combined into a composite score through a "fuzzy"-logic AND gate[32] that estimates whether both energies are sufficiently favorable. To partly address the possibility that mutations deform the backbone, we model each mutant in the context of the two wild type structures (which differ mainly in two interfacial Im loops and in a modest rigid-body reorientation[34]; **Fig. 1C**) and choose the best-

scoring model to represent each mutant (**Fig. 1E**, step I). Finally, to represent all mutational paths that traverse the sequence landscape, we construct a directed graph in which the nodes represent all mutants, and edges connect nodes that are separated by a single amino acid mutation. We also add edges through the codon bridges to connect the seven amino acids that cannot be exchanged by single-nucleotide mutations. The paths are not forced to go through the codon-bridging identities, however. Instead, edges go both through the bridges and circumventing them, thus ensuring that bridges are selected only if they contribute to lowering molecular strain. The graph edges are weighted according to the negative logarithm of the composite score of the predecessor node (**Fig. 1E**, step II). Thus, any path that connects the terminal nodes representing the extant pairs is a single-mutation trajectory by construction, and the sum of edge weights represents molecular strain along the path (lower weights correspond to lower strain). We then use a shortest-path algorithm[39] to find minimum-strain paths (see Methods). We also computed a path without including bridges to serve as a reference (**Fig. S2**) and three additional paths without bridges that used different parameter choices to construct the graph (**Fig. S3**). We found that the energy profile of the optimal path with codon-bridging mutations was substantially lower (more favorable) than those of any of the other four paths that excluded codon bridges (**Fig. 2B, S2,** and **S3**). Moreover, throughout most of this trajectory, the computed energies were lower than those of the wild type pairs, rising slightly above them in the few final mutants that adopt the backbone conformation of E9/Im9 (**Fig. 2B**).

We next tested all paths by experimental viability screening in *E. coli* to probe the ability of bacteria to produce the high-affinity pairs and to verify that the E proteins maintained their toxicity towards naive cells (that do not harbor any immunity) (**Fig. 1E**, III). Remarkably, experiments revealed that all 18 intermediates from the optimal path with bridging mutations were viable (**Fig. S1**), and that all endonuclease intermediates retained toxicity towards naive JM83 cells (**Fig. S6**). By contrast, the four alternative paths were not feasible, and each exhibited at least seven intermediates that were non-viable (**Fig. S2** and **S3**). We concluded that the minimum-strain path with the bridging mutations is an evolutionarily plausible trajectory linking the two wild type orthogonal colicin pairs through fully functional single-point amino acid mutations. Because our computational predictions are subject to uncertainty and error, we cannot rule out that viable paths exist within the sequence space that excludes bridges. Nonetheless, the fact that the atomistic calculations failed to identify a low-strain path in this sequence space and the predicted best candidate paths all contained inviable intermediates suggests that the bridges are indeed necessary to reduce the molecular strain of introducing mutually incompatible specificity-switching mutations.

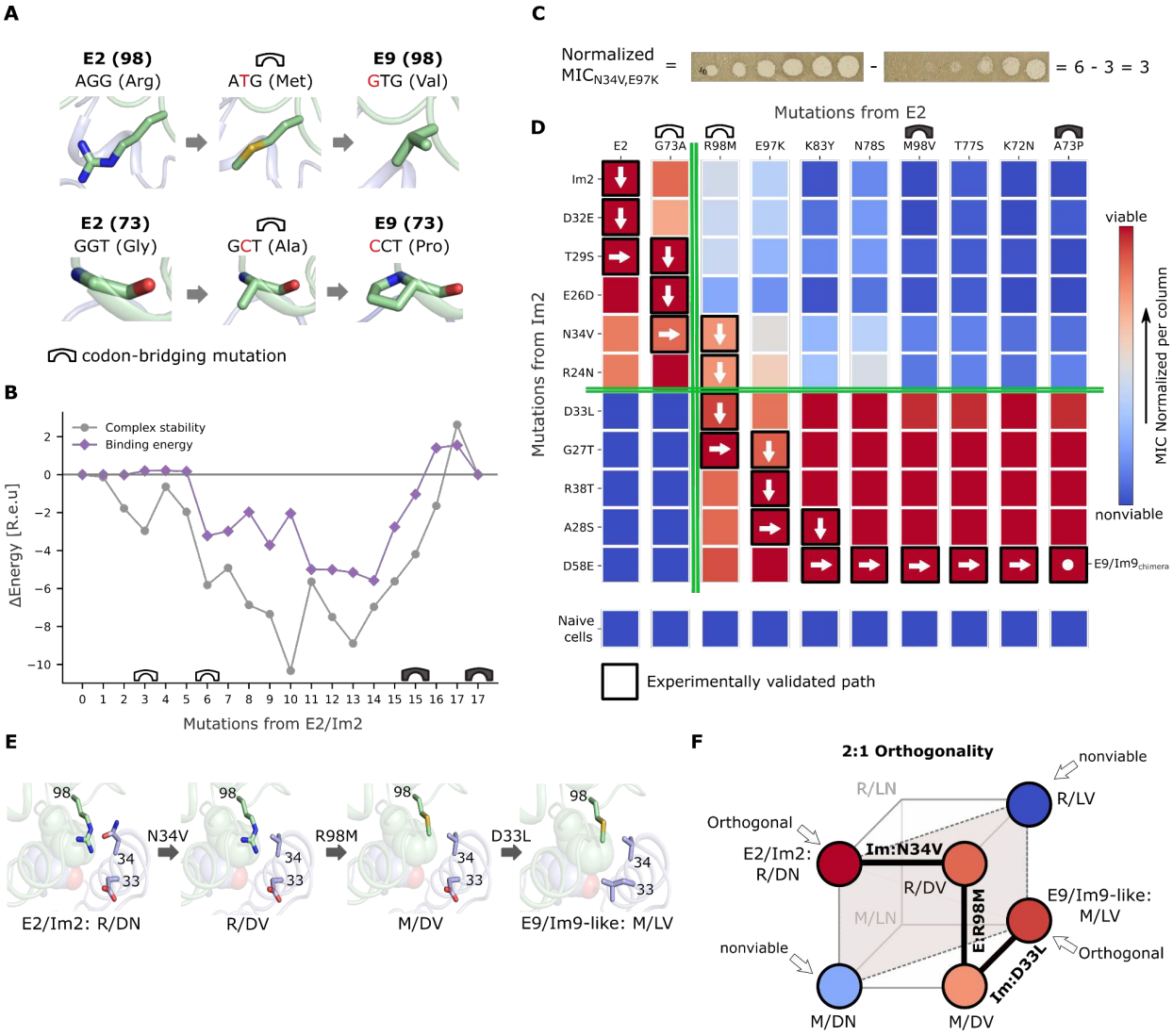


Figure 2. Experimental validation of an evolutionarily plausible path linking two orthogonal binding pairs. (A) Examples of codon-bridging mutations. DNA nucleotide mutations are indicated in red. (B) Binding and stability energy differences of the intermediates in the minimum-strain path relative to the two wild type pairs. An empty bridge icon represents the introduction of a bridging mutation, and a full bridge icon represents its resolution. (C) An example of the inferred normalized minimal inhibitory concentration (MIC). (D) Functional specificity map. Each endonuclease on the viable path was applied to bacteria harboring each immunity protein. Mutations from E2 accumulate from left-to-right and mutations from Im2 accumulate from top-to-bottom. A vertical green line marks the specificity switch on E and a horizontal one marks the specificity switch on Im. The mutations are completely reversible, and arrows are only visual guides. The MIC scores within the map are normalized to the minimal and maximal killing (MIC on naive cells) of each E (in each column), and higher viability corresponds to higher normalized MIC. The scores within the map are an average of two technical repeats (raw averages and standard deviations in **Table S5**). (E) Molecular models of the key mutations producing the orthogonality along the experimentally validated evolutionary path consistent with the model in panel F. The hotspot is represented in spheres. (F) General model of orthogonality in a bimolecular fitness landscape involving two sites in one molecule (Im protein) and one site in the other molecule (E toxin) defining what we call a 2:1 orthogonality model (see Supplementary Information and **Fig. S9**). The corners of the cube represent all the possible pairs of sequences, and colors are consistent with panel D. Thick lines represent the empirical path and dotted lines join the pairs whose fitness determines orthogonality of the pairs for interacting molecules.

An abrupt specificity switch

To investigate whether the specificity switch between the two colicin pairs occurs gradually through many small-effect mutations or abruptly through a few strong-effect mutations, we built a functional quantitative specificity map between E and Im variants along the viable path. We applied the eight intermediate ColE mutants, the wild type ColE2 and ColE9$_{chimera}$ in tenfold serial dilutions to JM83 cells transformed with the nine Im intermediates, wild type Im2, and Im9$_{chimera}$. We then visually estimated how protective each Im protein was with respect to each ColE according to the minimal inhibitory concentration (MIC; the tenfold dilution of ColE in which no clear zones were observed; **Fig. 2C**) and integrated the results into a functional specificity map (**Fig. 2D**). We also verified that all purified ColE/Im complexes exhibited similar concentrations (approx. 3 mg/ml) (**Fig. S7**) and that Im intermediates were similarly expressed (approx. 2 mg/ml) and stable up to 50 °C (**Fig. S8**).

Strikingly, the map reveals that orthogonality is preserved along most of the steps in the evolutionary path, and that as few as three substitutions are sufficient to induce a complete switch in specificity. Two of these substitutions take place at consecutive positions in the immunity protein (Asn34Val and Asp33Leu) and one in the E protein (Arg98Met). Im:Asn34Val abrogates a hydrogen bond to E:Arg98 and acts as an enabling mutation, allowing the mutated Im to bind the E:Arg98Met mutant (**Fig. 2E**). E:Arg98Met completely switches the specificity pattern of the endonuclease, interacting strongly with Im9-like variants but substantially less with the Im2-like variants (**Fig. 2D**). Subsequently, Im:Asp33Leu, a known specificity determinant[27,34], eliminates the Im protein recognition of E2-like endonucleases in favor of E9-like variants and completes the functional transition (**Fig. 2D**).

These results offer several insights that may generalize to other coevolving systems. First, the dramatic specificity switch occurs in only a few evolutionary steps — similar to observations in several molecular systems, though the path reconstructed here is much longer than any that have been previously characterized[4,21,25]. Second, the results also demonstrate how two incompatible molecular interaction networks (charged/polar *vs.* hydrophobic) could be exchanged one mutation at a time. Third, in Supplemental Information, we study general molecular mechanisms that produce orthogonal binders, concluding that they can be classified by the minimal number of mutations that are required for the emergence of orthogonality (**Fig. S9**). The mechanism of orthogonality in the experimentally validated path is what we call "2:1 orthogonality" (**Fig. 2F**). In this model, orthogonality necessarily evolves through a strict order of mutations starting with a mutation on one partner, then the other, and finally a mutation in the first partner that interacts with the first mutation, as observed in the empirical path (Im:Asn34Val→E:Arg98Met→Im:Asp33Leu) (**Fig. 2E** and **2F**). Consistent with this proposed mechanism, the two interacting sites in the same molecule (Im:33 and Im:34) form an epistatic interaction with the same site (E:98) and are indeed spatially close (**Fig. 2E**).

The specificity switch, however, is more complex when viewed relative to the four E mutants that

contain the Met98 codon-bridging mutation. Following the E:Arg98Met mutation and until the E:Met98Val mutation (E9 identity), the endonucleases present a reduced specificity barrier for the Im proteins, though they maintain nearly equal toxicity when applied to naive JM83 cells (**Fig. S6**). Thus, although the Im intermediates close to Im2 are less protective than those close to Im9 against endonucleases that carry Met98, the former nonetheless present a significant level of protection (**Fig. 2D**). The Met98-containing endonucleases are, therefore, multispecific relative to the very strict functional orthogonality observed in all the other endonucleases. E:Arg98Met is thus not only a specificity watershed but may also be an enabling mutation, lowering the specificity barrier and allowing the accumulation of additional E and Im mutations. Once the Val mutation is fixed, the specificity barrier is, once again, extremely strict.

The case presented in this evolutionary trajectory is one in which a minimal number of large-effect mutations (two) establishes a specificity watershed; yet, many surrounding mutations are needed to stabilize and enable the pair of mutations[40].

Stepwise resolution of epistatic interactions

The empirical specificity map shows that many mutations could have appeared in a different order. For example the map suggests that Im:Arg24Asn preceding rather than following E:Arg98Met might have produced an even higher-fitness path than the computed one. In fact, we can determine that >3,500 alternative and fully functional paths traverse the viable regions in the empirical map (red regions in **Fig. 2D**) by counting the number of viable paths within the set of empirically measured genotypes (see Methods). The existence of such alternatives raises questions about the possible ordering of mutations that cannot be answered on the basis of the empirical specificity map alone. We therefore used the empirical data from the specificity map (including cognate and noncognate pairs from **Fig. 2D**) to calibrate a predictor of normalized MIC scores (**Fig. 2C**) based on the Rosetta binding energy of each pair (Pearson r=0.92, p-value<10^{-16} in leave-one-out cross-validation; **Fig. S10**; see Methods). We then predicted normalized MIC scores for all 2.2 million possible E/Im pairs (including all codon bridges) and applied a technique for visualizing fitness landscapes[41] to produce a low-dimensional representation in which distances between pairs approximate the time it may take for a population to evolve from one pair to another under selection for high normalized MIC score (see Methods).

The visualization reveals that the predicted fitness landscape is shaped by three main sets of molecular interactions that separate sequences along the first two "Diffusion axes" (see methods) that define the low-dimensional representation (**Fig. 3A**). The first interaction separates the fitness landscape into three main groups of sequences due to an incompatibility between mutations at positions E:97 and Im:38. In E2/Im2, these positions (E2:Glu97/Im2:Arg38) form a saltbridge across the interface (**Fig. S11A**), whereas in pair 9 they form a polar contact (E9:Lys97/Im9:Thr38) (**Fig. S11A**). Averaging out mutations in all other interface positions shows that this separation is driven by the low average fitness of variants with one of the intermediate

mutational states, *i.e.*, E:Glu97/Im:Thr38, where E:Glu97 is not paired to a countercharge and is therefore strained (**Fig. 3B** & **S11A**). Such ordering of mutations increases the expected time it takes to evolve from one sequence to another, splitting the sequence space in the visualization.

Additionally, the high fitness sequences form an L-shaped pattern along the rim of each cluster in the visualization. The portion of the rim descending from E2/Im2 in the upper left corresponds to sequences that differ at the specificity-switching positions, E:98 and Im:33-34, and is mainly explained by incompatibilities between mutations at these positions that were already revealed by the experimental specificity map (**Fig. 3A**). In particular, **Fig. 3C** shows that several possible fit combinations against the background of the E2 identity E:Lys83 link the E2/Im2 sequence E2:Arg98/Im2:Asp33-Asn34 to the combination E:Met98/Im:Leu33-Val34 that exhibits E9/Im9-like specificity (**Fig. 2D**). Other combinations are predicted to have low average fitness, *e.g.* the combination E:Met98/Im:Asp33-Asn34 which was experimentally shown to be nonfunctional (**Fig. 2D**). Such low-fitness combinations restrict the order in which mutations can be accumulated, explaining why sequences that differ in these specificity-switching positions are separated in our visualizations. Notably, E:Met98/Im:Leu33-Val34 is predicted to be the fittest possible subsequence at these positions and context (**Fig. 3C**), explaining why the minimum-strain empirical path constructs this combination early and maintains it for most of its length.

The third set of interactions defines high-fitness sequences that ascend from the lower left to the upper right in the visualization (**Fig. 3A**). These are involved in the change in backbone conformation between E2/Im2 and E9/Im9 (**Fig. 1C)** that is associated with substantial changes in binding mechanism[34] **(Fig. S11 B&C)**. In the fitness landscape, these interface and backbone changes modify the sign of the effects of many mutations. For example, E:Lys72Asn and the adjacent E:Ala73Pro are both strongly deleterious at the start of this high-fitness region (**Fig. 3D**, left), but they later become only slightly deleterious individually and neutral in combination (**Fig. 3D**, center). Towards the end, with the addition of E:Thr77Ser (which relieves a steric clash with the Im9 helix 1 backbone, **Fig. S11D**), the double mutant becomes strongly beneficial (**Fig. 3D**, right). More generally, structure modeling suggests that the "L" shape of the high-fitness region arises because E:Lys83 is only compatible with a pair 2-like backbone (**Fig. S11C**) whereas E:Asn72 is only compatible with a pair 9-like backbone (**Fig. S11B**). As a result, the combination E:Asn72,Lys83 is incompatible with both structures leading to the large low-fitness regions in the top half of each of the three main clusters (**Fig. S12**).

We verified the robustness of our conclusions to modeling uncertainties. For example, while our analysis was performed assuming a quantitative fitness function (normalized MIC score), mutations that increase MIC score might not increase organismal fitness if fitness depends on survival/death in only a specific concentration of toxin. Nevertheless, we find that the structure of the fitness landscape and the main genetic interactions remain robust even when using a threshold-like fitness function along the whole range of reasonable cutoffs on the normalized MIC score

(**Fig. S13 and S14**). Our analysis also remains robust to noise added to the Rosetta energies before the calibration procedure (**Fig. S15**). This robustness is likely due to the fact that the major aspects of the landscape structure depend on a small set of grossly incompatible combinations of mutations (such as E:Glu97/Im:Thr38). Correctly identifying these incompatibilities does not depend strongly on the details of the molecular force field or the precise choice of fitness function.

The global view of the predicted fitness landscape suggests several insights in addition to those we inferred from the empirical specificity map (**Fig. 2D**). First, we find that the empirically validated path resolves each epistatic set of interacting sites one at a time, likely minimizing strain by not superimposing incompatibilities (**Fig. 3A,** see also **Fig. 2D)**. Second, by visualizing pairs that are orthogonal with respect to either E2/Im2 or E9/Im9, we find that each set is concentrated in a region of sequence space containing one of the extant protein pairs, (**Fig. 3E,F**) so that the evolution of functional orthogonality can be understood in terms of the population entering or exiting specific regions of sequence space. This analysis is robust to the thresholds used to define orthogonality (**Fig. S16** & **S17**). Third, the global analysis suggests other mutations that may impact specificity, such as Im:Arg38Thr, that were not seen to impact specificity in the empirical path (**Fig. S18 and S19**).

