# Conserved energetic changes drive function in an ancient protein fold

Malcolm L. Wells[1]*, Chenlin Lu[1]*, Daniel Sultanov[1]*, Kyle C. Weber[1]*, Zhen Gong[1], and Anum Glasgow[1†]
[1] Department of Biochemistry and Molecular Biophysics, Columbia University Irving Medical Center, New York, NY, United States
* These authors made equal contributions.
[†] Corresponding author. Email: ag4522@cumc.columbia.edu

## Abstract

Many protein sequences occupy similar three-dimensional structures known as protein folds. In nature, protein folds are well-conserved over the course of evolution, such that there are 100,000 times as many extant protein sequences than there are folds. Despite their common shapes, similar protein folds can adopt wide-ranging functions, raising the question: are protein folds so strongly conserved for the purpose of maintaining function-driving energetic changes in protein families? Here we show that a set of key energetic relationships in a family of bacterial transcription factors (TFs) is conserved using high-resolution hydrogen exchange/mass spectrometry, bioinformatics, X-ray crystallography, and molecular dynamics simulations. We compared the TFs to their anciently diverged structural homologs, the periplasmic binding proteins (PBPs), expecting that protein families that share the same fold and bind the same sugars would have similar energetic responses. Surprisingly, our findings reveal the opposite: the “energetic blueprints” of the PBPs and the TFs are largely distinct, with the allosteric network of the TFs evolving specifically to support the functional requirements of genome regulation, versus conserved interactions with membrane transport machinery in PBPs. These results demonstrate how the same fold can be adapted for different sense/response functions via family-specific energetic requirements – even when responding to the same chemical trigger. Understanding the evolutionarily conserved energetic blueprint for a protein family provides a roadmap for designing functional proteins *de novo*, and will help us treat aberrant protein behavior in conserved domains in disease-related drug targets, where engineering selectivity is challenging.

Over 700 million natural proteins (*1*) fold into only a few thousand spatial arrangements of secondary structures, or protein folds (*2*–*6*). Extant protein folds have been maintained over billions of years despite low sequence identity (*7*). To evolve a new function, a protein typically samples sequence changes that preserve its fold while varying its conformational dynamics (*8*). Sequence changes to a protein that disrupt essential allosteric regulation are disfavored (*9*). Thus, proteins are robust to mutations that do not disrupt essential dynamics or function, enabling sequence diversity while preserving the basic topology.

Unfortunately, methodological limitations in observing energetic relationships in proteins limit our capacity to identify the functionally important elements of the fold for most proteins. For example, protein functions may depend on perturbation-driven changes in the local stability of specific residues, secondary structures, intramolecular interactions, and quaternary interfaces. Consequently, we do not generally understand how new functions emerge from old folds.

Using a high-resolution integrative structural biology strategy, we determined the evolutionarily conserved, functionally important energetic relationships in two bacterial protein families that share the Venus Flytrap (VFT) fold. VFT-fold proteins are found across the domains of life (*7*, *10*, *11*). Five proteins from *E. coli* serve as our model system for probing energetic redistribution in the families: three dimeric LacI/GalR transcription factors (TFs) and two monomeric periplasmic binding proteins (PBPs). The TFs regulate sugar metabolism genes in the cytoplasm, while the PBPs are periplasmic transporters for the same sugars (*12*, *13*). The PBPs and TFs most likely evolved ligand binding specificity independently after a DNA binding domain (DBD) acquisition event that occurred before the divergence of eubacteria (*7*, *14*).

Despite their shared topologies, in experimentally solved structures, the sugar-induced conformational changes in the VFT domains of TFs are subtler than in PBPs due to the dimer state of the TFs (*12*, *13*, *15*). The family's namesake TF, the *lac* repressor (LacI), reweights specific intramolecular contacts to switch between its transcriptionally repressed and active functional states without dramatic changes to the VFT domain structure (*16*–*18*). The DBD of each TF has a helix-turn-helix (HTH) motif that is connected by a hinge helix to the VFT domain. TFs switch between their repressed and active states by allosterically restructuring their DBDs in response to binding inducer molecules in their VFT domains. We hypothesized that the VFT-HTH architecture might be so strongly conserved in the TFs because they require the DBD and VFT domain to be coupled via a minimal set of shared energetic relationships. Because the PBPs bind the same sugars, we hypothesized that they evolved similar sugar-driven energetic changes within the constraints of the VFT fold.

**LacI/GalR TFs and PBPs have distinct sequence conservation**

To investigate whether evolution of separate functions in PBPs versus LacI/GalR TFs precipitated amino acid changes in the VFT fold that are specific to either family, we built a gene tree based on 2,222 diverse LacI/GalR TF and PBP sequences using only their VFT domains (Figs. 1A, S1). Although the large number of sequences and the small number of inspected amino acid sites limit the resolution of the more internal branches, all TFs form a separate clade from all PBPs with high confidence (bootstrap support 100%), corroborating a previous study (*7*) and suggesting independent evolution of the VFT domain in the two lineages.

Comparing conservation scores (see **Methods**) for each protein family revealed distinct patterns of sequence conservation (Fig. 1B). While the PBPs are uniformly highly conserved,

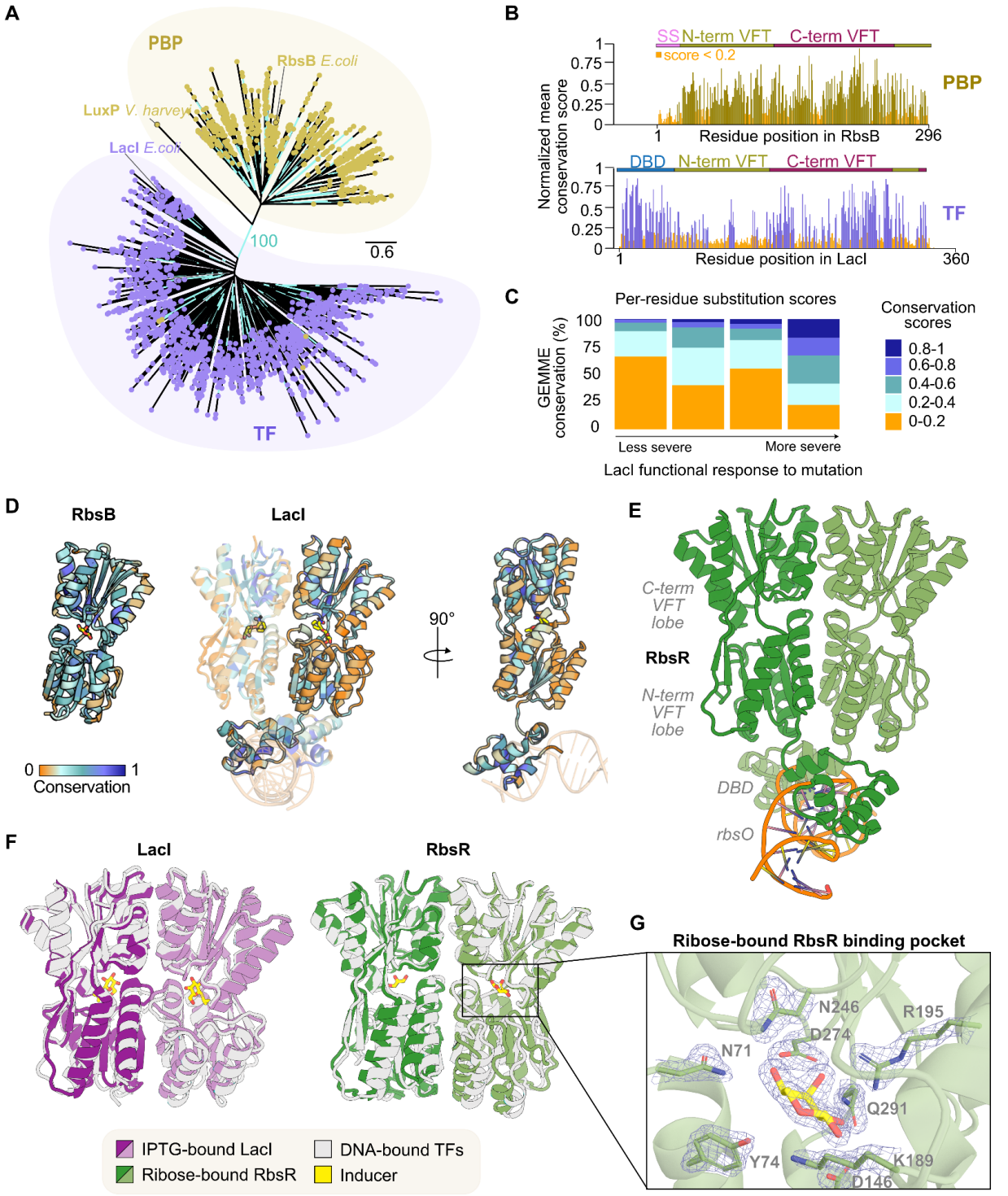


**Figure 1. Sequence conservation and structure of LacI/GalR TFs and PBPs. A**, Reconstructed gene tree shows distinct clades of LacI/GalR TFs versus PBPs. Bootstrap values >90% are in turquoise. **B**, Sequence conservation is higher and more uniform for PBPs than for TFs. **C**, TF sequence conservation echoes mutational data for LacI: more conserved positions are less tolerant to mutations, causing severe changes to LacI function. **D**, The TF dimer interface is among the least conserved regions of the fold (*e.g.*, LacI, *right)*. The homologous region is well-conserved for PBPs (*e.g.*, RbsB, *left*). **E**, We solved 2.1 Å X-ray crystal structures of the RbsR complex with operator DNA *rbsO* and **F**, with ribose (*right*). The inducer-bound states of LacI and RbsR VFT domains are structurally similar by X-ray crystallography. **G**, Zoomed view of key RbsR interactions with ribose with 2Fo–Fc maps for sidechains and ligand at 1σ.

the TFs exhibit "valleys" of higher sequence diversity. As expected, there is high conservation in the DBD. However, the N-terminal lobe of the VFT domain is more permissive to mutations than the C-terminal lobe (Fig. 1B). The conservation pattern recapitulates experimental phenotypic outcomes in LacI mutants (Fig. 1C) (*19*).

