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

Zhongyue Yang[1], Natalia Hajlasz[1], Adam H. Steeves[1], and Heather J. Kulik[1,*]

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

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 *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[80, 88-95]. Recent advances[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[96, 103]. This has motivated increasingly large-scale QM region treatments in QM/MM models of enzyme catalysis[35, 104-119], which have revealed unexpectedly large dependence of properties such as the favorability of proton or charge transfer[106], electric fields[35, 75], excitation energies[114-115, 120], bond critical points[117] and partial charges[116] on the selection of the QM region. These observations have motivated renewed interest in systematic methods for atom-economical QM region selection[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[124]. Recently, we carried out[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.[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 Cα atoms or dihedral angles) are coupled. Analysis of geometric coupling has been extensively applied to understand this conformational allostery in proteins.[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[129-131], interpret QM/MM simulations[76, 121, 125], or to guide QM method selection[132], it has not been applied to the charge coupling obtained from *ab initio* molecular dynamics (AIMD) of entire proteins.

While some QM effects can be incorporated using recent developments in polarizable force field modeling[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,[96, 103] simulations have revealed the importance of first principles to accurately describe unexpected structures[96] and to explain charge transfer[124] and polarization in water[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.[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 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)[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% α-helix and β-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[134], which consists of an N-terminus α-helix (residues 1–12) and C-terminus β-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).

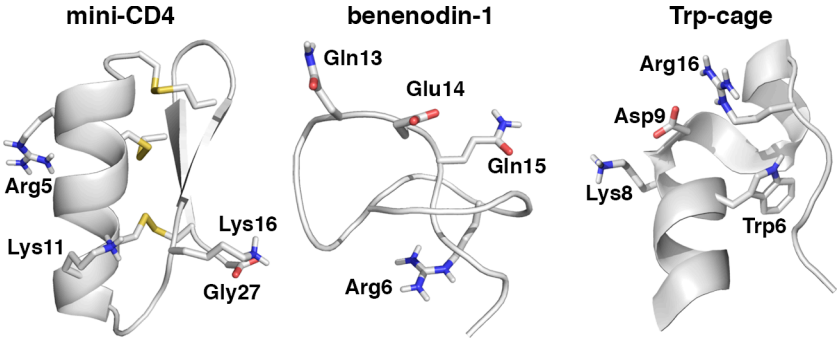


**Figure 1.** Cartoon structures (in white) for mini-CD4 (left, PDB ID: 1D5Q[134]), benenodin-1 (middle, PDB ID: 6B5W[135]), and Trp-cage (right, PDB ID: 1L2Y[136]), as obtained from their solution NMR structures. Representative polar and charged residue sidechains are labeled with

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[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[135], which is a naturally occurring[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[136], a representative designed peptide that has been widely used[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 $3_{10}$ 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 *ab initio* molecular dynamics (see Sec. 4).

**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, $q$(RES):

$$q(\mathrm{RES}) = \sum_{i}^{N_{\mathrm{at}} \in \mathrm{RES}} q_i \tag{0}$$

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 $q$(RES) values to quantify the flexibility of the charge distribution, and we estimate the relative deviation of $q$(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 $q$(RES) distributions, and they are the only residue class with consistently normally distributed $q$(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 α-helix and β-sheet has a significantly larger range (ca. 0.3 a.u.) than a Leu3 (range: 0.2 a.u.) in the α-helix (ESI Figure S3 and Table S8).

