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 provide mechanistically-detailed maps of information transmission within proteins in a ligand- and mutation-sensitive manner. On a PDZ domain and the entire human steroid receptor family, 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 or undergoing covalent modifications, thereby altering their subsequent molecular interactions to reprogram host behavior, phenotype, and fate. For example, the transcription factor estrogen receptor alpha (ERα) is a steroid receptor that amplifies angstrom-scale differences in ligands at nanomolar concentrations into morphological changes apparent within days (*1, 2*). Much of ERα behavior, including tissue-specific responses to breast cancer drugs and drug-resistant phenotypes of cancer mutants, depends on changes in conformational dynamics (*3, 4*) and disorder (*4, 5*) outside the ligand-binding site, without tertiary structure change. Such long-range modulation of a protein's ensemble of conformational states, called ensemble allostery (*6*), is a widespread mechanism by which protein activities are regulated (*7–12*). More broadly, conformational ensemble encoding may explain how evolutionary variation reprograms protein function and allostery while conserving tertiary structure, such as in the steroid receptor family of transcription factors (*13–15*).

It remains mysterious how proteins reliably encode signals in conformational ensembles, limiting fundamental understanding of biological information-processing and therapeutic engineering (*12, 16–18*). Even for the third PDZ domain (PDZ3) of PSD-95 – a small model protein for ensemble allostery – proposed allosteric networks do not predict experimental observables of protein-wide mutational sensitivities from saturation mutagenesis, peptide-binding thermodynamics from calorimetry (ITC), or state-specific dynamics from nuclear magnetic resonance (NMR) spectroscopy (*19–21*). Fundamentally, demystifying ensemble allostery remains challenging because existing biophysical frameworks (*22–26*) do not map conformational ensembles to functional predictions.

We address this challenge by introducing the conformational inference Maximum Information Spanning Tree (ciMIST) method, which constructs predictive intraprotein communication networks directly from atomistic simulations of molecular dynamics by leveraging two key protein properties: 1) the discrete nature of free energy wells in dihedral angle degrees of freedom apparent in structure (*27*) and dynamics (*28–31*) and correlated within residues (*27*), and 2) sparse couplings between residues suggested by saturation mutagenesis (*20,32*) and coevolution (*32–35*). ciMIST qualitatively differs from existing approaches by identifying ensemble-specific discrete conformational states

of entire residues from bond and dihedral angle fluctuations and constructing an information tree network that is the optimal estimator (*36–38*) of the joint probability distribution over these residue states. This network model enables calculation of entropy contributions of arbitrary residue subsets, facilitating interpretation and prediction of experimental data.

Applying ciMIST to large-scale simulations of PDZ3 and the ERα ligand binding domain – both in multiple ligand-bound and mutational states – we demonstrate that ciMIST predicts entropy differences consistent with calorimetry with no fit parameters while correlating with experimentally-measured differences in protein dynamics. This predictive power arises from capturing cooperative interactions which would otherwise remain hidden. For PDZ3, we show quantitatively that cooperative fluctuations are necessary to predict mutational allostery. For ERα, ciMIST reveals that cooperative interactions within a region enriched with cancer mutations lead to compensation of helix formation by increased loop entropy as the major discriminant between liganded states. ciMIST also accurately predicts sequence-specific quantitative effects, such as the positional propensities for allosteric effects in a saturation mutagenesis library of PDZ3 (*20*) and the phylogenetic relationships within the human steroid receptor family (*39*). Taken together, these results show how entropy-mediated ensemble-to-function network maps can be decoded from molecular dynamics.

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 (*20, 40*) provided two key insights that shaped our modeling choices. First, most sites are insignificantly 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, significant allosteric mutations change binding by just 1-3 kcal mol^{-1}. Together, these suggest a sparsity of allosteric networks and a absolute scale for contributions of single residues to binding thermodynamics. 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.

Communication through single paths translates to the assumption that any residue conformation influences conformational probabilities at another site only through conditional probabilities along a unique path of direct pairwise interactions. This property is expressed formally as $p(x_v|x_u) = \prod_{(u',v')\in \mathrm{Path}(u\to v)} p(x_{v'}|x_{u'})$, where $p(x_v|x_u)$ denotes the probability that residue v assumes conformation x_v when residue u is fixed in conformation x_u and $\mathrm{Path}(u \to v)$ denotes the set of residue pairs along the path from residue u to v. Because this relationship holds 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:

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

A probability distribution of this form is called a tree distribution, reflecting the structure of the factor graph, wherein each marginal $p(x_u)$ and pairwise $p(x_u, x_v)$ probability is identified with a node and an edge, respectively.

The tree structure enables an exact expression for the (Gibbs-Shannon) protein conformational entropy as

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

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. The tree producing the joint probability distribution with minimum Kullback-Leibler divergence from the true distribution is just the maximum spanning tree of a network in which the mutual informations are regarded as edge weights, i.e. the maximum information spanning tree (*36, 38*) (Materials and Methods). Therefore, maximum likelihood inference of the network structure requires only that we estimate mutual informations using the empirical frequencies of residue conformations and calculate the maximum information spanning tree.

Equation 2 admits a useful decomposition into single residue contributions:

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

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 scales from single residues to the entire protein by summing individual and pairwise residue contributions.

This leaves the problem of identifying residue conformations X_μ. 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 bond and dihedral angle configurational distributions of each residue observed in our simulations, and next clustering the mixture components using DBSCAN (*41*), merging ”mixed states” of high component density overlap and grouping configurations in sparsely-populated regions into a single residue state (effectively a ”free energy plateau”) (Materials and Methods). This allows the number of conformations of each residue to be learned directly from data (Materials and Methods). The largest number of conformations ciMIST identified for any residue across all systems studied in this paper was 47, with 90 percent of residues having 20 conformations or fewer (supplemental figure S2).

ciMIST differs from previous dihedral-based entropy approaches in two key ways: directly inferring ensemble-specific conformations of entire residues from the data rather than by grouping dihedral angles into pre-defined bins, and construction of ensemble-specific tree structures that optimally account for cooperative residue fluctuations. Our residue conformational inference is crucial because the use of empirical frequencies to estimate the discrete entropy and mutual information has an intrinsic bias which grows with the number of states, largely because of sampling fluctuations in low-probability states (*42*). Our residue entropy estimates are therefore less biased than those based on dihedral-binning approaches (*43*).

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

1. Sample configurations of each residue 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 by applying DBSCAN to an overlap-based distance-measure (Materials and Methods) to identify discrete states of each residue (minimum conformer probability: 0.01).

4. Calculate residue entropies and residue-pair mutual informations from joint probability distributions over discrete residue states.

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 (*44*), MDTraj (*45*), and NetworkX (*46*) libraries, and the nerfax (*47*) implementation of the mp-NERF algorithm (*48*) 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 (Materials and Methods). All trees shown are derived from the tree whose entropy was closest to the average over 21 runs. For all proteins, mutants, and ligand-states studied, the ciMIST tree was an excellent generative model of pairwise mutual informations not directly retained in the tree, neither missing nor creating spurious significant mutual information pairs (figure S5). We note that all analyses reported herein used the same DBSCAN and von Mises mixture model hyperparameter settings.

To generate molecular dynamics data from which we inferred ciMIST models, we employed parallel simulations, generating microseconds of data per thermodynamic state from between 4 and 6 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 ensemble, so that full thermodynamic cycles required between 6 and 8 weeks.

For the remainder of the paper, we report entropy changes in terms of their contributions to binding free energy, that is, by multiplying by RT, where R is the gas constant and $T = 298\ K$ (as it is for all calorimetry experiments which we analyze).

A unified dynamics-based model for allostery in PDZ3

The allosteric network of PDZ3 (figure 2A) has been extensively profiled. Mutagenesis, structure, and coevolution have shown how residues in the $\beta 2-\beta 3$ loop shape specificity (*20,32,49*). NMR and calorimetry show how the presence of a non-conserved helix (H3) stabilizes apo-state dynamics, reducing the entropy lost when a peptide (CRIPT) binds (*50*). Yet, a unified thermodynamic explanation of these phenomena has remained elusive (*21*).

We performed MD simulations (*51, 52*) (Materials and Methods, Supplemental Table S1) of $PDZ3^{302-402}$ and (henceforth: PDZ3) and $PDZ3^{302-395}$ (PDZ3Δ7ct, which lacks H3) in the presence and absence of CRIPT as well as simulations of the isolated CRIPT peptide. For both CRIPT-bound states, our models were based on PBD entry 5HEB, while apo states were constructed from 1BFE. Trajectories were downsampled at 1 ns^{-1} frequencies. Each ensemble was obtained by concatenating the final 1.75 microseconds from six independent replicate trajectories, corresponding to over 100 nanoseconds of equilibration time for each trajectory and resulting in 10.5 microseconds of simulation time per thermodynamic state. This constitutes exhaustive simulation of states in the experimentally-characterized thermodynamic cycle that revealed H3-mediated entropic control of the PDZ3-CRIPT interaction (*50*), while also providing a suitable point of comparison to mutagenesis experiments performed on $PDZ3^{312-394}$.

Applying ciMIST to these simulation data, we observed that the $\beta 1-\beta 2$ and $\beta 2-\beta 3$ loops exhibit particularly strong conformational couplings in all states (figure 2B), a number of which highlighted known interactions. The central information hub in the $\beta 1-\beta 2$ loop is ARG318, which is known to modulate binding kinetics in response to ionic strength (*53*). GLY330 acts as a $\beta 2-\beta 3$ loop network hub, consistent with its known mutational sensitivity (*20, 32, 49*) and the role of this site in shaping specificity for distinct peptide classes (*49*).

Using equation 3, we explored contributions of different regions of PDZ3 to conformational entropy differences (figure 2 C-D). The CRIPT peptide contributes a similar entropy loss, −3.35 kcal mol^{-1} upon binding to PDZ3Δ7−ct and −3.56 kcal mol^{-1} upon binding to PDZ3, imposing a baseline entropic penalty from peptide binding. PDZ3 partially offsets this with an increase of 1.63 kcal mol^{-1} in entropy contributed by C-terminal loop residues 399-402, an effect which is visible in randomly-sampled structures from the ensembles (figure 2 C). The $\beta 2-\beta 3$ loop contributes a

loss of entropy which is greater for PDZ3Δ7ct (-1.27 kcal mol^{-1}) than PDZ3 (-0.76 kcal mol^{-1}) due to loss of pre-stabilization by H3.

