Quantifying the Long-Range Coupling of Electronic Properties in Proteins with Ab Initio Molecular Dynamics

Zhongyue Yang, Natalia Hajlasz, Adam Steeves, Heather Kulik

Submitted date: 11/12/2020 • Posted date: 14/12/2020
Licence: CC BY-NC-ND 4.0
Citation information: Yang, Zhongyue; Hajlasz, Natalia; Steeves, Adam; Kulik, Heather (2020): Quantifying the Long-Range Coupling of Electronic Properties in Proteins with Ab Initio Molecular Dynamics. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.13368692.v1

A delicate interplay of covalent and noncovalent interactions gives proteins their unique ability to flexibly play numerous roles in cellular processes. This interplay is inherently quantum mechanical and highly dynamic in nature. To directly interrogate the evolving nature of the electronic structure of proteins, we carry out 100-ps-scale ab initio molecular dynamics simulations of three representative small proteins with range-separated hybrid density functional theory. We quantify the nature and length-scale of the coupling of residue-specific charge probability distributions in these proteins. While some nonpolar residues exhibit expectedly narrow charge distributions, most polar and charged residues exhibit broad, multimodal distributions. Even for nonpolar residues, we observe sequence-specific deviations corresponding to charge accumulation or depletion that would be challenging to capture in a fixed charge force field. We quantify the effect of residue–residue interactions on charge distributions first with linear cross-correlations. We then show how additional insight can be gained from evaluating the mutual information of charge distributions. We show that a significant number of residues couple most strongly with residues that are distant in both sequence and space over a range of secondary structures including α-helical, β-sheet, disulfide bridging, and lasso motifs. The mutual information analysis is necessary to capture coupling between some polar and charged residues. These analyses are expected to be broadly useful in understanding the mechanisms of long-range charge transfer in proteins and for determining what interactions require a quantum mechanical description for predictive simulation of enzyme mechanism and protein function.
ABSTRACT: A delicate interplay of covalent and noncovalent interactions gives proteins their unique ability to flexibly play numerous roles in cellular processes. This interplay is inherently quantum mechanical and highly dynamic in nature. To directly interrogate the evolving nature of the electronic structure of proteins, we carry out 100-ps-scale ab initio molecular dynamics simulations of three representative small proteins with range-separated hybrid density functional theory. We quantify the nature and length-scale of the coupling of residue-specific charge probability distributions in these proteins. While some nonpolar residues exhibit expectedly narrow charge distributions, most polar and charged residues exhibit broad, multimodal distributions. Even for nonpolar residues, we observe sequence-specific deviations corresponding to charge accumulation or depletion that would be challenging to capture in a fixed charge force field. We quantify the effect of residue–residue interactions on charge distributions first with linear cross-correlations. We then show how additional insight can be gained from evaluating the mutual information of charge distributions. We show that a significant number of residues couple most strongly with residues that are distant in both sequence and space over a range of secondary structures including α-helical, β-sheet, disulfide bridging, and lasso motifs. The mutual information analysis is necessary to capture coupling between some polar and charged residues. These analyses are expected to be broadly useful in understanding the mechanisms of long-range charge transfer in proteins and for determining what interactions require a quantum mechanical description for predictive simulation of enzyme mechanism and protein function.
1. Introduction

Proteins are ubiquitous in cellular processes and chemical transformations thanks to the structural flexibility and functional diversity imparted by the twenty natural amino acids that they comprise. Quantum mechanical (QM), non-covalent interactions play a critical role in the diverse structures and functions of proteins.\(^1\)\(^-\)\(^5\) Amino acid residues can form both stronger charge-assisted\(^6\)\(^-\)\(^{10}\) or low-barrier\(^1\)\(^,\)\(^{11}\)\(^-\)\(^{12}\) hydrogen bonds and salt bridges\(^13\) as well as weaker\(^14\)\(^-\)\(^{22}\) hydrogen bonds and dispersive\(^{23}\)\(^-\)\(^{27}\) interactions. The greater protein environment can shape the electric field of the active site to influence chemical bond formation\(^{28}\)\(^-\)\(^{36}\) as well as tune noncovalent interactions\(^{37}\)\(^-\)\(^{40}\) critical for catalytic action. As these inherently QM interactions transiently form and dissipate, proteins dynamically change their shape, e.g., in response to the presence of substrates, inhibitors, or solvent.\(^41\)\(^-\)\(^{48}\) The fastest timescales of the reorganization of the protein’s electronic structure cannot be readily resolved by most experimental techniques (e.g., NMR\(^49\)).

Computational, atomistic modeling provides essential insight into the dynamics\(^{41}\)\(^-\)\(^{42},\)\(^{50}\)\(^-\)\(^{54}\) and non-covalent interactions\(^{48}\)\(^,\)\(^{55}\)\(^-\)\(^{57}\) of proteins. Given the large size of proteins and timescale of rare, transient dynamical events, classical molecular mechanics (MM) force fields with fixed point charges are most frequently employed.\(^58\) While parameterization against QM or experiment has improved the fidelity of MM force fields, charge transfer and bond rearrangement cannot be faithfully modeled at the MM level. As an alternative, multi-scale QM/MM modeling\(^{59}\)\(^-\)\(^{70}\) can be fruitfully applied when one knows \textit{a priori} which portion of the protein or enzyme must be treated quantum mechanically. Unfortunately, QM/MM predictions can be strongly sensitive to QM region choice and averaging protocol\(^{71}\)\(^-\)\(^{76}\), boundary treatment\(^{65},\)\(^ {77}\)\(^-\)\(^{87}\), and embedding
method\textsuperscript{80, 88-95}. Recent advances\textsuperscript{96-102} in hardware and algorithms have made large-scale QM treatments (e.g., with hybrid density functional theory) tractable for the study of proteins\textsuperscript{96, 103}. This has motivated increasingly large-scale QM region treatments in QM/MM models of enzyme catalysis\textsuperscript{35, 104-119}, which have revealed unexpectedly large dependence of properties such as the favorability of proton or charge transfer\textsuperscript{106}, electric fields\textsuperscript{35, 75}, excitation energies\textsuperscript{114-115, 120}, bond critical points\textsuperscript{117} and partial charges\textsuperscript{116} on the selection of the QM region. These observations have motivated renewed interest in systematic methods for atom-economical QM region selection\textsuperscript{76, 121-123} for QM/MM properties obtained from single point energies and optimizations, but the application of these methods is still in its infancy in dynamics simulation\textsuperscript{124}. Recently, we carried out\textsuperscript{124} large-scale free energy simulations with ca. 500 atoms treated at the QM level with range-separated hybrid DFT and showed that catalysis-facilitating charge transfer at the active site was influenced by fluctuations in charge distributions of residues distant from the active site.\textsuperscript{124-125}

Proteins are not just flexible but undergo concerted changes in shape, meaning that the motions of residues (e.g., changes in positions of C\textalpha{} atoms or dihedral angles) are coupled. Analysis of geometric coupling has been extensively applied to understand this conformational allostery in proteins.\textsuperscript{126-128} Given that the interactions that govern dynamic protein structure and function are inherently quantum mechanical, an open question is the extent to which the QM charge distribution among protein residues varies dynamically, in close analogy to more well-understood dynamics of the classical nuclei in proteins. The same techniques that have provided valuable insight into concerted geometric motions in proteins, i.e., the linear cross-correlation and mutual information, may help to describe the length-scale and nature of electronic coupling in proteins. Although some analysis of electronic properties has been leveraged to understand
dynamic events in materials\textsuperscript{129-131}, interpret QM/MM simulations\textsuperscript{76, 121, 125}, or to guide QM method selection\textsuperscript{132}, it has not been applied to the charge coupling obtained from \textit{ab initio} molecular dynamics (AIMD) of entire proteins. While some QM effects can be incorporated using recent developments in polarizable force field modeling\textsuperscript{80, 88-95}, charge transfer and dynamical formation of charge-assisted hydrogen bonds remain challenging to describe. As small proteins have begun to be studied with a full QM treatment,\textsuperscript{96, 103} simulations have revealed the importance of first principles to accurately describe unexpected structures\textsuperscript{96} and to explain charge transfer\textsuperscript{124} and polarization in water\textsuperscript{99}. Therefore full AIMD simulation of proteins is expected to be important to accurately quantify QM charge-coupling dynamics. For example, when increasingly large QM regions were employed in QM/MM free energy simulations of enzyme catalysis, distinct nuclear and charge dynamics were observed in comparison to small QM regions.\textsuperscript{124} Fully QM modeling of peptides will be essential to rule out a potential role the boundary or embedding method could have played in this observation.

In this work, we turn our focus to the study of peptides for which we can sample the dynamical fluctuations in both their geometry and electronic structure with a fully QM description. Here, we focus on small peptides both with representative secondary structure motifs of larger globular proteins as well as less common structural elements. From the AIMD study of three proteins, we show that the charge distributions sampled during dynamics are broad, and that this breadth is associated with significant pairwise coupling of the charges between residues that are often distant in both space and sequence. Through these qualitative observations and quantification of the strength of these couplings, we present analysis aimed at understanding the potential role of QM charge coupling in protein structure and function.
2. Results and Discussion.

We curated small (ca. 20 residue) peptides that are large enough to possess characteristics of globular proteins (e.g., diverse secondary structural motifs) but small enough to ensure efficient sampling with hybrid DFT on the 100-ps timescale (see Sec. 4). By studying multiple small proteins with distinct secondary structural motifs, we aimed to ascertain the generality of observations of the coupling lengthscales for QM properties (i.e., partial charges) across diverse peptides. We used distinct search criteria to curate three peptides with available solution NMR structures (i.e., for correspondence between the experimental conditions and solvated protein simulation) from the protein data bank (PDB)\textsuperscript{133}.

First, we identified a peptide with highly stable secondary structure reinforced by disulfide bonds. A search for peptides with 20 to 30 residues, 20–50\% $\alpha$-helix and $\beta$-sheet content, and one to three disulfide bonds yielded 13 unique results (ESI Table S1). We selected the 27-residue mini-CD4, an engineered peptide relevant to HIV treatment\textsuperscript{134}, which consists of an N-terminus $\alpha$-helix (residues 1–12) and C-terminus $\beta$-sheet (residues 17–27) connected by a flexible loop (residues 13–16) and held together by three disulfide bonds (Figure 1 and ESI Tables S1–S2 and Figure S1).

![Figure 1](image)

**Figure 1.** Cartoon structures (in white) for mini-CD4 (left, PDB ID: 1D5Q\textsuperscript{134}), benenodin-1 (middle, PDB ID: 6B5W\textsuperscript{135}), and Trp-cage (right, PDB ID: 1L2Y\textsuperscript{136}), as obtained from their solution NMR structures. Representative polar and charged residue sidechains are labeled with
their residue number and three-letter code and shown in stick structures with nitrogen in blue, oxygen in red, sulfur in yellow, and hydrogen in white. The Gly27 residue of mini-CD4 is the C-terminal residue and contains the negatively charged carboxylate group.

Next, we searched for disordered peptides that lacked conventional secondary structure motifs, in particular lasso peptides\textsuperscript{137} that have a knotted structure, with a typical length of 10 to 20 residues. From 13 candidate lasso peptide structures in the PDB, we selected the 19-residue benenodin-1\textsuperscript{135}, which is a naturally occurring\textsuperscript{135} thermally activated rotaxane switch that we study in its lower-energy conformer (Figure 1 and ESI Tables S3–S4 and Figure S1). The lasso structure contains a ring (residues 1–8) that is closed by the isopeptide bond between the N-terminus of Gly1 and the sidechain of Asp8 residue, which makes both residues effectively neutral, through which a tail (residues 9–19) is threaded (Figure 1 and ESI Figure S1).

Finally, we selected the solution NMR structure of the 20-residue Trp-cage\textsuperscript{136}, a representative designed peptide that has been widely used\textsuperscript{138-140} as a model to study protein folding (Figure 1 and ESI Table S5). Trp-cage contains an α-helix (residues 1–8) much like mini-CD4 along with a hydrophobic core of residues in turn (residues 9–10) and a 3\textsubscript{10} helix (residues 11–14) centered around Trp6 along with a proline-rich tail (residues 15-20; Figure 1 and ESI Table S6). Unlike mini-CD4, the Trp-cage fold is stabilized only by non-covalent, hydrophobic interactions (Figure 1).

These three diverse small model proteins provide a platform for evaluating residue-specific and secondary-structure-specific trends in the coupling of electronic (i.e., partial charge) properties with sufficient sampling from fully \textit{ab initio} molecular dynamics (see Sec. 4).

\textbf{2a. Residue Charge Distributions.}

To quantify how electronic structure properties fluctuate during the AIMD trajectories, we computed the net partial charge sum on each residue, $q(\text{RES})$: 
by summing the Mulliken partial charges, \( q_i \), of all backbone and sidechain atoms within each amino acid residue, as in prior work\(^{35, 76, 124-125} \). Taking this sum over the entire residue minimizes sensitivity to partial charge scheme, yielding comparable results on test systems with alternative real space\(^{141-143} \) partitioning schemes (ESI Table S7). We calculate these \( q(\text{RES}) \) values to quantify the flexibility of the charge distribution, and we estimate the relative deviation of \( q(\text{RES}) \) from expected residue formal charges to quantify charge donation or accumulation (ESI Tables S8–S10). Summing instead over only sidechain atoms would yield qualitatively similar conclusions but at the cost of making it more challenging to identify if charge transfer is inter-residue (ESI Table S11 and Figure S2).

Overall, the by-residue charges of residues vary significantly during the simulation for all amino acids in the three proteins. As expected, nonpolar residues have the narrowest \( q(\text{RES}) \) distributions, and they are the only residue class with consistently normally distributed \( q(\text{RES}) \) distributions (Figure 2 and ESI Figures S3–S5). Most nonpolar distributions are comparably narrow, with the exception of specific cases that are likely driven more by residue context than sidechain identity. For example, Leu15 in the loop of mini-CD4 between the \( \alpha \)-helix and \( \beta \)-sheet has a significantly larger range (ca. 0.3 a.u.) than a Leu3 (range: 0.2 a.u.) in the \( \alpha \)-helix (ESI Figure S3 and Table S8).
Figure 2. Normalized charge distributions of the by-residue summed partial charges for representative amino acids: nonpolar Gly14 in mini-CD4 (top, left), polar Gln15 in benenodin-1 (top, right), negatively charged Asp9 in Trp-cage (bottom, left), and positively charged Lys16 in mini-CD4 (bottom, right). Dashed lines are shown for the expected formal charge of each residue along with a shaded gray region to indicate ± 0.05 a.u. around that value.

While the distribution widths are generally comparably narrow across nonpolar residues, distribution means can differ significantly from an expected neutral value, leading to time-averaged charges that vary within each amino acid identity (Figure 2 and ESI Figures S3–S5). Surprisingly, even the smallest Gly residues alternatively accumulate a net charge (e.g., Gly14 in mini-CD4, Gly5 in benenodin-1, or Gly15 in Trp-cage) or donate charge to the surrounding protein (e.g., Gly18 in mini-CD4 or Gly3 in benenodin-1, or Gly11 in Trp-cage), a behavior which would be challenging to capture with a fixed charge force field (Figure 2 and ESI Figures S3–S5). These differences are observed in even relatively proximal residues that share the same secondary structure unit (e.g., Gly3 and Gly5 are both in the lasso ring of benenodin-1), highlighting the importance of evaluating residue couplings (see Secs. 2b–2c) even for nonpolar residues. We generally observe both mean charge donation (e.g., Ile4 in Trp-cage or Leu13 in
mini-CD4) and accumulation (e.g., Ile10 in benenodin-1 or Leu3 in mini-CD4) for the amino acids for which we have several examples (ESI Figures S3–S5). Overall, slightly more charge transfer away from nonpolar residues is observed than charge accumulation, and residue-specific values appear largely insensitive to the nonpolar amino acid identity (ESI Figures S3–S5 and Tables S8–S10).

In comparison to nonpolar residues, polar residues are capable of forming directional hydrogen bonds, which we would expect to influence the QM charge distribution. Indeed some polar residues such as Gln15 in the benenodin-1 lasso tail sample fully bimodal distributions with two fully resolved peaks, one peak corresponding to case that accumulates charge and one that donates charge to the surroundings (Figure 2 and ESI Figures S6–S8). For all three proteins, the Gln residues (e.g., Gln7 or Gln20 in mini-CD4, Gln13 or Gln15 in benenodin-1, and Gln5 in Trp-cage) have the broadest, most clearly bimodal distributions for polar residues, whether in an α-helix in mini-CD4 or Trp-cage or the disordered loop in benenodin-1 (ESI Figures S1 and S6–S8). For hydroxyl-containing residues (e.g., Ser, Thr, or Tyr), the charge distributions are only slightly wider than those of the nonpolar residues, with select cases having asymmetric distributions with wider tails (e.g., Thr25 in the mini-CD4 tail or Thr12 in the benenodin-1 lasso tail) especially when in disordered secondary structure motifs (ESI Figures S1 and S6–S8 and Tables S8–S10). The hydroxyl-containing residues accumulate charge (e.g., Ser12 in mini-CD4) and donate charge (e.g., Tyr3 in Trp-cage) to comparable amounts, as was observed for nonpolar residues, in a manner that is likely governed by the residue context (ESI Figures S6 and S8).

When analyzing polar residues, we also include a number of special cases in the three proteins: i) Gly1 and Asp8 that form an isopeptide bond in benenodin-1, ii) prolines in both Trp-cage and benenodin-1, and iii) the six Cys that form disulfide bridges in mini-CD4. For the first
two categories of residues, geometric constraints due to covalent bonding in these residues appear correlated with narrow charge distributions comparable to those observed in nonpolar residues (ESI Figures S6 and S8). Proline is often categorized as a nonpolar residue but contains a polar amide bond, and we do generally observe it to have a wider charge distribution (e.g., Pro12 in Trp-cage or Pro18 in benenodin-1) than any nonpolar residue in the same protein but significantly narrower than the most variable polar residues (ESI Tables S8 and S10). Similar observations of a relatively narrow charge distribution hold for the Trp6 residue in Trp-cage, which is too bulky to form strong, directional interactions with its environment or move as rapidly as other residues (ESI Figure S8 and Table S10). Thus, transient, variable directional interactions (e.g., in Gln) are likely to produce residue-specific charge distributions and couplings that are most sensitive to local environments (see Secs. 2b-2c), but all polar and special residues exhibit significantly more variable charge distributions unless motion is constrained.

In comparison to neutral amino acids, we may expect positively or negatively charged residues to have the strongest sensitivity to through-space interactions and, thus, the broadest charge distributions. Indeed, significant charge transfer means that these residues seldom sample within 0.05 a.u. of their formal charges and very broad charge distributions are observed for representative positively charged (e.g., Lys16 in mini-CD4) and negatively charged (e.g., Asp9 in Trp-cage) residues (Figure 2 and ESI Figures S9–S11). Carboxylate-containing terminal residues or sidechains (e.g., Asp9 in Trp-cage or Glu14 in benenodin-1) tend to have a very broad, symmetric distribution with a mean charge transfer to the environment of at least 0.1 a.u., i.e., larger than the neutral residues (ESI Tables S8–S10).

Even after accounting for charged terminal residues, the total number of charged residues
across the three peptides is smaller (9 positive and 5 negative) than for polar (29) or nonpolar (23) residues, making it difficult to identify which trends are general for this class of residues (Figure 2 and ESI Figures S9–S11). Nevertheless, all three proteins have at least one Lys and one Arg that can be compared. For both residues, a bimodal distribution is generally present, with Arg always exhibiting charge transfer and having a small non-dominant second peak close to its expected formal charge (ESI Figures S9–S11). Lys behaves somewhat differently, with the relative heights of the asymmetric, bimodal distribution depending on the residue context: Lys16 in mini-CD4 and Lys17 in benenodin-1 have a higher peak around 0.85 a.u., whereas Lys11 in mini-CD4 favors the peak closer to the expected formal charge of 1.0 a.u. (ESI Figures S9–S10). Charge distributions of both Lys8 and Arg16 are least broad in Trp-cage, potentially due to less sampling time, but its N-terminal Asn1 exhibits as broad a distribution as the Cys1 terminus of mini-CD4 (ESI Figures S9–S11). Overall, it is evident that both charged and neutral residues exhibit significant variation in their charges during ab initio MD. Having recognized the extent of variation of the charges of individual residues, we next sought to explain the length-scales and mechanisms of charge accumulation or depletion by considering pairwise couplings of residue charges.

2b. Linear coupling of residue charge distributions.

To quantify the coupling of electronic properties between the residues of the protein, we computed the cross-correlation (CC)\textsuperscript{144-145} between the by-residue summed partial charges, $q(J)$ and $q(K)$, of residues $J$ and $K$ as:

$$\rho_{JK} = \frac{\sigma_{JK}}{\sigma_J \sigma_K}$$

where $\sigma_{JK}$ is the covariance between $q(J)$ and $q(K)$ and $\sigma_J$ or $\sigma_K$ are the standard deviations of
the individual charge distributions. The CC captures the linear dependence of charges between residue pairs. A high, negative CC likely suggests charge transfer between two residues, whereas a high positive value suggests both accumulate or lose charge in a coupled, albeit less physically intuitive manner.

For each of the three proteins, a range of both positive and negative CC values with magnitudes up to 0.4–0.8 are observed between all types of residues (Figure 3). There are slightly more (ca. 58–65%) negative CCs that are indicative of charge transfer than positive CC values, but the values for residue pairs with negative CCs are significantly larger in magnitude (i.e., few positive CCs exceed 0.2, ESI Tables S12–S14). For all three proteins, many of the strong (i.e., > |0.3|) negative CCs are between nearest-neighbor residues that are connected via the amide backbone (Figure 3 and ESI Tables S12–S14).

Overall, at least half of residues in all three proteins demonstrate the largest absolute CC with a nearest neighbor, with this effect most pronounced in Trp-cage where 95% of the strongest CCs are among nearest neighbors (Figure 3). The highly local coupling in Trp-cage may be due to the distinct methodology and shorter timescale over which it was simulated (see Sec. 4). Nevertheless, in both mini-CD4 and benenodin-1, numerous non-nearest-neighbor couplings are among the strongest including several examples where the highest-magnitude CCs are with more sequence-distant residue partners (e.g., Arg5–Ser9 in mini-CD4 and Arg6–Gln13 in benenodin-1, Figure 3 and ESI Tables S12–S13).
Figure 3. Matrix of signed cross-correlation (CC) values ranging from -0.60 (red) to +0.60 (blue) colored as in inset colorbar. Select matrix elements exceeding the range are capped to the extrema of the range. All residues are indicated by their single-letter code and number. The single strongest coupling for a given residue is indicated by a circle (black unless white is needed for contrast).

The cases with sequence-distant, strongest CCs appear dependent on the sidechain and character of the residue. Breaking down CCs by interactions between residue types, we observe that a greater percentage of strong CCs (i.e., > |0.3|) occur for charged–charged interactions (i.e., 10–30% of all pairs of that type in mini-CD4 or benenodin-1, ESI Tables S12–S14). As expected, no nonpolar–nonpolar residue interactions have very strong CCs in these proteins, but the presence of a charged residue in a charged–nonpolar interaction is sufficient to induce strong (i.e., > |0.3|) negative and positive couplings in both mini-CD4 and Trp-cage (ESI Tables S12–S14). The polar residues reside between these two limits, with these residues forming some strong CCs with residues of all types (ESI Tables S12–S14). Average and maximum values of CC magnitudes are not strongly sensitive to residue type, but they are, as expected, higher for charged and polar residue interactions than for those involving nonpolar residues (ESI Table S15). For the special case of the six Cys residues involved in stabilizing disulfide bonds in mini-CD4, we observe even lower average CC magnitudes than those for nonpolar residues, consistent
with earlier observations\textsuperscript{124-125} that sidechains that form strong bonds exhibit reduced cross-correlations (ESI Tables S15–S16).

Although small in number (ca. 10–16 or less than 10% overall), strong (i.e., > |0.3|) CC values are present in all three of the proteins. The free N-terminal (Cys1 with Gly21/Ser22/Phe23 in mini-CD4 or Asn1 with Asp9 in Trp-cage) or C-terminal (e.g., Gly27 with Leu15/Lys16 in mini-CD4, Met19 with Lys17 in benenodin-1) residues occur frequently in these top couplings (ESI Tables S17–S19). This high representation of terminal residues interacting especially with non-nearest-neighbor, charged residues is likely due to the charged terminus being positioned on a highly flexible portion of the protein. In addition to interactions with the terminal residues, the strongest non-nearest-neighbor couplings involve all residue types. These strong couplings include expected charged–polar or charged–charged interactions in the mini-CD4 α-helix (Arg5, Ser9, and Lys11, |0.31–0.36|) as well as between the lasso ring and tail of benenodin-1 (Arg6, Gln13, and Lys17, |0.35–0.41|, ESI Tables S17–S18). However, strong couplings are also apparent for polar–polar cases in the benenodin-1 lasso tail (Gln13–Gln15, |0.46|) or polar–nonpolar between α- and 3\textsubscript{10}-helices in Trp-cage (i.e., Gln5–Gly11, |0.31|) or between the α-helix and β-sheet of mini-CD4 Gln7–Gly17 (ESI Tables S17–S19). Little can thus be concluded about the role of secondary structure except when strong couplings are due to the secondary structure bringing them into close spatial proximity (i.e., the aligned Arg5, Ser9, and Lys11 in the α-helical turns of mini-CD4). Thus, if through-space interactions are important for the formation of strong coupling, residue charge and sidechain chemistry should be key.

Focusing on sidechain chemistry, we now compare whether trends that were evident in charge distributions also give rise to distinct couplings. We observed (see Sec. 2a) that polar Gln residues had broad bimodal charge distributions in comparison to the distributions for polar Ser
or Thr sidechains. Indeed, Gln7 in mini-CD4 has a strong CC with both Gly17 and Lys11 as well as a moderate CC (i.e., > |0.2|) with two additional (i.e., Cys6 and Cys1) residues (ESI Tables S13 and S17). The remaining Gln residues (e.g., Gln20 in mini-CD4, Gln13 in benenodin-1, and Gln5 in Trp-cage) behave similarly, forming moderate to strong coupling with a greater number of residues in comparison to other polar residues (i.e., Ser or Thr) in the same protein (ESI Table S20). We also previously noted distinct charge accumulation or depletion for specific residues, which were especially evident and surprising for the case of nonpolar Gly residues. Some Gly residues form unexpectedly strong couplings especially with Gln residues (e.g., Gln5–Gly11 in Trp-cage or Gln7–Gly17 in mini-CD4, Figure 3). However, the relationship between strong couplings and accumulation or charge loss is generally not obvious except in specific cases (e.g., Gly14 to C-terminal Gly27 in mini-CD4) that are strongly interacting with negatively charged residues (Figure 3).