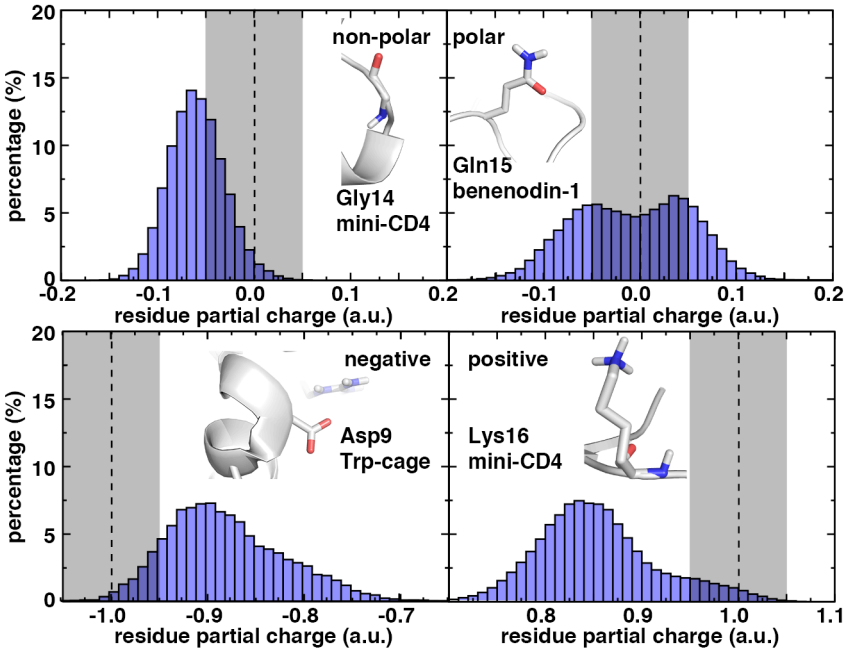


**Figure 2.** Normalized charge distributions of the by-residue summed partial charges for representative amino acids: nonpolar Gly14 in mini-CD4 (top, left), polar Gln15 in benenodin-1 (top, right), negatively charged Asp9 in Trp-cage (bottom, left), and positively charged Lys16 in mini-CD4 (bottom, 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)[144-145] between the by-residue summed partial charges, $q(J)$ and $q(K)$, of residues $J$ and $K$ as:

$$\rho_{JK} = \frac{\sigma_{JK}}{\sigma_J \sigma_K} \tag{0}$$

where $\sigma_{JK}$ is the covariance between $q(J)$ and $q(\mathrm{K})$ and $\sigma_J$ or $\sigma_\mathrm{K}$ 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).

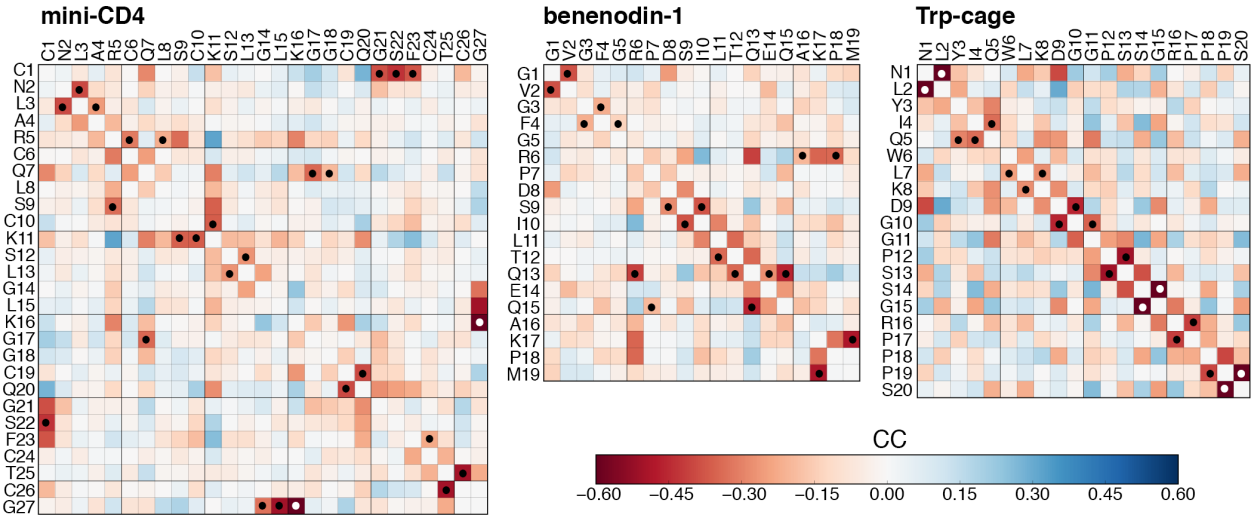


**Figure 3.** Matrix of signed cross-correlation (CC) values ranging from -0.60 (red) to +0.60 (blue) colored as in inset colorbar. Select matrix elements exceeding the range are capped to the extrema of the range. All residues are indicated by their single-letter code and number. The single strongest 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[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 $3_{10}$-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.

