Conformational Ensembles Reveal the Origins of Serine Protease Catalysis

Authors: Siyuan Du[1,2], Rachael C. Kretsch[1,3], Jacob Parres-Gold†, Elisa Pieri[2,4,5]‡, Vinícius Wilian D. Cruzeiro[2,4,5], Mingning Zhu[2,4,5], Margaux M. Pinney[1]§, Filip Yabukarski[1]¶, Jason P. Schwans[1]#, Todd J. Martínez[2,4,5], Daniel Herschlag[1,6,7]*

Affiliations:

[1] Department of Biochemistry, Stanford University; Stanford, 94305, United States.

[2] Department of Chemistry, Stanford University; Stanford, 94305, United States.

[3] Biophysics Program, Stanford University; Stanford, 94305, United States.

[4] The PULSE Institute, Stanford University, Stanford, CA 94305

[5] SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025

[6] Department of Chemical Engineering, Stanford University, Stanford, California 94305, United States.

[7] Stanford ChEM-H, Stanford University, Stanford, California 94305, United States.

†Current affiliation: Division of Biology and Biological Engineering and Division of Chemistry and Chemical Engineering, California Institute of Technology; Pasadena, 91125, United States

‡Current affiliation: Department of Chemistry, University of North Carolina at Chapel Hill; Chapel Hill, 27514, United States

§Current affiliation: Department of Biochemistry and Biophysics, University of California San Francisco; San Francisco, 94158, United States

¶Current affiliation: Bristol Myers Squibb; San Diego, 92121, United States.

#Current affiliation: Department of Chemistry and Biochemistry, California State University, Long Beach; Long Beach, 90840, United States.

*Corresponding author. Email: herschla@stanford.edu

One sentence summary: Ensemble–function analyses provide a quantitative model fully accounting for serine protease catalysis, reveal previously unknown conformational features that contribute to their catalysis, and identify a structural motif responsible for the catalytically competent conformation that has evolved in over 100 different enzymes.

Abstract: Despite a century of enzymology, we lack the ability to fully describe enzyme catalysis in molecular and energetic terms. Conformational ensembles built from 1231 structures of 17 serine proteases provide a comprehensive model that specifies the molecular features of their reaction path, reveals previously unknown catalytic features, and quantitates their catalytic contributions. The sum of these contributions fully accounts for the rate enhancement provided by serine proteases. These enzymes precisely position their reactants in destabilized conformers, creating a downhill energetic gradient that selectively favors the motions required for reaction while limiting off-pathway motions. A local catalytic motif, the "nucleophilic elbow", generates ground state destabilization and has evolved in 50 proteases and 52 additional enzymes spanning 32 distinct structural folds. Ensemble–function analyses provide experimentally based, quantitative models for enzyme catalysis based on simple physical and chemical properties and identify molecular features that may be used as ingredients in enzyme design.

Main Text:

Understanding enzyme catalysis has been a central goal of biochemistry, with broad implications spanning fundamental understanding of biology, medical interventions, and engineering enzymes of practical value. Over the past century, we have largely achieved the first cornerstone of this understanding—high-confidence chemical models for the roles of active site residues and cofactors for the vast majority of enzymes (*1–6*). Nevertheless, the second cornerstone has remained elusive—understanding and reproducing the rate enhancements of natural enzymes. Consequently, current efforts to design enzymes typically require additional cycles of directed evolution, and even then only occasionally approach the efficiency of natural enzymes (*7*, *8*). To achieve a comprehensive understanding, we need to advance from descriptive models to quantitative models that specify the molecular features involved in catalysis and their catalytic contributions. We meet these goals herein by supplanting traditional structure–function analysis with ensemble–function analysis.

Serine proteases are the classic example in biochemistry textbooks used to illustrate how enzymes perform catalysis (*9–12*). While these descriptions highlight the nucleophilic serine, the oxyanion hole, and the catalytic triad (Fig. 1A) (*13–18*), it is not apparent *how* these catalytic groups furnish the observed 10^9-fold rate enhancement (*17*, *19*) (Fig 1B). The hydroxyl group of the nucleophilic serine has nearly identical reactivity as water, so it alone does not provide catalytic advantages. The oxyanion hole typically uses amide N–H hydrogen bond donors, but amides form hydrogen bonds that are similar in strength as the hydrogen bonds from water (*20*). The histidine of the catalytic triad is expected to accelerate the reaction by acting as a general base, but not nearly to the rate enhancement achieved by serine proteases (*21*, *22*). There have been several proposals for additional catalytic contributions from the catalytic triad aspartate, but they have not been supported by subsequent experiments, and its role (or roles) remains to be determined (supplementary text S1).

Identifying enzyme features that provide catalysis and quantifying their catalytic contributions requires first knowing what changes occur along the reaction path and then determining the energetic consequences of these changes. Conformational ensembles, which encode distributions of states, are needed to achieve this level of understanding. Most fundamentally, distributions of states are needed to calculate entropy (*23*). The ability of enzymes to restrict motions of the reacting groups and position them for reaction is near-universally invoked in the literature as a catalytic feature (table S1) (*2*, *24–41*). However, the extent of enzyme positioning and the amount of catalytic contribution arising from its reduced conformational entropy remain undetermined, a deficit that could be overcome with ensemble information. In addition, ensembles of multiple reaction states are needed to determine what changes occur along the reaction path (*42*). Although one could identify differences between single, static X-ray structures of the same protein in different states, these differences may be real or may result from capturing random conformers within the same underlying distribution (*43*). Ensembles provide the statistical power to identify the changes that are significant and to quantitate their energetic differences. Here, we describe an experimentally-based *ensemble–function* approach that allowed us to identify previously unknown catalytic features, establish the physical origins of their catalysis, and quantify their catalytic contributions, leading to a quantitative model that fully accounts for serine protease catalysis.

Specifically, we brought together >1000 X-ray crystallographic structures of serine proteases to build "pseudo-ensembles" (*42*, *44–47*) and we asked the following questions: (1) *What molecular changes occur along the reaction path?* (2) *What are the energetic consequences of these changes?* (3) *How do the molecular and energetic features of the enzymatic reaction differ from those of the uncatalyzed reaction in solution?* and (4) *What*

elements of the enzyme architecture are responsible for these differences? Pseudo-ensembles in apo, ground state analog (GSA) and transition state analog (TSA) bound states revealed a shortened and low-entropy enzymatic reaction path, wherein a sidechain rotation is used to facilitate reaction and replaces much of the translational motion needed in solution. This reaction path has commonly evolved in serine proteases and in many additional enzymes catalyzing analogous reactions. By evaluating the energetic consequences of the molecular changes along the enzymatic and solution reaction paths, we identified enzyme features that provide catalysis and quantified their catalytic contributions. When summed, these features fully account for the reduction in reaction barrier provided by serine proteases. Further, our analyses revealed that a local structural motif referred to as the "nucleophilic elbow" (*48*) is responsible for the catalytic conformational features and has arisen in over 100 different enzymes across 32 distinct structural folds.

The ensemble–function approach has led to a quantitative catalytic model based on simple chemical and physical factors, providing a framework that connects complex biological processes to their underlying physicochemical principles. The in-depth, quantitative understanding obtained from this approach can transform how enzymology is taught in classrooms, and its simplicity and effectiveness in deriving quantitative biophysical models will make it broadly applicable to many future studies of macromolecular functions.

Results and discussion

Crystallographic data provides experimental ensemble information for serine proteases across reaction states

To obtain ensemble information for serine proteases, we leveraged the vast structural data available in the Protein Data Bank (PDB) to build "pseudo-ensembles" (*42*, *44–47*). These ensembles are built by bringing together repetitive structures of the same proteins so that they capture different states around the local minima of the overall conformational landscape. Pseudo-ensembles are in excellent agreement with nuclear magnetic resonance (NMR) order parameters (*46*) and residual dipolar coupling (*49*, *50*), and with multiconformer models obtained via room temperature X-ray diffraction (*42*), supporting the ability of pseudo-ensembles to capture a range of low energy states in crystalline and aqueous states. In particular, pseudo-ensembles provide three-dimensional atomic positions that is not accessible via NMR dynamics experiments and thus allow detailed interrogation of ensemble–function relationships. The extensive X-ray crystallographic data readily available in the PDB further allowed us to instantaneously obtain ensemble information for multiple enzymes. We observed consistent ensemble properties in the active sites of multiple serine proteases and in additional enzymes with analogous catalytic motifs, both within and across structural folds, leading to the discovery of the same catalytic features (discussed below). This consistency provides additional confidence in the conformational features and catalytic models obtained from the pseudo-ensembles.

Serine proteases are ideal for mechanistic pseudo-ensemble analysis, as thousands of X-ray structures representing multiple states along the enzymatic reaction path have been accumulated over decades of studies. To build pseudo-ensembles and to compare serine proteases evolved within and across structural folds, we collected 1231 wild-type structures of 17 serine proteases across four structural superfamilies (also referred to as "clans") from the PDB (Fig. 1C to D, figs. S1 to 4, and table S2). To examine serine protease ensembles across reaction states, we identified structures that are in the apo, GSA-bound, and TSA-bound states (fig. S5 and table S3). GSAs consist of non-covalent peptide (or peptide analog) inhibitors that bind in the same fashion as cognate substrates (supplementary text S2). TSAs are covalent inhibitors such as fluoromethyl ketones (*51*), boronic acids (*52–54*) and peptidyl aldehydes (*55*); they react

with the catalytic serine to form a tetrahedral intermediate (Fig. 1A) but do not react further. Tetrahedral intermediates resemble the geometry of the transition state, as found in prior quantum mechanics/molecular mechanics (QM/MM) studies of serine proteases and in our QM calculation for the reaction path in solution (fig. S6), thereby providing experimentally tractable models to mimic the transition state structure. Our analyses below using a range of bound GSAs and TSAs for multiple serine proteases and other enzymes revealed consistent conformational properties, suggesting that differences in the properties of individual analogs or crystal forms did not significantly distort the ensembles or the conclusions obtained[1]. The GSA- and TSA-bound ensembles provide information for the acylation step that is typically rate-limiting for amide substrates (*56–58*), thereby allowing us to relate ensembles to the rate enhancement ($k_{cat}/k_{solution}$) via comparisons to the uncatalyzed solution reaction. We first focused our analyses on clan PA serine proteases, which include trypsin, the serine protease with the greatest number of available structures (n = 446; Fig. 1C and fig. S1). Additional serine proteases evolved from different structural clans (SB, SE and SC; Fig. 1D, figs. S2 to 4) provided tests for whether the identified conformational features have convergently evolved.

Computational and crystallographic data provide information for the uncatalyzed solution reaction

Enzyme catalysis is a comparison of the enzyme-catalyzed reaction to the corresponding uncatalyzed reaction. To make this comparison, we used quantum mechanical calculations to obtain the reaction path for the analogous solution reaction. We also needed ensemble information about the solution reaction states to assess their energetics. To sample the ensemble distribution of reacting molecules in solution, we conducted molecular dynamics (MD) simulations of an analog of the substrate amide, *N*-methylacetamide (NMA) in water (with water acting as the nucleophile). Computational approaches provide all-atom information for a system, but they involve approximations and also may be limited in sampling (*59–62*). We therefore used a complementary, knowledge-based approach in which we assembled crystallographic data of small molecule interactions between amide and hydroxyl groups which resemble the reacting groups in the solution reaction. Knowledge-based distributions have been widely used to determine conformational preferences and derive energy functions (*63–100*), as pioneered by Ramachandran's maps of preferred protein backbone rotamers (*101*). These crystallographic distributions have been shown to quantitatively match those obtained from NMR experiments and QM calculations, supporting their ability to approximate ensembles of molecular interactions and determine conformational preferences (*74*, *102–105*) (see also supplementary text 3).

[1] Subtle changes on the scale of <0.1 Å were observed for the catalytic triad His•Asp hydrogen bonds for different TSA complexes. These changes are expected from their electronic properties and are consistent with NMR chemical shifts (supplementary text S8).

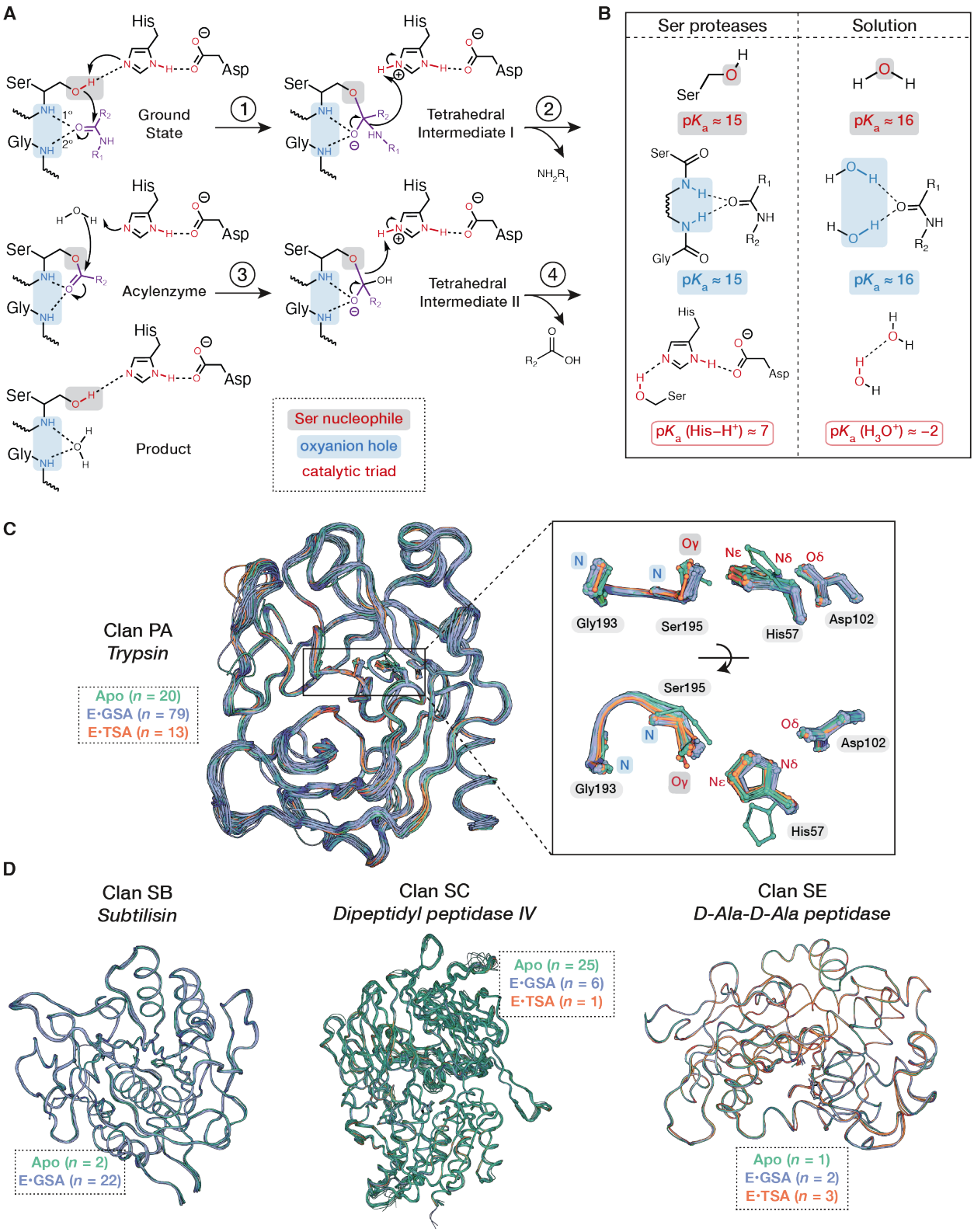

Fig. 1. Serine proteases. (A) The serine protease reaction mechanism and catalytic groups. The atoms involved in catalytic triad hydrogen bonds are colored in red, with the nucleophile shaded in gray; the oxyanion hole hydrogen bond donor atoms are colored and shaded in light blue. The 1° hydrogen bond donor is the backbone amide neighboring nucleophilic group (*106*). The 2°

hydrogen bond donor shown here is a glycine backbone amide, as found for trypsin and other clan PA serine proteases, but the 2° donor varies across different structural clans (table S2). Atoms from the substrate amide is colored in purple. **(B)** Comparison of chemical features of the enzyme catalytic groups compared to analogous interactions in the solution reaction. pK_a values are used as proxies for electron density (*107–110*). **(C)** Pseudo-ensemble of trypsin, a representative serine protease of clan PA, with the number of structures in the apo, GSA-bound and TSA-bound sub-ensembles annotated in the box with dashed outline. Structures were aligned on the Cα atoms of the protease chains using the highest resolution structure as the reference (table S3). Labeled atom names follow the PDB format (*111*). **(D)** Pseudo-ensembles of representative serine proteases across three additional structural clans: subtilisin (clan SB), dipeptidyl peptidase (clan SC) and D-Ala-D-Ala peptidase (clan SE), with the number of structures in their apo, GSA-bound and TSA-bound sub-ensembles annotated in the box with dotted outline. See figs. S1 to 4 for additional serine proteases in each clan and their active site ensembles.

Serine proteases precisely position reactants in the reactive orientation

Classic proposals of enzyme catalysis near-universally invoke the positioning of catalytic groups and substrates to facilitate reactions (table S1) (*2*, *24–41*). While it is easy to appreciate the advantage of bringing together substrates that are present at low concentrations in cells (micro to millimolar), the advantage, if any, for hydrolysis reactions is less clear, since water (55 M) is always nearby. To quantify the extent of positioning in serine proteases *versus* in solution, we compared the distributions of reactant positions in their ground state ensembles.

The position of the serine nucleophile relative to the substrate amide carbonyl is defined by three parameters, a distance (d_{attack}), an angle (α_{attack}) and a dihedral (ϕ_{attack}) (Fig. 2A). The observed distance and angles are narrowly distributed and highly similar for serine proteases across the four structural clans, with d_{attack} = 2.68 (*s.d.* = 0.14 Å), α_{attack} = 93° (*s.d.* = 7°) and ϕ_{attack} = 84° (*s.d.* = 8°) (Fig. 2B, fig. S7 and table S4)[2]. In contrast, the same parameters in solution, measured using the knowledge-based approach and MD simulation, were much broader (Fig. 2B and fig. S6). In addition, the distances found in enzymes are shorter than those in solution, a feature common to serine proteases across the four distinct structural clans. We return to this observation in later sections.

For the restricted positioning to aid catalysis, two additional criteria need to be met: (1) the nucleophile and the substrate must be positioned in reactive orientations[3] and (2) the motions needed for bond formation must be sufficiently unencumbered. As described below and in the following section, both criteria are met in serine protease active sites. Further, not only are the catalytically relevant motions unencumbered, but they are *favored* in comparison to other motions in the enzyme active site.

The observed enzyme angles α_{attack} (93°) and ϕ_{attack} (84°) agree with the preferred attack angles determined by Burgi and Dunitz from their analysis of small molecule O•C=O interactions ($\alpha_{\text{attack}} \approx 100°$, $\phi_{\text{attack}} \approx 90°$) (*112*) and those obtained from our QM calculations for the solution reaction path (α_{attack} = 82°, ϕ_{attack} = 84°; Fig. 2C, fig. S8 and table S4). The attack distance in the QM ground state (d_{attack} = 3.62 Å) agrees with the preferred distances observed in MD (3.46 Å, *s.d.* = 0.18 Å) and small molecule distributions (3.61 Å, *s.d.* = 0.54 Å), whereas the enzymatic d_{attack} is ~0.9 Å shorter (Fig. 2C and table S4), suggesting that the enzymatic reaction begins farther along the reaction coordinate.