Charged residues, which have the broadest distributions, may be expected to have strong couplings to a range of residues. All three proteins possess Lys and Arg residues, and, indeed, the two charged residues participate in a disproportionate number of the strong coupling cases for all three proteins (Figure 3 and ESI Tables S17–S19). Nevertheless, the Lys16 in mini-CD4 disproportionately couples only to Gly27 in a salt bridge, giving rise to one very strong (-0.60) CC value, whereas Lys11 forms strong couplings with four residues (i.e., Arg5, Gln7, Ser9, and Cys10, Figure 3 and ESI Table S17). In benenodin-1, a similar trend is observed where Lys17 forms its dominant strong CC to the carboxylate of the C-terminal Met19, whereas the more mobile Arg6 in the benenodin-1 ring forms strong CCs with Gln13, Lys17, and Pro18 (Figure 3 and ESI Table S18).

In most cases, charged and polar bulky residues have both the broadest, multi-peaked
distributions and greatest number of linear correlations with other protein residues, but exceptions are also apparent. Gln15 in benenodin-1 exhibits fewer moderately strong CCs with non-nearest-neighbor residues than Gln residues in the other three proteins (ESI Table S20). At first glance, this suggests that Gln15 behaves distinctly from other Gln residues, however examination of the joint $q$(RES) charge distributions highlights the limitations of linear CC evaluation (Figure 4). For the residues with normally distributed $q$(RES), the linear CC distinguishes when two residue charge distributions (e.g., Ser9–Ile10 in benenodin-1) are correlated and when they are uncorrelated (e.g., Phe4–Ile10 in benenodin-1, Figure 4). However, for the broader charge distributions, e.g., of the Gln residues, the presence of multiple peaks can complicate the use of a linear CC (Figure 4). While Arg6 appears to be more strongly correlated with Gln13 ($r = -0.41$) than Gln15 ($r = 0.18$) in benenodin-1, structure is apparent in the joint distribution between both sets of residue pairs (Figure 4). These observations motivate consideration of how the coupling of charge distribution probabilities can be quantified beyond the linear relationships captured by CCs.
Figure 4. Example joint distributions of $q$(RES) (in a.u.) for four pairs of residues in benenodin-1 with CC values shown in upper right insets: (top, left) Phe4–Ile10, (top, right) Ser9–Ile10, (bottom, left) Arg6–Gln13, and (bottom, right) Arg6–Gln15. The same color scale is used for the normalized histograms in all cases, with yellow indicating high density and purple indicating none. The same range is used for all axes, with the positively charged Arg shifted with respect to the neutral residues.

2c. Beyond linear couplings with mutual information.

Inspired by the use of information theoretic tools to understand coupled conformational dynamics of protein residues\textsuperscript{126-128}, we computed the mutual information (MI)\textsuperscript{126, 128, 146} to identify interactions between residue charge distributions not captured by a linear CC. The MI between the probability distributions, $p$, of $q(J)$ and $q(K)$ for residues $J$ and $K$ is computed as:

$$I(J;K) = \sum_j \sum_k p_{(J,K)}(j,k) \ln \left( \frac{p_{(J,K)}(j,k)}{p_J(j)p_K(k)} \right)$$

(0)

Here, $p_{(J,K)}(j,k)$ is the joint probability of the charge distributions, and $p_J(j)$ or $p_K(k)$ refer to the marginal probability distributions (see Sec. 4). To characterize the importance of nonlinear MI, we primarily compare the relative rank of MI and CC values for residue pairs, and we also estimate a linear component of the MI\textsuperscript{147-148} derived from the CC (ESI Text S1).\textsuperscript{148} In charge couplings, we expect the nonlinear MI perspective to be most important between the pairs of residues for which we have observed broader, multi-modal $q$(RES) distributions because the linear CC-derived term should be the sole component of the MI in the normal distribution limit\textsuperscript{147-148}.

Global trends are qualitatively consistent between MI and CC values for residue couplings, with the largest MI pairs also having high CC values (Figure 5 and ESI Figure S12). Hotspots in the CC matrix (e.g., Gln13–Gln15 in benenodin-1 or Ser9–Lys11 in mini-CD4) are confirmed in the MI matrix (Figures 3 and 5). Despite qualitative agreement, quantitative estimations of relative coupling strength differ between MI and the CC magnitudes (i.e., $|CC|$,
For each of the three proteins, both the Pearson’s $r$ and the Spearman’s rank correlation coefficient (SRCC) between the coupling strengths from the MI and |CC| are moderate (SRCC: 0.70–0.77 $r$: 0.8–0.88, ESI Figure S13). Consistent with this analysis, the nonlinear MI contribution is substantial for a large number of residue pairs in all three proteins (ESI Figures S14–S16).

**Figure 5.** Matrix of mutual information (MI) values ranging from 0.0 (white) to +0.25 (dark blue) colored as in inset colorbar. Select matrix elements exceeding the range are capped to the extrema of the range. All residues are indicated by their single-letter code and number. The cases for which the MI coupling percentile rank is $> 25\%$ above the |CC| rank are indicated by green circles in the lower triangle of the matrix, and the cases for which the MI coupling rank is $> 25\%$ below the |CC| rank are indicated by red circles in the upper triangle of the matrix.

For around one quarter of all residue–residue couplings, the percentile rank for the MI differs from that for the |CC| by more than 25%, with a comparable number for either direction (i.e., MI $> |\text{CC}|$ or vice versa, Figure 5 and ESI Figure S13). Generally, the most extreme disagreement in rank is observed for pairs with significant MI that had very low |CC|, whereas disagreements for the reverse are more moderate in nature (ESI Figure S13). Focusing on the pairs of residues that have relatively higher MI or |CC|, however, reveals the role of MI analysis in interpreting charge coupling (Figure 5 and ESI Tables S21–S27). In mini-CD4, the MI is relatively lower than CC disproportionately for nonpolar residues (i.e., Leu3, Ala4, Leu8, Leu13,
Gly17, or Gly18) especially for coupling to sequence-distant residues (Figure 5 and ESI Table S21). Some weakly coupled Cys residues that are constrained by disulfide bonds (i.e., Cys10 and Cys24) also have reduced MI in mini-CD4 (Figure 5 and ESI Table S21). Conversely, the MI is significantly enhanced relative to CC in mini-CD4 for the terminal and charged (e.g., Lys11, Lys16) and polar (e.g., Ser9, Ser12, Gln20, or Thr25) residues (Figure 5 and ESI Table S22). As an example, Lys11–Gly27 exhibits among the strongest MI values in mini-CD4 (0.113) placing it at the 97th percentile, whereas the low CC of this pair (-0.022) would have suggested much weaker coupling (Figure 5 and ESI Table S22).

The pairwise MI of residues in benenodin-1 exhibits similar trends to those observed for mini-CD4. The relative MI values of sequence-distant nonpolar residues are smaller, whereas the apparent coupling of polar (i.e., Gln13, Gln15) and charged (i.e., Arg6, Glu14, Lys17, and Met19) residues with other polar or nonpolar residues is stronger (Figure 5 and ESI Tables S23–S24). For example, Phe4–Ile10, which had a modest CC that placed it in the middle (i.e., 46th percentile) of all |CC| values is instead one of the weakest (i.e., 18th percentile) couplings from the MI perspective (Figures 4–5 and ESI Table S23). Gln15, which in benenodin-1 had been identified as having relatively lower CC strengths than other Gln residues, shows enhanced MI values relative to the CC picture, particularly with the isopeptide-bond-forming Asp8 as well as with Phe4 or Pro18 (Figure 5 and ESI Table S24). While the MI of Arg6 with Gln15 is lower than that with Gln13, the gap is reduced, and both Arg6–Gln13 and Arg–Gln15 are in the top 10% of all MI values for the benenodin-1 protein (Figures 4–5).

While the Trp-cage MI and CC couplings are qualitatively similar to each other, they exhibit residue-type-specific shifts in line with trends observed for mini-CD4 and benenodin-1 (Figures 3 and 5). For Trp-cage, this means an enhancement of MI relative to CC for both
terminal residues (i.e., Asn1 and Ser20) and the charged Arg16 while couplings to both Ile4 and Trp6 are significantly reduced (Figure 5 and ESI Tables S25–S26). Overall, the nearest-neighbor pairs are in good agreement between the linear CC and MI, whereas most differences arise from more distant residues (ESI Table S27). In all cases, the type of residues participating in the interaction appears to have a dominant effect over secondary structure or proximity (ESI Table S27). While for mini-CD4, MI is increased most over CC for intra-β-sheet pairs while it is reduced for β-sheet to α-helix interactions, MI for the α-helix to tail pairs shift both directions for Trp-cage (ESI Tables S21–S22 and S25–S26). Returning to residue type, we note that most cases where MI is enhanced involve at least one charged or polar residue for all proteins, whereas most of the cases where the linear CC is relatively smaller than the MI involve at least one nonpolar residue (ESI Table S27). Nevertheless, only of the two residues in the pair needs to be charged, meaning that significant nonlinear coupling can be observed between charged–nonpolar residue pairs (i.e., 25–33% of the outlier cases for the three proteins, ESI Table S27). The mutual information analysis therefore supports the observations from CC that sequence-distant residues have charge distributions that couple significantly but it captures different classes of interactions that are needed to describe the observed variability of residue charge distributions.

2d. Comparison of geometric length scales for electronic coupling.

Although moderate to high MI and CC has been observed for non-adjacent residue pairs, it may be expected that these couplings decay rapidly with increasing through-space distance. We evaluate residue pair separations by their AIMD-averaged center-of-mass (COM) distance, a quantity closely related to the average COM distance from the NMR ensembles and proportional to the shortest inter-residue distances (ESI Figure S17). For mini-CD4, the highest MI and CCs
are at short COM–COM separations, but MI and CC values of significant magnitude persist for distant residue pairs (Figure 6).

Figure 6. Dependence of MI and CC for mini-CD4 (top) and benenodin-1 (bottom) on the average center-of-mass (COM) distance between residues in a pair ($d$(COM-COM), in Å) during the AIMD simulation. The axis values are the same for both plots. The subset of residue pairs corresponding to charged–charged interactions are shown for the CC subpane in blue as shown in inset legend in the bottom pane. In the top MI subpane, two representative pairs are shown in red and annotated. These same residue pairs are shown schematically as sticks along with the remainder of the proteins in cartoon at right, with a subset of representative structures overlaid from AIMD. Atoms in the sidechains are colored as: blue for nitrogen, red for oxygen, white for carbon, and yellow for sulfur.

Somewhat surprisingly, the coupling of distant residue pairs is not exclusive to charged-charged interactions. For example, a distant Lys11–Gln20 charged–polar MI/CC is higher than that for equivalently distant charged interactions in mini-CD4 (Figure 6). Distinguishing short-range from long-range interactions with a cutoff of 10 Å COM–COM distances (corresponding to ca. 4 Å shortest-atom separations), we observe overall that CCs among charged residues are
roughly equivalent for both short-range and long-range residue pairs, excluding only the most extreme, short-range salt bridges (e.g., Lys16–Gly27, Figure 6 and ESI Figure S17). However, this observation is not specific to charged–charged pairs, as other classes of residue–residue interactions are also equivalently significant at both short- and long-range. Examining the Lys11–Gln20 pair more closely, we observe it samples a wide range of COM–COM distances (ca. 15–18 Å) depending on the orientation of the two sidechains, and the strong coupling of the charge distributions for this pair is indicative of a long-range cooperativity in the protein that is not mediated by any direct hydrogen bonding interaction (ESI Figure S18).

We observe similar behavior among charged–charged residues in the lasso peptide benenodin-1, except the CC and MI appear to exhibit slightly stronger dependence on distance (Figure 6). Despite this, the long-range Arg6–Gln13 coupling in benenodin-1 has an MI that exceeds many residue pairs at a comparable distance, even among the charged residues (Figure 6). This phenomenon suggests long-range cooperativity in the two residues’ charge distributions in a manner similar to the Lys11–Gln20 pair in mini-CD4 (Figure 6). The distance dependence of MI and CC in Trp-cage resides roughly between the other two proteins, with a significant long-range charged–charged interaction between terminal Asn1–Asp9 spanning the α-helix of the peptide but most other strong CCs corresponding to low-separation nearest-neighbor interactions (ESI Figure S19). Thus, the somewhat greater distance dependence of couplings in benenodin-1 may be due to the fact that the lasso structure brings more residues into mid-range proximity (i.e., 5–10 Å COM distance) but in a manner that prevents them from coupling in comparison to either mini-CD4 or Trp-cage (Figure 6 and ESI Figure S19).

While the long-range coupling of electronic properties is somewhat unexpected, long-range geometric couplings are well-established\textsuperscript{126-128} as important for understanding protein
dynamics. If electronic coupling could be inferred from geometric measures alone, one might be able to estimate electronic coupling from lower-cost (i.e., classical or semi-empirical) MD. However, the bulky, nonpolar residues that are frequently observed to display coupled geometric motion would likely show smaller electronic couplings (i.e., with CC or MI), challenging the notion that electronic coupling can be determined solely from geometric motion. Indeed, comparisons of geometric coupling and electronic coupling of residues yield limited correspondence (ESI Text S2 and Figures S20–S22). These studies support earlier observations of long-range coupling in QM/MM simulation of enzyme catalysis\textsuperscript{124-125} and emphasize the importance of continued study of the quantum mechanical mechanisms underlying this phenomenon.

3. Conclusions

We carried out fully \textit{ab initio} molecular dynamics simulation of three representative small proteins to quantify the nature and length-scale of the coupling of electronic properties in proteins. To cover both common protein features representative of larger proteins as well as less common ones, our three proteins included mini-CD4 and Trp-cage as well as a lasso peptide. We focused on the evaluation of charge distributions and their couplings since these are QM properties that are essential to the understanding of protein structure and function but challenging to capture with protein force fields. By analyzing the individual distributions of residue charge, we observed that while some nonpolar residues exhibited narrow charge distributions, most polar and charged residues exhibit very broad, multimodal distributions. Even in cases with narrow charge distributions (e.g., Gly), we noted sequence-specific deviations corresponding to charge accumulation or depletion that would be challenging to capture in a fixed-charge force field. Charged residues (e.g., Lys or Arg) exhibited wide charge distributions indicative of a large
degree of charge transfer with surrounding residues. Most surprisingly, among polar residues, Gln residues in all three proteins displayed broad, multimodal distributions that sampled both positive and negative partial charges.

To quantify residue–residue interactions to explain observed variations in residue charge distributions and to identify interactions that potentially require a full QM treatment, we computed both linear cross-correlations and the mutual information of these charge distributions. From the purely linear CC picture, we observed that a significant number of residues formed the strongest couplings with non-nearest-neighbor residues, especially for mini-CD4 and benenodin-1. In some cases, these strong couplings corresponded to clusters of polar and charged residues. Using mutual information analysis, we observed additional coupling between sequence-distant residues that would have been missed from the linear picture alone. We observed limited through-space-distance-dependence of strong couplings in mini-CD4, and somewhat stronger distance dependence in the constrained lasso peptide of benenodin-1 or Trp-cage. While the expected electrostatically driven, charged–charged CCs were strong and had limited distance dependence in all of the proteins, surprising polar–polar and polar–charged residue couplings were also significant at long-range. Analyzing the robustness and reproducibility of these couplings, both across other proteins and through more extensive independent dynamics, will be important in the future to develop a broad understanding of charge dynamics in proteins. We expect this charge coupling analysis to provide additional insight into the mechanistic role of the enzyme environment in catalysis and to aid assessment of method and embedding sensitivity in multi-scale modeling.

4. Computational Details

*Protein structure preparation and MM MD equilibration.* The representative, first
solution NMR structure for three peptides was obtained for simulation from the protein databank (PDB): the 27-residue globular protein mini-CD4 (PDB ID: 1D5Q), the 19-residue lasso peptide benenodin-1 (PDB ID: 6B5W), and the 20-residue globular protein Trp-cage (PDB ID: 1L2Y). Protonation states were assigned with the H++ webserver assuming a pH of 7.0 and a dielectric constant of 10.0 with all other defaults applied (ESI Tables S2, S4, and S6). Mini-CD4 was simulated with its three disulfide bonds at Cys1–Cys19, Cys6–Cys24, and Cys10–Cys26 intact, and benenodin-1 was simulated with a Gly1–Asp8 isopeptide bond (ESI Tables S2 and S4). The resulting peptide sizes and charges were: 367 atoms and a +3 net charge for mini-CD4, 282 atoms and neutral for benenodin-1, 304 atoms and a +1 net charge and for Trp-cage (ESI Tables S2, S4, and S6). All proteins have charged termini (i.e., C-terminal carboxylate and NH$_3^+$ for the N-terminus) except for the isopeptide-bond-forming N-terminus in benenodin-1.

Structures were prepared using the AMBER tleap utility for classical molecular dynamics (MD) equilibration with the AMBER ff14SB force field. Isopeptide bond parameters in benenodin-1 were obtained from the AMBER99 force field (ESI Table S28). The miniproteins were equilibrated in both explicit TIP3P water and with the implicit generalized Born solvent model with all defaults applied to assess the impact of solvent choice. All proteins were equilibrated using the GPU-accelerated PMEMD AMBER code as follows: i) 3000 minimization steps, ii) 10-ps NVT heating to 300 K with a Langevin thermostat with collision frequency of 1.0 ps$^{-1}$ and a random seed, iii) 250-ps NpT equilibration using the Berendsen barostat with a pressure relaxation time of 2 ps, and iv) a 100-ns NpT production run. The SHAKE algorithm was applied in combination with a 2-fs timestep. For the long-range electrostatics, the particle mesh Ewald method was used with a 10-Å real space
cutoff. The backbone atom root-mean-square deviation (RMSD) with respect to the starting NMR structure was used to validate choice of solvent. Implicit solvent was found to be suitable for mini-CD4 and benenodin-1 but not Trp-cage, which unfolded unless in explicit solvent (ESI Figure S23). All initial MD structures are provided in the ESI .zip file.

*Ab initio Molecular Dynamics (AIMD).* AIMD calculations were initiated from snapshots of the MM MD equilibration spaced 10 ns apart following slightly different protocols for the implicit solvent mini-CD4 and benenodin-1 and the explicitly solvated Trp-cage. All QM calculations were carried out with density functional theory (DFT) using range-separated hybrid functional ωPBeH\(^{161}\) (ω = 0.2 bohr\(^{-1}\)) and the 6-31G\(^{162}\) basis. The AIMD calculations employed a 0.5-fs timestep with a temperature of 300 K using a Langevin thermostat and a collision frequency of 3.3 ps\(^{-1}\). For mini-CD4 and benenodin-1, we carried out AIMD in an implicit conductor-like polarizable continuum (C-PCM) implicit solvation model\(^{163-164}\), as implemented\(^{103, 165}\) in TeraChem\(^97, 166\). These calculations used 1.2x Bondi’s van der Waals radii\(^{167}\) to construct the cavity in conjunction with \(\varepsilon = 80\) to model water. For these two proteins, ten independent 10 ps-AIMD simulations were initiated, and we discarded the first 15% of all AIMD simulations, retaining 85 ps for analysis per protein. This simulation length was validated by comparison of charge distribution properties obtained on shorter trajectories as well as from enhanced sampling\(^{168-169}\) (ESI Figures S24–S27 and Tables S29–S32). Semi-empirical dispersion\(^{170-171}\) was omitted from calculations after it was determined it had limited effect on computed electronic properties (ESI Figure S28).

For explicitly solvated Trp-cage, the TeraChem-AMBER interface\(^{113}\) was used to drive TeraChem for the QM portion and AMBER\(^{152}\) for the MM (i.e., TIP3P water molecules) component with SHAKE applied only to the TIP3P water. We selected 8 snapshots spaced 10 ns
apart from the production explicit solvent classical MD simulations. We used the cpptraj closestwater command to extract a 29-Å radius spherical droplet with 4001 water molecules, neutralize the sphere (i.e., add back the Cl\(^-\) ion where necessary), and define spherical boundary conditions with a 1.5 kcal/mol Å\(^2\) force constant applied (ESI Table S33). After re-equilibration with classical MD for 20 ps, AIMD was carried out at 298 K for 5 ps with a 0.5-fs timestep and a Langevin thermostat with a 1 ps\(^{-1}\) collision frequency. After discarding the first 15% of each trajectory, we obtained 34 ps for analysis. Starting structures for AIMD are provided in the ESI .zip file.

**Partial charges and analysis.** As in prior work\(^ {124} \), Mulliken partial charges were collected at each AIMD step and summed over all atoms, including the backbone atoms. Trends were comparable for sidechain-only sums or alternative partial charge schemes (ESI Tables S7 and S11 and Figure S2). The cross-correlations and mutual information of the charge distributions were evaluated in scikit-learn\(^ {172} \). The scikit-learn\(^ {172} \) estimates of mutual information between two continuous variables (here, charges) use non-parametric methods based on distances between nearest neighbors\(^ {173} \). After trial and error, the number of nearest neighbors was increased from its default (i.e., three) to 10.
ASSOCIATED CONTENT

**Electronic Supplementary Information.** Details of protein structure curation; effect of charge scheme and comparison to sidechain-only convention on charge distributions; overall statistics of charge distributions and residue-specific distributions; overall statistics of CC values in three proteins; specific CC attributes of disulfide Cys residues; list and counts of high CCs for each protein; details and statistics on total and linear contributions to the MI; comparison of MI and CC percentile rank for all couplings in the three proteins; summary of cases where percentile rank disagrees by >25% for MI and |CC|; summary of residues with MI and CC differences by type; analysis of COM-COM distances in NMR and AIMD along with shortest distances; example of distances sampled in AIMD; geometric coupling analysis; force field parameters for the isopeptide bond; RMSD analysis of solvated proteins; evaluation of MI convergence with REMD and with subsampled trajectory lengths as well as with and without D3 correction; and details of spherical droplet construction for Trp-cage. (PDF)

Starting structures for classical MD and AIMD of the three proteins. (ZIP)

AUTHOR INFORMATION

**Corresponding Author**

*email: hjkulik@mit.edu phone: 617-253-4584*

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT
H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, which supported this work. The authors also acknowledge an NEC Corporation Grant from the MIT Research Support Committee (for H.J.K. and Z.Y.). Z.Y. was supported in part by the Center for Enhanced Nanofluidic Transport, an Energy Frontier Research Center funded by the US Department of Energy, Office of Science, Basic Energy Sciences under Award DE-SC0019112. N.H. was supported by a Herchel Smith undergraduate research fellowship. This work was carried out in part using computational resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. This work used the XStream computational resource, supported by the National Science Foundation Major Research Instrumentation program (ACI-1429830).

References

1. Cleland, W. W.; Kreevoy, M. M., Low-Barrier Hydrogen Bonds and Enzymic Catalysis. *Science* **1994**, *264*, 1887-1890.
2. Pauling, L.; Corey, R. B.; Branson, H. R., The Structure of Proteins - 2 Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *Proc. Natl. Acad. Sci. U. S. A.* **1951**, *37*, 205-211.
3. Pauling, L.; Corey, R. B., Configurations of Polypeptide Chains with Favored Orientations around Single Bonds - 2 New Pleated Sheets. *Proc. Natl. Acad. Sci. U. S. A.* **1951**, *37*, 729-740.
4. Grutter, M. G.; Hawkes, R. B.; Matthews, B. W., Molecular-Basis of Thermostability in the Lysozyme from Bacteriophage-T4. *Nature* **1979**, *277*, 667-669.
5. Perutz, M. F.; Raidt, H., Stereochemical Basis of Heat-Stability in Bacterial Ferredoxins and in Hemoglobin-A2. *Nature* **1975**, *255*, 256-259.
6. Desiraju, G. R., A Bond by Any Other Name. *Angew. Chem., Int. Ed.* **2011**, *50*, 52-59.
7. Perrin, C. L.; Nielsen, J. B., “Strong” Hydrogen Bonds in Chemistry and Biology. *Annu. Rev. Phys. Chem.* **1997**, *48*, 511-544.
8. Gilli, P.; Pretto, L.; Bertolasi, V.; Gilli, G., Predicting Hydrogen-Bond Strengths from Acid–Base Molecular Properties. The pKa Slide Rule: Toward the Solution of a Long-Lasting Problem. *Acc. Chem. Res.* **2009**, *42*, 33-44.
9. Gilli, P.; Gilli, G., Hydrogen Bond Models and Theories: The Dual Hydrogen Bond Model and Its Consequences. *J. Mol. Struct.: THEOCHEM* **2010**, *972*, 2-10.
10. Gilli, P.; Pretto, L.; Gilli, G., Pa/pKa Equalization and the Prediction of the Hydrogen-Bond Strength: A Synergism of Classical Thermodynamics and Structural Crystallography. J. Mol. Struct.: THEOCHEM 2007, 844, 328-339.

11. Frey, P.; Whitt, S.; Tobin, J., A Low-Barrier Hydrogen Bond in the Catalytic Triad of Serine Proteases. Science 1994, 264, 1927-1930.

12. Zhou, S. M.; Wang, L., Unraveling the Structural and Chemical Features of Biological Short Hydrogen Bonds. Chem. Sci. 2019, 10, 7734-7745.

13. Kurczab, R.; Sliwa, P.; Rataj, K.; Kafel, R.; Bojarski, A. J., Salt Bridge in Ligand–Protein Complexes—Systematic Theoretical and Statistical Investigations. J. Chem. Inf. Model. 2018, 58, 2224-2238.

14. Yesselman, J. D.; Horowitz, S.; Brooks, C. L.; Trievel, R. C., Frequent Side Chain Methyl Carbon-Oxygen Hydrogen Bonding in Proteins Revealed by Computational and Stereochemical Analysis of Neutron Structures. Proteins: Struct., Funct., Bioinf. 2014, 83, 403-410.

15. Steiner, T.; Saenger, W., The Ordered Water Cluster in Vitamin-B-12 Coenzyme at 15 K Is Stabilized by C-H···O Hydrogen-Bonds. Acta Crystallogr., Sect. D: Biol. Crystallogr. 1993, 49, 592-593.