Interestingly, we found substantial sequence variability in the TF dimer interface, reflecting the biochemical diversity of interfaces across many TFs (Figs. 1D, S2) (*15*, *20–22*). Although several solvent-exposed helices in the TFs are also sequence-diverse, the variability in the dimer interface is surprising because the LacI/GalR TFs are obligate homodimers, with no known monomeric functional state. By contrast, the same region is highly conserved in PBPs (Fig. 1D), likely due to interactions with nutrient transporters (*23*). Differences in sequence conservation in the two families reflect their distinct functional constraints while sharing a fold.

**LacI/GalR TFs are structurally similar in different functional states**

Bioinformatics-derived sequence-level differences do not reveal the structural basis for ligand regulation in the TFs, which requires comparing high-resolution X-ray crystal structures of related TFs in different functional states (*24*, *25*). Of the 41 experimentally solved structures of full-length LacI/GalR TFs, 31 are solved to <3 Å resolution (Table S1). Among all 100 published TF structures that include at least the VFT domain, 83 are solved to <3 Å resolution. Overall, all TF VFT domain structures have <4 Å pairwise root-mean-square deviation (RMSD) across their protomers (Fig. S3). We are interested in TFs that activate transcription by binding small molecules, and of this subset, only the LacI VFT domain structure is available in both operator- and inducer-bound states (*26–28*, *15*). In summary, an evolutionarily conserved activation mechanism in the LacI/GalR family cannot be gleaned from the available structural information.

To properly compare LacI with another sugar-inducible LacI/GalR TF, we solved X-ray crystal structures of the ribose repressor (RbsR) in its inducer- and operator-bound forms at 2.1 Å resolution (Figs. 1E-G, S4, Table S2). RbsR is induced by ribose and has 33% sequence identity with LacI (*29*). While LacI controls lactose metabolism by binding the *lac* operator DNA (*lacO1*), RbsR controls ribose metabolism by binding the *rbs* operator DNA (*rbsO*).

Like the LacI complexes with inducer isopropyl β-D-1-thiogalactopyranoside (IPTG) and *lacO1*, RbsR-ribose and RbsR-*rbsO* have 1.58 Å backbone RMSD in the VFT domain (Fig. 1F). The C-terminal VFT lobe is almost identical between the operator- and inducer-bound states of both TFs (0.40 Å and 0.61 Å RMSD for RbsR and LacI, respectively). The two TFs are also highly structurally similar, with 1.6 Å RMSD in the VFT domain (Fig. 1E, F). The DNA-bound states for both TFs show the hinge helices buried in the minor groove of the operator, which is bent at 45° (Figs. 1E, S4-1). Notably, the extra base insertion between the palindromic repeats in *lacO1* compared to *rbsO* does not change the crystallographic DNA or protein conformation.

**Measuring the local stabilities of VFT protein functional states**

For transcriptional regulation to occur, the TFs must sample different conformations with different probabilities in each state, leading to *conformational ensembles* that are energetically distinct and not clear from crystal structures (*30*). To determine how the energy landscape is conserved in TFs (the "energetic blueprint"), we performed hydrogen exchange/mass spectrometry (HX/MS) on three paralogous sugar-inducible LacI/GalR TFs in their apo, DNA-

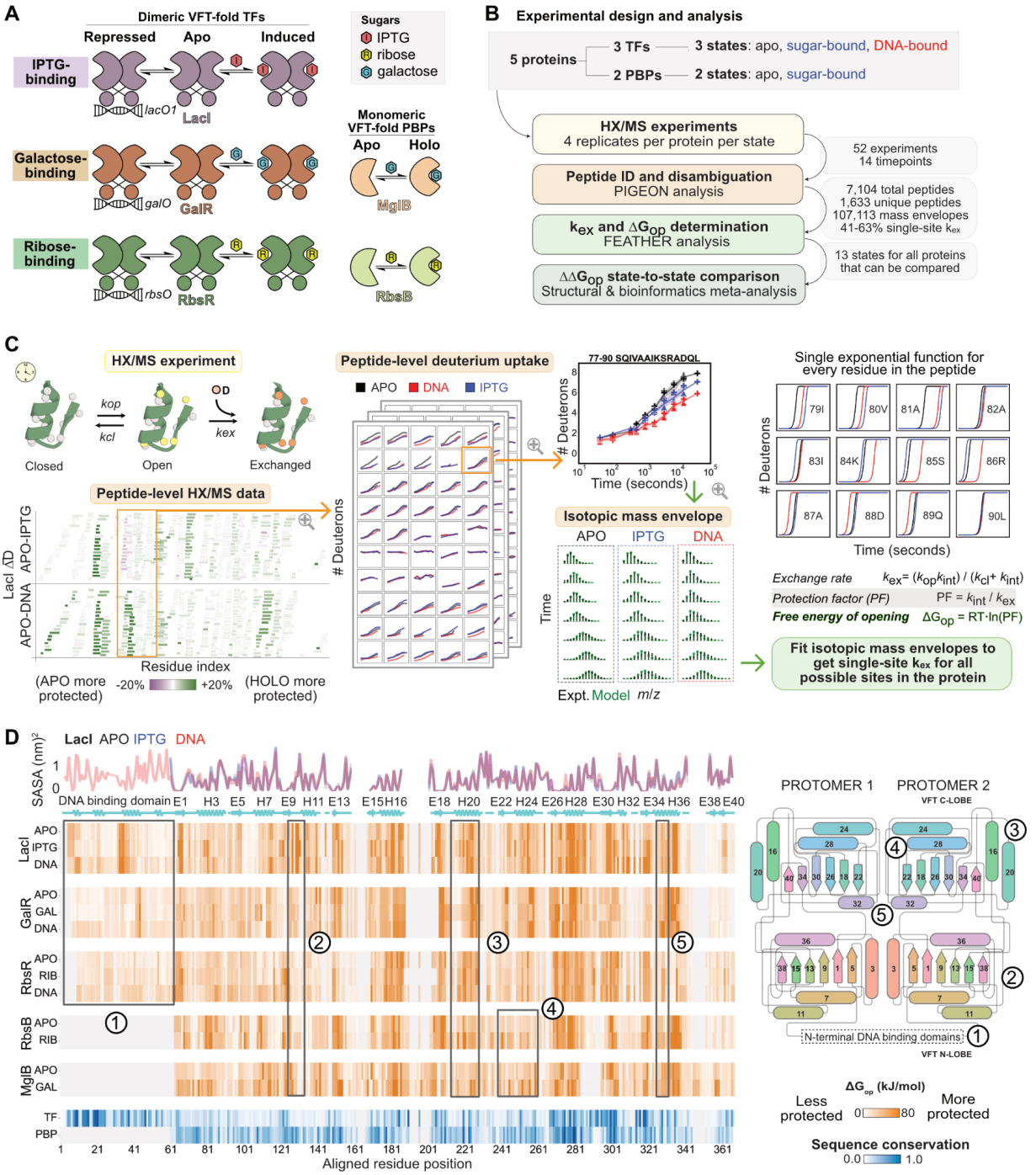


**Figure 2. High-resolution HX/MS. A**, We chose two PBPs and three TFs to study by HX/MS, including two PBP/TF pairs that respond to the same ligands. **B**, Experimental design and data summary. **C**, Overview of HX/MS experiment and analysis using PIGEON-FEATHER, resulting in site-resolved $\Delta G_{op}$ for many residues. **D**, $\Delta G_{op}$ heatmap for the proteins ordered by structurally aligned residues, compared to LacI solvent-accessible surface area (SASA) and TF and PBP sequence conservation. Gray marks an alignment gap. *Right:* VFT-fold structure schematic. Several regions are marked by matching numbers on the heatmap and schematic. (1) The DBD has higher $\Delta G_{op}$ in the operator-bound state than in apo or sugar-bound states in the TFs. (2) A binding pocket loop responds differently to sugar binding in each protein. (3) Sugar binding stabilizes helix 20 in each protein. (4) Sugar binding stabilizes a C-terminal VFT lobe segment in the PBPs. (5) Sugar binding stabilizes the N-terminal region of helix 36 for all proteins.

bound, and inducer-bound states: LacI, RbsR, and the galactose repressor (GalR) (*31*) (Figs. 2A, S5-S7, Tables S3-S5). GalR induces transcription of the *gal* operon by binding galactose and has 25% sequence identity to LacI and 31% to RbsR (*32*). To account for differences in hydrogen exchange among the TFs that might arise from binding different sugars, we also performed HX/MS on two PBPs that bind the same sugars as RbsR and GalR: ribose binding protein (RbsB) and galactose binding protein (MglB) (Figs. 2A, S8, Tables S6, S7). These five VFT proteins bind the sugars with affinities in the 100 nM to 1 mM range. Both TF/PBP pairs have 28% sequence identity in the VFT domain.