[2] The highest probability value (i.e. mode) of the distribution are reported for geometries measured from ensembles here and below, as they represent the lowest free energy state.

[3] Reactive configurations are defined as configurations in the energy landscape that lie within a local minimum along the lowest barrier pathway to the transition state.

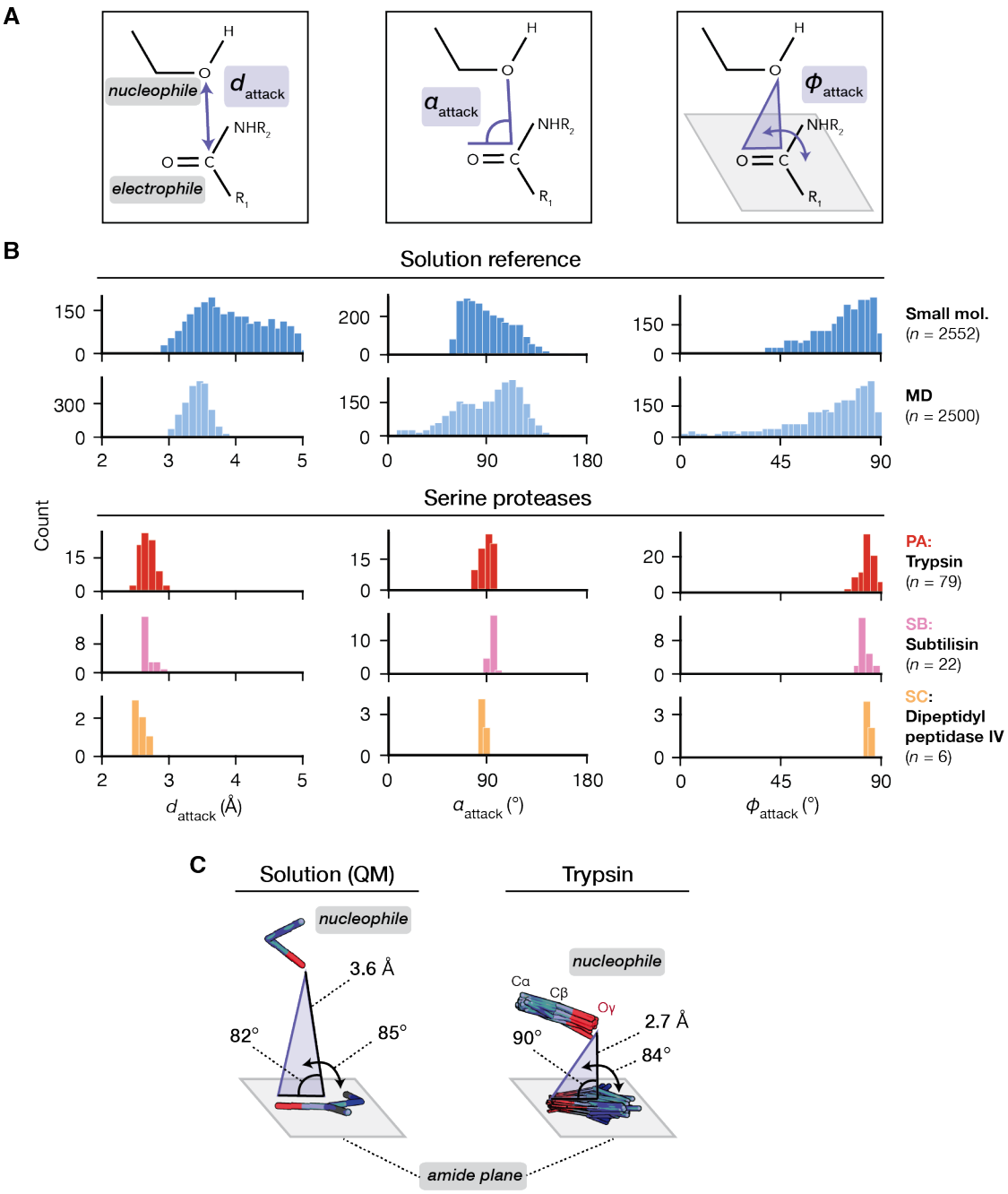

Fig. 2. Positioning of the reacting groups in serine proteases *versus* in solution. **(A)** The nucleophilic attack geometry is described by three parameters: the attack distance (d_{attack}), defined by the distance between the nucleophilic and the electrophilic atom; the attack angle (α_{attack}), defined by the angle between the nucleophilic atom, the electrophilic atom and the oxygen on the reactive carbonyl group; and ϕ_{attack}, defined by the dihedral angle between the two planes: (1) the plane defined by the nucleophilic atom, the electrophilic atom, and the oxygen on the amide carbonyl group (purple), and (2) the plane of the reactive amide carbonyl (gray). **(B)** Distributions of attack geometries in serine protease GSA-bound pseudo-ensembles *versus* in the reference ensembles for the solution reaction. The two reference distributions come from small molecule ("small mol.") hydroxyl•amide interactions or MD simulations of *N*-methyl-acetamide

(NMA) in water ("MD", 2500 snapshots). The differences in the tails of the MD and small molecule distributions arise from choosing only the water closest to the electrophile in each MD frame, which limited the sampling of conformations at longer distances in the MD distribution (see methods). The data for one representative protease from each structural clan are shown: trypsin for clan PA (red), subtilisin for clan SB (pink), and dipeptidyl peptidase for clan SC (yellow). Additional proteases within each clan are shown in fig. S7. **(C)** Ground state conformation of the reactants from QM calculation of the solution reaction compared to those observed in the trypsin pseudo-ensemble (n = 79). The trypsin pseudo-ensemble is locally aligned on the sidechain atoms of the catalytic triad. Ground state ensembles of additional proteases are shown in fig. S9. In QM, the nucleophile is ethanol, and the electrophilic amide is NMA.

Serine proteases provide a shortened and high-efficiency reaction path in comparison to the solution reaction

The notion that positioning can increase reaction probabilities is readily appreciated by considering two reactants constrained in a smaller box, equivalent to increasing their concentrations and thus their reaction probability. Reaction probabilities can be further increased by using a "box" complementary to the shape of the substrates (i.e., a binding site) so that the reacting groups of the substrates are more frequently aligned for reaction than in solution. However, complementary interactions can also limit reactant mobility and lower the probability of the motions needed for reaction. Hammes *et al.* articulated this point as follows: "For catalysis, flexible but not too flexible, as well as rigid but not too rigid, is essential. Specifically, the protein must be rigid enough to maintain the required structure, but flexible enough to permit atomic movements as the reaction proceeds." (*113*). Ensemble–function analysis allowed us to transform this general axiom into atomic-level models for the magnitude, direction, and energetics of the motions that are involved in an enzymatic reaction.

A sidechain rotation of the catalytic serine brings the reacting atoms closer in distance

To carry out an unbiased search for motions that may be part of the serine protease reaction path, we measured the backbone and sidechain rotameric angles of each residue in the GSA- and TSA-bound states and compared the distributions between these states (supplementary text S4). We also compared the apo and GSA-bound enzyme ensembles to determine the changes associated with substrate binding. From these comparisons, we derived a minimal model for the motions along the enzymatic reaction path.

Among all buried residues in trypsin, there are 370 rotameric angles, 26 of which have significantly different distributions upon substrate binding (apo *versus* GSA-bound state; $p < 0.05$ after multiple hypothesis correction), and 19 of which significantly changes upon nucleophilic attack (GSA- *versus* TSA-bound state) (Fig. 3A and fig. S10). Overall, the residues identified are spread throughout the structure rather than contiguous, suggesting no large-scale conformational changes or extensive motional coupling in the serine protease reaction[4].

[4] Many enzymes have much more extensive conformational changes upon substrate binding, such as loop closures and domain rotations to form the active site. These rearrangements are critical for catalysis by enzymes that undergo them as they are required to align catalytic groups, but these rearrangements are expected to occur on a slower timescale and thus prior to the sub-picosecond barrier crossing (*114*). The absence of such larger-scale motions for trypsin and several other serine proteases simplifies analysis of chemical catalysis, allowing us to identify and focus on motions that may be part of the reaction coordinate for covalent bond changes.

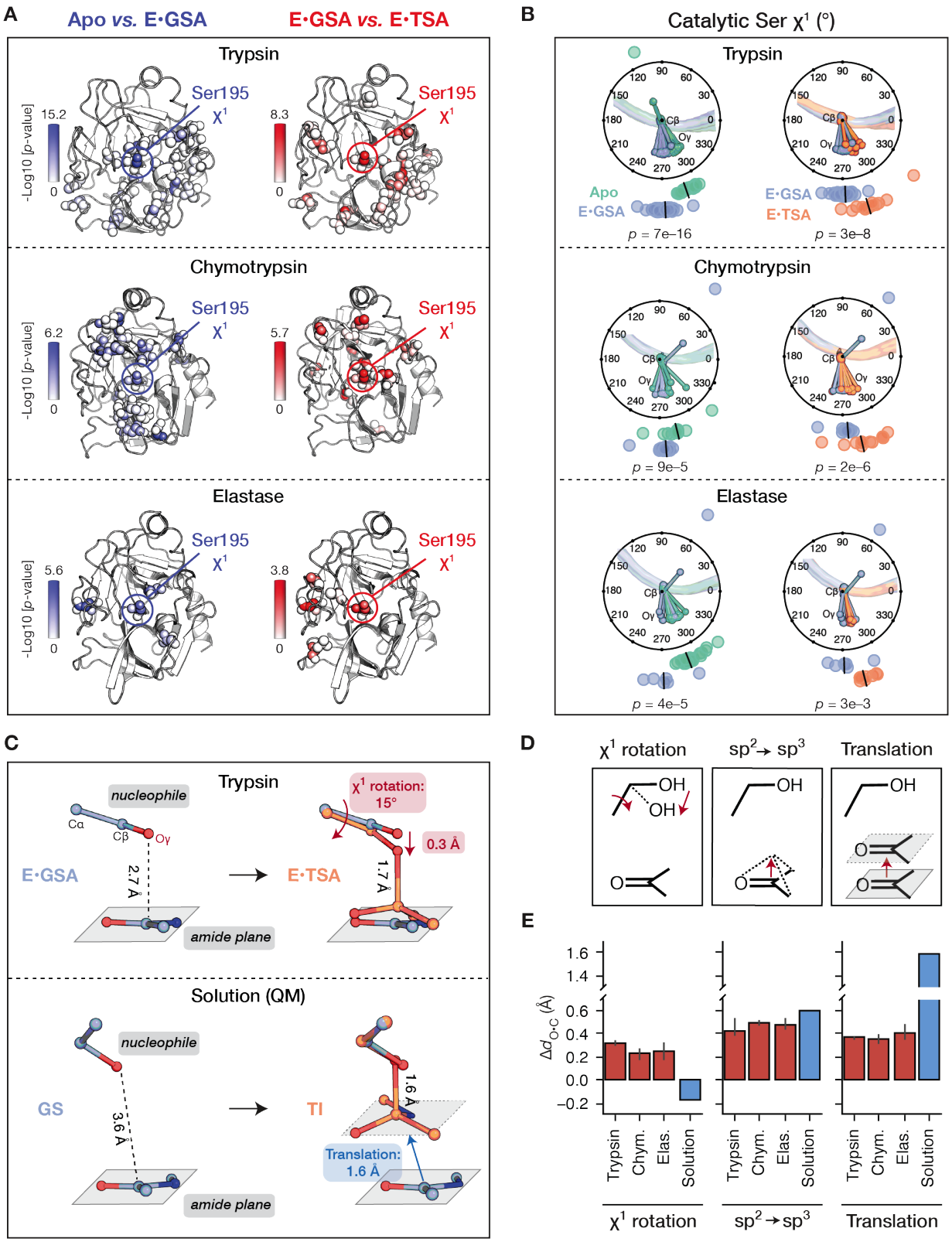

Fig. 3. The enzymatic and solution reaction path. (A) Trypsin, chymotrypsin and elastase structures colored by the significance [-Log10(*p*-value)] of the differences in rotameric angles between apo and GSA-bound (blue) and between GSA- and TSA-bound pseudo-ensembles (red), as determined by the Kolmogorov–Smirnov (K–S) test (*115*, *116*). Rotameric angles with *p*-

values <0.05 after multiple hypothesis correction are shown as spheres and colored; the color scales differ for each serine protease due to the large differences in *p*-values between the serine proteases. **(B)** Catalytic serine (Ser195) χ^1 angle distributions in serine protease pseudo-ensembles (apo: aqua; GSA-bound: blue; TSA-bound: orange). Black lines indicate medians of the distributions. *P*-values are determined by the K–S test; additional tests using the Mann–Whitney U test are shown in fig. S11. Ser195 sidechains are shown in Newman projections along the Cα–Cβ axis, aligned locally on the Ser195 backbone N, C, O and Ca and C*β* atoms. **(C)** Geometric changes of the reacting groups that accompany bond formation on trypsin and in the solution reaction. For trypsin, only one "center" structure from each pseudo-ensemble is shown (E•GSA: 3M7Q, E•TSA: 1TPP) for clearer visualization. The center structures were found by selecting the structure with the median Ser195 χ^1 value in the ensembles, and they were locally aligned on sidechain atoms of the catalytic triad. See fig. S15 for full pseudo-ensembles. For the QM calculated solution reaction, ground state (GS) and tetrahedral intermediate (TI) were aligned on the ethanol, imidazole and acetate atoms. **(D)** The catalytic serine χ^1 rotation, the sp^2 to sp^3 geometric transition of the substrate ($sp^2 \rightarrow sp^3$), and the translational motion between reacting groups shorten the distance between reacting atoms. **(E)** Contribution from each type of motion to the nucleophile•electrophile distance shortening ($\Delta d_{O \bullet C}$) calculated from trypsin, chymotrypsin and elastase pseudo-ensembles (red) *versus* in the QM-determined solution reaction (blue). Bar heights indicate median values and error bars indicate 95% confidence interval from the enzyme ensemble distributions (red).

Intriguingly, the sidechain rotation of the catalytic serine (Ser195 χ^1; rotation along the sidechain Cα–Cβ axis) showed highly significant changes in both the substrate binding (apo *vs.* GSA-bound) and the chemical step (GSA- *vs.* TSA-bound) (Fig. 3A and fig. S10). The shift in the Ser195 χ^1 angle is significant regardless of the types of TSAs that are bound (fig. S11). The same analysis carried out for chymotrypsin and elastase, two serine proteases in the same clan as trypsin (PA), also revealed significant changes in the catalytic serine χ^1 (Fig. 3A and fig. S10 to 12). This change was the only highly significant rotamer motion in both the binding and the chemical step *common to all three proteases* and was the most significant (or among the most significant) in all cases (Fig. 3A and fig. S12 to 13). The direction and magnitude of the rotation were also common[5]. Upon substrate binding, the sidechain of the catalytic serine rotates by −17°, and it rotates back by about the same amount (+13°) in going from the GSA-bound to the TSA-bound state (Fig. 3B and table S5).

Trypsin shortens the reaction path and uses a catalytic serine rotation to replace translational motion

The serine χ^1 rotation from the GSA- to the TSA-bound state brings the nucleophilic oxygen closer to the electrophilic carbon by ~0.3 Å, consistent with a motion along the reaction coordinate (Fig. 3C, fig. S14 to 16 and table S6)[6]. The distance between the serine nucleophile and the substrate amide carbon progresses from ~2.7 Å in the ground state to form a ~1.6 Å covalent bond[7], requiring a total of ~1.1 Å distance change. The rest of the distance shortening must derive from geometric changes on the substrate or relative translational motion between the reacting groups. Comparisons of the GSA- and TSA-bound pseudo-ensembles allowed us to quantify the distance changes that arise from these two factors, accounting for the rest ~0.8 Å motion needed for bond formation (Fig. 3D to E and table S6).

The substrate amide "pyramidalization" from a sp^2 carbonyl to a sp^3 tetrahedral intermediate contributes 0.42 ± 0.09 Å to bond formation in trypsin (table S6), measured by converting the pyramidal TSA-bound structures to mimic their planar unreacted state. This amount of change agrees with the prediction from an empirical relationship derived by Burgi *et al.* (*112*). Translational motion contributes 0.36 ± 0.22 Å, determined by aligning pairs of GSA-bound and TSA-bound trypsin structures locally at the active site. The same types and extents of motions were found for other clan PA serine proteases (Fig. 3E, figs S14 to 15 and table S6), suggesting conservation of the core dynamic features of the enzymatic reaction path.

QM calculations revealed a longer distance traveled by the reacting atoms in the solution reaction path (2.0 Å instead of 1.1 Å), as the reactants begin farther apart in the ground state (Fig. 3C). Substrate pyramidalization contributes the same amount in the enzymatic and solution reactions, as expected for a covalent change. Yet, the solution path does not involve bond

[5] Thrombin is an exception. Its catalytic serine exhibited smaller rotations that did not pass a statistical significance test (table S5). Thrombin's highly flexible active site (fig. S1 and 8) and its ability to respond to allosteric ligands may obscure changes that occur along the reaction path (*117*).

[6] There are likely additional motions that accompany the geometric and electrostatic changes along the reaction coordinate, as there are in solution. These may include changes too small to detect from our ensembles and changes that are idiosyncratic to each serine protease. Our minimal catalytic model assumes that these changes have similar energetic consequences as those in solution and thus do not differentially affect the enzymatic and uncatalyzed reactions. Our ability to fully account for serine protease catalysis suggests that this is the case, but of course requires further testing as does any model.

[7] The most commonly cited C–O covalent bond length is 1.43 Å (*118*), but C–O bond lengths vary depending on the molecular context and conformation (*119–121*). The C–O bond length of the tetrahedral intermediate was determined to be 1.60 Å in the QM calculations herein, and TSA-bound structures (where the attacked electrophile is carbon) gave similar bond lengths of 1.56 Å (*s.d.* = 0.19 Å) (table S7).

rotation and is instead dominated by a 1.6 Å translational motion, corresponding to 80% of the total reaction distance (Fig. 3C and E). Thus, the enzyme shortens the reaction coordinate and removes ~75% of the translational motion, replacing a portion of the three-dimensional translation with a bond rotation along a single dimension that appears to coincide with the reaction path. The serine protease active site therefore provides a low-entropy, high-probability reaction path (see also *Conclusions and Implications* and Fig. 8).

Common evolution of a short enzymatic reaction path via bond rotation across different structural contexts

We next turned to serine proteases evolved from four distinct structural clans (where there is sufficient structural data) to ask whether the structural convergence of their catalytic groups was accompanied by the convergence of their ensemble features and reaction paths. Rotameric analysis for clans SB, SE, and SC proteases revealed changes along the reaction cycle analogous to those observed for clan PA (Fig. 4A, figs. S17 to 19 and tables S5 to 6). The rotations observed for clan SB were statistically significant but smaller in magnitude than those for clan PA, possibly arising from loss of signal from averaging across multiple proteases, as is required due to the sparsity of data for individual clan SB members, or from a quantitatively smaller contribution of the bond rotation to the reaction coordinate. For clan SE, comparisons of a smaller number of structures suggest the same catalytic serine rotations, although there is insufficient structural data to carry out statistical tests (figs. S17 to 19 and tables S5 to 6).