16. Steiner, T.; Saenger, W., Role of C-H···O Hydrogen-Bonds in the Coordination of Water-Molecules - Analysis of Neutron-Diffraction Data. J. Am. Chem. Soc. 1993, 115, 4540-4547.

17. Derewenda, Z. S.; Derewenda, U.; Kobos, P. M., (His)C-Epsilon-H···O=C Hydrogen-Bond in the Active-Sites of Serine Hydroases. J. Mol. Biol. 1994, 241, 83-93.

18. Derewenda, Z. S.; Lee, L.; Derewenda, U., The Occurrence of C-H···O Hydrogen-Bonds in Proteins. J. Mol. Biol. 1995, 252, 248-262.

19. Iyer, A. H.; Krishna Deepak, R. N. V.; Sankararamakrishnan, R., Imidazole Nitrogens of Two Histidine Residues Participating in N-H···N Hydrogen Bonds in Protein Structures: Structural Bioinformatics Approach Combined with Quantum Chemical Calculations. J. Phys. Chem. B 2018, 122, 1205-1212.

20. Holcomb, M.; Adhikary, R.; Zimmermann, J.; Romesberg, F. E., Topological Evidence of Previously Overlooked Ni+1–H···Ni H-Bonds and Their Contribution to Protein Structure and Stability. J. Phys. Chem. A 2018, 122, 446-450.

21. Deepak, R. N. V. K.; Sankararamakrishnan, R., Unconventional N-H···N Hydrogen Bonds Involving Proline Backbone Nitrogen in Protein Structures. Biophys. J. 2016, 110, 1967-1979.

22. Luisi, B.; Orozco, M.; Sponer, J.; Luque, F. J.; Shakked, Z., On the Potential Role of the Amino Nitrogen Atom as a Hydrogen Bond Acceptor in Macromolecules. J. Mol. Biol. 1998, 279, 1123-1136.

23. Nishio, M.; Umezawa, Y.; Fantini, J.; Weiss, M. S.; Chakrabarti, P., CH − π Hydrogen Bonds in Biological Macromolecules. Phys. Chem. Chem. Phys. 2014, 16, 12648-12683.

24. Steiner, T.; Koellner, G., Hydrogen Bonds with π-Acceptors in Proteins: Frequencies and Role in Stabilizing Local 3D Structures. J. Mol. Biol. 2001, 305, 535-557.

25. Newberry, R. W.; Raines, R. T., The N→π * Interaction. Acc. Chem. Res. 2017, 50, 1838-1846.

26. Bartlett, G. J.; Woolfson, D. N., On the Satisfaction of Backbone-Carbonyl Lone Pairs of Electrons in Protein Structures. Protein Sci. 2016, 25, 887-897.
27. Bartlett, G. J.; Newberry, R. W.; VanVeller, B.; Raines, R. T.; Woolfson, D. N., Interplay of Hydrogen Bonds and N→π* Interactions in Proteins. *J. Am. Chem. Soc.* **2013**, 135, 18682-18688.
28. Vaissier, V.; Sharma, S. C.; Schaettle, K.; Zhang, T.; Head-Gordon, T., Computational Optimization of Electric Fields for Improving Catalysis of a Designed Kemp Eliminase. *ACS Catal.* **2017**, 8, 219-227.
29. Bhowmick, A.; Sharma, S. C.; Head-Gordon, T., The Importance of the Scaffold for De Novo Enzymes: A Case Study with Kemp Eliminase. *J. Am. Chem. Soc.* **2017**, 139, 18682-18688.
30. Suydam, I. T.; Snow, C. D.; Pande, V. S.; Boxer, S. G., Electric Fields at the Active Site of an Enzyme: Direct Comparison of Experiment with Theory. *Science* **2006**, 313, 200-204.
31. Lockhart, D. J.; Boxer, S. G., Electric Field Modulation of the Fluorescence from Rhodobacter Sphaeroides Reaction Centers. *Chem. Phys. Lett.* **1988**, 144, 243-250.
32. Fafarman, A. T.; Sigala, P. A.; Schwans, J. P.; Fenn, T. D.; Herschlag, D.; Boxer, S. G., Quantitative, Directional Measurement of Electric Field Heterogeneity in the Active Site of Ketosteroid Isomerase. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, 109, E299-E308.
33. Welborn, V. V.; Head-Gordon, T., Fluctuations of Electric Fields in the Active Site of the Enzyme Ketosteroid Isomerase. *J. Am. Chem. Soc.* **2019**, 141, 12487-12492.
34. Welborn, V. V.; Pestana, L. R.; Head-Gordon, T., Computational Optimization of Electric Fields for Better Catalysis Design. *Nat. Catal.* **2018**, 1, 649-655.
35. Yang, Z.; Liu, F.; Steeves, A. H.; Kulik, H. J., Quantum Mechanical Description of Electrostatics Provides a Unified Picture of Catalytic Action across Methyltransferases. *J. Phys. Chem. Lett.* **2019**, 10, 3779-3787.
36. Zoi, I.; Antoniou, D.; Schwartz, S. D., Electric Fields and Fast Protein Dynamics in Enzymes. *J. Phys. Chem. Lett.* **2017**, 8, 6165-6170.
37. Genna, V.; Marcia, M.; De Vivo, M., A Transient and Flexible Cation−π Interaction Promotes Hydrolysis of Nucleic Acids in DNA and RNA Nucleases. *J. Am. Chem. Soc.* **2019**, 141, 10770-10776.
38. Bootsma, A. N.; Doney, A. C.; Wheeler, S. E., Predicting the Strength of Stacking Interactions between Heterocycles and Aromatic Amino Acid Side Chains. *J. Am. Chem. Soc.* **2019**, 141, 11027-11035.
39. Zhang, J.; Kulik, H. J.; Martinez, T. J.; Klinman, J. P., Mediation of Donor–Acceptor Distance in an Enzymatic Methyl Transfer Reaction. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, 112, 7954-7959.
40. Mehmod, R.; Qi, H. W.; Steeves, A. H.; Kulik, H. J., The Protein’s Role in Substrate Positioning and Reactivity for Biosynthetic Enzyme Complexes: The Case of SyrB2/SyrB1. *ACS Catal.* **2019**, 9, 4930-4943.
41. Blaha-Nelson, D.; Krüger, D. M.; Szeler, K.; Ben-David, M.; Kamerlin, S. C. L., Active Site Hydrophobicity and the Convergent Evolution of Paraoxonase Activity in Structurally Divergent Enzymes: The Case of Serum Paraoxonase 1. *J. Am. Chem. Soc.* **2017**, 139, 1155-1167.
42. Crean, R. M.; Gardner, J. M.; Kamerlin, S. C. L., Harnessing Conformational Plasticity to Generate Designer Enzymes. *J. Am. Chem. Soc.* **2020**.
43. Verkhivker, G. M.; Agajanian, S.; Hu, G.; Tao, P., Allosteric Regulation at the Crossroads of New Technologies: Multiscale Modeling, Networks, and Machine Learning. *Front. Mol. Biosci.* **2020**, 7.
44. Kimura, S. R.; Hu, H. P.; Ruvinsky, A. M.; Sherman, W.; Favia, A. D., Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. *J. Chem. Inf. Model.* **2017**, *57*, 1388-1401.

45. Cournia, Z.; Allen, B. K.; Beuming, T.; Pearlman, D. A.; Radak, B. K.; Sherman, W., Rigorous Free Energy Simulations in Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60*, 4153-4169.

46. Nadig, G.; Van Zant, L. C.; Dixon, S. L.; Merz, K. M., Charge-Transfer Interactions in Macromolecular Systems: A New View of the Protein/Water Interface. *J. Am. Chem. Soc.* **1998**, *120*, 5593-5594.

47. Son, C. Y.; Yethiraj, A.; Cui, Q., Cavity Hydration Dynamics in Cytochrome C Oxidase and Functional Implications. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E8830-E8836.

48. Wang, L.; Fried, S. D.; Boxer, S. G.; Markland, T. E., Quantum Delocalization of Protons in the Hydrogen-Bond Network of an Enzyme Active Site. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 18454-18459.

49. Jarymowycz, V. A.; Stone, M. J., Fast Time Scale Dynamics of Protein Backbones: NMR Relaxation Methods, Applications, and Functional Consequences. *Chem. Rev.* **2006**, *106*, 1624-1671.

50. Lau, E. Y.; Bruice, T. C., Importance of Correlated Motions in Forming Highly Reactive near Attack Conformations in Catechol O-Methyltransferase. *J. Am. Chem. Soc.* **1998**, *120*, 12387-12394.

51. Horowitz, S.; Dirk, L. M. A.; Yesselman, J. D.; Nimtz, J. S.; Adhikari, U.; Mehl, R. A.; Scheiner, S.; Houtz, R. L.; Al-Hashimi, H. M.; Trievel, R. C., Conservation and Functional Importance of Carbon–Oxygen Hydrogen Bonding in AdoMet-Dependent Methyltransferases. *J. Am. Chem. Soc.* **2013**, *135*, 15536-15548.

52. Phatak, P.; Sumner, I.; Iyengar, S. S., Gauging the Flexibility of the Active Site in Soybean Lipoxygenase-1 (Slo-1) through an Atom-Centered Density Matrix Propagation (Admp) Treatment That Facilitates the Sampling of Rare Events. *J. Phys. Chem. B* **2012**, *116*, 10145-10164.

53. Lu, X.; Ovchinnikov, V.; Demapan, D.; Roston, D.; Cui, Q., Regulation and Plasticity of Catalysis in Enzymes: Insights from Analysis of Mechanochemical Coupling in Myosin. *Biochemistry* **2017**, *56*, 1482-1497.

54. Patra, N.; Ioannidis, E. I.; Kulik, H. J., Computational Investigation of the Interplay of Substrate Positioning and Reactivity in Catechol O-Methyltransferase. *PLoS ONE* **2016**, *11*, e0161868.

55. Parrish, R. M.; Thompson, K. C.; Martínez, T. J., Large-Scale Functional Group Symmetry-Adapted Perturbation Theory on Graphical Processing Units. *J. Chem. Theory Comput.* **2018**, *14*, 1737-1753.

56. Qi, H. W.; Kulik, H. J., Evaluating Unexpectedly Short Non-Covalent Distances in X-Ray Crystal Structures of Proteins with Electronic Structure Analysis. *J. Chem. Inf. Model.* **2019**, *59*, 2199-2211.

57. Vennelakanti, V.; Qi, H. W.; Mehmood, R.; Kulik, H. J., When Are Two Hydrogen Bonds Better Than One? Accurate First-Principles Models Explain the Balance of Hydrogen Bond Donors and Acceptors Found in Proteins. *Chem. Sci.* **2021**, *Accepted manuscript*.

58. Riniker, S., Fixed-Charge Atomic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J. Chem. Inf. Model.* **2018**, *58*, 565-578.
59. Field, M. J.; Bash, P. A.; Karplus, M., A Combined Quantum-Mechanical and Molecular Mechanical Potential for Molecular-Dynamics Simulations. *J. Comput. Chem.* **1990**, *11*, 700-733.

60. Bakowies, D.; Thiel, W., Hybrid Models for Combined Quantum Mechanical and Molecular Mechanical Approaches. *J. Phys. Chem.* **1996**, *100*, 10580-10594.

61. Mordasini, T. Z.; Thiel, W., Combined Quantum Mechanical and Molecular Mechanical Approaches. *Chimia* **1998**, *52*, 288-291.

62. Monard, G.; Merz, K. M., Combined Quantum Mechanical/Molecular Mechanical Methodologies Applied to Biomolecular Systems. *Acc. Chem. Res.* **1999**, *32*, 904-911.

63. Gao, J.; Truhlar, D. G., Quantum Mechanical Methods for Enzyme Kinetics. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467-505.

64. Rosta, E.; Klahn, M.; Warshel, A., Towards Accurate Ab Initio QM/MM Calculations of Free-Energy Profiles of Enzymatic Reactions. *J. Phys. Chem. B* **2006**, *110*, 2934-2941.

65. Lin, H.; Truhlar, D., QM/MM: What Have We Learned, Where Are We, and Where Do We Go from Here? *Theor. Chem. Acc.* **2007**, *117*, 185-199.

66. Warshel, A.; Levitt, M., Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103*, 227-249.

67. Senn, H. M.; Thiel, W., QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198-1229.

68. Acevedo, O.; Jorgensen, W. L., Advances in Quantum and Molecular Mechanical (QM/MM) Simulations for Organic and Enzymatic Reactions. *Acc. Chem. Res.* **2009**, *43*, 142-151.

69. Amaro, R. E.; Mulholland, A. J., Multiscale Methods in Drug Design Bridge Chemical and Biological Complexity in the Search for Cures. *Nat. Rev. Chem.* **2018**, *2*, 1-12.

70. Cui, Q., Perspective: Quantum Mechanical Methods in Biochemistry and Biophysics. *J. Chem. Phys.* **2016**, *145*, 140901.

71. Ryde, U., How Many Conformations Need to Be Sampled to Obtain Converged QM/MM Energies? The Curse of Exponential Averaging. *J. Chem. Theory Comput.* **2017**, *13*, 5745-5752.

72. Hu, L.; Soderhjelm, P.; Ryde, U., Accurate Reaction Energies in Proteins Obtained by Combining QM/MM and Large QM Calculations. *J. Chem. Theory Comput.* **2012**, *9*, 640-649.

73. Hu, L.; Söderhjelm, P. r.; Ryde, U., On the Convergence of QM/MM Energies. *J. Chem. Theory Comput.* **2011**, *7*, 761-777.

74. König, G.; Hudson, P. S.; Boresch, S.; Woodcock, H. L., Multiscale Free Energy Simulations: An Efficient Method for Connecting Classical MD Simulations to QM or QM/MM Free Energies Using Non-Boltzmann Bennett Reweighting Schemes. *J. Chem. Theory Comput.* **2014**, *10*, 1406-1419.

75. Mehmood, R.; Kulik, H. J., Both Configuration and QM Region Size Matter: Zinc Stability in QM/MM Models of DNA Methyltransferase. *J. Chem. Theory Comput.* **2020**, *16*, 3121-3134.

76. Karelina, M.; Kulik, H. J., Systematic Quantum Mechanical Region Determination in QM/MM Simulation. *J. Chem. Theory Comput.* **2017**, *13*, 563-576.

77. Eurenius, K. P.; Chatfield, D. C.; Brooks, B. R.; Hodoscek, M., Enzyme Mechanisms with Hybrid Quantum and Molecular Mechanical Potentials. I. Theoretical Considerations. *Int. J. Quantum Chem.* **1996**, *60*, 1189-1200.
78. Senn, H. M.; Thiel, W., QM/MM Studies of Enzymes. *Curr. Opin. Chem. Biol.* **2007**, *11*, 182-187.
79. Monari, A.; Rivail, J.-L.; Assfeld, X., Advances in the Local Self-Consistent Field Method for Mixed Quantum Mechanics/Molecular Mechanics Calculations. *Acc. Chem. Res.* **2012**, *46*, 596-603.
80. Wang, Y.; Gao, J., Projected Hybrid Orbitals: A General QM/MM Method. *J. Phys. Chem. B* **2015**, *119*, 1213-1224.
81. Murphy, R. B.; Philipp, D. M.; Friesner, R. A., A Mixed Quantum Mechanics/Molecular Mechanics (QM/MM) Method for Large Scale Modeling of Chemistry in Protein Environments. *J. Comput. Chem.* **2000**, *21*, 1442-1457.
82. Zhang, Y.; Lee, T.-S.; Yang, W., A Pseudobond Approach to Combining Quantum Mechanical and Molecular Mechanical Methods. *J. Chem. Phys.* **1999**, *110*, 46-54.
83. DiLabio, G. A.; Hurley, M. M.; Christiansen, P. A., Simple One-Electron Quantum Capping Potentials for Use in Hybrid QM/MM Studies of Biological Molecules. *J. Chem. Phys.* **2002**, *116*, 9578-9584.
84. von Lilienfeld, O. A.; Tavbernelli, I.; Rothlisberger, U.; Sebastiani, D., Variational Optimization of Effective Atom-Centered Potentials for Molecular Properties. *J. Chem. Phys.* **2005**, *122*, 14113.
85. Wang, B.; Truhlar, D. G., Combined Quantum Mechanical and Molecular Mechanical Methods for Calculating Potential Energy Surfaces: Tuned and Balanced Redistributed Charge Algorithm. *J. Chem. Theory Comput.* **2010**, *6*, 359-369.
86. Kairys, V.; Jensen, J. H., QM/MM Boundaries across Covalent Bonds: A Frozen Localized Molecular Orbital-Based Approach for the Effective Fragment Potential Method. *J. Phys. Chem. A* **2000**, *104*, 6656-6665.
87. Watanabe, H. C.; Cui, Q., Quantitative Analysis of QM/MM Boundary Artifacts and Correction in Adaptive QM/MM Simulations. *J. Chem. Theory Comput.* **2019**, *15*, 3917-3928.
88. Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr. R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T., Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549-2564.
89. Halgren, T. A.; Damm, W., Polarizable Force Fields. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236-242.
90. Thellamurege, N. M.; Hirao, H., Effect of Protein Environment within Cytochrome P450cam Evaluated Using a Polarizable-Embedding QM/MM Method. *J. Phys. Chem. B* **2014**, *118*, 2084-2092.
91. Nåbo, L. J.; Olsen, J. M. H.; Martinez, T. J.; Kongsted, J., The Quality of the Embedding Potential Is Decisive for Minimal Quantum Region Size in Embedding Calculations: The Case of the Green Fluorescent Protein. *J. Chem. Theory Comput.* **2017**, *13*, 6230-6236.
92. Ganguly, A.; Boulanger, E.; Thiel, W., Importance of MM Polarization in QM/MM Studies of Enzymatic Reactions: Assessment of the QM/MM Drude Oscillator Model. *J. Chem. Theory Comput.* **2017**, *13*, 2954-2961.
93. Loco, D.; Lagardère, L.; Caprascenna, S.; Lipparini, F.; Mennucci, B.; Piquemal, J.-P., Hybrid QM/MM Molecular Dynamics with AMOEBA Polarizable Embedding. *J. Chem. Theory Comput.* **2017**, *13*, 4025-4033.
94. Bondanza, M.; Nottoli, M.; Cupellini, L.; Lipparini, F.; Mennucci, B., Polarizable Embedding QM/MM: The Future Gold Standard for Complex (Bio) Systems? *Phys. Chem. Chem. Phys.* **2020**, *22*, 14433-14448.
95. Loco, D.; Lagardère, L.; Cisneros, G. A.; Scalmani, G.; Frisch, M.; Lipparini, F.; Mennucci, B.; Piquemal, J.-P., Towards Large Scale Hybrid QM/MM Dynamics of Complex Systems with Advanced Point Dipole Polarizable Embeddings. *Chem. Sci.* **2019**, *10*, 7200-7211.
96. Kulik, H. J.; Luehr, N.; Ufimtsev, I. S.; Martínez, T. J., Ab Initio Quantum Chemistry for Protein Structures. *J. Phys. Chem. B* **2012**, *116*, 12501-12509.
97. Ufimtsev, I. S.; Martínez, T. J., Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5*, 2619-2628.
98. Isborn, C. M.; Luehr, N.; Ufimtsev, I. S.; Martínez, T. J., Excited-State Electronic Structure with Configuration Interaction Singles and Tamm-Dancoff Time-Dependent Density Functional Theory on Graphical Processing Units. *J. Chem. Theory Comput.* **2011**, *7*, 1814-1823.
99. Ufimtsev, I. S.; Luehr, N.; Martínez, T. J., Charge Transfer and Polarization in Solvated Proteins from Ab Initio Molecular Dynamics. *J. Phys. Chem. Lett.* **2011**, *2*, 1789-1793.
100. Ochsenfeld, C.; Kussmann, J.; Lambrrecht, D. S., Linear-Scaling Methods in Quantum Chemistry. *Rev. Comput. Chem.* **2007**, *23*, 1.
101. Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R., Auxiliary Basis Sets for Main Row Atoms and Transition Metals and Their Use to Approximate Coulomb Potentials. *Theor. Chem. Acc.* **1997**, *97*, 119-124.
102. Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R., Auxiliary Basis Sets to Approximate Coulomb Potentials. *Chem. Phys. Lett.* **1995**, *240*, 283-290.
103. Liu, F.; Luehr, N.; Kulik, H. J.; Martínez, T. J., Quantum Chemistry for Solvated Molecules on Graphical Processing Units Using Polarizable Continuum Models. *J. Chem. Theory Comput.* **2015**, *11*, 3131-3144.
104. Flaig, D.; Beer, M.; Ochsenfeld, C., Convergence of Electronic Structure with the Size of the QM Region: Example of QM/MM NMR Shieldings. *J. Chem. Theory Comput.* **2012**, *8*, 2260-2271.
105. Hartman, J. D.; Neubauer, T. J.; Caulkins, B. G.; Mueller, L. J.; Beran, G. J., Converging Nuclear Magnetic Shielding Calculations with Respect to Basis and System Size in Protein Systems. *J. Biomol. NMR* **2015**, *62*, 327-340.
106. Roßbach, S.; Ochsenfeld, C., Influence of Coupling and Embedding Schemes on QM Size Convergence in QM/MM Approaches for the Example of a Proton Transfer in DNA. *J. Chem. Theory Comput.* **2017**, *13*, 1102-1107.
107. Sumowski, C. V.; Ochsenfeld, C., A Convergence Study of QM/MM Isomerization Energies with the Selected Size of the QM Region for Peptidic Systems. *J. Phys. Chem. A* **2009**, *113*, 11734-11741.
108. Fox, S. J.; Pittock, C.; Fox, T.; Tautermann, C. S.; Malcolm, N.; Skylaris, C. K., Electrostatic Embedding in Large-Scale First Principles Quantum Mechanical Calculations on Biomolecules. *J. Chem. Phys.* **2011**, *135*, 224107.
109. Liao, R. Z.; Thiel, W., Convergence in the QM-Only and QM/MM Modeling of Enzymatic Reactions: A Case Study for Acetylene Hydratase. *J. Comput. Chem.* **2013**, *34*, 2389-2397.
110. Sadeghian, K.; Flaig, D.; Blank, I. D.; Schneider, S.; Strasser, R.; Stathis, D.; Winnacker, M.; Carell, T.; Ochsenfeld, C., Ribose-Protonated DNA Base Excision Repair: A Combined Theoretical and Experimental Study. *Angew. Chem., Int. Ed.* **2014**, *53*, 10044-10048.

111. Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martinez, T. J., How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol O-Methyltransferase. *J. Phys. Chem. B* **2016**, *120*, 11381-11394.

112. Solt, I.; Kulhanek, P.; Simon, I.; Winfield, S.; Payne, M. C.; Csanyi, G.; Fuxreiter, M., Evaluating Boundary Dependent Errors in QM/MM Simulations. *J. Chem. Theory Comput.* **2009**, *113*, 5728-5735.

113. Isborn, C. M.; Goetz, A. W.; Clark, M. A.; Walker, R. C.; Martinez, T. J., Electronic Absorption Spectra from MM and Ab Initio QM/MM Molecular Dynamics: Environmental Effects on the Absorption Spectrum of Photoactive Yellow Protein. *J. Chem. Theory Comput.* **2012**, *8*, 5092-5106.

114. Provorse, M. R.; Peev, T.; Xiong, C.; Isborn, C. M., Convergence of Excitation Energies in Mixed Quantum and Classical Solvent: Comparison of Continuum and Point Charge Models. *J. Phys. Chem. B* **2016**, *120*, 12148-12159.

115. Milanese, J. M.; Provorse, M. R.; Alameda, E.; Isborn, C. M., Convergence of Computed Aqueous Absorption Spectra with Explicit Quantum Mechanical Solvent. *J. Chem. Theory Comput.* **2017**, *13*, 2159-2171.

116. Vanpoucke, D. E.; Oláh, J.; De Proft, F.; Van Speybroeck, V.; Roos, G., Convergence of Atomic Charges with the Size of the Enzymatic Environment. *J. Chem. Inf. Model.* **2015**, *55*, 564-571.

117. Morgenstern, A.; Jaszai, M.; Eberhart, M. E.; Alexandrova, A. N., Quantified Electrostatic Preorganization in Enzymes Using the Geometry of the Electron Charge Density. *Chem. Sci.* **2017**, *8*, 5010-5018.

118. Harris, T. V.; Szilagyi, R. K., Protein Environmental Effects on Iron - Sulfur Clusters: A Set of Rules for Constructing Computational Models for Inner and Outer Coordination Spheres. *J. Comput. Chem.* **2016**, *37*, 1681-1696.

119. Benediktsson, B.; Bjornsson, R., QM/MM Study of the Nitrogenase Mofe Protein Resting State: Broken-Symmetry States, Protonation States, and QM Region Convergence in the Femoco Active Site. *Inorg. Chem.* **2017**, *56*, 13417-13429.

120. Provorse Long, M. R.; Isborn, C. M., Combining Explicit Quantum Solvent with a Polarizable Continuum Model. *J. Phys. Chem. B* **2017**, *121*, 10105-10117.

121. Waller, M. P.; Kumbhar, S.; Yang, J., A Density - Based Adaptive Quantum Mechanical/Molecular Mechanical Method. *ChemPhysChem* **2014**, *15*, 3218-3225.

122. Summers, T.; Cheng, Q.; Palma, M.; Pham, D.-T.; Kelso III, D.; Edwin Webster, C.; DeYonker, N., Rational, Reproducible, and Rigorous Computational Enzymology: The Case of Catechol-O-Methyltransferase. *ChemRxiv, DOI:10.26434/chemrxiv.12756245.v1* **2020**.

123. Qi, H. W.; Karelina, M.; Kulik, H. J., Quantifying Electronic Effects in QM and QM/MM Biomolecular Modeling with the Fukui Function. *Acta Phys.-Chem. Sin.* **2018**, *34*, 81-91.

124. Kulik, H. J., Large-Scale QM/MM Free Energy Simulations of Enzyme Catalysis Reveal the Influence of Charge Transfer. *Phys. Chem. Chem. Phys.* **2018**, *20*, 20650-20660.