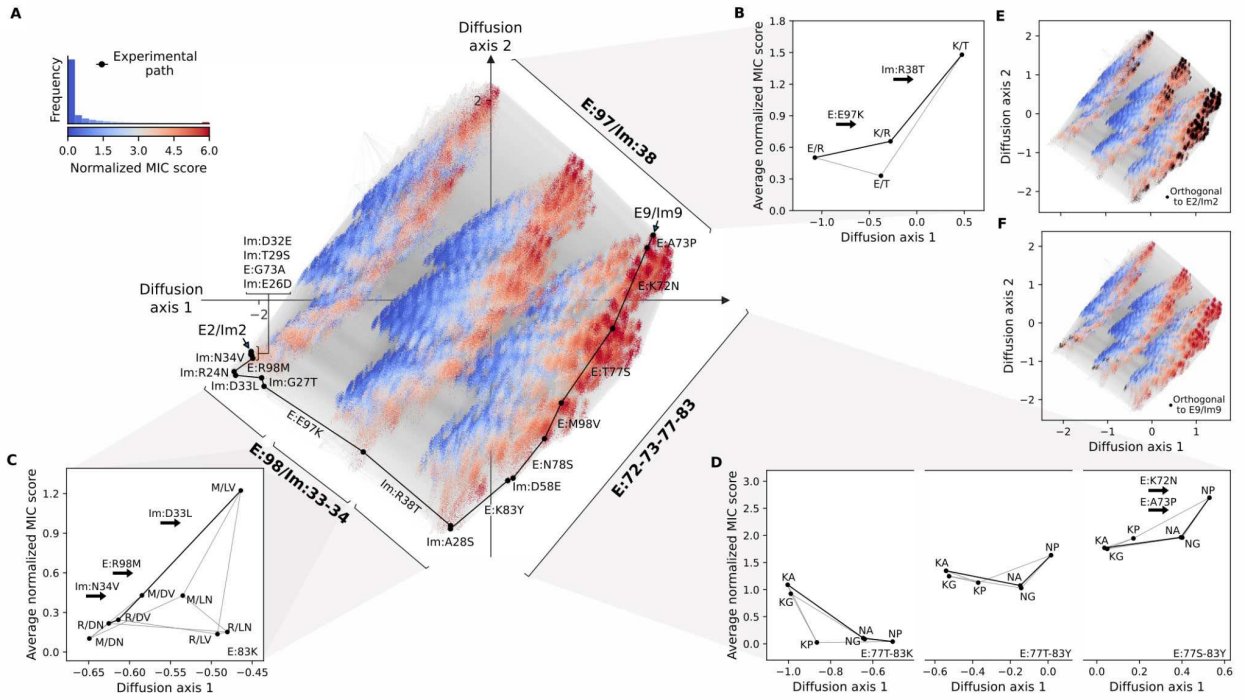


Figure 3. Computed global fitness landscape highlights the potential role of multiple epistatic interactions in effecting the specificity switch. (A) Visualization of the predicted fitness landscape. Distance between points represents the time to evolve between E/Im pairs under selection for high normalized MIC scores (see Methods). Nodes represent protein pairs and edges connect pairs that differ by a single amino acid substitution. Node color indicates normalized MIC score. Sequences with higher scores are plotted close to the viewer above to emphasize the relationships among functional sequences. Empirical path is highlighted in black. Inset shows the distribution of predicted normalized MIC scores across all possible pairs together with the color scale. Interactions between key

mutations are marked in the visualization and simplified in panels B-D depicting the networks of average MIC scores defined by the identities at the key sets of interacting positions. (B,C) Diagrams representing the average MIC score and average position at Diffusion axis 1 for each possible subsequence at positions E:97/Im:38 (B) and E:98/Im:33-34 in the E:Lys83 context (C). The empirical trajectory across each set of sites is highlighted by a thick line. (D) Diagrams representing the average MIC score and average position at Diffusion axis 1 for each possible subsequence at positions E:72-73 in the E:Lys83-Thr77 context on the left, E:Tyr83-Thr77 on the center, and in the E:Tyr83-Ser77 context on the right. (E,F) Visualization of the fitness landscape with overlaid black dots indicating pairs that are predicted to be orthogonal to pair E2/Im2 (E) and E9/Im9 (F). Pairs were defined to be orthogonal to either E2/Im2 or E9/Im9 if their predicted normalized MIC score was at least as high as that of the alternative extant pair minus 0.5, and if their combinations with E2 and Im2 or E9 and Im9, respectively, were at most 0.5 greater than the minimum of E2/Im9 and E9/Im2.

Adaptive mutations may drive functional divergence

The atomistic and graph-based calculations we employed are based solely on the physical constraints that a generic protein pair must meet to retain binding affinity. If these were the only determinants of fitness, we would expect natural populations to mutate neutrally among functional pairs. The fact that the pair forms a toxin-antitoxin system, however, can influence how populations evolve and functionally diverge. In this case, fitness depends not only on the ability of the immunity protein to neutralize its cognate endonuclease but also on whether the immunity protein can neutralize other endonucleases existing in the population and whether the cognate endonuclease can escape recognition by other immunity proteins, killing competing strains.

Given these considerations, we re-examined the functional specificity map (**Fig. 2D**) to investigate the direction, if any, in which the experimental path is adaptive or maladaptive. We tracked how the relative MIC score (Methods) changes along the path with respect to each extant pair. In this simplified analysis, an intermediate exhibits high fitness relative to the parental population if its Im protein resists the parental endonuclease and its endonuclease escapes resistance by the parental Im protein (Equation in Figure 4 caption). While any sequence in the path could act as a hypothetical ancestor, we started by assuming that evolution begins at the E2/Im2 pair or mutants that are phenotypically similar, and that it proceeds in a population dominated by such pairs (**Fig. 4**, purple line, Step 1). After a series of neutral mutations that do not affect binding specificity, the enabling mutation Im:Asn34Val expands the repertoire of endonucleases that the Im can neutralize, including E:Met98. At the same time, Im:Asn34Val maintains binding affinity towards the ancestral E2. The following mutation, E:Arg98Met (Step 2), introduces an E variant that is, for the first time, cytotoxic towards the parental bacteria harboring an Im2-like protein while its cognate Im protein protects against E2-like endonucleases (**Fig. 2D**). Thus, the E:Arg98Met mutation is not only tolerated but may have an adaptive advantage that would allow it to penetrate, persist and spread within the parental population. The E:Arg98Met mutant, however, does not form an optimal interaction with the Im protein prior to incorporating the Im:Asp33Leu mutation, and the Met98-containing endonucleases are only partially cytotoxic towards Im2-containing hosts (**Fig. 2D**). Thus, Im:Asp33Leu improves the viability of this pair relative to the E:Arg98Met, and in Step 3, the codon-bridging mutation is resolved to the E9 identity through E:Met98Val. At this

point, the endonuclease gains high cytotoxicity against the Im2-like harboring bacteria and completes the functional transition. In contrast, in a starting population dominated by an E9/Im9-like pair, the reverse trajectory passes through at least one maladaptive step (**Fig. 4**, gray line): Im:Leu33Asp reduces viability relative to the Im9-like proteins and is strongly susceptible to killing by parental E9-like endonucleases (**Fig. 2D**). In this simplified analysis, the adaptive mutations may provide a "ratchet"-like mechanism in which each adaptive step reduces the likelihood of the reverse mutation and increases the likelihood of accumulating further function-switching mutations.

There are limitations to our analyses, however. Although viability was measured *in vivo* and in the native host, experimental measures of functionality may not be completely faithful to viability in a wild population in nature. Despite these limitations, our results provide an example of how the specific nature of a protein-protein interaction, *i.e.*, a toxin-antitoxin system, may influence evolutionary paths, transforming what are *a priori* neutral mutations (with respect to binding) into adaptive or deleterious ones depending on the dominant genotypes in the population.

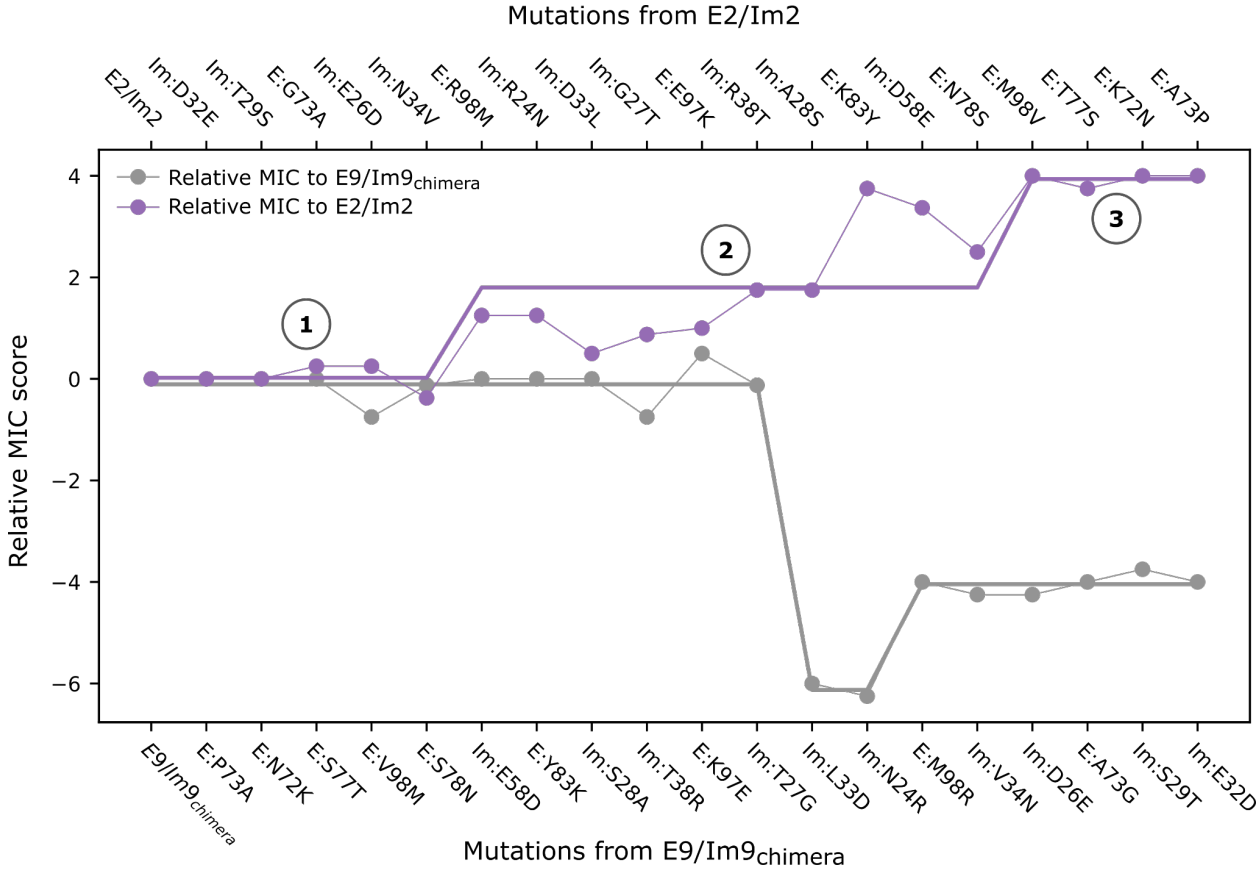


Fig. 4. A "ratchet-like" model for the evolution of ultrahigh specificity. The experimentally determined viability of each cognate pair on the path relative (**Fig. 2D**) to the extant E2/Im2 (purple line, top *x*-axis) or E9/Im9 (gray, bottom *x*-axis). Viability increases in two steps in the direction of accumulating mutations from E2/Im2 and decreases in the reverse direction. Straight lines are guides to the eye. Higher viability corresponds to higher normalized MIC score. Relative MIC score (relative viability score) is the difference in MIC scores between a mutant and the parental

wild type: $RM_i = (M_{E(i)/Im(i)} - M_{E(i)/Im(wt)}) - (M_{E(wt)/Im(wt)} - M_{E(wt)/Im(i)})$, where RM is the relative MIC score; M is the MIC score (see Methods); *i* is an index to the mutant; *wt* is the parental protein, E2/Im2 or E9/$Im9_{chimera}$.

Discussion

We showed how a charged protein-protein interface could transform, one-mutation-at-a-time, into a largely hydrophobic one without compromising stability, affinity and catalytic activity below their functional threshold in any intermediate step. To achieve this, we found that a codon-bridging mutation temporarily reduces the specificity barriers, permitting mutations that would otherwise reduce stability or binding affinity. Unlike previously reconstructed specificity switches where the intermediates were promiscuous "generalist" binders[2], however, here the specificity switch occurs suddenly, with the key bridging mutation E:Arg98Met eliminating the interaction with Im2 and creating a new interaction with Im9 in a single step. Furthermore, while the structure of the genetic code is known to constrain the possible paths to evolve a new specificity[12], our results provide an example in which the extra biochemical diversity provided by bridging mutations is likely essential to obtain a viable path between the orthogonal pairs.

A key question in the evolution of orthogonal binding pairs is how ultrahigh specificity evolves by a single-mutation trajectory without crossing a fitness valley. Our results provide a case-study in which each of the specificity-switching mutations are not only tolerated but may endow their host with a selective advantage relative to the parental population due to functional asymmetry in the interacting pair, as in a toxin-antitoxin system. This polarizes the function-altering evolutionary process, increasing the likelihood of selecting a long series of mutations, whereas the reverse mutations are counterselected. In other words, the functional asymmetry in the toxin-antitoxin system suggests preferred directions for the evolutionary process depending on specific environmental conditions.

This work introduces energy-based modeling as an approach to study both the overall shape and the fine details of a highly epistatic fitness landscape. The approach is general and can be implemented, in principle, to suggest evolutionarily plausible trajectories that link wild type proteins or inferred ancestral states, complementing and extending existing experimental and computational methods[2,3,12,15–19]. It allows full enumeration of elaborate molecular interaction networks and may provide detailed mechanistic insights on the evolution of microbial resistance mutations and the evolution of new activities in enzymes and binders informing how to design new protein activity.

Acknowledgments

In memory of Dan S Tawfik who made critical contributions to this work. We thank Ravit Netzer, Dina Listov and Rosalie Lipsh-Sokolik for helpful reading. Research in the Fleishman lab was supported by the Volkswagen Foundation (94747), the European Research Council through a Consolidator Award (815379), the Israel Science Foundation (1844), the Dr. Barry Sherman

Institute for Medicinal Chemistry, and a donation in memory of Sam Switzer. Research in the McCandlish lab was supported by NIH grant R35GM133613, an Alfred P. Sloan Research Fellowship, and additional funding from the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

Author contributions

ZA and SJF developed the computational framework for finding low-strain evolutionary paths. ZA performed all the experiments. SYH helped with the interpretation and understanding of the results. CMG and DMM developed the general formalism of orthogonality in fitness landscapes. CMG generated and analyzed the fitness landscape. ZA, CMG, DMM and SJF wrote the manuscript.

Competing interests: The authors declare no competing interests.

Methods

Atomistic modeling

We enumerated all possible combinations of mutations and modeled them in Rosetta on the backbones of E2-Im2 and E9-Im9 (PDB codes: 3u43 and 1emv, respectively). Each mutant was modeled using all-atom Rosetta calculations, including combinatorial sidechain packing, and backbone and side chain minimization subject to harmonic restraints on the Cα coordinates[42]. Each mutant was modeled on the structures of both E2-Im2 and E9-Im9. All intermediates were ranked according to a composite score based on "fuzzy"-logic design (W)[32] (Eq. 1) accounting for the difference of all-atom energy ($\Delta\Delta G_s$) and difference of binding energy ($\Delta\Delta G_b$) compared to the two wild type sequences.

$$(1) \qquad W = -ln\left(\frac{1}{1 + e^{0.7\times(\Delta\Delta G_S - 5)}}\right) - ln\left(\frac{1}{1+e^{0.7\times(\Delta\Delta G_b - 5)}}\right)$$

For each sequence, the higher (more favorable) composite score was selected from the two calculations.

Finding a least-frustration path using a graph-theoretical shortest-path algorithm

The sequence space comprised approximately 130,000 sequences in the initial four formulations (using a binary sequence space) and 2.4 x 10^6 for the sequence space with codon-bridging mutations. We computed a truncated graph by taking the top 5% nodes with the highest composite scores. Nodes were connected by an edge if they differed by one amino acid substitution and contained all the previous mutations. The weight of each edge was assigned as W (Eq. 1) of the preceding node. Then, the Bellman-Ford shortest-path algorithm[39] was applied to find the path with the lowest sum of scores (least strained). In the graph that contained amino acids that served as DNA bridging mutations, edges were added to connect mutants that could not be crossed with

single-nucleotide mutations.

In the preliminary work that resulted in the four low-strain paths that were not viable (**Fig. S2** and **S3**), the graphs were constructed using somewhat different ways but with the same scoring rules as described above. Briefly, in (1) the interface was divided into three spatially separated modules. Then the shortest path was found inside each module and between the modules. In (2) intermediates that exhibited more than 7 R.e.u. difference in system energy were eliminated. (3) intermediates with binding energy difference to wild type or binding strain difference to wild type above 2.5 and 1.5 R.e.u. respectively were eliminated and a graph was formed with the remaining nodes. In this trial, we applied another Rosetta filter, BindingStrain, which calculates the rotameric strain of monomers upon binding[43]. The intermediate scores in those three paths were calculated using Rosetta energy function, talaris14. And in (4) the graph was formed as the graph that contains the bridging mutations, but after eliminating the bridging nodes. In this fourth graph and the one with bridging mutations we used the new Rosetta energy function, ref15, to model and rank the intermediates. This energy function improves the treatment of electrostatics and solvation relative to the talaris14 energy function we previously used for the first three paths[38].