We found that the ciMIST network inferred from the ensemble of the CRIPT-bound state of PDZ3Δ7−ct quantitatively predicted allosteric mutational sensitivities in a deep mutational scan (*20*). We used our residue-level thermodynamic features from this tree, the marginal entropy and the sum of mutual information with all neighbors, as inputs to a Bayesian generalized linear model (Materials and Methods) for the expected number of allosteric mutations at each site. We obtained an excellent fit to saturation mutagenesis data (figure 2E). Posterior distributions of model coefficients showed that higher network mutual information positively correlated with allosteric mutational propensity (95% credible interval: $[0.70,\ 2.61]$), validating that our residue conformations and conformational couplings capture ensemble features that promote stable CRIPT-binding. On the other hand, the posterior distribution for the coefficient of residue marginal entropy concentrated mostly on negative values (95% credible interval: $[-1.30,\ 0.20]$), indicating that accounts of cooperative fluctuations are necessary to predict mutational sensitivities in flexible protein regions. 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 a single state without the need to simulate the dynamics of the mutants.

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 backbone order parameters between the apo states (Pearson $r = -0.69$, $p < 10^{-8}$), but not side-chain order parameters ($r = -0.09$, $p = 0.67$), (supplemental figure S3) (*50*). This partially supports the model suggested by NMR (*50*) attributing the change in binding entropies to differences in apo-state ensembles, with the main difference being that our analysis does not support a major role for residues whose side chain order parameters they measured. However, these residues exhibit low dynamic range in both the entropy and order parameter differences, making significant correlation between these inherently difficult to detect. Our calculations are in better quantitative agreement with the calorimetry data than the NMR entropy meter, which estimates conformational entropy from side chain order parameter changes (*54*).

A unified model for entropic allostery in estrogen receptor α.

Allosteric regulation of estrogen receptors is a major factor in human health and disease that shapes outcomes in multiple cell and tissue types (*5, 55–57*). 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 (*3–5*), homodimerization (*58, 59*), and stabilization of agonist helix 12 (H12) tertiary structure accommodating binding of steroid receptor coactivators (SRCs), enabling assembly of active transcriptional complexes on DNA (*60*) (figure 3A, Supplemental figure S1A-B). Inhibition by anti-estrogens used as breast cancer therapies triggers an alternative (antagonist) H12 tertiary structure, occluding SRC binding (*5, 61*). Without a bound ligand, wild-type ERα displays signatures of conformational disorder which are likely central to maintaining an inactive state (*5*). Mutations in the ”instability” (residues 460-471) and (”spring”, residues 529-538) regions engender hormone-independent stabilization of active homodimers, causing aggressive and drug-resistant breast cancer phenotypes (*4, 5, 62*). Therefore, both local and global conformational instabilities are central to ERα function. We applied ciMIST to the ERα-LBD (henceforth we refer to this as ERα for brevity) to reveal the cooperative interactions underlying these ligand- and mutation-dependent instabilities.

We performed simulations of ERα homodimers, reflecting nanomolar-scale homodimer affinity (*58*) and micromolar-scale receptor concentrations in experimental datasets (*3, 63, 64*). For each thermodynamic state, we again simulated six replicates and concatenated these to obtain a single conformational ensemble, resulting total simulation times of 5 and 6 microseconds for each (Materials and Methods, supplemental table S1). We then concatenated the ensembles of individual 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 capturing cooperativity between domains.

Predicting the effects of thermostabilizing mutations on ERα binding entropies

To understand the role of local instability in ERα, we first used ciMIST to compare human ERα to an engineered thermostable estrogen receptor (prsERα, figure 3A) which shares many of its allosteric features (*64*). prsERα was designed by introducing 22 PROSS-suggested (*65*) amino

acid substitutions into the ERα sequence outside of major functional sites (*64*). While conserving binding affinities and ligand-dependent structure, prsERα displays altered binding entropy for genistein (GEN), a phytosterol found in soybeans. Genistein is allosterically discriminated from estradiol by human ERα (*58*), eliciting weaker phenotypic changes (*1, 66*), a possible consequence of natural selection for precise estrogen signaling against a background of similar environmental ligands (*67*). We simulated human ERα (models based on PDB: 1QKU, 2QA8) and prsERα (PDB: 7NEL, 7NFB), each in estradiol- and genistein-bound states, and analyzed each ensemble with ciMIST.

The H9-H10 loop region is the most significant information hub in the ciMIST networks for all four ensembles, with PHE461 especially prominent (figure 3B-D). In both human receptor trees, PHE461 displays singularly strong couplings with H9-H10 loop residues, H10 residues, and with TYR459 in H9, but these connections are stronger and more numerous when the receptor is bound to genistein than estradiol. Driven by this mutual information, 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}. On the other hand, H10 has less helical order and contributes more conformational entropy when bound to genistein than estradiol. Thus, the cost of helix formation is thermodynamically compensated by increased H9-H10 loop entropy due to decoupling of PHE461 from H10. Qualitatively similar couplings are observed in prsERα trees, but mutational pre-stabilization of H10 eliminated the PHE461/H10 entropic compensation.

We next 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 (*63*). We found strong correlation (Pearson $r = 0.58$, $p = 8 \times 10^{-5}$) between these (figure 3E), validating that our models capture experimentally-relevant ensemble shifts. Moreover, entropy contribution differences correlated better with HDX changes than residue marginal entropy differences (Pearson $r = 0.31$, $p = 0.052$), supporting that the experimental dynamics are intrinsically cooperative.

ciMIST entropy differences between estradiol and genistein-bound MD ensembles are within 1-2 kcal mol^{-1} of experimental values (*64*) 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 ($T\Delta\Delta S$) falls within statistical uncertainty of experiment (figure 3F), as expected due to the subtraction of likely similar solvent and ligand entropies.

By using ciMIST to compare ERα to prsERα – a similar engineered thermostable receptor – we found a ligand-dependent entropic compensation mechanism in the disordered H9-H10 region of ERα allosterically discriminates between similar ligands. Quantitative agreement between ciMIST predictions and two independent experimental datasets strengthens confidence in this mechanism.

Rewiring of the ERα allosteric network in the Y537S mutant

Having shown that ciMIST can yield accurate models of mutant-specific binding entropies in ERα, we wanted to know whether ciMIST could identify significant differences in the apo-state conformational ensembles of the constituitively-active Y537S somatic breast cancer mutant (*4,5,68*) and wild-type ERα, whose structure has not been determined. We applied ciMIST to simulations of apo states of ERα_{Y537S} (PDB: 2B23) and wild-type ERα (homology model (*69*) based on PDB entries 2B23 and 4Q13). Agonist H12 position was maintained in both ensemble-averaged structures (figure 4A).

The wild-type ciMIST network featured direct connections between PHE461 and residues HIS524 and GLU353 (figure 4B, supplemental figure S4D), both of which form hydrogen bonds with bound ligands. 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 new information hubs.

Relative to wild-type, Y537S displayed increased PHE461 entropy contribution (from decreased PHE461-H10 coupling) and decreased H10 entroy contribution (from increased LYS472-H10 coupling). We also observed decreased entropy contributed by H3, H11, binding pocket residues, and the H11-H12 loop (figure 4C, supplemental figure S7C), consistent with *in vivo* Y537S activities comparable to those of the estradiol-bound wild-type (*4*). Conformational stabilization of the binding site and H12 positioning in Y537S correlated with decreased hydration of the hormone-binding pocket (Supplementary figure S4A-B). To our surprise, ciMIST predicts similar conformational entropies for these ensembles, suggesting that cooperative stabilization of H12 position and the protein core (figure 4C), rather than global stabilization, might confer hormone-independent Y537S

SRC-binding, transcriptional activation, and crystallization.

Local entropy compensation and global disorder in ERα regulation

To build a more comprehensive model validating the observed entropy compensation mechanism in ERα, we performed simulations of ERα bound to 4-hydroxy-tamoxifen and lasofoxifene (models based on PDB: 3ERT, 6VJD) (supplemental table S1). We then analyzed these with ciMIST and compared local and global changes in entropy from the wild-type apo ensemble to those we found for estradiol and genistein.

Protein conformational entropy differences from the apo state are between -8 and -14 kcal mol^{-1} for all liganded states (figure 5 C) – in the vicinity of the experimental binding entropies (*64*) 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 hydrophobic ligand desolvation (likely favorable), our calculations provides evidence that conformational entropy is a major determinant of measured binding entropies.

We applied principal component analysis to differences in residue entropy contributions between all ERα ensembles and the wild-type apo state (figure 5B-C). The leading principal component, accounting for 46.8 percent of variance in entropy contributions, captures PHE461/H10 entropy compensation, validating this as a general mechanism of ERα allosteric regulation (figure 5B) in both agonist and antagonist H12 conformations. The second component separates antagonist ligands from apo-Y537S, consistent with resistance of Y537S to anti-estrogen therapies (*4, 5*). The third component, onto which all ensemble models have positive loading, captures global receptor stabilization, including in the hormone-binding pocket.

We compared entropy contribution differences to differences in hydrogen-deuterium exchange (*3, 63*), this time using the wild-type apo state as a baseline. We found significant correlation (Pearson $r = 0.36$, $p = 3 \times 10^{-6}$) between entropy contribution and HDX differences (figure 5D) and weaker correlation using summed residue marginal entropies (Pearson $r = 0.3$, $p = 0.0001$). The most notable outliers from this correlation were in the beta sheet region (residues 403-408 and 403-410) for the two antagonist ligand states. While HDX is interpreted as measuring dynamics in ERα (*3, 4, 63, 70*), to our knowledge this is the first study to bridge the gap between HDX and calorimetry using an ensemble model. The correlation with experiment is notable considering that

neither our sampled MD ensembles nor ciMIST parameters were fit to these data.

Our analyses yield an ensemble model of entropic allostery in ERα. Our model quantitatively explains multiple observations – large losses of entropy (*64*) and of hydrogen-deuterium exchange (*3, 63*) – 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. Supporting entropic compensation in ERα are increased homodimerization and estrogen-independent activities conferred by somatic breast cancer mutations F461V and S463P (*62*). The H9-H10 region is also a site of proteolytic cleavage (*71*), which often depends on protein stability, suggesting ligand-dependent ERα PHE461/H10 entropic compensation may control ERα abundance. Thus, our approach may be valuable towards the design of therapeutics that accelerate receptor degradation through targeted conformational destabilization, which can show improved efficacy against mutations that resist other treatments (*70*). The consequences of these simulation data would have been non-obvious without ciMIST, which provided thermodynamic network models of cooperative residue dynamics that quantitatively explain mutant- and ligand-specific ERα function.