125. Yang, Z.; Mehmood, R.; Wang, M.; Qi, H. W.; Steeves, A. H.; Kulik, H. J., Revealing Quantum Mechanical Effects in Enzyme Catalysis with Large-Scale Electronic Structure Simulation. *React. Chem. Eng.* **2019**, *4*, 298-315.
126. Cortina, G. A.; Kasson, P. M., Excess Positional Mutual Information Predicts Both Local and Allosteric Mutations Affecting Beta Lactamase Drug Resistance. *Bioinformatics* 2016, 32, 3420-3427.
127. Guo, J.; Zhou, H.-X., Protein Allostery and Conformational Dynamics. *Chem. Rev.* 2016, 116, 6503-6515.
128. McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P., Quantifying Correlations between Allosteric Sites in Thermodynamic Ensembles. *J. Chem. Theory Comput.* 2009, 5, 2486-2502.
129. Zhou, G.; Chu, W.; Prezhdo, O. V., Structure Deformation Controls Charge Losses in Mapbi3: Unsupervised Machine Learning of Nonadiabatic Molecular Dynamics. *ACS Energy Lett.* 2020.
130. Wang, L.-P.; McGibbon, R. T.; Pande, V. S.; Martinez, T. J., Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways. *J. Chem. Theory Comput.* 2016, 12, 638-649.
131. Hutchings, M.; Liu, J.; Qiu, Y.; Song, C.; Wang, L.-P., Bond-Order Time Series Analysis for Detecting Reaction Events in Ab Initio Molecular Dynamics Simulations. *J. Chem. Theory Comput.* 2020, 16, 1606-1617.
132. Stein, C. J.; Reiher, M., Automated Selection of Active Orbital Spaces. *J. Chem. Theory Comput.* 2016, 12, 1760-1771.
133. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* 2000, 28, 235-242.
134. Vita, C.; Drakopoulos, E.; Vizzavona, J.; Rochette, S.; Martin, L.; Ménez, A.; Roumestand, C.; Yang, Y.-S.; Ylisastigui, L.; Benjouad, A.; Gluckman, J. C., Rational Engineering of a Miniprotein That Reproduces the Core of the Cd4 Site Interacting with Hiv-1 Envelope Glycoprotein. *Proc. Natl. Acad. Sci. U. S. A.* 1999, 96, 13091.
135. Zong, C.; Wu, M. J.; Qin, J. Z.; Link, A. J., Lasso Peptide Benenodin-1 Is a Thermally Actuated [1]Rotaxane Switch. *J. Am. Chem. Soc.* 2017, 139, 10403-10409.
136. Neidigh, J. W.; Fesinnmeyer, R. M.; Andersen, N. H., Designing a 20-Residue Protein. *Nat. Struct. Biol.* 2002, 9, 425.
137. Maksimov, M. O.; Pan, S. J.; Link, A. J., Lasso Peptides: Structure, Function, Biosynthesis, and Engineering. *Nat. Prod. Rep.* 2012, 29, 996-1006.
138. Kitazawa, S.; Fossat, M. J.; McCallum, S. A.; Garcia, A. E.; Royer, C. A., NMR and Computation Reveal a Pressure-Sensitive Folded Conformation of Trp-Cage. *J. Phys. Chem. B* 2017, 121, 1258-1267.
139. Day, R.; Paschek, D.; Garcia, A. E., Microsecond Simulations of the Folding/Unfolding Thermodynamics of the Trp-Cage Miniprotein. *Proteins: Struct., Funct., Bioinf.* 2010, 78, 1889-1899.
140. Zagrovic, B.; Pande, V., Solvent Viscosity Dependence of the Folding Rate of a Small Protein: Distributed Computing Study. *J. Comput. Chem.* 2003, 24, 1432-1436.
141. Becke, A. D., A Multicenter Numerical Integration Scheme for Polyatomic Molecules. *J. Chem. Phys.* 1988, 88, 2547-2553.
142. Fonseca Guerra, C.; Handgraaf, J. W.; Baerends, E. J.; Bickelhaupt, F. M., Voronoi Deformation Density (VDD) Charges: Assessment of the Mulliken, Bader, Hirshfeld, Weinhold, and VDD Methods for Charge Analysis. *J. Comput. Chem.* 2004, 25, 189-210.
143. Hirshfeld, F. L., Bonded-Atom Fragments for Describing Molecular Charge Densities. *Theor. Chim. Acta* 1977, 44, 129-138.
144. Bracewell, R., Pentagram Notation for Cross Correlation. The Fourier Transform and Its Applications. *New York: McGraw-Hill 1965*, 46, 243.
145. Ichiye, T.; Karplus, M., Collective Motions in Proteins: A Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and Normal Mode Simulations. *Proteins: Struct., Funct., Bioinf.* **1991**, *11*, 205-217.
146. Cover, T. M.; Thomas, J. A., *Elements of Information Theory*. John Wiley & Sons: 2012.
147. Smith, R., A Mutual Information Approach to Calculating Nonlinearity. *Stat* **2015**, *4*, 291-303.
148. Gel'fand, I. M.; Yaglom, A. M., Computation of the Amount of Information About a Stochastic Function Contained in Another Such Function. *Usp. Mat. Nauk* **1957**, *12*, 3-52.
149. Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V., H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Res.* **2012**, *40*, W537-W541.
150. Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A., H++: A Server for Estimating pKas and Adding Missing Hydogenos to Macromolecules. *Nucleic Acids Res.* **2005**, *33*, W368-W371.
151. Myers, J.; Grothaus, G.; Narayanan, S.; Onufriev, A., A Simple Clustering Algorithm Can Be Accurate Enough for Use in Calculations of pKs in Macromolecules. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 928-938.
152. D.A. Case, J. T. B., R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman Amber 2018, University of California, San Francisco. 2018.
153. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696-3713.
154. Wang, J.; Cieplak, P.; Kollman, P. A., How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comput. Chem.* **2000**, *21*, 1049-1074.
155. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926-935.
156. Jayaram, B.; Sprous, D.; Beveridge, D. L., Solvation Free Energy of Biomacromolecules: Parameters for a Modified Generalized Born Model Consistent with the Amber Force Field. *J. Phys. Chem. B* **1998**, *102*, 9571-9576.
157. Onufriev, A.; Case, D. A.; Bashford, D., Effective Born Radii in the Generalized Born Approximation: The Importance of Being Perfect. *J. Comput. Chem.* **2002**, *23*, 1297-1304.
158. Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with Amber on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542-1555.
159. Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with Amber on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878-3888.
160. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* 1977, 23, 327-341.
161. Rohrdanz, M. A.; Martins, K. M.; Herbert, J. M., A Long-Range-Corrected Density Functional That Performs Well for Both Ground-State Properties and Time-Dependent Density Functional Theory Excitation Energies, Including Charge-Transfer Excited States. *J. Chem. Phys.* 2009, 130, 054112.
162. Harishara, P. C.; Pople, J. A., Influence of Polarization Functions on Molecular-Orbital Hydrogenation Energies. *Theor. Chim. Acta* 1973, 28, 213-222.
163. Lange, A. W.; Herbert, J. M., A Smooth, Nonsingular, and Faithful Discretization Scheme for Polarizable Continuum Models: The Switching/Gaussian Approach. *J. Chem. Phys.* 2010, 133, 244111.
164. York, D. M.; Karplus, M., A Smooth Solvation Potential Based on the Conductor-Like Screening Model. *J. Phys. Chem. A* 1999, 103, 11060-11079.
165. Liu, F.; Sanchez, D. M.; Kulik, H. J.; Martinez, T. J., Exploiting Graphical Processing Units to Enable Quantum Chemistry Calculation of Large Solvated Molecules with Conductor-Like Polarizable Continuum Models. *Int. J. Quantum Chem.* 2019, 119, e25760.
166. Petachem. [http://www.petachem.com](http://www.petachem.com). (accessed April 29, 2020).
167. Bondi, A., Van Der Waals Volumes and Radii. *J. Phys. Chem.* 1964, 68, 441-451.
168. Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M., Replica-Exchange Monte Carlo Method for the Isobaric–Isothermal Ensemble. *Chem. Phys. Lett.* 2001, 335, 435-439.
169. Sugita, Y.; Okamoto, Y., Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* 1999, 314, 141-151.
170. Grimme, S.; Ehrlich, S.; Goerigk, L., Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* 2011, 32, 1456-1465.
171. Grimme, S., Density Functional Theory with London Dispersion Corrections. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* 2011, 1, 211-228.
172. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E., Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011, 12, 2825--2830.
173. Kraskov, A.; Stogbauer, H.; Grassberger, P., Estimating Mutual Information. *Phys. Rev. E* 2004, 69, 066138.
Electronic Supplementary Information for

Quantifying the long-range coupling of electronic properties in proteins with ab initio molecular dynamics

Zhongyue Yang¹, Natalia Hajlasz¹, Adam H. Steeves¹, and Heather J. Kulik¹,*

¹Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

Contents

Table S1 Results of mini-protein PDB curation: NMR solution analysis Page S3
Table S2 Protonation states and number of atoms of residues in mini-CD4 Page S4
Figure S1 Distribution of residues by type in peptide structures Page S5
Table S3 Results of lasso peptide PDB curation: NMR solution analysis Page S5
Table S4 Protonation states and number of atoms of residues in benenodin-1 Page S6
Table S5 Results of designed protein PDB curation: NMR solution analysis Page S6
Table S6 Protonation states and numbers of atoms of residues in Trp-cage Page S7
Table S7 Comparison of q(RES) for different charge partitioning schemes Page S8
Table S8 q(RES) distribution statistics for residues in mini-CD4 Page S9
Table S9 q(RES) distribution statistics for residues in benenodin-1 Page S10
Table S10 q(RES) distribution statistics for residues in Trp-cage Page S11
Table S11 Sidechain-only q(RES) distribution statistics for residues in mini-CD4 Page S12
Figure S2 Parity plots of sidechain vs whole-residue q(RES) statistics in mini-CD4 Page S13
Figure S3 q(RES) histograms of nonpolar residues in mini-CD4 Page S14
Figure S4 q(RES) histograms of nonpolar residues in benenodin-1 Page S15
Figure S5 q(RES) histograms of nonpolar residues in Trp-cage Page S16
Figure S6 q(RES) histograms of polar residues in mini-CD4 Page S17
Figure S7 q(RES) histograms of polar residues in benenodin-1 Page S18
Figure S8 q(RES) histograms of polar residues in Trp-cage Page S19
Figure S9 q(RES) histograms of charged residues in mini-CD4 Page S20
Figure S10 q(RES) histograms of charged residues in benenodin-1 Page S21
Figure S11 q(RES) histograms of charged residues in Trp-cage Page S22
Table S12 Summary of average and max. CC values for mini-CD4 Page S23
Table S13 Summary of average and max. CC values for benenodin-1 Page S24
Table S14 Summary of average and max. CC values for Trp-cage Page S25
Table S15 Avg., max. and min. CC magnitudes by type for all 3 proteins Page S26
Table S16 CC between Cys residues in mini-CD4 Page S27
Table S17 List of high CCs for mini-CD4 Page S27
Table S18 List of high CCs for benenodin-1 Page S28
Table S19 List of high CCs for Trp-cage Page S28
Table S20 Counts of high CC values for polar residues in all 3 proteins Page S29
Text S1 Decomposition of total MI into linear and nonlinear contributions Page S29
Figure S12 MI vs CC in all mini-proteins Page S30
Figure S13 Rank correlation between the total MI and its linear components Page S30
Figure S14 Distribution of MI for mini-CD4 Page S31
Figure S15 Distribution of MI for benenodin-1 Page S31
Figure S16 Distribution of MI for Trp-cage Page S32
Table S21 Cases where MI rank exceeds |CC| rank by >25% for mini-CD4
Table S22 Cases where |CC| rank exceeds MI rank by >25% for mini-CD4
Table S23 Cases where MI rank exceeds |CC| rank by >25% for benenodin-1
Table S24 Cases where |CC| rank exceeds MI rank by >25% for benenodin-1
Table S25 Cases where MI rank exceeds |CC| rank by >25% for Trp-cage
Table S26 Cases where |CC| rank exceeds MI rank by >25% for Trp-cage
Table S27 Summary of types of residues with MI and CC differences
Figure S17 Comparison of d(COM-COM) for NMR/AIMD and shortest distances
Figure S18 Distribution of COM-COM distances of Lys11-Gln20 in mini-CD4
Figure S19 MI and CC with d(COM-COM) for Trp-cage
Text S2 MI calculation of torsional degrees of freedom and analysis
Figure S20 MI of torsional angles among the residue pairs of mini-CD4
Figure S21 Comparison of percentile ranks of geometric and charge coupling
Figure S22 Example comparison of geometric and electronic MI
Table S28 Force field parameters for the isopeptide bond in benenodin-1
Figure S23 RMSDs of backbone atoms for implicit/explicit solvent MD trajectories
Figure S24 Comparison of REMD and AIMD charge MI on mini-CD4
Figure S25 Comparison of REMD and AIMD charge MI on benenodin-1
Figure S26 AIMD charge MI for mini-CD4 with trajectory length
Figure S27 AIMD charge MI for benenodin-1 with trajectory length
Table S29 AIMD charge for mini-CD4 with trajectory length
Table S30 AIMD charge for benenodin-1 with trajectory length
Table S31 Temperatures and details for REMD simulation
Table S32 Comparison of REMD and AIMD charge distribution
Figure S28 Comparison of REMD charge MI with and without D3 correction
Table S33 Details of spherical droplet construction for Trp-cage
References
**Table S1.** Summary of solution NMR structure analysis for 13 unique mini-protein structures with associated peer reviewed publications curated from the PDB in August 2019. The mini-proteins were required to have a well-folded construction containing 20-50% alpha helix and 20-50% beta sheet component, constrained by 1 to 3 disulfide bonds. The sequence length was set between 20 and 30. The parent PDB code is listed in the first column and “other structures” refers to other PDB codes affiliated with that same paper. Only one example from each paper is shown. Function is the putative function affiliated with that same paper. Only one example from each paper is shown. Function is the putative function of the relevant enzyme.

| PDB  | Res. Count | Year | Function                                                                 | Other structures | Ref.                        |
|------|------------|------|--------------------------------------------------------------------------|------------------|-----------------------------|
| 1D5Q | 27         | 2000 | Reproducing the core of the CD4 surface interacting with the HIV-1 envelope glycoprotein | No               | DOI: 10.1073/pnas.96.23.13091 |
| 6NW8 | 27         | 2019 | Toxin                                                                    | No               | DOI: 10.1016/j.toxicon.2019.06.013 |
| 6E5H | 28         | 2018 | Heterogeneous-backbone mimics of a designed disulfide-rich protein: Aib turn | 6E5K, 6E5J, 6E5I | DOI: 10.1002/cbic.201800558  |
| 5KVN | 27         | 2016 | Designed peptide NC. HEE_D1                                              | 5KX2, 5KX1, 5KX0, 5KWX, 5KWP, 5KWO, 5JI4, 5JHI, 5JG9, 2ND2, 2ND3 | DOI: 10.1038/nature19791 |
| 2RTY | 30         | 2014 | Neurotoxin                                                                | No               | DOI: 10.1093/molbev/msu038   |
| 2KUX | 30         | 2010 | The cyclotide kalata B5 from Oldenlandia affinis                         | No               | DOI: 10.1002/bip.21409       |
| 2KNM | 30         | 2009 | The cyclotide cycloviolacin O2                                            | 2KNN             | DOI: 10.1002/cbic.200900342  |
| 2KCG | 30         | 2009 | The cyclotide cycloviolacin O2                                            | 2KCH             | DOI: 10.1016/j.bpj.2009.06.032 |
| 2EQT | 28         | 2007 | Growth-blocking peptide of the armyworm                                   | 2EQQ, 2EOH       | DOI: 10.1074/jbc.M109.011148  |
| 1ZWU | 30         | 2005 | AcAMP2-like peptide with non natural beta-(2-naphthyl)-alanine residue    | 1ZUV, 1ZNT       | DOI: 10.1002/chem.200500367  |
| 1NBJ | 30         | 2003 | Cycloviolacin O1                                                          | 1NB1             | DOI: 10.1074/jbc.M211147200   |
| 1JLZ | 23         | 2002 | K(+)‐channel blocker from the scorpion Tityus cambridgei                  | No               | DOI: 10.1110/ps.33402         |
| 1ACW | 29         | 1997 | Natural scorpion peptide                                                  | No               | DOI: 10.1002/(SICI)1097-0134(199603)24:3<359::AID-PRCB |
Table S2. Protonation state, number of atoms, and classification of amino acids in mini-CD4 (PDB ID: 1D5Q). Three disulfide bridges are formed between the residues denoted Cyx: Cyx1-Cyx19, Cyx6-Cyx24, and Cyx10-Cyx26. The Cyx1 N-terminus carries a positive charge and so is classified as a charged residue.

| Res.  | # At | Charge | Type   |
|-------|------|--------|--------|
| Cyx1  | 12   | 1      | Charged|
| Asn2  | 14   | 0      | Polar  |
| Leu3  | 19   | 0      | Nonpolar|
| Ala4  | 10   | 0      | Nonpolar|
| Arg5  | 24   | 1      | Charged|
| Cyx6  | 10   | 0      | Polar  |
| Gin7  | 17   | 0      | Polar  |
| Leu8  | 19   | 0      | Nonpolar|
| Ser9  | 11   | 0      | Polar  |
| Cyx10 | 10   | 0      | Polar  |
| Lys11 | 22   | 1      | Charged|
| Ser12 | 11   | 0      | Polar  |
| Leu13 | 19   | 0      | Nonpolar|
| Gly14 | 7    | 0      | Nonpolar|
| Leu15 | 19   | 0      | Nonpolar|
| Lys16 | 22   | 1      | Charged|
| Gly17 | 7    | 0      | Nonpolar|
| Gly18 | 7    | 0      | Nonpolar|
| Cyx19 | 10   | 0      | Polar  |
| Gin20 | 17   | 0      | Polar  |
| Gly21 | 7    | 0      | Nonpolar|
| Ser22 | 11   | 0      | Polar  |
| Phe23 | 20   | 0      | Nonpolar|
| Cyx24 | 10   | 0      | Polar  |
| Thr25 | 14   | 0      | Polar  |
| Cyx26 | 10   | 0      | Polar  |
| Gly27 | 8    | -1     | Charged|
| Total | 367  | 3      |        |
**Figure S1.** Cartoon structures of three proteins (i.e., mini-CD4, left; benenodin-1, middle; and Trp-cage, right) studied in this work with the heavy atoms of sidechains shown as sticks. Residues in each protein are colored by type: positively charged (red), negatively charged (blue), polar (green), and nonpolar (gray). Residues in the isopeptide bond are colored light green in benenodin-1. Some residues of each protein are indicated with a single-letter code and number in the primary sequence. Terminal residues or other special residues with unconventional charge are indicated with an asterisk.

**Table S3.** Summary of solution NMR structure analysis for 13 lasso peptide structures with associated peer reviewed publications curated from the PDB in August 2019. The sequence length was constrained to range between 10 and 20. The parent PDB code is listed in the first column and “other structures” refers to other PDB codes affiliated with that same paper. Only one example from each paper is shown. Function is the putative function of the relevant enzyme.

| PDB     | Res. Count | Year | Function                  | Other structures | Ref.                        |
|---------|------------|------|---------------------------|------------------|-----------------------------|
| 6B5W    | 19         | 2017 | Thermally actuated rotaxane switch | STJ1, 5TJ0       | DOI: 10.1021/jacs.7b04830   |
| 6MW6    | 19         | 2019 | Antimicrobial citrocin     |                  | DOI: 10.1074/jbc.RA118.006494|
| 5XM4    | 17         | 2018 | Unknown                    |                  | DOI: 10.1016/j.tetlet.2017.07.06|
| 5T56    | 12         | 2016 | Catenanes                  |                  | DOI: 10.1021/jacs.6b09454   |
| 5GVO    | 18         | 2017 | Unknown                    |                  | DOI: 10.1002/ehoc.201601334  |
| 5JPL    | 17         | 2016 | Antibiotic                 |                  | DOI: 10.1038/nchembio.2319  |
| 5U16    | 18         | 2017 | Antimicrobial              | 5U17             | DOI: 10.1021/acschembio.6b01  |
| 2N6U    | 20         | 2015 | Unknown                    | 2N6V             | DOI: 10.1074/jbc.M115.694083  |
| 2N5C    | 15         | 2015 | Cell invasion              | No               | DOI: 10.1021/acs.joc.5b01878  |
| 2MLJ    | 18         | 2014 | Unknown                    | No               | DOI: 10.1039/C4SC01428F       |
| 2MAI    | 16         | 2014 | Antibiotic                 | No               | DOI: 10.1016/j.chembiol.2014.0  |
| 2M37    | 19         | 2013 | Unknown                    | No               | DOI: 10.1016/j.chembiol.2013.0  |
| 2LX6    | 19         | 2012 | Unknown                    | No               | DOI: 10.1021/ja308173b        |
Table S4. Protonation state, number of atoms, and classification of amino acids in benenodin-1 (PDB ID: 6B5W).

Gly1 and Asp8 in benenodin-1 form isopeptide bond, which means there is no N-terminus positive charge on Gly1, and Asp8 is also neutral.

| Res. | # At | Charge | Type  |
|------|------|--------|-------|
| Gly1 | 7    | 0      | Polar |
| Val2 | 16   | 0      | Nonpolar |
| Gly3 | 7    | 0      | Nonpolar |
| Phe4 | 20   | 0      | Nonpolar |
| Gly5 | 7    | 0      | Nonpolar |
| Arg6 | 24   | 1      | Charged |
| Pro7 | 14   | 0      | Polar  |
| Asp8 | 11   | 0      | Polar  |
| Ser9 | 11   | 0      | Polar  |
| Ile10| 19   | 0      | Nonpolar |
| Leu11| 19   | 0      | Nonpolar |
| Thr12| 14   | 0      | Polar  |
| Gln13| 17   | 0      | Polar  |
| Glu14| 15   | -1     | Charged |
| Gln15| 17   | 0      | Polar  |
| Ala16| 10   | 0      | Nonpolar |
| Lys17| 22   | 1      | Charged |
| Pro18| 14   | 0      | Polar  |
| Met19| 18   | -1     | Charged |
| Total | 282  | -1     |        |

Table S5. Summary of solution NMR structure analysis for 13 unique de novo designed protein structures with associated peer reviewed publications curated from the PDB in August 2019. The de novo designed proteins were required to have a sequence length between 20 and 22. The parent PDB code is listed in the first column and “other structures” refers to other PDB codes affiliated with that same paper. Only one example from each paper is shown. Function is the putative function of the relevant enzyme.

| PDB  | Res. Count | Year | Other structures | Ref.               |
|------|------------|------|------------------|--------------------|
| 1L2Y | 20         | 2002 | No               | DOI: 10.1038/nsb798 |
| 6D37 | 21         | 2019 | No               | DOI: 10.1002/bip.23260 |
| 5V2G | 20         | 2017 | No               | DOI: 10.1073/pnas.1710695114 |
| 2N8D | 21         | 2017 | No               | DOI: 10.1002/sml.201701316 |
| 5KX1 | 22         | 2016 | No               | DOI: 10.1038/nature19791 |
| 2M7C | 21         | 2013 | 2M7D            | DOI: 10.1039/C3RA43674H |
| 2LL5 | 22         | 2012 | No               | DOI: 10.1073/pnas.1121421109 |
| 2LDJ | 20         | 2011 | No               | DOI: 10.1021/ja205609c |
| 2JOF | 20         | 2008 | No               | DOI: 10.1093/protein/gzm082 |
| 2JO4 | 22         | 2008 | 2JO5            | DOI: 10.1073/pnas.0706876104 |
| 2ORU | 20         | 2007 | No               | DOI: 10.1002/cbic.200600565 |
| 1U0I | 20         | 2004 | No               | DOI: 10.1002/bip.20150 |
| 1JY9 | 20         | 2001 | No               | DOI: 10.1073/pnas.211536998 |
Table S6. Protonation state, number of atoms, and classification of amino acids in Trp-cage (PDB ID: 1L2Y)\(^3\).

| Res. | # At | Charge | Type    |
|------|------|--------|---------|
| Asn1 | 16   | 1      | Charged |
| Leu2 | 19   | 0      | Nonpolar|
| Tyr3 | 21   | 0      | Polar   |
| Ile4 | 19   | 0      | Nonpolar|
| Gln5 | 17   | 0      | Polar   |
| Trp6 | 24   | 0      | Polar   |
| Leu7 | 19   | 0      | Nonpolar|
| Tyr3 | 21   | 0      | Polar   |
| Asn1 | 16   | 1      | Charged |
| Leu2 | 19   | 0      | Nonpolar|
| Tyr3 | 21   | 0      | Polar   |
| Ile4 | 19   | 0      | Nonpolar|
| Gln5 | 17   | 0      | Polar   |
| Trp6 | 24   | 0      | Polar   |
| Leu7 | 19   | 0      | Nonpolar|
| Lys8 | 22   | 1      | Charged |
| Asp9 | 12   | -1     | Charged |
| Gly10| 7    | 0      | Nonpolar|
| Gly11| 7    | 0      | Nonpolar|
| Pro12| 14   | 0      | Polar   |
| Ser13| 11   | 0      | Polar   |
| Ser14| 11   | 0      | Polar   |
| Gly15| 7    | 0      | Nonpolar|
| Arg16| 24   | 1      | Charged |
| Pro17| 14   | 0      | Polar   |
| Pro18| 14   | 0      | Polar   |
| Pro19| 14   | 0      | Polar   |
| Ser20| 12   | -1     | Charged |

Total 304 1
Table S7. The average and std. of by-residue-summed charges in mini-CD4 for each amino acid computed with different partial charge schemes. These charges are averaged over 10 random snapshots from the AIMD trajectory. The four charge schemes compared are Mulliken, Voronoi deformation density (VDD)\(^4\), Becke\(^5\), and Hirshfeld\(^6\) charges. All charges were computed from the wavefunctions generated as described in the computational details and then post-processed in Multiwfn\(^7\).