HX/MS measures the rate of exchange of backbone amide hydrogens with deuterium, revealing the relative structuredness of regions of the protein in various functional states. Using PIGEON-FEATHER (*17*), a high-resolution HX/MS analysis method, we determined the single-residue free energies of opening ($\Delta G_{op}$) for all three TFs in three states (apo, sugar-bound, and operator-bound) and both PBPs in two states (apo and sugar-bound) (Fig. 2B, C, Tables S8-S12). The $\Delta G_{op}$ is a measure of local stability; the highest $\Delta G_{op}$ values in a protein correspond with its global free energy of unfolding (*33*). Ligand-driven shifts in $\Delta G_{op}$ report local energetic changes in the context of the protein ensemble.

The $\Delta G_{op}$ datasets show local and global protein responses to binding sugars and operators (Figs. 2D, S9, Supplemental Note 1), demonstrating that some effects occur only in PBPs or TFs (Fig. 2D-1, 2D-4), while others are conserved in all VFT-fold proteins (Fig. 2D-3, 2D-5). Some regions of the VFT fold are differentially stabilized by ligand binding in each protein. For example, a structurally conserved binding pocket loop (Fig. 2D-2) responds differently to each sugar in each protein, suggesting protein-specific evolution of a functional role for the loop to allow *E. coli* to tune ligand binding affinity and specificity. We discuss state-, family-, and fold-specific energetic changes in the next sections.

**Binding to operator DNA has long-range allosteric effects in all TFs**

To learn how the TFs respond to binding operator DNA, we compared their $\Delta G_{op}$ in operator-bound and apo states ($\Delta\Delta G_{op,\ DNA\text{-}APO}$) (Fig. 3A, B). As expected, the DBD of each TF has higher average $\Delta\Delta G_{op,\ DNA\text{-}APO}$ than the VFT domain due to interactions with DNA that structure the DBD (Fig. 3C) (*34*). Operator binding also stabilizes the hinge helix (*35*) and several VFT-DBD interactions (Fig. 3C, D).

Remarkably, binding to operator DNA also allosterically stabilizes the TFs far beyond the DBD. One example is the structuring of a C-terminal dimer interface helix (the “metastable helix”, H32 in Fig. 2D), which points to increased dimer stability in LacI/GalR TFs upon operator binding regardless of the DNA sequence, length, or symmetry (Fig. 3B-1, D-1). This region is identical in crystal structures of LacI and RbsR in all functional states (Figs. 1F, S4-3) and in structural models of GalR (*36*). The behavior of H32 exemplifies a functionally important, evolutionarily conserved shift in a protein’s ensemble that occurs without a crystallographic conformational change. While the regions flanking the metastable helix are conserved in the LacI/GalR MSA (Fig. 1D, Fig. S2), the helix itself is not, suggesting that evolution might tune this region to control DNA binding affinity and inducer sensitivity. In LacI, mutations to H32 affect the quaternary structure (*37*). By contrast, H32 is well-conserved in PBPs (Fig. 1D), likely because it forms part of the binding interface for membrane transporters.

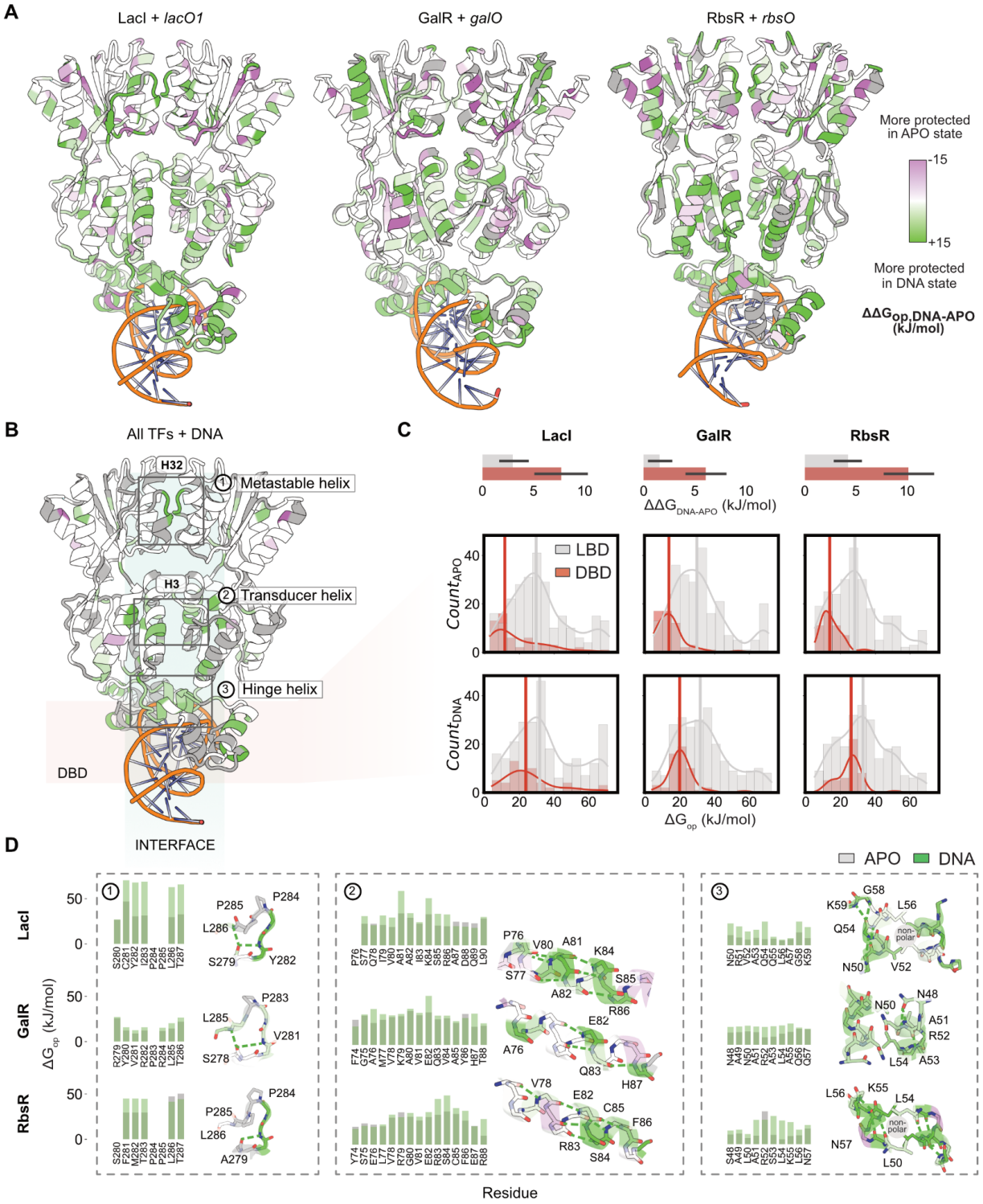


**Figure 3. Conserved allosteric effects of DNA binding in TFs. A**, Each TF responds to binding its own operator DNA with global changes in local stability. **B**, Conserved energetic effects of operator binding in the three TFs. **C**, Although there is global restructuring in the TFs, operator binding causes a larger increase in average $\Delta G_{op}$ in the DBD than the VFT domain. **D**, conserved increase in $\Delta G_{op}$ for each TF in key operator-responsive regions.

We also observed other conserved effects of operator binding at the dimer interface, including stabilization in the "transducer helix" H3 (Figs. 3D-2, S4-2), which bridges the binding pocket with the DBD and experiences ligand-dependent shifts in $\Delta G_{op}$ that are discussed further in the next section. Operator binding also stabilizes the central residue of the N-terminal VFT beta strand E5 in each TF, together with its hydrogen bond partner in the neighboring strand and in the DBD (Fig. 3B, D-3).

X-ray crystal structures of LacI-*lacO1* and RbsR-*rbsO* showed that TF binding bends operator DNA (Fig. 1E, S4-1). NMR spectroscopy on the LacI DBD corroborated this result and further demonstrated that DBD binding to off-target DNA maintains a linear DNA conformation (*38*). However, the structural explanation for DNA operator binding specificity by LacI remains mysterious because the behavior of the VFT domain in the presence of off-target DNA could not be studied using either technique. To explore LacI response to on-target vs. off-target DNA in its VFT domain, we compared peptide-level HX/MS of LacI in solution with *lacO1* (on-target, symmetric), *rbsO* (off-target, symmetric), and *scrO*, a scrambled version of the *lacO1* sequence (off-target, not symmetric) (Fig. S10, Table S13). While *rbsO* produced attenuated effects in the DBD compared to *lacO1*, only *lacO1* could cause the allosteric changes in the metastable helix, 70 Å from the DBD. These structural rearrangements, which are not observed in LacI incubated with weaker (*rbsO*) or nonspecific (*scrO*) DNA binding motifs, promote stronger binding to operator DNA as a demonstration of DNA sequence-induced protein allostery.

**Two VFT-fold protein families respond differently to binding the same sugars**

We reasoned that comparing $\Delta G_{op}$ for structurally aligned regions would highlight similarities and differences in how binding sugars redistributes the energy in TFs vs. PBPs to mediate distinct sense-response behaviors (Fig. 4) (*17*). Similarly, comparing $\Delta G_{op}$ for TFs and PBPs that bind the same ligands can reveal how the same fold can be adapted for different functional responses to the same ligand (Figs. 4, S11A-C). While sugar binding causes unique local effects in each protein due to each sugar's specific biochemical characteristics (Fig. 4C, D), we also observed long-range pan-VFT and family-specific patterns (Figs. 4B, E, S10D).

