Entropy tree networks of residue dynamics encode protein allostery

Kaitlin Trenfield[1,2,3] and Milo M. Lin[1,2,4,5*]

[1]Green Ctr. for Systems Biology, University of Texas Southwestern Medical Ctr., Dallas, TX, USA.

[2]Lyda Hill Dept. of Bioinformatics, University of Texas Southwestern Medical Ctr.

[3]Molecular Biophysics Graduate Program, University of Texas Southwestern Medical Ctr.

[4]Dept. of Biophysics, University of Texas Southwestern Medical Ctr.

[5]Ctr. for Alzheimer's and Neurodegenerative Diseases, University of Texas Southwestern Medical Ctr.

*Corresponding author. Email: Milo.Lin@UTSouthwestern.edu

Abstract

Proteins can sense signals and–in a process called allostery–transmit information to distant sites. Such information is often not encoded by a protein's average structure, but rather by its dynamics in a way that remains unclear. We show that maximum information tree networks learned from microseconds-long molecular dynamics simulations of a PDZ domain and the entire human steroid receptor family provide mechanistically-detailed maps of information transmission within proteins in a ligand- and mutation-sensitive manner. These networks quantitatively predict functionally relevant experimental datasets spanning multiple scales, including allosteric sensitivity across a saturation mutagenesis library, calorimetric binding entropies, and phylogenetic trees. These results suggest that a sparse network of entropic couplings encodes the dynamics-to-function map; functional reprogramming and diversification by ligand binding and evolution can modify this network without changing protein structure.

Introduction

Proteins sense their environments by binding ligands, which can alter downstream protein activity and gene expression to determine cell fate. For example, the steroid receptor estrogen receptor alpha (ERα) (figure S1A) amplifies angstrom-scale differences in ligands at nanomolar concentrations into morphological changes apparent within days (*1, 2*). The molecular basis for this sensory response is allostery—the communication of information from sites of signal detection to distant functional sites on the protein. Allosteric proteins serve as reproducible yet context-dependent communication channels which can be combined to form scalable signaling circuits (*3*).

Allostery suggests cooperative changes in conformational probabilities between signal reception sites and distant functional sites (*3–5*). Such cooperativity have been described using phenomenological models derived from experimental data (*3, 6–8*). Simplified thermodynamic models (*9–11*) clarify the nature of energetic interactions that foster long-range conformational coupling. In these models, the allosteric connections and conformations are typically assumed or known *a priori*. How to predict allosteric networks without such advance knowledge has remained a challenge.

Some allosteric encodings are structurally evident—the ligand binding-domain of ERα (ERα-LBD) distinguishes estrogens (e.g., PDB: 1QKU) from anti-estrogens (PDB: 3ERT) in its hormone-binding pocket by altering tertiary structure (”fold”; figure S1A-B) (*12–14*). The role of conformational dynamics in encoding protein function is less well-understood: how ERα allosterically discriminates estrogens from numerous environmental ligands that induce the same structural change (*15–17*), how mutations in its disordered ”instability” region modulate activities (figure S1A) (*14, 18*), and how these instabilities induce the ”off” state in the absence of estrogen (*14, 19*).

The limitations of static structural models become even more evident when explaining functional diversity between related signaling proteins despite strong structural conservation. For example, steroid receptors maintain identical folds despite performing distinct signaling roles. This pattern suggests that information processing, especially in signaling, emerges not from static structures, but from conformational ensembles that encode intramolecular communication via correlated dynamics (*5, 7, 20*). Evolution can therefore diversify proteins for specific functions by shaping dynamics; conversely, dynamics can in principle encode the evolutionary history within protein families (*21*).

Multiple lines of evidence suggest the existence of sparsely-connected communication networks in proteins. Saturation mutagenesis localizes functionally significant mutations primarily to binding sites and key allosteric positions (*22–25*). Coevolutionary analyses independently predict these sites from sparse networks of positional couplings (*26*). These approaches illuminate how correlated mutations on evolutionary timescales can reflect physical interactions within proteins, but do not directly capture the biophysical couplings encoded by correlated dynamics on physiological timescales.

Allosteric dynamics can localize to changes in specific dihedral angles, often without significant structural rearrangements. Energy barriers constrain these angles into discrete rotameric states, corresponding to geometric digital ”bits” that are robust to thermal fluctuations. Analyses of rotameric states in protein structures provide insight into molecular recognition and signaling (*27, 28*), but the limited number of structures in protein structure databases makes it difficult to infer whole-protein allosteric networks. This has motivated descriptions of allostery through rotameric entropies and mutual informations computed from molecular dynamics simulations (*29–31*), but frameworks that directly generate quantitative experimental predictions of protein function from simulation data have remained elusive. A key limitation was accounting for the strong intra-residue dihedral couplings to infer well-defined residue states from dihedral states.

Here, we address this challenge by constructing a functionally predictive intraprotein communication network, which we call the conformational inference Maximum Information Spanning Tree (ciMIST), using atomistic simulations. Our approach identifies protein-specific discrete conformational states of entire residues from dihedral angle fluctuations and constructs an information tree network (*32, 33*) that is the optimal estimator of the joint probability distribution over these residue states, enabling exact calculation of entropy contributions for arbitrary residue subsets. By applying ciMIST to large-scale simulations of the third PDZ domain from PSD95 (PDZ3) and the ERα ligand-binding domain – both in multiple ligand-bound and mutational states – we demonstrate that ciMIST accurately predicts entropy differences measured by calorimetry as well as experimentally-measured differences in protein dynamics. ciMIST residue features also quantitatively predict sequence-specific effects, such as the positional propensities for allosteric effects in a saturation mutagenesis library of PDZ3 and the phylogenetic relationships within the entire human steroid receptor family. This approach shows that modeling proteins as digital communica-

tion channels generates accurate predictions of protein allostery from equilibrium simulations of dynamical fluctuations.

ciMIST: Dynamics-based intra-protein information network

To identify allosteric networks from molecular dynamics requires global statistical models of conformational ensembles that are expressive enough to account for observed protein function yet sufficiently low-dimensional to alow reliable statistical inference. Recent saturation mutagenesis experiments probing binding thermodynamics of signaling protein domains (*24, 25*) provided two key insights that shaped our modeling choices. First, most sites are only weakly coupled to binding, with mutations typically changing affinities by an amount beneath the scale of thermal fluctuations ($|\Delta\Delta G_{bind}| < 0.6$ kcal mol^{-1}). Second, mutations with significant allosteric effects typically change binding by just 1-3 kcal mol^{-1}. Taken together, these suggest an absolute scale for contributions of single residues to binding thermodynamics and suggest that residues communicate through a sparse allosteric network. The sparsest network connecting all residues is one in which every residue pair communicates through a unique pathway – a hypothesis that we noticed has quantitative implications.

We denote $p(x_v)$ as the probability that residue v is in conformation x. Communication through single paths translates to the dynamical assumption that any residue conformation influences conformational probabilities at another site only through conditional distributions along a unique path of direct pairwise interactions, a property expressed formally as $p(x_v|x_u) = \prod_{(u',v')\in\text{Path}(u\to v)} p(x_{v'}|x_{u'})$. Because this relationship is assumed to hold for all residue pairs (x_u, x_v), global conformational probabilities for a protein of length L (which are joint probabilities of residue conformations) are determined by L residue marginal (node) probabilities and $L-1$ pairwise (edge) distributions:

$$p_T(x_1 \ldots x_L) = \prod_{u\in\text{residues}} p(x_u) \prod_{(u,v)\in\text{edges}} \frac{p(x_u, x_v)}{p(x_u)p(x_v)}. \tag{1}$$

A probability distribution of this form is called a tree distribution, reflecting the structure of the factor graph, wherein factors $\{p(x_u)\}$ and $\{p(x_u, x_v)\}$ are identified with nodes and edges, respectively.

The tree structure enables exact calculation of the Gibbs-Shannon conformational entropy of

the entire protein as

$$S_T(X_1 \ldots X_L) = \sum_{u \in \text{residues}} S(X_u) - \sum_{(u,v) \in \text{edges}} I(X_u, X_v), \tag{2}$$

where $S(X_u) = -\sum_{x_u} p(x_u) \log p(x_u)$ is the marginal entropy of residue u and $I(X_u, X_v) = \sum_{x_u, x_v} p(x_u, x_v) \log \frac{p(x_u, x_v)}{p(x_u)p(x_v)}$ denotes mutual information (MI) between residues u and v. Inference from data is also straightforward: the tree producing the joint probability distribution with minimum Kullback-Leibler divergence from the true distribution is just the spanning tree which maximizes $\sum_{(u,v) \in \text{branches}} I(X_u, X_v)$, i.e. the maximum information spanning tree (*32*) (Materials and Methods). We also have a useful decomposition of entropy into residue contributions:

$$S_u^{contrib} = S(X_u) - \frac{1}{2} \sum_{v \in \text{neighbors}(u)} I(X_u, X_v), \tag{3}$$

where neighbors(u) are residues connected to u by edges and the factor of $\frac{1}{2}$ ensures mutual information is split equally. This enables entropy analysis at all scales through simple summation of residue contributions, a fact we use to make comparisons to multiple experimental data sources.

This leaves the problem of identifying residue conformations. We identified conformations with free energy basins in the space of bond and dihedral angles defining the atomic positions within each residue. Free energy basins should satisfy two criteria: 1) they must be densely-populated regions of configurational space which are well-separated from each other 2) they must be sufficiently sampled so that their contributions to entropies and mutual informations are well-resolved. We satisfied both criteria by first fitting von Mises mixture models (with 100 components) to the configurational distributions of residues observed in our simulations, and next clustering the mixture components using DBSCAN (*34*), merging ”mixed states” of high component density overlap and grouping configurations in sparsely-populated regions into a single state (effectively a ”free energy plateau”) (Supplemental Text). This allows the number of conformations of each residue to be learned directly from data while reducing bias of our entropy estimates (Materials and Methods). ciMIST differs from previous rotamer-based approaches in three key ways: finding discrete states of entire residues rather than dihedral angles, defining fold-specific states for each residue, and construction of maximally predictive tree structures.

In summary, starting with MD trajectory data, the algorithm proceeds as follows:

1. Sample residue configurations parameterized by bond and dihedral angles
2. Fit von Mises mixture models to the joint probability space over all angles for each residue in the protein (100 components).
3. Cluster mixture components using DBSCAN to identify free energy basins corresponding to the discrete states of each residue (minimum probability: 0.01).
4. Calculate residue entropies and pairwise mutual informations from joint probability distributions over discrete residue states between all pairs of residues.
5. Construct the maximum information spanning tree.

We implemented this algorithm as ciMIST (conformational inference Maximum Information Spanning Tree, figure 1) in Python, utilizing the JAX (*35*), MDTraj (*36*), and NetworkX (*37*) libraries, and the nerfax (*38*) implementation of the mp-NERF algorithm (*39*) for conversion to internal coordinates.

To quantify robustness of our results to mixture model parameter initialization, we employed multiple initializations and developed an uncertainty estimation framework accounting for their variance (see Materials and Methods). Unless otherwise stated, estimates shown are derived from the tree whose entropy was closest to the average entropy over 21 runs. For all proteins, mutants, and ligand-states studied, the ciMIST tree was an excellent generative model of pairwise mutual information, neither missing nor creating spurious significant mutual information pairs (figure S4). To generate molecular dynamics data from which we inferred ciMIST models, we employed parallel simulations, generating microseconds of data from between 4 and 7 parallel trajectories per system (for PDB codes, simulation times, and numbers of replicates, see Supplemental Table S1). This translated to a requirement of between 1 and 2 weeks of simulation time on modern graphics processing units per thermodynamic state, so that full thermodynamic cycles required between 6 and 8 weeks in total. However, as we will demonstrate, dominant features of the network architectures are robust to ligands and mutations and also predictive of the most significant changes in conformational entropy, making it possible to gain useful insight into intraprotein communication from a single set of parallel trajectories.

A unified dynamics-based model for allostery in PDZ3

The allosteric network of PDZ3, the third PDZ domain of PSD95, (figure 2A) has been extensively profiled. Site-directed and saturation mutagenesis, structure, and coevolution have shown how residues in the $\beta 2 - \beta 3$ loop shape specificity (*22, 24, 40*). Nuclear magnetic resonance (NMR) and calorimetry have shown how a non-conserved helix (H3) stabilizes the apo state, reducing the entropy lost when a peptide (CRIPT) binds (*41*).

We performed MD simulations (*42, 43*) (Materials and Methods, Supplemental Table S1) of $\text{PDZ3}^{302-402}$ and (henceforth: PDZ3) and $\text{PDZ3}^{302-395}$ (PDZ3Δ7ct, which lacks H3) in the presence and absence of CRIPT as well as simulations of the isolated CRIPT peptide. This constitutes an exhaustive simulation of all protein states needed to characterize the H3-mediated entropic effect, while also providing a suitable point of comparison to mutagenesis experiments performed on $\text{PDZ3}^{312-394}$.

Applying ciMIST to these data, we observed that the $\beta 1 - \beta 2$ and $\beta 2 - \beta 3$ loops act as information hubs in all protein states (figure 2B). Paths in the ciMIST networks highlighted well-known interactions (supplemental figure S2B). A direct path from THR-2 of CRIPT to HIS372 to ILE336 (in both CRIPT-bound states) and branches connecting HIS372, GLY329, and GLY330 (in PDZ3-CRIPT), reflect evolutionary and functional couplings underlying specificity for distinct classes of peptides distinguished by the residue at the -2 position of the peptide (*22, 40*). Also acting as an information hub is ARG318, which mediates the dependence of binding kinetics on ionic conditions (*44*).

Using equation 3, we explored the contributions of different regions of PDZ3 to conformational entropy differences and interpreted these using randomly-sampled structures from our simulations (figure 2 C-D). The CRIPT peptide contributes a constant loss of 3.6 kcal mol^{-1} upon binding, imposing a baseline entropic penalty from the collapse of its unbound random coil ensemble into a quasi-linear geometry in the PDZ binding groove. PDZ3 partially offsets this with an increase of 1.46 kcal mol^{-1} in entropy contributed by C-terminal loop residues 399-402. The $\beta 2 - \beta 3$ loop contributes a loss of entropy which is greater for PDZ3Δ7ct (-1.4 kcal mol^{-1}) than PDZ3 (-0.76 kcal mol^{-1}) due to the loss of its pre-stabilization by H3 and consistent with the disorder of residues 330-332 in the crystal structure of $\text{PDZ3}^{302-392}$ (PDB: 6QJD) which is not observed in

other structures of PDZ3 (*45*).