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

The steroid receptor family consists of estrogen receptors α and β, and four ketosteroid receptors: progesterone, androgen, glucocorticoid and mineralocorticoid receptors. These master transcription factors most likely evolved from an ancestral estrogen receptor through serial genome duplications (*39*). Despite adopting similar structures in the presence of their activating hormones, their allosteric mechanisms have diverged. While all active transcriptional complexes are dimers, the interface is not conserved, and thermodynamic cooperativity is observed at different stages of transcriptional complex assembly (*13–15*). Past MD analysis of dynamic contact paths to helix 12 support a mixture of conservation and divergence in ketosteroid receptor dynamics (*72*), but an analysis without prior assumption of an allosteric site and which includes estrogen receptors has not been done.

Having characterized the ERα-LBD signaling network, we ran MD simulations of the ligand binding domains of the five remaining steroid receptors, each bound to an endogenous agonist

hormone, and inferred trees with ciMIST (figure 6A). ERβ was simulated as a homodimer, while each all ketosteroid receptors were simulated as monomers, consistent with solution-state measurements showing dimerization affinities greatly exceeding physiological receptor concentrations for hormone-bound ketosteroid receptors, but not estrogen receptors (*13–15*). We used biotite (*73*) to align steroid receptor sequences and build a phylogenetic tree using the UPGMA algorithm, recovering the accepted phylogeny (*39, 74*). Then, using the sequence alignment to align the vectors of ciMIST-computed residue entropy contributions, we calculated Pearson correlation coefficients between residue entropy contributions for all receptor-pairs and clustered these matrices using the UPGMA algorithm, as we had done with the sequence identity. We found residue entropy contributions diverged in a manner that mirrored evolutionary divergences. In contrast, the phylogenetic tree was not captured by root-mean-squared-fluctuations (RMSF) nor by template modeling scores (TM-scores) of the crystal structures (figure 6B).

We also found that principal component analysis of aligned entropy contributions separated receptors by phylogeny along the first two components (figure 6C). The first principal component (44% of variance) separates estrogen from ketosteroid receptors; this is believed to be the first major evolutionary transition within the family (*39*). The second component (19% of variance) separates corticosteroid receptors from androgen and progesterone receptors – the second evolutionary branch point (*39*).

These results hint at a general principle: divergent information-processing networks reflect the evolutionary history of novel functional adaptations even as protein tertiary structure is conserved. Such divergence may be as necessary to precise steroidal coordination of development as are steroid-binding specificities themselves: just as estrogen signaling must differentiate estrogen from related ligands, active receptors must be distinguishable from each other to enact distinct transcriptional programs. Consistent with this, we find that principal components of the entropy contributions reflect functional reprogramming at evolutionary branch points.

Conclusions and Implications

The adage that protein structure encodes function neatly summarizes, but can obscure, the remarkable track record of structural models obtained from experimental data as basis for the design,

interpretation, and prediction of experiments – effectively, as theories of protein function. Here, we show that sparse networks of cooperative residue flucutations observed in microseconds of MD simulations can provide unified models of experimental data not explainable by structure-only models for protein domains of differing sizes and complexity. The predictive power of our features across multiple datasets and proteins shows that the conformational entropy networks we resolve are not statistical noise, but encodings of protein function in the simulated protein dynamics.

Our network models of ensemble allostery reveal interactions and structural mechanisms that reprogram protein function without major average tertiary structure change, a phenomenon theorized four decades ago (*75*) which has proven difficult to understand (*6, 12, 18*). Our ciMIST models of apo and liganded states of ERα illustrate a surprising range of protein conformational entropy which can be explained by broadly-distributed entropy changes. Our expectation is that a tree approximation should accurately estimate entropy differences which are dominated by a small fraction of pairwise differences in a general combinatorial expansion of the joint entropy in terms of entropies of all subsets of variables (*76, 77*). Biophysically, significant changes in three-way (and higher) entropies might require larger conformational rearrangements than we see here, such as global loss of tertiary structure. This raises questions about the nature of disordered protein states and suggest some ”separation of concerns” between tertiary structure and allosteric signaling.

Evolutionarily successful signaling protein folds may be more analogous to circuit boards – physical scaffolds capable of supporting multiple computations – than to circuits themselves, with allostery programmed by surface mutations. Conserved and coevolving residues, predictive of structural contacts (*34*), could contribute to allostery less through conformational change than by providing rigid packing constraints around which flexible residues shift in response to signals. Consistent with this view, mutagenesis of an engineered photosensitive dihydrofolate reductase (*78*) and a tetracycline repressor (*79*) – found allostery-tuning mutations primarily at non-conserved sites. Allostery as a low-order phenomenon would also entail that high-order epistasis (*80–82*), to the extent that it exists (*83*), mostly captures folding and expression effects rather than allosteric regulatory sites. Extending family-level analysis of entropy networks to new protein families may further explicate the relationship between evolutionary and conformational couplings. ciMIST analysis of laboratory-evolved proteins, particularly descendants of reconstructed ancestral protein sequences (*74*) or AI-designed sequences (*84*), might clarify relationships between stability,

evolvability, and allostery (*49, 79, 85–90*).

By providing models of conformational entropy which are predictive at multiple scales, ciMIST may be broadly relevant to clarifying the role of conformational entropy in information-processing proteins. Theoretical (*91*) and empirical considerations suggest this role should be substantial. Steroid receptors associate with heat shock protein complexes in vivo when their ligands are absent, indicating conformational instability (*92*). Antibody interactions exhibit large entropy losses coincident with broadly-distributed losses of hydrogen-deuterium exchange (*93*), while cryo-EM supports a role for antibody conformational entropy in antigen recognition (*17*). Across the tree of life, protein disorder correlates to information-processing: disordered protein regions are more prevalent in multicellular than unicellular organisms and more prevalent transcriptional regulation than any other process (*94–96*). Our analysis of ERα leads us to speculate that these may be examples of a tendency for proteins to lose conformational entropy with information acquisition, with more complex communications requiring larger entropy reservoirs from which to draw.

The effectiveness of our description of proteins as networks of discrete-state residue sensors suggests that it may be useful to conceptualize amino acids as communicating through digital conformational codes. If evolution has generated protein conformational ensembles which are sophisticated encoders of their environments, then these codes might appear unstructured to unaided viewers, like encrypted messages (*97*). We anticipate that approaches such as ours, which connect statistical regularities of conformational ensembles to functional predictions, will aid in the decryption of this code.

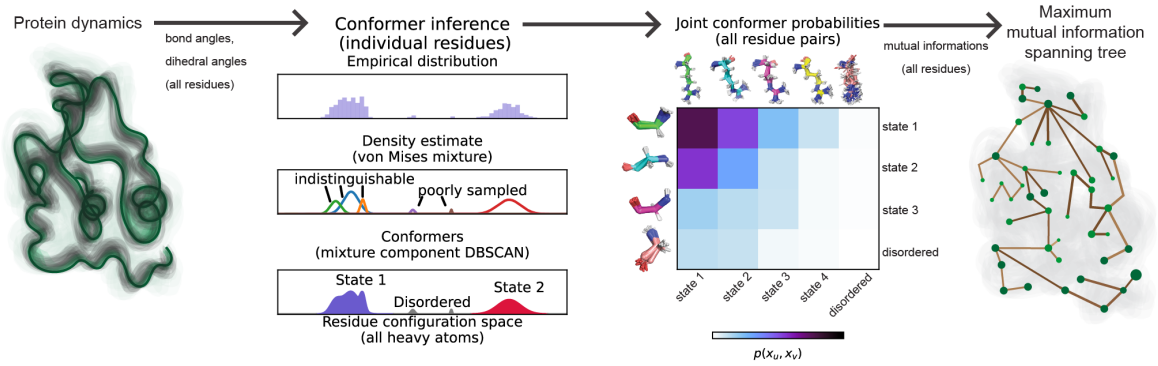

figure 1: **ciMIST infers maximum information spanning trees linking single-residue conformations into a global statistical model.** All frames of a molecular dynamics ensemble are mapped to internal coordinates. For each residue, a separate von Mises mixture model is fit to the observed residue configurations, 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. Conformational probabilities and mutual informations are computed for each pair of residues residue, and the maximum information spanning tree is then found.