The hinge loops of PBPs stabilize their sugar-bound conformational state (Fig. S12) (*13*), whereas the TFs show no conformational changes in these loops in apo vs. sugar-bound structures. Accordingly, although the PBPs and TFs bind the same sugars, they share few conserved effects by HX/MS (Figs. 4, S11C). All hinge loops are stabilized in both PBPs; some positions have $\Delta\Delta G_{op,\ HOLO\text{-}APO}$>20 kJ/mol, on par with the strongest effects in the TFs that arise from protein-sugar interactions (Fig. 4B, C).

Each TF is globally stabilized by inducer binding (Fig. 4D-F), particularly in the binding pocket, pocket-periphery, and dimer interface. The N-terminal VFT dimer interface experiences pervasive inducer-triggered allosteric structuring: part of the transducer helix is stabilized, and strands E5, E1, and E9 unite to form a super-sheet with a hydrogen bond pattern distinct from the apo and operator-bound states (Fig. S4-4, S13). The differences in how inducer and operator binding stabilize this region suggest that TF induction includes a strand slipping mechanism that favors an energetically similar yet configurationally distinct beta sheet with a different strand register (Figs. S4-4, S11A, S13). Strand slippage is unlikely to occur spontaneously because it requires a substantial energetic disruption to other local structural elements, but was previously implicated in mediating conformational changes in a membrane

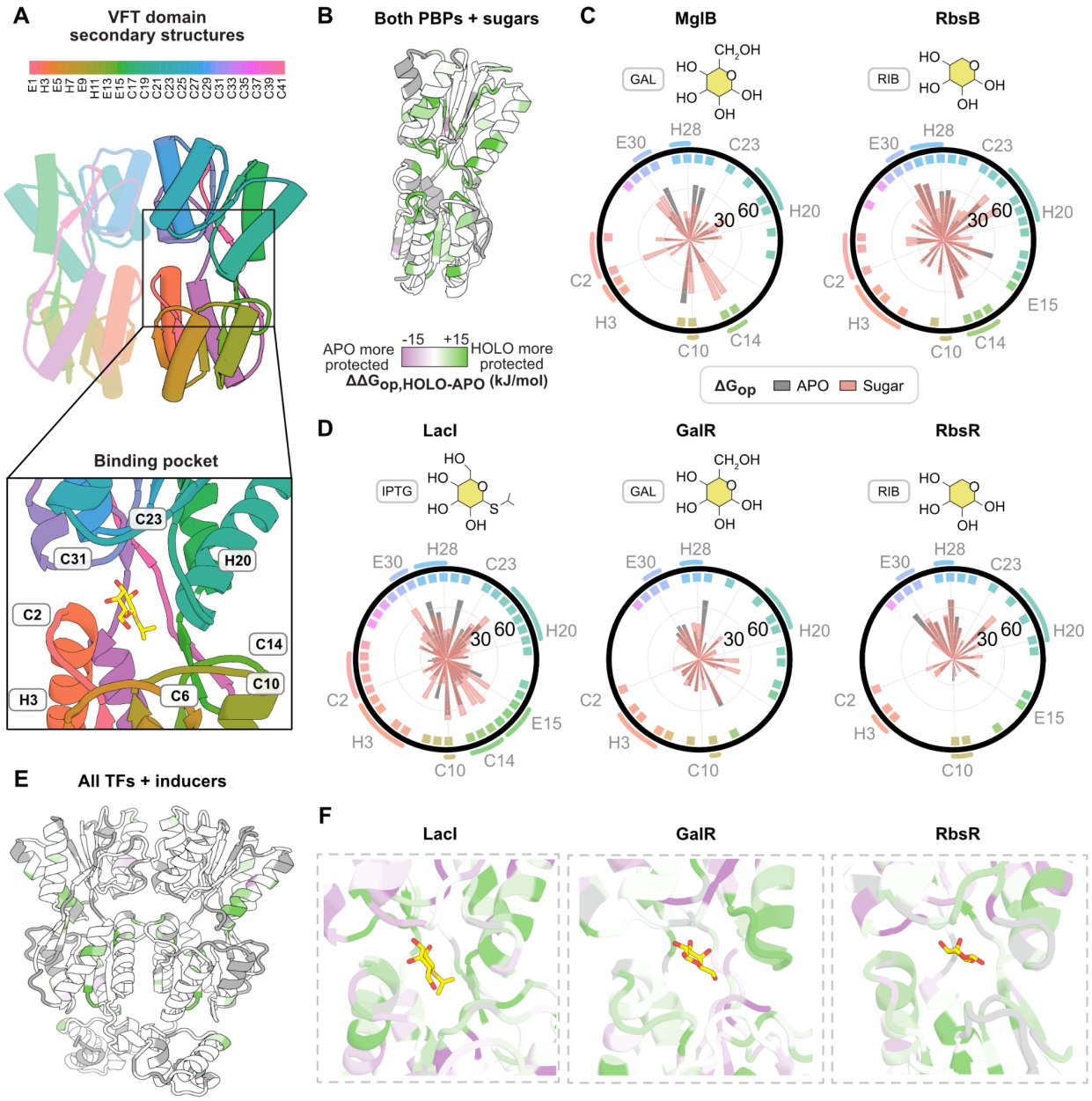


**Figure 4. Ligand binding in PBPs and TFs. A**, Color-coded secondary structures of the VFT domain. **B**, Conserved energetic effects of sugar binding in the two PBPs. Green regions are stabilized by sugar binding in both PBPs. **C**, Energetic redistribution in the PBPs upon sugar binding. The radar plots represent the ligand binding site. Each box at the radar plot edge, colored by secondary structure as in **A**, corresponds to a different amino acid lining the binding pocket. The spokes, colored by functional state, indicate the $\Delta G_{op}$ for these amino acids. For example, in MglB, residues in loop C14 are stabilized by sugar binding (increased $\Delta G_{op}$). This loop is colored green in **A**. **D**, Energetic redistribution in the TFs upon sugar binding, colored as in **C**. **E**, Conserved energetic effects of inducer binding in the three TFs. **F**, Molecular details of the TF binding pockets. **E** and **F** are colored by $\Delta\Delta G_{op,\ HOLO\text{-}APO}$, as in **B.**

protein (*39*). Helices 16 and 20 at the binding pocket periphery are also critical to inducer binding, corroborating mutational studies of LacI (Fig. 4F) (*16*, *19*). Finally, four positions are destabilized in response to inducer binding in all three TFs to compensate for locally increased helical stability.

To understand the baseline requirements for sugar binding in the VFT fold, we collated conserved effects in all five proteins (Fig. S11D). All proteins are stabilized by sugar binding in the binding pocket helices ($\Delta\Delta G_{op,\ HOLO\text{-}APO}$>10 kJ/mol). However, while there are many family-specific energetic changes, we observed few pan-VFT effects, which underscores how the fold is uniquely co-opted by each family.

**Water molecules structure the induced states of LacI/GalR TFs**

HX/MS shows that inducer binding stabilizes the transducer helix in the binding pocket of each TF (Fig. 4D), but crystal structures suggest that this requires water-mediated hydrogen bonds (Fig. S4-2) (*15*, *16*). We were curious whether structural water molecules, defined as waters that form three or more hydrogen bonds with protein or inducer atoms, are evolutionarily conserved to maintain the induced TF ensemble (*16*). To explore this possibility, we used RosettaECO (Explicit Consideration of cOordinated water) (*40*) to model structural waters in the induced states of several sugar-inducible TFs for which experimental evidence of sugar-binding and induction is available. We identified a high-confidence water molecule in the binding pocket of sugar-binding LacI/GalR TFs near the dimer interface, bridging the sugar with the transducer helix (Fig. S14). In support of the models, this water is found with very low b-factors in RbsR-ribose (b=11.73) and LacI-IPTG (b=23.80) crystal structures (Fig. S4-2) (*15*). As an interesting exception, in the trehalose repressor (TreR), the phosphoryl group of the inducer trehalose-6-phosphate acts as the bridge (*20*) (Fig. S14C), whereas unphosphorylated trehalose is an anti-inducer that stabilizes binding on the operator DNA (*41*). The LacI anti-inducer ONPF similarly disrupts the water network established by IPTG (*16*).

Using the crystal structures of LacI-IPTG and RbsR-ribose and our model for GalR-galactose, we performed atomistic MD simulations as an orthogonal method to explore how structural waters contribute to the induced TF ensembles (Fig. 5B). We extracted the waters within 3.5 Å of the sugars and performed a clustering analysis to find clusters with >80% occupancy (Fig. S15). Echoing the RosettaECO results, the simulations identified the same water near the transducer helix to have high occupancy in all TFs (Fig. 5C). The highest duration/occupancy for this structural water was in LacI, due to two hydroxyl groups in IPTG positioned towards the dimer interface, rather than only one in ribose; in GalR, it has near 100% occupancy. The water does not appear in operator-bound TFs (Fig. S4-2) (*27*), and we expect only transient waters in the binding pockets of apo TFs. These data support the hypothesis that the sugar-inducible TFs evolved to maintain not just the VFT fold or functionally important conformational changes, but also specific interactions with water that contribute to the induced state ensembles. PBPs also make a similar water-mediated interaction with sugars (Fig S16), suggesting that the pan-VFT-fold stabilization of the N-terminus of the transducer helix upon sugar binding relies partially on a structured water molecule.