Clan SC, also known as the α/*β* hydrolase fold (*122*), contains enzymes with a catalytic triad and oxyanion hole like all serine proteases but uses a different oxyanion hole hydrogen bond donor that requires the serine to sit in a different rotameric state (*trans vs. gauche–,* as defined in Fig. 4A) (*106*). The clan SC catalytic serines undergo a rotameric cycle analogous to the other protease clans, but in the opposite direction, as is needed to bring the nucleophilic oxygen closer to the electrophilic carbon for bond formation (Fig. 4A to B and fig. S20). The same catalytic outcome from this opposite motion strongly supports convergent evolution of the serine rotation to provide a common catalytic role.

The same reaction path was also found for non-protease members of clan SC that carry out nucleophilic attack on carbonyl centers. For acetylcholinesterase, a clan SC member with sufficient structures to evaluate its catalytic cycle, the reaction initiates from a short 2.42 Å (*s.d.* = 0.16 Å) distance in the ground state and involves the same ~15° catalytic serine rotation (Fig. 4B). Additional non-protease enzymes from clan SC, when combined, also revealed short attack distances (fig. S21) and the serine rotation (Fig. 4A).

Overall, the conservation and convergent evolution of the enzymatic reaction paths strongly suggests common catalytic strategies that are provided by common conformational features. In the next sections, we quantitatively assessed the catalytic contribution from these and additional conformational features.

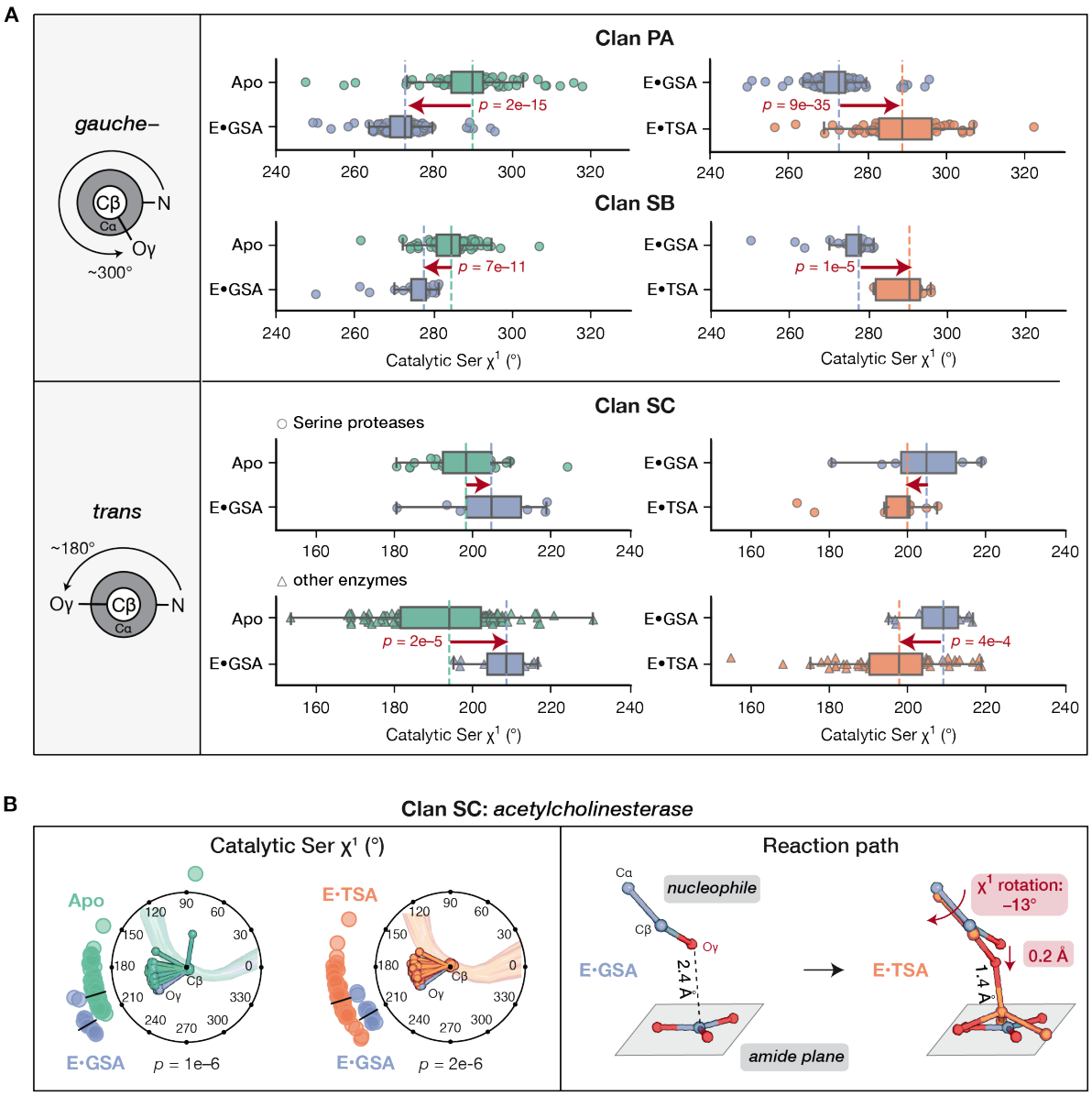

Fig. 4. Common features of the enzymatic reaction paths across serine proteases and non-protease enzymes from multiple structural clans. (A) Rotameric changes of the catalytic serine χ^1 across reaction states from serine proteases and additional enzymes from distinct structural clans. The reactive rotameric state for each clan is shown (*gauche–* for clan PA and SB; *trans* for clan SC). Dashed lines indicate the median value of each distribution (apo: aqua; GSA-bound: blue; TSA-bound: orange). The full distributions including rare outliers in unreactive rotamer states are shown in fig. S15; these outliers were included for *p*-value calculations. Arrows indicate the direction of change; *p*-values from the K–S tests are given for statistically significant changes ($p < 0.05$). See table S5 for the list of individual enzymes included in each distribution. **(B)** Geometric changes of acetylcholinesterase across reaction states. Left: catalytic serine χ^1 angle distribution in the apo, GSA-bound and TSA-bound states; right: alignment of the center structures from each pseudo-ensemble shown (E•GSA: 2GYW, E•TSA:2HA0). See fig. S17 for full pseudo-ensembles. See Fig. 3B to C legend for details.

A comprehensive, quantitative model for serine protease catalysis

The functional groups in serine protease active sites have chemical properties similar to water, with the exception of the His general base that abstracts a proton from serine to increase its reactivity and the catalytic triad His•Asp hydrogen bond that is absent in the solution reaction (Fig. 1B). Empirical linear free energy relationships (LFERs) from model systems provided estimates for the catalytic contributions from these two chemical features. The general base histidine and its positioning are estimated to provide 5.1 ± 2.6 kcal/mol catalysis based on LFERs from amide hydrolysis reactions (*123*, *124*) and effective molarities of intramolecular general base catalysis (*125–128*), corresponding to a ~10^4-fold rate enhancement (supplementary text S5, fig. S22 and table S8). The His•Asp hydrogen bond strengthens in going from the ground to the transition state as positive charge accumulates on the histidine, giving an estimated 0.8 kcal/mol catalytic contribution, as described below (*Catalysis from the catalytic triad hydrogen bonds*). While there are uncertainties associated with these values, they are far insufficient to account for the reduction in reaction barrier provided by serine proteases (5.9 kcal/mol predicted *vs*. 13.3 kcal/mol observed[8]) (*17*), strongly suggesting the presence of additional catalytic contributions.

We determined the energetic consequences from each of the conformational features of the enzymatic and solution reaction paths identified above. These calculations established the energetic origins for the catalysis provided by these novel features. As exemplified by trypsin below, the sum of their energetics quantitatively accounts for the remaining 7.4 kcal/mol catalysis (Fig. 5A). The same conformational features and catalytic contributions were found for multiple serine proteases across distinct structural folds, providing additional confidence in the catalytic features and suggesting that the same features have arisen multiple times in evolution.

Catalysis from the positioning of reactants

Our comparison of the ground state ensembles of the reactants in trypsin active sites *versus* in solution showed that the enzyme distributions are much more constrained (Fig. 2B). The higher level of positioning on the enzymes can contribute to catalysis by increasing the probability of reaching the highly restricted transition state (Fig. 5B). To quantify the difference in positioning, we considered the relative nucleophile•amide positions in three-dimensional space and calculated the conformational entropy (S_{conf}) associated with the enzyme and solution distributions, using the formalism of Gibbs entropy (*132–143*) (Fig. 5C). The trypsin and the "solution" distribution from small molecules differed in $T{\cdot}S_{conf}$ (T = 298 K) by 2.2 ± 0.2 kcal/mol (Fig. 5B to C and table S9)[9]; the same difference was found between the trypsin and the solution distribution from our MD simulation (fig. S23). Bootstrap resampling and microstate binning controls indicated that the data provided consistent entropy values (fig. S23 and table S9).

As the enzymatic groups are positioned in the reactive orientation for nucleophilic attack (Fig. 2C), the greater conformational restriction (lower entropy) on the enzyme relative to solution is expected to increase the probability of reaction. The 2.2 kcal/mol difference in the $T{\cdot}S_{conf}$ values corresponds to a ~40-fold rate enhancement from reactant positioning, despite the presence of 55 M water for the solution reaction (Fig. 5A, pink).

---

[8] Catalysis of trypsin was measured to be 13.3 kcal/mol for the amide substrate Z-Gly-Pro-Arg-AMC; other serine proteases gave similar k_{cat} values (within 10-fold) for amide substrates with their preferred specificity, such as subtilisin (*19*), chymotrypsin (*129*), α–Lytic protease (*130*), elastase (*131*), *et al*.

[9] For each energetic calculation, mean is reported followed by the standard deviation after the ± sign.

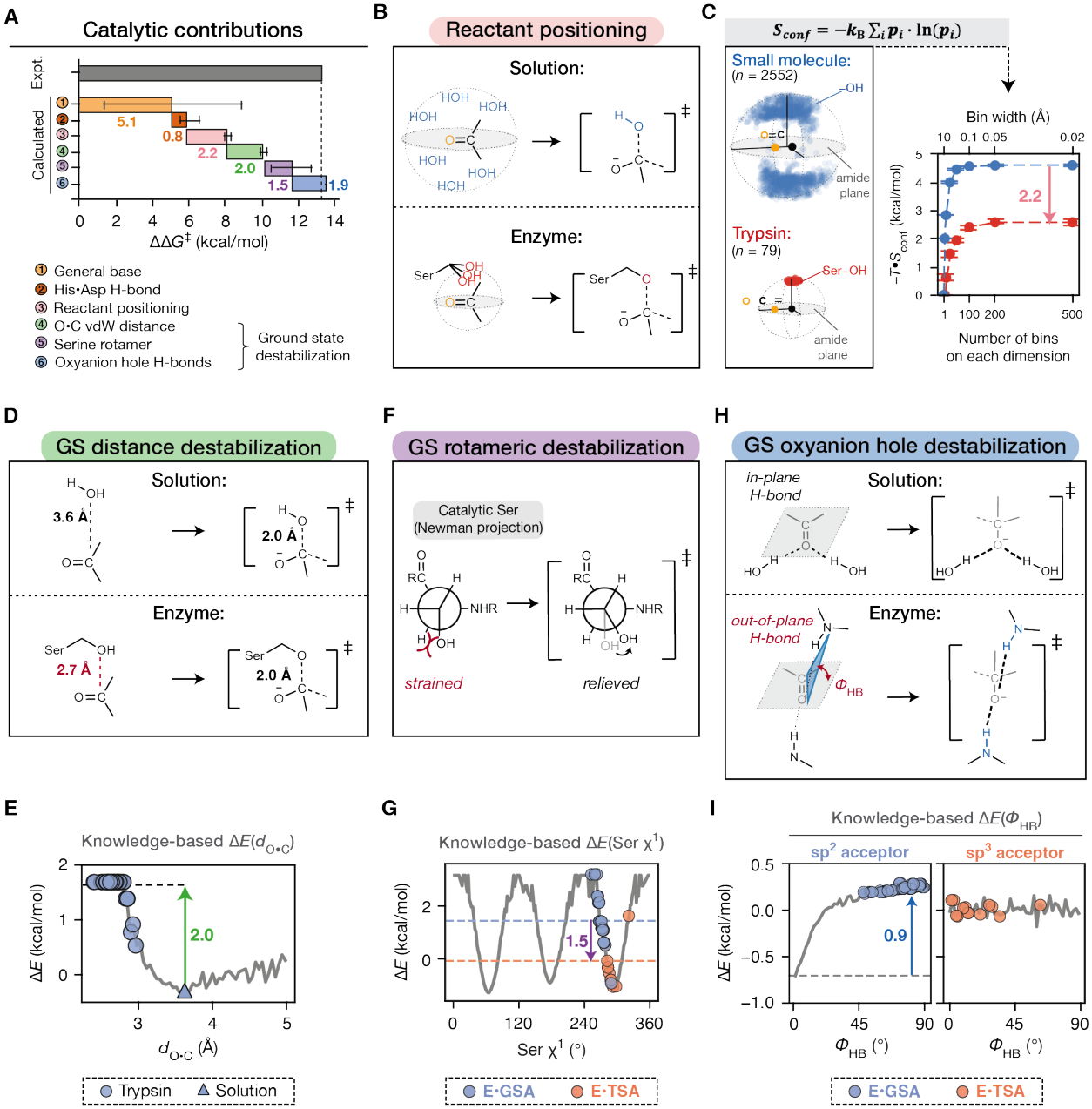

Fig. 5. Catalytic features and quantitative estimates for their energetic contributions to serine protease catalysis, derived from the trypsin pseudo-ensemble data. (See fig. S25 to 26 and S30 for the same features in other serine proteases.) **(A)** The sum of the energetic contribution from each identified catalytic feature accounts for the total rate enhancement from trypsin [13.3 kcal/mol from experiment (*19*) (gray bar and dashed line) *versus* 13.5 kcal/mol from our calculations (colored bars, with each color corresponding to one catalytic feature)]. The error bars for general base catalysis and the His•Asp hydrogen bond are calculated by using a range of measured LFERs (see supplementary text S5 and 7); the error bar for reactant positioning indicate the standard deviation from entropy calculations using a range of bin sizes from (2 Å)^3 to (0.02 Å)^3 (see 5C and table S9). The error bars for ground state destabilization mechanisms indicate the standard deviations from the calculated energies shown in 3E, G and I (see also tables S10 to 11 and 13). **(B)** The positioning of reactants on the enzyme compared to their near-free motions in solution provides entropic advantages to get to the highly restricted

transition state. **(C)** Left: trypsin pseudo-ensemble nucleophile•amide distributions (red) *versus* from small molecule hydroxyl•amide (blue) distributions. The spherical distributions use a coordinate system where the electrophilic carbon is at the origin and the substrate amide plane aligns with the xy plane. Calculations were performed by splitting the distributions into a range of number of bins (microstates) from 1 to 500 on each dimension, corresponding to bin sizes from (10 Å)^3 to (0.02 Å)^3. Right: conformational entropies ($-T \bullet S_{conf}$ at $T = 298$ K) calculated from these distributions using the formalism of Gibbs' entropy (equation shown in the gray shaded box, where k_B is the Boltzmann constant and p_i is the probability of observing the system at a microstate i). Error bars indicate standard deviations obtained via bootstrapping (table S9). **(D)** Ground state destabilization of the reacting atom distance ($d_{O \bullet C}$) on the enzyme, a catalytic feature that is not present in solution. **(E)** Catalytic contribution from ground state destabilization of the short O•C van der Waals interaction. The knowledge-based energy function for O•C van der Waals distance ($d_{O \bullet C}$) built from 2,552 small molecule structures of amide•hydroxyl interactions (Fig. 2B) is shown as the gray solid line. The distances observed in the trypsin GSA-bound pseudo-ensemble (blue circles) and the solution ground state distance from our QM calculation (blue triangle) were mapped on the energy function. The black dashed line indicates the mean of calculated energies from the trypsin GSA-bound ensemble. The green arrow and number indicate the difference between the energies calculated from trypsin and that from solution QM. **(F)** The catalytic serine is in a partially eclipsed, destabilized rotamer in the ground state that is relieved in the transition state. **(G)** Catalytic contribution from ground state destabilization of the serine sidechain rotamer. The knowledge-based energy function for serine χ^1 built from 41,237 serine residues from the BBDEP2010 database (*144*) is shown as the gray solid line. Catalytic serine χ^1 values from the trypsin pseudo-ensembles were mapped on the energy function (GSA-bound: blue; TSA-bound: orange). Dashed blue and orange lines indicate the mean of the calculated energies for all members in each pseudo-ensemble. The purple arrow and value indicate the difference between the means of the GSA-bound and TSA-bound energies. **(H)** Solution amide•carbonyl hydrogen bonds prefer an in-plane orientation whereas the enzymatic ground state is destabilized with out-of-plane oxyanion hole hydrogen bonds. The planarity of a hydrogen bond is quantified by ϕ_{HB}, the dihedral angle between the two planes: (1) the plane defined by the donor (N), the acceptor (carbonyl O) and the carbonyl C (blue) and (2) the carbonyl plane (gray). **(I)** Catalytic contribution from ground state destabilization of the 1° oxyanion hole hydrogen bond. The knowledge-based energy functions of ϕ_{HB} built from 60,018 small molecule amide•carbonyl hydrogen bonds (sp^2 acceptor) and from 5,823 hydrogen bonds between amide and sp^3 oxygens (sp^3 acceptor) are shown as gray solid lines. The ϕ_{HB} values from trypsin pseudo-ensembles for the 1° oxyanion hole hydrogen bonds were mapped on the energy function (GSA-bound: blue; TSA-bound: orange). See fig. S30 for the 2° hydrogen bonds. The blue arrow represents the amount of destabilization in the ground state.

Catalysis from ground state rotameric and distance destabilization

Our ensemble analyses uncovered a short nucleophile•electrophile distance and rotation of the catalytic serine, conformational features that are present in the enzymatic reaction but not in the solution reaction (Fig. 3C). To determine their energetic consequences, we constructed knowledge-based energy functions for each molecular feature from small molecule crystallographic data, using standard Boltzmann statistics to convert observed geometric distributions into relative energies (supplementary text S3).

The nucleophile•electrophile distance on serine proteases is much shorter than the favored O•C van der Waals interaction distance (2.68 Å *vs.* 3.61 Å) (Fig. 2B to C, fig. S7 and table S4) and is thus destabilized. To quantify the amount of destabilization, we constructed a knowledge-based energy function from the distance distribution of small molecule hydroxy•amide interactions (Fig. 5E). This energy function gave a preferred distance of 3.61 Å, matching the distance found in the solution ground state from MD and QM (Fig. 2B to C and table S4). The shorter nucleophile•electrophile distances in trypsin corresponds to a destabilization of 2.0 $\pm$ 0.2 kcal/mol (table S10). Bootstrap analyses resampling datasets of smaller sample sizes from the small molecule data generated consistent energy functions that gave the same amount of destabilization, providing confidence for the statistical stability of these functions (fig. S24A to B). However, since high energy regions are less sufficiently sampled in crystallographic distributions, the destabilization may be higher.