| Res.  | Mulliken     | VDD       | Becke       | Hirshfeld   |
|-------|--------------|-----------|-------------|-------------|
| Cys1* | 0.889±0.024  | 0.646±0.045| 0.745±0.150 | 0.633±0.053 |
| Asn2  | 0.135±0.025  | 0.202±0.036| -0.034±0.097| 0.266±0.051 |
| Leu3  | 0.052±0.022  | 0.069±0.033| -0.013±0.168| 0.078±0.036 |
| Ala4  | 0.037±0.023  | 0.037±0.030| -0.112±0.165| 0.048±0.031 |
| Arg5  | 0.916±0.030  | 0.897±0.035| 1.016±0.085 | 0.886±0.046 |
| Cys6  | -0.002±0.016 | 0.000±0.023| 0.107±0.263 | -0.005±0.029|
| Gin7  | 0.030±0.021  | 0.027±0.020| -0.184±0.298| 0.039±0.026 |
| Leu8  | 0.034±0.033  | 0.047±0.036| 0.090±0.088 | 0.039±0.033 |
| Ser9  | -0.018±0.037 | -0.045±0.047| 0.009±0.119 | -0.076±0.055|
| Cys10 | -0.074±0.034 | -0.036±0.026| -0.196±0.140| -0.018±0.024|
| Lys11 | 0.916±0.041  | 0.859±0.059| 1.244±0.082 | 0.803±0.058 |
| Ser12 | -0.027±0.033 | -0.064±0.037| -0.142±0.398| -0.076±0.033|
| Leu13 | -0.014±0.030 | -0.020±0.037| 0.206±0.432 | -0.050±0.040|
| Gly14 | -0.040±0.029 | -0.027±0.023| -0.032±0.208| -0.015±0.028|
| Leu15 | -0.010±0.025 | 0.007±0.025 | 0.108±0.066 | -0.002±0.024|
| Lys16 | 0.980±0.034  | 0.936±0.054| 0.942±0.110 | 0.916±0.063 |
| Gly17 | -0.005±0.025 | -0.024±0.024| 0.058±0.127 | -0.037±0.019|
| Gly18 | 0.051±0.023  | 0.052±0.029 | 0.009±0.188 | 0.049±0.030 |
| Cys19 | -0.059±0.017 | -0.031±0.018| 0.081±0.159 | -0.011±0.022|
| Gln20 | 0.019±0.027  | 0.013±0.029 | -0.116±0.175| 0.012±0.025 |
| Gly21 | 0.066±0.020  | 0.106±0.034 | 0.019±0.163 | 0.134±0.033 |
| Ser22 | 0.027±0.029  | 0.040±0.021 | -0.102±0.140| 0.042±0.021 |
| Phe23 | 0.043±0.032  | 0.022±0.046 | 0.040±0.108 | 0.053±0.053 |
| Cys24 | -0.018±0.044 | 0.000±0.048 | -0.055±0.128| 0.015±0.044 |
| Thr25 | 0.032±0.019  | 0.040±0.027 | 0.059±0.495 | 0.019±0.033 |
| Cys26 | -0.101±0.028 | -0.109±0.030| 0.146±0.491 | -0.117±0.034|
| Gly27*| -0.860±0.020 | -0.656±0.037| -0.896±0.102| -0.625±0.046|
Table S8. The mean, charge transfer (difference of formal charge and mean), standard deviation (std.), min., and max. of the Mulliken by-residue-summed \( \tilde{q}(\text{RES}) \) distribution (in a.u.) from 85 ps of production AIMD for each residue in mini-CD4 including backbone atoms. The residue three letter code, index, identity (i.e., polar, charged, nonpolar), and expected formal charge of each amino acid are also listed. All Cys residues are in disulfide bridges and annotated as Cyx. Terminal residues are charged and indicated as such. We also summarize at the bottom the average of charge transfer values, average of the absolute of charge transfer values, and the charge distribution std. and range (max-min) over residues of four types: negatively charged (neg. charge), positively charged (pos. charge), polar, and nonpolar.

| index | residue | identity | Formal charge | mean | Charge transfer | std. | min | max |
|-------|---------|----------|---------------|------|-----------------|------|-----|-----|
| 1     | CYX     | charged  | 1             | 0.877| -0.123          | 0.045| 0.737| 1.027|
| 2     | ASN     | polar    | 0             | 0.137| 0.137           | 0.033| -0.005| 0.262|
| 3     | LEU     | nonpolar | 0             | 0.055| 0.055           | 0.028| -0.051| 0.167|
| 4     | ALA     | nonpolar | 0             | 0.038| 0.038           | 0.031| -0.091| 0.165|
| 5     | ARG     | charged  | 1             | 0.937| -0.663          | 0.051| 0.791| 1.109|
| 6     | CYX     | polar    | 0             | -0.004| -0.040        | 0.030| -0.115| 0.110|
| 7     | GLN     | polar    | 0             | 0.042| 0.042           | 0.056| -0.126| 0.242|
| 8     | LEU     | nonpolar | 0             | 0.015| 0.015           | 0.032| -0.114| 0.152|
| 9     | SER     | polar    | 0             | -0.004| -0.004        | 0.042| -0.178| 0.161|
| 10    | CYX     | polar    | 0             | -0.034| -0.034        | 0.031| -0.156| 0.088|
| 11    | LYS     | charged  | 1             | 0.935| -0.665          | 0.072| 0.667| 1.101|
| 12    | SER     | polar    | 0             | -0.040| -0.040        | 0.036| -0.170| 0.102|
| 13    | LEU     | nonpolar | 0             | -0.023| -0.023        | 0.029| -0.156| 0.095|
| 14    | GLY     | nonpolar | 0             | -0.049| -0.049        | 0.029| -0.156| 0.094|
| 15    | LEU     | nonpolar | 0             | -0.008| -0.008        | 0.037| -0.139| 0.155|
| 16    | LYS     | charged  | 1             | 0.862| -0.138          | 0.062| 0.624| 1.080|
| 17    | GLY     | nonpolar | 0             | -0.007| -0.007        | 0.032| -0.130| 0.116|
| 18    | GLY     | nonpolar | 0             | 0.048| 0.048           | 0.027| -0.054| 0.152|
| 19    | CYX     | polar    | 0             | -0.015| -0.015        | 0.033| -0.148| 0.120|
| 20    | GLN     | polar    | 0             | 0.000| 0.000           | 0.041| -0.158| 0.165|
| 21    | GLY     | nonpolar | 0             | 0.037| 0.037           | 0.031| -0.078| 0.161|
| 22    | SER     | polar    | 0             | 0.020| 0.020           | 0.032| -0.116| 0.144|
| 23    | PHE     | nonpolar | 0             | 0.065| 0.065           | 0.032| -0.059| 0.179|
| 24    | CYX     | polar    | 0             | -0.026| -0.026        | 0.027| -0.133| 0.090|
| 25    | THR     | polar    | 0             | -0.008| -0.008        | 0.039| -0.145| 0.163|
| 26    | CYX     | polar    | 0             | -0.059| -0.059        | 0.033| -0.201| 0.067|
| 27    | GLY     | charged  | 1             | -0.792| 0.208         | 0.062| -1.044| -0.630|

Summary of mean charge transfer and statistics averaged over residues

| Residue type | # | avg. | avg. of abs. | max. | min. | avg. std. | avg. range |
|--------------|---|------|--------------|------|-----|-----------|------------|
| neg. charge  | 1 | 0.208| 0.208        | --   | --  | 0.062     | 0.414      |
| pos. charge  | 4 | -0.097| 0.097       | -0.063| -0.138| 0.058     | 0.375      |
| polar        | 12| 0.001| 0.025        | 0.137| -0.059| 0.036     | 0.280      |
| nonpolar     | 10| 0.017| 0.035        | 0.065| -0.049| 0.031     | 0.246      |
Table S9. The mean, charge transfer (difference of formal charge and mean), standard deviation (std.), min., and max. of the Mulliken by-residue-summed q(RES) distribution (in a.u.) from 85 ps of production AIMD for each residue in benenodin-1 including backbone atoms. The residue three letter code, index, identity (i.e., polar, charged, nonpolar), and expected formal charge of each amino acid are also listed. Gly1 and Asp8 form an isopeptide bond, and they are annotated by *. Since they have a net formal charge of zero, these two residues are also classified as polar. The C-terminal residue is charged and indicated as such. We also summarize at the bottom the average of charge transfer value s, average of the absolute of charge transfer values, and the charge distribution std. and range (max-min) over residues of four types: negatively charged (neg. charge), positively charged (pos. charge), polar, and nonpolar.

| index | residue | identity | Formal charge | mean | Charge transfer | std. | min  | max  |
|-------|---------|----------|---------------|------|----------------|------|------|------|
| 1     | GLY*    | polar    | 0             | 0.012| -0.012         | 0.031| -0.103| 0.138|
| 2     | VAL     | nonpolar | 0             | 0.047| -0.047         | 0.029| -0.086| 0.175|
| 3     | GLY     | nonpolar | 0             | 0.045| -0.045         | 0.027| -0.060| 0.139|
| 4     | PHE     | nonpolar | 0             | -0.017| 0.017          | 0.031| -0.130| 0.118|
| 5     | GLY     | nonpolar | 0             | -0.054| 0.054          | 0.027| -0.162| 0.049|
| 6     | ARG     | charged  | 1             | 0.808| 0.192          | 0.060| 0.619| 1.041|
| 7     | PRO     | polar    | 0             | 0.051| -0.051         | 0.031| -0.104| 0.160|
| 8     | ASP*    | polar    | 0             | 0.010| -0.010         | 0.032| -0.118| 0.131|
| 9     | SER     | polar    | 0             | 0.003| -0.003         | 0.030| -0.110| 0.123|
| 10    | ILE     | nonpolar | 0             | -0.045| 0.045          | 0.028| -0.153| 0.063|
| 11    | LEU     | nonpolar | 0             | 0.006| -0.006         | 0.027| -0.121| 0.119|
| 12    | THR     | polar    | 0             | -0.028| 0.028         | 0.030| -0.148| 0.083|
| 13    | GLN     | polar    | 0             | -0.005| 0.005         | 0.067| -0.222| 0.175|
| 14    | GLU     | charged  | -1            | -0.903| -0.997       | 0.054| -1.074| -0.712|
| 15    | GLN     | polar    | 0             | -0.008| 0.008         | 0.059| -0.194| 0.165|
| 16    | ALA     | nonpolar | 0             | 0.017| -0.017        | 0.029| -0.116| 0.132|
| 17    | LYS     | charged  | 1             | 0.871| 0.129         | 0.059| 0.669| 1.114|
| 18    | PRO     | polar    | 0             | -0.011| 0.011        | 0.037| -0.140| 0.128|
| 19    | MET     | charged  | -1            | -0.799| -0.201      | 0.039| -0.960| -0.646|

Summary of mean charge transfer averaged over residues

| Residue type | # | avg. | avg. of abs. | max. | min. | avg. std. | avg. range |
|--------------|---|------|-------------|------|------|-----------|------------|
| neg. charge  | 2 | -0.149| -0.097       | -0.201| 0.047| 0.338     |
| pos. charge  | 2 | 0.161| 0.192       | 0.129| 0.060| 0.433     |
| polar        | 8 | -0.003| 0.016       | 0.028| -0.051| 0.040     | 0.280     |
| nonpolar     | 7 | 0.000| 0.054       | -0.047| 0.028| 0.232     |
Table S10. The mean, charge transfer (difference of formal charge and mean), standard deviation (std.), min., and max. of the Mulliken by-residue-summed $q$(RES) distribution (in a.u.) from 34 ps of production AIMD for each residue in Trp-cage including backbone atoms. The residue three-letter code, index, identity (i.e., polar, charged, nonpolar), and expected formal charge of each amino acid are also listed. The terminal residues are charged and indicated as such. We also summarize at the bottom the average of charge transfer values, average of the absolute of charge transfer values, and the charge distribution std. and range (max-min) over residues of four types: negatively charged (neg. charge), positively charged (pos. charge), polar, and nonpolar.

| index | residue | identity | Formal charge | mean | Charge transfer | std. | min | max |
|-------|---------|----------|---------------|------|-----------------|------|-----|-----|
| 1     | ASN     | charged  | 1             | 1.055 | -0.055          | 0.067 | 0.906 | 1.217 |
| 2     | LEU     | nonpolar | 0             | 0.086 | -0.086          | 0.045 | -0.062 | 0.216 |
| 3     | TYR     | polar    | 0             | 0.081 | -0.081          | 0.033 | -0.046 | 0.182 |
| 4     | ILE     | nonpolar | 0             | 0.026 | -0.026          | 0.032 | -0.083 | 0.139 |
| 5     | GLN     | polar    | 0             | -0.005 | 0.005           | 0.054 | -0.169 | 0.192 |
| 6     | TRP     | polar    | 0             | -0.011 | 0.011           | 0.030 | -0.146 | 0.111 |
| 7     | LEU     | nonpolar | 0             | -0.041 | 0.041           | 0.035 | -0.164 | 0.101 |
| 8     | LYS     | charged  | 1             | 0.920 | 0.080           | 0.050 | 0.700 | 1.105 |
| 9     | ASP     | charged  | -1            | -0.880 | -0.120          | 0.059 | -1.045 | -0.669 |
| 10    | GLY     | nonpolar | 0             | -0.021 | 0.021           | 0.036 | -0.137 | 0.103 |
| 11    | GLY     | nonpolar | 0             | 0.040 | -0.040          | 0.033 | -0.073 | 0.147 |
| 12    | PRO     | polar    | 0             | 0.016 | -0.016          | 0.040 | -0.131 | 0.175 |
| 13    | SER     | polar    | 0             | -0.048 | 0.048           | 0.043 | -0.204 | 0.096 |
| 14    | SER     | polar    | 0             | -0.074 | 0.074           | 0.043 | -0.242 | 0.069 |
| 15    | GLY     | nonpolar | 0             | -0.050 | 0.050           | 0.044 | -0.192 | 0.088 |
| 16    | ARG     | charged  | 1             | 0.904 | 0.096           | 0.049 | 0.723 | 1.078 |
| 17    | PRO     | polar    | 0             | 0.051 | -0.051          | 0.033 | -0.097 | 0.166 |
| 18    | PRO     | polar    | 0             | -0.019 | 0.019           | 0.037 | -0.166 | 0.117 |
| 19    | PRO     | polar    | 0             | -0.041 | 0.041           | 0.033 | -0.203 | 0.104 |
| 20    | SER     | charged  | -1            | -0.987 | -0.013          | 0.029 | -1.108 | -0.874 |

Summary of mean charge transfer and statistics averaged over residues

| Residue type | # | avg. of abs. | max. | min. | avg. std. | avg. range |
|--------------|---|--------------|------|------|-----------|------------|
| neg. charge  | 2 | -0.067       | -0.13 | -0.120 | 0.044    | 0.305     |
| pos. charge  | 3 | 0.040        | 0.077 | 0.096 | -0.055   | 0.055     |
| polar        | 9 | 0.006        | 0.038 | 0.074 | -0.081   | 0.039     |
| nonpolar     | 6 | -0.007       | 0.044 | -0.086 | 0.037    | 0.251     |
Table S11. Side-chain-only sums (excluding backbone C, O, N, H and N-terminal or C-terminal groups) for mini-CD4. The mean, charge transfer (formal charge-mean), standard deviation (std.), min., max., and range (max-min) of the distribution of by-sidechain-summed Mulliken partial charges (in a.u.) from 85 ps of production AIMD are shown for each amino acid in mini-CD4 with the abbreviation ‘SC’. The range of the full residue with backbone is also shown as ‘range BB’ for comparison. The residue three-letter code, index, and identity (i.e., polar, charged, nonpolar) are also listed. All Cys residues are in disulfide bridges and annotated as Cyx. Without the N-terminal or C-terminal groups, charge distributions of residues 1 and 27 are the most affected, and their residue type is changed to polar and nonpolar, respectively.

| Index | Residue | Identity | Mean SC | Std. SC | Min SC | Max SC | Range SC | Range BB |
|-------|---------|----------|---------|---------|--------|--------|----------|----------|
| 1     | CYX     | charged  | 0.331   | 0.024   | 0.238  | 0.420  | 0.181    | 0.290    |
| 2     | ASN     | polar    | 0.409   | 0.027   | 0.295  | 0.506  | 0.211    | 0.260    |
| 3     | LEU     | nonpolar | 0.359   | 0.019   | 0.286  | 0.444  | 0.158    | 0.215    |
| 4     | ALA     | nonpolar | 0.338   | 0.018   | 0.272  | 0.421  | 0.148    | 0.256    |
| 5     | ARG     | charged  | 1.252   | 0.042   | 1.127  | 1.389  | 0.263    | 0.318    |
| 6     | CYX     | polar    | 0.253   | 0.020   | 0.167  | 0.325  | 0.157    | 0.225    |
| 7     | GLN     | polar    | 0.352   | 0.048   | 0.214  | 0.533  | 0.319    | 0.368    |
| 8     | LEU     | nonpolar | 0.334   | 0.021   | 0.252  | 0.420  | 0.167    | 0.266    |
| 9     | SER     | polar    | -0.112  | 0.036   | -0.241 | 0.047  | 0.288    | 0.339    |
| 10    | CYX     | polar    | 0.246   | 0.021   | 0.165  | 0.331  | 0.166    | 0.244    |
| 11    | LYS     | charged  | 1.261   | 0.067   | 1.043  | 1.394  | 0.351    | 0.434    |
| 12    | SER     | polar    | -0.113  | 0.031   | -0.227 | -0.010 | 0.217    | 0.272    |
| 13    | LEU     | nonpolar | 0.342   | 0.018   | 0.258  | 0.400  | 0.141    | 0.251    |
| 14    | GLY     | nonpolar | 0.487   | 0.015   | 0.430  | 0.555  | 0.125    | 0.250    |
| 15    | LEU     | nonpolar | 0.323   | 0.021   | 0.245  | 0.401  | 0.155    | 0.294    |
| 16    | LYS     | charged  | 1.196   | 0.059   | 0.996  | 1.393  | 0.397    | 0.456    |
| 17    | GLY     | nonpolar | 0.513   | 0.017   | 0.445  | 0.571  | 0.125    | 0.238    |
| 18    | GLY     | nonpolar | 0.510   | 0.014   | 0.457  | 0.568  | 0.111    | 0.206    |
| 19    | CYX     | polar    | 0.284   | 0.022   | 0.177  | 0.376  | 0.199    | 0.268    |
| 20    | GLN     | polar    | 0.329   | 0.034   | 0.203  | 0.450  | 0.247    | 0.323    |
| 21    | GLY     | nonpolar | 0.498   | 0.014   | 0.444  | 0.548  | 0.104    | 0.239    |
| 22    | SER     | polar    | -0.070  | 0.019   | -0.161 | 0.010  | 0.171    | 0.260    |
| 23    | PHE     | nonpolar | 0.346   | 0.024   | 0.252  | 0.438  | 0.186    | 0.238    |
| 24    | CYX     | polar    | 0.267   | 0.020   | 0.181  | 0.342  | 0.161    | 0.223    |
| 25    | THR     | polar    | 0.311   | 0.034   | 0.207  | 0.422  | 0.215    | 0.301    |
| 26    | CYX     | polar    | 0.271   | 0.023   | 0.188  | 0.360  | 0.172    | 0.268    |
| 27    | GLY     | charged  | 0.450   | 0.019   | 0.368  | 0.516  | 0.148    | 0.414    |
Figure S2. Parity plots of distribution properties for by-residue charges for only the sidechain (SC) vs with backbone atom (BB) for the 27 residues in mini-CD4: (left) std. properties and (right) range properties, both in a.u.. The N-terminal and C-terminal residues are shown as open circles and excluded from the linear regression fit (shown as a red dashed line), whereas the remaining symbols are shown filled. A black dotted parity line is also shown. The $R^2$ value for both properties is shown in the inset bottom right.
Figure S3. Normalized histograms of the by-residue-summed Mulliken partial charges, $q_{\text{RES}}$ (in e), for the nonpolar residues (i.e., Ala4, Gly14, Gly17, Gly18, Gly21, Leu3, Leu8, Leu13, Leu15, and Phe23) in mini-CD4 obtained over 85 ps of production MD. The range of all graphs is $[-0.2 \, \text{e}, 0.2 \, \text{e}]$, the bin width of the histograms is 0.01 e, and a zero value for the charge is indicated as a black vertical bar. A consistent set of coloring for all nonpolar residues is used, as indicated in the legend.
Figure S4. Normalized histograms of the by-residue-summed Mulliken partial charges, $q_{RES}$ (in e), for the nonpolar residues (i.e., Ala16, Gly3, Gly5, Ile10, Leu11, Phe4, and Val2) in benenodin-1 obtained over 85 ps of production MD. The range of all graphs is [-0.2 e,0.2 e], the bin width of the histograms is 0.01 e, and a zero value for the charge is indicated as a black vertical bar. A consistent set of coloring for all nonpolar residues is used, as indicated in the legend.
Figure S5. Normalized histograms of the by-residue-summed Mulliken partial charges, $q_{RES}$ (in e), for the nonpolar residues (i.e., Gly10, Gly11, Gly15, Ile4, Leu2, and Leu7) in Trp-cage obtained over 34 ps of production MD. The range of all graphs is [-0.2 e, 0.2 e], the bin width of the histograms is 0.01 e, and a zero value for the charge is indicated as a black vertical bar. A consistent set of coloring for all nonpolar residues is used, as indicated in the legend.
**Figure S6.** Normalized histograms of the by-residue-summed Mulliken partial charges, $q_{\text{RES}}$ (in e), for the polar residues (i.e., Asn2, Cyx6, Cyx10, Cyx19, Cyx24, Cyx26, Gln7, Gln20, Ser9, Ser12, Ser22, and Thr25) in mini-CD4 obtained over 85 ps of production MD. The range of all graphs is $[-0.3 \text{ e}, 0.3 \text{ e}]$ (i.e., larger than for the nonpolar residues), the bin width of the histograms is 0.01 e, and a zero value for the charge is indicated as a black vertical bar. A consistent set of coloring for all polar residues is used, as indicated in the legend. The designation Cyx refers to a cysteine residue participating in a disulfide bond.
Figure S7. Normalized histograms of the by-residue-summed Mulliken partial charges, $q_{\text{RES}}$ (in e), for the polar residues (i.e., Gln13, Gln15, Pro7, Pro18, Ser9, and Thr12) in benenodin-1 as well as the two polar (i.e., uncharged) residues participating in the isopeptide bond (i.e., Asp8 and Gly1, indicated with an asterisk) obtained over 85 ps of production MD. The range of all graphs is [-0.3 e, 0.3 e] (i.e., larger than for the nonpolar residues), the bin width of the histograms is 0.01 e, and a zero value for the charge is indicated as a black vertical bar. A consistent set of coloring for all polar residues is used, as indicated in the legend.
Figure S8. Normalized histograms of the by-residue-summed Mulliken partial charges, $q_{RES}$ (in e), for the polar residues (i.e., Gln5, Pro12, Pro17, Pro18, Pro19, Ser13, Ser14, Trp6, and Tyr3) in Trp-cage obtained over 34 ps of production MD. The range of all graphs is [-0.3 e, 0.3 e] (i.e., larger than for the nonpolar residues), the bin width of the histograms is 0.01 e, and a zero value for the charge is indicated as a black vertical bar. A consistent set of coloring for all polar residues is used, as indicated in the legend.
Figure S9. Normalized histogram of the by-residue-summed Mulliken partial charges, $q_{RES}$ (in e), for the positively charged residues (i.e., Arg5, Lys11, Lys16) as well as the charged terminal (indicated by *) Cyx1 (N-terminus) and Gly27 (C-terminus) residues in mini-CD4 obtained over 85 ps of production MD. The range of positively charged residue graphs is [0.65 e, 1.1 e] and is [-1.1 e, -0.65 e] for negatively charged residues (i.e., larger than for the polar or nonpolar residues), the bin width of the histograms is 0.01 e, and the expected formal charge value for the residue is indicated as a black vertical bar (i.e., at +1 or -1 for positively and negatively charged residues, respectively). A consistent set of coloring for all charged residues is used, with the N-terminal and C-terminal residues colored orange and purple, respectively, regardless of sidechain identity, as indicated in the legend.
Figure S10. Normalized histograms of the by-residue-summed Mulliken partial charges, $q_{RES}$ (in e), for the positively charged residues (i.e., Arg6, Lys17, and Glu14) as well as the charged terminal (indicated by *) Met19 (C-terminus) residue in benenodin-1 obtained over 85 ps of production MD. The N-terminal residue in benenodin-1 participates in the isopeptide bond and is not charged. The range of positively charges residue graphs is [0.65 e, 1.1 e] and is [-1.1 e, -0.65 e] for negatively charged residues (i.e., larger than for the polar or nonpolar residues), the bin width of the histograms is 0.01 e, and the expected formal charge value for the residue is indicated as a black vertical bar (i.e., at +1 or -1 for positively and negatively charged residues, respectively). A consistent set of coloring for all charged residues is used, with the N-terminal and C-terminal residues colored orange and purple, respectively, regardless of sidechain identity, as indicated in the legend.
Figure S11. Normalized histograms of the by-residue-summed Mulliken partial charges, $q_{\text{RES}}$ (in e), for the positively or negatively charged residues (i.e., Lys8, Arg16, Asp9) as well as the charged terminal (indicated by *) Asn1 (N-terminus) and Ser20 (C-terminus) residues in Trp-cage obtained over 34 ps of production MD. The N-terminal residue in benenodin-1 participates in the isopeptide bond and is not charged. The range of positively charges residue graphs is [0.65 e, 1.1 e] and is [-1.1 e, -0.65 e] for negatively charged residues (i.e., larger than for the polar or nonpolar residues), the bin width of the histograms is 0.01 e, and the expected formal charge value for the residue is indicated as a black vertical bar (i.e., at +1 or -1 for positively and negatively charged residues, respectively). A consistent set of coloring for all charged residues is used, with the N-terminal and C-terminal residues colored orange and purple, respectively, regardless of sidechain identity, as indicated in the legend.
### Table S12. Summary of mini-CD4 overall CC values above magnitude thresholds indicated at top along with percentages in each type: overall, nearest neighbors, and for specific residue interaction types and counts indicated in inset headings.