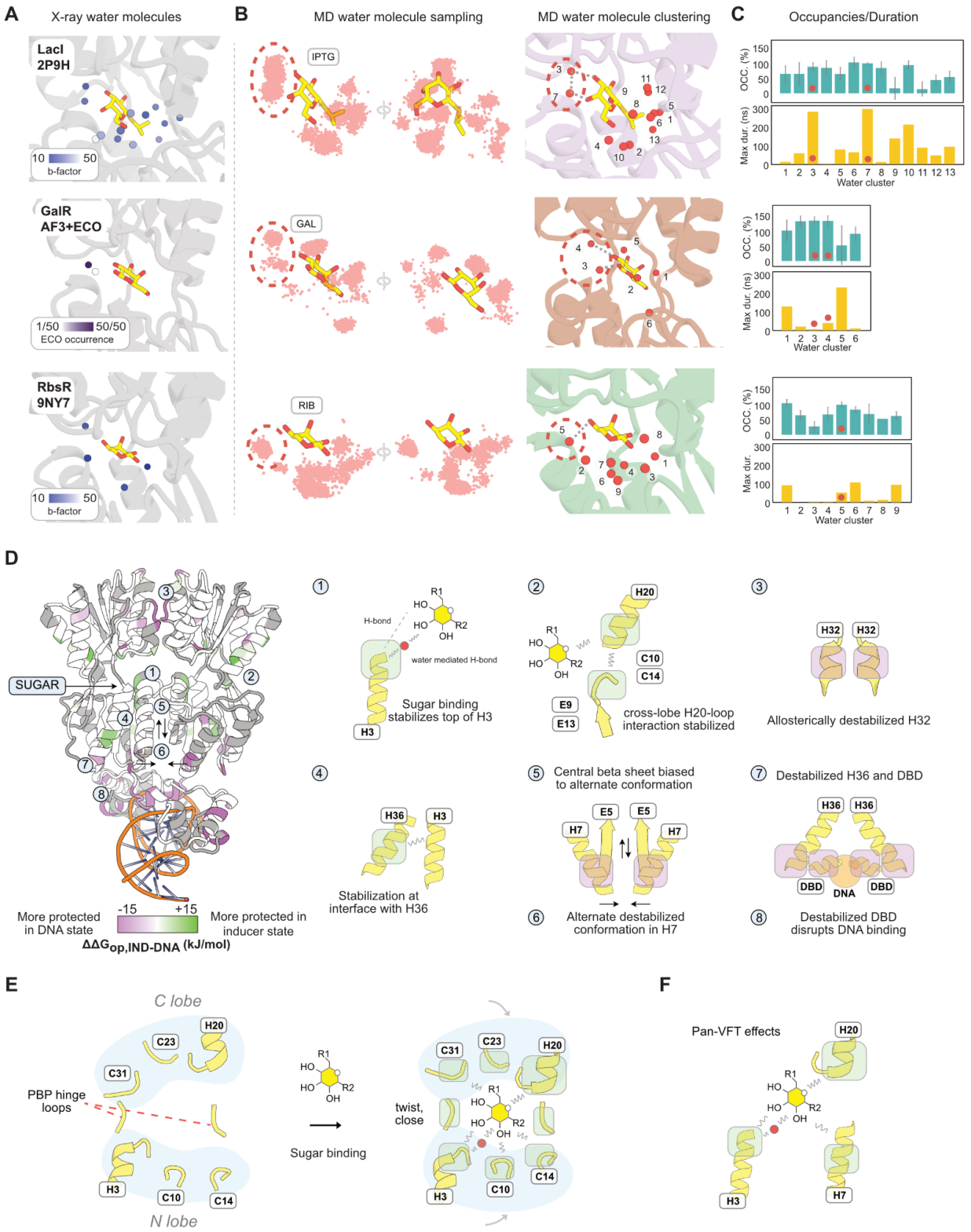


**Figure 5. Energetic changes in TFs vs. PBPs. A**, Binding pocket waters in TF structures and models. **B,** Waters sampled in MD simulations cluster together. **C**, Durations and occupancies for each water cluster in MD simulations. Red dots correspond to circled structural waters in **B**. **D**, Model for induction in LacI/GalR TFs based on conserved $\Delta\Delta G_{op,\ IND\text{-}DNA}$ (*left*), with molecular switches labeled and schematized on the *right*. **E**, Conserved mechanism of sugar response in both PBPs. **F**, Pan-VFT energetic relationships center on sugar interactions with binding pocket helices.

## Molecular switches toggle between TF functional states

PBPs respond to sugars with a hinge-mediated conformational change that is visible in structures (Fig. S12). Sugar binding stabilizes them in a "closed" conformational ensemble that they can already sample in the apo state, with locked inter-lobe loops forming a perfect binding interface for sugar transporters. But in TFs, the operator- and sugar-bound states have incompatible energetic requirements along the dimer and VFT-DBD interfaces, and the transition between states is invisible in structural models. Ultimately, the TFs evolved an energetic blueprint that is not a smaller-scale version of PBP response to sugars as we originally hypothesized, but which is instead unique to the family (Figs. 5D, S17).

A set of "molecular switches" transition the LacI/GalR TFs to the induced state by restructuring the hinge helix (*35*), but also, crucially, the VFT dimer interface, VFT cross-lobe interactions, and the VFT-DBD interface (Figs. 5D, S17, Supplemental Note 2). Inducer- and operator-driven conformational changes at switch positions have substantial $\Delta\Delta G_{op,DNA\text{-}ind}$ of 10-20 kJ/mol. Inducer binding to the DNA-bound TFs rigidifies the transducer helix at its N-terminus at the expense of the stability of the C-terminus, stabilizes inter-lobe interactions, and destabilizes the metastable helix, like an "undo" button for operator binding. Simultaneously, the hydrogen bond network in the N-terminal VFT beta sheet is rearranged, and several helices adjacent to the DBD are restructured. In the VFT-DBD interface, inducer binding stabilizes H36 residues to reduce hydrophobic interactions with the loop that joins the hinge helix to the VFT domain, thereby destabilizing the DBD. Simultaneously, reduced hydrogen bonding between H7 and the hinge helices releases them from the minor groove of the operator DNA, linearizing it. In support of this mechanism, mutations in this region modulate LacI-operator DNA affinity (*19*, *42–44*). Destabilization of the DBD is conserved in the induced state, resulting in decreased $\Delta G_{op}$ at recognition helix positions key to operator specificity (*38*). Additionally, nonspecific charged interactions with DNA are destabilized, as are DBD helical packing interactions. While sugars stabilize some of the binding pocket helices in both PBPs and TFs, few of these long-range effects that define the energetic blueprint of TFs are manifested in PBPs (Fig. 5E, F).

## Discussion

Coordinating switching between operator DNA- and specific sugar-bound states for LacI/GalR TFs is a life-or-death necessity for bacteria. Here, we discovered how the distribution of ensemble energies in the LacI/GalR TFs is dramatically different when bound to operators vs. sugars, though their X-ray crystal structures are similar. We found specific residues in the TF family that behave as molecular switches by occupying different, mutually exclusive energies in different functional states. By the evolutionary conservation of these switches and their state-specific interactions with protein, ligand, and water molecules, the TFs undergo a shift in their conformational ensembles that forbids prolonged operator DNA binding simultaneously with inducer binding. This is the invisible family secret that enables transcriptional regulation.

The VFT fold is not unique to LacI/GalR TFs. By comparing the TFs to their distant cousins, the PBPs, we determined how one protein fold performs two different functions in response to the same small molecule ligands. This yields a quantitative description of how specific energetic relationships persisted and adapted over the evolutionary timescale to protect functionally important conformational changes from being lost, even as new chemical triggers

challenged the remarkable plasticity of the fold. We can use this blueprint to rationally and precisely engineer VFT-fold sensors, which has historically proven challenging for both TFs and PBPs despite their promise in biotechnology (*18*, *45–48*).

The story of conserved energetic relationships in a classic TF and its cousins demonstrates how regulators of the genome have evolved alongside the genome itself, and that the energetic blueprint for the family, buried in the evolutionary record, is key to designing transcriptional regulators *de novo*. Previous efforts to design proteins that bind to DNA sequences specifically and with high affinity are mired by the repetitive degeneracy of DNA structures and a bias towards modeling linear DNA. In one recent effort to design DNA-binding proteins, fourteen sequence-specific protein binders were identified from screening more than 140,000 design models, many of which were designed with an experimentally solved target DNA structure available (*49*). Perhaps designing transcriptional regulators *de novo* requires considering two-way allostery: how the TF changes the conformational ensemble of DNA, and how different DNA sequences reshape the conformational ensemble of the TF.

**Acknowledgements**

We thank members of the Glasgow Lab for fruitful discussions, Dr. Wayne Hendrickson for advice on X-ray crystallography, Dr. Naomi Latorraca for feedback on the manuscript, and Dr. Rinat Abzalimov at the Advanced Science Research Center at the City University of New York in his role as the MS Facility Manager. We thank the staff at NYX beamline at NSLS-II for their support during data collection. We acknowledge support from the National Institutes of Health (R00GM135529 to AG). KCW was funded by a National Science Foundation Graduate Research Fellowship.

**Data availability statement**

The HX/MS deuterium uptake plots, FEATHER-derived centroid fits to HX/MS data, and a PyMOL session file showing protein-specific differences in $\Delta G_{op}$ are available in the Supplementary Materials. The HX/MS data have been deposited to the ProteomeXchange Consortium via the PRIDE (*50*) partner repository with the dataset identifier PXD062370. The X-ray crystal structures are uploaded to the PDB with codes 9NY7 and 9NY8. The AlphaFold 3-RosettaECO models, MD trajectories, conservation score matrices, gene trees, MSA files, and X-ray diffraction images are available in Zenodo (dataset identifier 10.5281/zenodo.15091462).