This ground state destabilization is expected to be relieved in the transition state and the high-energy tetrahedral intermediate where partial and full covalent bonds are formed with the same bond lengths as in solution (Fig. 5D). Prior QM/MM studies of serine proteases (*145–147*) and our solution QM calculations give the same transition state nucleophile•electrophile distance (2.0 Å); TSA-bound structures also gave the same bond lengths as that of the intermediate state in our QM calculation (1.56 Å *vs.* 1.60 Å; table S7). Therefore, the ground state destabilization of the O•C van der Waals reduces the reaction barrier by an estimated 2.0 kcal/mol, corresponding to a 30-fold rate enhancement (Fig. 5E); other serine proteases gave the same amount of destabilization (fig. S25 and table S10).

The energetic consequence of the catalytic serine rotation was determined using an analogous approach (Fig. 5F to G and fig. S24C to D). A knowledge-based energy function for serine sidechain χ^1 angles was built using 41,237 serine residues from high quality PDB structures (*144*), reproducing the well-known preferences for staggered over eclipsed conformations (*148*, *149*) (Fig. 5G). Mapping the catalytic serine χ^1 angles from pseudo-ensembles onto this energy function revealed a semi-eclipsed serine sidechain conformation in the ground state that is destabilized by 1.5 $\pm$ 1.1 kcal/mol (table S11). In the TSA-bound state, the serine rotates to the favored staggered conformation, thereby relieving the destabilization (Fig. 5G, fig. S26 and table S11). In contrast, no significant bond rotations occur for the nucleophile in the corresponding QM solution reaction (Fig. 3C). Thus, the relief of the enzymatic rotameric ground state destabilization contributes 1.5 kcal/mol to catalysis, corresponding to a rate enhancement of ~10-fold (Fig. 5A, purple).

Catalysis from ground state destabilization of oxyanion hole hydrogen bonds

Previous studies by Simón and Goodman (*150*, *151*) found that enzyme oxyanion hole hydrogen bonds are generally out-of-plane from the carbonyl groups of the bound ligands and therefore unable to interact with the oxygen lone pairs in optimal orientations in the ground state (Fig. 5H). Their computation also suggested that this destabilization was lessened in the oxyanionic transition state where no lone pair directionality is expected. Our analysis of

individual serine proteases with carefully curated GSA-bound structures supports the observation of out-of-plane hydrogen bonds at the oxyanion hole for trypsin, other members of its structural clan (PA), serine proteases from other clans (supplementary text S6, figs. S27 to 32 and tables S12 to 13). These proteases have preferred dihedral angles (ϕ_{HB}) of 76° (*s.d.* = 10°) and 76° (*s.d.* = 5°) for the two oxyanion hole hydrogen bonds, where ϕ_{HB} = 0° is in-plane and ϕ_{HB} = 90° is orthogonal to the carbonyl plane (Fig. 5H and fig. S28). As expected, knowledge-based distribution for small molecule amide•carbonyl hydrogen bonds donated to the sp^2 hybridized oxygen showed a preference for in-plane orientations (fig. S28 and table S12). The energy function for ϕ_{HB} derived from this distribution gave a destabilization of 0.94 ± 0.03 and 0.95 ± 0.02 kcal/mol from the suboptimal orientation for the two oxyanion hole hydrogen bonds in trypsin (Fig. 5I, sp^2); similar values were obtained for other serine proteases (fig. S30 and table S13).

In contrast, no such destabilization is present for the TSA-bound states, as there is no angular preference for hydrogen bonds between amide–NH and sp^3-hybridized oxygens (Fig. 5I, sp^3 and fig. S29 to 30), presumably reflecting the near symmetric electron distributions around sp^3 oxyanions. Thus, the two destabilized oxyanion hole hydrogen bonds in trypsin provide a catalytic contribution of 1.9 kcal/mol, corresponding to a rate enhancement of 25-fold (Fig. 5A, blue). Given its occurrence in multiple serine proteases (supplementary text S6, figs. S28 and 30 and tables S12 and 13C), the oxyanion hole ground state hydrogen bond destabilization appears to be conserved and convergently evolved, like the serine χ^1 rotameric destabilization and the shortened reaction distance.

Catalysis from the catalytic triad hydrogen bonds

Unlike the oxyanion hole hydrogen bonds, the catalytic triad hydrogen bonds (His•Ser and His•Asp) involve groups that are stronger hydrogen bond donors or acceptors than water, and therefore can provide catalysis based on their inherent properties without geometric destabilization. The catalytic serine and histidine undergo proton transfer concerted with the nucleophilic attack (*152*); thus, any catalytic contribution from the His•Ser hydrogen bond is in principle included as part of the general base catalysis term described above (supplementary text S5 and fig. S22). The His•Asp hydrogen bond strengthens going from the ground to the transition state as the histidine becomes protonated. The catalytic triad Asp is a stronger hydrogen bond acceptor than water, due to the higher electron density (pK_a) on a carboxylate than water oxygen atom. As the increase in hydrogen bond strength is sensitive to the electron density on the donor and acceptor, we expect the His•Asp hydrogen bond strengthens more than a His•water hydrogen bond during the reaction. The increase in the His•Asp hydrogen bond strength can be estimated from the pK_a values of the hydrogen bonded partners using a linear free energy relationship for hydrogen bond energetics (*153–155*), which gives a 0.8 kcal/mol contribution or a 4-fold rate enhancement (supplementary text S7 and fig. S33).

The catalytic triad hydrogen bonds could provide additional catalysis via geometric ground state destabilization, as was observed for the oxyanion hole hydrogen bonds. However, pseudo-ensemble comparisons suggest little or no geometric distortion for either hydrogen bonds (supplementary text S8, figs. S34 to 43 and tables S14 to 17). While prior studies emphasized the potential catalytic role of a short His•Asp hydrogen bond (supplementary text S1), the short His•Asp hydrogen bond (2.71 Å, *s.d.* = 0.11 Å) in the TSA-bound state matches the optimal distance predicted from analogous imidazole•carboxylate hydrogen bonds in small molecules (2.69 Å, *s.d.* = 0.14 Å) and the average His•Asp hydrogen bonds in the PDB (2.73 Å, *s.d.* = 0.39 Å) (table S14A), providing no indication of special geometric properties that facilitate catalysis.

The serine protease catalytic ledger

Our ensemble–function analyses allowed us to assemble an energetic ledger for the serine protease *versus* solution reaction. Catalytic contributions from (1) general base catalysis, (2) the catalytic triad His•Asp hydrogen bond, (3) positioning of the reactants, and (4) ground state destabilization of the serine rotamer, (5) the reactant distance, and (6) the oxyanion hole hydrogen bonds sum to fully account for serine protease catalysis (Fig. 5A). We expect these contributions to be additive, as each is obtained from a local, individual property (supplementary text S9 and fig S44).

This catalytic ledger provides a *minimal* complete model for serine protease catalysis that sums to 13.5 $\pm$ 4.0 kcal/mol that is within error of the experimentally measured catalysis of 13.3 kcal/mol (*17*). This model relies on GSA-bound and TSA-bound complexes to provide adequate information about the actual reaction states. While no tetrahedral analog perfectly matches a transition state, the resemblance between their geometries found in QM (fig. S6) suggests that this approximation is reasonable; the catalytic contribution from destabilizing the ground state will be smaller if relieved to a lesser extent in the actual transition state than with the bound tetrahedral analogs. More generally, the consistency of conformational features found with multiple TSAs and across multiple serine proteases strongly supports the catalytic features identified herein, as it would be unlikely for the same features to repeatedly arise from noise or model errors across many, widely varying examples.

The ability to detect catalytic features also depends on the size and quality of the available ensemble information. We cannot rule out additional changes that are too small to be resolved by our pseudo-ensembles. Nevertheless, these pseudo-ensembles detected subtle differences in catalytic triad hydrogen bond lengths (~0.1 Å) with different types of bound TSAs. These differences are consistent with NMR chemical shift data (supplementary text S8 and fig. S43), supporting the ability of enzyme pseudo-ensembles to accurately identify subtle conformational differences.

Our rotameric analyses revealed only a small number of changes beyond the active site (Fig. 3A), and these changes are not conserved (fig. S13) and yield minimal energetic consequences (fig. S45). In line with a minimal model, we do not invoke additional, more complex catalytic mechanisms involving long-range effects [e.g. electrostatic field (*29*, *31*, *156–158*), or ultra-fast events associated with barrier crossing (*159–163*)]. Future minimal models for other enzymes that account for less than the observed catalysis may suggest the presence of these or other additional catalytic contributions that would require further investigation. Regardless, all models, whether seemingly complete or not and whether from experiment or computation, require independent experiments that test incisive predictions from these models.

In the following sections, we explore the conformational interplay responsible for the catalytic features. These analyses provide additional independent support for the serine protease catalytic model and led us to a structural motif, the "nucleophilic elbow", that underlies the generality of catalytic features across serine proteases and in many additional enzymes.

Enzyme architectural elements that underlie positioning and ground state destabilization

Extensive catalysis (7.4 kcal/mol) is provided by the conformational disposition of serine proteases to constrain their reactants in a destabilized ground state and to relieve the destabilizations in the transition state. We wanted to understand *how* these locally destabilized conformers are maintained in the global energy minimum. In particular, we sought to identify the enzyme interactions that are responsible for enforcing the catalytic conformations in trypsin and

determine whether their combined energy favors the observed local destabilizations of the catalytic groups.

Destabilization of the unreactive rotameric states of the catalytic serine

Free serine has three local minima ("wells") for its sidechain rotamer that are nearly isoenergetic (Fig. 5G). Yet, 24 of the 25 members of the trypsin apo pseudo-ensemble have their catalytic serines in the reactive rotameric well (96%; fig. S46), suggesting destabilization of the alternative wells. Modeling rotational states of the serine sidechain revealed possible origins for their destabilization. Formation of the unreactive *gauche*+ conformer breaks the hydrogen bond with His57 and results in a steric clash with the Cys58–Cys42 disulfide bond (Fig. 6A). This disulfide bond is >97% conserved for clan PA serine proteases and gives deleterious rate effects upon mutation (*164*), supporting a catalytic role. Formation of the other unreactive conformer (*trans*) results in a steric clash with H57 of the catalytic triad, and relief of that clash would break the catalytic triad Ser•His and His•Asp hydrogen bonds (Fig. 6A). Indeed, the sole outlier structure adopting this rotamer has His57 rotated out of the active site and lacks these hydrogen bonds (fig. S46).

Destabilization within the reactive rotamer of the catalytic serine

Upon binding GSAs, the serine sidechain in trypsin rotates 18° to a semi-eclipsed, destabilized state within the *gauche*– well (Fig. 3B). Without this rotation, the serine oxygen (nucleophile) would be only ~2.3 Å away from the amide carbon (electrophile), a highly unfavorable distance (Fig. 6B and fig. S47A). Rotating the serine sidechain increases this O•C distance to ~2.7 Å, partially relieving the distance destabilization. Supporting this interplay, GSA-bound pseudo-ensembles revealed a correlation between the serine χ^1 angles and the O•C distances (Pearson $r = -0.5$, $p = 2e{-6}$), such that the rotation about χ^1 to destabilized states lengthens (and lessens the destabilization in) the O•C distance (Fig. 6C and fig. S48). Further rotation in the *gauche*– well to the boundary of this rotameric well (240°) would increase the distance to ~3.4 Å to nearly completely relieve the van der Waals destabilization (fig. S47A), yet this further rotation does not occur. Knowledge-based energy functions suggest that the cost of the rotameric destabilization is greater than the energy returned from fully relieving the distance destabilization (+4.2 kcal/mol *vs.* –2.4 kcal/mol; fig. S47B).

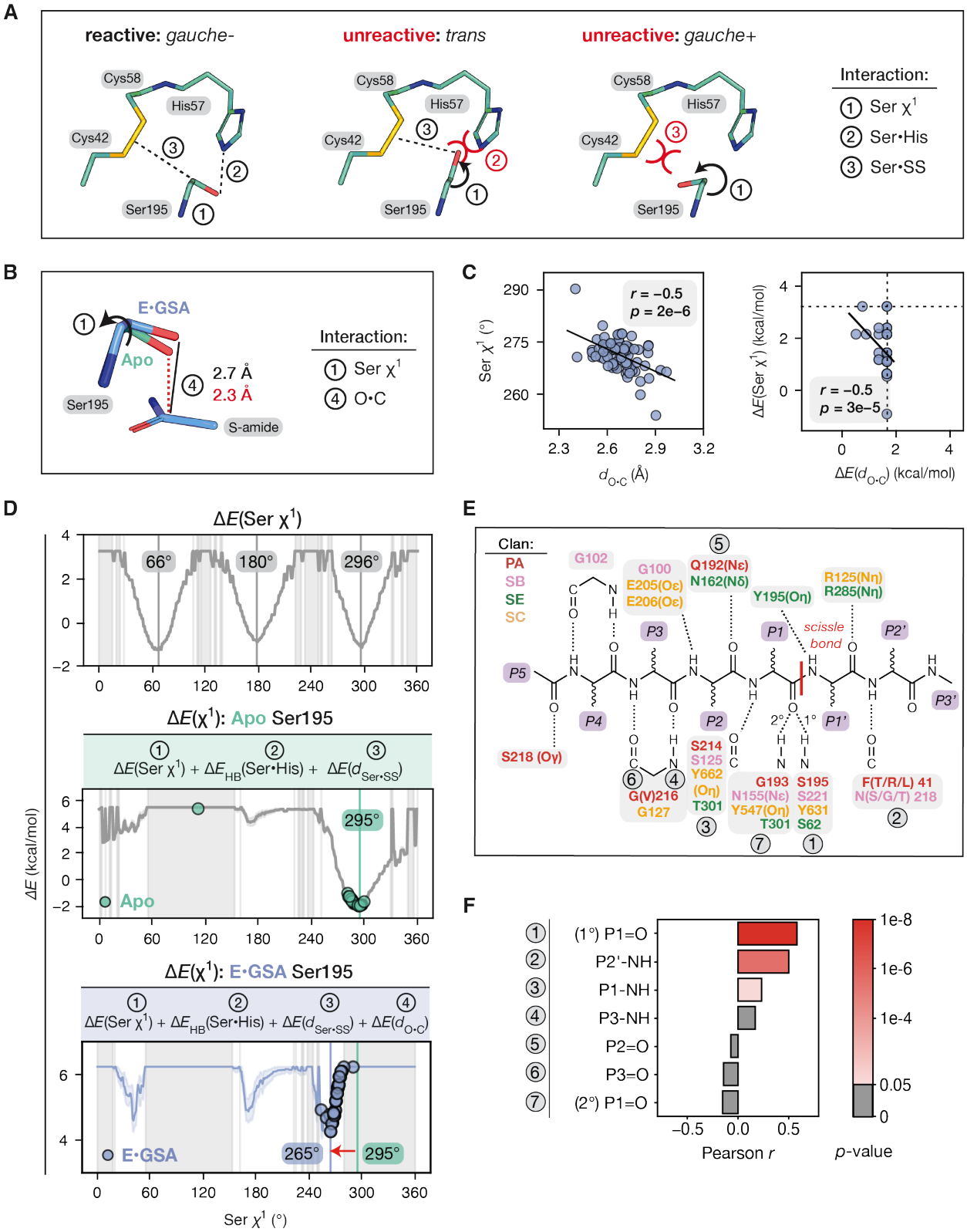

Fig. 6. Interactions at the active site and the substrate binding pocket enforce ground state destabilization. (A) Interactions surrounding the catalytic serine restrict the catalytic serine to the reactive *gauche*– rotamer well (trypsin, PDB: 5MNE). These interactions are (1) the catalytic serine sidechain χ^1 rotameric state [Ser χ^1], (2) the catalytic triad Ser•His hydrogen bond

[Ser•His], and (3) the van der Waals contacts between the catalytic serine Oγ and the sulfur atoms of a nearby disulfide bond between Cys42 and Cys58 [Ser•SS]. **(B)** Local alignment of the apo (PDB: 5MNE) and GSA-bound (PDB: 4XOJ) trypsin on atoms of the catalytic triad residues (Ser195, His57 and D102). The O•C van der Waals interaction (4) would be ~2.3 Å if the serine sidechain did not rotate when the substrate binds instead of the observed ~2.7 Å in the GSA-bound ensembles. **(C)** Left: correlation between the catalytic serine χ^1 and the nucleophile•electrophile distance ($d_{O\bullet C}$) in the trypsin GSA-bound ensemble (left). Right: correlation between their energies calculated from knowledge-based energy functions. The dotted lines indicate the upper limits of the knowledge-based energy functions (where geometries are highly unfavorable and therefore have no available crystallographic data). *r*, Pearson correlation coefficient; *p*-values indicate the significance of the correlation. **(D)** Knowledge-based energy functions for a free serine [ΔE(Ser χ^1)] built from 41,237 serines collected from the PDB (gray), for the trypsin catalytic serine in the apo state [$\Delta E(\chi^1)$: Apo Ser195] (aqua) and for that in the GSA-bound state [$\Delta E(\chi^1)$: GSA-bound Ser195] (blue). The energy functions for trypsin Ser195 were obtained by modeling the rotation of the Ser195 and calculating the sum of the individual interaction energies, as shown in the equations in the shaded boxes. The Ser195 χ^1 angles observed in the apo and GSA-bound pseudo-ensembles were mapped onto the corresponding energy functions and shown as circles (apo: aqua; GSA-bound: blue). Vertical lines indicate the χ^1 values at the local minima of the energy functions. Gray shaded areas in the plots indicate upper limits of the energy functions. **(E)** Binding interactions to the substrate backbone amide groups found in serine proteases from four structural clans. Substrate peptide positions (P and P') were annotated and shaded in light purple; enzyme positions are shaded in light gray and colored by their structural clans (PA: red; SB: pink; SE: green; SC: yellow). The bond being cut ("scissile bond") is marked by a red line. Hydrogen bonds between substrate backbone amide and enzyme groups are indicated by dotted lines. Numbers are assigned to each interaction (gray shaded circles) and referred to in 7F. Enzyme residue numbers are taken from the representative protease of each clan shown in Fig. 1C; atom names are specified if the sidechain, instead of the backbone amide of the enzyme residue, is involved in the binding interaction. Positions that are not evolutionarily conserved within each clan are indicated by one-letter coded amino acids separated by slashes. **(F)** Pearson correlation coefficients (*r*) for binding interaction lengths *versus* nucleophile•electrophile distance ($d_{O\bullet C}$) in the trypsin GSA-bound ensemble. Binding interactions that significantly correlate with $d_{O\bullet C}$ (*p-value* ≤ 0.05) were colored by their significance, with higher significance indicated by darker red colors; binding interactions that do not have a significant linear correlation with $d_{O\bullet C}$ (*p-value* > 0.05) were colored gray. The circled numbers on the left indicate the interactions annotated in 6E.

The balance of forces at the active site favors locally destabilized conformers

To further test the effectiveness of knowledge-based energy functions derived from small molecule and protein crystallographic structures, we combined the energies calculated for trypsin active site interactions using these functions and asked how well it reproduced the active site conformation observed in pseudo-ensembles. We modeled the rotation of the catalytic serine sidechain and evaluated energetic consequences from the changes in the surrounding interactions. For the apo state, these interactions and their corresponding energy terms are (1) the serine χ^1 rotamer, $\Delta E(\text{Ser } \chi^1)$, (2) the Ser•His hydrogen bond, $\Delta E_{\text{HB}}(\text{Ser•His})$ and (3) the serine•disulfide contact, $\Delta E(d_{\text{Ser•SS}})$ (Fig. 6A and fig. S49 to 50). For the GSA-bound state, an additional term contributes, (4) the van der Waals contact between the reacting atoms [$\Delta E(d_{\text{O•C}})$; Fig. 6B and fig. S49 to 50).