Number of viable paths in the experimental specificity map

We compute the number of viable directed, shortest-path (no reversions) evolutionary trajectories based on the data from the empirical specificity map (**Fig. 2D**). We note first that the space of viable intermediates in the specificity map is roughly separated by a bottleneck into two large grid-like blocks (2 x 6 and 8 x 5). The number of paths between two corners of an n by k grid is given by:

$$\frac{(n+k-2)!}{(n-1)!(k-1)!}$$

The total number of paths through both blocks is the product of the number of paths in each block. However, the calculation on the first block is slightly more complicated, because the path can exit via the mutation E:Arg98Met before or after Im:Arg24Asn. As a result, the total number of paths through the first block is equal to the sum of the number of paths between the corners of a 2x6 grid and the number of paths between the corners of a 2x5 grid. We thus obtain the following for the total number of unique paths:

$$\left(\frac{(6+2-2)!}{(6-1)!(2-1)!} + \frac{(5+2-2)!}{(5-1)!(2-1)!}\right) \times \frac{(5+8-2)!}{(5-1)!(8-1)!} = 11 \times 330 = 3{,}630.$$

Cloning

pET21d plasmid harboring ColE2 wild type gene (full gene including R, T, and C domains)

followed by Im2 wild type gene (separated by a 2-bp frameshift) with a C-terminal His_6-tag[34] was used as the basis for cloning. Synthetic genes for the two wild types and the intermediates of the computed least-strain path were codon-optimized for efficient *E. coli* expression and custom synthesized as linear fragments by TWIST Bioscience. The genes were swapped with the wild type gene by restriction-free (RF) cloning[44,45]. The plasmids were transformed into T7 Express *lysY/I^q E.coli* cells (NEB), and DNA was extracted for Sanger sequencing to validate accuracy. Validated plasmids of the intermediates were also tested for viability by transformation to BL21 DE3 *E. coli* cells. An intermediate pair was considered viable if a lawn of bacteria was observed in the plate. Three colonies were picked for repeat sequencing to validate that no additional mutations were introduced. Sequence-validated colonies of BL21 *E. coli* were stored in glycerol stock at -80℃ for expression and purification.

Protein expression and purification

BL21 (DE3) cultures were grown in Luria Broth (LB) with 100 mM ampicillin medium at 37 ˚C to OD_{600}=0.6–0.8 and induced with 1 mM IPTG at 16 °C for 20 hrs. Cells were harvested and stored at −20 °C. Pellet was resuspended in 10 ml lysis buffer in 200 ml culture containing 50 mM Tris (pH 8.6), 50 mM NaCl, 10 mM imidazole, 2 mM $MgCl_2$, benzonase (1:500), and protease inhibitor (1:1000) sonicated in 3 cycles of 20 sec On and 40sec Off (1 min On-time) and centrifuged as previously described[34]. The supernatant was loaded onto a column packed with 1.2 ml Ni-NTA beads for 200ml culture, equilibrated with lysis buffer, and incubated at 4°C for 1hr for binding, then washed with lysis buffer containing 20 mM imidazole, and eluted with lysis buffer containing 500 mM imidazole. Protein complexes were dialyzed overnight in 50 mM Tris (ph 8.6), 200 mM NaCl, and 2 mM $MgCl_2$. Protein complexes were stored at 4°C for a week and for longer times at -80°C.

Im proteins were purified using the same protocol as above with minor changes. The pH for the buffers was adjusted to 7.3, they were grown in 50 ml culture and all quantities were adjusted accordingly. Melting curves for the purified Im proteins were measured using Tyco NT.6 (Nanotemper Technologies Inc) at a rate of 30 °C/min.

All protein concentrations were measured using absorbance at 280nm. The concentrations for the ColE/Im complexes range between 2-3.5 mg/ml and Im protein concentrations range between 1.9-3 mg/ml.

ColE/Im *in vivo* assay (spot-test)

The biological activities of ColE and Im proteins were determined using a modified form of the spot-test described in refs. [46,47]. Briefly, ColE-sensitive *E. coli* JM83 (Addgene) cells were transformed with a control vector (empty pET21d) or one of the Im proteins and incubated in LB containing 100 mM Ampicillin at 37°C overnight. Then, 1 ml of culture was plated on agar plates

to create a lawn of bacteria. Purified ColE/Im complexes (concentration between 2-4 mg/ml) were 6 times diluted in 10-fold serial dilutions and 8 µl aliquots were dropped on the plates containing bacteria. Plates were incubated overnight at 37°C. Clear zones represent cell death and the absence of clear zones indicates biological protection conferred by the Im. The minimal inhibitory concentration (MIC; the concentration dilution of ColE in which no clear zones were observed) was set as the last visually apparent drop (0, 1, 2, 3, 4, 5, or 6). The normalized MIC score is then computed as the difference between the maximal killing (the number of drops against an empty plasmid) and the MIC. All viability assays were conducted using two technical repeats and averages and standard deviations are shown in **Fig. 2D** and **Table S5**.

Calibrating normalized MIC scores against Rosetta binding energies

In this section, we aimed to define a model that allows us to transform the computationally predicted binding energies for each E/Im pair sequence i into the more biologically relevant scale of the normalized MIC score that takes into account the order of magnitude increase in the tolerated concentration of E toxin. The binding energy of each E/Im pair was calculated using both the E2/Im2 and E9/Im9 backbones. The minimum of the two is chosen for each pair, taking into account a free parameter γ corresponding to the difference in binding energy between the wildtype E9/Im9 and E2/Im2 structures:

$$\Delta\Delta G_{binding,i} = min(\Delta\Delta G_{E2/Im2,i}, \Delta\Delta G_{E9/Im9,i} + \gamma)$$

Because the Im can only increase but not decrease the tolerance, the normalized MIC score cannot be lower than 0. To take into account, we model the expected normalized MIC score ($\bar{y}_i$) as a function of the computationally predicted wild-type normalized binding energy $\Delta\Delta G_{binding,i}$:

$$\bar{y}_i = log(1 + e^{\alpha + \beta\Delta\Delta G_{binding,i}})$$

However, we cannot measure directly $\bar{y}_i$ but only obtain data with experimental errors. Moreover, as we have data only for six concentrations, our observed normalized MIC score will not be higher than the number of tested concentrations, but the real underlying score could indeed be larger. Thus, we define a piecewise likelihood function for dependence on the observed y_i, together with the standard deviation σ for unexplained variation:

$$p(y_i \mid \bar{y}_i, \sigma) = p(y_i \geq 6 \mid \bar{y}, \sigma) \text{ if } y_i = 6$$
$$p(y_i \mid \bar{y}_i, \sigma) = p(y_i \leq 0 \mid \bar{y}, \sigma) \text{ if } y_i < 0$$
$$p(y_i \mid \bar{y}_i, \sigma) = p(y_i \mid \bar{y}, \sigma) \text{ if } 0 < y_i < 6$$

The data from the specificity map computed in **Fig. 2D** show consistent normalized MIC scores across the two biological replicates (Pearson r=0.99, **Fig. S10A**). We fit the model by maximizing the likelihood using the Nelder-Mead algorithm; that is, we find the combination of values for $\alpha, \beta, \gamma, \sigma$ that maximizes the likelihood, *i.e.*, the probability of observing the data under our calibration model. Using the maximum likelihood estimate $(\hat{\alpha} = 2.31, \hat{\beta} = -0.91, \hat{\gamma} = -9.46, \hat{\sigma} = 1.24)$, we can make predictions given the computed binding energies for any E/Im pair. While our

predictions can yield values beyond 6, in order to have predicted normalized MIC scores in the same range as the empirical path, scores were truncated to 6. Performance in both training and leave-one-out held out data is excellent (cross-validated Pearson r=0.92, **Fig. S10B,C**). Moreover, the vast majority of the data lie within the range of binding energies spanned by the experimentally validated data, suggesting that we can effectively interpolate the normalized MIC scores for nearly the full Im/E sequence space (**Fig. S10D**).

Fitness landscape visualization

Visualization method as previously described[41]. Briefly, we construct a model of molecular evolution where a population evolves via single amino acid substitutions and the rate at which each possible substitution becomes fixed in the population reflects its selective advantage or disadvantage relative to the currently fixed sequence. More specifically, in our model the rate of evolution from sequence i to any mutationally adjacent sequence j is given by

$$Q_{ij} = \frac{S_{ij}}{1 - e^{S_{ij}}}$$

where S_{ij} is the scaled selection coefficient (population size times the selection coefficient of j relative to i), time is measured relative to the amino acid mutation rate (each possible amino acid mutation occurs at rate 1), and the total leaving rate from each sequence i is given by

$$Q_{ii} = -\sum_{j \neq i} Q_{ij}$$

In the current context, $S_{ij} = c(f_j - f_i)$, where f_i is the predicted normalized MIC score at sequence i. For this analysis we choose c so that the equilibrium expected normalized MIC score is 2, corresponding to a roughly threefold increase relative to the expected normalized MIC score of 0.76 under neutrality, and close to the predicted normalized MIC score of the wild-type E2/Im2 of 2.40.

Given the rate matrix Q for our evolutionary model, we then construct the visualization by using the subdominant right eigenvectors associated with the smallest magnitude non-zero eigenvalues of this rate matrix as coordinates for the low dimensional representation of the landscape, where each such coordinate defines one of the "diffusion axes" used in the visualization. This produces a visualization that reflects the long-term barriers to diffusion in sequence space, and, in particular, clusters of sequences in the visualization correspond to sets of initial states from which the evolutionary model approaches its stationary distribution in the same manner, and multi-peaked fitness landscapes appear as broadly separated clusters with one peak in each cluster. Moreover, by scaling the axes appropriately, as is done here, these axes can be given units of $\sqrt{time}$ and it can be shown that the resulting distances reflect evolutionary times under this model. In particular, using these coordinates, the squared Euclidean distance between arbitrary sequences i and j optimally approximates (in a specific sense) the sum of the expected time to evolve from i to j and the expected time to evolve from j to i. See ref. [41] for details.

References

1. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**, 383–386 (2007).

2. Aakre, C. D. *et al.* Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594–606 (2015).

3. Levin, K. B. *et al.* Following evolutionary paths to protein-protein interactions with high affinity and selectivity. *Nat. Struct. Mol. Biol.* **16**, 1049–1055 (2009).

4. Siddiq, M. A., Hochberg, G. K. & Thornton, J. W. Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr. Opin. Struct. Biol.* **47**, 113–122 (2017).

5. McClune, C. J. & Laub, M. T. Constraints on the expansion of paralogous protein families. *Curr. Biol.* **30**, R460–R464 (2020).

6. Meenan, N. A. G. *et al.* The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10080–10085 (2010).

7. Papadakos, G., Wojdyla, J. A. & Kleanthous, C. Nuclease colicins and their immunity proteins. *Q. Rev. Biophys.* **45**, 57–103 (2012).

8. Kuriyan, J., Konforti, B. & Wemmer, D. *The molecules of life: Physical and chemical principles*. (Garland Science, 2012).

9. Mohammadi, M., Olsen, S. K. & Ibrahimi, O. A. Structural basis for fibroblast growth factor receptor activation. *Cytokine Growth Factor Rev.* **16**, 107–137 (2005).

10. Itoh, N. & Ornitz, D. M. Evolution of the Fgf and Fgfr gene families. *Trends Genet.* **20**, 563–569 (2004).

11. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).

12. Podgornaia, A. I. & Laub, M. T. Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673–677 (2015).

13. Bridgham, J. T., Carroll, S. M. & Thornton, J. W. Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312**, 97–101 (2006).

14. Welin, M. & Nordlund, P. Understanding specificity in metabolic pathways—Structural biology of human nucleotide metabolism. *Biochem. Biophys. Res. Commun.* **396**, 157–163

(2010).

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 (2017).

16. McClune, C. J., Alvarez-Buylla, A., Voigt, C. A. & Laub, M. T. Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space. *Nature* **574**, 702–706 (2019).

17. Poelwijk, F. J., Socolich, M. & Ranganathan, R. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature Communications* vol. 10 Preprint at https://doi.org/10.1038/s41467-019-12130-8 (2019).

18. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).

19. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).

20. Ding, D. *et al.* Co-evolution of interacting proteins through non-contacting and non-specific mutations. *Nat Ecol Evol* **6**, 590–603 (2022).

21. Nocedal, I. & Laub, M. T. Ancestral reconstruction of duplicated signaling proteins reveals the evolution of signaling specificity. *Elife* **11**, (2022).

22. Hochberg, G. K. A. & Thornton, J. W. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu. Rev. Biophys.* **46**, 247–269 (2017).

23. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* **2**, e00631 (2013).

24. Harms, M. J. & Thornton, J. W. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* **14**, 559–571 (2013).

25. Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution: the functional synthesis. *Nat. Rev. Genet.* **8**, 675–688 (2007).

26. Cramer, W. A., Sharma, O. & Zakharov, S. D. On mechanisms of colicin import: the outer membrane quandary. *Biochem. J* **475**, 3903–3915 (2018).

27. Li, W. *et al.* Highly Discriminating Protein–Protein Interaction Specificities in the Context of a Conserved Binding Energy Hotspot. *J. Mol. Biol.* **337**, 743–759 (2004).

28. Keeble, A. H., Kirkpatrick, N., Shimizu, S. & Kleanthous, C. Calorimetric dissection of

colicin DNase--immunity protein complex specificity. *Biochemistry* **45**, 3243–3254 (2006).

29. Kirkup, B. C. & Riley, M. A. Antibiotic-mediated antagonism leads to a bacterial game of rock-paper-scissors in vivo. *Nature* **428**, 412–414 (2004).

30. Kortemme, T. *et al.* Computational redesign of protein-protein interaction specificity. *Nat. Struct. Mol. Biol.* **11**, 371–379 (2004).

31. Netzer, R. *et al.* Ultrahigh specificity in a network of computationally designed protein-interaction pairs. *Nat. Commun.* **9**, 5286 (2018).

32. Warszawski, S., Netzer, R., Tawfik, D. S. & Fleishman, S. J. A 'fuzzy'-logic language for encoding multiple physical traits in biomolecules. *J. Mol. Biol.* **426**, 4125–4138 (2014).

33. Keeble, A. H. *et al.* Experimental and computational analyses of the energetic basis for dual recognition of immunity proteins by colicin endonucleases. *J. Mol. Biol.* **379**, 745–759 (2008).

34. Wojdyla, J. A., Fleishman, S. J., Baker, D. & Kleanthous, C. Structure of the ultra-high-affinity colicin E2 DNase-Im2 complex. *J. Mol. Biol.* **417**, 79–94 (2012).

35. Kosloff, M., Travis, A. M., Bosch, D. E., Siderovski, D. P. & Arshavsky, V. Y. Integrating energy calculations with functional assays to decipher the specificity of G protein-RGS protein interactions. *Nat. Struct. Mol. Biol.* **18**, 846–853 (2011).

36. Sharp, C., Bray, J., Housden, N. G., Maiden, M. C. J. & Kleanthous, C. Diversity and distribution of nuclease bacteriocins in bacterial genomes revealed using Hidden Markov Models. *PLoS Comput. Biol.* **13**, e1005652 (2017).

37. Kondrashov, D. A. & Kondrashov, F. A. Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* **31**, 24–33 (2015).

38. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).

39. Bang-Jensen, J. & Gutin, G. Z. *Digraphs: Theory, Algorithms and Applications*. (Springer, 2008).

40. Ortlund, E. A., Bridgham, J. T., Redinbo, M. R. & Thornton, J. W. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* **317**, 1544–1548 (2007).

41. McCandlish, D. M. Visualizing fitness landscapes. *Evolution* **65**, 1544–1558 (2011).

42. Khersonsky, O. *et al.* Automated Design of Efficient and Functionally Diverse Enzyme

Repertoires. *Mol. Cell* **72**, 178–186.e5 (2018).

43. Fleishman, S. J., Khare, S. D., Koga, N. & Baker, D. Restricted sidechain plasticity in the structures of native proteins and complexes. *Protein Sci.* **20**, 753–757 (2011).

44. Unger, T., Jacobovitch, Y., Dantes, A., Bernheim, R. & Peleg, Y. Applications of the Restriction Free (RF) cloning procedure for molecular manipulations and protein expression. *J. Struct. Biol.* **172**, 34–44 (2010).

45. Erijman, A., Dantes, A., Bernheim, R., Shifman, J. M. & Peleg, Y. Transfer-PCR (TPCR): a highway for DNA cloning and protein engineering. *J. Struct. Biol.* **175**, 171–177 (2011).

46. Wallis, R. *et al.* In vivo and in vitro characterization of overproduced colicin E9 immunity protein. *Eur. J. Biochem.* **207**, 687–695 (1992).

47. Wallis, R. *et al.* Protein-Protein Interactions in Colicin E9 DNase-Immunity Protein Complexes. 2. Cognate and Noncognate Interactions That Span the Millilmolar to Femptomolar Affinity Range. *Biochemistry* vol. 34 13751–13759 Preprint at https://doi.org/10.1021/bi00042a005 (1995).

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

49. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

50. O'Meara, M. J. *et al.* A Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **11**, 609–622 (2015).

51. Lite, T.-L. V. *et al.* Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. *Elife* **9**, (2020).

52. Crona, K. Rank orders and signed interactions in evolutionary biology. *Elife* **9**, (2020).

53. Conrad, M. The geometry of evolution. *Biosystems.* **24**, 61–81 (1990).

Supplementary information

Supplementary tables

Position	Chain	E2/Im2	E9-Im9	Bridging mutation
24	**A**	R	N	K
26	**A**	E	D	
27	**A**	G	T	S
28	**A**	A	S	
29	**A**	T	S	
32	**A**	D	E	
33	**A**	D	L	V
34	**A**	N	V	D
38	**A**	R	T	
58	**A**	D	E	
72	**B**	K	N	
73	**B**	G	P	A
77	**B**	T	S	
78	**B**	N	S	
83	**B**	K	Y	N
97	**B**	E	K	
98	**B**	R	V	M

Supplementary Table 1. The 17 interfacial positions that differ between E2/Im2 and E9/Im9 and the resulting expanded sequence space after adding single codon-bridging amino acid to the two parental identities. Position numbering and chains according to E2/Im2 PDB entry: 3u43.