**Methods**

**Calculation of conservation scores**

*Multiple sequence alignment (MSA)*. The MSA was generated using MMSeqs2 (*52*) separately for TFs (LacI as the query) and PBPs (RbsB as the query) with bacterial sequences from UniRef90 as the search database, using the following command line:

*mmseqs filtertaxseqdb ./Uniref90_DB UniRef90_bacteria --taxon-list 2 -s 6 --max-seqs 20000 -a 1 -e 0.01 --min-aln-len 150 --allow-deletion 1 --threads 60*.

*Conservation scores.* We used the global epistatic model GEMME (*53*) which takes into account both MSA and evolutionary history of the sequences using the MSA generated above. We used the following command line with the provided docker:

*docker run -ti --rm --mount type=bind,source=$PWD,target=/project elodielaine/gemme:gemme*
then
*python2.7 $GEMME_PATH/gemme.py MSA.fasta -r input -f MSA.fasta*

The combined prediction scores were normalized within each group of proteins (TFs or PBPs) such that the highest conservation score is 1.

**Gene tree construction**
For the input sequence alignment, we combined MSAs from the previous step and added the sequence for LuxP (Uniprot P54300) to initially use as an outgroup (*7*). We used only the VFT portion of the MSA and further excluded positions with >10% gaps. We used IQTree (*54*) for the tree construction. We estimated a substitution model using the following command line:

*iqtree -s ./MSA.fasta -m MFP -B 1000 -bnni -T AUTO .*

We chose Q.pfam+R10 (*55*) as the best-fit model based on Bayesian information criterion (BIC) (*56*) and ran the following command:

*iqtree -s ./MSA.fasta -m Q.pfam+R10 -B 1000 -bnni -T AUTO -safe -cptime 600 .*
Structural alignment of TFs
We collected experimentally solved structures of LacI/GalR TFs from the PDB (*57*) using the following query:

*( ( Lineage Name = "GalR/LacI-like bacterial regulator" AND Annotation Type = "SCOP" ) OR ( Lineage Name HAS ANY OF WORDS "LacI,GalR" AND Annotation Type = "ECOD" ) OR ( Lineage Identifier = "IPR046335" AND Annotation Type = "InterPro" ) ) AND Polymer Entity Sequence Length > 90*

The pairwise RMSD was calculated using US-align (*58*).

**Plasmid construction**
DNA sequences encoding RbsR, and GalR were synthesized as double-stranded genes by IDT DNA and cloned into a pET9a vector with an N-terminal 6× histidine tag and TEV protease sequence using a Golden Gate strategy with type IIs restriction enzyme BsaI (*59*). In the case of LacI, we used a previously described plasmid with the same architecture (*16*, *17*). The MglB and RbsB expression plasmids were also constructed with N-terminal 6× histidine tag and TEV site using the same strategy in the same vector, but the genes were amplified by polymerase chain reaction from the *E. coli* DH10B genome. All genetic constructs were confirmed by sequencing. All gene sequences are available in Table S15.

**Protein expression and purification**
Proteins were purified as described previously (*17*), with the exception that GalR buffers for cell lysis, purification, and storage included 600 mM NaCl to avoid protein aggregation (*60*).

**X-ray crystallography sample preparation**
To prepare the *rbsO*-RbsR sample, self-complementary single-stranded 30-base-pair primers encoding *rbsO* (5′-GTGGGTCAGCGAAACGTTTCGCTGATGGAG-3′) were synthesized (IDT DNA) and annealed using a previously described protocol (*16*). RbsR dimer (10 mg/mL) was then incubated at 4 °C for 12 h with a two-fold molar excess of the annealed DNA. After incubation, the mixture was centrifuged (14,000 rpm, 4 °C, 10 min), and crystals were grown by hanging-drop vapor diffusion using 400 µL of the protein–DNA complex solution and 300 µL of reservoir solution containing 0.2 M sodium sulfate, 0.1 M bis-tris-propane (pH 6.5), and 20% PEG 3350 (PACT Premier 2-20) at 20 °C.

To prepare the ribose-RbsR sample, RbsR (10.1 mg/mL) was incubated with a six-fold molar excess of ribose (1.3 mM) for 12 h at 4 °C and subsequently centrifuged (14,000 rpm, 4 °C, 10 min). The resulting mixture was crystallized by hanging-drop vapor diffusion using 400 µL of the protein–ribose complex solution and 300 µL of reservoir solution containing 0.2 M sodium malonate (pH 4.0) and 20% PEG 3350 (Pegion2 A5) at 20 °C.

**X-ray crystallography data collection and model building**
Diffraction data were collected at the NYX beamline 19-ID of the National Synchrotron Light Source II (NSLS-II) at Brookhaven National Laboratory. The raw diffraction HDF5 images were processed using Xia2 (*61*) with DIALS (*62*) pipeline, as implemented in CCP4 suite (*63*), for ribose-bound RbsR structure and XDS (*64*) for *rbsO*-bound structure to perform indexing, integration and initial scaling. Data scaling and merging were carried out using AIMLESS or STARANISO (*65*). Molecular replacement was performed with Phaser (*66*), using an AlphaFold 2 model (*67*) as the search model for the ribose-bound RbsR structure, and an AlphaFold 3 model (*36*) as the input model for the *rbsO*-bound structure. Coot (*68*) was used for manual modelling. Structural refinement was conducted using both Phenix.refine (*69*) and REFMAC (*70*). Final validation and refinement were completed for both using PDB-Redo (*71*).

**HX/MS sample preparation and experiment**
Sample preparation was performed as described previously (*17*), except for the following variables: sample buffer composition, protein stock composition, and quench composition, as specified in Tables S3-S7. All proteins were purified by size exclusion chromatography, confirmed to be >95% pure by SDS-PAGE, concentrated to 5-10 µM, and flash-frozen in aliquots. HX/MS experiments were performed by diluting freshly purified proteins or thawed aliquots by tenfold in an equivalent Tris-NaCl $D_2O$ buffer at pH 8.0. IPTG, galactose, and ribose were added to protein samples at 10 mM. Operator DNAs were added at twofold molar excess of TF dimer concentrations. HX experiments were performed as described previously for LacI (*17*) except for the following variables: reaction and injection volumes, and protease columns, specified in Tables S3-S7. After incubation in $D_2O$ buffer, the exchange reaction was quenched by 1.9-fold sample dilution in quench buffer (3% acetonitrile, 1% formic acid in $H_2O$, with GuHCl concentrations noted in Tables S3-S7 for all proteins, all MS-grade, Fisher Scientific).

**HX/MS fully deuterated sample preparation**

Fully deuterated controls were prepared as described previously (*17*) (Tables S3-S7). Of note, due to some observed aggregation, we removed GuHCl from the preparation of fully deuterated control samples for GalR samples. We allowed them to exchange at 37 ºC for at least 24 hours and used a quench buffer containing GuHCl as in the other HX experiments.

The protein stock solutions were diluted ten-fold in deuterated buffer and exchanged at 15 ºC for the specified time, then diluted in quench solution (75 µl, 3 M GuHCl, 3% acetonitrile, 1% formic acid in $H_2O$, all MS-grade, Fisher Scientific) of equal volume to the deuterated buffer at 2 ºC. All but 10 µl of quenched reaction was injected into the protease column at 7 ºC. Solvent A (3% acetonitrile, 0.15% formic acid in $H_2O$) was pumped (UltiMate 3000, ThermoFisher Scientific) at 150 µL/min through the protease column. Peptides eluted from the protease column were trapped on a 1x10 mm C18 column (Hypersil GOLD, 3 µM pore size, ThermoFisher Scientific).

**HX/MS data analysis**

MS/MS peak picking and deconvolution, disambiguation by peak picking, peptide disambiguation by PIGEON, FEATHER rate fitting, PIGEON error analysis, and quality control of all fit spectra were performed as described previously for LacI (*17*). Conservation among FEATHER-derived $\Delta G_{op}$ was determined by treating as nonzero only state-state differences that exceeded our significance threshold and occurred in the same direction for all proteins in the comparison set. We calculated the mean of these positional $\Delta\Delta G_{op}$, wherever mean$\{\Delta\Delta G_{op}\}$ > 2 kJ/mol. For a given protein, state-state differences at each site were considered significant if the "mini-peptide" (the set of residues including that site covered indistinguishably by a set of fit exchange rates) was the same for both states and satisfied $mean\{\Delta G_{op,a,i}-\Delta G_{op,b,i}\} > std\{\Delta G_{op,a,i}-\Delta G_{op,b,i}\}$, where the $\{\Delta G_{op,a,i}\}$ are the sorted $\Delta G_{op}$ corresponding to that site for each exchange rate in the mini-peptide, in state *a*. In all structural models presented in this work that are colored by $\Delta\Delta G_{op}$ conserved among 2+ proteins, we calculated the average positional $\Delta\Delta G_{op}$ for those proteins, if the selection criteria described above were met. Otherwise, we colored those positions white to indicate that the exchange behavior was not conserved. Gray coloring indicates a lack of data for at least one protein or state, or high standard deviation (noise) in the significance calculation. The full sets of $\Delta G_{op}$, their standard deviations, and their resolutions are available for all proteins in Tables S8-S12.