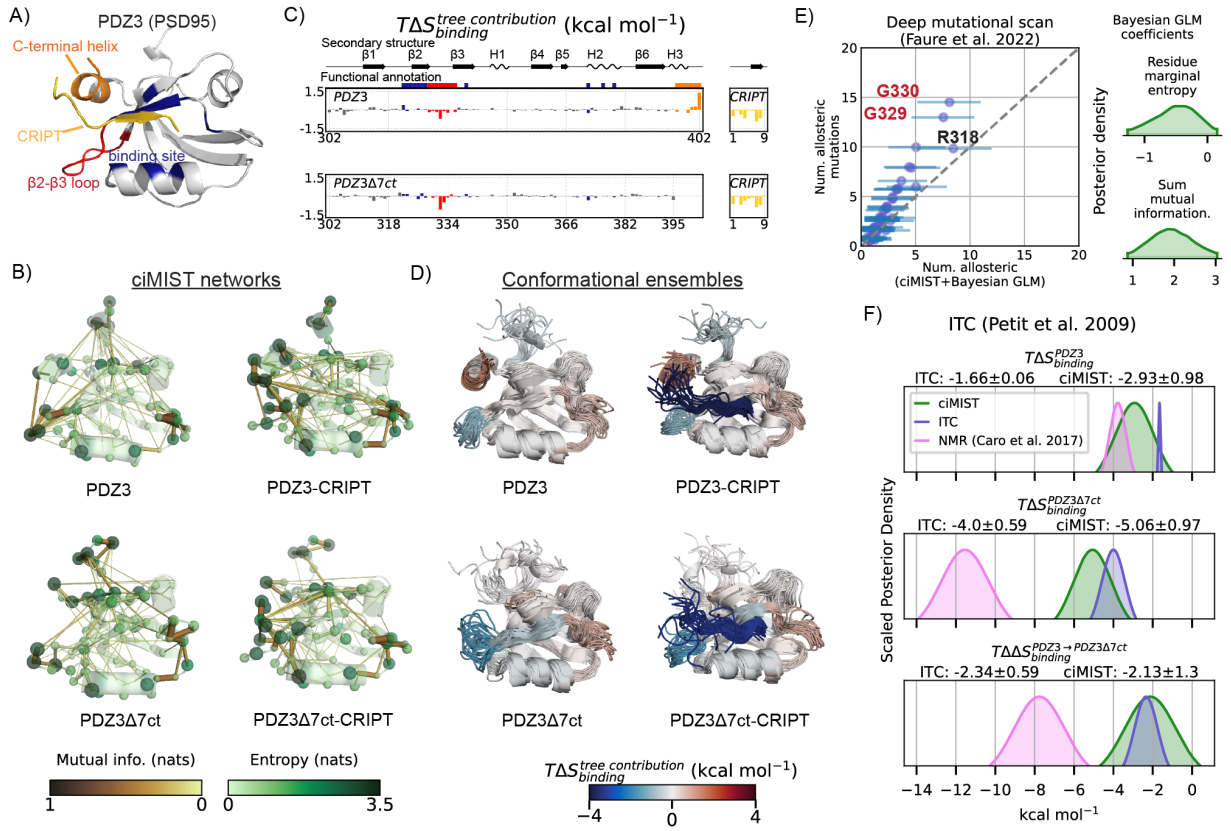

figure 2: **Connecting molecular dynamics to experimental signatures of allostery in PDZ3.** A) Structure of the third PDZ domain (PDZ3) from PSD95 bound to the CRIPT peptide. B) Maximum information spanning trees for PDZ3 (top) and PDZ3Δ7ct in apo (left) and CRIPT-bound (right) states. Sphere volumes of Cα atoms (shades of green) are proportional to residue marginal entropies, while volumes of links (shades of brown) are proportional to mutual information. C) Contributions of individual residues to the conformational entropy difference between apo and CRIPT-bound states. 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 classified as a single element. E) Prediction of allosteric mutational propensities in saturation mutagenesis from ciMIST residue thermodynamic features using a Bayesian GLM. Residue marginal entropies and summed mutual informations with tree neighbors from the CRIPT-bound state of PDZ3Δ7-ct were used as predictors. F) Posterior densities for whole-protein conformational entropy differences calculated by ciMIST, entropies measured with ITC (*50*), and conformational entropy differences estimated by NMR (*54*). ITC and NMR posteriors are Gaussians with reported means and standard deviations. $T = 298K$.

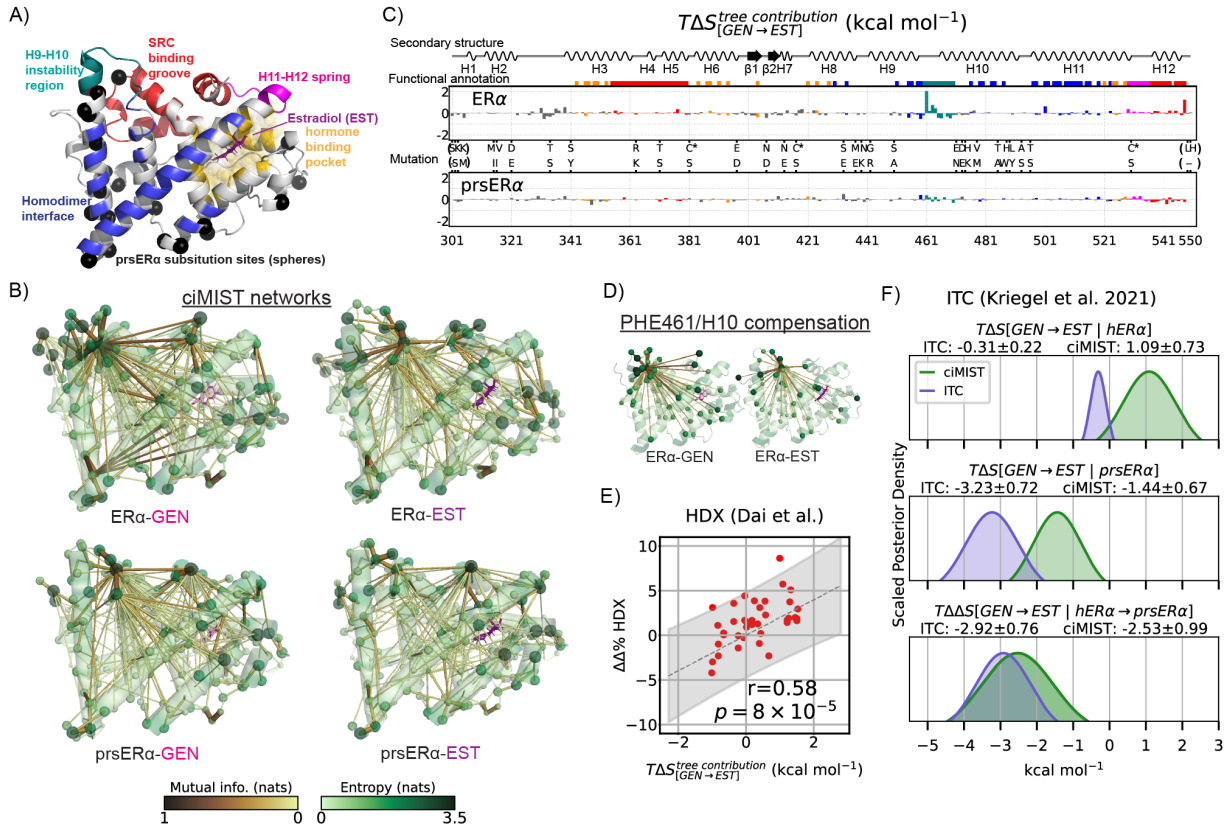

figure 3: **Entropic effects of 22 stabilizing mutations in ERα.** A) View of a single subunit of ERα with functional annotation. B) ciMIST networks for the ERα (top) and a prsERα (bottom) bound to genistein (left, pink sticks) and estradiol (right, purple sticks). Sphere volumes of Cα atoms (shades of green) are proportional to residue marginal entropies, while link volumes (shades of brown) are proportional to mutual information. Note that while simulations were of homodimers, ciMIST was applied to the aggregate statistics of both subunits. C) Differences in residue entropy contributions between genistein- and estradiol-bound states for ERα (top) and the prsERα (bottom). C* denotes positions of cysteine-serine substitutions made for calorimetry (*64*) but not in our simulations. Parentheses denote positions not present in the calorimetry experiment. Regions are color-coded as in (A) D) Connections in ciMIST networks between PHE461 and other residues for human ERα ensembles. E) Correlation between differences in contributions of protein segments to total conformational entropy and average percent differences in deuteration levels (*3, 63*) between estradiol and genistein-bound states of human ERα. F) Estimated posterior densities of entropy differences from ITC (*64*) and calculated using ciMIST. $T = 298K$. Uncertainties for ITC are estimated using the standard deviation of three replicates. $T = 298K$.

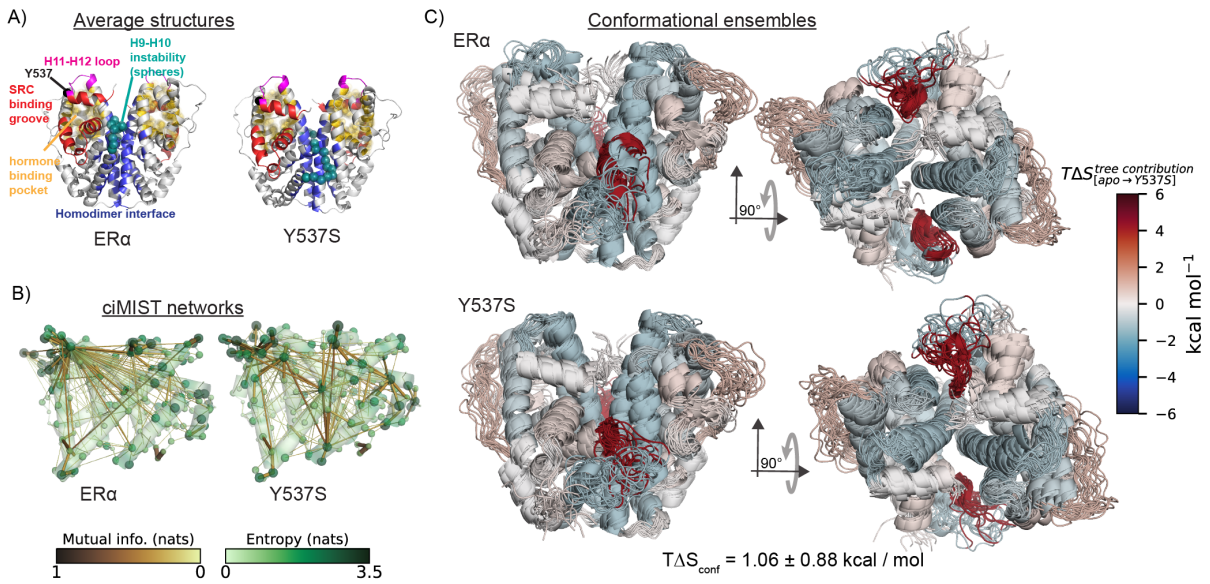

figure 4: **Redistribution of entropy in the unliganded (apo) state of the Y537S estrogen receptor α.** A) Average structure observed in simulations of human ERα-LBD homodimer and the Y537s variant. Structures were computed by superimposing all frames onto the first and then computing the average positions 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. $T = 298K$.