E2/Im2 wt	MESKRNKPGKATGKGKPVGDKWLDDAGKDSGAPIPDRIADKLRDKEFKNFDDFRKKFWEEV SKDPDLSKQF**KG**SNK**T**N**I**QKG**K**APFARKKDQVGGR**E**RFELHHDKPISQDGGVYDMNNIRVTT PKRHIDIHRGK/MELKHSISDYTEAEFLEFVKKIC**R**A**EGAT**EE**DD**NKLV**R**EFERLTEHPDGSDLI YYPR**D**DREDSPEGIVKEIKEWRAANGKSGFKQGLE
E9/Im9 wt	MESKRNKPGKATGKGKPVGDKWLDDAGKDSGAPIPDRIADKLRDKEFKSFDDFRKAVWEEV SKDPELSKNL**NP**SNK**SS**VSKG**Y**SPFTPKNQQVGGR**KV**YELHHDKPISQGGEVYDMDNIRVTTP KRHIDIHRGK/MELKHSISDYTEAEFLQLVTTIC**N**A**DTSS**EE**ELV**KLV**T**HFEEMTEHPSGSDLIY YPK**E**GDDDSPSGIVNTVKQWRAANGKSGFKQGLE
E9/Im9 chimera	MESKRNKPGKATGKGKPVGDKWLDDAGKDSGAPIPDRIADKLRDKEFK**N**FDDFRK**KF**WEEV SKDP**D**LSK**QFNP**SNK**SS**I**Q**KG**YA**PFA**R**KK**D**QVGGR**KV**YELHHDKPISQ**D**GG**V**YDM**N**NIRVTT PKRHIDIHRGK/MELKHSISDYTEAEFL**E**F**V**KK**I**C**N**A**DTSS**EE**ELV**KLV**TE**FE**RL**TEHP**D**GSDLI YYP**RE**D**RE**DSP**E**GIV**KEI**K**E**WRAANGKSGFKQGLE

Supplementary Table 2. The sequences of the wild type colicins E2/Im2 and E9/Im9 and the sequence of E9/Im9$_{chimera}$. The 17 interfacial positions are marked in bold and orange font. The 33 changes outside of the interface are marked on the sequence of E9/Im9$_{chimera}$ in red font.

Accession #	72	73	77	78	83	97	98
E2 wt	K	G	T	N	K	E	R
E9 wt	N	P	S	S	Y	K	V
EHW1683184	S	*	A	G	*	*	*
EAO6252864	S	*	A	G	*	*	*
EBP0523054	S	*	*	G	*	*	*
WP_216691228	*	*	*	*	*	*	*
WP_042969090	S	*	*	G	*	*	*
EET0447945	S	*	*	G	*	*	*
WP_149013229	S	*	*	G	*	*	*
WP_001684089	S	*	*	G	*	*	*
WP_139571364	S	*	*	G	*	*	*
HAY0237856	*	D	*	*	*	*	*
MCD3753904	*	*	*	*	*	*	*
CAA25609	*	*	*	*	*	*	*
EGO0637768	*	*	*	*	*	*	*
WP_078194592	*	*	*	*	*	*	*
EJN0406533	*	*	*	*	*	*	*
WP_136767443	*	*	*	*	*	*	*
PDU80939	*	*	*	*	*	*	*
EFU5689591	*	*	*	*	*	*	*
WP_096845574	*	*	*	*	*	*	*
EFM7227381	*	*	*	*	*	*	*
HBA5420398	*	*	*	*	*	*	*
EFJ1015634	*	*	*	*	*	*	*
MBZ9087339	*	*	*	*	*	*	*
MBW0659285	*	*	*	*	*	*	*
HAW1183305	*	D	*	*	*	*	*
EDD1161141	*	D	*	*	*	*	*
WP_141070407	*	D	*	*	*	*	*

Supplementary Table 3. Natural diversity in ColE interfacial positions. E9 homologs were collected from the nr database using protein BLAST with an 80% sequence identity cutoff. The sequences were clustered using CD-HIT[48] with 99% cutoff. The resulting 28 unique sequences were aligned with MUSCLE[49]. E2-like identities are shown in green and E9-like identities in blue.

Accession #	24	26	27	28	29	32	33	34	38	58
Im2 wt	R	E	G	A	T	D	D	N	R	D
Im9 wt	N	D	T	S	S	E	L	V	T	E
WP_171149985	E	*	*	E	*	*	H	*	A	*
WP_045973231	D	*	*	E	*	*	H	*	D	*
WP_086956852	D	*	*	E	*	*	H	H	*	*
WP_047963342	D	*	*	E	*	*	H	*	*	A
WP_166540812	*	T	*	*	N	*	H	K	*	*
EAR6896980	*	*	*	*	*	*	H	K	*	*
ECN5337130	*	*	*	*	*	*	H	K	*	*
HAS8350052	*	*	*	*	*	*	H	K	*	*
WP_193830442	*	*	*	*	*	*	H	K	*	*
WP_143345855	L	*	*	*	*	*	*	*	D	*
WP_139766971	*	*	*	*	*	*	*	*	*	-
WP_025672426	*	*	*	*	*	*	*	*	*	*
EYX73168	*	*	*	*	*	*	*	*	*	*
BAI39401	*	*	*	*	*	*	*	*	*	*
WP_000861935	*	*	*	*	*	*	*	*	*	*
EHV4529714	*	*	*	*	*	*	*	*	*	*
HAW1068311	*	*	*	*	*	*	*	*	*	*
HAG5762455	*	*	*	*	*	*	*	*	*	*
WP_032326735	*	*	*	*	*	*	*	*	*	*
EGD6078999	*	*	*	*	*	*	*	*	*	*
HAU9627203	*	*	*	*	*	*	*	*	*	*
EGT2301001	*	*	*	*	*	*	*	*	*	*
EGE7374601	*	*	*	*	*	*	*	*	*	*
EFI9877364	*	*	*	*	*	*	*	*	*	*
WP_096167538	*	*	*	*	*	*	*	T	*	*
EET6468166	*	*	*	*	*	*	*	*	*	*
EFL5137489	*	*	*	*	*	*	*	*	*	*

EFJ0587121	*	*	*	*	*	*	*	*	*	*
HAO1928242	*	*	*	*	*	*	*	*	*	*
WP_000861936	*	*	*	*	*	*	*	*	*	*
WP_106642707	*	*	*	*	*	*	*	*	*	*
WP_109043749	*	*	*	*	*	*	E	*	*	*
2WPT_A	*	*	*	*	*	*	*	*	*	*
EDP6088977	*	*	*	*	*	*	*	*	*	*
WP_136770323	*	*	*	*	*	*	*	*	*	*
WP_077108249	*	*	*	*	*	*	*	*	*	*
WP_000420693	*	*	*	*	*	*	*	*	*	*
WP_176563135	*	*	*	*	*	*	*	*	*	*
WP_153443728	*	*	*	*	*	*	*	*	*	*
WP_256488494	*	*	*	*	*	*	*	*	*	*
WP_097455943	*	*	*	*	*	*	*	*	*	*
WP_228106873	*	*	*	*	*	*	*	*	*	*
WP_112861050	*	*	*	*	*	*	*	T	*	*
HAZ3685192	*	*	*	*	*	*	*	*	*	*
WP_000420692	*	*	*	*	*	*	*	*	*	*
WP_035600465	S	*	*	*	*	*	*	D	*	*
ECJ2224819	S	*	*	I	*	*	*	T	*	*
MBA2139178	S	*	*	*	*	*	*	*	*	*
EAW1321402	S	*	*	*	*	*	*	*	*	*
WP_262955291	S	*	*	*	*	*	*	*	*	*
WP_015097263	D	S	A	R	*	*	*	*	L	*
WP_001684088	*	*	*	*	*	*	Q	*	S	*
HCB8098444	*	*	*	*	*	*	Q	*	S	*
WP_042969089	*	*	*	*	*	*	Q	*	S	*
WP_077168007	*	*	*	*	*	*	Q	*	N	*
WP_261232817	*	*	*	*	*	*	*	D	D	*
HBH7052729	*	*	*	*	*	*	*	I	*	K

EFB6209868	*	*	*	*	*	*	*	*	*	*
1BXI_A	*	*	*	*	*	*	*	*	*	*
2GZI_A	*	*	*	*	*	*	*	A	*	*
MBZ9013147	*	*	*	*	*	*	*	*	A	*
AAA23076	*	N	*	*	*	*	*	*	*	*
2GZF_A	*	*	*	*	*	*	*	*	*	*
EET1502589	*	*	*	*	*	*	*	*	*	*
1FR2_A	*	*	*	*	*	*	*	*	*	*
2GYK_A	*	*	*	*	*	*	*	*	*	*
2GZE_A	*	*	*	*	*	*	*	*	*	*
WP_012644887	*	*	*	*	*	*	*	*	*	*
2GZG_A	*	*	*	*	*	*	*	*	*	*
3GKL_C	D	*	A	T	*	*	*	D	*	*
3GJN_A	D	*	A	T	*	*	*	D	*	*
EAA8037545	*	*	*	*	*	*	Q	*	*	*
EDS5484017	*	*	*	*	*	*	Q	I	*	*
EDW8121861	*	*	*	*	*	*	Q	I	*	*
EAB7210983	*	*	*	*	*	*	Q	I	*	*
EAO6252865	*	*	*	*	*	*	Q	I	*	*
ECG8633828	*	*	*	*	*	*	Q	I	*	*
WP_008912668	*	*	*	E	N	*	Q	I	S	*
WP_036952821	*	*	*	*	N	*	Q	I	*	*
WP_004914275	*	*	*	*	N	*	Q	I	*	*
EJD6670373	*	*	*	*	N	*	Q	I	*	*
WP_210838631	*	*	*	*	N	*	Q	I	*	*
WP_123382475	*	*	*	*	*	*	Q	I	*	*
MBO2879508	*	*	*	*	*	*	Q	I	*	*
HCE8607293	*	*	*	*	*	*	Q	I	*	*
WP_198633470	*	*	*	*	*	*	Q	I	*	*

WP_094960452	*	*	*	*	*	*	Q	I	*	*

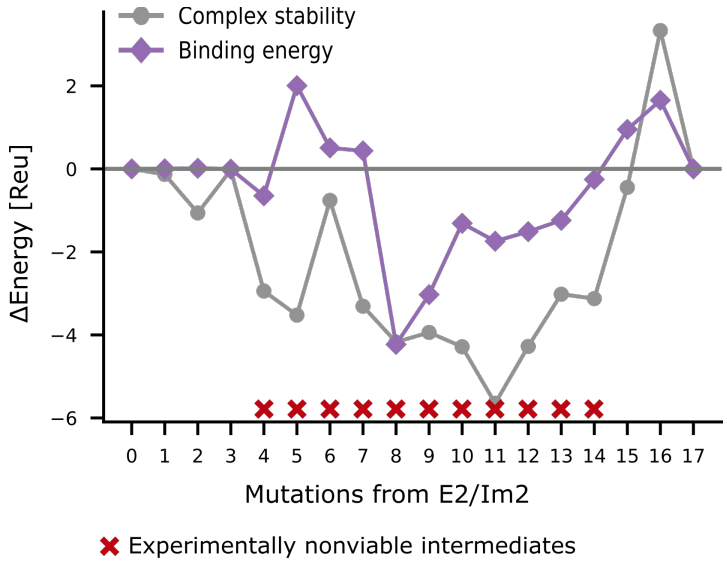
Supplementary Table 4. Natural diversity in Im interfacial positions. Sequence selection and analysis as in Supplemental Table 3, except that here, we used a sequence identity cutoff of 65% to collect Im9 homologs.

Im/E	E2	G73A	R98M	E97K	K83Y	N78S	M98V	T77S	K72N	A73P ($E9_{chimera}$)	E9
Im2	4±0	4.5±0	2.87±0.18	3.75±0.35	5.75±0.35	4.5±0.71	6±0	5.75±0.35	6±0	6±0	6±0
D32E	4±0	4.75±0.35	3±0	3.75±0.35	5.37±0.18	4.25±0.35	6±0	5.75±0.35	6±0	5.75±0.35	5.5±0
T29S	4±0	4.25±0.35	3±0	4.37±0.18	5±0	4.25±0.35	5.75±0.35	5.75±0.35	5.87±0.18	6±0	5.37±0.18
E26D	4±0	4.25±0.35	4±0.35	5±0	5.75±0.35	4.75±0.35	5.87±0.18	5.75±0.35	6±0	6±0	5.37±0.53
N34V	4.4±0.18	4.5±0	1.25±1.06	3±0	4.25±0.35	3.25±0.35	5.25±0.35	5.25±0.35	5.5±0.35	5.5±0	5.5±0.35
R24N	4.4±0.18	4.25±0.35	1.25±0.35	2.5±0	4±0	2.75±0.35	5.375±0.53	5.25±0.35	5.5±0.71	5.25±0.35	5.25±0.35
D33L	6±0	6±0	0.37±0.53	1±1.41	0±0	0±0	0.25±0.35	0.25±0.35	0±0	0.25±0.35	3±0
G27T	6±0	6±0	0±0	0.75±0.35	0±0	0±0	0±0	0±0	0±0	0±0	1.75±0.35
R38T	6±0	6±0	0.75±0.35	0±0	0±0	0±0	0±0	0±0	0±0	0±0	1.5±0
A28S	6±0	6±0	0.75±0.35	0±0	0±0	0±0	0±0	0±0	0±0	0±0	1.5±0
D58E ($Im9_{chimera}$)	6±0	6±0	0.5±0	0±0	0±0	0±0	0±0	0±0	0±0	0±0	0.75±0.35
Im9	6±0	6±0	4±0	3±0	0±0	0±0	2.25±0.35	2±1.41	1.75±1.78	2±1.41	3±0
Naive cells	6±0	6±0	5.25±0.35	6±0	5.875±0.18	5.25±0.35	6±0	6±0	6±0	6±0	6±0

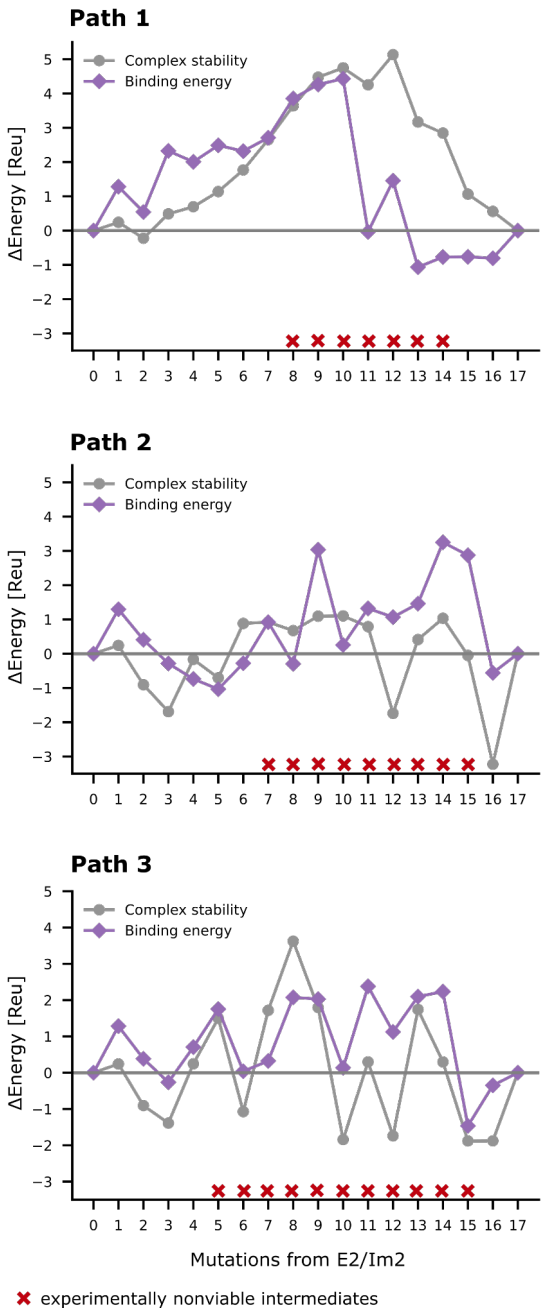
Supplementary Table 5. Average MIC scores and standard deviations of two technical repeats of all ColE mutants on the evolutionary trajectory, applied to all Im proteins (vertical axis). The MIC scores are the number of killing zones (clear zones of cell death) that were observed for each pair (higher numbers translate to lower viability). Each drop corresponds to tenfold dilution of the purified ColE/Im pair. For example, four drops means that the last clear zone was observed at 10^{-3} dilution of the ColE/Im pair. Normalized MIC scores were calculated by subtracting the maximal killing (MIC on naive cells). These averages were used for the calculations of the specificity map in **Fig. 2D**. The mutations from E2/Im2 identity to E9/$Im9_{chimera}$ (E:A73P/Im:D58E) accumulate from the top-left corner to the bottom-right corner as in Figure 2D. A row and a column of wild type Im9 and E9 were added.

Supplementary figures

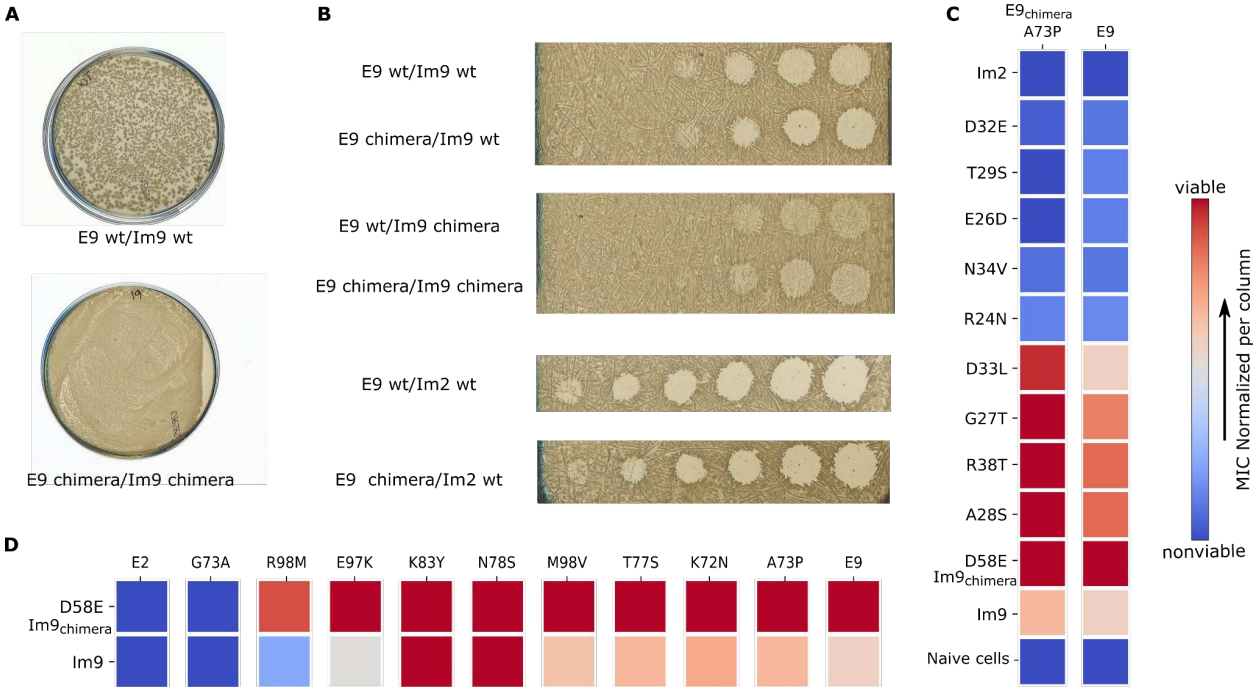
Supplementary Figure 1. Growth experiment for the intermediate of the minimum-strain path. E2/Im2, E9/Im9, negative control (E/Im design that does not grow in BL21 cells), the 18 intermediates and E9/Im9$_{chimera}$ (as marked above each plate) were cloned into BL21 *E.coli* cells and plated on selective agar plates (**Fig. 1E** step III). All clones exhibited similar growth except of the negative control that no growth was inspected.