**Structural water prediction using RosettaECO**

The experimental structures or AlphaFold 3 (AF3) (*36*) models were first relaxed using the Rosetta biomolecular modeling suite (*72*). The experimental structures were relaxed with backbone and sidechain constraints using the following command line to produce three models for each input file:

*./relax.default.linuxgccrelease -s input.pdb -out:suffix .constrain.relax -nstruct 3 -relax:default_repeats 5 -out:path:pdb ./ -out:path:score ./ -in:file:extra_res_fa ./tre.params -constrain_relax_to_start_coords true -relax:coord_constrain_sidechains true*

Three models were produced for each input protein. The structural model with lowest score in Rosetta Energy Units (REU) was chosen as the best model. For the AF3 models, we performed unconstrained relaxation using the command:

*./relax.default.linuxgccrelease -s input.pdb -out:suffix .relax -nstruct 3 -relax:default_repeats 5 -out:path:pdb ./ -out:path:score ./ -in:file:extra_res_fa ./param_fle.params*

The ligand was specified through the "param_file.params" file where needed. To solvate the models using Rosetta Explicit Consideration of cOordinated water (RosettaECO) (*40*), we used an .xml protocol (see Data availability) to generate 50 solvated models per input unsolvated relaxed model:

*./rosetta_scripts.default.linuxgccrelease -parser:protocol solvate.xml -s input.pdb -nstruct 50 -beta_nov16 -out:suffix .solvate -out:path:pdb ./ -in:file:extra_res_fa ./param_file.params*

Water molecules within 0.5 Å of each other were clustered and reported in the final solvated model (see Data availability; the number of waters in each model is reported as "b-factor" in the final .pdb file).

**Molecular dynamics simulations**

Five sugar-bound VFT-fold systems (LacI-IPTG, GalR-galactose, RbsR-ribose, MglB-galactose, and RbsB-ribose) were built for MD simulations using the following structures by PDB ID: 2P9H, AF3 model, 9NY7, 1GLG, and 2DRI as the input models, respectively. The CHARMM36 force field was applied and the simulations were performed in the GROMACS package (version 2022). We energy minimized the systems using the steepest descent minimization method until the maximum force was less than 1000 kJ/mol/nm$^2$, followed by 100 ps restrained MD simulations in NVT ensemble by constraining the heavy atoms of the protein and ligand at 298.15 K. The resulting simulation system was further equilibrated for a 1000 ps NPT simulation at 1 atm and 298.15 K with a 1000 kJ/mol/nm$^2$ position restraints on the protein and ligand. Finally, we conducted the NPT production MD simulations using the same conditions without restraints, with a time step of 2 fs, nonbonded cutoff of 12 Å, and particle-mesh Ewald long-range electrostatics. Five 300 ns simulation replicates were collected for each system. The MD trajectories were analyzed using MDAnalysis (*73*), VMD (*74*), or MDTraj (*75*).

**Water clustering analysis**

The trajectories from the production runs for each system were aligned to its initial structure based on the protein backbone atoms, and the coordinates of water molecules located within 3.5 Å of the ligands were extracted. These coordinates were subsequently clustered using the *k-means* algorithm implemented in the scikit-learn (*76*) Python package. The number of the clusters was determined by the elbow method.

**References**

1. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives,

Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
2. C. Chothia, One thousand families for the molecular biologist. *Nature* **357**, 543–544 (1992).
3. H. Li, C. Tang, N. S. Wingreen, Are protein folds atypical? *Proceedings of the National Academy of Sciences* **95**, 4987–4990 (1998).
4. P. Cossio, A. Trovato, F. Pietrucci, F. Seno, A. Maritan, A. Laio, Exploring the Universe of Protein Structures beyond the Protein Data Bank. *PLOS Computational Biology* **6**, e1000957 (2010).
5. B. Chitturi, S. Shi, L. N. Kinch, N. V. Grishin, Compact Structure Patterns in Proteins. *Journal of Molecular Biology* **428**, 4392–4412 (2016).
6. S. Minami, N. Kobayashi, T. Sugiki, T. Nagashima, T. Fujiwara, R. Tatsumi-Koga, G. Chikenji, N. Koga, Exploration of novel αβ-protein folds through de novo design. *Nature Structural and Molecular Biology* **30**, 1132–1140 (2023).
7. K. Fukami-Kobayashi, Y. Tateno, K. Nishikawa, Parallel evolution of ligand specificity between LacI/GalR family repressors and periplasmic sugar-binding proteins. *Molecular Biology and Evolution* **20**, 267–277 (2003).
8. J. A. Kaczmarski, M. C. Mahawaththa, A. Feintuch, B. E. Clifton, L. A. Adams, D. Goldfarb, G. Otting, C. J. Jackson, Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme. *Nature Communications* **11**, 5945 (2020).
9. K. R. Torgeson, M. W. Clarkson, D. Granata, K. Lindorff-Larsen, R. Page, W. Peti, Conserved conformational dynamics determine enzyme activity. *Science Advances* **8**, eabo5546 (2022).
10. J. Cao, S. Huang, J. Qian, J. Huang, L. Jin, Z. Su, J. Yang, J. Liu, Evolution of the class C GPCR Venus flytrap modules involved positive selected functional divergence. *BMC Evolutionary Biology* **9**, 67 (2009).
11. L. Swint-Kruse, K. S. Matthews, Allostery in the LacI/GalR family: variations on a theme. *Current Opinion in Microbiology* **12**, 129–137 (2009).
12. B. H. Shilton, M. M. Flocco, M. Nilsson, S. L. Mowbray, Conformational Changes of Three Periplasmic Receptors for Bacterial Chemotaxis and Transport: The Maltose-, Glucose/Galactose- and Ribose-binding Proteins. *Journal of Molecular Biology* **264**, 350–363 (1996).
13. S. L. Mowbray, A. J. Björkman, Conformational changes of ribose-binding protein and two related repressors are tailored to fit the functional need. *Journal of Molecular Biology* **294**, 487–499 (1999).
14. D. J. Parente, L. Swint-Kruse, Multiple Co-Evolutionary Networks Are Supported by the Common Tertiary Scaffold of the LacI/GalR Proteins. *PLOS ONE* **8**, e84398 (2013).
15. R. Daber, S. Stayrook, A. Rosenberg, M. Lewis, Structural Analysis of Lac Repressor Bound to Allosteric Effectors. *Journal of Molecular Biology* **370**, 609–619 (2007).
16. A. Glasgow, H. T. Hobbs, Z. R. Perry, M. L. Wells, S. Marqusee, T. Kortemme, Ligand-specific changes in conformational flexibility mediate long-range allostery in the lac repressor. *Nature Communications* **14**, 1179 (2023).
17. C. Lu, M. L. Wells, A. Reckers, A. Glasgow, Site-resolved energetic information from HX/MS experiments. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.08.04.606547.
18. P. Kröger, S. Shanmugaratnam, N. Ferruz, K. Schweimer, B. Höcker, A comprehensive binding study illustrates ligand recognition in the periplasmic binding protein PotF. *Structure* **29**, 433-443.e4 (2021).
19. P. Markiewicz, L. G. Kleina, C. Cruz, S. Ehret, J. H. Miller, Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as “spacers” which do not require a specific sequence. *Journal of Molecular Biology* **240**, 421–433 (1994).

20. U. Hars, R. Horlacher, W. Boos, W. Welte, K. Diederichs, Crystal structure of the effector-binding domain of the trehalose-repressor of Escherichia coli, a member of the LacI family, in its complexes with inducer trehalose-6-phosphate and noninducer trehalose. *Protein Science* **7**, 2511–2521 (1998).
21. M. A. Schumacher, A. Glasfeld, H. Zalkin, R. G. Brennan, The X-ray Structure of the PurR-Guanine-purF Operator Complex Reveals the Contributions of Complementary Electrostatic Surfaces and a Water-mediated Hydrogen Bond to Corepressor Specificity and Binding Affinity. *Journal of Biological Chemistry* **272**, 22648–22653 (1997).
22. M. A. Schumacher, G. S. Allen, M. Diel, G. Seidel, W. Hillen, R. G. Brennan, Structural basis for allosteric control of the transcription regulator CcpA by the phosphoprotein HPr-Ser46-P. *Cell* **118**, 731–741 (2004).
23. M. L. Oldham, D. Khare, F. A. Quiocho, A. L. Davidson, J. Chen, Crystal structure of a catalytic intermediate of the maltose transporter. *Nature* **450**, 515–521 (2007).
24. J. J. Birktoft, D. M. Blow, Structure of crystalline α-chymotrypsin: V. The atomic structure of tosyl-α-chymotrypsin at 2 Å resolution. *Journal of Molecular Biology* **68**, 187–240 (1972).
25. J. A. Endicott, M. E. M. Noble, L. N. Johnson, The Structural Basis for Control of Eukaryotic Protein Kinases. *Annual Review of Biochemistry* **81**, 587–613 (2012).
26. A. M. Friedman, T. O. Fischmann, T. A. Steitz, Crystal Structure of lac Repressor Core Tetramer and Its Implications for DNA Looping. *Science* **268**, 1721–1727 (1995).
27. M. Lewis, G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, P. Lu, Crystal Structure of the Lactose Operon Repressor and Its Complexes with DNA and Inducer. *Science* **271**, 1247–1254 (1996).
28. C. E. Bell, M. Lewis, A closer view of the conformation of the Lac repressor bound to operator. *Nature Structural Biology* **7**, 209–214 (2000).
29. C. a. Mauzy, M. a. Hermodson, Structural and functional analyses of the repressor, RbsR, of the ribose operon of Escherichia coli. *Protein Science* **1**, 831–842 (1992).
30. V. J. Hilser, An Ensemble View of Allostery. *Science* **327**, 653–654 (2010).
31. B. von Wilcken-Bergmann, B. Müller-Hill, Sequence of galR gene indicates a common evolutionary origin of lac and gal repressor in Escherichia coli. *Proceedings of the National Academy of Sciences* **79**, 2427–2431 (1982).
32. A. Majumdar, S. Adhya, Probing the structure of gal operator-repressor complexes. Conformation change in DNA. *Journal of Biological Chemistry* **262**, 13258–13262 (1987).
33. Y. Bai, T. R. Sosnick, L. Mayne, S. W. Englander, Protein Folding Intermediates: Native-State Hydrogen Exchange. *Science* **269**, 192–197 (1995).
34. C. G. Kalodimos, N. Biris, A. M. J. J. Bonvin, M. M. Levandoski, M. Guennuegues, R. Boelens, R. Kaptein, Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. *Science* **305**, 386–389 (2004).
35. C. A. E. M. Spronk, M. Slijper, J. H. van Boom, R. Kaptein, R. Boelens, Formation of the hinge helix in the lac represser is induced upon binding to the lac operator. *Nature Structural and Molecular Biology* **3**, 916–919 (1996).
36. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
37. A. Schmitz, U. Schmeissner, J. H. Miller, Mutations affecting the quaternary structure of the lac repressor. *Journal of Biological Chemistry* **251**, 3359–3366 (1976).