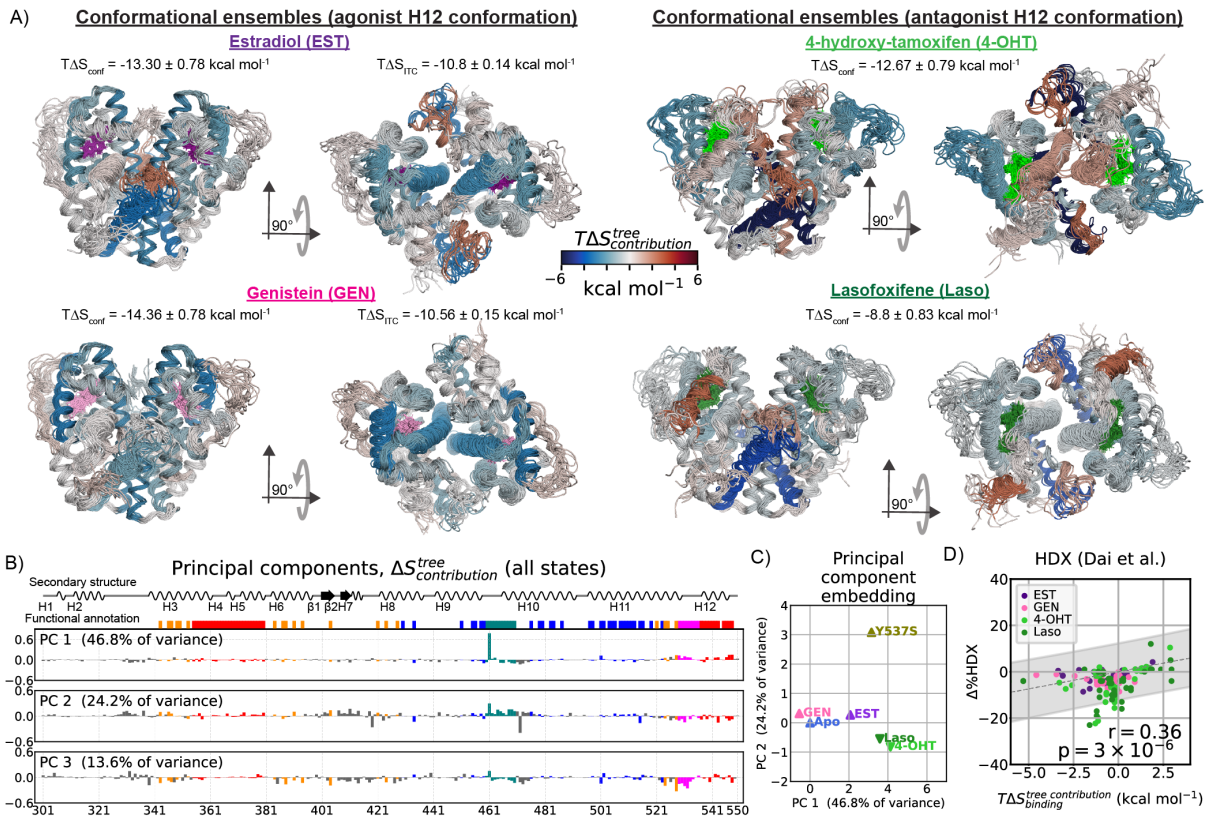

figure 5: **Entropic allostery in estrogen receptor α.** A) 20 randomly selected snapshots from simulations of ERα-LBD homodimers bound to estradiol (purple), 4-hydroxytamoxifen (lime green), genistein (pink), lasofoxifene (forest green). Secondary structural elements are colored by their contribution 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 (*64*), 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 (*3, 63*). The gray band demarcates 95% prediction intervals. $T = 298K$.

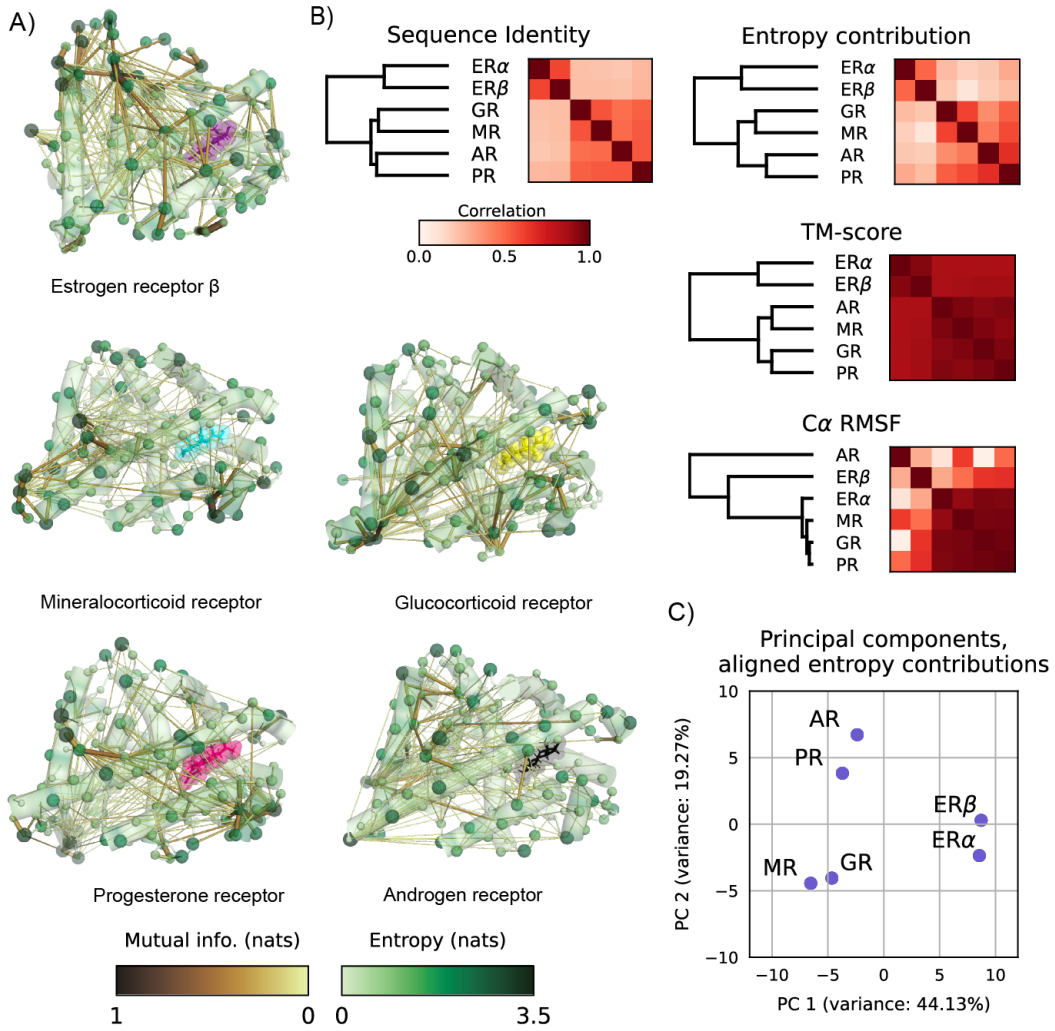

figure 6: **Evolutionary divergence of entropic contributions in the steroid receptor ligand-binding domains.** A) Maximum information spanning trees for human steroid receptor ligand-binding domains bound to endogenous ligands. Structures are oriented as ERα is in figure 3A. Sphere volumes of Cα atoms (shades of green) are proportional to residue marginal entropies, while volumes of links (shades of brown) are proportional to mutual information. 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. 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.

4. 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.

5. 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.

6. 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/.

7. D. Kern, E. Zuiderweg, The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology* **13**, 748–757 (2003), doi:10.1016/j.sbi.2003.10.008, https://www.sciencedirect.com/science/article/pii/S0959440X03001660.

8. K. Frederick, M. Marlow, K. Valentine, A. Wand, Conformational entropy in molecular recognition by proteins. *Nature* **448**, 325–329 (2007), number: 7151 Publisher: Nature Publishing Group, doi:10.1038/nature05959, https://www.nature.com/articles/nature05959.

9. S. Tzeng, C. Kalodimos, Protein activity regulation by conformational entropy. *Nature* **488**, 236–240 (2012), bandiera_abtest: a Cg_type: Nature Research Journals Number: 7410 Primary_atype: Research Publisher: Nature Publishing Group, doi:10.1038/nature11271, https://www.nature.com/articles/nature11271.

10. R. Nussinov, C. Tsai, Allostery in Disease and in Drug Discovery. *Cell* **153**, 293–305 (2013), publisher: Elsevier, doi:10.1016/j.cell.2013.03.034, https://www.cell.com/cell/abstract/S0092-8674(13)00391-7.

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

12. A. Kornev, Self-organization, entropy and allostery. *Biochemical Society Transactions* **46**, 587–597 (2018), doi:10.1042/BST20160144.

13. 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), publisher: American Chemical Society, doi:10.1021/bi2017156, https://doi.org/10.1021/bi2017156.

14. 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.

15. K. Connaghan, Q. Yang, M. Miura, A. Moody, D. Bain, Homologous steroid receptors assemble at identical promoter architectures with unique energetics of cooperativity. *Proteins: Structure, Function, and Bioinformatics* **82**, 2078–2087 (2014),

_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.24563, doi:10.1002/prot.24563, https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.24563.

16. K. Skriver, F. Theisen, B. Kragelund, Conformational entropy in molecular recognition of intrinsically disordered proteins. *Current Opinion in Structural Biology* **83**, 102697 (2023), doi:10.1016/j.sbi.2023.102697, https://www.sciencedirect.com/science/article/pii/S0959440X23001719.

17. H. Mikolajek, *et al.*, Correlation between the binding affinity and the conformational entropy of nanobody SARS-CoV-2 spike protein complexes. *Proceedings of the National Academy of Sciences* **119**, e2205412119 (2022), publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.2205412119, https://www.pnas.org/doi/full/10.1073/pnas.2205412119.

18. S. Wankowicz, J. Fraser, Advances in uncovering the mechanisms of macromolecular conformational entropy. *Nature Chemical Biology* **21**, 623–634 (2025), publisher: Nature Publishing Group, doi:10.1038/s41589-025-01879-3, https://www.nature.com/articles/s41589-025-01879-3.

19. C. Gautier, L. Laursen, P. Jemth, S. Gianni, Seeking allosteric networks in PDZ domains. *Protein Engineering, Design and Selection* **31**, 367–373 (2018), doi:10.1093/protein/gzy033, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6508479/.

20. 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.

21. S. Gianni, P. Jemth, Allostery Frustrates the Experimentalist. *Journal of Molecular Biology* **435**, 167934 (2023), doi:10.1016/j.jmb.2022.167934, https://www.sciencedirect.com/science/article/pii/S0022283622005617.

22. J. Monod, J. Wyman, J. Changeux, On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology* **12**, 88–118 (1965), doi:10.1016/S0022-2836(65)80285-6, https://www.sciencedirect.com/science/article/pii/S0022283665802856.

23. 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.

24. A. Van, J. Durrant, L. Votapka, R. Amaro, Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis. *Journal of Chemical Theory and Computation* **10**, 511–517 (2014), doi:10.1021/ct4008603, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3958135/.

25. Y. Karami, T. Bitard-Feildel, E. Laine, A. Carbone, "Infostery" analysis of short molecular dynamics simulations identifies highly sensitive residues and predicts deleterious mutations. *Scientific Reports* **8**, 16126 (2018), number: 1 Publisher: Nature Publishing Group, doi:10.1038/s41598-018-34508-2, https://www.nature.com/articles/s41598-018-34508-2.

26. 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].

27. M. Shapovalov, R. Dunbrack, A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure (London, England : 1993)* **19**, 844–858 (2011), doi:10.1016/j.str.2011.03.019, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3118414/.

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

29. C. Towse, M. Akke, V. Daggett, The Dynameomics Entropy Dictionary: A Large-Scale Assessment of Conformational Entropy across Protein Fold Space. *The Journal of Physical Chemistry B* **121**, 3933–3945 (2017), publisher: American Chemical Society, doi:10.1021/acs.jpcb.7b00577, https://doi.org/10.1021/acs.jpcb.7b00577.