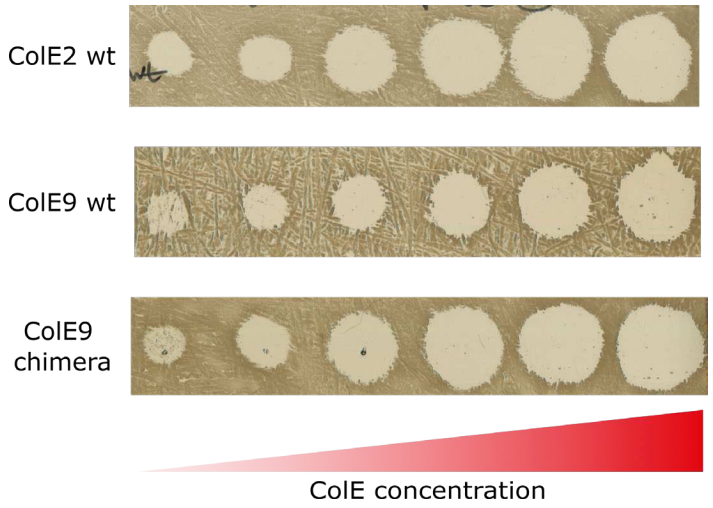
Supplementary Figure 2. The least-strain path was obtained from a graph computed with Rosetta energy function 15 (ref15) [38] and all bridging mutations were eliminated. The difference in complex stability and binding energy to the wild types is shown (*y*-axis*)* for each intermediate from the least-strain path. On the *x*-axis are marked the experimental results for growth in BL21, in which the (x) sign is for intermediates that did not show any growth.



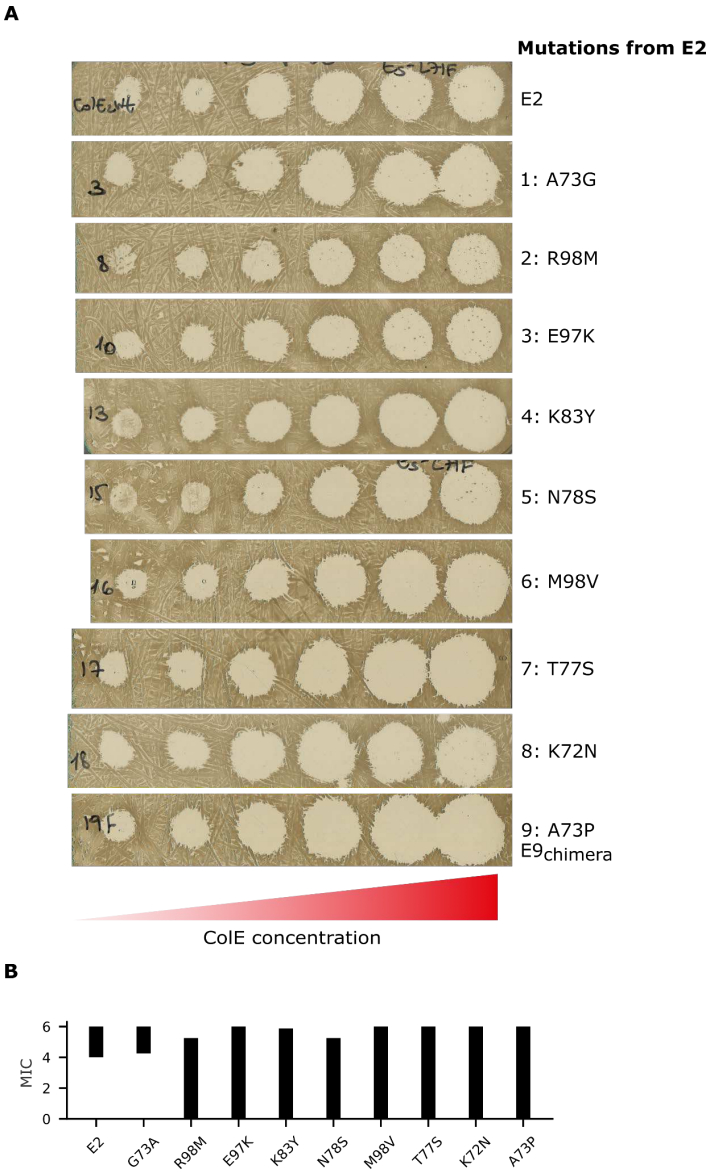
Supplementary Figure 3. Energy difference to wild types (E2/Im2 and E9/Im9) of complex stability and binding energy (*y*-axis*)* is shown for three low-strain paths from the first three trials, that were obtained from graphs computed with an older version of the Rosetta energy function, talaris14[50]. The experimental results for growth on selective agar plates after transformation into BL21 are shown by the red x sign for intermediates that did not show any growth and are hence considered to be nonviable. See Methods for a description of the differences between the three attempts.



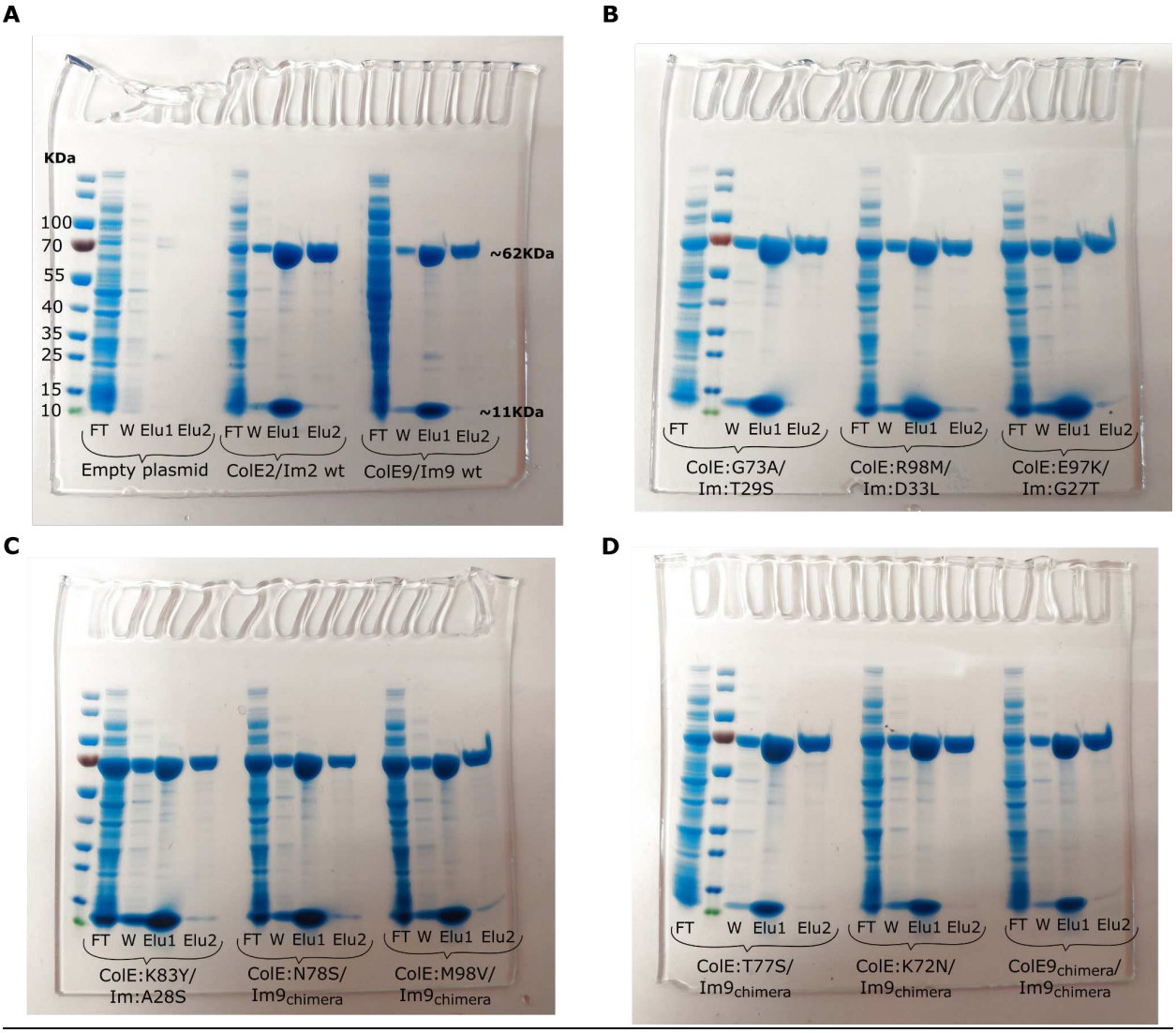
Supplementary Figure 4. Comparison between E9/Im9$_{chimera}$ and the wild-type E9/Im9 pair (A) The two pairs grow in BL21 cells as bacterial lawns. (B) *In vivo* activity testing shows that the two E9 variants induce similar killing levels and three Im variants (Im2, Im9 and Im9$_{chimera}$) provide similar levels of protection. Note that Im9$_{chimera}$ (with interface residues as in Im9 and the rest as in Im2) provides better protection against both E proteins than Im9. (C) E9 and E9$_{chimera}$ induce similar levels of killing towards all Im proteins. (D) Im9 and Im9$_{chimera}$ provide a similar level of protection. Note that Im9$_{chimera}$ provides a higher level of protection than Im9 (similar to the observation in panel B). We also note that Im9$_{chimera}$ exhibits higher stability than Im9 and Im2 as shown in Fig. S7.



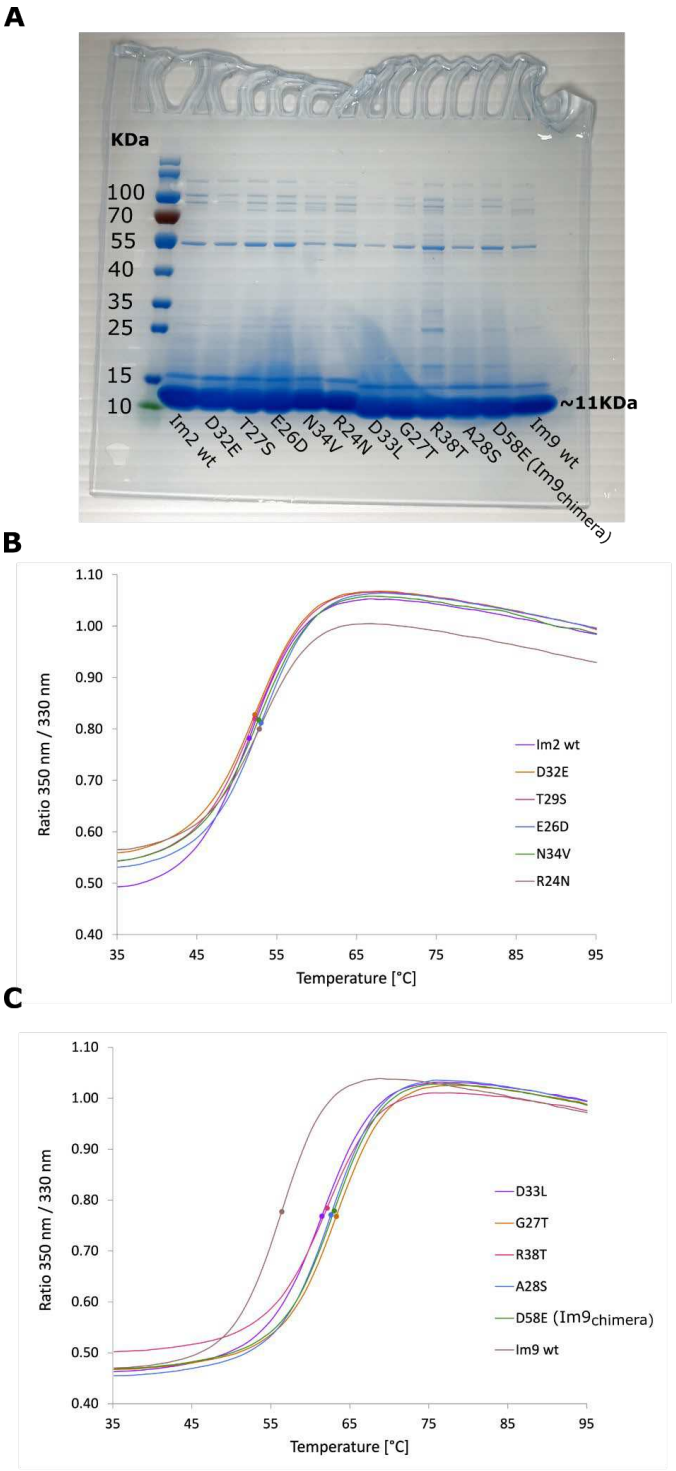
Supplementary Figure 5. In vivo activity of ColE9/Im9$_{chimera}$ compared to the two wild type complexes ColE2/Im2 and ColE9/Im9. The ColE/Im complexes were tested in an in vivo activity assay as explained in the methods. Here, they were applied to naive JM83 cells that do not harbor any Im protein. The ColE/Im complexes were diluted in six ten-fold serial dilutions. Meaning the concentration of the ColE decreases as marked by the gradient at the bottom. The ColEs exert killing in all concentrations that were tested.



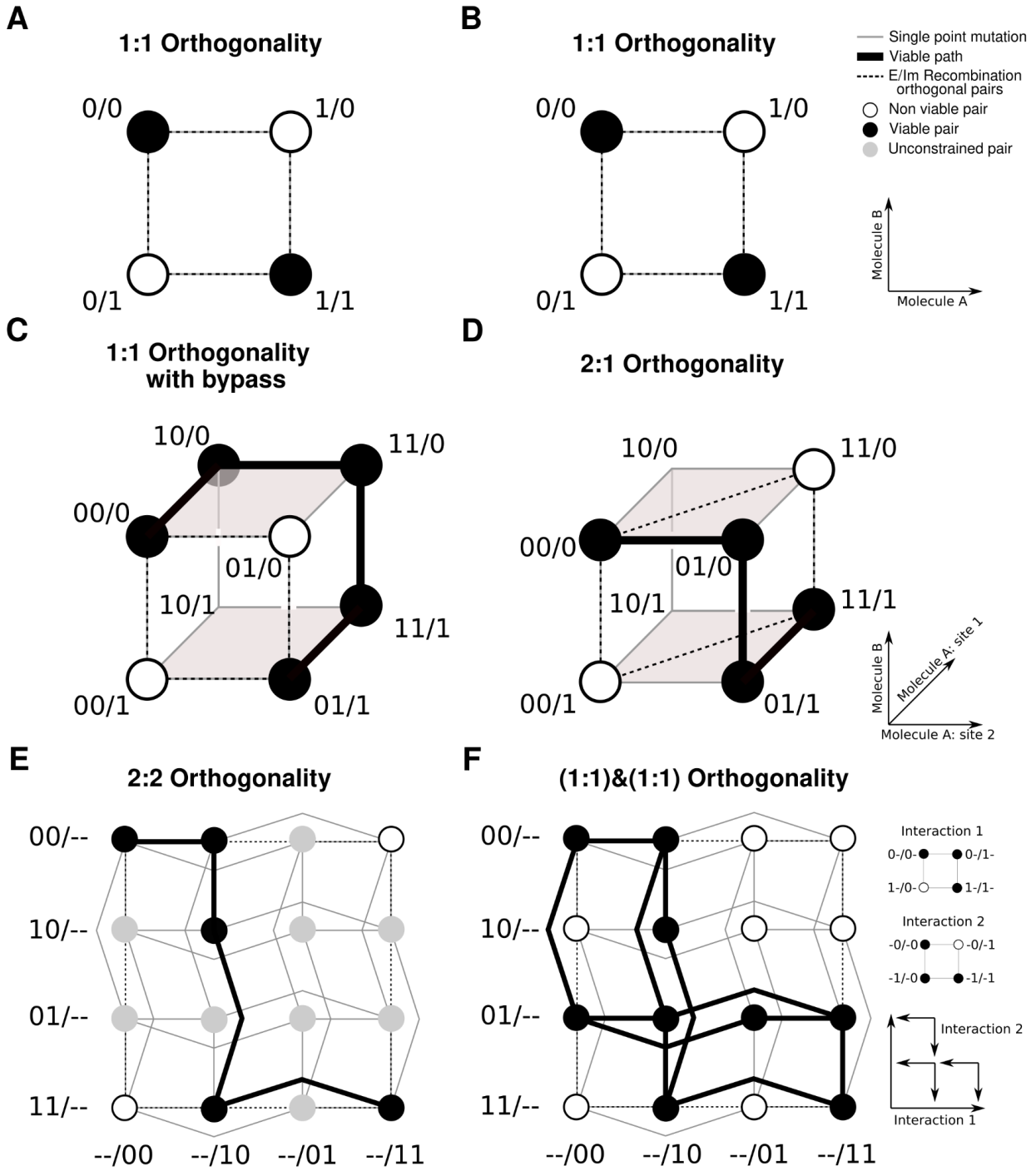
Supplementary Figure 6. (A) The activity of viable path ColEs. The 10 ColEs of the viable path were purified together with their cognate Im. The purified complexes were dropped in six tenfold serial dilutions against JM83 naive cells that do not harbor any Im protein (empty plasmid). All ColEs intermediates retained their toxicity. (B) The killing range of each of the 10 E proteins on the viable path. Maximal killing is the MIC when the E protein is applied to naive cells and the minimal killing is the lowest MIC which was observed when applying the E protein on the Im proteins from the viable path. Wild type E2 and the Gly73Ala mutant exhibit extreme killing efficiencies even against cognate Im2.