The apo energy function, defined by summing energy terms (1) to (3), predicts an energy minimum in the reactive rotamer conformation *gauche*–, as is observed in the apo pseudo-ensembles (Fig. 6D). With the substrate bound, the O•C van der Waals term shifts the energy minimum by 30˚, consistent with the rotation to destabilized serine rotamers in the GSA-bound state (Fig. 6D; predicted: 265˚; observed median: 272˚). The smaller observed rotation compared to the prediction may result from small rearrangements in the active site surroundings not accounted for in our model. Overall, our results indicate that interactions in and around the active site provide constraints to generate local destabilization, and that the sum of their energetics estimated by knowledge-based energy functions reproduces the preferred states observed in pseudo-ensembles. This agreement also provides support for the use of knowledge-based energy functions to recapitulate and explain the conformational preferences observed in experimental ensembles.

MD simulations of the apo and GSA-bound trypsin also reproduced the experimentally observed rotameric shift within the serine *gauche*– well. It also predicted 1:1 population of the reactive *gauche*– rotamer and the unreactive *gauche*+ rotamer in the apo state (fig. S51), which are observed in 1:24 ratio in GSA-bound trypsin structures (fig. S46). While this discrepancy with the experimental findings may seem to be large, it corresponds to an overall error of only ~1–2 kcal/mol. Additional quantitative comparisons of conformational changes observed in simulations *versus* in pseudo-ensembles, along with independent ensemble data from NMR and room-temperature X-ray crystallography, will help reveal the origin of these differences.

Interactions that constrain the reactive substrate amide

The knowledge-based energy analyses above assumed that the substrate is fixed and thus cannot escape the observed distance and rotameric destabilization. The ability to reproduce the observed conformational preferences with this assumption implies that the substrate is indeed highly constrained and led us to consider what interactions might be responsible. Many serine proteases have broad specificity (table S2); therefore, interactions to the substrate peptide backbone instead of specific sidechains could be critical for positioning the substrate (Fig. 6E). Supporting this notion, GSA-bound trypsin structures with preferred P1-site Arg and Lys and non-preferred sidechains, including hydrophobic and negatively charged groups, all showed destabilized nucleophile•electrophile distances and serine χ^1 angles (fig. S52); in contrast, GSA-bound structures lacking the hydrogen bond on either side of the reactive amide backbone (P1–NH or P2′–NH) have longer nucleophile•electrophile distances (fig. S53 to 54), partially relieved from the distance destabilization. Correlations are observed between the lengths of the substrate binding interactions shown in Fig. 6E and the nucleophile•electrophile distance, providing additional support for the roles of backbone interactions in positioning the substrate amide, with

the strongest correlations observed for one of the oxyanion hole (primary, 1°) and two additional amide hydrogen bonds donated to the –NH groups of the substrate P1 and P2′ sites (Fig. 6F, supplementary text S10 and figs. S55 to 57).

The nucleophilic elbow underlies oxyanion hole ground state destabilization

The nucleophilic elbow was discovered in α/β hydrolyses (*48*, *122*) and was subsequently found in serine and cysteine proteases[10] (*106*, *166*). This motif forms a pseudo-ring structure with the bound substrate carbonyl, via the serine nucleophile and its backbone amide (Fig. 7A, "N-type") or that of its neighboring residue ("N+1"-type, fig. S58). As ring structures can be subject to geometric constraints, we reasoned that the nucleophilic elbow pseudo-ring may be responsible for the ground state destabilization observed for the 1° oxyanion hole hydrogen bond (Fig. 5I), especially as the noncovalent interactions within the nucleophilic elbow have specific orientational preferences (Fig. 7B). Such an effect would be analogous to ring strain commonly found in small molecule covalent rings (*167*), but in this case with a mixture of covalent and non-covalent connections and thus likely smaller energetic consequences.

To evaluate this possibility, we searched the CSD for small molecule interactions with atomic arrangements that mimic the pseudo-ring structure of the nucleophilic elbow (Fig. 7B and fig. S58). The conformational distribution from 1294 small molecule "nucleophilic elbow" structures (N-type) revealed an overall preference for in-plane hydrogen bonds ($\phi_{HB} = 0°$), as found for unconstrained, intermolecular amide•carbonyl hydrogen bonds (Fig. 7C and fig. S59). However, the subset of structures exhibiting the reactive conformation ($\phi_{attack} \geq 70°$; $n = 278$) has instead out-of-plane, destabilized hydrogen bond orientations (Fig. 7C). Stated another way, conformers that simultaneously align reacting atoms for reaction and form a favorable hydrogen bond are low probability, high energy states (Fig. 7D). As a result, the alignment of reacting groups for nucleophilic attack in serine proteases concomitantly misaligns the oxyanion hole hydrogen bonds, resulting in ground state destabilization that is used in catalysis. Similar constraints were also found for N+1-type nucleophilic elbows (fig. S60).

The other serine protease oxyanion hole hydrogen bond (2°) is also out-of-plane and destabilized (fig. S28 and 30). Small molecule data indicates that when the 1° oxyanion hydrogen bond is out-of-plane, an out-of-plane 2° hydrogen bond is also favored (fig. S61), perhaps due to steric hindrance between the hydrogen bond donors. Analogous analyses for the acylenzyme state also revealed geometric constraints and oxyanion hole destabilization; thus, the nucleophilic elbow may also facilitate the deacylation step (figs. S62).

Our observations suggest that the nucleophilic elbow provides a simple and efficient catalytic strategy—aligning the catalytic serine and the substrate for reaction simultaneously destabilizes the oxyanion hole hydrogen bonds, thereby augmenting catalysis.

[10] "Nucleophilic elbow" was first discovered as a conserved sheet-turn-helix motif in α/β hydrolyses where the nucleophilic residue is located (*48*, *165*). This motif allows the backbone amide adjacent to the nucleophilic residue to be used as an oxyanion hole hydrogen bond donor (fig. S58). Similar atomic arrangements were later found across many serine and cysteine proteases (*106*, *166*). Here, we use the term "nucleophilic elbow" generically to refer to this pseudo-ring atomic arrangement connecting the nucleophile and the oxyanion hole, extending the term "nucleophilic elbow" beyond its original usage for α/β hydrolyses.

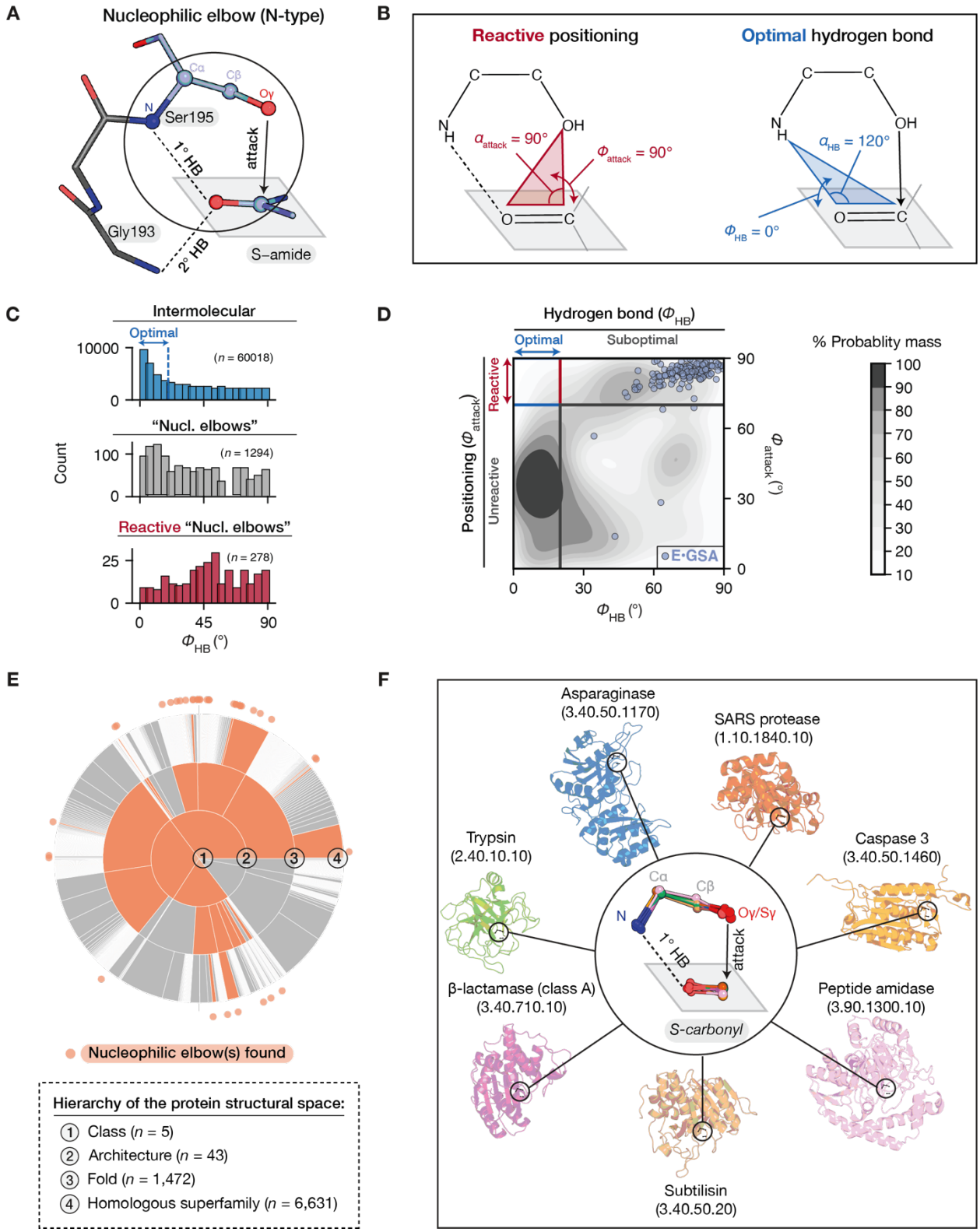

Fig. 7. The nucleophilic elbow underlies oxyanion hole hydrogen bond destabilization and has repeatedly evolved across many protein structures. (A) The nucleophilic elbow (N-type) in GSA-bound trypsin (PDB: 3M7Q). See fig. S58 for the N+1-type. **(B)** Reactive configuration of the reacting groups (red; α_{attack} and ϕ_{attack} defined in Fig. 2A) and optimal hydrogen bond geometry (blue; α_{HB}, defined as the angle between the hydrogen bond donor N, and the O and C

atoms on the acceptor carbonyl group, and ϕ_{attack}, defined in Fig. 5H). **(C)** Distributions of ϕ_{HB} in intermolecular amide•carbonyl hydrogen bonds (blue), small molecule N-type nucleophilic-elbow-like structures (gray), and its subset where the reactant orientations are reactive (defined as $\phi_{attack} \geq 70°$, red). The blue arrow indicates the region of $\phi_{HB} \leq 20°$ and is assigned as "optimal". **(D)** Knowledge-based conformational map for the hydrogen bond orientations (ϕ_{HB}) and nucleophile•amide attack orientation (ϕ_{attack}) in small molecule N-type nucleophilic-elbow-like motifs (n = 1294). The intensity of gray indicates the cumulative probability mass (%) at each contour level. Blue circles indicate the observed ϕ_{HB} and ϕ_{attack} in GSA-bound serine proteases with N-type nucleophilic elbows. **(E)** Positions of protein structures where nucleophilic elbows are found among the hierarchy of structural classification of the known protein structural space. The structural hierarchy was taken from the CATH database (*168*) and was constructed from 536,613 structural domains in the PDB. Levels of the hierarchy and the numbers of different types of structures at each level are annotated in the dashed-lined box. The area of each slice at each level is proportional to the number of its members. The protein structural space at class, architecture and fold levels is colored by whether the nucleophilic elbow is present (orange) *verses* absent (gray). At the homologous superfamily level, only superfamilies with nucleophilic elbow-bearing enzymes are shown as orange circles. **(F)** Alignment of the nucleophilic elbow in its GSA-bound state from 15 different enzymes (inner circle). These enzymes come from 7 different structural folds and the overall structures of one representative member from each fold are shown on the outside. The enzymes shown have N-type nucleophilic elbows and their serine rotamers are in the *gauche–* state. (See fig. S65 for additional alignments for the N-type nucleophilic elbows in the GSA-bound, TSA-bound and acylenzyme states and in different serine rotameric states, as well as the same alignments performed for N+1-type nucleophilic elbows.)

Repeated evolution of the nucleophilic elbow

Catalysis from the nucleophilic elbow derives from local bonds and interactions. We therefore expected that this motif could be used in additional enzymes with different global structures and may have evolved multiple times. In particular, the analysis above predicts that the nucleophilic elbow can provide catalytic advantages for enzymes that perform nucleophilic attack and stabilize an oxyanionic covalent intermediate via hydrogen bonds. We searched the Mechanism and Catalytic Site Atlas (M-CSA) database (*6*) for enzymatic reactions matching these criteria and found 92 nucleophilic residues from 91 unique enzymes that carry out three types of reactions: nucleophilic addition on carbonyl compounds (n = 86), nucleophilic aromatic substitution (n = 1), and S_N2 reaction on phosphoryl compounds (n = 5). Nucleophilic elbows were found in 69 of the 91 enzymes, including 50 that perform non-protease reactions (table S18 and fig. S63). Combined with the serine and cysteine proteases described herein (n = 16; table S2) and in prior work by Buller and Townsend (n = 18; table S19), the nucleophilic elbow was found in 102 different enzymes (fig. S63).

We examined the distribution of the nucleophilic elbow-bearing enzymes across the known protein structural space using the CATH database (*168*). This set of 102 enzymes spans 32 structural folds (2% of all known structural folds) spread across distinct branches in the structural space (Fig. 7E and fig. S64). Therefore, it appears that the nucleophilic elbow has repeatedly evolved from multiple structural origins, presumably a result of the simplicity and the local nature of this motif, combined with its catalytic prowess (see below).

Nucleophilic elbow-bearing enzymes from multiple folds exploit common catalytic features

The wide presence of the nucleophilic elbow suggests an ability to maintain its catalytic function for multiple reactions when it is inserted in different structural scaffolds. To test whether the nucleophilic elbows can provide catalysis in different enzymes and whether they follow the same catalytic strategies, we turned to the enzymes that perform nucleophilic addition reactions on carbonyl compounds, as they share similar mechanistic constraints. Within the set of 102 nucleophilic elbow-bearing enzymes, there are 66 enzymes (from 22 structural folds) performing this type of reaction. They perform nucleophile attack on an array of carbonyl compounds beyond peptide amides, including lactams (n = 3), amino acid sidechains (Gln or Asn; n = 8), esters (n = 16), thioesters (n = 9), carboxylate (n = 1), and aldehydes (n = 5) (table S22). The simplest prediction is that nucleophilic elbows provide the same catalytic advantages across these enzymes and reactions, which we assessed below.

As the catalytic function of the nucleophilic elbow originates from its conformational properties, we assessed the nucleophilic elbow geometries for enzymes performing nucleophilic addition on carbonyls where GSA-bound and acylenzyme structures are available (34 enzymes from 13 structural folds, including 10 non-protease enzymes; table S21). Despite the diversity in overall structures, the nucleophilic elbow atoms of these enzymes are well aligned (Fig. 7F and fig. S65), with root-mean-square deviations from 0.16 to 0.23 Å (fig. S66 and table S21). In their GSA-bound states, the nucleophilic elbows from most of the non-protease enzymes achieve the same reactive configuration as observed for serine proteases, with ϕ_{attack} = 84° (*s.d.* = 21°) and α_{attack} = 88° (*s.d.* = 11°) and exhibit short, destabilized distances (2.77 Å, *s.d.* = 0.31 Å) (fig. S67 and table S23). In addition, their oxyanion hole hydrogen bonds are out-of-plane from the substrate carbonyl groups and destabilized, with ϕ_{HB} = 84 ° (*s.d.* = 20°) and 83° (*s.d.* = 29°) (fig. S68 and table S23), as they are in the serine proteases. Thus, reactive positioning and ground state destabilization mechanisms appear to be common to nucleophilic elbow enzymes.

Conversely, we predicted that nucleophilic elbows would not be present in enzymes where structural or mechanistic constraints prevent the formation and catalytic use of its pseudo-

ring structure, in particular in the following scenarios: (1) use of a nucleophilic residue with more than one rotatable sidechains (e.g. His, Glu); (2) formation of an oxyanion that is multiple covalent bonds from the electrophile instead of adjacent, as occurs in nucleophilic aromatic substitution; (3) use of the backbone nitrogen of the nucleophilic residue as a general base instead of a hydrogen bond donor, an alternative strategy used by N-terminal proteases where the nitrogen is an amine instead of an amide; or (4) the nucleophile attack occurs internally at a neighboring residue such that its amide group cannot be used for hydrogen bonding. Eighteen of the 23 enzymes that do not use a nucleophilic elbow follow these criteria (table S24), whereas no nucleophilic elbow-bearing enzymes have these mechanistic features. The remaining five enzymes have limited structural and mechanistic data so we cannot evaluate potential mechanistic reasons for their lacking a nucleophilic elbow. Two of these enzymes use stronger hydrogen bond donors (Lys, Arg), suggesting possible alternative strategies to achieve catalysis from an oxyanion hole. Finally, and intriguingly, nucleophilic elbows were also found in two phosphoryl transfer enzymes in our search, suggesting broader catalytic capabilities of this motif (fig. S69).

Conclusions and implications

Our ensemble-function investigation has identified catalytic features that were previously unknown and established a comprehensive, quantitative model for serine proteases, a model grounded in basic concepts of physics and chemistry—bond rotamers, van der Waals interactions, hydrogen bonds, and conformational entropy. We show that the identified catalytic features are recurrent in Nature, having commonly evolved in serine proteases from multiple distinct structural clans and in many enzymes catalyzing analogous reactions.

One of the most vexing questions in understanding enzymes has been the dichotomy between the potential catalysis from positioning and the motions that are needed for reactions to occur. Fig. 8 illustrates this paradox and how serine proteases have arrived at an elegant solution. An overly positioned enzyme (Enzyme 1) would hinder rather than facilitate reaction because the ground state is restricted to a deep energy minimum, rendering excursions to the transition state less likely than in solution. Uniformly increasing flexibility around this state would allow the reaction to proceed but would also increase the probability of occupying off-pathway conformations (Fig. 8, Enzyme 2). Serine proteases overcome this inefficiency, at least in part, by restricting motions to a bond rotation that aligns with the reaction path (Fig. 8, Enzyme 3). This rotation replaces the less efficient translational motion that dominates the solution reaction path, providing a low-entropy pathway for the reaction to proceed.