In addition to qualitative consistency with functional studies, our identified information hubs quantitatively predict positional probabilities that mutations allosterically influence binding in a deep mutational scan (*24*). We used our residue-level thermodynamic features, the marginal entropy and the sum of mutual information with all tree neighbors, as inputs to a Bayesian generalized linear model (Supplemental Text) for the expected number of allosteric mutations at each site. Using only these two parameters, we obtained an excellent fit to saturation mutagenesis data (figure 2E). To our knowledge, this is the first time that protein-wide allosteric sensitivity to mutations has been quantitatively predicted by protein dynamics; notably, these predictions are based on the equilibrium fluctuations of the wild-type protein without the need to simulate the dynamics of the mutants. The posterior distributions of model coefficients also showed that higher network mutual information correlated with increased allosteric mutation propensity (95% credible interval: $[0.9357,\ 2.993]$), validating that our residue conformations and network structures are capturing features of the conformational ensemble that promote a stable bound conformation.

Conformational entropy differences predicted by ciMIST are in good agreement with measured binding entropies of PDZ3 and PDZ3Δ7-ct (figure 2F). Differences in residue entropy contributions also correlate significantly with reported changes in NMR dynamics between the apo states (Pearson $r = 0.63$, $p = 3 \times 10^{-10}$) (*41*). This partially supports the model originally suggested based on NMR (*41*) attributing the change in binding entropies to differences in apo state conformational flexibility, with the main difference being that our analysis does not suggest a major role for residues whose side chain order parameters they measured. Our models are in better agreement with the calorimetry data than the NMR entropy meter, which estimates conformational entropy from changes in these side chain order parameters (*46*).

A unified dynamics-based model for allostery in estrogen receptor α.

Allosteric regulation of estrogen receptors is a major factor in human health and disease (*14, 47–49*). Activation of ERα occurs when estrogens such as estradiol (EST) bind in the hormone-binding pocket, inducing a number of changes in the ligand-binding domain (LBD): increased global stability (*14, 19, 50*), increased homodimerization (*51, 52*), and stabilization of an agonist

helix 12 (H12) conformation, accommodating binding of steroid receptor coactivators (SRCs), enabling assembly of active transcriptional complexes on DNA (*53*) (figure 3A, Supplemental figure S1A-B). Inhibition by anti-estrogens used as breast cancer therapies triggers an alternative (antagonist) H12 fold, occluding SRC binding (*12, 14*). Mutations in the ”instability” (residues 460-471) and (”spring”, residues 529-538) regions bring about hormone-independent stabilization of active homodimers, causing breast cancer cell proliferation and resistance of tumors to anti-estrogen therapies (*14, 18, 19*). The wild-type ERα displays signatures of disorder and has resisted crystallization (*14*). Therefore, both local and global conformational instabilities are central to ERα function. We applied ciMIST to the ERα ligand-binding domain (henceforth: ERα for brevity) to reveal the thermodynamic mechanism behind these ligand- and mutation-dependent instabilities in terms of cooperative residue interactions.

We performed simulations of ERα homodimers, reflecting nanomolar-scale affinity (*51*) and the micromolar-scale receptor concentrations in experimental datasets (*16, 50, 54*). We concatenated trajectories of the single subunits, so that all conformational distributions and couplings are those observed within a single subunit, but averaged over the two copies in the homodimer, improving statistical resolution at the cost of modeling interdomain communication.

Predicting the effects of thermostabilizing mutations on ERα binding entropies

To gain insight into the role of local instability in ERα function, we first used ciMIST to compare human ERα to a engineered thermostable estrogen receptor (prsERα, figure 3A) which shares many of its allosteric features (*54*). prsERα was designed by introducing 22 PROSS-suggested (*55*) ERα amino acid substitutions outside of major allosteric sites (*54*). While conserving binding affinities and ligand-dependent tertiary structure, prsERα has an altered binding entropy for genistein (GEN), a phytosterol found in soybeans which is allosterically discriminated from estradiol by human ERα (*51*). We performed simulations of human ERα and prsERα, each in estradiol- and genistein-bound states, and analyzed them with ciMIST to reveal the relationship of ERα instability to allosteric specificity.

The H9-H10 loop region is the most significant information hub in all four states, with PHE461 playing an especially prominent role (figure 3B-D). In both states of the human receptor, PHE461 displays singularly strong couplings with residues in the H9-H10 loop, H10, with TYR459 in H9,

but these connections are stronger and more numerous when the receptor is bound to genistein than to estradiol. Due to this difference in coupling strengths, PHE461 contributes 2.06 kcal mol^{-1} more entropy when bound to estradiol than genistein, even while its marginal entropy is greater by just 0.1 kcal mol^{-1} – an entropic compensation effect. Qualitatively similar couplings are observed in prsERα, but pre-stabilization by mutations in H10 blocked communication between these residues and PHE461, eliminating entropic compensation.

To investigate whether this predicted compensatory effect reflected a true difference between these states, we first compared differences in the entropy contributions of segments (computed by summing equation 3 over segments in each state) with published differences in hydrogen-deuterium exchange (HDX) between estradiol and genistein-bound states of ERα (*16*). We found strong correlation (Pearson $r = 0.58$, $p = 8 \times 10^{-5}$) between these (figure 3E), suggesting that our dynamics-based model is capturing experimentally-relevant conformational shifts. Moreover, changes in entropy contributions correlated better with changes in HDX than changes in summed residue marginal entropies (Pearson $r = 0.31$, $p = 0.052$), further supporting our finding of entropic compensation and underscoring the importance of cooperative interactions in experimental dynamics.

ciMIST entropy differences between estradiol and genistein-bound states are within 1-2 kcal mol^{-1} of experimental values (*54*) for both ERα and prsERα, although exact agreement should not be expected here because our method estimates only protein entropies without accounting for differences in ligand or solvent entropies (figure 3F). However, our calculated difference in how the two proteins respond to estradiol compared to genistein ($\Delta\Delta S$) falls within statistical uncertainty of experimental measurement (figure 3F), as expected due to the subtraction of likely similar solvent and ligand entropies.

By using ciMIST to compare ERα to prsERα – a highly similar receptor optimized for thermostability – we show how ERα uses localized changes in conformational entropy to allosterically discriminate an endogenous estrogen from a similar ligand present in the environment, a challenge likely faced by estrogen receptors for much of their evolutionary history (*56*). We found that this is achieved using entropic compensation of H10 ordering with PHE461, which can be effected by specific ligands.

Rewiring of the ERα allosteric network in the Y537S mutant

We next applied ciMIST to simulations of apo states of ERα_{Y537S} and wild-type ERα. Because no crystal structure of an unliganded wild-type exists, we seeded its simulations with a homology model (*57*) built using apo states of the Y537S and D538G mutants (PDB codes: 2B23, 4Q13) as templates. On average, agonist H12 positioning was maintained by both Y537S and the wild-type (figure 4A).

The wild-type ciMIST network featured direct connections between PHE461 and binding site residues HIS524 and GLU353 (figure 4B, supplemental figure S3D). Both of these form hydrogen bonds with bound ligands, indicating that ligands engage allosteric pathways present in the equilibrium fluctuations of the apo receptor. The ciMIST network of the Y537S mutant was noticeably different, consistent with its altered allostery (figure 4B); in particular, many PHE461 connections were weakened or lost, while LYS472 and ARG434 emerged as information hubs.

Relative to the wild-type, the Y537S mutant displayed increased entropy contributed by PHE461, decreased entropy contributed by H10 driven by increased network mutual information from LYS472, decreased entropy contributed by H3, H11, binding pocket residues, and the H11-H12 loop (figure 4C, supplemental figure S6C). Increased stability of the binding pocket and H12 positioning in Y537S correlated to decreased hydration of the hydrophobic hormone-binding pocket (Supplementary figure S3A), with extensive pocket hydration accumulation preceding the largest wild-type H12 repositionings (Supplementary figure S3B). Interestingly, ciMIST predicts similar conformational entropies for the two apo states, suggesting that stabilization of H12 positioning and the protein core (figure 4C), rather than global stabilization, underlies hormone-independent Y537S activities and crystal formation.

Local entropy compensation and global disorder in ERα regulation

To build a more comprehensive model validating the observed entropy compensation mechanism in ERα, we next compared ciMIST trees derived from multiple additional liganded states to the wild-type apo state. In addition to networks previously inferred from estradiol and genistein-bound states, we performed simulations of ERα bound to 4-hydroxy-tamoxifen and lasofoxifene, both of which induce an antagonist conformation, and analyzed these with ciMIST.

For all liganded states we computed differences in residue entropy contributions, entropic contributions of secondary structural elements, and total protein conformational entropies relative to the wild-type apo state. Differences in total conformational entropies from the apo state are between -8 and -14 kcal mol^{-1} for all liganded states (figure 5 C). These differences are in the vicinity of the experimental binding entropies (*54*) of -10.56 ± 0.15 (genistein) and -10.88 ± 0.16 kcal mol^{-1} (estradiol). While a complete account of the binding entropy would require an estimate of the contributions from the desolvation of the hydrophobic ligands (likely favorable), our calculations provide quantitative evidence that conformational entropy loss is a major determinant of measured binding entropies.

We applied principal component analysis to the set of differences in residue entropy contributions of these liganded states plus the Y537S apo state from the wild-type apo state (figure 5B-C). The leading principal component, accounting for 46.8 percent of variance in conformational entropy differences, captures H10 ordering and its PHE461 compensation, indicating that this entropy compensation mechanism we first observed when changing from genistein to estradiol (figure 3C) is a general regulatory feature of ERα (figure 5B). The second component separates antagonist ligands from the apo-Y537S state, consistent with increased resistance of Y537S to anti-estrogen therapies. The third component, onto which all states have a positive loading, captures global stabilization of the receptor, including in the hormone-binding pocket.

We compared to differences in hydrogen-deuterium exchange (*16, 50*) using the wild-type apo state as a baseline. We again found strong correlation (Pearson $r = 0.36$, $p = 3 \times 10^{-6}$) between changes in entropy contributions and changes in HDX (figure 5D) and weaker correlation using summed residue marginal entropies (Pearson $r = 0.3$, $p = 0.0001$). The most notable outliers were found in segments of the beta sheet region (residues 403-408 and 403-410) for the two antagonist ligand states, possibly indicating a breakdown of this linear relationship for larger HDX differences. While HDX is interpreted as measuring dynamics in ERα (*16, 19, 50, 58*), to our knowledge this is the first study to quantitatively correlate HDX to binding thermodynamics through changes in conformational entropy. The agreement with experiment is notable considering that neither our sampled MD ensembles nor ciMIST parameters were fit to these data.

Taken together, our analyses yield a comprehensive dynamics-based model of entropic allostery in ERα. Our model quantitatively explains multiple observations – large losses of entropy (*54*) and of

hydrogen-deuterium exchange (*16, 50*) – as consequences of the loss of conformational entropy upon ligand-binding. ciMIST makes the new prediction that ligands are distinguished by PHE461/H10 entropy compensation, providing a framework that may explain distinct responses evoked by ligands which trigger the same H12 conformation. Supporting a role for this entropic compensation in ERα activation are the increased homodimerization and estrogen-independent activities conferred by the F461V and S463 mutants found in breast cancer patients (*18*). Also notable is that H9-H10 is a major site of protease cleavage (*59*), suggesting that ligands may regulate ERα abundance through their effects on PHE461. The consequences of these simulation data would have been far from obvious without ciMIST, which provided thermodynamic models of intraprotein communication, revealing changes in cooperative residue dynamics that quantitatively explain mutant- and ligand-specific order-disorder transitions underlying ERα function.

Decoding the phylogenetic history of the steroid receptor family from residue-level entropy maps

The steroid receptor family consists of the progesterone receptor, androgen receptor, estrogen receptors α and β, and the glucocorticoid and mineralocorticoid receptor. Evolutionary analysis suggests that these master transcriptional regulators evolved from an ancestral receptor – most likely an estrogen receptor – through a series of genome duplications (*60*). Despite adopting the same fold in the presence of their activating hormones, the details of allosteric regulation appear to have diverged. First, although the active transcriptional complexes are dimers, the dimer interface is not itself conserved within the family. Second, thermodynamic cooperativity is observed at different stages of assembly of the active transcriptional complex (*61, 62*). Past analysis of contact paths to helix 12 in molecular dynamics of the ketosteroid receptors suggest a mixture of conservation and divergence in the dynamics (*21*), but an analysis without prior assumption of an allosteric site and which includes the estrogen receptors has not been done.

Having already characterized the ERα-LBD signalling network, we ran MD simulations of the ligand binding domains of the five remaining steroid receptors, each bound to its endogenous agonist hormone, and inferred trees with ciMIST (figure 6A). ERβ was simulated as a homodimer, and each of the four ketosteroid receptors was simulated as a monomer. We aligned the sequences

and built a phylogenetic tree using the UPGMA algorithm, which recovered the widely-accepted phylogeny. Then, using the sequence alignment to align the vectors of entropy contributions, we calculated similarities between proteins using the correlations between entropy contributions at the alignment positions. We repeated this for the RMSF of the Cα atoms in the simulations. We then clustered these matrices using the UPGMA algorithm, as we had done with the sequence identity. We found that the entropy contributions were not conserved at the level of the family, but rather diverged in a manner that mirrored their evolutionary divergences. This was not the case for the RMSF, which proved highly sensitive to fluctuations in the termini, nor for the template modeling scores (TM-scores) of the crystal structures, which has very little dynamic range (figure 6B).

We also found that principal component analysis of the aligned entropy contributions separated the receptors by phylogeny along the first two principal components (figure 6C), which collectively captured almost 65% of the variance in the distribution of conformational entropy along the aligned sequences. The first principal component, which captures 44% of the variance, separates estrogen from ketosteroid receptors, which is believed to be the first evolutionary transition within the family. The second component captures about 19% of the variance and separates the corticosteroid receptors from androgen and progesterone receptors.