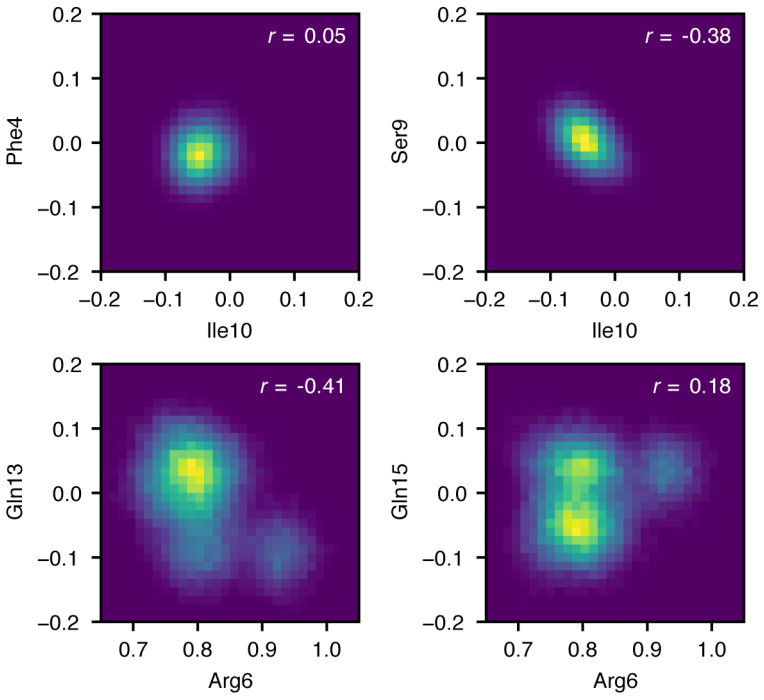

**Figure 4.** Example joint distributions of $q$(RES) (in a.u.) for four pairs of residues in benenodin-1 with CC values shown in upper right insets: (top, left) Phe4–Ile10, (top, right) Ser9–Ile10, (bottom, left) Arg6–Gln13, and (bottom, right) Arg6–Gln15. The same color scale is used for the normalized 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[126-128], we computed the mutual information (MI)[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:

$$I(J;K) = \sum_j \sum_k p_{(J,K)}(j,k) \ln\left(\frac{p_{(J,K)}(j,k)}{p_J(j)p_K(k)}\right) \quad (0)$$

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[147-148] derived from the CC (ESI Text S1).[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[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|,

ESI Figure S13). 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).

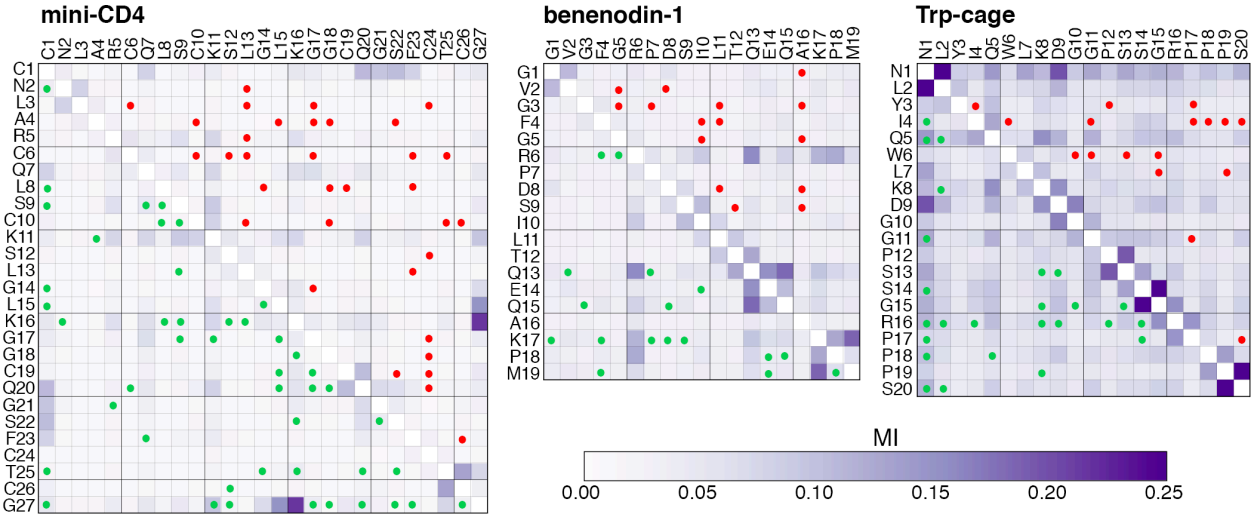


**Figure 5.** Matrix of mutual information (MI) values ranging from 0.0 (white) to +0.25 (dark blue) colored as in inset colorbar. Select matrix elements exceeding the range are capped to the extrema of the range. All residues are indicated by their single-letter code and number. The cases for which 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 > |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 97$^{th}$ 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., 46$^{th}$ percentile) of all |CC| values is instead one of the weakest (i.e., 18$^{th}$ 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 lengthscales 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).