|                | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 |
|----------------|-----|-----|-----|-----|-----|-----|-----|
| **All**        | 351 |     |     |     |     |     |     |
| neg.           | 217 | 66  | 31  | 15  | 4   | 3   | 1   |
| pos.           | 134 | 29  | 5   | 1   | 0   | 0   | 0   |
| neg. %         | 61.8 | 18.8 | 8.8 | 4.3 | 1.1 | 0.9 | 0.3 |
| pos. %         | 38.2 | 8.3  | 1.4 | 0.3 | 0.0 | 0.0 | 0.0 |
| **Nearest neighbors** | 26 |     |     |     |     |     |     |
| neg.           | 22  | 17  | 12  | 5   | 1   | 1   | 0   |
| pos.           | 4   | 1   | 0   | 0   | 0   | 0   | 0   |
| neg. %         | 84.6 | 65.4 | 46.2 | 19.2 | 3.8 | 3.8 | 0.0 |
| pos. %         | 15.4 | 3.8  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| **Non-NN**     | 325 |     |     |     |     |     |     |
| neg.           | 195 | 49  | 19  | 10  | 3   | 2   | 1   |
| pos.           | 130 | 28  | 5   | 1   | 0   | 0   | 0   |
| neg. %         | 62.5 | 18.4 | 3.9 | 3.3 | 2.0 | 0.0 | 0.0 |
| pos. %         | 37.5 | 8.6  | 2.0 | 0.0  | 0.0  | 0.0  | 0.0  |
| **Charged-charged** | 10 |     |     |     |     |     |     |
| neg.           | 7   | 4   | 2   | 1   | 1   | 1   | 1   |
| pos.           | 3   | 3   | 1   | 1   | 1   | 0   | 0   |
| pos %          | 70.0 | 40.0 | 20.0 | 10.0 | 10.0 | 10.0 | 10.0  |
| neg %          | 30.0 | 30.0 | 10.0 | 10.0 | 0.0  | 0.0  | 0.0  |
| **Charged-polar** | 35 |     |     |     |     |     |     |
| neg.           | 23  | 10  | 7   | 4   | 1   | 0   | 0   |
| pos.           | 12  | 4   | 1   | 0   | 0   | 0   | 0   |
| pos %          | 65.7 | 28.6 | 20.0 | 11.4 | 2.9 | 0.0  | 0.0  |
| neg %          | 34.3 | 71.4 | 2.9 | 0.0  | 0.0  | 0.0  | 0.0  |
| **Charged-nonpolar** | 50 |     |     |     |     |     |     |
| neg.           | 30  | 16  | 6   | 4   | 1   | 1   | 0   |
| pos.           | 20  | 8   | 3   | 0   | 0   | 0   | 0   |
| pos %          | 60.0 | 32.0 | 12.0 | 8.0 | 2.0 | 2.0 | 0.0  |
| neg %          | 40.0 | 68.0 | 6.0 | 0.0  | 0.0  | 0.0  | 0.0  |
| **Polar-polar** | 21 |     |     |     |     |     |     |
| neg.           | 11  | 3   | 1   | 0   | 0   | 0   | 0   |
| pos.           | 10  | 1   | 0   | 0   | 0   | 0   | 0   |
| pos %          | 52.4 | 14.3 | 4.8 | 0.0  | 0.0  | 0.0  | 0.0  |
| neg %          | 47.6 | 85.7 | 5.2 | 0.0  | 0.0  | 0.0  | 0.0  |
| **Polar-nonpolar** | 70 |     |     |     |     |     |     |
| neg.           | 44  | 11  | 5   | 2   | 0   | 0   | 0   |
| pos.           | 26  | 5   | 0   | 0   | 0   | 0   | 0   |
| pos %          | 62.9 | 15.7 | 7.1 | 2.9 | 0.0  | 0.0  | 0.0  |
| neg %          | 37.1 | 84.3 | 7.1 | 0.0  | 0.0  | 0.0  | 0.0  |
| **Nonpolar-nonpolar** | 45 |     |     |     |     |     |     |
| neg.           | 25  | 4   | 2   | 0   | 0   | 0   | 0   |
| pos.           | 20  | 1   | 0   | 0   | 0   | 0   | 0   |
| pos %          | 55.6 | 8.9  | 4.4 | 0.0  | 0.0  | 0.0  | 0.0  |
| neg %          | 44.4 | 91.1 | 9.6 | 0.0  | 0.0  | 0.0  | 0.0  |
| **Disulfide-disulfide** | 15 |     |     |     |     |     |     |
| neg.           | 13  | 2   | 0   | 0   | 0   | 0   | 0   |
| pos.           | 2   | 1   | 0   | 0   | 0   | 0   | 0   |
| pos %          | 86.7 | 13.3 | 0.0 | 0.0  | 0.0  | 0.0  | 0.0  |
| neg %          | 13.3 | 8.7  | 0.0 | 0.0  | 0.0  | 0.0  | 0.0  |
Table S13. Summary of benedolin-I overall CC values above magnitude thresholds indicated at top along with percentages in each type: overall, nearest neighbors, and for specific residue interaction types and counts indicated in inset headings.

|       | 0.0| 0.1| 0.2| 0.3| 0.4| 0.5| 0.6|
|-------|----|----|----|----|----|----|----|
| All   | 171|    |    |    |    |    |    |
| neg   | 112| 43 | 17 | 10 | 3  | 0  | 0  |
| pos   | 59 | 13 | 3  | 0  | 0  | 0  | 0  |
|       | 66.5%| 25.1%| 9.9%| 5.8%| 1.8%| 0.0%| 0.0%|
| Nearest neighbors | 19|    |    |    |    |    |    |
| neg   | 17 | 15 | 11 | 5  | 0  | 0  | 0  |
| pos   | 2  | 0  | 0  | 0  | 0  | 0  | 0  |
|       | 89.5%| 78.9%| 57.9%| 26.3%| 0.0%| 0.0%| 0.0%|
| Non-NN | 152|    |    |    |    |    |    |
| neg   | 95 | 28 | 6  | 5  | 3  | 0  | 0  |
| pos   | 57 | 13 | 3  | 0  | 0  | 0  | 0  |
|       | 62.5%| 18.4%| 3.9%| 3.3%| 2.0%| 0.0%| 0.0%|
| Charged-charged | 6|    |    |    |    |    |    |
| neg   | 5  | 4  | 2  | 2  | 1  | 0  | 0  |
| pos   | 1  | 1  | 0  | 0  | 0  | 0  | 0  |
|       | 88.3%| 66.7%| 33.3%| 33.3%| 16.7%| 0.0%| 0.0%|
| Charged-nonpolar | 28|    |    |    |    |    |    |
| neg   | 18 | 9  | 6  | 3  | 1  | 0  | 0  |
| pos   | 14 | 5  | 0  | 0  | 0  | 0  | 0  |
|       | 56.3%| 28.1%| 18.8%| 9.4%| 3.1%| 0.0%| 0.0%|
| Polar-polar | 28|    |    |    |    |    |    |
| neg   | 20 | 4  | 0  | 0  | 0  | 0  | 0  |
| pos   | 8  | 1  | 1  | 0  | 0  | 0  | 0  |
|       | 71.4%| 14.3%| 0.0%| 0.0%| 0.0%| 0.0%| 0.0%|
| Polar-nonpolar | 56|    |    |    |    |    |    |
| neg   | 33 | 13 | 3  | 3  | 0  | 0  | 0  |
| pos   | 23 | 2  | 0  | 0  | 0  | 0  | 0  |
|       | 58.9%| 23.2%| 5.4%| 5.4%| 0.0%| 0.0%| 0.0%|
| Nonpolar-nonpolar | 21|    |    |    |    |    |    |
| neg   | 16 | 4  | 2  | 0  | 0  | 0  | 0  |
| pos   | 5  | 0  | 0  | 0  | 0  | 0  | 0  |
|       | 76.4%| 19.0%| 9.5%| 0.0%| 0.0%| 0.0%| 0.0%|
| neg   | 23.8%| 0.0%| 0.0%| 0.0%| 0.0%| 0.0%| 0.0%|
Table S14. Summary of Trp-cage overall CC values above magnitude thresholds indicated at top along with percentages in each type: overall, nearest neighbors, and for specific residue interaction types and counts indicated in inset headings.

|       | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 |
|-------|-----|-----|-----|-----|-----|-----|-----|
| **All** |     |     |     |     |     |     |     |
| neg   | 111 | 67  | 36  | 13  | 7   | 4   | 3   |
| pos   | 79  | 35  | 12  | 1   | 0   | 0   | 0   |
|       | 58.4% | 35.3% | 18.9% | 6.8% | 3.7% | 2.1% | 1.6% |
|       | 41.6% | 18.4% | 6.3%  | 0.5% | 0.0% | 0.0% | 0.0% |
| **Nearest neighbors** |     |     |     |     |     |     |     |
| neg   | 19  | 19  | 17  | 11  | 7   | 4   | 3   |
| pos   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |
|       | 100.0% | 100.0% | 89.5% | 57.9% | 36.8% | 21.1% | 15.8% |
|       | 0.0%  | 0.0%  | 0.0%  | 0.0%  | 0.0%  | 0.0%  | 0.0%  |
| **Non-NN** |     |     |     |     |     |     |     |
| neg   | 92  | 48  | 19  | 2   | 0   | 0   | 0   |
| pos   | 79  | 35  | 12  | 1   | 0   | 0   | 0   |
|       | 53.8% | 28.1% | 11.1% | 1.2%  | 0.0%  | 0.0%  | 0.0%  |
|       | 46.2% | 20.5% | 7.0%  | 0.6%  | 0.0%  | 0.0%  | 0.0%  |
| **Charged-charged** |     |     |     |     |     |     |     |
| neg   | 4   | 3   | 2   | 1   | 0   | 0   | 0   |
| pos   | 6   | 1   | 0   | 0   | 0   | 0   | 0   |
| pos % | 40.0% | 30.0% | 20.0% | 10.0% | 0.0%  | 0.0%  | 0.0%  |
| neg % | 60.0% | 10.0% | 10.0% | 0.0%  | 0.0%  | 0.0%  | 0.0%  |
| **Charged-polar** |     |     |     |     |     |     |     |
| neg   | 30  | 20  | 8   | 2   | 2   | 1   | 1   |
| pos   | 15  | 8   | 2   | 0   | 0   | 0   | 0   |
| pos % | 66.7% | 44.4% | 17.8% | 4.4%  | 4.4%  | 2.2%  | 2.2%  |
| neg % | 33.3% | 17.8% | 4.4%  | 0.0%  | 0.0%  | 0.0%  | 0.0%  |
| **Charged-nonpolar** |     |     |     |     |     |     |     |
| neg   | 17  | 11  | 8   | 3   | 2   | 1   | 1   |
| pos   | 13  | 8   | 4   | 1   | 0   | 0   | 0   |
| pos % | 56.7% | 36.7% | 26.7% | 10.0% | 6.7%  | 3.3%  | 3.3%  |
| neg % | 43.3% | 63.7% | 33.3% | 3.3%  | 0.0%  | 0.0%  | 0.0%  |
| **Polar-polar** |     |     |     |     |     |     |     |
| neg   | 21  | 11  | 7   | 3   | 2   | 1   | 0   |
| pos   | 15  | 4   | 0   | 0   | 0   | 0   | 0   |
| pos % | 58.3% | 30.6% | 19.4% | 8.3%  | 5.6%  | 2.8%  | 0.0%  |
| neg % | 41.7% | 11.1% | 0.0%  | 0.0%  | 0.0%  | 0.0%  | 0.0%  |
| **Polar-nonpolar** |     |     |     |     |     |     |     |
| neg   | 29  | 15  | 7   | 3   | 1   | 1   | 1   |
| pos   | 25  | 11  | 4   | 0   | 0   | 0   | 0   |
| pos % | 53.7% | 27.8% | 13.0% | 5.6%  | 1.9%  | 1.9%  | 1.9%  |
| neg % | 46.3% | 72.8% | 87.0% | 94.4% | 98.1% | 99.1% | 100.0% |
| **Nonpolar-nonpolar** |     |     |     |     |     |     |     |
| neg   | 10  | 7   | 4   | 1   | 0   | 0   | 0   |
| pos   | 5   | 3   | 1   | 0   | 0   | 0   | 0   |
| pos % | 66.7% | 46.7% | 26.7% | 6.7%  | 0.0%  | 0.0%  | 0.0%  |
| neg % | 33.3% | 53.3% | 73.3% | 93.3% | 99.4% | 99.4% | 100.0% |
Table S15. Overall average, maximum, and minimum cross-correlation absolute magnitude (i.e., unsigned) for each protein grouped by type. Pro residues are included in polar. For mini-CD4, only the charged Cys1 is included in analysis for the standard residue types, and that residue and the remaining Cys residues are included in the Cys-Cys category shown at the bottom of the table, meaning that CC values are summarized for only 246 of the 351 possible results in mini-CD4.

|                | mini-CD4 | benenodin-1 | Trp-cage |
|----------------|----------|-------------|----------|
| **charged-charged** |          |             |          |
| number         | 10       | 6           | 10       |
| average        | 0.1886   | 0.1169      | 0.1276   |
| max            | 0.6037   | 0.4831      | 0.3962   |
| min            | 0.0023   | 0.0072      | 0.0007   |
| **charged-polar** |         |             |          |
| number         | 35       | 32          | 45       |
| average        | 0.1262   | 0.1136      | 0.1420   |
| max            | 0.4329   | 0.4092      | 0.6285   |
| min            | 0.0050   | 0.0066      | 0.0027   |
| **charged-nonpolar** |       |             |          |
| number         | 50       | 28          | 30       |
| average        | 0.1186   | 0.0755      | 0.1811   |
| max            | 0.5060   | 0.2668      | 0.8052   |
| min            | 0.0011   | 0.0013      | 0.0136   |
| **polar-polar** |          |             |          |
| number         | 21       | 28          | 36       |
| average        | 0.0670   | 0.1132      | 0.1171   |
| max            | 0.2449   | 0.4593      | 0.5020   |
| min            | 0.0044   | 0.0013      | 0.0014   |
| **polar-nonpolar** |        |             |          |
| number         | 70       | 56          | 54       |
| average        | 0.0799   | 0.0788      | 0.1342   |
| max            | 0.3938   | 0.3828      | 0.6066   |
| min            | 0.0024   | 0.0002      | 0.0157   |
| **nonpolar-nonpolar** |     |             |          |
| number         | 45       | 21          | 15       |
| average        | 0.0547   | 0.0742      | 0.1404   |
| max            | 0.2533   | 0.2962      | 0.3523   |
| min            | 0.0026   | 0.0072      | 0.0011   |
| **Cys-Cys**    |          |             |          |
| number         | 15       | 0           | 0        |
| average        | 0.052    | --          | --       |
| max            | 0.193    | --          | --       |
| min            | 0.002    | --          | --       |
| **total**      | 246      | 171         | 190      |
**Table S16.** Cross-correlation values between Cys residues in mini-CD4. The maximum absolute value and minimum absolute value are also shown. The CC values that correspond to disulfide-linked residue pairs are shown in red.

| Cys1  | Cys6   | Cys10  | Cys19  | Cys24  | Cys26  |
|-------|--------|--------|--------|--------|--------|
| Cyx1  | -0.0054| 0.1361 | -0.0783| -0.0394| -0.1927|
| Cyx6  | -0.0639| -0.0186| -0.0020| -0.0089|
| Cyx10 | 0.1361 | -0.0639| 0.0295 | -0.0224| -0.1000|
| Cyx19 | -0.0783| 0.0186 | 0.0614 | 0.0213 |
| Cyx24 | -0.0394| -0.0224| -0.0614| -0.0062|
| Cyx26 | -0.1927| 0.0639 | 0.1361 | 0.0783 | 0.0614 | 0.1927|

max(abs) 0.1927 0.0639 0.1361 0.0783 0.0614 0.1927

min(abs) 0.0054 0.0020 0.0224 0.0186 0.0062 0.0062

**Table S17.** The 16 residue pairs in mini-CD4 with cross-correlation absolute values above |0.30|, with the residue 3 letter code and number shown along with the absolute value of the CC, the rank of that coupling for each residue, their distance in primary sequence, whether the residues are nearest neighbors, the type of each residue (terminal residues are charged), and the secondary sequence (SS) element that each residue belongs to in the protein.

| Res 1 | Res 2 | |CC| |Rank #1| |Rank #2| |seq. diff.| |NN?| |type 1| |type 2| |SS 1| |SS 2|
|-------|-------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lys16 | Gly27 | 0.6037 | 1 | 1 | | | | | | | | | | | |
| Leu15 | Gly27 | 0.5060 | 1 | 2 | | | | | | | | | | | |
| Thr25 | Cyx26 | 0.5006 | 1 | 1 | 1 | | | | | | | | | | |
| Cyx1  | Ser22 | 0.4329 | 1 | 1 | 21 | | | | | | | | | | |
| Cyx19 | Gln20 | 0.3963 | 1 | 1 | 1 | | | | | | | | | | |
| Asn2  | Leu3  | 0.3938 | 1 | 1 | 1 | | | | | | | | | | |
| Cyx1  | Gly21 | 0.3855 | 2 | 1 | 20 | | | | | | | | | | |
| Cyx1  | Phe23 | 0.3765 | 3 | 1 | 22 | | | | | | | | | | |
| Cyx10 | Lys11 | 0.3628 | 1 | 1 | 1 | | | | | | | | | | |
| Ser9  | Lys11 | 0.3589 | 1 | 2 | 2 | | | | | | | | | | |
| Gln7  | Gly17 | 0.3439 | 1 | 1 | 10 | | | | | | | | | | |
| Arg5  | Ser9  | 0.3334 | 1 | 2 | 4 | | | | | | | | | | |
| Gly14 | Gly27 | 0.3271 | 1 | 3 | 13 | | | | | | | | | | |
| Arg5  | Lys11 | 0.3143 | 2 | 3 | 6 | | | | | | | | | | |
| Arg5  | Cyx6  | 0.3111 | 3 | 1 | 1 | | | | | | | | | | |
| Gln7  | Lys11 | 0.3016 | 2 | 4 | 4 | | | | | | | | | | |

Page S27
Table S18. The 10 residue pairs in benenodin-1 with cross-correlation absolute values above |0.30|, with the residue 3 letter code and number shown along with the absolute value of the CC, the rank of that coupling for each residue, their distance in primary sequence, whether the residues are nearest neighbors, the type of each residue (terminal residues are charged), and the secondary sequence (SS) element that each residue belongs to in the protein.

| Res 1 | Res 2 | |CC| |Rank #1|Rank #2|seq. diff.|NN?|type 1|type 2|SS 1|SS 2 |
|-------|-------|---|---|---|---|---|---|---|---|---|---|---|
|Lys17  |Met19  |0.4831|1|1|2| N |charged |terminal | tail | C-terminus |
|Gln13  |Gln15  |0.4593|1|1|2| N |polar |polar |tail |tail |
|Arg6   |Gln13  |0.4092|1|2|7| N |charged |polar |ring |tail |
|Gly1   |Val2   |0.3828|1|1|1|Y |terminal |polar |nonpolar |ring |ring |
|Ser9   |Ile10  |0.3766|1|1|1|Y |charged |polar |nonpolar |tail |tail |
|Arg6   |Lys17  |0.3615|2|2|11|N |charged |charged |ring |tail |
|Thr12  |Gly15  |0.3446|3|1|12|N |charged |polar |polar |tail |tail |
|Leu11  |Pro18  |0.3213|3|2|Y |charged |nonpolar |tail |tail |

Table S19. The 14 residue pairs in Trp-cage with cross-correlation absolute values above |0.30|, with the residue 3 letter code and number shown along with the absolute value of the CC, the rank of that coupling for each residue, their distance in primary sequence, whether the residues are nearest neighbors, the type of each residue (terminal residues are charged), and the secondary sequence (SS) element that each residue belongs to in the protein.

| Res 1 | Res 2 | |CC| |Rank #1|Rank #2|seq. diff.|NN?|type 1|type 2|SS 1|SS 2 |
|-------|-------|---|---|---|---|---|---|---|---|---|---|---|
|Asn1   |Leu2   |0.8052|1|1|1|Y |terminal |nonpolar |N-terminus |alpha helix |
|Pro19  |Ser20  |0.6285|1|1|1|Y |nonpolar |terminal |tail |C-terminus |
|Ser14  |Gly15  |0.6066|1|1|1|Y |polar |nonpolar |3-10 helix |tail |
|Pro12  |Ser13  |0.5020|1|1|1|Y |nonpolar |polar |3-10 helix |3-10 helix |
|Asp9   |Gly10  |0.4550|1|1|1|Y |charged |nonpolar |turn |turn |
|Arg16  |Pro17  |0.4170|1|1|1|Y |charged |nonpolar |tail |tail |
|Pro18  |Pro19  |0.4016|1|2|1|Y |nonpolar |nonpolar |tail |tail |
|Asn1   |Asp9   |0.3962|2|2|8|N |terminal |charged |N-terminus |turn |
|Ser13  |Ser14  |0.3801|2|2|1|Y |polar |polar |3-10 helix |3-10 helix |
|Gly10  |Gly11  |0.3523|2|1|1|Y |nonpolar |nonpolar |turn |3-10 helix |
|Ile4   |Gln5   |0.3361|1|1|1|Y |nonpolar |polar |alpha helix |alpha helix |
|Gly15  |Arg16  |0.3112|2|2|1|Y |nonpolar |charged |tail |tail |
|Gln5   |Gly11  |0.3101|2|2|6|N |polar |nonpolar |alpha helix |3-10 helix |
|Leu2   |Asp9   |0.3019|2|3|7|N |nonpolar |charged |alpha helix |turn |
Table S20. Counts of absolute CC values above specific thresholds for polar Gln, Ser, and Thr residues in mini-CD4, benenodin-1, and Trp-cage.

| Protein  | Residue | CC > |0.3| | CC > |0.2| | CC > |0.2| (non-NN) |
|----------|---------|------|-----|-----|------|-----|------|-----|--------|
| mini-CD4 | Asn2    | 1    | 1   | 0   |      |     |      |     |
| mini-CD4 | Gln7    | 2    | 4   | 3   |      |     |      |     |
| mini-CD4 | Ser9    | 2    | 2   | 2   |      |     |      |     |
| mini-CD4 | Gln20   | 1    | 6   | 4   |      |     |      |     |
| mini-CD4 | Thr25   | 1    | 3   | 1   |      |     |      |     |
| benenodin-1 | Ser9 | 1    | 4   | 2   |      |     |      |     |
| benenodin-1 | Thr12 | 2    | 2   | 0   |      |     |      |     |
| benenodin-1 | Gln13 | 3    | 6   | 4   |      |     |      |     |
| benenodin-1 | Gln15 | 1    | 2   | 0   |      |     |      |     |
| Trp-cage  | Gln5    | 2    | 8   | 7   |      |     |      |     |
| Trp-cage  | Ser13   | 2    | 4   | 2   |      |     |      |     |
| Trp-cage  | Ser14   | 2    | 7   | 5   |      |     |      |     |

Text S1. Description of linear and nonlinear MI contributions.

In addition to CC, we compute the mutual information (MI)\(^8\)\(^-\)\(^10\) between the probability distributions, \(p\), of \(q(J)\) and \(q(K)\) for residues \(J\) and \(K\) as:

\[
I(J;K) = \sum_j \sum_k p(J,K)(j,k) \ln \left( \frac{p(J,K)(j,k)}{p(J)p(K)} \right)
\]  

(1)

To delineate the contribution of nonlinear terms to the MI, we can approximately decompose it into a linear component\(^11\)\(^-\)\(^12\) derived from the CC (i.e., \(r_{JK}\)) as:

\[
I(J;K)_{\text{linear}} = -\frac{1}{2} \ln \left( 1 - r_{JK}^2 \right)
\]  

(2)

The remaining nonlinear MI terms can then be estimated as the difference between the linear approximation and the total MI:

\[
I(J;K)_{\text{nonlinear}} = I(J;K)_{\text{total}} - I(J;K)_{\text{linear}}
\]  

(3)

The linear MI term makes the sole contribution to the MI if both charge residue sums follow normal bivariate distributions.\(^12\)
**Figure S12.** The CC vs MI in mini-CD4 (gray circles), benenodin-1 (red circles), and Trp-cage (green circles). These values are compared to the assumption for a purely linear MI expression (black line). Significant deviations from the linear approximation are observed, especially for cases with positive CC values.

**Figure S13.** Comparison of the percentile rank for the absolute value of the CC vs. the total MI for mini-CD4 (left), benenodin-1 (middle), and Trp-cage (right) evaluated for all residue pairs within each protein (i.e., from 0 to 100\(^{th}\)). A gray dashed parity line is shown along with the Spearman’s rank (SRCC) and the Pearson’s \( r \) correlation (i.e., for the actual values, not percentile) for the two quantities. Any points with percentile ranks that differ more than 25% between the two quantities are shown in red.
Figure S14. Normalized histograms of the total MI (left), nonlinear contribution to the MI (middle), and linear contribution (right) for all residue couplings in mini-CD4. A bin width of 0.005 is used for all graphs. The integrated count each histogram is shown as a black dotted line. The value corresponding to the top 10% of couplings is indicated by a green dashed line.