These results provide a quantitative map of a single manifestation of a more general evolutionary principle: divergent information-processing endows novel function even as 3D protein structure itself is conserved. Notably, such divergence is as much a necessity for mammalian development as the distinct hormone-binding capabilities themselves: just as estrogen signaling must differentiate estrogen from related ligands, active states of these receptors must be distinguishable from each other to coordinate distinct transcriptional programs. Consistent with this, we find that the principal components of the entropy contributions reflect the evolutionary reprogramming of function at evolutionary branch points. Through this branching, the conserved fold comes to encode function in a qualitatively new way: not through the encoding of a single behavior, but through capacity for evolutionary reprogramming, without major structural rearrangement, in support of novel behaviors.

Conclusions and Implications

There is growing appreciation for the need to move beyond protein structure-function relationships towards an understanding of how protein function is encoded in conformational ensembles. For two protein domains of differing size and complexity, we provided a unified, quantitatively-predictive model of thermodynamic binding entropies and ly-measured protein dynamics. The positive correlation between residue mutual information and allosteric mutational frequencies in PDZ and the recapitulation of steroid receptor phylogeny further demonstrate that these conformational entropies are not noise, but encodings of protein function.

Our results show how the diversity of protein function is achievable with a single protein fold. The unliganded state of ERα provides a particularly striking example: we find that the ciMIST model accounts for a difference of over 10 kcal mol^{-1} of conformational entropy of the apo state relative to the liganded states, consistent with the measured calorimetric binding entropy. While this does not conclusively prove that our sampled ensemble constitutes the true ensemble of unliganded ERα, it demonstrates the range of conformational entropy achievable through the broadening of the occupancy distributions of the existing discrete residue states without invoking new residue conformations to explain the disordered state.

The ciMIST network may be broadly relevant to signaling proteins, including those that display ligand-mediated order-disorder transitions. For example, steroid receptors associate with heat shock protein complexes in vivo when their ligands are absent (*63*). Likewise, the apo states of extracellular regions of G protein-coupled receptors are typically disordered in structural models found in the Protein Data Bank. Immunoglobulins, which must select antigens from a large pool of possible proteins, participate in interactions exhibiting large and negative binding entropies coincident with broadly-distributed losses of hydrogen-deuterium exchange (*64*). More broadly, bioinformatic analyses across the tree of life shows a robust correlation between protein disorder and complex decision-making: disordered protein regions are more prevalent in multicellular organisms than in unicellular organisms and more prevalent in transcriptional regulation than in any other process (*65–67*). Phosphorylation of disordered protein regions is known to be crucial for the specificity of transcriptional responses in bacteria (*68*), hinting at a deep evolutionary link between protein conformational entropy and information processing. Our analysis of ERα leads us to wonder

whether each of these might be understood as examples of a general tendency for information acquisition by proteins to coincide with loss of conformational entropy – just as the entropy of a communication channel is reduced when a message is fixed, with more complex communications requiring greater entropy reductions – and provides a quantitative, structurally-interpretable model of this for one important signaling protein.

We modeled proteins as tree-structured networks of discrete-state amino acid sensors, which we arrived at through pragmatic considerations. The use of a network structure with unique paths between residues enabled exact predictions of entropies, while network structures with loops would have necessitated approximations (beyond the model structure itself) that would have introduced ambiguity in our predictions. We defined residues as basic sensory units because we judged this would simplify comparison to data, given that the single residue, via mutagenesis, is a common scale at which differences in protein function are resolved.

What can we gather about proteins from the apparent effectiveness of this framework? From the perspective of binding thermodynamics, we expect that a tree approximation will accurately estimate entropy differences which are dominated by a small number of pairwise differences in the combinatorial expansion of the entropy as a sum of entropies of residue subsets (single residues, pairs, triplets, and so on). Biophysically, it may be that large differences in three-way (and higher) entropies are not realizable without larger conformational rearrangements than we see here, such as global protein unfolding. This is suggestive of some ”separation of concerns” with respect to the maintenance of folded states and allosteric signaling.

Evolutionarily successful signaling protein folds might then be more analogous to circuit boards – physical scaffolds capable of supporting multiple computations – than to circuits themselves, with allostery specialized through mutations at surface sites. Conserved and coevolving residues, typically packed in protein cores, might then contribute to allostery less through conformational change than through packing and rigidity, as constraints around which higher-entropy residues shift in response to signals. Consistent with this view, the mutational landscapes of two allosteric proteins – an engineered chimeric light-sensitive dihydrofolate reductase (*69*) and a tetracycline repressor (*70*) – revealed allostery-tuning mutations primarily at non-conserved sites. Allosteric regulation as a low-order interactive phenomenon would also entail that high-order mutational epistasis (*71–73*), to the extent that it exists (*74*), mostly captures cooperative interactions necessary for folding rather

than perturbations that tune allosteric regulation. Extending family-level analysis of conformational entropy networks, especially to protein families in which evolutionary correlations reflect orthogonal selective forces (*75*), can further explicate the relationship between evolutionary and conformational couplings. It would be especially interesting to know how entropic networks are rewired in laboratory evolution, especially in experiments seeded with reconstructed ancestral protein sequences (*76*) or sequences designed using generative deep learning (*77*), which might clarify relationships between protein stability, evolvability, and allostery (*40, 70, 78–83*).

The effectiveness of our description of proteins as networks of discrete-state residue sensors raises the question of whether it may be useful to conceptualize amino acids as communicating through a digital conformational code, no less fundamental than other biochemical alphabets, but rather more cryptic. We anticipate that approaches such as ours, which connect dynamical correlations between local conformations to experimental observables, will further decrypt the nature of this code.

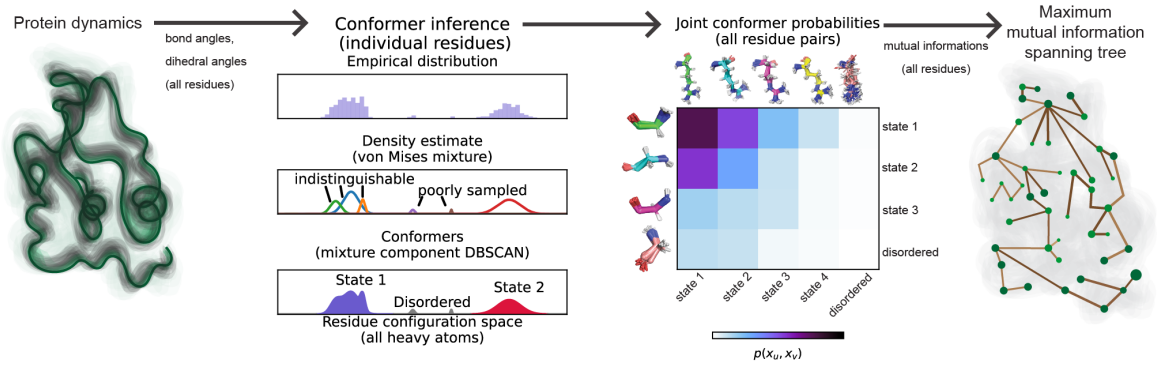

figure 1: **ciMIST infers maximum information spanning trees linking single-residue conformations into a global statistical model.** All frames of a molecular dynamics ensemble are mapped to internal coordinates. For each residue, a separate von Mises mixture model is fit to the observed residue configurations, parameterized using bond and dihedral angles. Mixture components for each residue are coarse-grained by applying DBSCAN, which identifies residue conformations by merging components whose densities overlap, and maps all low-probability components to a single disordered state. Pairwise conformational probabilities and mutual informations are computed for each residue, and the maximum information spanning tree is then found.

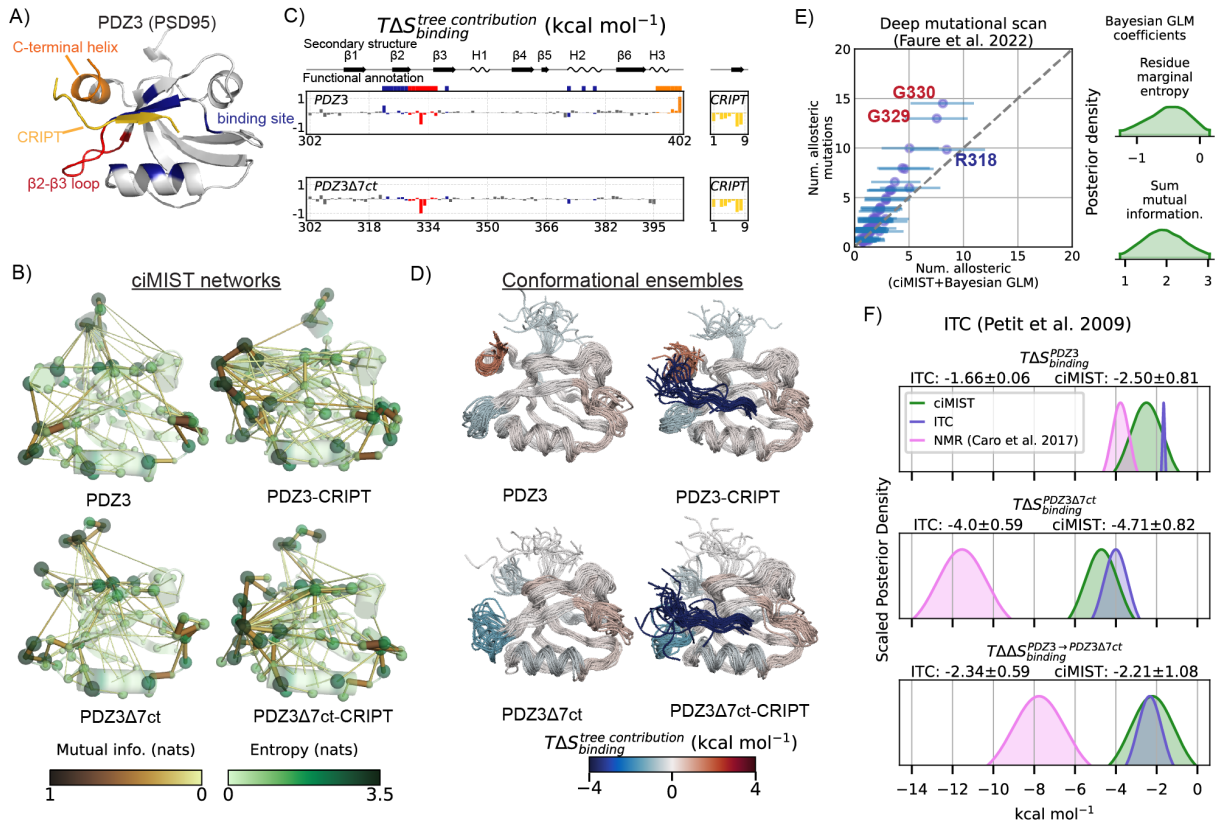

figure 2: **Connecting molecular dynamics to experimental signatures of allostery in PDZ3.** A) Structure of the third PDZ domain (PDZ3) from PSD95 bound to the CRIPT peptide. B) Maximum information spanning trees for PDZ3 (top) and PDZ3Δ7ct. C) Contributions of individual residues to binding conformational entropy. Regions of of the protein are color-coded as in (A). D) 20 randomly selected snapshots from simulations of PDZ3 (top) and PDZ3Δ7ct (bottom) with secondary structural elements colored by contribution to the binding conformational entropy. CRIPT is labelled as a single element. E) Comparison between numbers of allosteric mutations predicted using a Bayesian generalized linear model (GLM) with residue marginal entropies and total mutual information with neighbors in the PDZ3Δ7-ct tree as predictor variables (x-axis) and observed (y-axis) in a deep mutational scan (*24*) (left) and posterior distributions coefficients from the Bayesian GLM (right), with positive values indicating positive contributions to likelihood of allostericmutation. F) Comparison of posterior densities of calculated differences in conformational entropies, thermodynamic entropies measured with ITC (*41*), estimated using an NMR entropy meter (*46*). $T = 298K$.

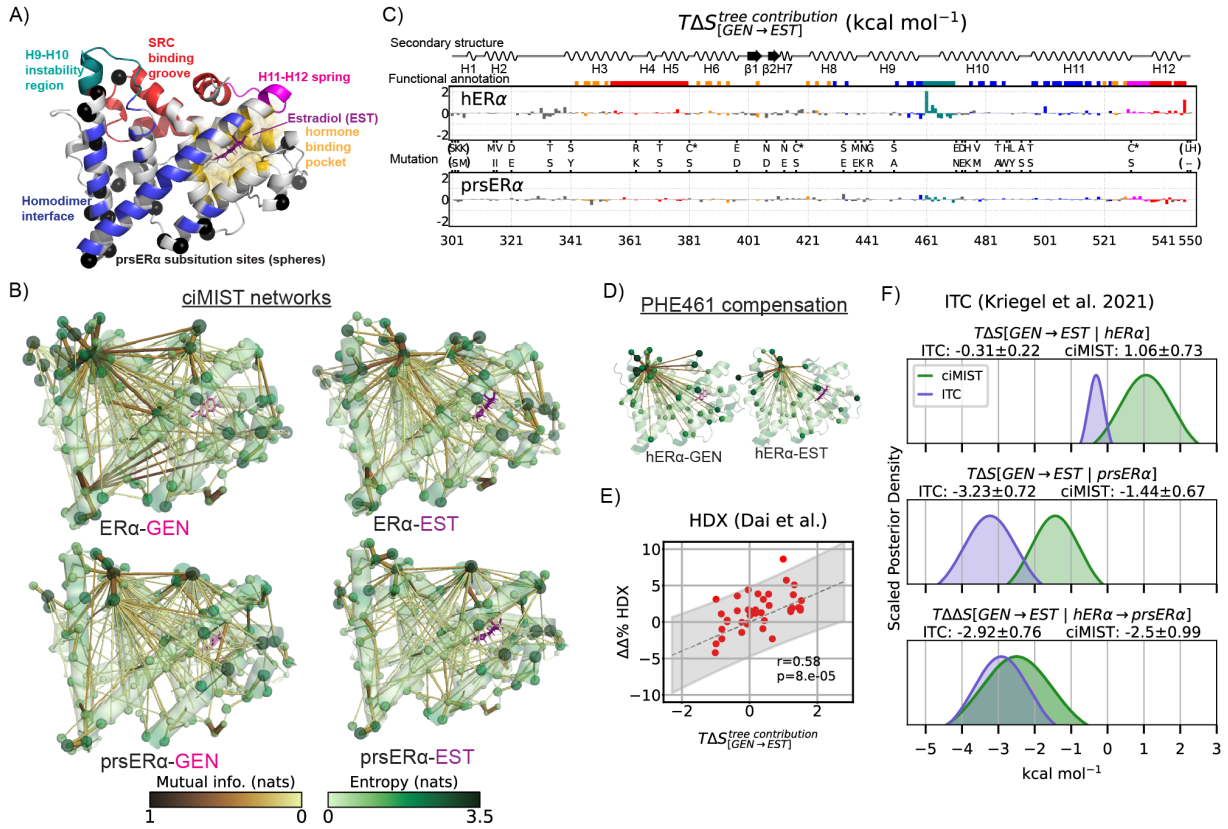

figure 3: **Entropic effects of 22 stabilizing mutations in ERα.** A) View of a single subunit of ERα with functional annotations. B) Maximum information spanning trees obtained by applying ciMIST to the marginal distribution of a single subunit of human ERα (top) and a PROSS-optimized estrogen receptor (bottom) bound to genistein (left) and estradiol (right). Genistein and estradiol are displayed as pink and purple sticks, respectively. C) Changes in the contributions of each residue to the total protein conformational entropy between genistein- and estradiol-bound states for human ERα (top) and the PROSS-optimized estrogen receptor (bottom). C* denotes residues which were cysteines in our simulations but serines in calorimetry. Parentheses denote positions not present in the calorimetry experiment (performed on residues 304-548). D) Connections in the maximum information spanning trees between PHE461 and other residues for both states of human ERα. E) Correlation between contributions of protein segments to the difference in conformational entropy and average percent differences in deuteration levels (*16, 50*) between estradiol and genistein-bound states of human ERα. F) Estimated posterior densities from ITC (*54*) and calculated using ciMIST. $T = 298K$. Uncertainties for ITC are estimated using the standard deviation of three replicates.