Supplementary Figure 7. Quantifying ColE/Im complex expression levels. In each gel there are three samples. FT - flow through from the Ni-NTA column; W - wash step (see Methods); Elu1 - eluted complex of ColE/Im; Elu2 - elution of the ColE alone. An unfolding step was done to separate the ColE from the Im and the flow through was collected. For the *in vivo* assay only the complexes were used to check the full biological activity. (A) Empty plasmid as a control, wild type ColE2/Im2 and wild type E9/Im9. (B-D) intermediate ColE/Im complexes. The expected sizes of the ColE and Im proteins are marked in (A) on the right hand side and the ladder sizes on the left.

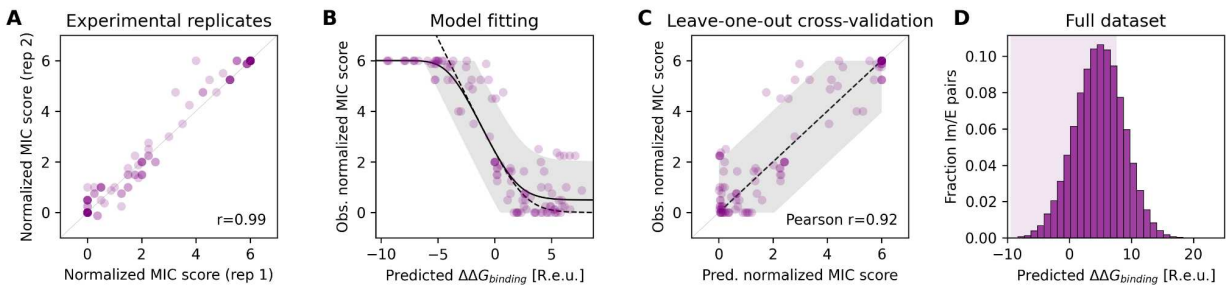


Supplementary Figure 8. Expression levels of Im proteins. (A) Wild type Im2 and Im9, intermediate immunity proteins, and $Im9_{chimera}$ expressed in BL21 cells and purified. All Im proteins exhibit similar expression levels. (B-C) Thermal melting for all Im proteins. All Im proteins exhibit an inflection above 50 °C and an apparent two-state behavior. The sizes of the Im protein is marked in (A) on the right hand side and the ladder sizes on the left.

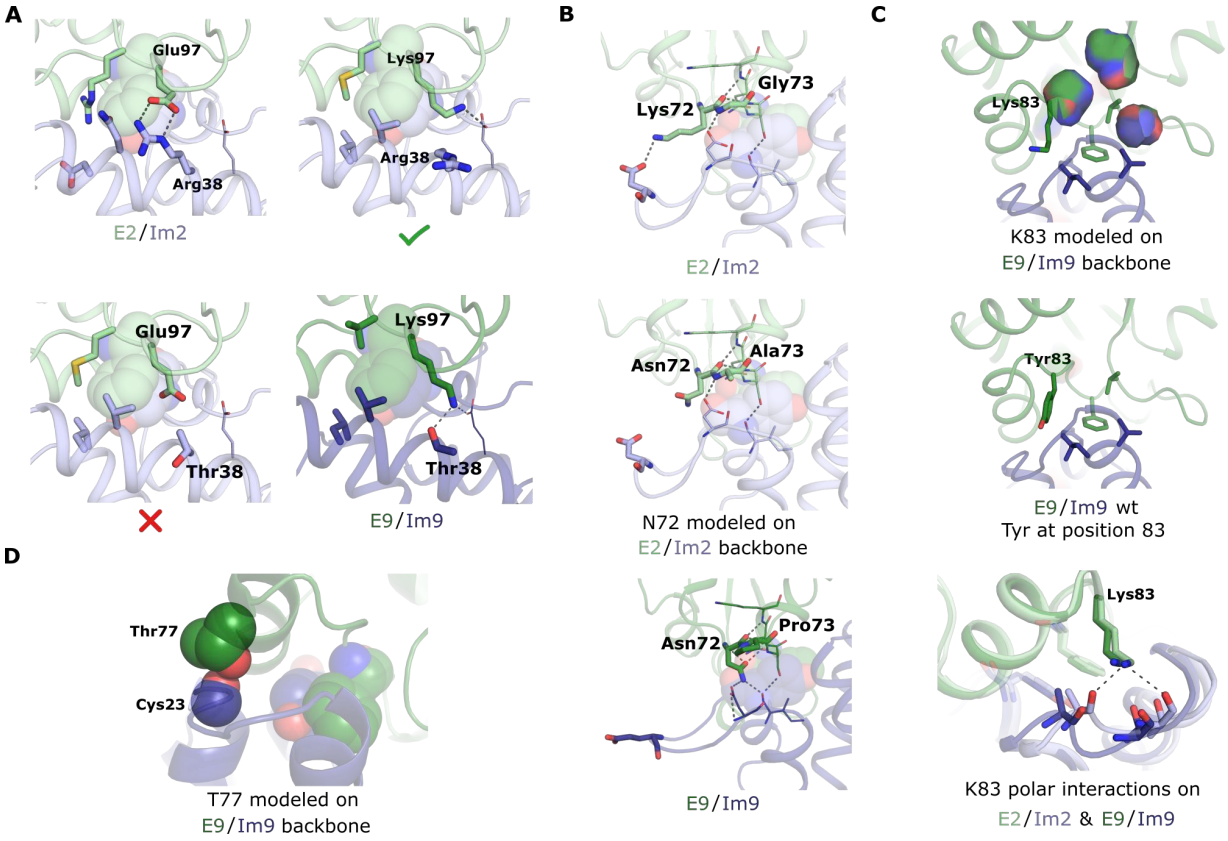


Supplementary Figure 9. Models of orthogonality in fitness landscapes for several distinct *a:b* types, where *a* and *b* are the number of mutations differentiating each of the orthogonal sequences in each molecule. Dotted lines join the four pairs that are involved in the orthogonality, directly reachable through recombination, and characterizing the required "rectangular perturbation" (see

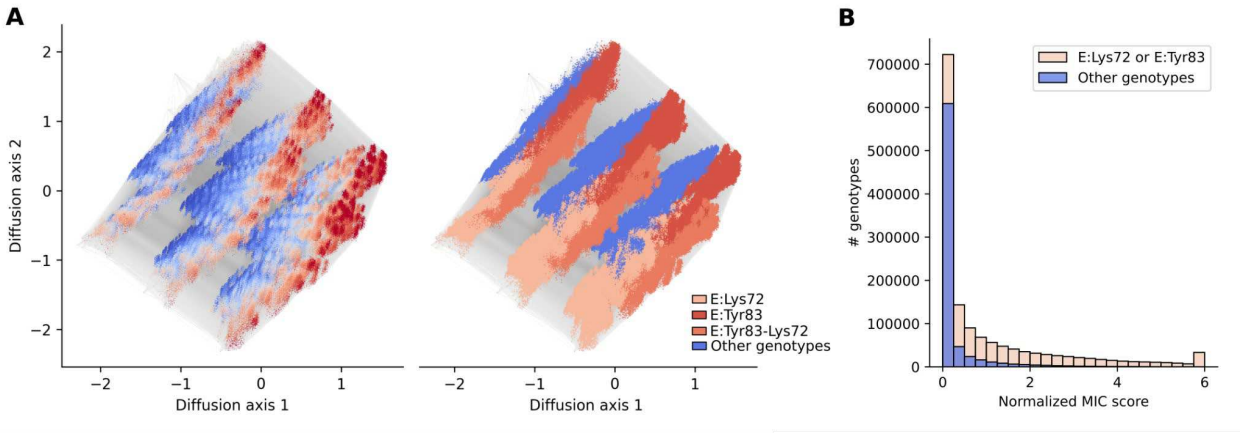
Supplementary Information). Thick black lines represent possible viable paths between the orthogonal pairs in each model. (A) General condition for orthogonality and the genotypes involved for any fitness landscape. (B) Model for 1:1 orthogonality, in which the fitness landscape is fully determined by the orthogonality. (C) 1:1 orthogonality model with a viable path provided by an extra site in molecule 1 that serves as an extradimensional bypass. (D) 2:1 orthogonality model with one of the possible viable paths between the orthogonal pairs. (E) 2:2 orthogonality model showing one possible path between the orthogonal pairs. (F) 2:2 orthogonality model resulting from the combination of two simple 1:1 non orthogonal models with a single incompatibility and the resulting patterns of connectivity of viable pairs. The inset shows how interaction 1 involves the first position from each molecule while interaction 2 involves the second position from each molecule, and illustrates how these two underlying interactions combine to form the set of viable pairs shown in the larger figure



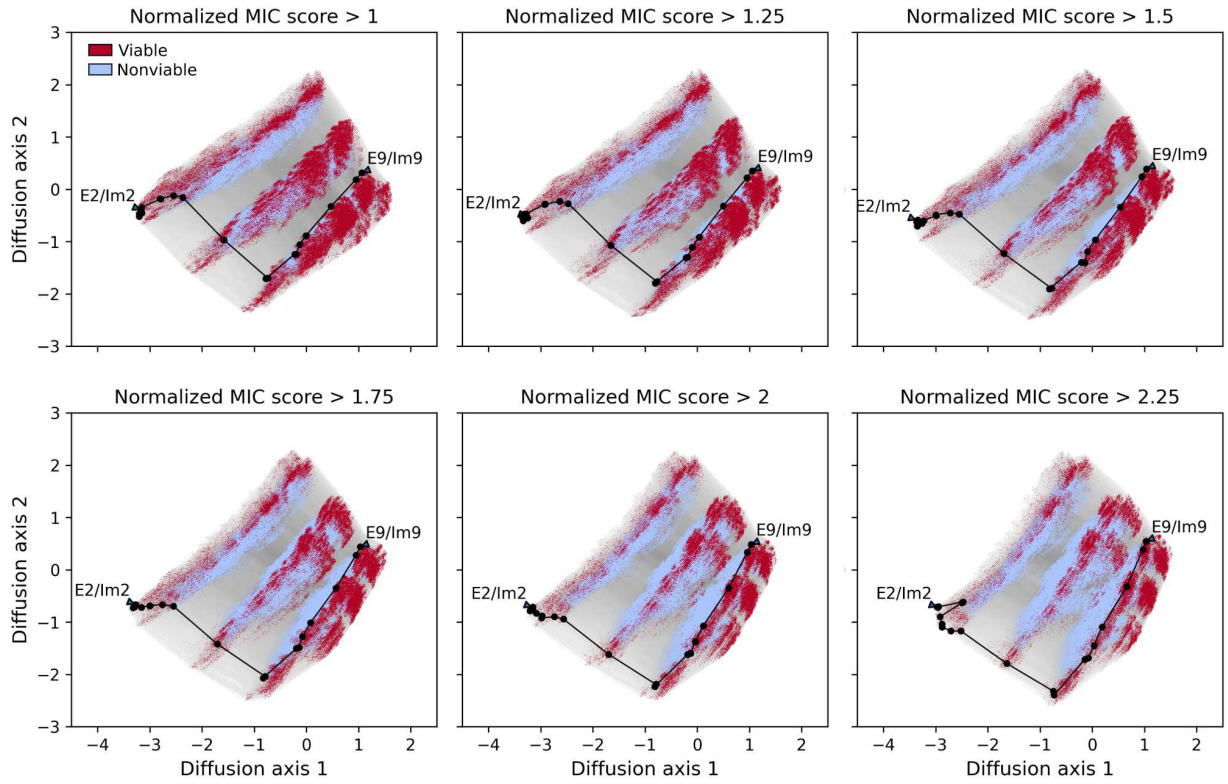
Supplementary Figure 10. Calibration of computational predictions using the specificity map measurements (Fig. 2D). (A) The correspondence between two biological replicates across all Im/E pairs along the experimental path. (B) Experimentally measured normalized MIC scores versus computationally predicted binding energies relative to E2/Im2. The dashed line represents the predictions for the theoretical normalized MIC scores in absence of experimental error. The solid line represents the expected measurement for the normalized MIC score taking into account our noise model and the saturation at a normalized MIC score of 6, corresponding to survival at the greatest tested E concentration. The shaded area represents the interval in which 95% of the measurements are expected to lie under the maximum likelihood estimates (91% of the data lies within those intervals). (C) Scatter plot of the observed normalized MIC scores against the calibrated computational predictions in leave-one-out settings. The shaded area represents the interval in which 95% of the measurements are expected to lie under the maximum likelihood estimates (91% of the data lies within those intervals). (D) Histogram showing the distribution of the adjusted computational predictions of the binding energy relative to E2/Im2 in the full dataset containing all E/Im pairs in the sequence space. The majority of the data lie within the range of energies used for the calibration procedure as shown by the shaded area.



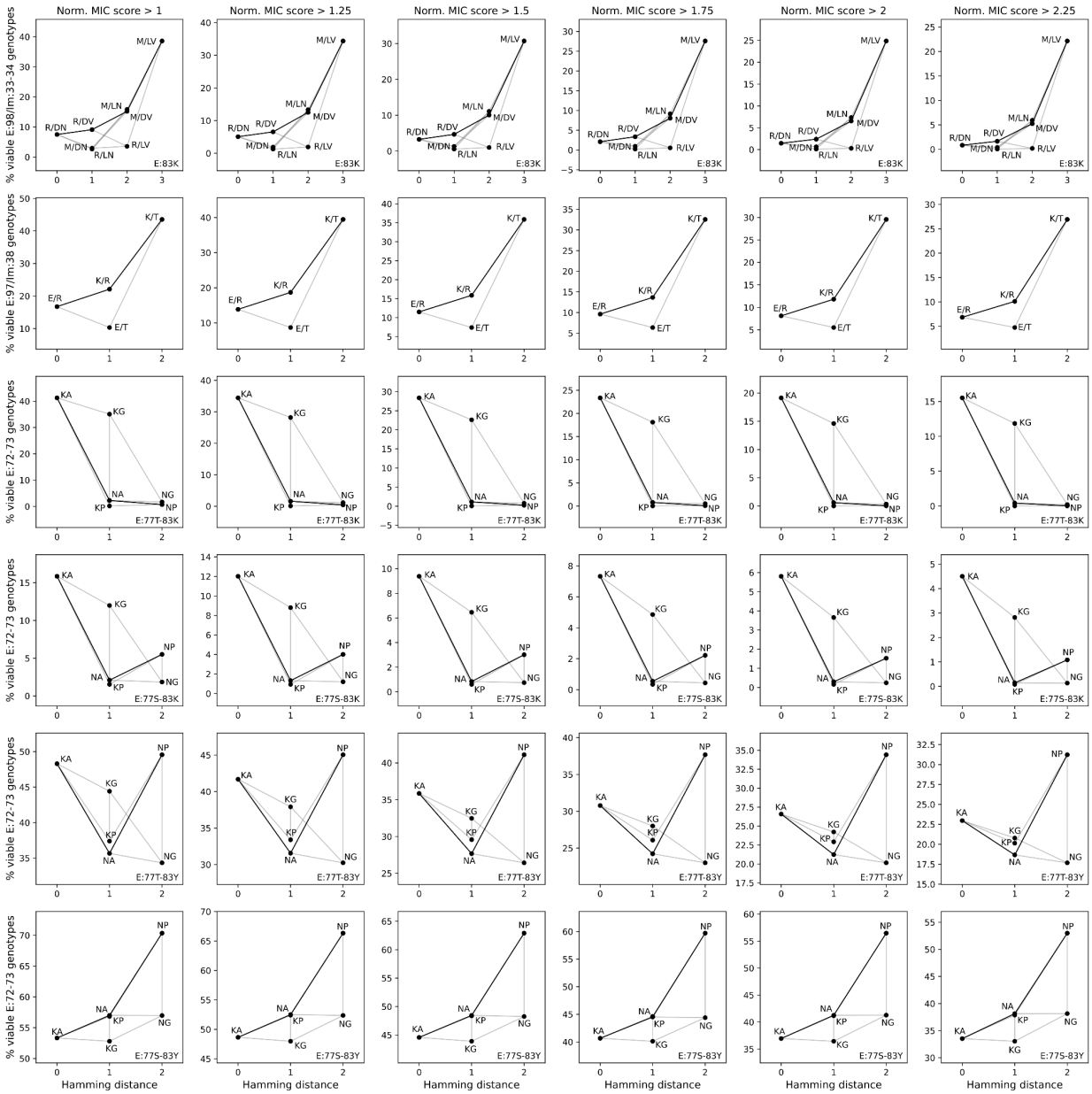
Supplementary Figure 11. Incompatibilities in volume, electrostatics, and backbone conformation separate mutants in the visualization of the fitness landscape (Fig. 3A). (A) The interaction between positions E:97 and Im:38. E2/Im2 harbors a salt bridge (top-left); the viable intermediate (marked with a green checkmark) forms an apolar interaction between hydrophobic groups on the side chains of E:Lys97 and Im:Arg38 with both positive charges facing away (top-right); the nonviable intermediate (marked with a red x) exhibits an unsatisfied and therefore strained E:Glu97; finally, in the wild type interface of E9/Im9, E:Lys97 forms stabilizing polar interactions with surrounding amino acids (bottom-right). (B) E2:Lys72 forms a salt bridge across the interface in E2/Im2. While E9:Asn72 modeled on the backbone of E2/Im2 abrogates a hydrogen-bond network that is observed in the E9/Im9 interface. (C) E2:Lys83 is incompatible with the E9/Im9 backbone, exhibiting interface cavities (shown as surfaces). The E9:Tyr83 fills these cavities. By contrast, E:Lys83 is compatible with the E2/Im2 backbone forming interfacial hydrogen bonds with backbone carbonyls. These bonds cannot form in the E9/Im9 backbone due to the increased distance between the two partners in this state. (D) The E2:Thr77 sidechain modeled in spheres on the E9/Im9 backbone. The rigid and bulky side chain overlaps with the immunity backbone. All models were generated and relaxed using the same atomistic Rosetta protocol used to generate the empirical path.