Figure S15. Normalized histograms of the total MI (left), nonlinear contribution to the MI (middle), and linear contribution (right) for all residue couplings in benenodin-1. A bin width of 0.005 is used for all graphs. The integrated count of each histogram is shown as a black dotted line. The value corresponding to the top 10% of couplings is indicated by a green dashed line.
Figure S16. Normalized histograms of the total MI (left), nonlinear contribution to the MI (middle), and linear contribution (right) for all residue couplings in Trp-cage. A bin width of 0.005 is used for all graphs. The integrated count of each histogram is shown as a black dotted line. The value corresponding to the top 10% of couplings is indicated by a green dashed line.
**Table S21.** Residue pairs in mini-CD4 for which the MI % rank is > 25% **below** the |CC| % rank (i.e., the CC is stronger than the MI). The MI and CC values, % rank for each, the difference, residue type, and secondary structure type are shown.

| Res1  | Res2  | MI % rank | MI % rank | | % diff | type 1     | type 2     | SS1     | SS2     |
|-------|-------|-----------|-----------|-------|---------|-----------|----------|---------|---------|
| Cxy24 | Gly18 | 0.010     | 16.0%     | -0.084| 67.0%   | polar     | nonpolar  | beta sheet| beta sheet|
| Leu13 | Asn2  | 0.009     | 7.7%      | -0.058| 52.4%   | -44.7%    | polar     | loop     | alpha helix|
| Leu13 | Cxy10 | 0.009     | 10.8%     | 0.063 | 55.6%   | -44.7%    | nonpolar  | loop     | alpha helix|
| Gly17 | Cxy6  | 0.012     | 32.5%     | 0.106 | 74.4%   | -41.9%    | nonpolar  | loop     | alpha helix|
| Cxy24 | Gly17 | 0.009     | 12.5%     | 0.061 | 54.4%   | -41.9%    | polar     | nonpolar | beta sheet beta sheet|
| Cxy6  | Leu3  | 0.011     | 21.1%     | -0.071| 60.7%   | -39.6%    | polar     | alpha helix| alpha helix|
| Gly17 | Gly14 | 0.011     | 22.2%     | -0.070| 60.1%   | -37.9%    | nonpolar  | polar    | loop     |
| Gly14 | Leu8  | 0.009     | 11.4%     | -0.052| 48.7%   | -37.3%    | nonpolar  | polar    | loop     |
| Ser12 | Cxy6  | 0.008     | 2.8%      | -0.037| 39.3%   | -36.5%    | polar     | polar    | alpha helix alpha helix|
| Ser22 | Cxy19 | 0.011     | 40.7%     | -0.117| 76.9%   | -36.2%    | polar     | loop     | alpha helix|
| Gly18 | Ala4  | 0.007     | 0.9%      | 0.035 | 36.8%   | -35.9%    | polar     | nonpolar | nonpolar |
| Cxy26 | Phe23 | 0.013     | 36.5%     | 0.097 | 72.1%   | -35.6%    | polar     | nonpolar | beta sheet beta sheet|
| Phe23 | Cxy6  | 0.010     | 17.4%     | 0.058 | 52.7%   | -35.3%    | nonpolar  | polar    | beta sheet alpha helix|
| Gly18 | Cxy10 | 0.011     | 19.4%     | 0.061 | 54.7%   | -35.3%    | nonpolar  | polar    | beta sheet alpha helix|
| Cxy24 | Leu3  | 0.009     | 12.3%     | -0.047| 46.2%   | -33.9%    | polar     | nonpolar | beta sheet alpha helix|
| Gly18 | Leu8  | 0.009     | 9.4%      | 0.042 | 43.0%   | -33.6%    | nonpolar  | nonpolar | beta sheet alpha helix|
| Thr25 | Cxy6  | 0.010     | 17.1%     | -0.054| 49.6%   | -32.5%    | polar     | polar    | beta sheet alpha helix|
| Cxy26 | Cxy10 | 0.014     | 41.9%     | -0.100| 72.9%   | -31.1%    | polar     | polar    | beta sheet alpha helix|
| Phe23 | Leu8  | 0.012     | 33.0%     | -0.078| 63.2%   | -30.2%    | nonpolar  | nonpolar | beta sheet alpha helix|
| Cxy10 | Cxy6  | 0.011     | 26.2%     | -0.064| 56.4%   | -30.2%    | polar     | alpha helix alpha helix|
| Leu13 | Leu3  | 0.009     | 12.0%     | 0.041 | 41.6%   | -29.6%    | nonpolar  | loop     | alpha helix|
| Gly17 | Ala4  | 0.015     | 47.0%     | 0.113 | 75.5%   | -28.5%    | nonpolar  | nonpolar | beta sheet alpha helix|
| Gly17 | Leu3  | 0.011     | 23.6%     | -0.057| 51.9%   | -28.2%    | nonpolar  | nonpolar | beta sheet alpha helix|
| Cxy24 | Cxy19 | 0.011     | 26.8%     | -0.061| 55.0%   | -28.2%    | polar     | polar    | beta sheet beta sheet|
| Leu13 | Arg5  | 0.015     | 45.0%     | -0.098| 72.4%   | -27.4%    | nonpolar  | charged | loop alpha helix|
| Thr25 | Cxy10 | 0.011     | 19.1%     | -0.047| 46.4%   | -27.4%    | polar     | beta sheet alpha helix|
| Cxy24 | Gln20 | 0.015     | 46.4%     | -0.101| 73.5%   | -27.1%    | polar     | beta sheet beta sheet|
| Cxy19 | Leu8  | 0.011     | 21.4%     | 0.051 | 48.1%   | -26.8%    | polar     | nonpolar | beta sheet alpha helix|
| Leu13 | Cxy6  | 0.009     | 9.1%      | 0.034 | 35.6%   | -26.5%    | nonpolar  | loop     | alpha helix|
| Cxy10 | Ala4  | 0.011     | 22.8%     | -0.053| 49.0%   | -26.2%    | polar     | nonpolar | alpha helix alpha helix|
| Phe23 | Leu13 | 0.015     | 45.6%     | -0.096| 71.5%   | -25.9%    | nonpolar  | nonpolar | beta sheet loop|
| Leu15 | Ala4  | 0.011     | 24.5%     | 0.055 | 50.4%   | -25.9%    | nonpolar  | loop     | alpha helix|
| Ser22 | Ala4  | 0.009     | 7.1%      | 0.031 | 33.0%   | -25.9%    | polar     | nonpolar | beta sheet alpha helix|
| Cxy24 | Ser12 | 0.009     | 6.8%      | 0.030 | 31.9%   | -25.1%    | polar     | polar    | beta sheet alpha helix|
**Table S22.** Residue pairs in mini-CD4 for which the MI % rank is > 25% **above** the |CC| % rank (i.e., the MI is stronger than the CC). The MI and CC values, % rank for each, the difference, residue type, and secondary structure type are shown.

| Res1  | Res2  | MI % rank | CC | CCI % rank | % diff | type 1     | type 2   | SS1     | SS2     |
|-------|-------|-----------|----|------------|--------|------------|----------|---------|---------|
| Lys11 | Ala4  | 0.019     | 61.5% | 0.006 | 8.8% | 52.7% charged | nonpolar | alpha helix | alpha helix |
| Lys16 | Asn2  | 0.016     | 49.6% | -0.008 | 9.4% | 40.2% charged | polar | loop | alpha helix |
| Lys16 | Leu8  | 0.024     | 68.9% | 0.014 | 15.4% | 53.6% charged | nonpolar | loop | alpha helix |
| Lys16 | Ser9  | 0.036     | 80.6% | -0.005 | 6.3% | 74.4% charged | polar | loop | alpha helix |
| Lys16 | Ser12 | 0.018     | 58.4% | -0.025 | 25.1% | 33.3% charged | polar | loop | alpha helix |
| Lys16 | Leu13 | 0.016     | 48.1% | -0.006 | 7.4% | 40.7% charged | nonpolar | loop | loop |
| Leu8  | Cys1  | 0.013     | 35.0% | -0.003 | 4.0% | 31.1% nonpolar | terminal | alpha helix | N-terminus |
| Leu13 | Ser9  | 0.015     | 46.2% | 0.003 | 3.4% | 42.7% nonpolar | polar | loop | alpha helix |
| Gly14 | Cys1  | 0.014     | 39.9% | -0.001 | 0.9% | 39.0% nonpolar | terminal | loop | N-terminus |
| Leu15 | Cys1  | 0.026     | 73.8% | 0.048 | 47.3% | 26.5% nonpolar | terminal | loop | N-terminus |
| Leu15 | Gly14 | 0.044     | 84.9% | -0.022 | 23.1% | 61.8% nonpolar | terminal | loop | loop |
| Gly17 | Ser9  | 0.020     | 63.2% | -0.026 | 27.4% | 35.9% nonpolar | polar | beta sheet | alpha helix |
| Gly17 | Lys11 | 0.019     | 60.7% | 0.014 | 15.7% | 45.0% nonpolar | charged | beta sheet | alpha helix |
| Gly17 | Leu15 | 0.021     | 65.5% | -0.006 | 8.0% | 57.5% nonpolar | nonpolar | beta sheet | loop |
| Gly18 | Lys16 | 0.025     | 70.9% | -0.031 | 32.8% | 38.2% charged | beta sheet | loop |
| Gly21 | Arg5  | 0.024     | 69.2% | -0.029 | 31.1% | 38.2% nonpolar | charged | beta sheet | alpha helix |
| Gly21 | Lys11 | 0.024     | 70.7% | 0.046 | 44.4% | 26.2% nonpolar | charged | beta sheet | alpha helix |
| Gly21 | Leu15 | 0.019     | 61.0% | -0.026 | 26.5% | 34.5% nonpolar | nonpolar | beta sheet | loop |
| Phe23 | Gln7  | 0.021     | 67.0% | 0.032 | 33.3% | 33.6% nonpolar | polar | beta sheet | alpha helix |
| Asn2  | Cys1  | 0.043     | 84.3% | 0.025 | 25.6% | 58.7% polar | terminal | alpha helix | N-terminus |
| Ser9  | Cys1  | 0.027     | 74.9% | -0.029 | 30.8% | 44.2% polar | terminal | alpha helix | N-terminus |
| Ser9  | Gln7  | 0.054     | 89.7% | 0.024 | 24.8% | 65.0% polar | polar | alpha helix | alpha helix |
| Ser9  | Leu8  | 0.019     | 60.1% | 0.017 | 16.8% | 43.3% polar | nonpolar | alpha helix | alpha helix |
| Cys10 | Leu8  | 0.013     | 37.3% | -0.008 | 10.3% | 27.1% polar | nonpolar | alpha helix | alpha helix |
| Cys10 | Ser9  | 0.018     | 56.7% | -0.025 | 25.9% | 30.8% polar | polar | alpha helix | alpha helix |
| Cys19 | Leu15 | 0.017     | 54.1% | -0.017 | 17.1% | 37.0% polar | nonpolar | beta sheet | loop |
| Cys19 | Gly17 | 0.012     | 31.9% | 0.005 | 6.0% | 25.9% polar | nonpolar | beta sheet | beta sheet |
| Gln20 | Cys6  | 0.012     | 29.1% | -0.001 | 1.1% | 27.9% polar | polar | beta sheet | alpha helix |
| Gln20 | Leu15 | 0.026     | 72.9% | 0.032 | 34.5% | 38.5% polar | nonpolar | beta sheet | loop |
| Gln20 | Gly17 | 0.042     | 84.0% | -0.066 | 57.3% | 26.8% polar | nonpolar | beta sheet | beta sheet |
| Gln20 | Gly18 | 0.022     | 67.5% | 0.041 | 41.9% | 25.6% polar | nonpolar | beta sheet | beta sheet |
| Ser22 | Lys16 | 0.020     | 65.2% | -0.032 | 34.8% | 30.5% polar | charged | beta sheet | loop |
| Ser22 | Gly21 | 0.046     | 85.8% | -0.046 | 45.9% | 39.9% polar | nonpolar | beta sheet | beta sheet |
| Thr25 | Cys1  | 0.020     | 65.0% | 0.012 | 14.0% | 51.0% polar | terminal | beta sheet | N-terminus |
| Thr25 | Gly14 | 0.018     | 58.7% | 0.018 | 19.4% | 39.3% polar | nonpolar | beta sheet | loop |
| Thr25 | Lys16 | 0.048     | 86.6% | 0.017 | 18.5% | 68.1% polar | charged | beta sheet | loop |
| Thr25 | Gln20 | 0.026     | 73.5% | -0.046 | 44.7% | 28.8% polar | polar | beta sheet | beta sheet |
| Thr25 | Ser22 | 0.014     | 40.5% | 0.007 | 9.1% | 31.3% polar | polar | beta sheet | beta sheet |
| Cys26 | Ser12 | 0.012     | 28.8% | 0.001 | 0.0% | 28.8% polar | polar | beta sheet | alpha helix |
| Gly27 | Cys1  | 0.057     | 90.6% | -0.022 | 23.4% | 67.2% terminal | terminal | beta sheet | N-terminus |
| Gly27 | Lys11 | 0.113     | 97.7% | -0.002 | 2.0% | 95.7% terminal | charged | beta sheet | alpha helix |
| Gly27 | Ser12 | 0.026     | 73.2% | 0.045 | 44.2% | 29.1% terminal | polar | beta sheet | alpha helix |
| Gly27 | Gly17 | 0.047     | 86.3% | 0.056 | 51.3% | 35.0% terminal | nonpolar | beta sheet | beta sheet |
| Gly27 | Gly18 | 0.016     | 50.4% | 0.021 | 21.9% | 28.5% terminal | nonpolar | beta sheet | beta sheet |
| Gly27 | Gln20 | 0.050     | 88.0% | -0.075 | 61.8% | 26.2% terminal | polar | beta sheet | beta sheet |
| Gly27 | Ser22 | 0.016     | 52.1% | -0.012 | 13.7% | 38.5% terminal | polar | beta sheet | beta sheet |
| Gly27 | Phe23 | 0.024     | 70.1% | -0.027 | 28.5% | 41.6% terminal | nonpolar | beta sheet | beta sheet |
| Gly27 | Cys26 | 0.038     | 82.3% | 0.004 | 5.7% | 76.6% terminal | polar | beta sheet | beta sheet |

---

**Page S34**
Table S23. Residue pairs in benenodin-1 for which the MI % rank is > 25% below the (CC) % rank (i.e., the CC is stronger than the MI). The MI and CC values, % rank for each, the difference, residue type, and secondary structure type are shown.

| Res1 | Res2 | MI   | MI %rank | CC  | CCI %rank | % diff | type 1   | type 2   | SS1 | SS2 |
|------|------|------|----------|-----|-----------|--------|----------|----------|-----|-----|
| Pro7 | Gly3 | 0.027| 16.4%    | -0.082| 62.0%     | -45.6% | polar    | nonpolar | ring | ring |
| Ala16| Gly5 | 0.029| 17.5%    | -0.074| 57.3%     | -39.8% | nonpolar | nonpolar | tail | ring |
| Ala16| Gly1 | 0.025| 24.6%    | 0.046 | 60.8%     | -36.3% | nonpolar | terminal | tail | ring |
| Asp8 | Val2 | 0.024| 5.8%     | -0.033| 40.4%     | -34.5% | polar    | nonpolar | ring | ring |
| Ala16| Asp8 | 0.027| 26.3%    | -0.086| 58.5%     | -32.2% | nonpolar | polar    | tail | ring |
| Leu11| Gly3 | 0.027| 15.2%    | -0.053| 45.6%     | -30.4% | nonpolar | nonpolar | tail | ring |
| Ala16| Gly3 | 0.032| 51.5%    | -0.139| 80.7%     | -29.2% | nonpolar | nonpolar | tail | ring |
| Ile10| Phe4 | 0.027| 18.1%    | 0.054 | 46.2%     | -28.1% | nonpolar | nonpolar | tail | ring |
| Leu11| Phe4 | 0.024| 1.8%     | 0.034 | 29.8%     | -28.1% | nonpolar | nonpolar | tail | ring |
| Gly5 | Gly3 | 0.028| 1.2%     | 0.065 | 28.7%     | -27.5% | nonpolar | nonpolar | ring | ring |
| Ile10| Gly5 | 0.027| 25.7%    | -0.074| 53.2%     | -27.5% | nonpolar | nonpolar | tail | ring |
| Ala16| Ser9 | 0.026| 9.4%     | 0.037 | 36.8%     | -27.5% | nonpolar | polar    | tail | tail |
| Leu11| Asp8 | 0.028| 8.8%     | -0.079| 35.7%     | -26.9% | nonpolar | polar    | tail | ring |
| Thr12| Ser9 | 0.025| 6.4%     | -0.035| 32.2%     | -25.7% | polar    | polar    | tail | tail |
| Gly5 | Val2 | 0.026| 31.6%    | 0.042 | 56.7%     | -25.1% | nonpolar | nonpolar | ring | ring |

Table S24. Residue pairs in benenodin-1 for which the MI % rank is > 25% above the (CC) % rank (i.e., the MI is stronger than the CC). The MI and CC values, % rank for each, the difference, residue type, and secondary structure type are shown.

| Res1 | Res2 | MI   | MI %rank | CC  | CCI %rank | % diff | type 1   | type 2   | SS1 | SS2 |
|------|------|------|----------|-----|-----------|--------|----------|----------|-----|-----|
| Met19| Pro18| 0.034| 87.1%    | 0.034| 7.6%      | 79.5%  | terminal | polar    | tail | tail |
| Pro18| Glu14| 0.038| 73.7%    | -0.012| 10.5%     | 63.2%  | polar    | charged  | tail | tail |
| Met19| Glu14| 0.029| 73.1%    | 0.006 | 12.9%     | 60.2%  | terminal | charged  | tail | tail |
| Gln13| Val2 | 0.030| 66.7%    | 0.004 | 11.7%     | 55.0%  | polar    | nonpolar | tail | ring |
| Lys17| Pro7 | 0.036| 59.6%    | -0.023| 6.4%      | 53.2%  | charged  | polar    | tail | ring |
| Pro18| Gin15| 0.031| 71.3%    | 0.015 | 18.1%     | 53.2%  | polar    | polar    | tail | tail |
| Lys17| Phe4 | 0.049| 64.9%    | -0.070| 22.2%     | 42.7%  | charged  | nonpolar | ring | ring |
| Gin13| Pro7 | 0.036| 63.7%    | -0.030| 26.3%     | 37.4%  | polar    | polar    | tail | ring |
| Glu14| Ile10| 0.034| 39.8%    | 0.007 | 2.3%      | 37.4%  | charged  | nonpolar | tail | tail |
| Arg6 | Phe4 | 0.032| 37.4%    | 0.015 | 4.7%      | 32.7%  | charged  | nonpolar | ring | ring |
| Gin15| Asp8 | 0.030| 47.4%    | 0.008 | 16.4%     | 31.0%  | polar    | polar    | tail | ring |
| Lys17| Asp8 | 0.035| 39.2%    | 0.036 | 8.2%      | 31.0%  | charged  | polar    | tail | ring |
| Lys17| Ser9 | 0.030| 62.0%    | -0.001| 33.9%     | 28.1%  | charged  | polar    | tail | tail |
| Lys17| Gly1 | 0.042| 58.5%    | 0.011 | 30.4%     | 28.1%  | charged  | terminal | tail | ring |
| Gin15| Gly3 | 0.042| 33.3%    | -0.013| 5.3%      | 28.1%  | polar    | nonpolar | tail | ring |
| Arg6 | Gly5 | 0.041| 81.9%    | -0.017| 54.4%     | 27.5%  | charged  | nonpolar | ring | ring |
| Met19| Phe4 | 0.062| 41.5%    | 0.007 | 15.8%     | 25.7%  | terminal | nonpolar | tail | ring |
Table S25. Residue pairs in Trp-cage for which the MI % rank is > 25\% below the |CC| % rank (i.e., the CC is stronger than the MI). The MI and CC values, % rank for each, the difference, residue type, and secondary structure type are shown.

| Res1 | Res2 | MI  | MI % rank | CC  | CC % rank | % diff | type 1    | type 2   | SS1 | SS2 |
|------|------|-----|-----------|-----|-----------|--------|-----------|----------|-----|-----|
| Ser20| Pro17| 0.043| 8.4%      | -0.110| 50.5%     | -42%   | terminal  | polar    | C-terminus | tail |
| Trp6 | Ile4 | 0.046| 13.7%     | -0.118| 54.7%     | -41%   | polar     | nonpolar  | alpha helix | alpha helix |
| Pro17| Gly11| 0.041| 4.2%      | -0.096| 44.7%     | -41%   | polar     | nonpolar  | tail | 3-10 helix |
| Ile4 | Tyr3 | 0.052| 30.5%     | -0.176| 69.5%     | -39%   | nonpolar  | polar     | alpha helix | alpha helix |
| Ser20| Ile4 | 0.048| 18.9%     | 0.119 | 55.8%     | -37%   | terminal  | nonpolar  | C-terminus | alpha helix |
| Pro17| Tyr3 | 0.050| 24.7%     | -0.137| 61.1%     | -36%   | polar     | polar     | tail | alpha helix |
| Pro17| Ile4 | 0.047| 15.8%     | 0.113 | 51.6%     | -36%   | polar     | nonpolar  | tail | alpha helix |
| Gly15| Leu7 | 0.048| 20.5%     | 0.114 | 52.1%     | -32%   | nonpolar  | nonpolar  | tail | alpha helix |
| Gly11| Ile4 | 0.066| 54.2%     | 0.239 | 85.8%     | -32%   | nonpolar  | nonpolar  | 3-10 helix | alpha helix |
| Pro12| Tyr3 | 0.048| 18.4%     | -0.107| 48.9%     | -31%   | polar     | polar     | 3-10 helix | alpha helix |
| Gly15| Trp6 | 0.043| 7.9%      | 0.084 | 38.4%     | -31%   | nonpolar  | polar     | tail | alpha helix |
| Gly11| Trp6 | 0.045| 12.6%     | -0.090| 42.6%     | -30%   | nonpolar  | polar     | 3-10 helix | alpha helix |
| Pro18| Ile4 | 0.048| 17.4%     | -0.099| 46.3%     | -29%   | polar     | nonpolar  | tail | alpha helix |
| Pro19| Ile4 | 0.042| 6.3%      | -0.077| 35.3%     | -29%   | polar     | nonpolar  | tail | alpha helix |
| Gly10| Trp6 | 0.046| 14.7%     | 0.091 | 43.2%     | -28%   | nonpolar  | polar     | turn | alpha helix |
| Ser13| Trp6 | 0.048| 19.5%     | 0.105 | 47.4%     | -28%   | polar     | polar     | 3-10 helix | alpha helix |
| Pro19| Leu7 | 0.045| 11.1%     | 0.086 | 38.9%     | -28%   | polar     | nonpolar  | tail | alpha helix |
Table S26. Residue pairs in Trp-cage for which the MI % rank is > 25% above the |CC| % rank (i.e., the MI is stronger than the CC). The MI and CC values, % rank for each, the difference, residue type, and secondary structure type are shown.

| Res1  | Res2  | MI   | MI % rank | CC  | |CC| % rank | % diff | type 1 | type 2 | SS1 | SS2 |
|-------|-------|------|-----------|-----|--------|---------|--------|--------|--------|-----|-----|
| Gln5  | Asn1  | 0.141| 94.7%     | -0.003| 2.6%   | 92% polar| terminal| alpha helix| alpha helix |
| Arg16 | Asn1  | 0.112| 88.4%     | -0.001| 0.5%   | 88% charged| terminal| tail | alpha helix |
| Ser20 | Asn1  | 0.115| 89.5%     | 0.046| 16.8%  | 73% terminal| terminal| C-terminus| alpha helix |
| Arg16 | Ser14 | 0.070| 60.5%     | -0.015| 5.3%   | 55% charged| polar | tail | 3-10 helix |
| Arg16 | Lys8  | 0.077| 68.4%     | 0.029| 14.7%  | 54% charged| charged | tail | alpha helix |
| Arg16 | Asp9  | 0.079| 70.5%     | 0.049| 22.6%  | 48% charged| charged | tail | turn |
| Pro17 | Asn1  | 0.071| 63.7%     | -0.041| 16.3%  | 47% polar | terminal| tail | alpha helix |
| Lys8  | Leu2  | 0.073| 65.3%     | 0.047| 19.5%  | 46% charged| nonpolar| alpha helix| alpha helix |
| Arg16 | Leu2  | 0.080| 72.1%     | -0.072| 33.2%  | 39% charged| nonpolar| tail | alpha helix |
| Pro18 | Gln5  | 0.059| 45.8%     | 0.018| 7.9%   | 38% polar | polar | tail | alpha helix |
| Ser10 | Asp9  | 0.063| 49.5%     | 0.023| 12.1%  | 37% polar | charged | 3-10 helix| turn |
| Gly11 | Asn1  | 0.126| 92.1%     | 0.119| 55.3%  | 37% nonpolar| terminal| 3-10 helix| alpha helix |
| Ser20 | Leu2  | 0.059| 45.3%     | -0.019| 8.4%   | 37% terminal| nonpolar| C-terminus| alpha helix |
| Arg16 | Pro12 | 0.057| 38.9%     | -0.007| 3.2%   | 36% charged| polar | tail | 3-10 helix |
| Ile4  | Asn1  | 0.079| 71.1%     | -0.079| 36.3%  | 35% nonpolar| terminal| alpha helix| alpha helix |
| Pro18 | Asn1  | 0.098| 84.2%     | 0.112| 51.1%  | 33% polar | terminal| tail | alpha helix |
| Pro17 | Ser14 | 0.055| 36.8%     | 0.013| 4.2%   | 33% polar | polar | tail | 3-10 helix |
| Gly15 | Ser13 | 0.065| 53.2%     | 0.049| 21.1%  | 32% nonpolar| polar | tail | 3-10 helix |
| Pro19 | Lys8  | 0.053| 34.2%     | -0.003| 2.1%   | 32% polar | charged | tail | alpha helix |
| Gly15 | Gly10 | 0.052| 30.0%     | 0.001| 1.1%   | 29% nonpolar| nonpolar| tail | turn |
| Ser14 | Asn1  | 0.105| 86.3%     | -0.130| 57.9%  | 28% polar | terminal| 3-10 helix| alpha helix |
| Arg16 | Ile4  | 0.055| 35.8%     | -0.018| 7.4%   | 28% charged| nonpolar| tail | alpha helix |
| Ser13 | Lys8  | 0.066| 54.7%     | 0.063| 27.4%  | 27% polar | charged | 3-10 helix| alpha helix |
| Gly15 | Lys8  | 0.064| 50.0%     | 0.053| 24.7%  | 25% nonpolar| charged| tail | alpha helix |

Table S27. Summary of cases where CC and MI differ in percentile rank by > 25% (either |CC| > MI or vice versa) for all three proteins by residue sidechain type as well as whether the residues are nearest neighbors.

| Type               | mini-CD4 | benenodin-1 | Trp-cage |
|--------------------|----------|-------------|----------|
|                    | MI 25% > | MI 25% > | MI 25% > |
|                    | | ICC| ICC| ICC| |
| charged-charged    | 2        | 0           | 1        | 0 | 4 | 0 |
| charged-polar      | 12       | 0           | 6        | 0 | 9 | 1 |
| charged-nonpolar   | 13       | 1           | 5        | 0 | 7 | 1 |
| polar-polar        | 6        | 9           | 3        | 1 | 2 | 3 |
| polar-nonpolar     | 12       | 14          | 2        | 6 | 2 | 10 |
| nonpolar-nonpolar  | 3        | 10          | 0        | 8 | 1 | 2 |
| total              | 48       | 34          | 17       | 15 | 25 | 17 |
| nearest neighbor   | 6        | 0           | 2        | 0 | 0 | 1 |
| non-nearest neighbor| 42      | 34          | 15       | 15 | 25 | 16 |
Figure S17. (Left) Average residue pair COM-COM distances from the AIMD 85 ps trajectory (in Å) vs. from the 20-structure NMR ensemble (in Å) of mini-CD4. (right) The average center-of-mass distances between residue pairs (AIMD average $d_{\text{COM-COM}}$, in Å) vs the average minimum distance between any atoms in that pair of residues (average $d_{\text{min}}$, in Å) obtained over the full production length of AIMD simulation in mini-CD4 (gray open circles), benenodin-1 (blue open circles), and Trp-cage (black open circles). The overall correlation (Pearson’s $r$) is 0.96 for all data together, 0.97 for only mini-CD4, 0.93 for benenodin-1, and 0.94 for Trp-cage. A cutoff of 4.3 Å for the average $d_{\text{min}}$ values is selected to distinguish pairs of residues that are considered short-range from the long-range pairs. This minimum average distance corresponds to an average $d_{\text{COM-COM}}$ of 10 Å.