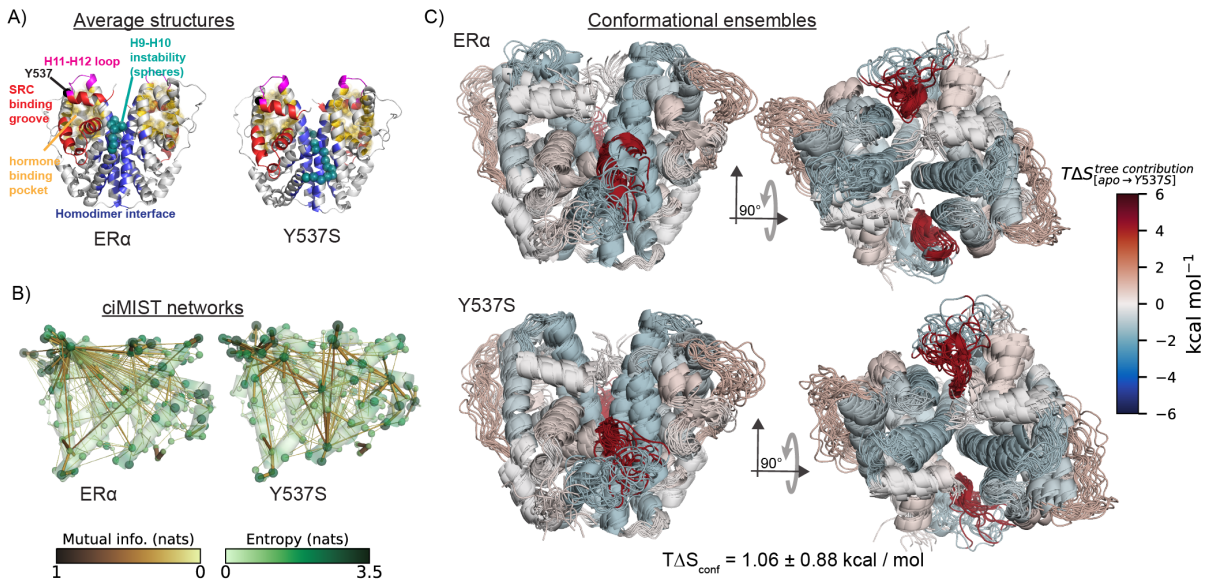

figure 4: **Redistribution of entropy in the unliganded (apo) state of the Y537S estrogen receptor α.** A) Average structure observed in simulations of human ERα-LBD homodimer and the Y537s variant. Structures were computed by superimposing all frames onto the first and then computing the average positions of atoms over all frames. B) Maximum information spanning trees for the unliganded wild-type and Y537S variants of ERα. C) 20 randomly selected snapshots from MD simulations of unliganded wild-type and Y537S variants of ERα. Secondary structural elements are colored by their contributions to the difference in entropy between mutant and wild-type states, $T(S_{\text{Y537S}} - S_{\text{wild-type}})$. $T\Delta S_{conf}$ is the ciMIST calculation of difference in conformational entropy of Y537S from the wild-type. $T = 298K$.

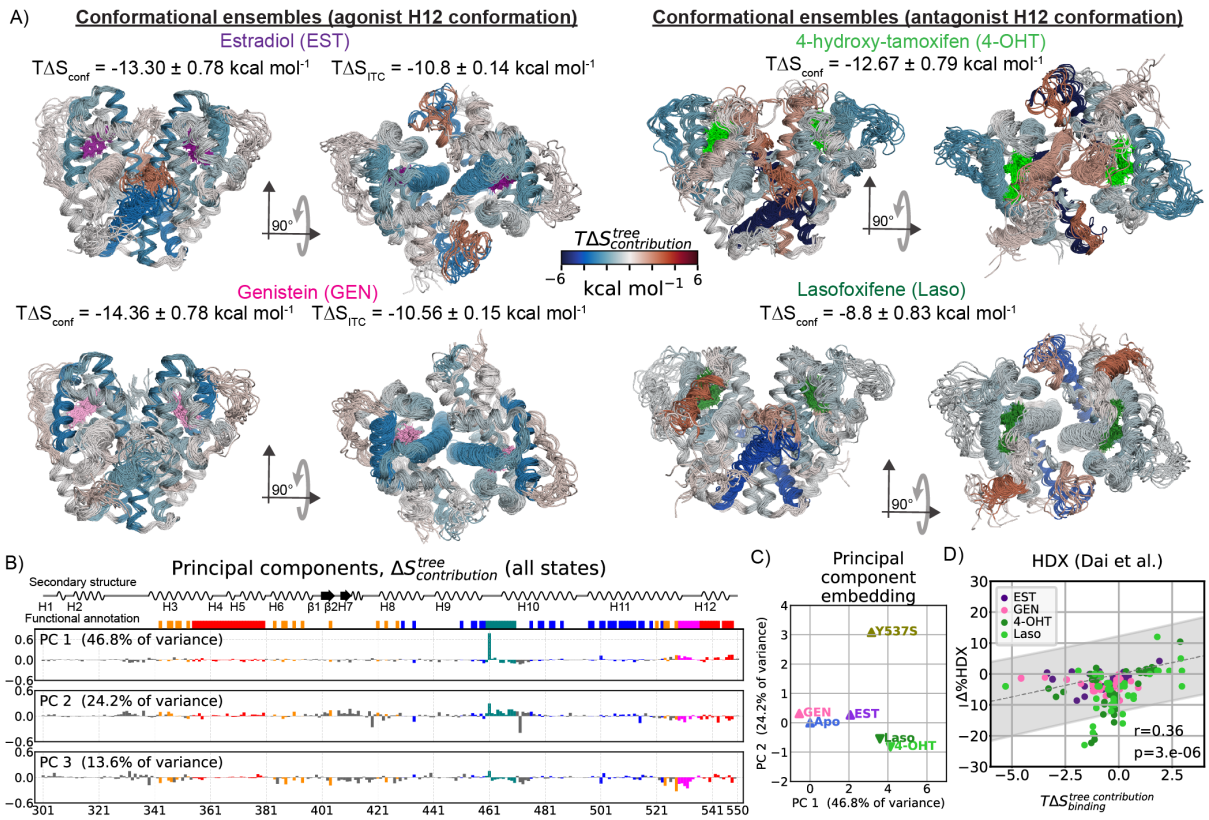

figure 5: **Entropic allostery in estrogen receptor α.** A) 20 randomly selected snapshots from simulations of ERα-LBD homodimers bound to estradiol (purple), 4-hydroxytamoxifen (lime green), genistein (pink), lasofoxifene (forest green). Secondary structural elements are colored by their contribution to the difference in entropy from the wild-type apo state. $T\Delta S_{conf}$ is the ciMIST calculation of change in conformational entropy relative to the apo state, with posterior standard deviation as the stated uncertainty. $T\Delta S_{ITC}$ denotes experimentally-measured binding entropy (*54*), for which uncertainties are standard deviations of three replicates. B) First three principal components of changes in entropy contributions of the set of liganded states shown here as well as the Y537S apo state from the wild-type apo state. C) Projection of residue entropy contribution vectors for ERα states onto the first two principal components. D) Correlation between average contributions of residues in protein segments to binding conformational entropy and average percent difference in deuteration levels between apo and liganded states (*16, 50*). The gray band demarcates 95% prediction intervals. $T = 298K$.

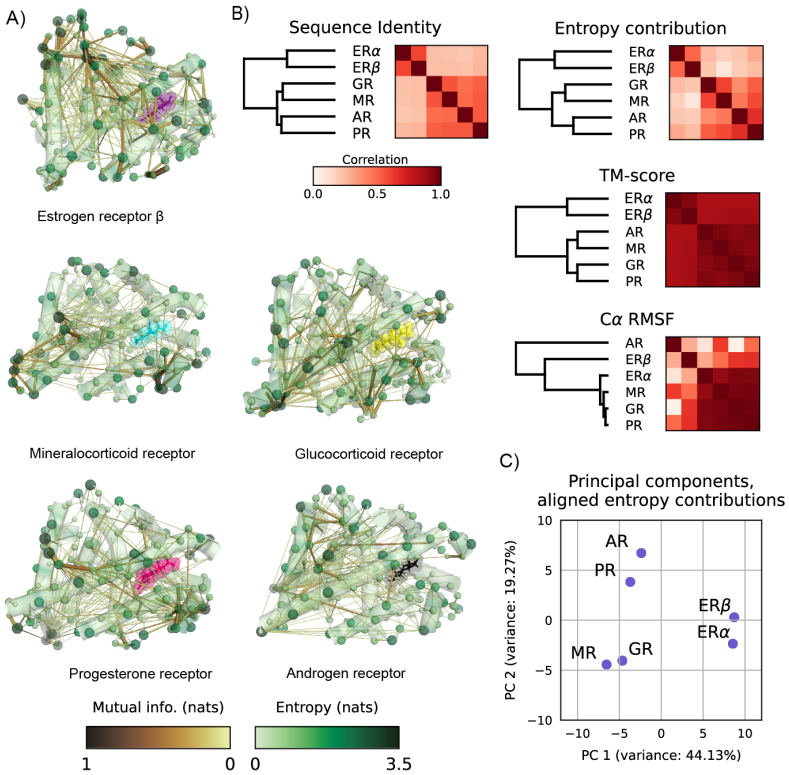

figure 6: **Evolutionary divergence of entropic contributions in the steroid receptor ligand-binding domains.** A) Maximum information spanning trees for human steroid receptors ligand-binding domains bound to endogenous ligands. Structures are oriented as ERα is in figure 3A. B) Human steroid receptor phylogeny and application of the UPGMA clustering algorithm to four different similarity matrices calculated for human steroid receptors. C) The first two principal components of the aligned entropy contributions.

References and Notes

1. C. Hsieh, R. Santell, S. Haslam, W. Helferich, Estrogenic effects of genistein on the growth of estrogen receptor-positive human breast cancer (MCF-7) cells in vitro and in vivo. *Cancer Research* **58**, 3833–3838 (1998).

2. S. Hewitt, J. Couse, K. Korach, Estrogen receptor transcription and transactivation Estrogen receptor knockout mice: what their phenotypes reveal about mechanisms of estrogen action. *Breast Cancer Research* **2**, 345 (2000), doi:10.1186/bcr79, https://doi.org/10.1186/bcr79.

3. J. Monod, J. Changeux, F. Jacob, Allosteric proteins and cellular control systems. *Journal of Molecular Biology* **6**, 306–329 (1963), doi:10.1016/s0022-2836(63)80091-1.

4. A. Cooper, D. Dryden, Allostery without conformational change. *European Biophysics Journal* **11**, 103–109 (1984), doi:10.1007/BF00276625, https://doi.org/10.1007/BF00276625.

5. H. Motlagh, J. Wrabl, J. Li, V. Hilser, The ensemble nature of allostery. *Nature* **508**, 331–339 (2014), doi:10.1038/nature13001, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4224315/.

6. F. Philip, G. Kadamur, R. Silos, J. Woodson, E. Ross, Synergistic Activation of Phospholipase C-β3 by Gαq and G$\beta\gamma$ Describes a Simple Two-State Coincidence Detector. *Current Biology* **20**, 1327–1335 (2010), publisher: Elsevier, doi:10.1016/j.cub.2010.06.013, https://www.cell.com/current-biology/abstract/S0960-9822(10)00717-7.

7. J. Li, *et al.*, Genetically tunable frustration controls allostery in an intrinsically disordered transcription factor. *eLife* **6**, e30688 (2017), doi:10.7554/eLife.30688.

8. Z. Liu, T. Gillis, S. Raman, Q. Cui, A parameterized two-domain thermodynamic model explains diverse mutational effects on protein allostery. *eLife* **12**, RP92262 (2024), publisher: eLife Sciences Publications, Ltd, doi:10.7554/eLife.92262, https://doi.org/10.7554/eLife.92262.

9. A. Atilgan, *et al.*, Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophysical Journal* **80**, 505–515 (2001), doi:10.1016/S0006-3495(01)76033-X, https://linkinghub.elsevier.com/retrieve/pii/S000634950176033X.

10. E. Rouviere, R. Ranganathan, O. Rivoire, Emergence of Single- versus Multi-State Allostery. *PRX Life* **1**, 023004 (2023), publisher: American Physical Society, doi:10.1103/PRXLife.1.023004, https://link.aps.org/doi/10.1103/PRXLife.1.023004.

11. E. Rouviere, O. Rivoire, R. Ranganathan, A Unified View of Allostery (2024), doi:10.48550/arXiv.2411.06519, http://arxiv.org/abs/2411.06519, arXiv:2411.06519 [q-bio].

12. A. Shiau, *et al.*, The Structural Basis of Estrogen Receptor/Coactivator Recognition and the Antagonism of This Interaction by Tamoxifen. *Cell* **95**, 927–937 (1998), publisher: Elsevier, doi:10.1016/S0092-8674(00)81717-1, https://www.cell.com/cell/abstract/S0092-8674(00)81717-1.