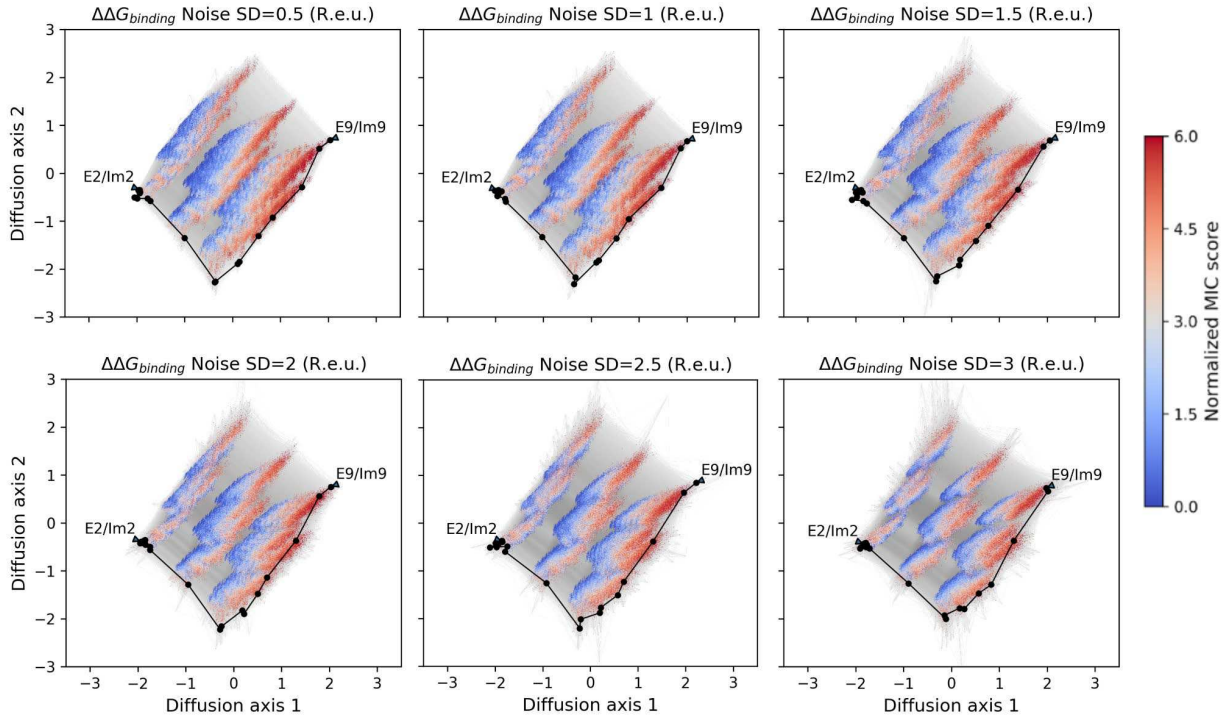
Supplementary Figure 12. Positions E:72 and E:83 define the regions of sequence space containing high fitness sequences. (A) Landscape colored by predicted normalized MIC score (left) and colored by the presence of E:Lys72 and/or E:Tyr83. (B) Histogram of predicted normalized MIC scores for sequences containing E:Lys72 and/or E:Tyr83 shown with histogram for all other sequences. The vast majority of predicted high fitness sequences contain at least one of E:Lys72 or E:Tyr83. Histograms for the two sets of genotypes are overlaid rather than stacked.



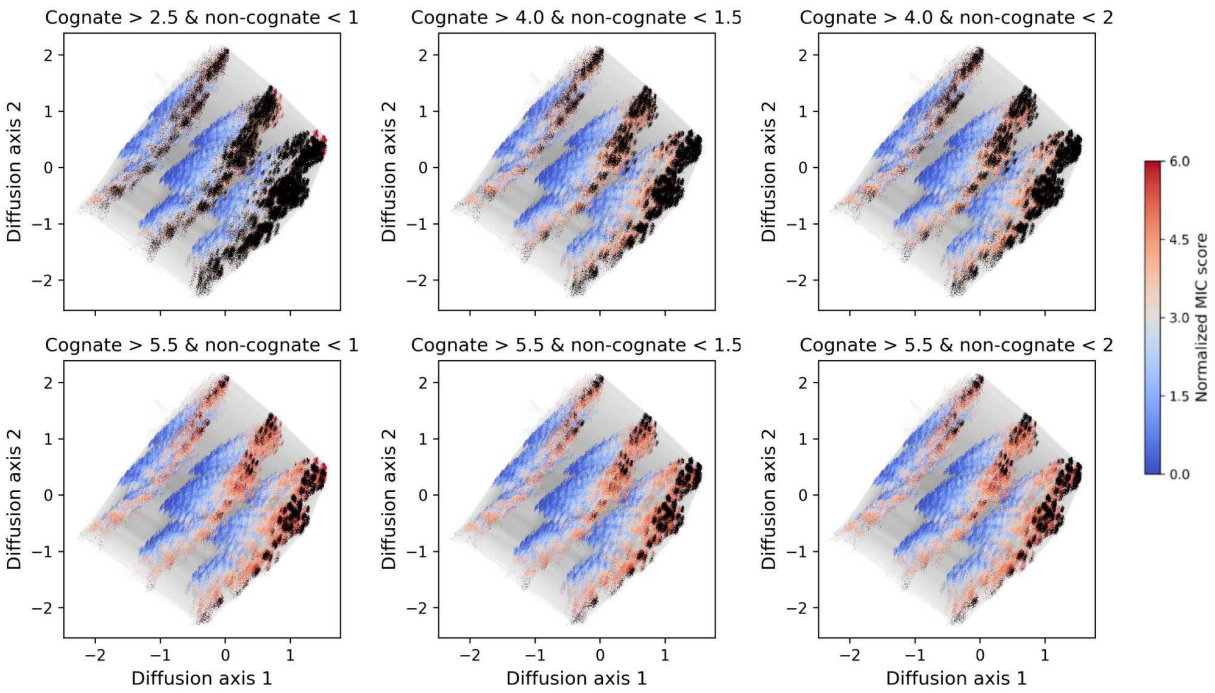
Supplementary Figure 13. Landscape structure is robust to a threshold-based binary fitness function over a wide range of possible cutoffs. Each shows the visualization of a threshold-based binary fitness landscape that arises by considering cell survival in an E toxin concentration given by a range of thresholds (i.e. all cutoffs less than the normalized MIC of the wild-type pairs E2/Im2 and E9/Im9 which have a minimum normalized MIC of approximately 2.5 and greater than the normalized MIC of the non-cognate pairs E9/Im2 and E2/Im9 that have a maximum normalized MIC of approximately 0.9). The experimentally validated path from Fig. 2 is highlighted in black and E2/Im2 and E9/Im9 wild-type positions are indicated with arrows. Parameter c for the visualizations was chosen to have an odds ratio of 9 when comparing the fraction of time that the population spends at viable pairs under purifying selection with that under neutral evolution.



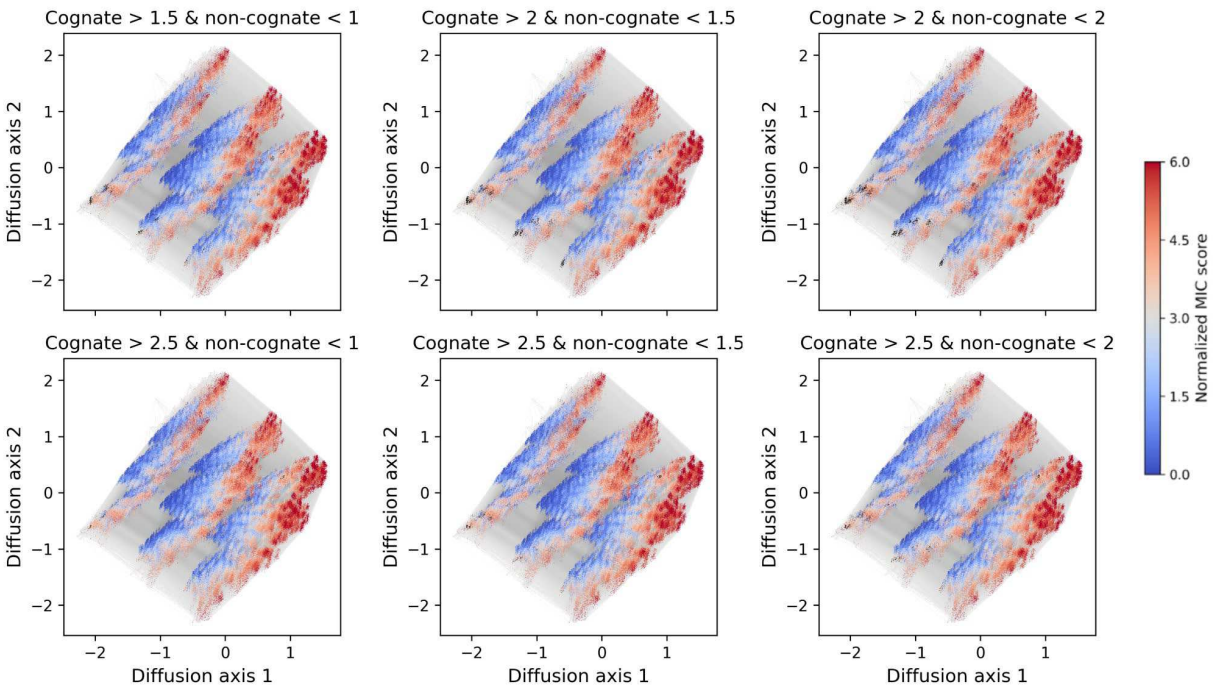
Supplementary Figure 14. Genetic interactions are robust to a threshold-based binary fitness function over a wide range of possible cutoffs. Each panel represents a sub-landscape obtained by averaging out positions not involved in the specific genetic interaction (shown along rows) and in the specified sequence context (shown on the bottom right corner of each panel, if any) as in Figure 3. The columns correspond to different tested thresholds along the range of reasonable viability cutoffs (*i.e.,* all cutoffs less than the normalized MIC of the wild-type pairs E2/Im2 and E9/Im9 which have a minimum normalized MIC of approximately 2.5 and greater than the normalized MIC of the non-cognate pairs E9/Im2 and E2/Im9 that have a maximum normalized MIC of approximately 0.9). Thick black lines represent the mutations that were included in the experimentally validated path.



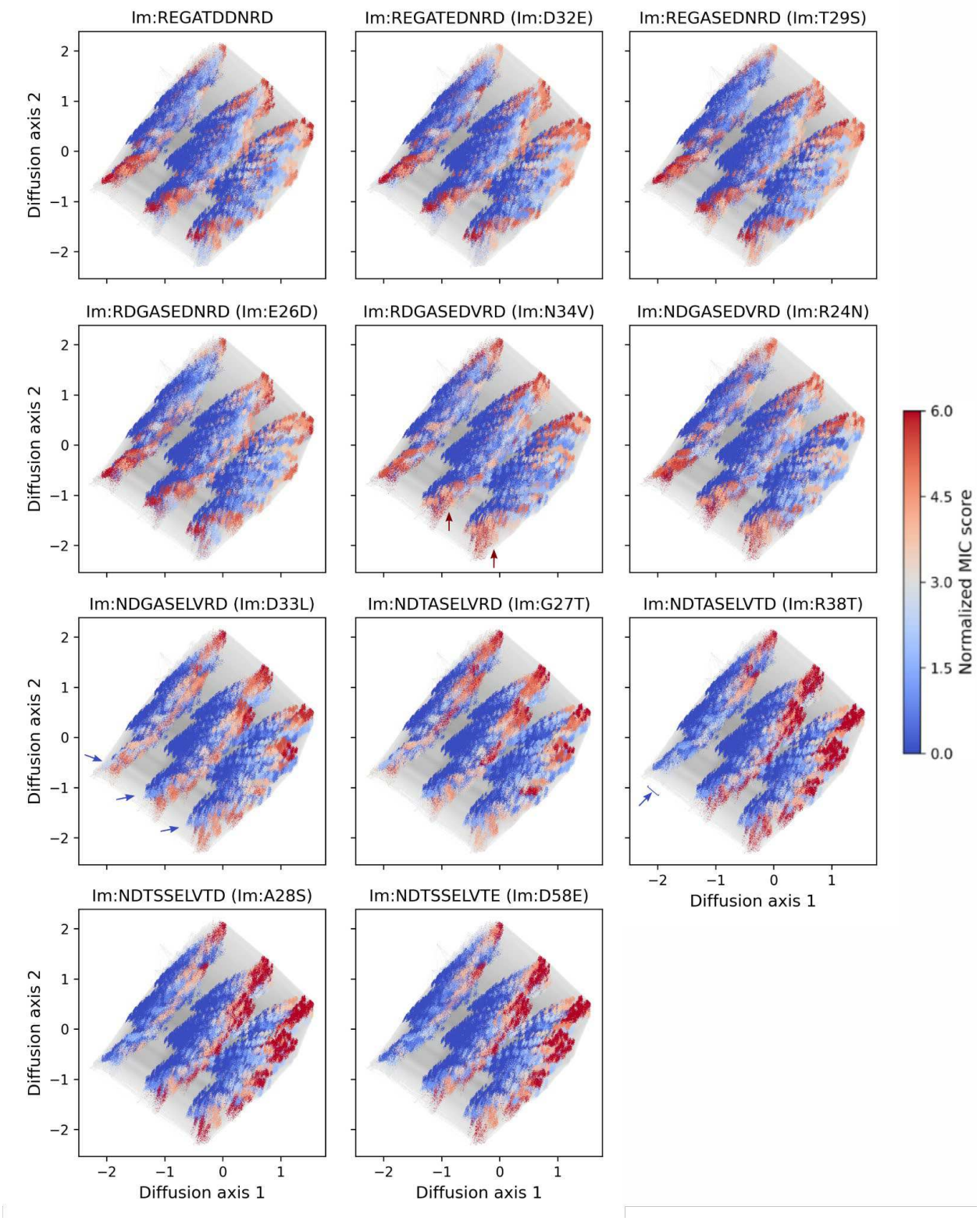
Supplementary Figure 15. Landscape structure is robust to noise in the Rosetta binding energies. Each shows the visualization of a landscape generated by adding a normally distributed error with standard deviations ranging from 0.5 to 3 R.e.u. to the predicted binding energies. The experimentally validated path from Fig. 2 is highlighted in black and E2/Im2 and E9/Im9 wild-type positions are indicated with arrows. To account for changes in the distribution of Normalized MIC scores after introducing noise in the binding energies when computing the coordinates of the visualization, parameter *c* was chosen to have an increase of 3.5 fold when comparing the ratio of mean(Normalized MIC score) / (6 - mean(Normalized MIC score)) in the stationary phase for a population evolving under purifying selection with that under neutral evolution.



Supplementary Figure 16. Orthogonal pairs to E2/Im2 are robust to different cutoffs for the cognate and non-cognate interactions. Each panel represents the low dimensional visualization of the binding landscapes with genotypes that are orthogonal to E2/Im2 overlaid in black, where orthogonality is defined based on the indicated thresholds on the normalized MIC score for cognate and non-cognate interactions. Note that the number of sequences that are orthogonal to E2/Im2 decrease as the thresholds for viability and non-viability change, but the orthogonal pairs are always concentrated in the same region regardless of the threshold used.

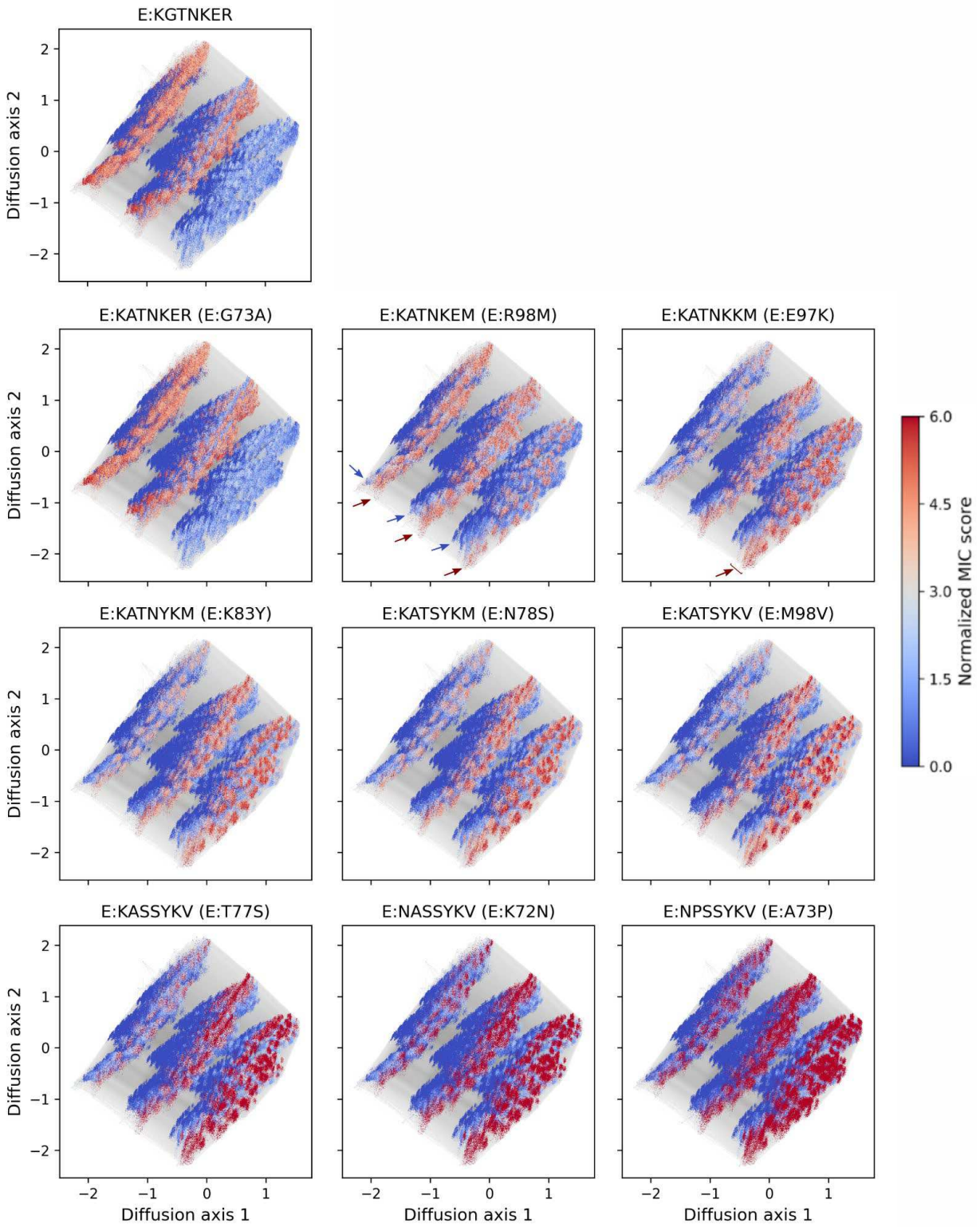


Supplementary Figure 17. Orthogonal pairs to E9/Im9 are robust to different cutoffs for the cognate and non-cognate interactions. Each panel represents the low dimensional visualization of the binding landscapes with genotypes that are orthogonal to E9/Im9 overlaid in black, where orthogonality is defined based on the indicated thresholds on the normalized MIC score for cognate and non-cognate interactions. Note that the number of sequences that are orthogonal to E9/Im9 decrease as the thresholds for viability and non-viability change, but the orthogonal pairs are always concentrated in the same region regardless of the threshold used.