Figure S18. Normalized distribution of COM-COM distances (in Å) of the residue pair Lys11-Gln20 in mini-CD4 over the full AIMD trajectory. Bin sizes are 0.25 Å in width. The vertical red line corresponds the average 16.9-Å distance observed over the trajectory. Representative configuration insets are shown for shorter (i.e., 16.1 Å) and longer (i.e., 18.2 Å) COM-COM distances that differ by the orientation of sidechains (in this case, Gln20 is oriented away while Lys11’s position is less significantly changed). Structures are shown with the full mini-CD4 protein in light gray cartoon, and the two key residues are shown as sticks, with carbon in gray, nitrogen in blue, oxygen in red, and hydrogen in white.
**Figure S19.** Dependence of MI and CC for Trp-cage with the average center-of-mass (COM) distance between residues in a pair ($d$(COM-COM), in Å) during the AIMD simulation. The x-axis values are the same for both plots. The subset of residue pairs corresponding to charged-charged interactions are shown for the CC subpane in blue as shown in inset legend in the bottom pane. In the top MI subpane, two representative pairs are shown in red and annotated. These same residue pairs are shown schematically as sticks along with the remainder of the proteins in cartoon at right, with a subset of representative structures overlaid from AIMD. Atoms in the sidechains are colored as: blue for nitrogen, red for oxygen, and white for carbon. One point with MI of 0.6 has been truncated. This point has a CC of -0.8 and is at 5 Å.

**Text S2.** Description of geometric coupling.
The PARENT program was employed to compute the geometric MI of torsional angles on discrete histograms with 32 bins, selected by trial and error. Representative MI values for the dihedral angle coupling are shown in Figure S20. We investigated the correlation between percentile ranks of geometric and electronic MI for residue pairs to determine if one could be inferred from the other. We determined the correspondence of the rank between the two quantities was limited (Figure S21). To highlight the distinctive features of the two coupling types, we specifically compare two residues pairs in mini-CD4 with one involving strong coupling in geometric motion but weak in charge fluctuation (Leu15-Phe23) and vice versa for the other pair (Arg5-Ser9). Since Leu15 and Phe23 are nonpolar and have centers of mass separated by 15.3 Å, this pair has weak electronic coupling (MI = 41st percentile and CC = 0.04, Figure S22). Although Leu15 and Phe23 are on the opposite sides of the β-sheet, the β-sheet fold leads to coupled motion (MI = 98th percentile and CC = 0.62, Figure S22). In contrast, Arg5 and Ser9 are spatially close and can form hydrogen bonds between the guanidinium sidechain of Arg and the hydroxyl group of Ser (Figure S22). The interaction leads to significant charge coupling (MI = 94th percentile and CC = -0.33, Figure S22). Despite the close hydrogen bonding interaction, backbone geometric coupling between Arg5 and Ser9 is very weak (MI = 1st percentile and CC = 0.01, Figure S22). The differences between geometric and electronic coupling are also observed in benenodin-1, i.e., Asp8-Ile10 (electronic MI = 14th and geometric MI = 95th percentile) and Lys17-Met19 (electronic MI = 100th and geometric MI = 18th percentile). Representative examples in Trp-cage are Tyr3-Pro19 (electronic MI = 2nd and geometric MI = 91st percentile) and Asn1-Ser20 (electronic MI = 90th and geometric MI = 1st percentile). These results indicate the difficulty of inferring electronic coupling based on geometric motions from classical MD.
Figure S20. Mutual information (in $k_B$) coupling of the geometric (i.e., dihedral motions) properties obtained over the production AIMD trajectory of mini-CD4 vs. the COM-COM distance obtained as an average over the AIMD trajectory between the relevant residue pairs.

Figure S21. Comparison of the percentile rank of each pair of residues based on total geometric, dihedral motion MI (in $k_B$) vs. the total by-residue-summed charge-based MI (in $k_B$) for mini-CD4 (left), benenodin-1 (middle), and Trp-cage (right). In all cases, the percentile rank will range from 0–100% for a single protein, even if the individual values of MI differ across proteins or quantity being compared. The symbol of each residue pair indicates the type of interaction: charged-charged and charged-polar (blue circles), polar-polar (red circles), nonpolar-nonpolar (gray circles), charged-nonpolar (green triangles), and polar-nonpolar (light orange triangles), all as indicated in top legend. The overall rank correlation for the two quantities is low for all three proteins.
Figure S22. Parity plots of by-residue partial charge sums ($q$, in a.u.) vs geometric root mean squared deviations (RMSD, in Å) of all atoms in residues for mini-CD4 residues pairs: Leu15-Phe23 (top) and Arg5-Ser9 (bottom). Each point is obtained equally spaced at 0.1 ps increments over one 4.25 ps AIMD trajectory. The geometric RMSD was computed with respect to the representative NMR structure. Insets show the superposition of residue pairs colored (from red to blue) to represent their change in position over the full production trajectory. The cross-correlation (CC) and mutual information (MI) percentile rank are shown in inset for the geometric or partial charge quantities, respectively. The MI percentile rank is obtained with respect to the relevant quantity over all residue pairs in mini-CD4 for either the geometric or partial charge quantities.
Table S28. Parameters for the isopeptide bond in benenodin-1 obtained from the AMBER99 force field\textsuperscript{14}.

| Labels          | Parameters | Notes                                           |
|-----------------|------------|-------------------------------------------------|
| Bond            |            |                                                 |
| N3-CO           | 490        | 1.335 same as C-N JCC,7,(1986),230; AA          |
| Angle           |            |                                                 |
| CX-N3-CO        | 50         | 121.9 same as C-N-CT AA general                |
| H-N3-CO         | 50         | 120 AA general, gln, asn, changed based on NMA n modes |
| N3-CO-2C        | 70         | 116.6 AA general                               |
| N3-CO-O2        | 80         | 122.9 AA general                               |
| Dihedral        |            |                                                 |
| CX-N3-CO-2C     | 4 10       | 180 2 AA,NMA                                   |
| CX-N3-CO-O2     | 4 10       | 180 2 AA,NMA                                   |
| H-N3-CO-O2      | 1 2.5      | 180 -2 JCC,7,(1986),230                       |
| H-N3-CO-2C      | 1 2 0 1    | J.C.cistrans-NMA DE                            |
| N3-CO-2C-CX     | 6 0 0 2    | JCC,7,(1986),230                               |
| N3-CO-2C-HC     | 1 0.08     | 180 3 Junmei et al, 1999                      |
| H-N3-CO-2C      | 1 2 0 1    | J.C.cistrans-NMA DE                            |
Figure S23. Root mean square deviations (RMSDs, in Å) of classical MD trajectories in implicit Generalized-Born (black) and explicit TIP3P (red) solvent for mini-CD4 (top), benenodin-1 (middle), and Trp-cage (bottom) over a 10 ns trajectory. For explicit water, a periodic rectangular box with at least 10 Å TIP3P$^{15}$ water buffer for mini-CD4 and benenodin-1 and 20 Å buffer for Trp-cage was employed. After neutralization with Cl$^-$, final system sizes for each
protein and explicit water box were 7,291 atoms for mini-CD4, 6,681 atoms for benenodin-1, and 21,548 atoms for Trp-cage.

**Figure S24.** The MI of mini-CD4 charges from 85 ps of AIMD (left) and 32 ps of RE-AIMD (right). All residue numbers correspond to the residues in the mini-CD4 sequence, and the individual matrix elements are colored according to the colorbar shown at right.

**Figure S25.** The MI of benenodin-1 charges from 85 ps of AIMD (left) and 32 ps of RE-AIMD (right). All residue numbers correspond to the residues in the benenodin-1 sequence, and the individual matrix elements are colored according to the color bar shown at right.
Figure S26. The MI of mini-CD4 charges from different lengths of AIMD starting from the initial 10 simulations and including 25 ps (top, left), 50 ps (top, right), 75 ps (bottom, left), or 85 ps (i.e., the full length of the trajectory used for analysis, bottom, right). All residue numbers correspond to the residues in the mini-CD4 sequence, and the individual matrix elements are colored according to the colorbar shown at right.
Figure S27. The MI of benenodin-1 charges from different lengths of AIMD starting from the initial 10 simulations and including 25 ps (top, left), 50 ps (top, right), 75 ps (bottom, left), or 85 ps (i.e., the full length of the trajectory used for analysis, bottom, right). All residue numbers correspond to the residues in the benenodin-1 sequence, and the individual matrix elements are colored according to the colorbar shown at right.
Table S29. The mean and standard deviation (std.) of individual residue charges in mini-CD4 from different lengths of AIMD starting from the initial 10 simulations and including 25 ps, 50 ps, 75 ps, or 85 ps. For each time length, mean absolute error (MAE) of the mean residue charges was computed with respect to the 85 ps data set, as shown at the bottom of the table. All Cys residues are in disulfide bridges and annotated as Cyx.

| Index | Residue | 25 ps | 50 ps | 75 ps | 85 ps |
|-------|---------|-------|-------|-------|-------|
|       |         | mean  | std.  | mean  | std.  | mean  | std.  | mean  | std.  | mean  | std.  |
| 1     | CYX     | 0.885 | 0.046 | 0.866 | 0.043 | 0.876 | 0.046 | 0.877 | 0.045 |
| 2     | ASN     | 0.133 | 0.034 | 0.136 | 0.033 | 0.136 | 0.033 | 0.137 | 0.033 |
| 3     | LEU     | 0.054 | 0.028 | 0.056 | 0.028 | 0.057 | 0.028 | 0.055 | 0.028 |
| 4     | ALA     | 0.036 | 0.033 | 0.037 | 0.032 | 0.037 | 0.032 | 0.038 | 0.031 |
| 5     | ARG     | 0.946 | 0.055 | 0.948 | 0.052 | 0.941 | 0.052 | 0.937 | 0.051 |
| 6     | CYX     | -0.002| 0.030 | -0.005| 0.029 | -0.005| 0.030 | -0.004| 0.030 |
| 7     | GLN     | 0.043 | 0.057 | 0.051 | 0.052 | 0.047 | 0.056 | 0.042 | 0.056 |
| 8     | LEU     | 0.013 | 0.031 | 0.011 | 0.032 | 0.015 | 0.032 | 0.015 | 0.032 |
| 9     | SER     | -0.008| 0.041 | -0.011| 0.042 | -0.003| 0.042 | -0.004| 0.042 |
| 10    | CYX     | -0.035| 0.030 | -0.035| 0.031 | -0.033| 0.031 | -0.034| 0.031 |
| 11    | LYS     | 0.947 | 0.067 | 0.942 | 0.070 | 0.932 | 0.073 | 0.935 | 0.072 |
| 12    | SER     | -0.041| 0.036 | -0.038| 0.035 | -0.042| 0.036 | -0.040| 0.036 |
| 13    | LEU     | -0.025| 0.028 | -0.023| 0.028 | -0.023| 0.028 | -0.023| 0.029 |
| 14    | GLY     | -0.048| 0.030 | -0.051| 0.028 | -0.050| 0.029 | -0.049| 0.029 |
| 15    | LEU     | -0.005| 0.037 | -0.007| 0.035 | -0.009| 0.038 | -0.008| 0.037 |
| 16    | LYS     | 0.875 | 0.067 | 0.852 | 0.061 | 0.857 | 0.058 | 0.862 | 0.062 |
| 17    | GLY     | -0.012| 0.032 | -0.009| 0.032 | -0.009| 0.032 | -0.007| 0.032 |
| 18    | GLY     | 0.046 | 0.026 | 0.048 | 0.026 | 0.049 | 0.027 | 0.048 | 0.027 |
| 19    | CYX     | -0.015| 0.034 | -0.012| 0.033 | -0.013| 0.032 | -0.015| 0.033 |
| 20    | GLN     | -0.001| 0.039 | -0.006| 0.040 | 0.001 | 0.041 | 0.000 | 0.041 |
| 21    | GLY     | 0.037 | 0.031 | 0.034 | 0.029 | 0.036 | 0.030 | 0.037 | 0.031 |
| 22    | SER     | 0.015 | 0.032 | 0.020 | 0.031 | 0.019 | 0.032 | 0.020 | 0.032 |
| 23    | PHE     | 0.062 | 0.033 | 0.068 | 0.032 | 0.066 | 0.032 | 0.065 | 0.032 |
| 24    | CYX     | -0.026| 0.027 | -0.027| 0.027 | -0.027| 0.027 | -0.026| 0.027 |
| 25    | THR     | 0.004 | 0.045 | -0.001| 0.041 | -0.006| 0.039 | -0.008| 0.039 |
| 26    | CYX     | -0.066| 0.033 | -0.062| 0.034 | -0.060| 0.033 | -0.059| 0.033 |
| 27    | GLY     | -0.814| 0.072 | -0.797| 0.061 | -0.789| 0.063 | -0.792| 0.062 |
| MAE   | 0.005 | 0.003 | 0.002 | --   | --   | --   | --   | --   | --   |
Table S30. The mean and standard deviation (std.) of individual residue charges in benenodin-1 from different lengths of AIMD starting from the initial 10 simulations and including 25 ps, 50 ps, 75 ps, or 85 ps. For each time length, mean absolute error (MAE) of the mean residue charges was computed with respect to the 85 ps data set, as shown at the bottom of the table. The Gly1-Asp8 residues form an isopeptide bond, as indicated with a *. Only the C-terminnal Met19 is charged.

| Index | Residue | 25 ps | 50 ps | 75 ps | 85 ps |
|-------|---------|-------|-------|-------|-------|
|       |         | mean  | std.  | mean  | std.  | mean  | std.  | mean  | std.  |
| 1     | GLY*    | 0.018 | 0.031 | 0.013 | 0.032 | 0.013 | 0.031 | 0.012 | 0.031 |
| 2     | VAL     | 0.046 | 0.031 | 0.049 | 0.030 | 0.046 | 0.029 | 0.047 | 0.029 |
| 3     | GLY     | 0.044 | 0.028 | 0.045 | 0.027 | 0.044 | 0.027 | 0.045 | 0.027 |
| 4     | PHE     | -0.013| 0.032 | -0.015| 0.031 | -0.017| 0.030 | -0.017| 0.031 |
| 5     | GLY     | -0.052| 0.028 | -0.054| 0.028 | -0.054| 0.027 | -0.054| 0.027 |
| 6     | ARG     | 0.812 | 0.068 | 0.801 | 0.058 | 0.809 | 0.062 | 0.808 | 0.060 |
| 7     | PRO     | 0.052 | 0.031 | 0.050 | 0.031 | 0.051 | 0.031 | 0.051 | 0.031 |
| 8     | ASP*    | 0.011 | 0.032 | 0.010 | 0.032 | 0.010 | 0.032 | 0.010 | 0.032 |
| 9     | SER     | 0.002 | 0.031 | 0.003 | 0.030 | 0.004 | 0.030 | 0.003 | 0.030 |
| 10    | ILE     | -0.042| 0.029 | -0.046| 0.028 | -0.044| 0.028 | -0.045| 0.028 |
| 11    | LEU     | 0.008 | 0.027 | 0.007 | 0.027 | 0.006 | 0.027 | 0.006 | 0.027 |
| 12    | THR     | -0.021| 0.030 | -0.023| 0.030 | -0.028| 0.030 | -0.028| 0.030 |
| 13    | GLN     | -0.019| 0.063 | -0.007| 0.062 | 0.001 | 0.065 | -0.005| 0.067 |
| 14    | GLU     | -0.910| 0.056 | -0.918| 0.056 | -0.908| 0.054 | -0.903| 0.054 |
| 15    | GLN     | -0.008| 0.056 | 0.000 | 0.059 | -0.008| 0.059 | -0.008| 0.059 |
| 16    | ALA     | 0.018 | 0.030 | 0.019 | 0.029 | 0.018 | 0.029 | 0.017 | 0.029 |
| 17    | LYS     | 0.894 | 0.073 | 0.882 | 0.067 | 0.872 | 0.061 | 0.871 | 0.059 |
| 18    | PRO     | -0.019| 0.035 | -0.011| 0.036 | -0.013| 0.037 | -0.011| 0.037 |
| 19    | MET     | -0.821| 0.042 | -0.806| 0.043 | -0.801| 0.039 | -0.799| 0.039 |
| MAE   |         | 0.005 | 0.003 | 0.001 |       | --     |        |        |        |
Replica exchange\textsuperscript{16-20} (RE)-AIMD was carried out to study the effect of enhanced sampling on property distributions using TeraChem. A total of eight replicas were used with temperatures ranging from 269.5 to 570.9 K to accelerate dynamics at 300 K. These values were obtained from a temperature generator tool to produce an estimated exchange probability of 0.2.\textsuperscript{18} These RE-AIMD simulations were run for 8 ps for each replica, resulting a total of 32 ps of 300 K data for analysis.

| Replica | Temperature (K) |
|---------|----------------|
| 1       | 269.5          |
| 2       | 300.0          |
| 3       | 334.0          |
| 4       | 371.8          |
| 5       | 413.9          |
| 6       | 460.7          |
| 7       | 512.9          |
| 8       | 570.9          |

Table S32. The mean and standard deviation (std.) of individual residue charges in mini-CD4 from 32 ps of RE-AIMD computed with and without empirical D3 dispersion correction. Some charged residues are found to be more sensitive (i.e., have more distinct charge distributions) to D3 corrections, including Cys1, Lys11, Lys16, and Gly27. All Cys residues are in disulfide bridges and annotated as Cyx.

| Index | Residue | $\omega$PBEh/6-31G | $\omega$PBEh-D3/6-31G | abs. diff. |
|-------|---------|---------------------|----------------------|------------|
|       |         | mean | std. | mean | std. | mean | std. | mean | std. |
| 1     | CYX     | 0.886 | 0.042 | 0.849 | 0.059 | 0.036 | 0.018 |
| 2     | ASN     | 0.118 | 0.039 | 0.141 | 0.036 | 0.022 | 0.002 |
| 3     | LEU     | 0.049 | 0.033 | 0.067 | 0.031 | 0.018 | 0.002 |
| 4     | ALA     | 0.035 | 0.036 | 0.048 | 0.036 | 0.013 | 0.000 |
| 5     | ARG     | 0.946 | 0.055 | 0.940 | 0.045 | 0.006 | 0.010 |
| 6     | CYX     | 0.001 | 0.033 | 0.000 | 0.037 | 0.000 | 0.004 |
| 7     | GLN     | 0.087 | 0.049 | 0.071 | 0.062 | 0.015 | 0.013 |
| 8     | LEU     | 0.027 | 0.038 | 0.022 | 0.036 | 0.005 | 0.002 |
| 9     | SER     | -0.004 | 0.040 | -0.010 | 0.042 | 0.006 | 0.002 |
| 10    | CYX     | -0.041 | 0.032 | -0.035 | 0.036 | 0.006 | 0.004 |
| 11    | LYS     | 0.920 | 0.049 | 0.888 | 0.071 | 0.032 | 0.022 |
| 12    | SER     | -0.052 | 0.040 | -0.059 | 0.039 | 0.007 | 0.002 |
| 13    | LEU     | -0.025 | 0.033 | -0.035 | 0.036 | 0.010 | 0.002 |
| 14    | GLY     | -0.054 | 0.029 | -0.053 | 0.033 | 0.001 | 0.003 |
| 15    | LEU     | -0.019 | 0.039 | 0.011 | 0.041 | 0.030 | 0.002 |
| 16    | LYS     | 0.928 | 0.083 | 0.863 | 0.047 | 0.065 | 0.036 |
| 17    | GLY     | -0.010 | 0.031 | 0.000 | 0.032 | 0.010 | 0.001 |
| 18    | GLY     | 0.035 | 0.031 | 0.055 | 0.031 | 0.021 | 0.000 |
| 19    | CYX     | -0.027 | 0.041 | 0.004 | 0.041 | 0.031 | 0.000 |
| 20    | GLN     | 0.004 | 0.036 | -0.019 | 0.042 | 0.023 | 0.007 |
| 21    | GLY     | 0.042 | 0.035 | 0.031 | 0.033 | 0.011 | 0.002 |
| 22    | SER     | 0.030 | 0.034 | 0.029 | 0.041 | 0.001 | 0.007 |
| 23    | PHE     | 0.049 | 0.037 | 0.076 | 0.043 | 0.027 | 0.005 |
| 24    | CYX     | -0.019 | 0.031 | -0.024 | 0.037 | 0.005 | 0.006 |
| 25    | THR     | -0.014 | 0.039 | 0.007 | 0.043 | 0.022 | 0.004 |
| 26    | CYX     | -0.058 | 0.035 | -0.071 | 0.035 | 0.013 | 0.000 |
| 27    | GLY     | -0.832 | 0.047 | -0.797 | 0.060 | 0.035 | 0.013 |
**Figure S28.** The MI of mini-CD4 charges from 32 ps of production RE-AIMD computed without (left) and with (right) empirical D3 dispersion correction\(^{21-22}\). All residue numbers correspond to the residues in the mini-CD4 sequence, and the individual matrix elements are colored according to the colorbar shown at right.

**Table S33.** Details of spherical droplet for explicit water QM/MM modeling of Trp-cage extracted from periodic all-MM simulations with TIP3P water. The center coordinate refers to the position obtained from the Trp-cage center of mass used for identifying the closest waters to create the droplet.

| Description                  | Amount                        |
|------------------------------|-------------------------------|
| Trp-cage protein atoms       | 304                           |
| Counterion (Cl)              | 1                             |
| Sphere (center, Å)           | 32.557, 30.895, 27.586         |
| Sphere (radius, Å)           | 29.0                           |
| # water molecules/atoms       | 4,001/12,003                  |
| Total # atoms                | 12,308                        |
References
1. Vita, C.; Drakopoulou, E.; Vizzavona, J.; Rochette, S.; Martin, L.; Ménez, A.; Roumestand, C.; Yang, Y.-S.; Ylisastigui, L.; Benjouad, A.; Gluckman, J. C., Rational Engineering of a Miniprotein That Reproduces the Core of the Cd4 Site Interacting with Hiv-1 Envelope Glycoprotein. Proc. Natl. Acad. Sci. U. S. A. 1999, 96, 13091.
2. Zong, C.; Wu, M. J.; Qin, J. Z.; Link, A. J., Lasso Peptide Benenodin-1 Is a Thermally Actuated [1]Rotaxane Switch. J. Am. Chem. Soc. 2017, 139, 10403-10409.
3. Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H., Designing a 20-Residue Protein. Nat. Struct. Biol. 2002, 9, 425.
4. Fonseca Guerra, C.; Handgraaf, J. W.; Baerends, E. J.; Bickelhaupt, F. M., Voronoi Deformation Density (VDD) Charges: Assessment of the Mulliken, Bader, Hirshfeld, Weinhold, and VDD Methods for Charge Analysis. J. Comput. Chem. 2004, 25, 189-210.
5. Becke, A. D., A Multicenter Numerical Integration Scheme for Polyatomic Molecules. J. Chem. Phys. 1988, 88, 2547-2553.
6. Hirshfeld, F. L., Bonded-Atom Fragments for Describing Molecular Charge Densities. Theor. Chim. Acta 1977, 44, 129-138.
7. Lu, T.; Chen, F., Multiwf2: A Multifunctional Wavefunction Analyzer. Journal of computational chemistry 2012, 33, 580-592.
8. Cover, T. M.; Thomas, J. A., Elements of Information Theory. John Wiley & Sons: 2012.
9. Cortina, G. A.; Kasson, P. M., Excess Positional Mutual Information Predicts Both Local and Allosteric Mutations Affecting Beta Lactamase Drug Resistance. Bioinformatics 2016, 32, 3420-3427.
10. McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P., Quantifying Correlations between Allosteric Sites in Thermodynamic Ensembles. J. Chem. Theory Comput. 2009, 5, 2486-2502.
11. Smith, R., A Mutual Information Approach to Calculating Nonlinearity. Stat 2015, 4, 291-303.
12. Gel'fand, I. M.; Yaglom, A. M., Computation of the Amount of Information About a Stochastic Function Contained in Another Such Function. Usp. Mat. Nauk 1957, 12, 3-52.
13. Fleck, M.; Polyansky, A. A.; Zagrovic, B., Parent: A Parallel Software Suite for the Calculation of Configurational Entropy in Biomolecular Systems. J. Chem. Theory Comput. 2016, 12, 2055-2065.
14. Wang, J.; Cieplak, P.; Kollman, P. A., How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? J. Comput. Chem. 2000, 21, 1049-1074.
15. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. J. Chem. Phys. 1983, 79, 926-935.
16. Hukushima, K.; Nemoto, K., Exchange Monte Carlo Method and Application to Spin Glass Simulations. Journal of the Physical Society of Japan 1996, 65, 1604-1608.
17. Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M., Replica-Exchange Monte Carlo Method for the Isobaric–Isothermal Ensemble. Chem. Phys. Lett. 2001, 335, 435-439.
18. Patriksson, A.; van der Spoel, D., A Temperature Predictor for Parallel Tempering Simulations. Physical Chemistry Chemical Physics 2008, 10, 2073-2077.
19. Sugita, Y.; Okamoto, Y., Replica-Exchange Molecular Dynamics Method for Protein Folding. Chem. Phys. Lett. 1999, 314, 141-151.
20. van der Spoel, D.; Seibert, M. M., Protein Folding Kinetics and Thermodynamics from Atomistic Simulations. *Physical Review Letters* **2006**, *96*, 238102.
21. Grimme, S., Density Functional Theory with London Dispersion Corrections. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 211-228.
22. Grimme, S.; Ehrlich, S.; Goerigk, L., Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456-1465.
Other files

SI_Data120120.zip (3.74 MiB)  view on ChemRxiv  download file