13. M. Gangloff, *et al.*, Crystal Structure of a Mutant hERα Ligand-binding Domain Reveals Key Structural Features for the Mechanism of Partial Agonism*. *Journal of Biological Chemistry* **276**, 15059–15065 (2001), doi:10.1074/jbc.M009870200, https://www.sciencedirect.com/science/article/pii/S0021925819567989.

14. J. Katzenellenbogen, C. Mayne, B. Katzenellenbogen, G. Greene, S. Chandarlapaty, Structural underpinnings of oestrogen receptor mutations in endocrine therapy resistance. *Nature Reviews Cancer* **18**, 377–388 (2018), number: 6 Publisher: Nature Publishing Group, doi:10.1038/s41568-018-0001-z, https://www.nature.com/articles/s41568-018-0001-z.

15. K. Nettles, *et al.*, NFκB selectivity of estrogen receptor ligands revealed by comparative crystallographic analyses. *Nature Chemical Biology* **4**, 241–247 (2008), number: 4 Publisher: Nature Publishing Group, doi:10.1038/nchembio.76, https://www.nature.com/articles/nchembio.76.

16. S. Dai, *et al.*, Unique Ligand Binding Patterns between Estrogen Receptor α and β Revealed by Hydrogen-Deuterium Exchange. *Biochemistry* **48**, 9668–9676 (2009), publisher: American Chemical Society, doi:10.1021/bi901149t, https://doi.org/10.1021/bi901149t.

17. J. Gertz, T. E. Reddy, K. E. Varley, M. J. Garabedian, R. M. Myers, Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner. *Genome Research* **22** (11), 2153–2162 (2012), doi:10.1101/gr.135681.111.

18. S. Irani, *et al.*, Somatic estrogen receptor α mutations that induce dimerization promote receptor activity and breast cancer proliferation. *The Journal of Clinical Investigation* **134** (2024), publisher: American Society for Clinical Investigation, doi:10.1172/JCI163242, `https://www.jci.org/articles/view/163242#B15`.

19. S. Fanning, *et al.*, Estrogen receptor alpha somatic mutations Y537S and D538G confer breast cancer endocrine resistance by stabilizing the activating function-2 binding conformation. *eLife* **5**, e12792 (2016), publisher: eLife Sciences Publications, Ltd, doi:10.7554/eLife.12792, `https://doi.org/10.7554/eLife.12792`.

20. D. Boehr, R. Nussinov, P. Wright, The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology* **5**, 789–796 (2009), number: 11 Publisher: Nature Publishing Group, doi:10.1038/nchembio.232, `https://www.nature.com/articles/nchembio.232`.

21. N. Dube, S. Khan, R. Sasse, C. Okafor, Identification of an Evolutionarily Conserved Allosteric Network in Steroid Receptors. *Journal of Chemical Information and Modeling* **63**, 571–582 (2023), publisher: American Chemical Society, doi:10.1021/acs.jcim.2c01096, `https://doi.org/10.1021/acs.jcim.2c01096`.

22. R. McLaughlin, F. Poelwijk, A. Raman, W. Gosal, R. Ranganathan, The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142 (2012), number: 7422 Publisher: Nature Publishing Group, doi:10.1038/nature11500, `https://www.nature.com/articles/nature11500`.

23. P. Bandaru, *et al.*, Deconstruction of the Ras switching cycle through saturation mutagenesis. *eLife* **6**, e27810 (2017), publisher: eLife Sciences Publications, Ltd, doi:10.7554/eLife.27810, `https://doi.org/10.7554/eLife.27810`.

24. A. Faure, *et al.*, Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022), number: 7904 Publisher: Nature Publishing Group, doi:10.1038/s41586-022-04586-4, https://www.nature.com/articles/s41586-022-04586-4.

25. C. Weng, A. Faure, A. Escobedo, B. Lehner, The energetic and allosteric landscape for KRAS inhibition. *Nature* **626**, 643–652 (2024), publisher: Nature Publishing Group, doi:10.1038/s41586-023-06954-0, https://www.nature.com/articles/s41586-023-06954-0.

26. T. Hopf, *et al.*, Mutation effects predicted from sequence co-variation. *Nature Biotechnology* **35**, 128–135 (2017), number: 2 Publisher: Nature Publishing Group, doi:10.1038/nbt.3769, https://www.nature.com/articles/nbt.3769.

27. B. North, A. Lehmann, R. Dunbrack, A new clustering of antibody CDR loop conformations. *Journal of Molecular Biology* **406**, 228–256 (2011), doi:10.1016/j.jmb.2010.10.030.

28. M. Parker, J. Meyer, E. Golemis, J. Dunbrack, Delineating the RAS Conformational Landscape. *Cancer Research* **82**, 2485–2498 (2022), doi:10.1158/0008-5472.CAN-22-0804, https://doi.org/10.1158/0008-5472.CAN-22-0804.

29. M. Lin, Timing correlations in proteins predict functional modules and dynamic allostery. *Journal of the American Chemical Society* **138**, 5036–5043 (2016).

30. S. Singh, G. Bowman, Quantifying Allosteric Communication via Both Concerted Structural Changes and Conformational Disorder with CARDS. *Journal of Chemical Theory and Computation* **13**, 1509–1517 (2017), doi:10.1021/acs.jctc.6b01181, https://pubs.acs.org/doi/10.1021/acs.jctc.6b01181.

31. L. Manley, M. Lin, Kinetic and thermodynamic allostery in the Ras protein family. *Biophysical journal* **122**, 3882–3893 (2023).

32. C. Chow, C. Liu, Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory* **14**, 462–467 (1968), doi:10.1109/TIT.1968.1054142, http://ieeexplore.ieee.org/document/1054142/.

33. B. King, B. Tidor, MIST: Maximum Information Spanning Trees for dimension reduction of biological data sets. *Bioinformatics* **25**, 1165–1172 (2009), doi:10.1093/bioinformatics/btp109, https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp109.

34. M. Ester, H. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996).

35. J. Bradbury, *et al.*, JAX: composable transformations of Python+NumPy programs (2018), http://github.com/jax-ml/jax.

36. R. McGibbon, *et al.*, MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **109**, 1528–1532 (2015), doi:10.1016/j.bpj.2015.08.015, https://www.sciencedirect.com/science/article/pii/S0006349515008267.

37. A. Hagberg, D. Schult, P. Swart, Exploring Network Structure, Dynamics, and Function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)* p. 5 (2008).

38. O. Dutton, F. Hoffmann, K. Tamiola, NeRFax: An efficient and scalable conversion from the internal representation to Cartesian space (2022), doi:10.1101/2022.05.25.493427, https://www.biorxiv.org/content/10.1101/2022.05.25.493427v2, pages: 2022.05.25.493427 Section: New Results.

39. E. Alcaide, S. Biderman, A. Telenti, M. Maher, MP-NeRF: A massively parallel method for accelerating protein structure reconstruction from internal coordinates. *Journal of Computational Chemistry* **43**, 74–78 (2022), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.26768, doi:10.1002/jcc.26768, https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.26768.

40. A. Raman, K. White, R. Ranganathan, Origins of Allostery and Evolvability in Proteins: A Case Study. *Cell* **166**, 468–480 (2016), doi:10.1016/j.cell.2016.05.047, https://linkinghub.elsevier.com/retrieve/pii/S009286741630650X.

41. C. Petit, J. Zhang, P. Sapienza, E. Fuentes, A. Lee, Hidden dynamic allostery in a PDZ domain. *Proceedings of the National Academy of Sciences* **106**, 18249–18254 (2009), publisher: National Academy of Sciences Section: Biological Sciences, doi:10.1073/pnas.0904492106, https://www.pnas.org/content/106/43/18249.

42. J. Maier, *et al.*, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **11**, 3696–3713 (2015), publisher: American Chemical Society, doi:10.1021/acs.jctc.5b00255, https://doi.org/10.1021/acs.jctc.5b00255.

43. P. Eastman, *et al.*, OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* **13**, e1005659 (2017), publisher: Public Library of Science, doi:10.1371/journal.pcbi.1005659, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005659.

44. C. Chi, A. Engstrom, S. Gianni, M. Larsson, P. Jemth, Two Conserved Residues Govern the Salt and pH Dependencies of the Binding Reaction of a PDZ Domain*. *Journal of Biological Chemistry* **281**, 36811–36818 (2006), doi:10.1074/jbc.M607883200, https://www.sciencedirect.com/science/article/pii/S0021925820720272.

45. A. Camara-Artigas, J. Murciano-Calles, J. Martínez, Conformational changes in the third PDZ domain of the neuronal postsynaptic density protein 95. *Acta Crystallographica. Section D, Structural Biology* **75**, 381–391 (2019), doi:10.1107/S2059798319001980.

46. J. Caro, *et al.*, Entropy in molecular recognition by proteins. *Proceedings of the National Academy of Sciences* **114**, 6563–6568 (2017), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.1621154114, https://www.pnas.org/doi/full/10.1073/pnas.1621154114.

47. Y. Li, *et al.*, A mutant form of ERα associated with estrogen insensitivity affects the coupling between ligand binding and coactivator recruitment. *Science Signaling* **13**, eaaw4653 (2020), publisher: American Association for the Advancement of Science, doi:10.1126/scisignal.aaw4653, https://www.science.org/doi/full/10.1126/scisignal.aaw4653.

48. J. Abramowitz, V. Tangpricha, Hormonal Management for Transfeminine Individuals. *Clinics in Plastic Surgery* **45**, 313–317 (2018), doi:10.1016/j.cps.2018.03.003, https://www.sciencedirect.com/science/article/pii/S0094129818300166.

49. R. Langley, *et al.*, Transdermal oestradiol for androgen suppression in prostate cancer: long-term cardiovascular outcomes from the randomised Prostate Adenocarcinoma Transcutaneous Hormone (PATCH) trial programme. *The Lancet* **397**, 581–591 (2021), publisher: Elsevier, doi:10.1016/S0140-6736(21)00100-8, https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00100-8/abstract.

50. S. Dai, *et al.*, Prediction of the tissue-specificity of selective estrogen receptor modulators by using a single biochemical method. *Proceedings of the National Academy of Sciences* **105**, 7171–7176 (2008), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.0710802105, https://www.pnas.org/doi/full/10.1073/pnas.0710802105.

51. R. Paulmurugan, A. Tamrazi, T. Massoud, J. Katzenellenbogen, S. Gambhir, In Vitro and in Vivo Molecular Imaging of Estrogen Receptor α and β Homo- and Heterodimerization: Exploration of New Modes of Receptor Regulation. *Molecular Endocrinology* **25**, 2029–2040 (2011), doi:10.1210/me.2011-1145, https://doi.org/10.1210/me.2011-1145.

52. J. Nwachukwu, *et al.*, Predictive features of ligand-specific signaling through the estrogen receptor. *Molecular Systems Biology* **12**, 864 (2016), publisher: John Wiley & Sons, Ltd, doi:10.15252/msb.20156701, https://www.embopress.org/doi/full/10.15252/msb.20156701.

53. W. Huang, *et al.*, Multidomain architecture of estrogen receptor reveals interfacial cross-talk between its DNA-binding and ligand-binding domains. *Nature Communications* **9** (1), 3520 (2018), number: 1 Publisher: Nature Publishing Group, doi:10.1038/s41467-018-06034-2, https://www.nature.com/articles/s41467-018-06034-2.

54. M. Kriegel, H. Wiederanders, S. Alkhashrom, J. Eichler, Y. Muller, A PROSS-designed extensively mutated estrogen receptor α variant displays enhanced thermal stability while retaining native allosteric regulation and structure. *Scientific Reports* **11**, 10509 (2021),

number: 1 Publisher: Nature Publishing Group, doi:10.1038/s41598-021-89785-1, https://www.nature.com/articles/s41598-021-89785-1.

55. J. Weinstein, A. Goldenzweig, S. Hoch, S. Fleishman, PROSS 2: a new server for the design of stable and highly expressed protein variants. *Bioinformatics* **37**, 123–125 (2021), doi:10.1093/bioinformatics/btaa1071, https://doi.org/10.1093/bioinformatics/btaa1071.

56. M. Baker, R. Lathe, The promiscuous estrogen receptor: Evolution of physiological estrogens and response to phytochemicals and endocrine disruptors. *The Journal of Steroid Biochemistry and Molecular Biology* **184**, 29–37 (2018), doi:10.1016/j.jsbmb.2018.07.001, https://linkinghub.elsevier.com/retrieve/pii/S0960076018303571.

57. B. Webb, A. Sali, Protein Structure Modeling with MODELLER (2021), doi:10.1007/978-1-0716-0892-0_14, https://doi.org/10.1007/978-1-0716-0892-0_14.

58. J. Min, *et al.*, Dual-mechanism estrogen receptor inhibitors. *Proceedings of the National Academy of Sciences* **118**, e2101657118 (2021), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.2101657118, https://www.pnas.org/doi/full/10.1073/pnas.2101657118.

59. D. Seielstad, K. Carlson, P. Kushner, G. Greene, J. Katzenellenbogen, Analysis of the Structural Core of the Human Estrogen Receptor Ligand Binding Domain by Selective Proteolysis/Mass Spectrometric Analysis. *Biochemistry* **34**, 12605–12615 (1995), publisher: American Chemical Society, doi:10.1021/bi00039a016, https://doi.org/10.1021/bi00039a016.

60. J. Thornton, Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proceedings of the National Academy of Sciences* **98**, 5671–5676 (2001), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.091553298, https://www.pnas.org/doi/10.1073/pnas.091553298.

61. A. Moody, M. Miura, K. Connaghan, D. Bain, Thermodynamic Dissection of Estrogen Receptor–Promoter Interactions Reveals That Steroid Receptors Differentially Partition Their Self-Association and Promoter Binding Energetics. *Biochemistry* **51**, 739–749 (2012), pub-

lisher: American Chemical Society, doi:10.1021/bi2017156, https://doi.org/10.1021/bi2017156.

62. J. Robblee, M. Miura, D. Bain, Glucocorticoid Receptor–Promoter Interactions: Energetic Dissection Suggests a Framework for the Specificity of Steroid Receptor-Mediated Gene Regulation. *Biochemistry* **51**, 4463–4472 (2012), publisher: American Chemical Society, doi:10.1021/bi3003956, https://doi.org/10.1021/bi3003956.