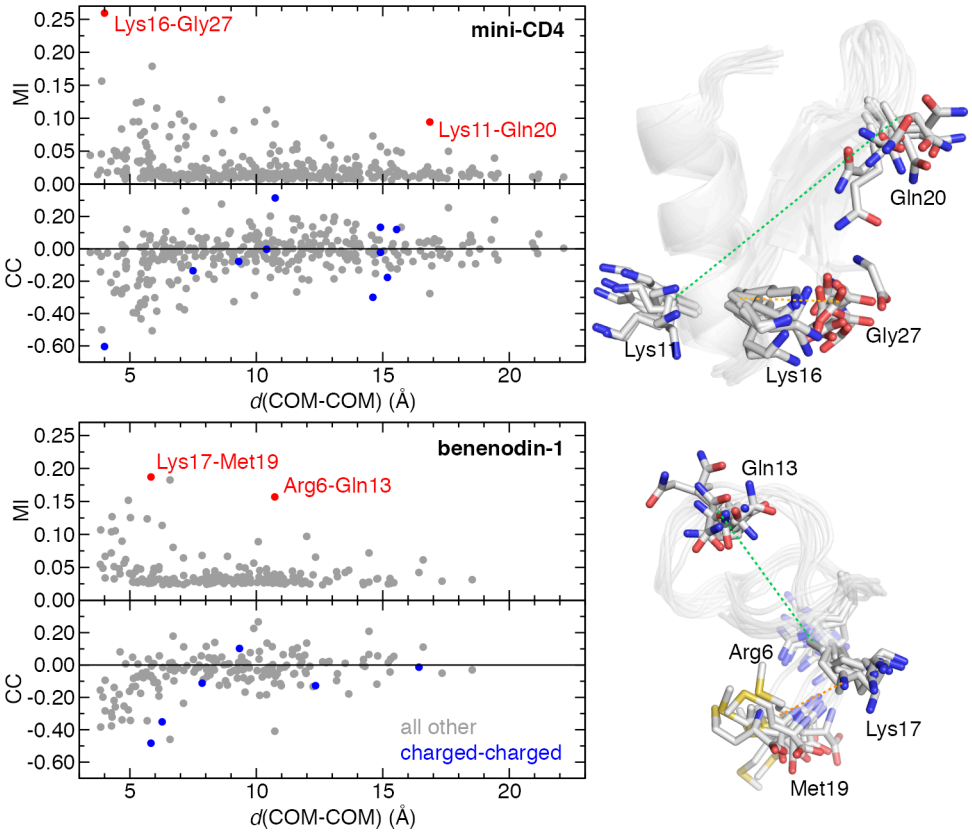


**Figure 6.** Dependence of MI and CC for mini-CD4 (top) and benenodin-1 (bottom) on the average center-of-mass (COM) distance between residues in a pair (*d*(COM-COM), in Å) during the AIMD simulation. The axis values are the same for both plots. The subset of residue pairs corresponding to charged–charged 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[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[124-125] and emphasize the importance of continued study of the quantum mechanical mechanisms underlying this phenomenon.

**3. Conclusions**

We carried out fully *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),[134] the 19-residue lasso peptide benenodin-1 (PDB ID: 6B5W),[135] and the 20-residue globular protein Trp-cage (PDB ID: 1L2Y)[136]. Protonation states were assigned with the H++ webserver[149-151] 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[134] was simulated with its three disulfide bonds at Cys1–Cys19, Cys6–Cys24, and Cys10–Cys26 intact, and benenodin-1[135] 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[152] tleap utility for classical molecular dynamics (MD) equilibration with the AMBER ff14SB force field[153]. Isopeptide bond parameters in benenodin-1 were obtained from the AMBER99 force field[154] (ESI Table S28). The miniproteins were equilibrated in both explicit TIP3P[155] water and with the implicit generalized Born solvent model[156-157] with all defaults applied to assess the impact of solvent choice. All proteins were equilibrated using the GPU-accelerated PMEMD AMBER code[158-159] 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[160] 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] ($\omega = 0.2$ 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.; Nielson, 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.; Śliwa, 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 Hydrolases. *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*, 5793-5800.
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. Mehmood, 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 Atomistic 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.; Tavernelli, 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.; Martínez, 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.; Caprasecca, 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.; Martinez, 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.; Martinez, 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.; Lambrecht, 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. Phys. Chem. B* **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.-Chim. 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.; 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.
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.; Fesinmeyer, 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 Hydrogens 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. Harihara, 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. (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.; St\"ogbauer, H.; Grassberger, P., Estimating Mutual Information. *Phys. Rev. E* **2004,** *69*, 066138.