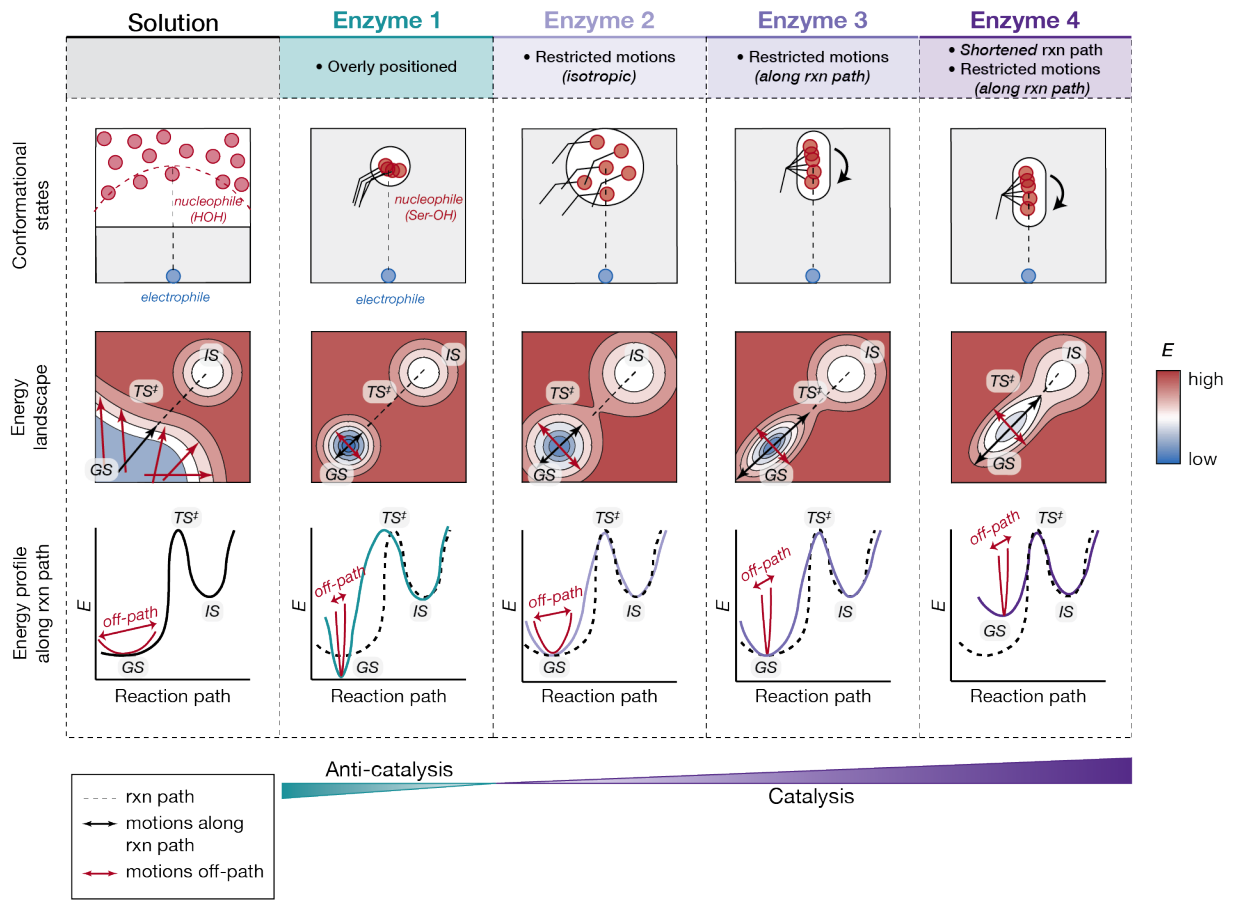

Fig. 8. Synergy of positioning and motion in serine protease catalysis. A solution reaction and four different schematic enzyme examples are shown for comparison. The top row shows the positioning between the nucleophile (red) and the electrophile (blue) for each reaction. The second row shows a two-dimensional energy landscape for the corresponding reaction. Dashed lines indicate the shortest reaction path; black arrows indicate motions along this reaction path whereas red arrows show a subset of possible off-pathway motions. The contour levels are colored by their energies, with blue to red corresponding to energies from low to high. The bottom row shows the one-dimensional "energy profile" of each reaction, corresponding to the slice of the energy landscape along the shortest reaction path. The energy wells and arrows colored in red represent a different "slice" of the energy landscape that is not along the reaction path ("off-path"). GS, ground state; TS‡, transition state; IS, intermediate state.

This synergy of positioning and motion is achieved by ground state destabilization. A partially eclipsed serine rotamer and a shorter-than-ideal van der Waals distance in the ground state allow the enzymatic reaction to commence from a higher energy state that is further along the reaction coordinate than the ground state in solution (Fig. 8, Enzyme 4). Therefore, the enzyme creates a downhill energetic gradient that selectively favors the serine rotation and bond formation and partially offsets the energetic barrier to reaction. By positioning reactants in reactive, destabilized conformers in the ground state and relieving them in the transition state, serine proteases promote the motions needed for reaction and reduce the reaction barrier both entropically and enthalpically[11].

Textbook and literature descriptions of enzyme catalysis often emphasize the convergent evolution of the Ser•His•Asp catalytic triad across serine proteases (*11*, *12*, *169–171*). Our analyses revealed an even broader generalization and its underlying mechanistic and evolutionary significance. Nature has repeatedly evolved the nucleophilic elbow motif, a motif that connects the nucleophile and the oxyanion hole and restrains their conformations in a catalytically adventitious manner. Additional groups are recruited to provide general base catalysis, the remainder of the oxyanion hole, and substrate binding and positioning; these groups differ in different enzymes and folds, and only a subset contains the classic Ser•His•Asp catalytic triad (*172*, *173*). Understanding how nature has combined and elaborated upon catalytic motifs to construct active sites that enable enormous rate enhancements may lead to insights for enzyme design.

The ensemble–function approach developed here differs fundamentally from traditional site-directed mutagenesis studies that assign "catalytic contributions" based on the rate decrease upon mutation. Knowing that catalysis is greatly diminished when a group is replaced by another is different from *knowing how it contributes when present*. In particular, the challenges in interpreting mutagenesis data and relating them to catalytic mechanisms arise from the high degree of interconnectedness between catalytic features, which is expected to yield complex, non-additive mutational effects (*174*). For example, in serine proteases, mutation of the catalytic triad histidine eliminates general base catalysis but may also lessen positioning and destabilization of the nucleophilic serine. Mutation of the catalytic triad aspartate yields large deleterious effects [10^3–10^5-fold (*17*, *19*, *175*)][12], consistent with roles beyond the direct catalytic contribution from hydrogen bond strengthening (4-fold). These roles presumably include facilitating general base catalysis via tuning the histidine pK_a (*176*), positioning the histidine for proton transfer, and aligning the nucleophilic serine (via the histidine) and the substrate for reaction.[13] Further, the magnitude of rate decrease varies for different mutations,

---

[11] We note that the entropic and enthalpic advantages from positioning and ground state destabilization described here are different from the *observed* reduction in activation enthalpy or entropy, as additional enzyme interactions and solvent interactions may lead to complex enthalpy-entropy compensation. In the classic account of the *Circe effect*, Jencks described this phenomenon as follows: "…This entropic advantage will appear in the observed free energy of activation even if the difference in entropy is masked by compensating entropy and enthalpy changes caused by solvation effects…Because there is no way to estimate the magnitude of these compensating changes *a priori*, the observed enthalpy and entropy of a particular reaction do not provide a direct measure of either the change in free energy resulting from the solvent effect or the intrinsic change in the entropy of the reactants." (*26*) Our ensemble analyses provide a means to estimate this positioning and its entropic advantage.

[12] 10^3–fold for Asp to Cys, 10^4–fold for Asp to Ser or Asn and 10^5–fold for Asp to Ala (*17*, *19*, *175*).

[13] The catalytic triad Asp may play a role in positioning the nucleophilic serine and enforcing its ground state destabilization, as the serine appears to escape the reactive, destabilized conformer when the hydrogen bond between the catalytic Asp and His is lost (fig. S46). In clan PA serine proteases, Asp102 also forms a hydrogen bond with Ser214, which interacts with the substrate P1' amide. Since the hydrogen bond between Ser214 and P1'–NH

suggesting that additional perturbations on the active site are introduced by the more deleterious mutations (*177*). Instead of a comparison between wild-type and mutant enzymes, our ensemble–function analyses compare the enzymatic and the uncatalyzed solution reaction, and thus can identify individual catalytic features and quantify their contributions to catalysis.

The ensemble–function approach provides a link between traditional experimental approaches—such as site-directed mutagenesis—and computational approaches and will greatly expand the understanding that can be achieved. Comparisons of the wild-type and mutant enzyme ensembles will allow deep investigations for the origins of mutational effects and provide quantitative predictions that can be tested via site-directed mutagenesis. Computational studies have provided models that reproduce the observed overall rate enhancements [e.g. (*147*, *178*)] and known mutational effects [e.g. (*179–181*)]. Applying the ensemble–function approach to analyze computational ensemble data will allow researchers to identify and quantify specific catalytic mechanisms and harness the full, combined potential of these approaches. The iterative process of model generation via ensemble–function analyses and subsequent experimental tests will lead to more comprehensive and reliable catalytic models.

Most generally, conformational ensembles provide the ability to connect molecules to free energy, a connection that is needed to define enzyme catalysis and all aspects of protein function. The ensemble–function approach will allow enzymology to transition from qualitative descriptions to *quantitative* models and may be of value in understanding still more complex processes, including allostery (*182*) and the coordination of molecular machines (*183*). Its ability to generate models based on simple physiochemical factors promises to transform biochemistry education, allowing students and researchers to uncover and appreciate the fundamental concepts of physics and chemistry that underlie the behavior of complex biomolecules.

showed a stronger correlation with the nucleophile•electrophile distance (Fig. 6F), it is possible that the mutation of Asp102 also disrupts substrate positioning and results in additional rate decrease.

References and Notes

1. C. Walsh, *Enzymatic Reaction Mechanisms* (W. H. Freeman, 1979).

2. W. P. Jencks, *Catalysis in Chemistry and Enzymology* (Courier Corporation, 1987).

3. R. B. Silverman, *Organic Chemistry of Enzyme-Catalyzed Reactions, Revised Edition* (Academic Press, 2002).

4. J. McMurry, T. P. Begley, T. P. Begley, *The Organic Chemistry of Biological Pathways* (Roberts and Company Publishers, 2005).

5. J. D. Fischer, G. L. Holliday, J. M. Thornton, The CoFactor database: organic cofactors in enzyme catalysis. *Bioinformatics*. **26**, 2496–2497 (2010).

6. A. J. M. Ribeiro, G. L. Holliday, N. Furnham, J. D. Tyzack, K. Ferris, J. M. Thornton, Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* **46**, D618–D623 (2018).

7. W. S. Mak, J. B. Siegel, Computational enzyme design: transitioning from catalytic proteins to enzymes. *Curr. Opin. Struct. Biol.* **27**, 87–94 (2014).

8. S. L. Lovelock, R. Crawshaw, S. Basler, C. Levy, D. Baker, D. Hilvert, A. P. Green, The road to fully programmable protein catalysis. *Nature*. **606**, 49–58 (2022).

9. R. L. Miesfeld, M. M. McEvoy, *Biochemistry (Second International Student Edition)* (W. W. Norton, 2021).

10. N. D. L., A. L. Lehninger, D. L. Nelson, M. M. Cox, Institute for Molecular Genetics David L Nelson, University Michael M Cox, *Lehninger Principles of Biochemistry* (Macmillan, 2005).

11. D. Voet, J. G. Voet, C. W. Pratt, *Fundamentals of Biochemistry: Life at the Molecular Level* (John Wiley & Sons, 2016).

12. J. L. Tymoczko, J. M. Berg, University Lubert Stryer, *Biochemistry* (W. H. Freeman, ed. 2, 2012).

13. D. M. Blow, Structure and mechanism of chymotrypsin. *Acc. Chem. Res.* **9**, 145–152 (1976).

14. B. W. Matthews, P. B. Sigler, R. Henderson, D. M. Blow, Three-dimensional structure of tosyl-alpha-chymotrypsin. *Nature*. **214**, 652–656 (1967).

15. P. B. Sigler, D. M. Blow, B. W. Matthews, R. Henderson, Structure of crystalline α-chymotrypsin: II. A preliminary report including a hypothesis for the activation mechanism. *J. Mol. Biol.* **35**, 143–164 (1968).

16. J. D. Robertus, J. Kraut, R. A. Alden, J. J. Birktoft, Subtilisin; a stereochemical mechanism involving transition-state stabilization. *Biochemistry*. **11**, 4293–4303 (1972).

17. P. Carter, J. A. Wells, Dissecting the catalytic triad of a serine protease. *Nature*. **332**, 564–568 (1988).

18. P. Bryan, M. W. Pantoliano, S. G. Quill, H. Y. Hsiao, T. Poulos, Site-directed mutagenesis and the role of the oxyanion hole in subtilisin. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 3743–3745 (1986).

19. D. R. Corey, C. S. Craik, An investigation into the minimum requirements for peptide hydrolysis by mutation of the catalytic triad of trypsin. *J. Am. Chem. Soc.* **114**, 1784–1790 (1992).

20. I. M. Klotz, J. S. Franzen, Hydrogen bonds between model peptide groups in solution. *J. Am. Chem. Soc.* **84**, 3461–3466 (1962).

21. R. M. Smith, D. E. Hansen, The pH-rate profile for the hydrolysis of a peptide bond. *J. Am. Chem. Soc.* **120**, 8910–8913 (1998).

22. R. S. Brown, A. J. Bennet, H. Slebocka-Tilk, Recent perspectives concerning the mechanism of H3O+- and hydroxide-promoted amide hydrolysis. *Acc. Chem. Res.* **25**, 481–488 (1992).

23. L. Boltzmann, *Vorlesungen über Gastheorie* (J.A. Barth (A. Meiner), 1910).

24. J. B. S. Haldane, *Enzymes* (M.I.T. Press, 1965).

25. L. Pauling, Molecular Architecture and Biological Reactions. *Chem. Eng. News*. **24**, 1375–1377 (1946).

26. W. P. Jencks, Binding Energy, Specificity, and Enzymic Catalysis: The Circe Effect. *Advances in Enzymology - and Related Areas of Molecular Biology* (2006), pp. 219–410.

27. B. L. Vallee, R. J. Williams, Metalloenzymes: the entatic nature of their active sites. *Proc. Natl. Acad. Sci. U. S. A.* **59**, 498–505 (1968).

28. M. I. Page, W. P. Jencks, Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. *Proc. Natl. Acad. Sci. U. S. A.* **68**, 1678–1683 (1971).

29. A. Warshel, M. Levitt, Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**, 227–249 (1976).

30. A. Warshel, Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *J. Biol. Chem.* **273**, 27035–27038 (1998).

31. A. Warshel, P. K. Sharma, M. Kato, Y. Xiang, H. Liu, M. H. M. Olsson, Electrostatic basis for enzyme catalysis. *Chem. Rev.* **106**, 3210–3235 (2006).

32. S. Hur, T. C. Bruice, The mechanism of catalysis of the chorismate to prephenate reaction by the Escherichia coli mutase enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 1176–1181 (2002).

33. S. Hur, T. C. Bruice, Enzymes do what is expected (chalcone isomerase versus chorismate mutase). *J. Am. Chem. Soc.* **125**, 1472–1473 (2003).

34. S. Hur, T. C. Bruice, Comparison of formation of reactive conformers (NACs) for the Claisen rearrangement of chorismate to prephenate in water and in the E. coli mutase: the efficiency of the enzyme catalysis. *J. Am. Chem. Soc.* **125**, 5964–5972 (2003).

35. D. R. Storm, D. E. Koshland, A Source for the Special Catalytic Power of Enzymes: Orbital Steering*. *Proceedings of the National Academy of Sciences*. **66**, 445–452 (1970).

36. A. Dafforn, D. E. Koshland Jr, Proximity, entropy and orbital steering. *Biochem. Biophys. Res. Commun.* **52**, 779–785 (1973).

37. A. D. Mesecar, B. L. Stoddard, D. E. Koshland Jr, Orbital steering in the catalytic power of enzymes: small structural changes with large catalytic consequences. *Science*. **277**, 202–206 (1997).

38. F. M. Menger, U. V. Venkataram, Proximity as a component of organic reactivity. *J. Am. Chem. Soc.* **107**, 4706–4709 (1985).

39. F. M. Menger, On the source of intramolecular and enzymatic reactivity. *Acc. Chem. Res.* **18**, 128–134 (1985).

40. F. M. Menger, F. Nome, Interaction vs Preorganization in Enzyme Catalysis. A Dispute That Calls for Resolution. *ACS Chem. Biol.* **14**, 1386–1392 (2019).

41. A. S. Mildvan, Mechanism of enzyme action. *Annu. Rev. Biochem.* **43**, 357–399 (1974).

42. F. Yabukarski, J. T. Biel, M. M. Pinney, T. Doukov, A. S. Powers, J. S. Fraser, D. Herschlag, Assessment of enzyme active site positioning and tests of catalytic mechanisms through X-ray-derived conformational ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33204–33215 (2020).

43. F. Yabukarski, T. Doukov, M. M. Pinney, J. T. Biel, J. S. Fraser, D. Herschlag, Ensemble-function relationships to dissect mechanisms of enzyme catalysis. *Sci Adv*. **8**, eabn7738 (2022).

44. D. M. van Aalten, D. A. Conn, B. L. de Groot, H. J. Berendsen, J. B. Findlay, A. Amadei, Protein dynamics derived from clusters of crystal structures. *Biophys. J.* **73**, 2891–2896 (1997).

45. V. Zoete, O. Michielin, M. Karplus, Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J. Mol. Biol.* **315**, 21–52 (2002).

46. R. B. Best, K. Lindorff-Larsen, M. A. DePristo, M. Vendruscolo, Relation between native ensembles and experimental structures of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 10901–10906 (2006).

47. P. V. Burra, Y. Zhang, A. Godzik, B. Stec, Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 10505–10510 (2009).

48. D. L. Ollis, E. Cheah, M. Cygler, B. Dijkstra, F. Frolow, S. M. Franken, M. Harel, S. J. Remington, I. Silman, J. Schrag, The alpha/beta hydrolase fold. *Protein Eng.* **5**, 197–211 (1992).

49. Y. Shen, A. Bax, Synergism between x-ray crystallography and NMR residual dipolar couplings in characterizing protein dynamics. *Struct Dyn*. **10**, 040901 (2023).

50. Y. Shen, A. J. Robertson, A. Bax, Validation of X-ray Crystal Structure Ensemble Representations of SARS-CoV-2 Main Protease by Solution NMR Residual Dipolar Couplings. *J. Mol. Biol.* **435**, 168067 (2023).

51. B. Imperiali, R. H. Abeles, Inhibition of serine proteases by peptidyl fluoromethyl ketones. *Biochemistry*. **25**, 3760–3767 (1986).

52. D. A. Matthews, R. A. Alden, J. J. Birktoft, S. T. Freer, J. Kraut, X-ray crystallographic study of boronic acid adducts with subtilisin BPN' (Novo). A model for the catalytic transition state. *J. Biol. Chem.* **250**, 7120–7126 (1975).

53. C. A. Kettner, R. Bone, D. A. Agard, W. W. Bachovchin, Kinetic properties of the binding of alpha-lytic protease to peptide boronic acids. *Biochemistry*. **27**, 7682–7688 (1988).

54. R. Bone, A. B. Shenvi, C. A. Kettner, D. A. Agard, Serine protease mechanism: structure of an inhibitory complex of alpha-lytic protease and a tightly bound peptide boronic acid. *Biochemistry*. **26**, 7609–7614 (1987).

55. G. D. Brayer, L. T. Delbaere, M. N. James, C. A. Bauer, R. C. Thompson, Crystallographic and kinetic investigations of the covalent complex formed by a specific tetrapeptide aldehyde and the serine protease from Streptomyces griseus. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 96–100 (1979).

56. B. Zerner, M. L. Bender, The kinetic consequences of the acyl-enzyme mechanism for the reactions of specific substrates with chymotrypsin. *J. Am. Chem. Soc.* **86**, 3669–3674 (1964).