63. E. Sanchez, Chaperoning steroidal physiology: lessons from mouse genetic models of Hsp90 and its cochaperones. *Biochimica Et Biophysica Acta* **1823**, 722–729 (2012), doi:10.1016/j.bbamcr.2011.11.006.

64. Y. Yamaguchi, *et al.*, The Fab portion of immunoglobulin G has sites in the CL domain that interact with Fc gamma receptor IIIa. *mAbs* **14**, 2038531 (2022), doi:10.1080/19420862.2022.2038531.

65. J. Ward, J. Sodhi, L. McGuffin, B. Buxton, D. Jones, Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *Journal of Molecular Biology* **337**, 635–645 (2004), doi:10.1016/j.jmb.2004.02.002, https://www.sciencedirect.com/science/article/pii/S0022283604001482.

66. H. Xie, *et al.*, Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *Journal of Proteome Research* **6**, 1882–1898 (2007), doi:10.1021/pr060392u.

67. A. Dunker, S. Bondos, F. Huang, C. Oldfield, Intrinsically disordered proteins and multicellular organisms. *Seminars in Cell & Developmental Biology* **37**, 44–55 (2015), doi:10.1016/j.semcdb.2014.09.025, https://www.sciencedirect.com/science/article/pii/S1084952114002924.

68. S. Brodsky, *et al.*, Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity. *Molecular Cell* **79**, 459–471.e4 (2020), doi:10.1016/j.molcel.2020.05.032, https://www.sciencedirect.com/science/article/pii/S109727652030352X.

69. J. McCormick, M. Russo, S. Thompson, A. Blevins, K. Reynolds, Structurally distributed surface sites tune allosteric regulation. *eLife* **10**, e68346 (2021), publisher: eLife Sciences Publications, Ltd, doi:10.7554/eLife.68346, https://doi.org/10.7554/eLife.68346.

70. M. Leander, Y. Yuan, A. Meger, Q. Cui, S. Raman, Functional plasticity and evolutionary adaptation of allosteric regulation. *Proceedings of the National Academy of Sciences* **117**, 25445–25454 (2020), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.2002613117, https://www.pnas.org/doi/10.1073/pnas.2002613117.

71. Z. Sailer, M. Harms, Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. *Genetics* **205**, 1079–1088 (2017), doi:10.1534/genetics.116.195214.

72. Z. Sailer, M. Harms, High-order epistasis shapes evolutionary trajectories. *PLOS Computational Biology* **13**, e1005541 (2017), publisher: Public Library of Science, doi:10.1371/journal.pcbi.1005541, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005541.

73. T. Dupic, A. Phillips, M. Desai, Protein sequence landscapes are not so simple: on reference-free versus reference-based inference. *bioRxiv: The Preprint Server for Biology* p. 2024.01.29.577800 (2024), doi:10.1101/2024.01.29.577800.

74. Y. Park, B. Metzger, J. Thornton, The simplicity of protein sequence-function relationships. *Nature Communications* **15**, 7953 (2024), publisher: Nature Publishing Group, doi:10.1038/s41467-024-51895-5, https://www.nature.com/articles/s41467-024-51895-5.

75. N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* **138**, 774–786 (2009), doi:10.1016/j.cell.2009.07.038, https://www.sciencedirect.com/science/article/pii/S0092867409009635.

76. J. Thornton, E. Need, D. Crews, Resurrecting the Ancestral Steroid Receptor: Ancient Origin of Estrogen Signaling. *Science* **301**, 1714–1717 (2003), doi:10.1126/science.1086185, https://www.science.org/doi/10.1126/science.1086185.

77. J. Dauparas, *et al.*, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022), publisher: American Association for the Ad-

vancement of Science, doi:10.1126/science.add2187, https://www.science.org/doi/10.1126/science.add2187.

78. M. Kirschner, J. Gerhart, Evolvability. *Proceedings of the National Academy of Sciences* **95**, 8420–8427 (1998), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.95.15.8420, https://www.pnas.org/doi/10.1073/pnas.95.15.8420.

79. J. Bloom, S. Labthavikul, C. Otey, F. Arnold, Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences* **103**, 5869–5874 (2006), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.0510098103, https://www.pnas.org/doi/full/10.1073/pnas.0510098103.

80. V. Hilser, E. Thompson, Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proceedings of the National Academy of Sciences* **104**, 8311–8315 (2007), publisher: National Academy of Sciences Section: Biological Sciences, doi:10.1073/pnas.0700329104, https://www.pnas.org/content/104/20/8311.

81. J. Kohn, K. Deshpande, E. Ortlund, Deciphering Modern Glucocorticoid Cross-pharmacology Using Ancestral Corticosteroid Receptors. *The Journal of Biological Chemistry* **287**, 16267–16275 (2012), doi:10.1074/jbc.M112.346411, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3351298/.

82. L. Wheeler, S. Lim, S. Marqusee, M. Harms, The thermostability and specificity of ancient proteins. *Current Opinion in Structural Biology* **38**, 37–43 (2016), doi:10.1016/j.sbi.2016.05.015, https://www.sciencedirect.com/science/article/pii/S0959440X16300501.

83. H. Strobel, E. Horwitz, J. Meyer, Viral protein instability enhances host-range evolvability. *PLoS genetics* **18**, e1010030 (2022), doi:10.1371/journal.pgen.1010030.

84. W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey, M. Klein, Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926–935 (1983), doi:10.1063/1.445869, https://doi.org/10.1063/1.445869.

85. J. Fass, *et al.*, Quantifying Configuration-Sampling Error in Langevin Simulations of Complex Molecular Systems. *Entropy* **20**, 318 (2018), number: 5 Publisher: Multidisciplinary Digital

Publishing Institute, doi:10.3390/e20050318, https://www.mdpi.com/1099-4300/20/5/318.

86. J. Wang, R. Wolf, J. Caldwell, P. Kollman, D. Case, Development and testing of a general amber force field. *Journal of Computational Chemistry* **25**, 1157–1174 (2004), doi:10.1002/jcc.20035.

87. J. Wang, W. Wang, P. Kollman, D. Case, Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* **25**, 247–260 (2006), doi:10.1016/j.jmgm.2005.12.005, https://www.sciencedirect.com/science/article/pii/S1093326305001737.

88. E. Archer, I. Park, J. Pillow, Bayesian Entropy Estimation for Countable Discrete Distributions. *Journal of Machine Learning Research* **15**, 2833–2868 (2014), http://jmlr.org/papers/v15/archer14a.html.

89. M. Homan, A. Gelman, The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623 (2014).

90. J. Hunter, Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9**, 90–95 (2007), doi:10.1109/MCSE.2007.55.

91. L. Schrödinger, W. DeLano, PyMOL, http://www.pymol.org/pymol.

92. K. Thyng, C. Greene, R. Hetland, H. Zimmerle, S. DiMarco, True Colors of Oceanography: Guidelines for Effective and Accurate Colormap Selection. *Oceanography* **29**, 9–13 (2016), doi:10.5670/oceanog.2016.66, https://tos.org/oceanography/article/true-colors-of-oceanography-guidelines-for-effective-and-accurate-colormap.

93. M. Mirdita, *et al.*, ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679–682 (2022), publisher: Nature Publishing Group, doi:10.1038/s41592-022-01488-1, https://www.nature.com/articles/s41592-022-01488-1.

Acknowledgments

The authors would like to thank Madhav Mani, Kimberly Reynolds, Prashant Mishra, and Marielle Russo for feedback that significantly improved the manuscript, and Mark Kriegel and Yves Muller for sharing raw data on individual calorimetry replicates.

Funding: This work was supported by the NIH R35GM150897-01 and Sloan Foundation G-2024-22449 (M.M.L.) and a NIH molecular biophysics T32 GM131963(K.T.).

Author contributions: K.T. and M.M.L. conceived the work, K.T. performed simulations and analysis, K.T. and M.M.L wrote the paper.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: Code used for calculations will be deposited on GitHub (link to be added).

Supplementary materials

Materials and Methods

Supplementary Text

Figs. S1 to S6

Table S1

References *(7-93)*

Supplementary Materials for

Entropy tree networks of residue dynamics encode protein allostery

Kaitlin Trenfield, Milo Lin*

*Corresponding author. Email: Milo.Lin@UTSouthwestern.edu

This PDF file includes:

Materials and Methods

Supplementary Text

figures S1 to S6

Table S1

Materials and Methods

Molecular dynamics simulations

All simulations presented here were performed using OpenMM7 using the Amberff14SB force field (*42*), TIP3P water model (*84*), Langevin integrator (*85*) with a 2 fs timestep, and a Monte Carlo barostat using Tesla V100 32GB GPUs on the UT Southwestern high performance computing cluster. Ligands were parameterized using the GAFF in Antechamber (*86, 87*). All systems were solvated in boxes with a minimum of 1.5 nanometer separation between solute and box edge. Neutralizing ions were added, followed by additional sodium and potassium ions chosen to approximate a 50 mM NaCl buffer for all PDZ domain and CRIPT simulations and a 170 mM NaCl buffer for all steroid receptors. All systems were energy-minimized using OpenMM default settings, subject to a harmonic restraint of 100 kcal mol^{-1} on positions of all alpha carbons. Equilibration proceeded by heating at constant volume from 98K to 298K while relaxing harmonic restraints to zero at a constant rate over the course of 1 nanoseconds, followed by an equilibration time of 100 nanoseconds in an NPT ensemble. All PDZ domain and CRIPT simulations were sampled at 1 nanosecond intervals. Steroid receptor simulations were sampled at 0.5 nanosecond intervals. Parallel trajectories were run for each system and concatenated into a single trajectory prior to analysis. Details of simulations are provided in Supplemental Table S1.

Inferring residue states by clustering von Mises mixture model components with DBSCAN

A configuration of a residue is a vector, $\mathbf{\Theta} = [\theta_1 ... \theta_M]^T$, whose entries are the bond angles and torsion angles needed to specify the positions of its atoms. We model the probability distribution over configurations using a mixture of von Mises distributions,

$$p(\mathbf{\Theta}) = \sum_{k=1}^{K} p(k) \prod_{m=1}^{M} \frac{\exp\left(\kappa_{km} \cos(\theta_m - \mu_{km})\right)}{I_0(\kappa_{km})}, \tag{S1}$$

where $K = 100$ for all residue types, with the aim of choosing a number of mixture components that will greatly exceed the number of modes of each distribution. We fit the mixture model, initialized with the spherical k-means^{++} algorithm, to the configurations observed in the trajectory in the usual way, by maximizing the likelihood using the expectation-maximization algorithm, to

obtain parameter sets $\{p(k)\}$, $\{\mu_{km}\}$, and $\{\kappa_{km}\}$, which are the posterior probabilities, location parameters, and inversion dispersion parameters of each von Mises component.

To arrive at a set of states that is adapted to the observed statistics, rather than pre-determined by a chosen number of clusters, we turn to the DBSCAN algorithm. However, instead of applying DBSCAN to the configurations themselves, we apply DBSCAN to the mixture components, and then assign configurations to DBSCAN clusters using the mixture model. To use DBSCAN, we need only to supply a metric that measures the distance between components, a cutoff distance, and the minimum probability mass of a so-called "core point". For the metric, we used a measure of the overlap between mixture components,

$$\mathcal{D}(k, k') = 1 - \frac{\sum_t \sqrt{p(k|\mathbf{\Theta}(t))p(k|\mathbf{\Theta}(t))}}{\sqrt{p(k)p(k')}}, \tag{S2}$$

where $p(k|\mathbf{\Theta}(t))$ is the posterior probability that the component generating observation $\mathbf{\Theta}(t)$ is k, which is calculated from the component likelihoods using Bayes' theorem, and $\langle ... \rangle$ denotes the expected value.

The distance parameter for DBSCAN, for each residue, to one minus the standard deviation of the pairwise overlaps, $\sqrt{p(k|\mathbf{\Theta}(t))p(k|\mathbf{\Theta}(t))}$. We set the minimum mass parameter, which is the minimum probability of a core point in the present application, in DBSCAN to 0.01. The number of observations for all of our trajectories was greater than $10,000$, so this also guarantees that for any pair of residues, the expected number of counts, $\langle n(x_u, x_v) \rangle$, is greater than 1 for any pair of core points under the hypothesis of independence, $p(x_u, x_v) = p(x_u)p(x_v)$. We note that this is a lower bound; in practice, few states have such small numbers of counts.

The output of this step is a set of conformations, C, which is a partition of the von Mises mixture components. Each observation, $\mathbf{\Theta}(t)$, is classified to the state with the maximum likelihood of generating it.

Maximum likelihood tree inference

If we have a distribution, $p(X_1 ... X_L)$, which is not necessarily a tree, then the tree minimizing the information loss (called the Kullback-Leibler divergence, D_{KL}) from this distribution is just the spanning tree of $X_1 ... X_L$ with the maximum mutual information, which can be shown with some

algebra:

$$
\begin{aligned}
D_{KL}[p||p_T] &= \sum_{x_1 \ldots x_L} p(x_1 \ldots x_L) \log \frac{p(x_1 \ldots x_L)}{p_T(x_1 \ldots x_L)}, \\
&= -S(X_1 \ldots X_L) - \sum_{x_1 \ldots x_L} p(x_1 \ldots x_L) \log p_T(x_1 \ldots x_L), \\
&= -S(X_1 \ldots X_L) - \sum_{x_1 \ldots x_L} p(x_1 \ldots x_L) \left(\sum_{u \in \text{Residues}} \log p(x_u) + \sum_{(u,v) \in \text{edges}} \log \frac{p(x_u, x_v)}{p(x_u) p(x_v)} \right), \\
&= -S(X_1 \ldots X_L) + \sum_{u \in \text{Residues}} S(X_u) - \sum_{(u,v) \in \text{edges}} I(X_u, X_v),
\end{aligned}
$$

where the last line follows from the fact that the single and pairwise marginals of p_T are equal to those of p. The first two terms on the last line are the sum of the marginal entropies, $S(X_u)$, of the residues minus the true entropy of p. The only term that depends on the structure of the tree is the last one which is the summed mutual informations. The tree which minimizes the information loss from the true distribution is therefore the one which maximizes the sum of mutual informations. Moreover, the Kullback-Leibler divergence is the expected difference in log-likelihood per observation between the true distribution and the tree; therefore, the tree that generates the joint distribution closest to the true distribution is the maximum information spanning tree whose component distributions, $\{p(x_u)\}$ and $\{p(x_u, x_v)\}$, are equal to their counterparts in the distribution $p(x_1 \ldots x_L)$.