38. C. A. Spronk, A. M. Bonvin, P. K. Radha, G. Melacini, R. Boelens, R. Kaptein, The solution structure of Lac repressor headpiece 62 complexed to a symmetrical lac operator. *Structure* **7**, 1483–1492 (1999).
39. K. Makabe, S. Yan, V. Tereshko, G. Gawlak, S. Koide, β-Strand Flipping and Slipping Triggered by Turn Replacement Reveal the Opportunistic Nature of β-Strand Pairing. *Journal of the American Chemical Society* **129**, 14661–14669 (2007).
40. R. E. Pavlovicz, H. Park, F. DiMaio, Efficient consideration of coordinated water molecules improves computational protein-protein and protein-ligand docking discrimination. *PLOS Computational Biology* **16**, e1008103 (2020).
41. R. Horlacher, W. Boos, Characterization of TreR, the major regulator of the Escherichia coli trehalose system. *Journal of Biological Chemistry* **272**, 13026–13032 (1997).
42. C. M. Falcon, K. S. Matthews, Glycine Insertion in the Hinge Region of Lactose Repressor Protein Alters DNA Binding*. *Journal of Biological Chemistry* **274**, 30849–30857 (1999).
43. L. G. Kleina, J. H. Miller, Genetic studies of the *lac* repressor. *Journal of Molecular Biology* **212**, 295–318 (1990).
44. P. Campitelli, L. Swint-Kruse, S. B. Ozkan, Substitutions at Nonconserved Rheostat Positions Modulate Function by Rewiring Long-Range, Dynamic Interactions. *Molecular Biology and Evolution* **38**, 201–214 (2021).
45. M. A. Dwyer, H. W. Hellinga, Periplasmic binding proteins: a versatile superfamily for protein engineering. *Current Opinion in Structural Biology* **14**, 495–504 (2004).
46. N. D. Taylor, A. S. Garruss, R. Moretti, S. Chan, M. A. Arbing, D. Cascio, J. K. Rogers, F. J. Isaacs, S. Kosuri, D. Baker, S. Fields, G. M. Church, S. Raman, Engineering an allosteric transcription factor to respond to new ligands. *Nature Methods* **13**, 177–183 (2016).
47. S. Tungtur, S. M. Egan, L. Swint-Kruse, Functional consequences of exchanging domains between LacI and PurR are mediated by the intervening linker sequence. *Proteins: Structure, Function, and Bioinformatics* **68**, 375–388 (2007).
48. D. H. Richards, S. Meyer, C. J. Wilson, Fourteen Ways to Reroute Cooperative Communication in the Lactose Repressor: Engineering Regulatory Proteins with Alternate Repressive Functions. *ACS Synthetic Biology* **6**, 6–12 (2017).
49. C. J. Glasscock, R. Pecoraro, R. McHugh, L. A. Doyle, W. Chen, O. Boivin, B. Lonnquist, E. Na, Y. Politanska, H. K. Haddox, D. Cox, C. Norn, B. Coventry, I. Goreshnik, D. Vafeados, G. R. Lee, R. Gordan, B. L. Stoddard, F. DiMaio, D. Baker, Computational design of sequence-specific DNA-binding proteins. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.09.20.558720.
50. Y. Perez-Riverol, J. Bai, C. Bandla, D. García-Seisdedos, S. Hewapathirana, S. Kamatchinathan, D. J. Kundu, A. Prakash, A. Frericks-Zipper, M. Eisenacher, M. Walzer, S. Wang, A. Brazma, J. A. Vizcaíno, The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research* **50**, D543–D552 (2022).
51. Y. Bai, J. S. Milne, L. Mayne, S. W. Englander, Primary structure effects on peptide group hydrogen exchange. *Proteins: Structure, Function, and Bioinformatics* **17**, 75–86 (1993).
52. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026–1028 (2017).
53. E. Laine, Y. Karami, A. Carbone, GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Molecular Biology and Evolution* **36**, 2604–2619 (2019).
54. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).
55. B. Q. Minh, C. C. Dang, L. S. Vinh, R. Lanfear, QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. *Systematic Biology* **70**, 1046–1060 (2021).

56. G. Schwarz, Estimating the Dimension of a Model. *The Annals of Statistics* **6**, 461–464 (1978).
57. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
58. C. Zhang, M. Shine, A. M. Pyle, Y. Zhang, US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods* **19**, 1109–1115 (2022).
59. C. Engler, R. Gruetzner, R. Kandzia, S. Marillonnet, Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes. *PLoS ONE* **4**, e5553 (2009).
60. E. D. Agerschou, G. Christiansen, N. P. Schafer, D. J. Madsen, D. E. Brodersen, S. Semsey, D. E. Otzen, The transcriptional regulator GalR self-assembles to form highly regular tubular structures. *Scientific Reports* **6**, 27672 (2016).
61. G. Winter, xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography* **43**, 186–190 (2010).
62. G. Winter, D. G. Waterman, J. M. Parkhurst, A. S. Brewster, R. J. Gildea, M. Gerstel, L. Fuentes-Montero, M. Vollmar, T. Michels-Clark, I. D. Young, N. K. Sauter, G. Evans, DIALS: implementation and evaluation of a new integration package. *Acta Crystallographica Section D* **74**, 85–97 (2018).
63. J. Agirre, M. Atanasova, H. Bagdonas, C. B. Ballard, A. Baslé, J. Beilsten-Edmands, R. J. Borges, D. G. Brown, J. J. Burgos-Mármol, J. M. Berrisford, P. S. Bond, I. Caballero, L. Catapano, G. Chojnowski, A. G. Cook, K. D. Cowtan, T. I. Croll, J. É. Debreczeni, N. E. Devenish, E. J. Dodson, T. R. Drevon, P. Emsley, G. Evans, P. R. Evans, M. Fando, J. Foadi, L. Fuentes-Montero, E. F. Garman, M. Gerstel, R. J. Gildea, K. Hatti, M. L. Hekkelman, P. Heuser, S. W. Hoh, M. A. Hough, H. T. Jenkins, E. Jiménez, R. P. Joosten, R. M. Keegan, N. Keep, E. B. Krissinel, P. Kolenko, O. Kovalevskiy, V. S. Lamzin, D. M. Lawson, A. A. Lebedev, A. G. W. Leslie, B. Lohkamp, F. Long, M. Malý, A. J. McCoy, S. J. McNicholas, A. Medina, C. Millán, J. W. Murray, G. N. Murshudov, R. A. Nicholls, M. E. M. Noble, R. Oeffner, N. S. Pannu, J. M. Parkhurst, N. Pearce, J. Pereira, A. Perrakis, H. R. Powell, R. J. Read, D. J. Rigden, W. Rochira, M. Sammito, F. Sánchez Rodríguez, G. M. Sheldrick, K. L. Shelley, F. Simkovic, A. J. Simpkin, P. Skubak, E. Sobolev, R. A. Steiner, K. Stevenson, I. Tews, J. M. H. Thomas, A. Thorn, J. T. Valls, V. Uski, I. Usón, A. Vagin, S. Velankar, M. Vollmar, H. Walden, D. Waterman, K. S. Wilson, M. D. Winn, G. Winter, M. Wojdyr, K. Yamashita, The CCP4 suite: integrative software for macromolecular crystallography. *Acta Crystallographica Section D* **79**, 449–461 (2023).
64. W. Kabsch, XDS. *Acta Crystallographica Section D* **66**, 125–132 (2010).
65. STARANISO anisotropy & Bayesian estimation server. https://staraniso.globalphasing.org/cgi-bin/staraniso.cgi.
66. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. *Journal of Applied Crystallography* **40**, 658–674 (2007).
67. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
68. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D* **60**, 2126–2132 (2004).
69. P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, P. D. Adams, Towards

automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D* **68**, 352–367 (2012).
70. P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D* **67** (2011).
71. R. P. Joosten, F. Long, G. N. Murshudov, A. Perrakis, The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213–220 (2014).
72. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J. Gray, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13**, 3031–3048 (2017).
73. N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, O. Beckstein, MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry* **32**, 2319–2327 (2011).
74. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33–38 (1996).
75. R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, V. S. Pande, MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **109**, 1528–1532 (2015).
76. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).