Supplementary Figure 18. Predicted patterns of Im specificity along the experimentally validated viable path between Im2 and Im9. The color scale shows the predicted normalized MIC scores of the Im protein variants along the path when combined with the E subsequence of

every E/Im pair in the landscape. Note that the specificities remain mostly unchanged after each mutation except the key specificity changing mutations: Asn34Val, Asp33Leu and Arg38Thr. Arrows point to regions of sequence space for which the mutation increases (red) or decreases (blue) the normalized MIC score when combined with the E sequence of every pair. To avoid showing interactions with the E proteins of inviable pairs, the score for the interaction with the E protein of any pair with a normalized MIC score below 1 was set to 0.



Supplementary Figure 19. Predicted patterns of E specificity along the experimentally validated viable path between E2 and E9. The color scale shows the predicted normalized MIC scores of the E protein variants along the path when combined with the Im subsequence of every

E/Im pair in the landscape. Specificities remain mostly unchanged after each mutation except the key specificity changing mutations: Arg98Met and Glu97Lys. Arrows point to regions of sequence space for which the mutation increases (red) or decreases (blue) the normalized MIC score when combined with the E sequence of every pair. To avoid showing interactions with the Im protein of inviable pairs the score for the interaction with the Im protein of any pair with a normalized MIC score below 1 was set to 0.

Mathematical theory of functional orthogonality in bi-molecular fitness landscapes

In the following section, we aim to formalize the concepts of specificity and orthogonality. Previous work has focused on studying the effects of mutations on cognate and non cognate interactions [51], but here we study these concepts from the more general fitness landscape perspective in order to understand the necessary and sufficient conditions for functional orthogonality to arise in a binding landscape. Moreover, these simple models of orthogonality-encoding landscapes lead to testable predictions about the fitness of unobserved sequences and to a better understanding of the basic underlying molecular mechanisms based on the required patterns of genetic interaction

A pair of molecules A and B can be expressed by the concatenation of the sequences encoding each of the two molecules as a combined genotype A/B, such that the resulting sequence space of all possible pairs is obtained by combining each possible A variant with every B. As every molecule consists of at least 1 site, the sequence of a pair necessarily has length $l \geq 2$. We define the fitness of each pair A_i/B_j as a binary outcome (viable/nonviable) depending on whether the A_i and B_j variants can interact with each other. The specificity of A_i is defined by the set of B_j's such that A_i/B_j is viable and the specificity of B_j is defined as the set of A_i's such that A_i/B_j is viable. Therefore, two variants of A, A_i and A_j, can have different binding specificities depending on the subsets of B molecules with which they interact. In the case that these subsets of B differ, we say that there is a specificity switch along the series of mutations separating A_i from A_j. The existence of such a specificity switch implies that the effect of a mutation in one of the partners A depends on the alleles at the other partner B; that is, epistatic interactions across the binding interface are required for having different specificities in a binding landscape.

Two pairs of molecules A_i/B_j and A_k/B_m, are defined to be orthogonal when both pairs are viable but B_j cannot bind A_k and B_m cannot bind A_i, or in other words, both pairs are viable but their recombinant genotypes (A_i/B_m and A_k/B_j) are nonviable (**Fig. S9A**), a pattern of epistasis that has previously been referred to as a type of "rectangular perturbation"[52]. Importantly, the rectangular perturbation has to involve sites at both sides of the interface. Therefore, regardless of the number of possible combinations of mutations that differ between the two pairs, orthogonality depends only on the fitness of four very specific genotypes. These genotypes are contained within the biallelic landscape defined by all possible allelic combinations present in the orthogonal pairs A_i/B_j and A_k/B_m. That is, we can disregard alleles not present in these two pairs as well as sites that match in the two orthogonal pairs and see how orthogonality arises based only on the number of sites, *a* and *b,* that differ in each of the molecules A and B, respectively, into different *a*:*b* mechanisms or types of orthogonality.

While orthogonality only constrains the fitness values of four possible genotypes, we often study homologous orthogonal pairs. Thus, if we assume that the binding of the two ancestral molecules was required for survival, we expect that the two extant orthogonal genotypes are additionally connected by a path of viable intermediate pairs, imposing additional constraints on the viability

of certain genotypes. Since we are only considering the biallelic sublandscape defining orthogonality, we can characterize the number of plausibly viable shortest paths between orthogonal pairs by counting first the number of nonviable paths, *i.e.*, that go through the nonviable pairs and subtract that number from the total number $(a+b)!$ of possible paths (and $b!$ indicates b factorial). The number of nonviable paths can be easily calculated as the product of the number of paths from each nonviable pair to each of the orthogonal pairs. As the nonviable pairs result from recombination of the orthogonal pairs, we can reach them by mutating all sites in each molecule separately, so they are exactly a and b mutations away from each orthogonal pair. Thus, the total number of shortest paths between the orthogonal pairs that go through nonviable pairs is 2 times $a!b!$, and thus the number of viable paths is $(a+b)!-2a!b!$.

In the following section, we systematically analyze the fitness landscapes under various *a:b* types of orthogonality and the different paths that connect the two orthogonal pairs.

1:1 Orthogonality. In the simplest possible scenario, two orthogonal sequences differ only in one site in each of the molecules. In these settings, the fitness landscape is completely specified by the orthogonality condition directly (0/0 and 1/1 are viable and their recombinants 0/1 and 1/0 are nonviable) and requires reciprocal sign epistasis. However, there is no possible viable path between the two orthogonal pairs (**Fig. S9B**). This matches our expectations of 1+1! possible paths with 2 times 1!1! paths going through the non viable pairs. However, adding an allele or a site to a 1:1 orthogonal landscape can provide a new sequence context that is no longer constrained by the condition for orthogonality and in which a viable path can exist via a mechanism sometimes known as an extradimensional bypass[53]. The addition of a second allele allows the viable path 0/0-0/2-1/2-1/1 through a bridging mutation, while adding a third site provides the viable path 0/00-0/01-1/01-1/11-1/10 with a reversion (**Fig. S9C**). Thus, adding an extra site or allele is the only way for evolving 1:1 orthogonality, but can, in general, provide a way to evolve any *a*:*b* model of orthogonality.

2:1 Orthogonality. In the next simplest case, the two orthogonal genotypes differ by a=2 sites in one of the molecules and b=1 site in the other molecule, resulting in a landscape encoding 2:1 orthogonality. In this case, 00/0 and 11/1 are constrained to be viable while their recombinants 11/0 and 00/1 are necessarily nonviable. Note that, in contrast to 1:1 orthogonality, the recombinants can be at a larger Hamming distance from the orthogonal pairs; *e.g.*, 00/1 is two mutations away from 11/1. Despite these constraints, there are now four unconstrained genotypes that can potentially provide a viable path between the two orthogonal pairs that avoids the nonviable ones *e.g.* 00/0-01/0-01/1-11/1 (**Fig. S9D**). More generally, out of the 3!=6 possible paths, $2 \cdot 2!1! = 4$ go through nonviable genotypes, leaving only two plausible viable paths between the orthogonal pairs. Both of these paths require the mutation in the single-site molecule to happen between the mutations in the two-site molecule, and thus these paths only differ by which of the mutations in the two-site molecule takes place first.

There are two main implications of this mechanism: (1) We show below that this mechanism

requires epistatic interactions between the mutation in the one site molecule and each of the sites in the two-site molecule. Therefore, we expect the two amino acid positions in the two-site molecule to be located close to each other in the three-dimensional structure, since both must interact with the same site in the one-site molecule. (2) Mutations in the viable path take place in a very specific order, since mutating first the single-site molecule will directly take us to one of the nonviable genotypes. Therefore, specificity switching mutations in the two-site molecule need to take place first and last in the path before and after mutation in the single-site molecule.

3:1 Orthogonality. A way to keep expanding our previous model is by adding another site to the two-site molecule, constraining 000/0 and 111/1 to be viable and 000/1 and 111/0 to be nonviable. The number of plausible viable paths between the orthogonal genotypes keeps increasing as there are more unconstrained pairs in the landscape, in the sense that their fitness can be freely assigned without affecting orthogonality. The number of nonviable paths is $2 \cdot 3!1!=12$ in 3:1 orthogonality leaving 12 potentially viable paths. As in 2:1 orthogonality, or more generally, any mechanism of orthogonality with a single site in one of the molecules, the orthogonality constraint implies that the one-site mutation cannot be the first or last mutation to take place in a viable path, so that the mutation in the one-site molecular will always be flanked by mutations in the multi-site molecule both before and after.

2:2 Orthogonality. While the total Hamming distance between the two orthogonal pairs is the same as in the 3:1 orthogonality mechanism, the pairs with constrained fitness values are arranged differently in sequence space (00/00 and 11/11 as viable, and 00/11 and 11/00 as nonviable). As a result, the number of possible viable paths between the orthogonal pairs are increased (**Fig. S9E**) with a total of $(2+2)!-2 \cdot 2!2!=16$ potentially viable paths. However, having for the first time more than one site in each molecule opens the possibility of creating orthogonal landscapes by a simple combination of non-orthogonal landscapes at subsets of sites through a boolean function. In particular, suppose we have two specificity switching (but not orthogonal) biallelic landscapes consisting of 3 viable combinations and only one nonviable intermediate (either 0/1 or 1/0) forming a 1:1 incompatibility. If we define the viable pairs as those requiring both subsets to be viable (AND function) and specifically select 0/1 and 1/0 as non-viable in the two interactions, we obtain a four-site 2:2 orthogonal fitness landscape in a type of orthogonality that we called (1:1)&(1:1). Importantly, 00/11 and 11/00 are non-viable, as each of them contains one of the non-viable intermediates in each subset of sites, or what is equivalent, containing one of the specific alleles in one of the 1:1 incompatibilities. Moreover, the orthogonal pairs (00/00 and 11/11) are connected, not only by a single path, but they occupy the corners of a 3x3 grid of viable pairs providing 6 different viable paths between the orthogonal pairs (**Fig. S9F**).

Here, we aimed to describe and characterize some of the simpler models of orthogonality, but mechanisms involving a larger number of sites are certainly possible. As shown for the 2:2 model, a natural and parsimonious way in which orthogonality involving a larger number of sites appears is by combining smaller intermolecular interactions or incompatibilities that do not necessarily encode orthogonality independently. This phenomenon generates orthogonality through the AND

function, which makes the recombinant pairs that contain only a subset of the required interactions nonviable. Another simple way is by having additional neutral sites differing between the orthogonal pairs. This would increase the number of pairs that are orthogonal to each pair as well as the viable paths between them, as the number of contexts in which specificity switching mutations can happen would increase. Thus, even if two orthogonal pairs differ in a relatively large number of sites in their interface, if there are only a few specificity-switching mutations, these simple mechanisms or combinations of them can realistically explain the evolution of orthogonality and the geometry of the fitness landscape. For example, in the empirical path, the closest orthogonal pairs differ by 4 mutations, but it corresponds to a 2:1 orthogonality because one of the mutations (Im:Arg24Asn) is largely neutral.

Moreover, and as orthogonal landscapes involve more sites, different mechanisms driving orthogonality could coexist, *e.g.*, if instead of combining two intermolecular biallelic interactions as in the (1:1)&(1:1) model (**Fig. S9F**), we replace one of those interactions by a 2:1 orthogonality landscape, the resulting (2:1)&(1:1) landscape shows both 2:1 orthogonality within one set of sites, but could simultaneously show 2:2 orthogonality through a (1:1)&(1:1) mechanism involving only two of the 3 sites required for the 2:1 orthogonality. As a consequence, under selection for orthogonality to a particular pair in a fitness landscape with coexisting modes of orthogonality *e.g.* (2:1)&(1:1), orthogonality could first arise via one mechanism, *e.g.*, 2:1, which requires fewer mutations, but be maintained at some point via another mechanism, *e.g.* 2:2, without ever losing orthogonality. In fact, the predicted fitness landscape (**Fig. 3A**) is largely consistent with a (2:1)&(1:1) orthogonality where the subsets of interacting sites are E:98/Im:33-34 and E:97/Im:38. The empirical path switches specificity through the 2:1 orthogonality mechanism alone, but, after resolving the second specificity-changing interaction E:97/Im:38, the Im:Asp33Leu mutation may be reversed without losing orthogonality to E2/Im2 (though with potential loss in affinity and stability), because orthogonality is predicted to be maintained via a 2:2 mechanism involving E:98/Im:34 and E:97/Im:38. Another simple class of models with coexisting modes of orthogonality can arise from combining more than 2 intermolecular interactions in a more complex way *e.g.* requiring at least 2 viable interactions such as in a (1:1)&((1:1)|(1:1)) model, where "|" indicates logical OR. In this case, we have 3 intermolecular interactions but 2 of them are sufficient for the interaction to happen. Then, we can easily see how orthogonality can arise first via 4 mutations affecting the 2 first interactions, but after mutations involving the 3rd interaction take place, the mutations at the second set of sites can be reverted without losing orthogonality to the original pair.

Epistasis in orthogonal landscapes. In this section we study how the rectangular interaction imposed by the orthogonality constraint and the viable path requirement impact local double and triple mutant epistatic coefficients in the fitness landscape. We use the definition of background-dependent epistatic coefficient with our binary fitness values:

$$\varepsilon_2 = (f_{11} - f_{01}) - (f_{10} - f_{00}).$$

Thus for the 1:1 orthogonality model we have $\varepsilon_2 = 2$. If we allow a viable path through a bypass mutation at a third bi-allelic site, it creates an opportunity for a 3-way interaction, which corresponds to the difference in pairwise epistatic coefficients between parallel faces, giving rise to the local three-way epistatic coefficient:

$$\varepsilon_3 = \Big((f_{11/1} - f_{11/0}) - (f_{10/1} - f_{10/0})\Big) - \Big((f_{01/1} - f_{01/0}) - (f_{00/1} - f_{00/0})\Big).$$

Thus, we can ask whether a 3-way interaction is necessary to have a viable path between the orthogonal pairs. Under the 1:1 orthogonality mechanism with a bypass mutation, there is a single unconstrained genotype: 10/1 or 11/0 depending on the viable path we choose; so we can easily calculate the 3-way epistatic coefficient as a function of this gentoype's fitness:

$$\varepsilon_3 = \Big((1 - 1) - \big(f_{10/1} - 1\big)\Big) - \big((1 - 0) - (0 - 1)\big) = -1 - f_{10/1} < 0\,.$$

Since $f_{10/1} \in \{0, 1\}$, the 3-way epistatic coefficient must be negative, and therefore a 3-way interaction is required to allow a viable path between the orthogonal pairs through a 3rd site bypass.

The 2:1 orthogonality model with viable path constrains the fitnesses of 6 genotypes in the landscape spanning two adjacent faces of the cube, and therefore establishes a requirement on the pairwise epistatic coefficients of each of them. Specifically, it implies that the one site molecule establishes a pairwise interaction with each site at the other molecule with pairwise epistatic coefficients of opposite signs.

$$(f_{11/1} - f_{11/0}) - (f_{01/1} - f_{01/0}) = 1$$

$$(f_{00/1} - f_{00/0}) - (f_{01/1} - f_{01/0}) = -1$$

We next ask if 2:1 orthogonality requires a 3-way interaction. As before, we calculate the 3-way epistatic coefficient, which now depends on the fitnesses of the 2 remaining unconstrained genotypes:

$$\varepsilon_3 = \Big((1 - 0) - \big(f_{10/1} - f_{10/0}\big)\Big) - \big((1 - 1) - (0 - 1)\big) =$$

$$= f_{10/0} - f_{10/1}$$

Thus, we can see that 2:1 orthogonality is compatible with a landscape without 3-way interactions as long as mutating the one-site molecule is neutral when paired with the 10 background in the two site molecule $\big(f_{10/0} = f_{10/1}\big)$. That still leaves open the possibility of an *intra*molecular pairwise interaction, which would happen if $f_{10/0} = f_{10/1} = 1$ but not if $f_{10/0} = f_{10/1} = 0$. The former allows having two viable paths between the orthogonal pairs, while the latter minimizes the number of required epistatic interactions to have 2:1 orthogonality with a single viable path. Thus, the 2:1 orthogonality mechanism requires a pair of inter-molecular pairwise epistatic

interactions but is also compatible with the presence of an intramolecular pairwise epistatic interaction in the two-site molecule or the presence of a three-way epistatic interaction across both molecules.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation1223.docx