Posterior distribution of the entropy

Optimizing the likelihood of a mixture model using the expectation-maximization algorithm does not lead to a global optimum, but rather to a fixed point of the update rule determined by the randomly initialized parameters. Random variation in the mixture model parameters propagates through all subsequent steps in the analysis to produce random variation in the final estimate of the conformational entropy. We therefore derived a heuristic method for estimating the distribution of plausible conformational entropies from a single tree which encapsulates this uncertainty. Although our treatment of uncertainty is somewhat heuristic, we find empirically that it captures the variability associated with random initializations.

We model the posterior distribution of the entropy conditional on a single tree, T, using a normal distribution,

$$p(S|T) = \mathcal{N}(\mu_T, \sigma_T^2).$$

Postrior mean. The posterior mean, μ_T, is the expected entropy of the tree distribution given the observed counts $\{\mathbf{n}_u\}$, $\{\mathbf{n}_{uv}\}$ of single-residue and residue-pair states for branches included in the tree. This is just the sum of expectations of residue entropies and branch mutual informations. We compute these expectations with respect to Dirichlet distributions parameterized by the counts. The expected value of the entropy of a realization of a Dirichlet distribution parameterized by a vector of counts, $\mathbf{n}$, is given by the expression (*88*):

$$\langle S|\mathbf{n}\rangle = \psi_0(N+1) - \sum_{k=1}^{K} \frac{n_k}{N}\psi_0(n_k+1). \tag{S3}$$

where n_k denotes the k-th state, K is its dimension, N is the number of states, N is the total number of observations, and $\psi_0(*)$ is the polygamma function of order zero (also called the digamma function), defined as $\psi_n(z) = \frac{d^{n+1}}{dz^{n+1}}\left(\ln \int_0^\infty t^{z-1}e^{-t}dt\right)$. It is straightforward to then compute the expected mutual informations as

$$\langle I|\mathbf{n}_{uv}\rangle = \langle S|\mathbf{n}_u\rangle + \langle S|\mathbf{n}_v\rangle - \langle S|\mathbf{n}_{uv}\rangle, \tag{S4}$$

leading to the expression for the posterior mean entropy of the tree:

$$\mu_T = \sum_{u\in\text{residues}} \langle S|\mathbf{n}_u\rangle - \sum_{uv\in\text{branches}} \langle I|\mathbf{n}_{uv}\rangle. \tag{S5}$$

Posterior variance. In estimating the uncertainty of the entropy we hope to capture the expected squared deviation of μ_T from its (unknown) true value, which is the value that would be obtained by initializing the program many times and averaging. Our posterior variance aims to account for three sources of uncertainty 1) fluctuations in the number of observed counts for single and pairwise residue states that arise due to finite sampling 2) uncertainty due to the use of the use of the maximum likelihood estimator in determining the tree structure 3) uncertainty about the entropy that arises due to uncertainty about the optimal tree structure itself. This led us to the following decomposition of the variance:

$$\sigma_T^2 = \sigma_{count}^2 + \sigma_{bias}^2 + \sigma_{structure}^2.$$

The counting variance is accounted for using the following expression for the expected squared entropy of a realization from a Dirichlet distribution (*88*),

$$\langle S^2|\mathbf{n}\rangle = \sum_{i\neq j} \frac{n_i n_j}{N(N+1)} A_{ij} + \sum_i \frac{n_i(n_i+1)}{N(N+1)} B_i, \tag{S6}$$

$$A_{ij} = \Big(\psi_0(n_j+1) - \psi_0(N+2)\Big), \Big(\psi_0(n_i+1) - \psi_0(N+2)\Big) - \psi_1(N+2) \tag{S7}$$

$$B_i = \Big(\psi_0(n_i+2) - \psi_0(N+2)\Big)^2 + \psi_1(n_i+2) - \psi_1(N+2), \tag{S8}$$

from which the variance is obtained as $\langle S^2|\mathbf{n}\rangle - \langle S|\mathbf{n}\rangle^2$. We approximate the counting fluctuations in the residue entropies and mutual informations as independent, so that their contributions to the total variance are simply additive.

We estimate the bias component of the expected squared deviation as

$$\sigma^2_{bias} = (\langle S|T\rangle - (\hat{S}_{MLE}|T))^2,$$

where $(\hat{S}_{MLE}|T)$ is the entropy of the tree distribution when each residue entropy and mutual information are calculated using the plug-in estimator $\hat{S}(\mathbf{n}) = -\sum_k \frac{n_k}{N} \log \frac{n_k}{N}$.

We emphasize that we do not attempt to make explicit bias corrections, but only to obtain an estimate of its contribution to the mean-squared error. We give a few justifications for this. First, we observed that the estimated biases for the individual entropies and mutual informations are typically not much larger than the estimated standard deviations, which led us to expect that approximating the bias as a component of the variance might lead to more accurate interval estimates than attempting an explicit correction. Second, the true shape of the posterior distribution of the entropy is unknown, so incorporating an estimate of bias into the uncertainty reflects some degree of uncertainty about which estimator is most appropriate given the data without imposing any prior (or hyperprior) distribution on the counts vector. Third, and most importantly, our broader goal is to accurately capture differences in entropy, as total protein entropies are not observable. Towards this end, we considered explicit bias correction to be undesirable, as it is not generally possible to distinguish between bias that due to differing numbers of states and a genuine entropy difference due to a change in the number of occupied free energy wells. We therefore judged it more appropriate to account for bias as a source of uncertainty than to attempt corrections.

The structural uncertainty aims to capture the variation in the entropy that originates from changes in the structure of the maximum likelihood tree that arise from different partitions of the configurational landscapes of residues, which result in correlated differences in the presence of branches connecting distinct residues. In doing this, we also account, to some extent, for the possibility of spurious mutual information between residues that arises due to undersampling of slow degrees of freedom. Let $\hat{S}_{-u}$ denote the entropy estimate obtained by deleting the u-th residue and estimating the entropy by applying the Chow-Liu algorithm to the remaining residues and $\bar{S}_{-u} = \frac{1}{L}\sum_{l=1}^{L}\hat{S}_{-u}$ denote the empirical average of these replicates. We estimate the structural variance as the empirical variance of all replicates formed this way:

$$\sigma^2_{structure} = \frac{1}{L}\sum_{u=1}^{L}(\hat{S}_{-u} - \bar{S}_{-u})^2.$$

In spite of the heuristic nature of this decomposition, the credibility intervals produced this way appear to have good frequentist coverage in our sample, with 93.7% of the single-tree 95% credible intervals containing the average entropy over 21 fits for a representative set of eight thermodynamic states (Supplemental figure S5).

Bayesian model for allosteric mutation predictions.

We modelled the number of allosteric mutations $N_u^{\text{allosteric}}$ at each site using a binomial distribution,

$$N_u^{\text{allosteric}} \sim \text{Binomial}(N_u^{\text{mut}}, p_u),$$

where, N_u^{mut} is the number of point mutations observed at site u reported in the sequencing data (*24*) and p_u is the latent variable we hope to infer. For most residues, all 19 possible mutations were observed, but for a minority of residues only 17 or 18 mutant alleles were observed. This heterogeneity in allele counts is a significant source of uncertainty, as we cannot know for certain whether alleles were not observed because they were extremely deleterious to protein function (and if they were, whether this due to binding or abundance) or because the mutations simply happened not to arise in the population. To account for this, we modelled p_u, the site-specific probability of allosteric mutation as a random draw from the beta distribution which is determined by the predictors $\mathbf{x}_u = [1, S_u, \sum_{v\,\in N_u(T)} I_{uv}]^T$ according to

$$p_u^{\text{MLE}} = \text{logit}(\beta_u^T \mathbf{x}_u) = \frac{1}{1 + e^{-\beta_u^T \mathbf{x}_u}},$$
$$p_u \sim \text{Beta}\left(N_u^{\text{mut}} \cdot p_u^{\text{MLE}} + \frac{1}{2},\ N_u^{\text{mut}} \cdot (1 - p_u^{\text{MLE}}) + \frac{1}{2}\right).$$

This is similar to a logit-binomial regression model, which would be obtained by setting $p_u = p_u^{\text{MLE}}$. Instead, we have replaced the maximum likelihood step by simulating a Bayesian inference step using a Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$ (Jeffreys) prior. We used a multivariate normal distribution for the coefficient vector β with a diagonal covariance matrix and a weakly-informative Half-Cauchy prior with scale parameter equal to 25 for the variance.

Posterior distributions were estimated for the model parameters using Markov Chain Monte Carlo (MCMC) with 5000 draws using the No U-turn sampler (*89*).

Data Visualization

Plots were made in Python using Matplotlib (*90*). Protein visualization was done in PyMOL (*91*). Colormaps displayed on proteins are drawn from the cmocean libary (*92*).

Supplementary Text

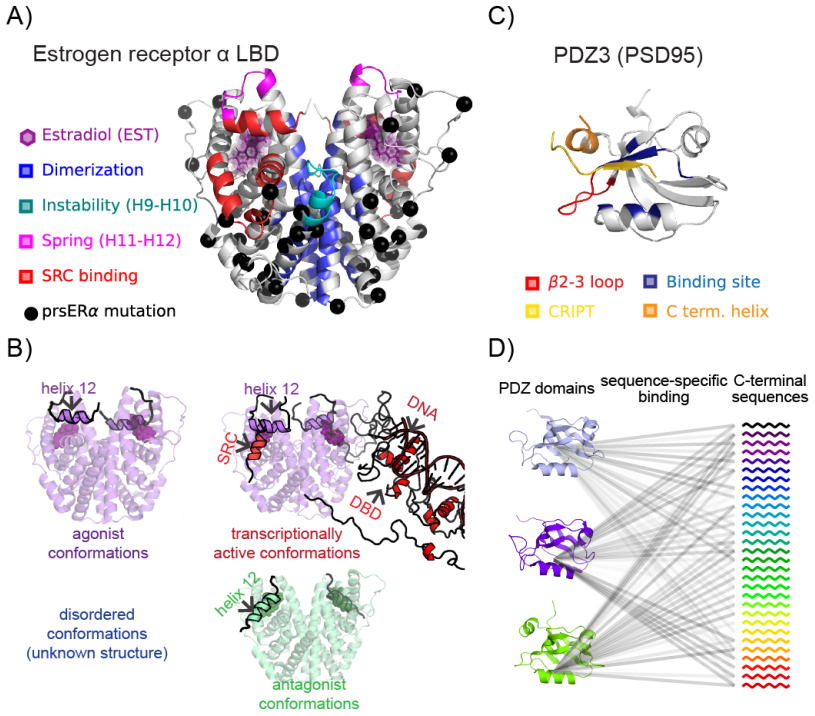

figure S1: **Allostery in estrogen receptor α and PDZ3** A) Estrogen receptor α ligand-binding domain (LBD, PDB: 1QKU) homodimer bound to estradiol. Sites colored in blue, teal, pink, and red influence activities from outside the hormone binding site. Sites shown in black spheres are those at which stabilizing mutations were introduced to create an engineered thermostable estrogen receptor. B) Major conformations of ERα. Binding of estradiol (purple) in the ligand-binding pocket stabilizes homodimers with an inward-facing H12 conformation poised for binding to steroid receptor coactivators (SRCs), leading to the formation of an active transcriptional complex on DNA estrogen response elements. Binding of selective estrogen receptor modulators (SERMs; green) induces an H12 conformation oriented away from the homodimer interface, precluding the binding of SRCs. The unliganded state is thought to be disordered, but is not known. Structures displayed are based on PDB entries 1QKU (agonist), 3ERT (antagonist), and SASBDB entry SASDDU8 (DNA bound). C) Structure of the third PDZ domain (PDZ3, PDB: 5HEB) from PSD95. The CRIPT peptide, binding site, and allosteric sites are labelled. D) Schematic illustrating an interaction network showing distinct profiles of specificities of PDZ domains for C-terminal peptide sequences Models shown are based on PDB entries 1KEF, 3ZRT, and 5HEB (top to bottom).

Table S1: **Details of molecular dynamics simulations.** State refers to the thermodynamic state that we modelled. PDB refers to the PDB code of the model(s) used to create the starting structure. Parentheses demarcate structures used as templates for a homology models in Modeller (m) or ColabFold (a) (*93*). Other than for the Y537S and prsERα simulations, mutations were made to restore the UniProt sequence corresponding to the PDB id. Residues not present in PDB structures shown were modelled using ”molecular dynamics”-based refinement using the Modeller software. All replicate trajectories were allowed to run for one week of clock time and are of similar length within each state.

State	PDB	Residues	Total time (μs)	Replicates
PDZ3	1BFE	302-402	10.7	6
PDZ3Δ7ct	1BFE	302-395	10.5	6
PDZ3Δ7ct-CRIPT	5HEB	302-395 (A), 1-9 (B)	11.4	7
PDZ3-CRIPT	5HEB	302-402 (A), 1-9 (B)	11.5	7
CRIPT	5HEB	1-9 (B)	12.5	4
ERα (apo)	(2B23,4Q13)^m	301-550 (A, B)	5.1	6
ERα-C3D	6VJD	301-550 (A, B)	4.9	6
ERα-EST	1QKU	301-550 (A, B)	5.5	6
ERα-GEN	2QA8	301-550 (A, B)	5.3	6
ERα-OHT	(3ERT)^a	301-550 (A, B)	5.3	6
ERα-Y537S (apo)	2B23	301-550 (A, B)	5.1	6
prsERα-EST	7NEL	302-548 (A, B)	5.3	6
prsERα-GEN	7NFB	302-548 (A, B)	5.7	6
AR-DHT	1T7T	669-920	5.3	4
ERβ-EST	3OLS	261-500 (A, B)	4.0	4
GR-HCY	(4P6X)^m	529-776	6.1	4
MR-AS4	2AA2	737-984	6.5	4
PR-STR	1A28	678-933	5.5	4

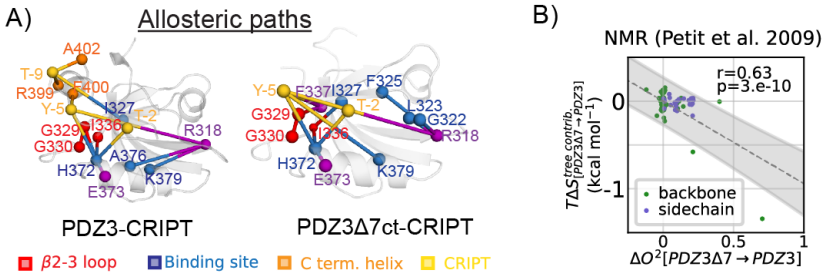

figure S2: **Allosteric pathways and correlations with experimental dynamics in PDZ3** A) Paths in trees connecting residues in the binding site (dark blue) the $\beta 2-\beta 3$ loop (red), CRIPT (gold), and the C-terminal helix (orange). Other residues on paths are shown in purple. B) Correlation between differeences and residue entropy contributions and S^2 order parameters (*41*) between PDZ3 and PDZ3Δ7-ct.