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

31. S. Singh, G. Bowman, Quantifying Allosteric Communication via Both Concerted Structural Changes and Conformational Disorder with CARDS. *Journal of Chemical Theory and Com-*

putation **13**, 1509–1517 (2017), doi:10.1021/acs.jctc.6b01181, https://pubs.acs.org/doi/10.1021/acs.jctc.6b01181.

32. 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.

33. 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.

34. D. Marks, *et al.*, Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE* **6**, e28766 (2011), publisher: Public Library of Science, doi:10.1371/journal.pone.0028766, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766.

35. 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.

36. 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/.

37. K. Klinkner, C. Shalizi, M. Camperi, Measuring Shared Information and Coordinated Activity in Neuronal Networks, in *Advances in Neural Information Processing Systems* (2005), https://papers.nips.cc/paper_files/paper/2005/hash/12a1d073d5ed3fa12169c67c4e2ce415-Abstract.html.

38. 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.

39. 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.

40. 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.

41. 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).

42. L. Paninski, Estimation of Entropy and Mutual Information. *Neural Computation* **15**, 1191–1253 (2003), doi:10.1162/089976603321780272, https://direct.mit.edu/neco/article/15/6/1191-1253/6731.

43. S. Genheden, U. Ryde, Will molecular dynamics simulations of proteins ever reach equilibrium? *Physical Chemistry Chemical Physics* **14**, 8662–8677 (2012), publisher: The Royal Society of Chemistry, doi:10.1039/C2CP23961B, https://pubs.rsc.org/en/content/articlelanding/2012/cp/c2cp23961b.

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

45. 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.

46. 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).

47. 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.

48. 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.

49. 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.

50. 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.

51. 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.

52. 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.

53. 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.

54. 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.

55. 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.

56. 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.

57. 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.

58. 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.

59. 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.

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

61. 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.

62. 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.

63. 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.

64. 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.

65. 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.

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

67. 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.

68. 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.

69. 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.

70. 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.

71. 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.

72. 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.

73. P. Kunzmann, *et al.*, Biotite: new tools for a versatile Python bioinformatics library. *BMC Bioinformatics* **24**, 236 (2023), doi:10.1186/s12859-023-05345-6, https://doi.org/10.1186/s12859-023-05345-6.

74. 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.

75. 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.

76. T. Cover, J. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience) (2006).

77. B. Killian, J. Yundenfreund, M. Gilson, Extraction of configurational entropy from molecular simulations via an expansion approximation. *The Journal of Chemical Physics* **127**, 024107 (2007), publisher: American Institute of Physics, doi:10.1063/1.2746329, `https://aip.scitation.org/doi/10.1063/1.2746329`.

78. 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`.

79. 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`.

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

81. 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`.

82. 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.

83. 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`.

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

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

85. 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.

86. 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.

87. 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.

88. 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/.

89. 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.

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

91. C. Shannon, A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 (1948), conference Name: The Bell System Technical Journal, doi:10.1002/j.1538-7305.1948.tb01338.x, https://ieeexplore.ieee.org/document/6773024.

92. 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.

93. 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.

94. 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.

95. 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.

96. 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.

97. M. Lachmann, M. Newman, C. Moore, The physical limits of communication or Why any sufficiently advanced technology is indistinguishable from noise. *American Journal of Physics* **72**, 1290–1293 (2004), doi:10.1119/1.1773578, https://doi.org/10.1119/1.1773578.

98. 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.

99. 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.

100. 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.

101. 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`.

102. D. Arthur, S. Vassilvitskii, k-means++: The Advantages of Careful Seeding .

103. 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`.

104. 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).

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

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

107. 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`.

108. 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, Roland Dunbrack, 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 *(7-108)*

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 S7

Table S1

Materials and Methods

Molecular dynamics simulations

All simulations were performed using OpenMM7 using the Amberff14SB force field (*51*), TIP3P water model (*98*), Langevin integrator (*99*) 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 (*100, 101*). 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 nanosecond, 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, including simulation times, are provided in Supplemental Table S1.

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

We define a configuration of a residue as a vector, $\mathbf{\Theta} = [\theta_1 ... \theta_M]^T$, whose entries are the bond angles and torsion angles (expressed in units of radians) needed to specify the positions of atoms belonging to that residue. We used both bond and torsion angles in our analysis because we found that setting bond angles equal to their average values distorted global protein structures upon converting from internal coordinates (lengths, bond angles, torsion angles) to Cartesian. On the other hand, setting bond lengths equal to their average resulted in very little distortion; therefore, we did not use these to define residue configurations.

For each residue, our data are an ensemble $\{\mathbf{\Theta}_1 ... \mathbf{\Theta}_N\}$ of N residue configurations obtained by concatenating frames from all MD replicates for that thermodynamic state (combination of protein and ligand). We model the probability density as a function of residue configuration $\mathbf{\Theta}$ using a

mixture of von Mises distributions:

$$p(\mathbf{\Theta}) = \sum_{k=1}^{K} p(k) \prod_{m=1}^{M} \frac{\exp\left[\kappa_{km} \cos(\theta_m - \mu_{km})\right]}{2\pi I_0(\kappa_{km})}. \tag{S1}$$

Here, the sum is over the set of K mixture components, $p(k)$ is the posterior probability of the k-th component (with $\sum_{k=1}^{K} p(k) = 1$), the product is over the M bond and torsion angles used to place the residue's atoms, θ_m represents the m-th angle (from the aforementioned set), μ_{km} and κ_{km} denote the mode and precision parameters (respectively) of the k-th von Mises component along the m-th angular coordinate, and $I_0(\kappa_{km})$ is a modified Bessel function of order 0 guaranteeing independent normalization of each component along each dimension. Note that the value of M depends on the residue type.

We note that each of the K component distributions is a product of one-dimensional von Mises distributions. Although this means that fluctuations about the individual component modes are treated as independent, the use of multiple mixture components allows this model to capture irregularly-shaped probability densities.

We chose $K = 100$ (independently of residue type) with the aim of choosing a number of mixture components greatly exceeding the number of modes of each distribution, so that the mixture models operate in the regime of density estimation, rather than peak detection.

To initialize parameters, we seeded von Mises modes using the k-means^{++} algorithm (*102*) with a cosine distance between individual configurations. We initialize precision parameters using random samples from uniform distributions.

Fitting proceeded by iterating the expectation-maximization algorithm until the log-likelihood function converged to within 0.01 nats per observation, obtaining distinct mixture models of the configurational probability density for each residue in the protein.

To arrive at residue states adapted to the observed probability density, rather by a predetermined selection of the number of clusters, we applied DBSCAN to the mixture components. We used a distance function defined in terms of overlap between mixture components:

$$\mathcal{D}(k, l) = 1 - \frac{\sum_{n=1}^{N} \sqrt{p(k|\mathbf{\Theta}_n) p(l|\mathbf{\Theta}_n)}}{\sqrt{p(k)p(l)}}. \tag{S2}$$

Here, $p(k|\mathbf{\Theta}_n)$ is the posterior probability that observation $\mathbf{\Theta}_n$ was generated by the k-th mixture

component (sometimes called the ”responsibility” of component k for observation n), assuming the mixture model is the generative model for the data. These posterior probabilities are calculated from the component likelihoods using Bayes' theorem. The sum runs over all N frames in the MD ensemble. This distance function, which is related to the Bhattacharya divergence and the Hellinger distance, takes on a value of 0 when k and l index the same mixture component and approaches a limiting value of 1 as two distributions become more sharply concentrated on distinct points. The quantity $\sum_n \sqrt{p(k|\mathbf{\Theta}_n)p(l|\mathbf{\Theta}_n)}$, which measures the overlap between distributions, can be informally interpreted as estimating the number of observations of ”mixed states” in the MD ensemble – snapshots which could have been generated by sampling from either component. Thus, clustering components using this distance metric reduces the number of observations which cannot be unambiguously assigned to a single state.

For each residue, we set the distance parameter for DBSCAN to one minus the standard deviation of all pairwise overlaps, $\sqrt{p(k|\mathbf{\Theta}_n)p(l|\mathbf{\Theta}_n)}$. We found that this resulted in sensible partitions of the configurational spaces by inspecting lower-dimensional marginal distributions; in particular, distinct Ramachandran regions could be identified for loop glycines – a feat which was not straightforward to achieve using other clustering methods.

We set the minimum mass parameter in DBSCAN to 0.01. In the present application, this corresponds to a minimum probability mass required to classify a region of continuously high probability density as a conformation. The number of observations for all of our trajectories was greater than $10,000$, so this requirement 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 conformations (x_u, x_v) under the hypothesis of independence: $p(x_u, x_v) = p(x_u)p(x_v)$. Hence, we can be more confident that pairs of conformations which we do not observe in simultaneously are somewhat energetically inconsistent with one another (rather than unobserved by mere chance) than we would be with no coarse-graining; controlling chance fluctuations in low-probability states tames a major source of error in entropy estimation (*42*).

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 conformation with the maximum likelihood of generating it, which is the sum of the likelihoods for all mixture components mapped to that conformation.

Maximum likelihood tree inference

If we have a distribution, $p(X_1 \ldots X_L)$, 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 $X_1 \ldots X_L$ 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,

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

Posterior 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 (*103*):

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

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 (*103*),

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

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

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

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

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

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 S6).

Bayesian model for allosteric mutation predictions.

We extracted the number of allosteric mutations at each site from the supplemental data provided in (*20*). We adopted the definition of allosteric mutation used by the authors: a mutation whose effect on CRIPT-binding was significantly greater than the effect of a random binding site mutation on CRIPT-binding. We modeled the number of allosteric mutations 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 total number of point mutations reported at site u in the sequencing data (*20*), $N_u^{\text{allosteric}}$ is the number of allosteric mutations reported at this site, and p_u is a latent variable which is our estimated probability that a random mutation at site u will be allosteric.

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 the reason that alleles are absent from the sequencing data; alleles may have been unobserved they were extremely deleterious to protein function (and if they were, whether this due to binding or abundance), because the mutations simply happened not to arise in the population, or due to sequencing errors. To account for this, we modeled p_u, the site-specific probability of allosteric mutation as a random draw from the beta distribution which determined by the predictors

$$\mathbf{x}_u = \left[S_u, \sum_{v \in N_u(T)} I_{uv}, 1\right]^T,$$

where S_u is the marginal entropy of residue u, $\sum_{v \in N_u(T)} I_{uv}$ is the sum of all mutual informations between residue u and its neighbors $N_u(T)$ in the tree, and the final term represents an intercept. These features are related to our estimate of the number of allosteric mutations according to

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

where $\boldsymbol{\beta}$ is the vector of model coefficients. 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. The use of the Beta distribution allows us to integrate over uncertainty in p_u that arises due to unobserved alleles.