57. J. McConn, E. Ku, A. Himoe, K. G. Brandt, G. P. Hess, Investigations of the chymotrypsin-catalyzed hydrolysis of specific substrates. V. Determination of pre-steady state kinetic parameters for specific substrate esters by stopped flow techniques. *J. Biol. Chem.* **246**, 2918–2925 (1971).

58. J. Fastrez, A. R. Fersht, Demonstration of the acyl-enzyme mechanism for the hydrolysis of peptides and anilides by chymotrypsin. *Biochemistry*. **12**, 2025–2034 (1973).

59. F. Himo, Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions. *J. Am. Chem. Soc.* **139**, 6780–6786 (2017).

60. S. F. Sousa, A. J. M. Ribeiro, R. P. P. Neves, N. F. Brás, N. M. F. S. A. Cerqueira, P. A. Fernandes, M. J. Ramos, Application of quantum mechanics/molecular mechanics methods in the study of enzymatic reaction mechanisms. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **7**, e1281 (2017).

61. S. Ahmadi, L. B. Herrera, M. Chehelamirani, J. Hostaš, S. Jalife, D. R. Salahub, Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review. *International Journal of Quantum Chemistry*. **118** (2018), p. e25558.

62. H. M. Senn, W. Thiel, QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed Engl.* **48**, 1198–1229 (2009).

63. H.-B. Bürgi, J. D. Dunitz, *Structure Correlation* (VCH, 1994).

64. J. Kroon, J. A. Kanters, J. G. C. M. van Duijneveldt-van De Rijdt, F. B. van Duijneveldt, J. A. Vliegenthart, O-H · O Hydrogen bonds in molecular crystals a statistical and quantum-chemical analysis. *J. Mol. Struct.* **24**, 109–129 (1975).

65. P. Murray-Rust, J. P. Glusker, Directional hydrogen bonding to sp2- and sp3-hybridized oxygen atoms and its relevance to ligand-macromolecule interactions. *J. Am. Chem. Soc.* **106**, 1018–1025 (1984).

66. R. Taylor, O. Kennard, Crystallographic evidence for the existence of CH.cntdot..cntdot..cntdot.O, CH.cntdot..cntdot..cntdot.N and CH.cntdot..cntdot..cntdot.Cl hydrogen bonds. *J. Am. Chem. Soc.* **104**, 5063–5070 (1982).

67. R. Taylor, O. Kennard, Hydrogen-bond geometry in organic crystals. *Acc. Chem. Res.* **17**, 320–326 (1984).

68. G. A. Jeffrey, J. Mitra, Three-center (bifurcated) hydrogen bonding in the crystal structures of amino acids. *J. Am. Chem. Soc.* **106**, 5546–5553 (1984).

69. G. A. Jeffrey, H. Maluszynska, A survey of the geometry of hydrogen bonds in the crystal structures of barbiturates, purines and pyrimidines. *J. Mol. Struct.* **147**, 127–142 (1986).

70. G. A. Jeffrey, Crystallographic studies of carbohydrates. *Acta Crystallogr. B*. **46 (Pt 2)**, 89–103 (1990).

71. A. Gavezzotti, Statistical analysis of some structural properties of solid hydrocarbons. *J. Am. Chem. Soc.* **111**, 1835–1843 (1989).

72. A. Nangia, Conformational polymorphism in organic crystals. *Acc. Chem. Res.* **41**, 595–604 (2008).

73. F. H. Allen, W. D. S. Motherwell, Applications of the Cambridge Structural Database in organic chemistry and crystal chemistry. *Acta Crystallogr. B*. **58**, 407–422 (2002).

74. F. H. Allen, S. E. Harris, R. Taylor, Comparison of conformer distributions in the crystalline state with conformational energies calculated by ab initio techniques. *J. Comput. Aided Mol. Des.* **10**, 247–254 (1996).

75. R. Taylor, Life-science applications of the Cambridge Structural Database. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 879–888 (2002).

76. Z. Rappoport, S. E. Biali, M. Kaftory, Application of the structural correlation method to ring-flip processes in benzophenones. *J. Am. Chem. Soc.* **112**, 7742–7748 (1990).

77. K. A. Brameld, B. Kuhn, D. C. Reuter, M. Stahl, Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis. *J. Chem. Inf. Model.* **48**, 1–24 (2008).

78. A. J. Cruz-Cabeza, F. H. Allen, Conformation and geometry of cyclopropane rings having π-acceptor substituents: a theoretical and database study. *Acta Crystallogr. B*. **67**, 94–102 (2011).

79. S. J. Cottrell, T. S. G. Olsson, R. Taylor, J. C. Cole, J. W. Liebeschuetz, Validating and understanding ring conformations using small molecule crystallographic data. *J. Chem. Inf. Model.* **52**, 956–962 (2012).

80. J. C. Cole, S. Wiggin, F. Stanzione, New insights and innovation from a million crystal structures in the Cambridge Structural Database. *Struct Dyn*. **6**, 054301 (2019).

81. H. K. Ganguly, G. Basu, Conformational landscape of substituted prolines. *Biophys. Rev.* **12**, 25–39 (2020).

82. M. J. Sippl, Calculation of conformational ensembles from potentials of mena force: An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883 (1990).

83. M. J. Sippl, Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided Mol. Des.* **7**, 473–501 (1993).

84. S. Tanaka, H. A. Scheraga, Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*. **9**, 945–950 (1976).

85. S. Tanaka, H. A. Scheraga, Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proc. Natl. Acad. Sci. U. S. A.* **72**, 3802–3806 (1975).

86. M.-Y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 (2006).

87. R. Samudrala, J. Moult, An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275**, 895–916 (1998).

88. Y. Xia, M. Levitt, Extracting knowledge-based energy functions from protein structures by error rate minimization: Comparison of methods using lattice model. *J. Chem. Phys.* **113**, 9318–9330 (2000).

89. E.-A. D. Amir, N. Kalisman, C. Keasar, Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. *Proteins*. **72**, 62–73 (2008).

90. S. Miyazawa, R. L. Jernigan, Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*. **18**, 534–552 (1985).

91. T. Hamelryck, M. Borg, M. Paluszewski, J. Paulsen, J. Frellsen, C. Andreetta, W. Boomsma, S. Bottaro, J. Ferkinghoff-Borg, Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS One*. **5**, e13714 (2010).

92. T. Kortemme, A. V. Morozov, D. Baker, An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein–Protein Complexes. *J. Mol. Biol.* **326**, 1239–1259 (2003).

93. K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, D. Baker, Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*. **34**, 82–95 (1999).

94. J. Skolnick, A. Kolinski, A. Ortiz, Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins: Structure, Function, and Genetics*. **38** (2000), pp. 3–16.

95. H. Zhou, J. Skolnick, GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* **101**, 2043–2052 (2011).

96. J. Park, K. Saitou, ROTAS: a rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. *BMC Bioinformatics*. **15**, 307 (2014).

97. D. Gront, A. Kolinski, A new approach to prediction of short-range conformational propensities in proteins. *Bioinformatics*. **21**, 981–987 (2005).

98. M. Borg, T. Hamelryck, J. Ferkinghoff-Borg, "On the Physical Relevance and Statistical Interpretation of Knowledge-Based Potentials" in *Bayesian Methods in Structural Bioinformatics*, T. Hamelryck, K. Mardia, J. Ferkinghoff-Borg, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012), pp. 97–124.

99. D. T. Jones, W. R. Taylor, J. M. Thornton, A new approach to protein fold recognition. *Nature*. **358**, 86–89 (1992).

100. J. U. Bowie, R. Lüthy, D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure. *Science*. **253**, 164–170 (1991).

101. G. N. Ramachandran, V. Sasisekharan, Conformation of polypeptides and proteins. *Adv. Protein Chem.* **23**, 283–438 (1968).

102. A. J. Cruz-Cabeza, J. W. Liebeschuetz, F. H. Allen, Systematic conformational bias in small-molecule crystal structures is rare and explicable. *CrystEngComm*. **14**, 6797–6811 (2012).

103. A. V. Morozov, T. Kortemme, K. Tsemekhman, D. Baker, Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 6946–6951 (2004).

104. B. K. Rai, V. Sresht, Q. Yang, R. Unwalla, M. Tu, A. M. Mathiowetz, G. A. Bakken, Comprehensive Assessment of Torsional Strain in Crystal Structures of Small Molecules and Protein-Ligand Complexes using ab Initio Calculations. *J. Chem. Inf. Model.* **59**, 4195–4208 (2019).

105. G. M. Clore, J. Kuszewski, Chi(1) rotamer populations and angles of mobile surface side chains are accurately predicted by a torsion angle database potential of mean force. *J. Am. Chem. Soc.* **124**, 2866–2867 (2002).

106. A. R. Buller, C. A. Townsend, Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E653-61 (2013).

107. K. C. Gross, P. G. Seybold, Substituent effects on the physical properties and pKa of phenol. *Int. J. Quantum Chem.* **85**, 569–579 (2001).

108. D. S. Dwyer, Electronic properties of amino acid side chains: quantum mechanics calculation of substituent effects. *BMC Chem. Biol.* **5**, 2 (2005).

109. K. C. Gross, P. G. Seybold, Z. Peralta-Inga, J. S. Murray, P. Politzer, Comparison of Quantum Chemical Parameters and Hammett Constants in Correlating pK_a Values of Substituted Anilines. *The Journal of Organic Chemistry*. **66** (2001), pp. 6919–6925.

110. K. C. Gross, P. G. Seybold, Substituent effects on the physical properties and pKa of aniline. *Int. J. Quantum Chem.* **80**, 1107–1115 (2000).

111. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, The protein data bank: A computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.* **185**, 584–591 (1978).

112. H. B. Bürgi, J. D. Dunitz, E. Shefter, Chemical reaction paths. IV. Aspects of O··· C= O interactions in crystals. *Acta Crystallogr. B*. **30**, 1517–1527 (1974).

113. G. G. Hammes, S. J. Benkovic, S. Hammes-Schiffer, Flexibility, diversity, and cooperativity: pillars of enzyme catalysis. *Biochemistry*. **50**, 10422–10430 (2011).

114. V. C. Nashine, S. Hammes-Schiffer, S. J. Benkovic, Coupled motions in enzyme catalysis. *Curr. Opin. Chem. Biol.* **14**, 644–651 (2010).

115. A. Kolmogorov, Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* **4**, 83–91 (1933).

116. N. Smirnov, Table for Estimating the Goodness of Fit of Empirical Distributions. *aoms*. **19**, 279–281 (1948).

117. E. Di Cera, M. J. Page, A. Bah, L. A. Bush-Pelc, L. C. Garvey, Thrombin allostery. *Phys. Chem. Chem. Phys.* **9**, 1291–1306 (2007).

118. L. E. Sutton, *Tables of Interatomic Distances and Configuration in Molecules and Ions: Supplement, 1956-1959* (Chemical Society, 1965).

119. P. G. Jones, A. J. Kirby, Linear relationship between bond length and reactivity. *J. Chem. Soc. Chem. Commun.*, 288–289 (1979).

120. A. J. Briggs, R. Glenn, P. G. Jones, A. J. Kirby, P. Ramaswamy, Bond length and reactivity. Stereoelectronic effects on bonding in acetals and glucosides. *J. Am. Chem. Soc.* **106**, 6200–6206 (1984).

121. F. H. Allen, A. J. Kirby, Bond length and reactivity. Variable length of the carbon-oxygen single bond. *J. Am. Chem. Soc.* **106**, 6197–6200 (1984).

122. M. Holmquist, Alpha beta-hydrolase fold enzymes structures, functions and mechanisms. *Curr. Protein Pept. Sci.* **1**, 209–235 (2000).

123. R. L. Schowen, G. W. Zuorick, Amide Hydrolysis. Superimposed General Base Catalysis in the Cleavage of Anilides1. *J. Am. Chem. Soc.* **88**, 1223–1225 (1966).

124. J. J. Morris, M. I. Page, Structure–reactivity relationships and the mechanism of general base catalysis in the hydrolysis of a hydroxy-amide. Concerted breakdown of a tetrahedral intermediate. *J. Chem. Soc., Perkin Trans. 2* (1980), pp. 685–692.

125. A. J. Kirby, G. J. Lloyd, Structure and efficiency in intramolecular and enzymic catalysis: intramolecular general base catalysis. Hydrolysis of monoaryl malonates. *J. Chem. Soc., Perkin Trans. 2*, 1753 (1976).

126. A. J. Kirby, "Effective Molarities for Intramolecular Reactions" in *Advances in Physical Organic Chemistry*, V. Gold, D. Bethell, Eds. (Academic Press, 1980), vol. 17, pp. 183–278.

127. A. J. Kirby, Enzyme mechanisms, models, and mimics. *Angew. Chem. Int. Ed Engl.* **35**, 706–724 (1996).

128. A. J. Kirby, A. J. Kirby, D. H. Williams, Molecular recognition of transition states. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*. **345**, 67–76 (1997).

129. L. Hedstrom, L. Szilagyi, W. J. Rutter, Converting trypsin to chymotrypsin: the role of surface loops. *Science*. **255**, 1249–1253 (1992).

130. D. M. Epstein, R. H. Abeles, Role of serine 214 and tyrosine 171, components of the S2 subsite of alpha-lytic protease, in catalysis. *Biochemistry*. **31**, 11216–11223 (1992).

131. J. Bieth, B. Spiess, C. G. Wermuth, The synthesis and analytical use of a highly sensitive and convenient substrate of elastase. *Biochem. Med.* **11**, 350–357 (1974).

132. N. Trbovic, J.-H. Cho, R. Abel, R. A. Friesner, M. Rance, A. G. Palmer 3rd, Protein side-chain dynamics and residual conformational entropy. *J. Am. Chem. Soc.* **131**, 615–622 (2009).

133. A. A. Polyansky, R. Zubac, B. Zagrovic, "Estimation of Conformational Entropy in Protein–Ligand Interactions: A Computational Perspective" in *Computational Drug Discovery and Design*, R. Baron, Ed. (Springer New York, New York, NY, 2012), pp. 327–353.

134. S. Genheden, M. Akke, U. Ryde, Conformational Entropies and Order Parameters: Convergence, Reproducibility, and Transferability. *J. Chem. Theory Comput.* **10**, 432–438 (2014).

135. O. Edholm, H. J. C. Berendsen, Entropy estimation from simulations of non-diffusive systems. *Mol. Phys.* **51**, 1011–1028 (1984).

136. M. Karplus, J. N. Kushick, Method for estimating the configurational entropy of macromolecules. *Macromolecules*. **14**, 325–332 (1981).

137. J. Brady, M. Karplus, Configuration entropy of the alanine dipeptide in vacuum and in solution: a molecular dynamics study. *J. Am. Chem. Soc.* **107**, 6103–6105 (1985).

138. R. M. Levy, M. Karplus, J. Kushick, D. Perahia, Evaluation of the configurational entropy for proteins: application to molecular dynamics simulations of an α-helix. *Macromolecules*. **17**, 1370–1374 (1984).

139. M. Karplus, T. Ichiye, B. M. Pettitt, Configurational entropy of native proteins. *Biophys. J.* **52**, 1083–1085 (1987).

140. O. Caldararu, R. Kumar, E. Oksanen, D. T. Logan, U. Ryde, Are crystallographic B-factors suitable for calculating protein conformational entropy? *Phys. Chem. Chem. Phys.* **21**, 18149–18160 (2019).

141. A. J. Doig, M. J. Sternberg, Side-chain conformational entropy in protein folding. *Protein Sci.* **4**, 2247–2251 (1995).

142. J. Novotny, R. E. Bruccoleri, F. A. Saul, On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. *Biochemistry*. **28**, 4735–4749 (1989).

143. S. D. Pickett, M. J. Sternberg, Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol.* **231**, 825–839 (1993).

144. M. V. Shapovalov, R. L. Dunbrack Jr, A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*. **19**, 844–858 (2011).

145. T. Rungrotmongkol, P. Decha, P. Sompornpisut, M. Malaisree, P. Intharathep, N. Nunthaboot, T. Udommaneethanakit, O. Aruksakunwong, S. Hannongbua, Combined QM/MM

mechanistic study of the acylation process in furin complexed with the H5N1 avian influenza virus hemagglutinin's cleavage site. *Proteins*. **76**, 62–71 (2009).

146. T. Ishida, S. Kato, Role of Asp102 in the catalytic relay system of serine proteases: a theoretical study. *J. Am. Chem. Soc.* **126**, 7111–7118 (2004).

147. T. Ishida, S. Kato, Theoretical perspectives on the reaction mechanism of serine proteases: the reaction free energy profiles of the acylation process. *J. Am. Chem. Soc.* **125**, 12035–12048 (2003).

148. J. D. Kemp, K. S. Pitzer, The Entropy of Ethane and the Third Law of Thermodynamics. Hindered Rotation of Methyl Groups. *J. Am. Chem. Soc.* **59**, 276–279 (1937).

149. R. M. Pitzer, The barrier to internal rotation in ethane. *Acc. Chem. Res.* **16**, 207–210 (1983).

150. L. Simón, J. M. Goodman, Enzyme catalysis by hydrogen bonds: the balance between transition state binding and substrate binding in oxyanion holes. *J. Org. Chem.* **75**, 1831–1840 (2010).

151. L. Simón, J. M. Goodman, Hydrogen-bond stabilization in oxyanion holes: grand jeté to three dimensions. *Org. Biomol. Chem.* **10**, 1905–1913 (2012).

152. M. L. Bender, G. A. Hamilton, Kinetic isotope effects of deuterium oxide on several α-chymotrypsin-catalyzed reactions. *J. Am. Chem. Soc.* **84**, 2570–2576 (1962).

153. J. Hine, Structural effects on rates and equilibriums. XV. Hydrogen-bonded intermediates and stepwise mechanisms for proton-exchange reactions between oxygen atoms in hydroxylic solvents. *J. Am. Chem. Soc.* **94**, 5766–5771 (1972).

154. N. Stahl, W. P. Jencks, Hydrogen bonding between solutes in aqueous solution. *J. Am. Chem. Soc.* **108**, 4196–4205 (1986).

155. S.-O. Shan, D. Herschlag, "[11] Hydrogen bonding in enzymatic catalysis: Analysis of energetic contributions" in *Methods in Enzymology* (Academic Press, 1999), vol. 308, pp. 246–276.

156. A. Warshel, Energetics of enzyme catalysis. *Proc. Natl. Acad. Sci. U. S. A.* **75**, 5250–5254 (1978).

157. S. D. Fried, S. Bagchi, S. G. Boxer, Extreme electric fields power catalysis in the active site of ketosteroid isomerase. *Science*. **346**, 1510–1514 (2014).

158. S. D. Fried, S. G. Boxer, Electric Fields and Enzyme Catalysis. *Annu. Rev. Biochem.* **86**, 387–415 (2017).

159. J. P. Klinman, Dynamically achieved active site precision in enzyme catalysis. *Acc. Chem. Res.* **48**, 449–456 (2015).

160. Y. Cha, C. J. Murray, J. P. Klinman, Hydrogen tunneling in enzyme reactions. *Science*. **243**, 1325–1330 (1989).

161. M. J. Knapp, K. Rickert, J. P. Klinman, Temperature-dependent isotope effects in soybean lipoxygenase-1: correlating hydrogen tunneling with protein dynamics. *J. Am. Chem. Soc.* **124**, 3865–3874 (2002).

162. S. Hammes-Schiffer, Hydrogen Tunneling and Protein Motion in Enzyme Reactions. *ChemInform*. **37** (2006), , doi:10.1002/chin.200618274.