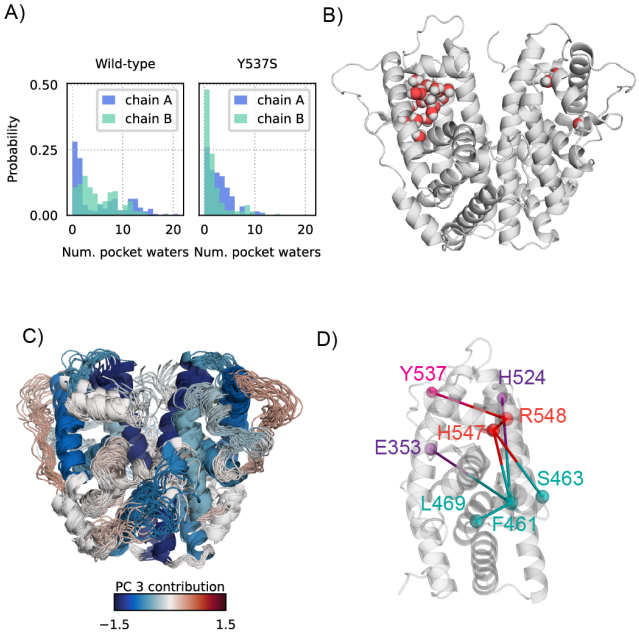

figure S3: A) Histograms of the number of waters in the binding pocket as defined by the convex hull of its atoms in each frame. Differences between the wild-type and Y537S pocket hydrations are significant with $p < 10^{-6}$. B) A snapshot of the wild-type apo state showing a lifted H12 conformation and hydrated binding pocket. C) 20 random snapshots drawn from our sampled conformational ensemble for the wild-type apo state, with secondary structural elements colored by their contribution to the third principal component, which represents the ordering of the receptor common to all ligand binding. Blue regions are generally ordered by ligand-binding, while red ones are disordered. D) Paths in the ciMIST tree connected the H9-H10 region, H12, Y537, and two binding pocket residues which form hydrogen bonds with estradiol.

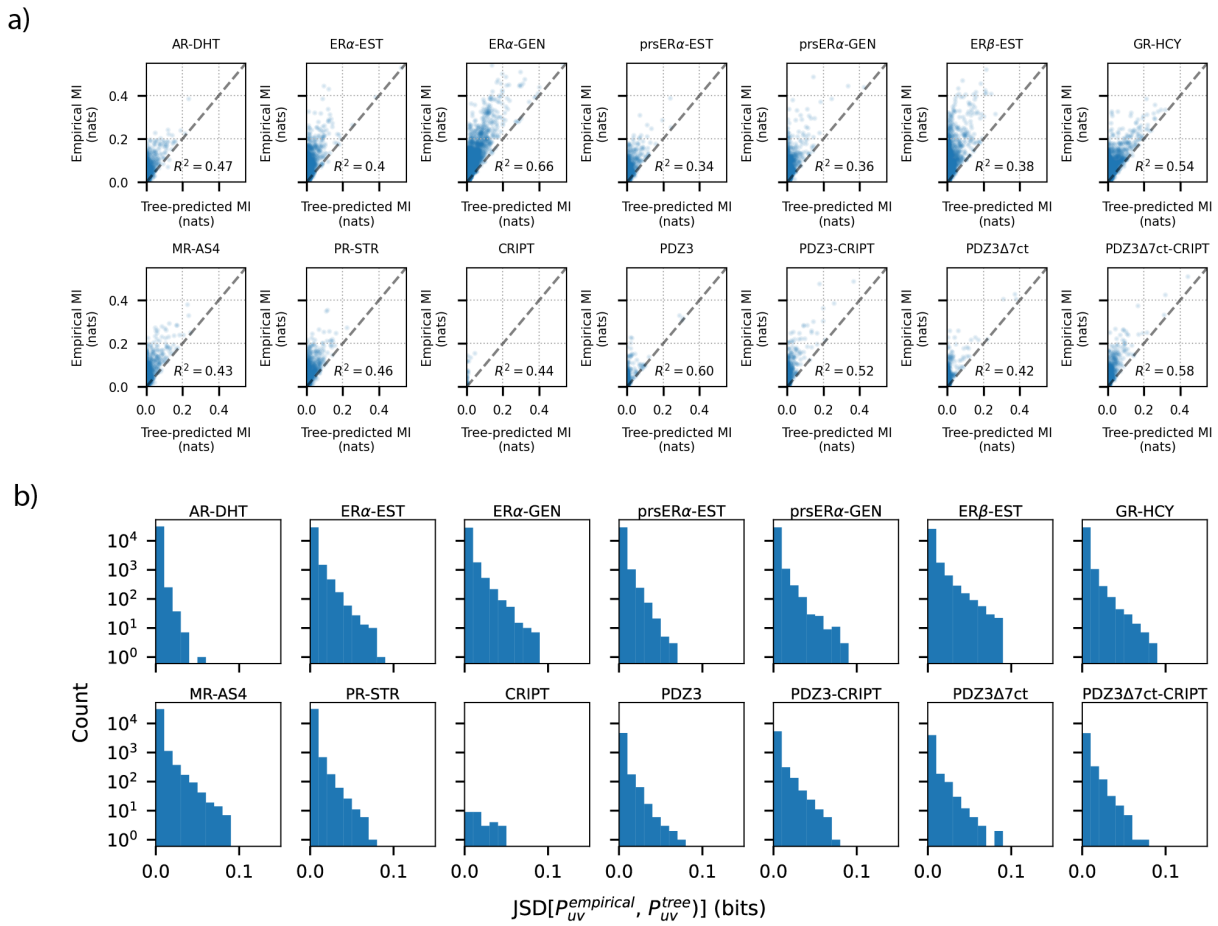

figure S4: **Consistency of tree-predicted pairwise conformational distributions with empirical conformational distributions for pairs of residues not retained in the tree.** A) Correlation between tree-predicted and empirical (maximum likelihood) mutual information for edges not included in the maximum information spanning trees for all systems analyzed in this work. B) Histograms showing frequecies of Jensen-Shannon divergences (JSD) between tree-predicted and empirical pairwise conformational distributions. JSDs are expressed in units of bits so that the maximum possible value is 1.

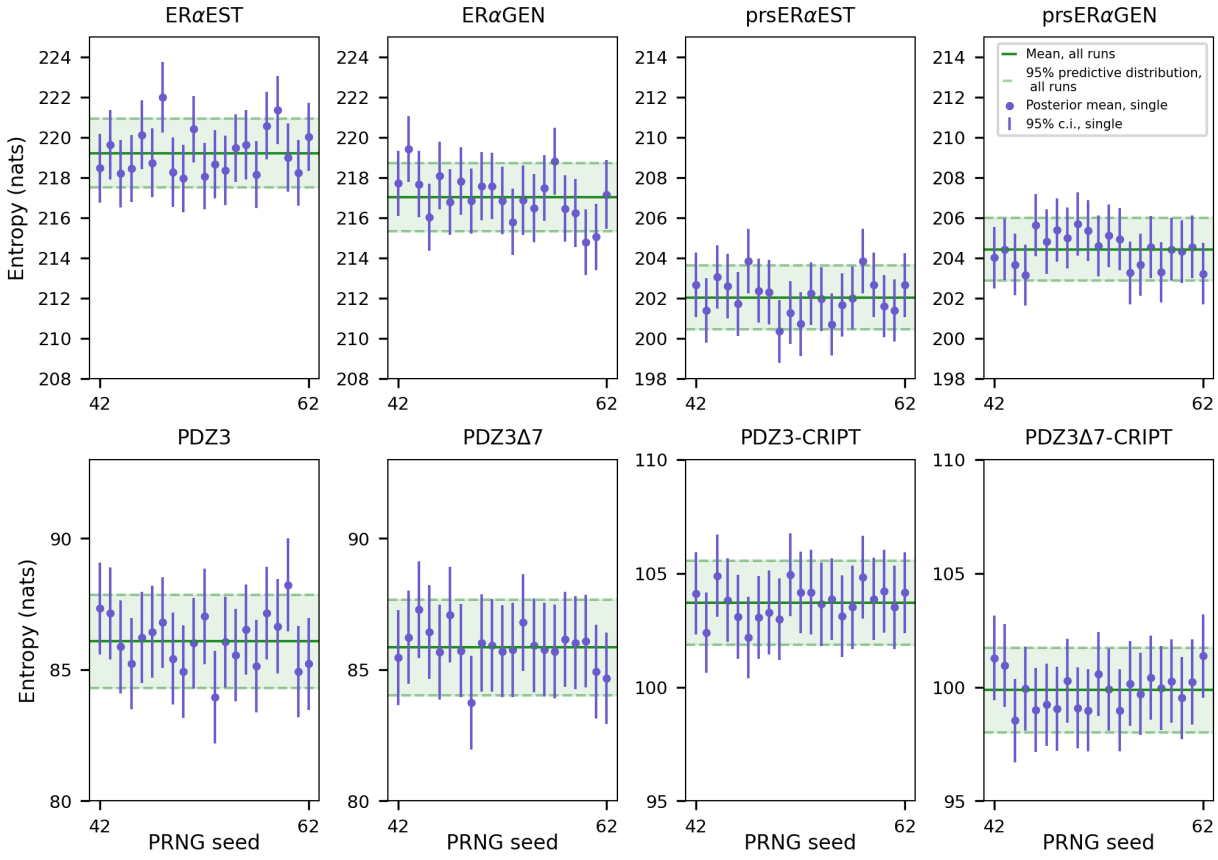

figure S5: Uncertainty quantification for tree entropies. Shown are the average entropies obtained from running the program with 21 different pseudo-random number generator (PRNG) seeds, the 95% posterior predictive distribution estimated from these runs and the posterior means and credibility intervals from the each run.

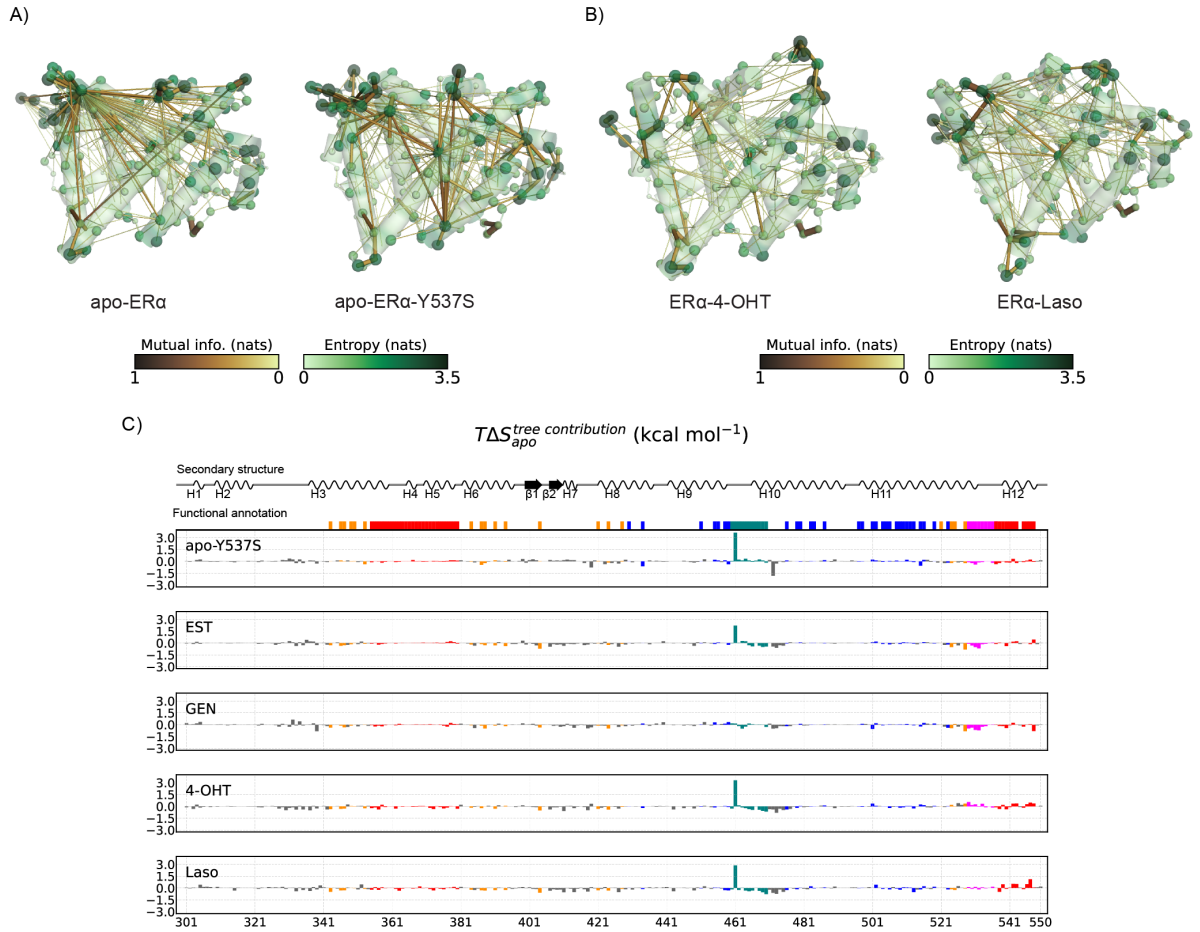

figure S6: a) Maximum information spanning trees for apo states of wild-type ERα and the Y537S variant. b) Maximum information spanning trees for ERα bount to anti-estrogen ligands 4-hydroxy-tamoxifen (4-OHT) and lasofoxifene (Laso). c) Differences in residue entropy contributions from the wild-type apo state of ER alpha for the Y537S mutant and all liganded states.