We used a multivariate normal distribution as a prior for our coefficient vector $\boldsymbol{\beta}$ with a diagonal covariance matrix and a weakly-informative Half-Cauchy prior with scale parameter equal to 25 for the variance. The model was estimated using Markov Chain Monte Carlo (MCMC) with 5000 draws using the No U-turn sampler (*104*).

Data Visualization

Plots were made in Python using Matplotlib (*105*). Protein visualization was done in PyMOL (*106*). Colormaps displayed on proteins are drawn from the cmocean libary (*107*).

Supplementary Text

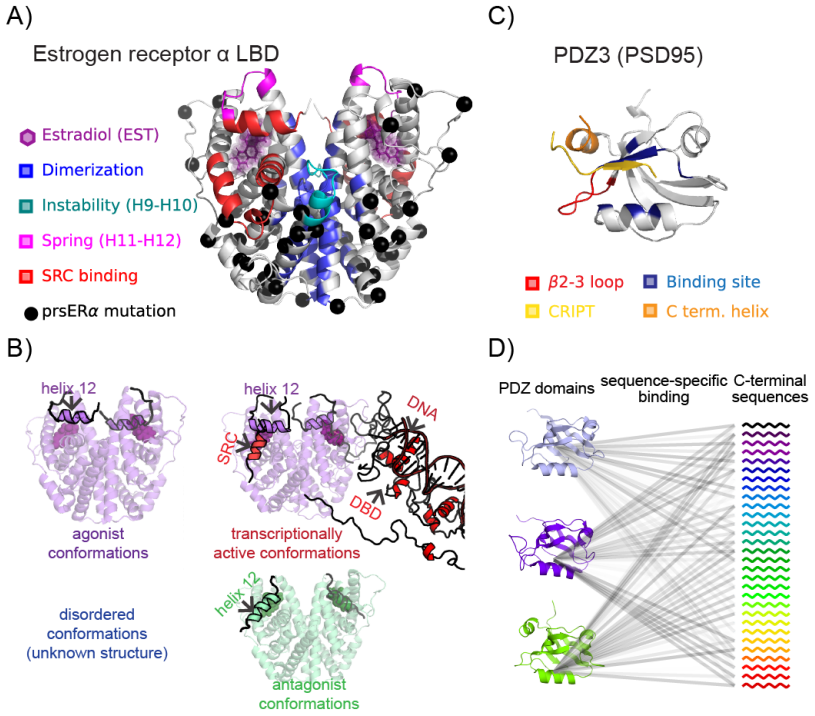

figure S1: **Allostery in estrogen receptor α and PDZ3** A) Estrogen receptor α ligand-binding domain (LBD, PDB: 1QKU) homodimer bound to estradiol. Sites colored in blue, teal, pink, and red influence activities from outside the hormone binding site. Sites shown in black spheres are those at which stabilizing 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).

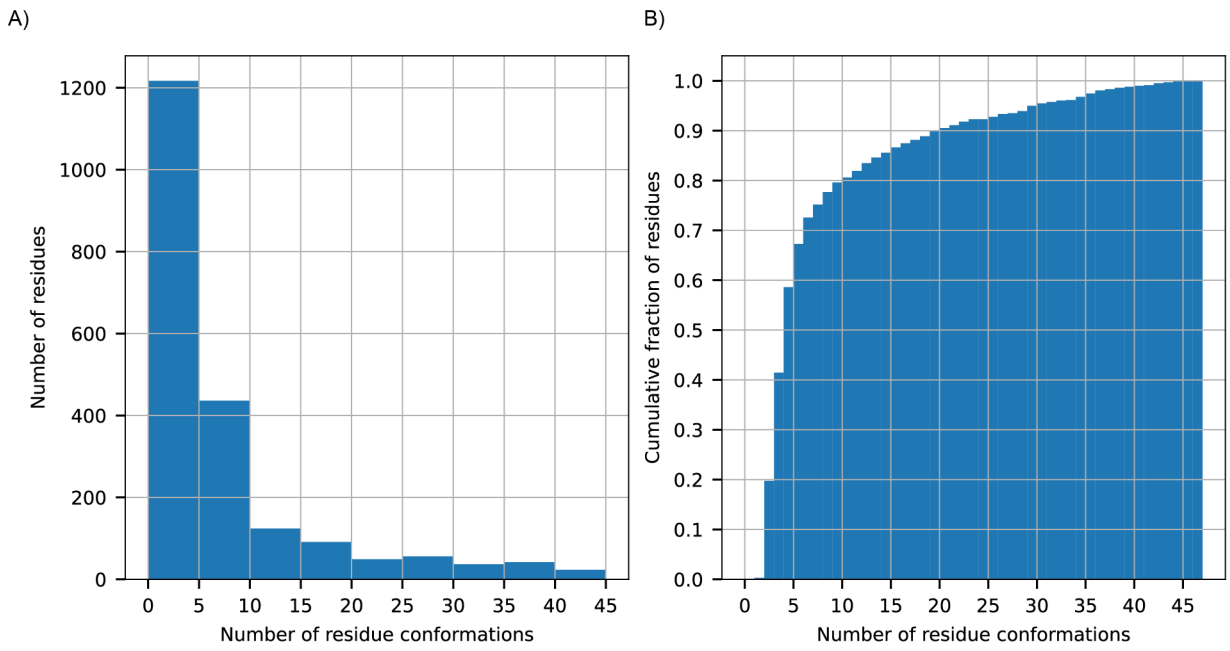

figure S2: **Frequencies of numbers of residue conformations across all ensembles analyzed in this work.** A) Histogram showing the absolute number of times our conformational inference algorithm inferred different numbers of residue conformations across all ensembles analyzed in this work. Note we use a bin width of five for the histogram. B) Cumulative fraction of times our conformational inference algorithm inferred a number of residue conformations less than or equal to a given value. Note the bin width (x-axis) is one.

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) (*108*). 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.

State	PDB	Residues	Total time (μs)	Replicates
PDZ3	1BFE	302-402	10.5	6
PDZ3Δ7ct	1BFE	302-395	10.5	6
PDZ3Δ7ct-CRIPT	5HEB	302-395 (A), 1-9 (B)	10.5	6
PDZ3-CRIPT	5HEB	302-402 (A), 1-9 (B)	10.5	6
CRIPT	5HEB	1-9 (B)	10.5	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

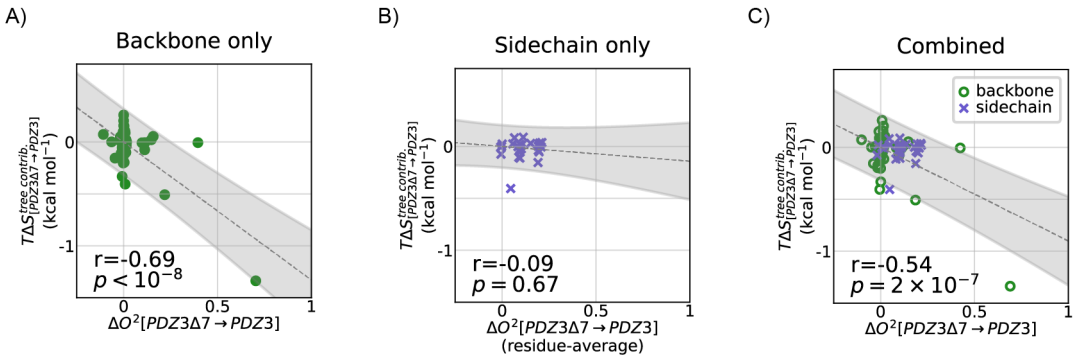

figure S3: **Correlation of residue entropy contributions with NMR** A) Correlation between differences in residue entropy contributions and changes in backbone O^2 order parameters (*50*) between PDZ3 and PDZ3Δ7-ct. B) Correlation between differences and residue entropy contributions and changes in side-chain O^2 order parameters (*50*) between PDZ3 and PDZ3Δ7-ct. C) Correlation between differences and residue entropy contributions and the combined set of O^2 order parameters (*50*) between PDZ3 and PDZ3Δ7-ct. Side-chain order parameters are residue-averaged.

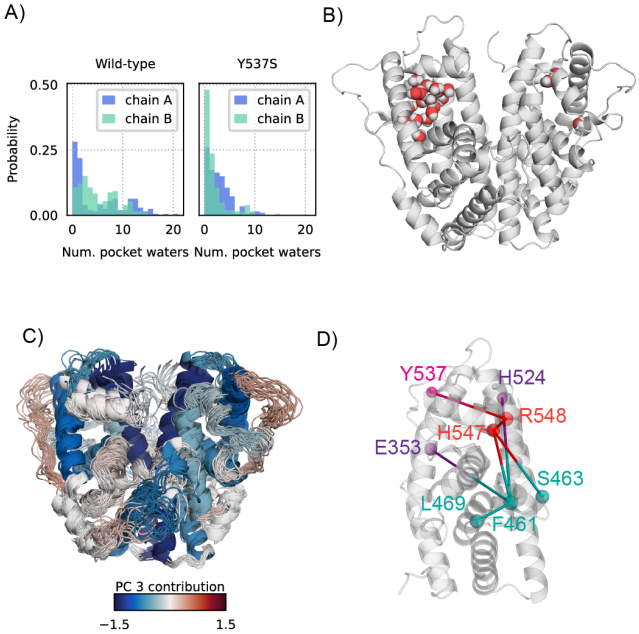

figure S4: A) Histograms of the number of waters in the binding pocket as defined by the convex hull of its atoms in each frame. Differences between the wild-type and Y537S pocket hydrations are significant with $p < 10^{-6}$. B) A snapshot of the wild-type apo state showing a lifted H12 conformation 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.

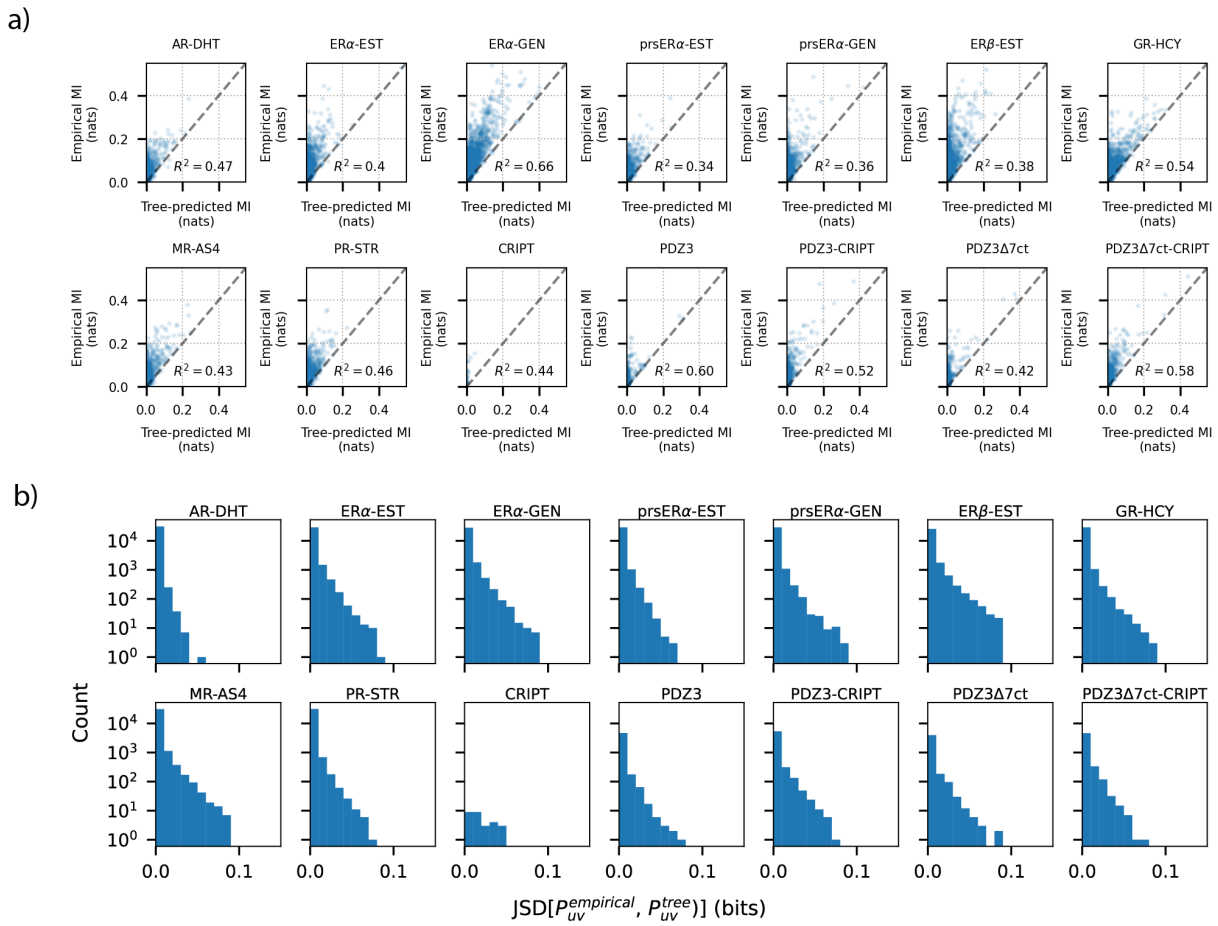

figure S5: **Consistency of tree-predicted pairwise conformational distributions with empirical conformational distributions for pairs of residues not retained in the tree.** A) Correlation between tree-predicted and empirical (maximum likelihood) mutual information for edges not included in the maximum 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.

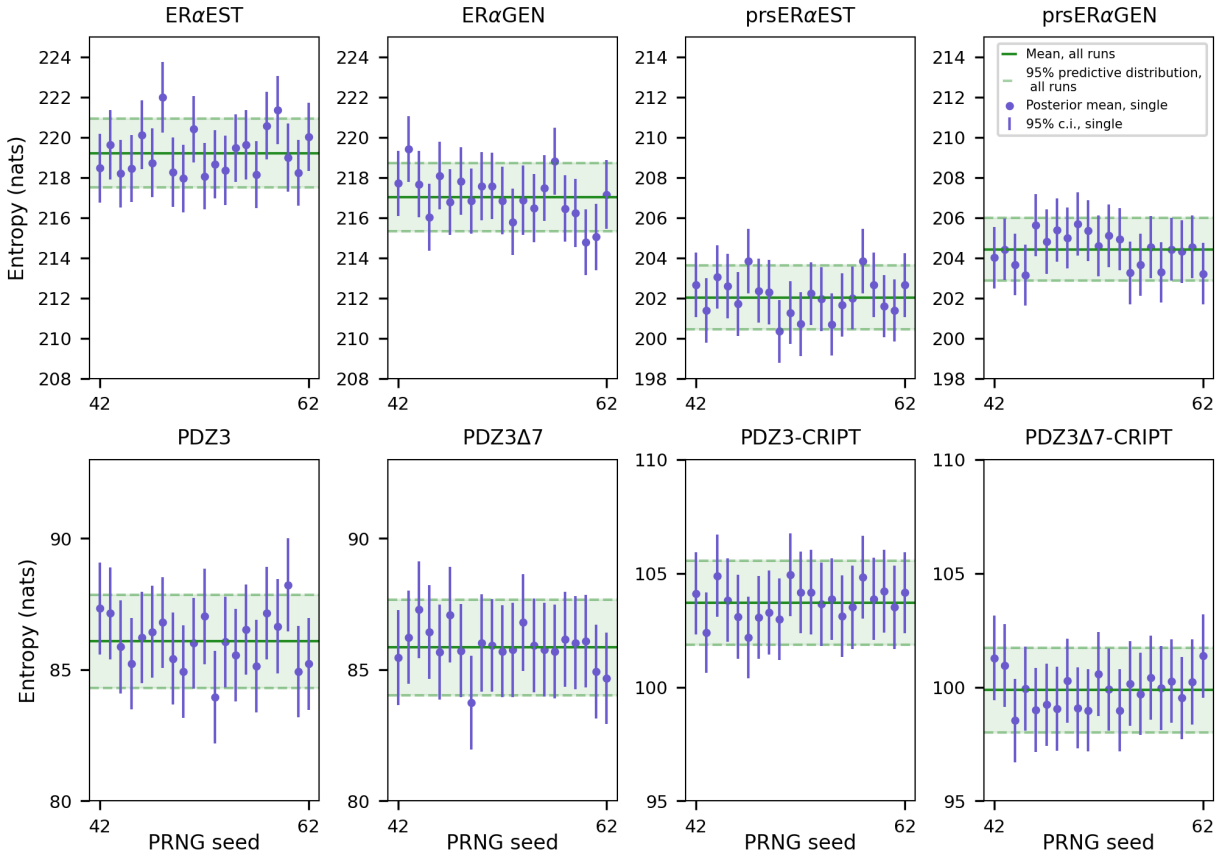

figure S6: Uncertainty quantification for tree entropies. Shown are the average entropies obtained from running the program with 21 different pseudo-random number generator (PRNG) seeds, the 95% posterior predictive distribution estimated from these runs and the posterior means and credibility intervals from the each run.

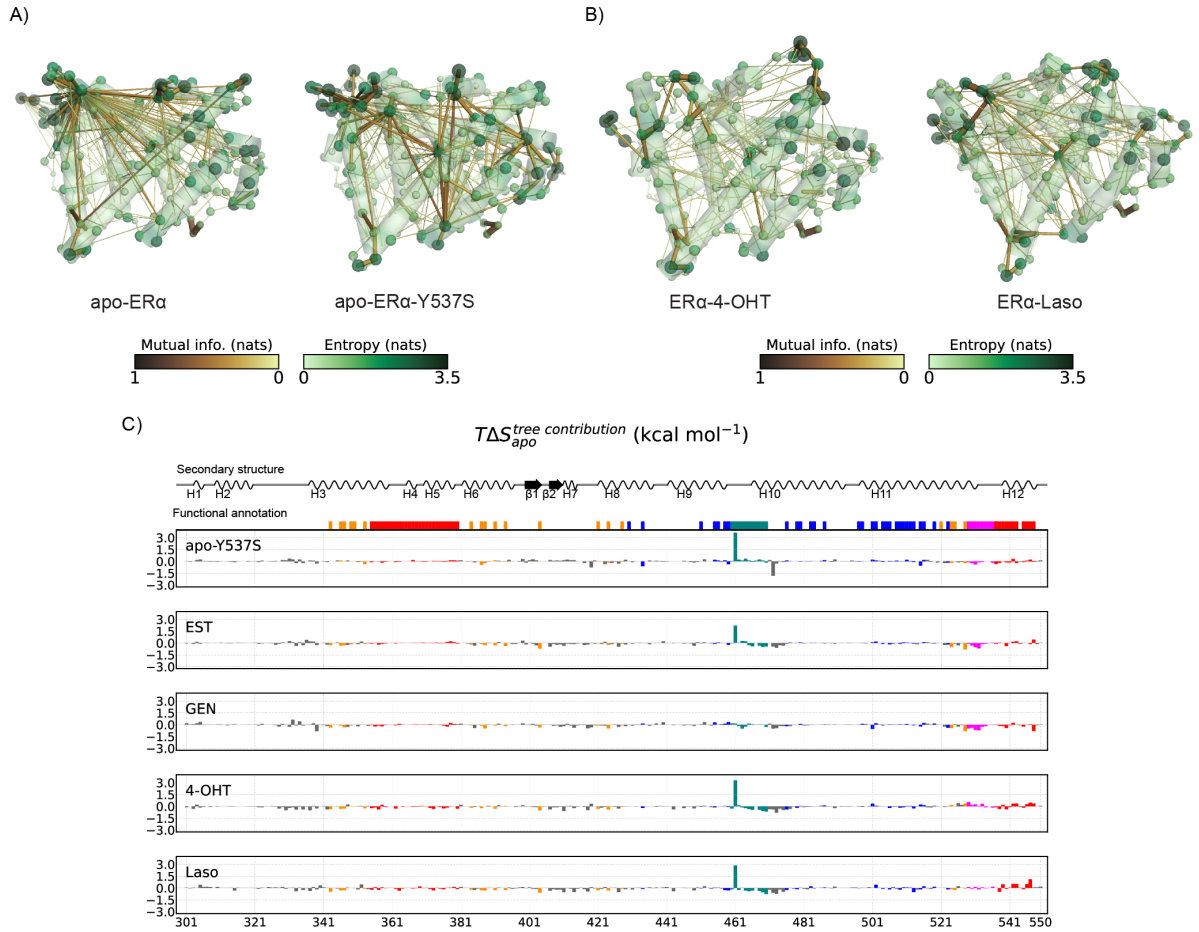

figure S7: a) Maximum information spanning trees for apo states of wild-type ERα and the Y537S variant. b) Maximum information spanning trees for ERα bount to anti-estrogen ligands 4-hydroxy-tamoxifen (4-OHT) and lasofoxifene (Laso). c) Differences in residue entropy contributions from the wild-type apo state of ER alpha for the Y537S mutant and all liganded states.