163. S. D. Schwartz, V. L. Schramm, Enzymatic transition states and dynamic motion in barrier crossing. *Nat. Chem. Biol.* **5**, 551–558 (2009).

164. T. T. Baird Jr, W. D. Wright, C. S. Craik, Conversion of trypsin to a functional threonine protease. *Protein Sci.* **15**, 1229–1238 (2006).

165. M. Nardini, B. W. Dijkstra, α/β Hydrolase fold enzymes: the family keeps growing. *Curr. Opin. Struct. Biol.* **9**, 732–737 (1999).

166. H. J. Korza, M. Bochtler, Pseudomonas aeruginosa LD-carboxypeptidase, a serine peptidase with a Ser-His-Glu triad and a nucleophilic elbow. *J. Biol. Chem.* **280**, 40802–40812 (2005).

167. L. N. Ferguson, Ring strain and reactivity of alicycles. *J. Chem. Educ.* **47**, 46 (1970).

168. M. Knudsen, C. Wiuf, The CATH database. *Hum. Genomics*. **4**, 207 (2010).

169. P. F. Gherardini, M. N. Wass, M. Helmer-Citterich, M. J. E. Sternberg, Convergent evolution of enzyme active sites is not a rare phenomenon. *J. Mol. Biol.* **372**, 817–845 (2007).

170. H. H. Zakon, Convergent evolution on the molecular level. *Brain Behav. Evol.* **59**, 250–261 (2002).

171. R. F. Doolittle, Convergent evolution: the need to be explicit. *Trends Biochem. Sci.* **19**, 15–18 (1994).

172. G. Dodson, A. Wlodawer, Catalytic triads and their relatives. *Trends Biochem. Sci.* **23**, 347–352 (1998).

173. O. D. Ekici, M. Paetzel, R. E. Dalbey, Unconventional serine proteases: variations on the catalytic Ser/His/Asp triad configuration. *Protein Sci.* **17**, 2023–2037 (2008).

174. F. Sunden, A. Peck, J. Salzman, S. Ressl, D. Herschlag, Extensive site-directed mutagenesis reveals interconnected functional units in the alkaline phosphatase active site. *Elife*. **4** (2015), doi:10.7554/eLife.06181.

175. D. R. Corey, M. E. McGrath, J. R. Vasquez, R. J. Fletterick, C. S. Craik, An alternate geometry for the catalytic triad of serine proteases. *J. Am. Chem. Soc.* **114**, 4905–4907 (1992).

176. C. S. Craik, S. Roczniak, C. Largman, W. J. Rutter, The catalytic role of the active site aspartic acid in serine proteases. *Science*. **237**, 909–913 (1987).

177. S. Sprang, T. Standing, R. J. Fletterick, R. M. Stroud, J. Finer-Moore, N. H. Xuong, R. Hamlin, W. J. Rutter, C. S. Craik, The three-dimensional structure of Asn102 mutant of trypsin: role of Asp102 in serine protease catalysis. *Science*. **237**, 905–909 (1987).

178. R. V. Stanton, M. Peräkylä, D. Bakowies, P. A. Kollman, Combined ab initio and Free Energy Calculations To Study Reactions in Enzymes and Solution: Amide Hydrolysis in Trypsin and Aqueous Solution. *J. Am. Chem. Soc.* **120**, 3448–3457 (1998).

179. J. K. Hwang, A. Warshel, Semiquantitative calculations of catalytic free energies in genetically modified enzymes. *Biochemistry*. **26**, 2669–2673 (1987).

180. S. N. Rao, U. C. Singh, P. A. Bash, P. A. Kollman, Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin. *Nature*. **328**, 551–554 (1987).

181. A. Warshel, G. Naray-Szabo, F. Sussman, J. K. Hwang, How do serine proteases really work? *Biochemistry*. **28**, 3629–3637 (1989).

182. J. Liu, R. Nussinov, Allostery: An overview of its history, concepts, methods, and applications. *PLoS Comput. Biol.* **12**, e1004966 (2016).

183. D. Thirumalai, C. Hyeon, P. I. Zhuravlev, G. H. Lorimer, Symmetry, rigidity, and allosteric signaling: From monomeric proteins to molecular machines. *Chem. Rev.* **119**, 6788–6821 (2019).

184. P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M. J. L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. **25**, 1422–1423 (2009).

185. C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, The Cambridge Structural Database. *Acta Crystallogr B Struct Sci Cryst Eng Mater*. **72**, 171–179 (2016).

186. Daylight Theory Manual (2011) (available at https://www.daylight.com/dayhtml/doc/theory/index.html).

187. I. Y. Chernyshov, I. V. Ananyev, E. A. Pidko, Revisiting van der Waals radii: From comprehensive structural analysis to knowledge-based classification of interatomic contacts. *Chemphyschem*. **21**, 370–376 (2020).

188. C. J. Williams, D. C. Richardson, J. S. Richardson, The importance of residue-level filtering and the Top2018 best-parts dataset of high-quality protein residues. *Protein Sci.* **31**, 290–300 (2022).

189. D. A. Case, H. M. Aktulga, K. Belfon, I. Y. Ben-Shalom, J. T. Berryman, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, G. A. Iii, V. W. D. Cisneros, T. A. Cruzeiro, R. E. Darden, G. Duke, M. K. Giambasu, H. Gilson, A. W. Gohlke, R. Goetz, S. Harris, S. A. Izadi, K. Izmailov,

M. C. Kasavajhala, E. Kaymak, A. King, T. Kovalenko, T. S. Kurtzman, S. Lee, P. Legrand, C. Li, J. Lin, T. Liu, R. Luchko, M. Luo, V. Machado, M. Man, K. M. Manathunga, Y. Merz, O. Miao, G. Mikhailovskii, H. Monard, K. A. Nguyen, A. Ohearn, F. Onufriev, S. Pan, R. Pantano, A. Qi, D. R. Rahnamoun, A. Roe, C. Roitberg, S. Sagui, A. Schott-Verdugo, J. Shajan, C. L. Shen, N. R. Simmerling, J. Skrynnikov, J. Smith, R. C. Swails, J. Walker, J. Wang, H. Wang, R. M. Wei, X. Wolf, Y. Wu, Y. Xiong, D. M. Xue, S. York, P. A. Zhao, Amber 2022. *Nancy_Zhu*. **10**, 37–38 (2022).

190. P. M. Zimmerman, Growing string method with interpolation and optimization in internal coordinates: method and examples. *J. Chem. Phys.* **138**, 184102 (2013).

191. A. Klamt, G. Schüürmann, COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2*, 799–805 (1993).

192. I. S. Ufimtsev, T. J. Martínez, Quantum chemistry on graphical processing units. 1. Strategies for two-electron integral evaluation. *J. Chem. Theory Comput.* **4**, 222–231 (2008).

193. I. S. Ufimtsev, T. J. Martinez, Quantum chemistry on graphical processing units. 2. Direct self-consistent-field implementation. *J. Chem. Theory Comput.* **5**, 1004–1015 (2009).

194. I. S. Ufimtsev, T. J. Martinez, Quantum chemistry on graphical processing units. 3. Analytical energy gradients, geometry optimization, and first principles molecular dynamics. *J. Chem. Theory Comput.* **5**, 2619–2628 (2009).

195. C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q. Wu, C. Simmerling, ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).

196. S. Izadi, R. Anandakrishnan, A. V. Onufriev, Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **5**, 3863–3871 (2014).

197. J.-P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).

198. P. P. Ewald, Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.* **369**, 253–287 (1921).

199. W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. **22**, 2577–2637 (1983).

200. G. Wouter, C. Touw, J. Baakman, A. H. Tim, E. Beek, R. P. Krieger, G. Joosten, A series of PDB related databases for everyday needs. *Nucleic Acids Research*. **43**, D364–D368 (2015).

201. H. B. Mann, D. R. Whitney, On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **18**, 50–60 (1947).

202. Schrödinger, LLC, The PyMOL Molecular Graphics System, Version 1.8 (2015).

203. S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, P. A. Kollman, THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13**, 1011–1021 (1992).

204. M. Souaille, B. Roux, Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput. Phys. Commun.* **135**, 40–57 (2001).

205. M. W. Hunkapiller, S. H. Smallcombe, D. R. Whitaker, J. H. Richards, Ionization Behavior of the Histidine Residue in the Catalytic Triad of Serine Proteases. *Journal of Biological Chemistry*. **248** (1973), pp. 8306–8308.

206. M. W. Hunkapiller, S. H. Smallcombe, D. R. Whitaker, J. H. Richards, Carbon nuclear magnetic resonance studies of the histidine residue in alpha-lytic protease. Implications for the catalytic mechanism of serine proteases. *Biochemistry*. **12**, 4732–4743 (1973).

207. W. W. Cleland, M. M. Kreevoy, Low-barrier hydrogen bonds and enzymic catalysis. *Science*. **264**, 1887–1890 (1994).

208. P. A. Frey, S. A. Whitt, J. B. Tobin, A low-barrier hydrogen bond in the catalytic triad of serine proteases. *Science*. **264**, 1927–1930 (1994).

209. P. A. Frey, Low-barrier hydrogen bonds. *Science*. **268** (1995), p. 189.

210. J. A. Gerlt, M. M. Kreevoy, W. Cleland, P. A. Frey, Understanding enzymic catalysis: the importance of short, strong hydrogen bonds. *Chem. Biol.* **4**, 259–267 (1997).

211. W. W. Cleland, P. A. Frey, J. A. Gerlt, The low barrier hydrogen bond in enzymatic catalysis. *J. Biol. Chem.* **273**, 25529–25532 (1998).

212. G. Robillard, R. G. Shulman, High resolution nuclear magnetic resonance studies of the active site of chymotrypsin. II. Polarization of histidine 57 by substrate analogues and competitive inhibitors. *J. Mol. Biol.* **86**, 541–558 (1974).

213. W. W. Bachovchin, Confirmation of the assignment of the low-field proton resonance of serine proteases by using specifically nitrogen-15 labeled enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 7948–7951 (1985).

214. A. A. Kossiakoff, S. A. Spencer, Neutron diffraction identifies His 57 as the catalytic base in trypsin. *Nature*. **288**, 414–416 (1980).

215. W. W. Bachovchin, W. Y. Wong, S. Farr-Jones, A. B. Shenvi, C. A. Kettner, Nitrogen-15 NMR spectroscopy of the catalytic-triad histidine of a serine protease in peptide boronic acid inhibitor complexes. *Biochemistry*. **27**, 7689–7697 (1988).

216. T. C. Liang, R. H. Abeles, Complex of .alpha.-chymotrypsin and N-acetyl-L-leucyl-L-phenylalanyl trifluoromethyl ketone: structural studies with NMR spectroscopy. *Biochemistry*. **26** (1987), pp. 7603–7608.

217. J. Lin, C. S. Cassidy, P. A. Frey, Correlations of the basicity of His 57 with transition state analogue binding, substrate reactivity, and the strength of the low-barrier hydrogen bond in chymotrypsin. *Biochemistry*. **37**, 11940–11948 (1998).

218. J. Lin, W. M. Westler, W. W. Cleland, J. L. Markley, P. A. Frey, Fractionation factors and activation energies for exchange of the low barrier hydrogen bonding proton in peptidyl trifluoromethyl ketone complexes of chymotrypsin. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14664–14668 (1998).

219. P. A. Sigala, E. A. Ruben, C. W. Liu, P. M. B. Piccoli, E. G. Hohenstein, T. J. Martínez, A. J. Schultz, D. Herschlag, Determination of Hydrogen Bond Structure in Water versus Aprotic Environments To Test the Relationship Between Length and Stability. *J. Am. Chem. Soc.* **137**, 5730–5740 (2015).

220. D. Herschlag, M. M. Pinney, Hydrogen Bonds: Simple after All? *Biochemistry*. **57**, 3338–3352 (2018).

221. G. Robillard, R. G. Shulman, High resolution nuclear magnetic resonance studies of the active site of chymotrypsin. I. The hydrogen bonded protons of the "charge relay" system. *J. Mol. Biol.* **86**, 519–540 (1974).

222. W. W. Bachovchin, R. Kaiser, J. H. Richards, J. D. Roberts, Catalytic mechanism of serine proteases: reexamination of the pH dependence of the histidyl 1J13C2-H coupling constant in the catalytic triad of alpha-lytic protease. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 7323–7326 (1981).

223. E. L. Ash, J. L. Sudmeier, E. C. De Fabo, W. W. Bachovchin, A low-barrier hydrogen bond in the catalytic triad of serine proteases? Theory versus experiment. *Science*. **278**, 1128–1132 (1997).

224. C. N. Fuhrmann, M. D. Daugherty, D. A. Agard, Subangstrom crystallography reveals that short ionic hydrogen bonds, and not a His-Asp low-barrier hydrogen bond, stabilize the transition state in serine protease catalysis. *J. Am. Chem. Soc.* **128**, 9086–9102 (2006).

225. T. Tamada, T. Kinoshita, K. Kurihara, M. Adachi, T. Ohhara, K. Imai, R. Kuroki, T. Tada, Combined high-resolution neutron and X-ray analysis of inhibited elastase confirms the active-site oxyanion hole but rules against a low-barrier hydrogen bond. *J. Am. Chem. Soc.* **131**, 11033–11040 (2009).

226. S. O. Shan, S. Loh, D. Herschlag, The energetics of hydrogen bonds in model systems: implications for enzymatic catalysis. *Science*. **272**, 97–101 (1996).

227. M. Kunitz, J. H. Northrop, ISOLATION FROM BEEF PANCREAS OF CRYSTALLINE TRYPSINOGEN, TRYPSIN, A TRYPSIN INHIBITOR, AND AN INHIBITOR-TRYPSIN COMPOUND. *J. Gen. Physiol.* **19**, 991–1007 (1936).

228. S. Ranasinghe, D. P. McManus, Structure and function of invertebrate Kunitz serine protease inhibitors. *Dev. Comp. Immunol.* **39**, 219–227 (2013).

229. D. E. Bowman, Fractions Derived from Soy Beans and Navy Beans Which Retard Tryptic Digestion of Casein. *Proc. Soc. Exp. Biol. Med.* **57**, 139–140 (1944).

230. Y. Birk, The Bowman-Birk inhibitor. Trypsin- and chymotrypsin-inhibitor from soybeans. *International Journal of Peptide and Protein Research*. **25** (2009), pp. 113–131.

231. L. A. Kazal, D. S. Spicer, R. A. Brahinsky, Isolation of a Crystalline Trypsin Inhibitor-Anticoagulant Protein from Pancreas1a. *J. Am. Chem. Soc.* **70**, 3034–3040 (1948).

232. V. Rimphanitchayakit, A. Tassanakajon, Structure and function of invertebrate Kazal-type serine proteinase inhibitors. *Dev. Comp. Immunol.* **34**, 377–386 (2010).

233. S. Murao, S. Sato, S-SI, a new alkaline protease inhibitor from Streptomyces albogriseolns S-3253. *Agric. Biol. Chem.* **36**, 160–163 (1972).

234. S. Sato, S. Murao, Isolation and Crystallization of Microbial Alkaline Protease Inhibitor, S-SI. *Agric. Biol. Chem.* **37**, 1067–1074 (1973).

235. M. Laskowski Jr, I. Kato, Protein inhibitors of proteinases. *Annu. Rev. Biochem.* **49**, 593–626 (1980).

236. E. S. Radisky, D. E. Koshland Jr, A clogged gutter mechanism for protease inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 10316–10321 (2002).

237. E. Zakharova, M. P. Horvath, D. P. Goldenberg, Functional and Structural Roles of the Cys14–Cys38 Disulfide of Bovine Pancreatic Trypsin Inhibitor. *J. Mol. Biol.* **382**, 998–1013 (2008).

238. M. L. Colgrave, M. J. L. Korsinczky, R. J. Clark, F. Foley, D. J. Craik, Sunflower trypsin inhibitor-1, proteolytic studies on a trypsin inhibitor peptide and its analogs. *Biopolymers*. **94**, 665–672 (2010).

239. H. J. Schirra, R. F. Guarino, M. A. Anderson, D. J. Craik, Selective Removal of Individual Disulfide Bonds within a Potato Type II Serine Proteinase Inhibitor from Nicotiana alata Reveals Differential Stabilization of the Reactive-Site Loop. *J. Mol. Biol.* **395**, 609–626 (2010).

240. J. E. Swedberg, S. J. de Veer, K. C. Sit, C. F. Reboul, A. M. Buckle, J. M. Harris, Mastering the canonical loop of serine protease inhibitors: enhancing potency by optimising the internal hydrogen bond network. *PLoS One*. **6**, e19302 (2011).

241. J. D. McBride, A. B. Brauer, M. Nievo, R. J. Leatherbarrow, The role of threonine in the P2 position of Bowman-Birk proteinase inhibitors: studies on P2 variation in cyclic peptides encompassing the reactive site loop. *J. Mol. Biol.* **282**, 447–458 (1998).

242. J. Dasgupta, S. Khamrui, J. K. Dattagupta, U. Sen, Spacer Asn determines the fate of Kunitz (STI) inhibitors, as revealed by structural and biochemical studies on WCI mutants. *Biochemistry*. **45**, 6783–6792 (2006).

243. U. C. Marx, M. L. J. Korsinczky, H. J. Schirra, A. Jones, B. Condie, L. Otvos Jr, D. J. Craik, Enzymatic cyclization of a potent bowman-birk protease inhibitor, sunflower trypsin inhibitor-1, and solution structure of an acyclic precursor peptide. *J. Biol. Chem.* **278**, 21782–21789 (2003).

244. E. Zabłotna, K. Kaźmierczak, A. Jaśkiewicz, M. Stawikowski, G. Kupryszewski, K. Rolka, Chemical Synthesis and Kinetic Study of the Smallest Naturally Occurring Trypsin Inhibitor SFTI-1 Isolated from Sunflower Seeds and Its Analogues. *Biochem. Biophys. Res. Commun.* **292**, 855–859 (2002).

245. M. Laskowski, R. W. Sealock, "11 Protein Proteinase Inhibitors—Molecular Aspects" in *The Enzymes*, P. D. Boyer, Ed. (Academic Press, 1971), vol. 3, pp. 375–473.

246. A. Kashima, Y. Inoue, S. Sugio, I. Maeda, T. Nose, Y. Shimohigashi, X-ray crystal structure of a dipeptide-chymotrypsin complex in an inhibitory interaction. *Eur. J. Biochem.* **255**, 12–23 (1998).

247. Y. Shimohigashi, I. Maeda, T. Nose, K. Ikesue, H. Sakamoto, T. Ogawa, Y. Ide, M. Kawahara, T. Nezu, Y. Terada, K. Kawano, M. Ohno, Chymotrypsin inhibitory conformation induced by amino acid side chain–side chain intramolecular CH/π interaction. *J. Chem. Soc. Perkin 1*, 2479–2485 (1996).

248. K. T. Simons, C. Kooperberg, E. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225 (1997).

249. M. Gerstein, R. B. Altman, Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* **251**, 161–175 (1995).

250. J. L. Radkiewicz, C. L. Brooks, Protein Dynamics in Enzymatic Catalysis: Exploration of Dihydrofolate Reductase. *J. Am. Chem. Soc.* **122**, 225–231 (2000).

251. J. A. Irving, J. C. Whisstock, A. M. Lesk, Protein structural alignments and functional genomics. *Proteins*. **42**, 378–382 (2001).

252. A. Amadei, A. B. Linssen, H. J. Berendsen, Essential dynamics of proteins. *Proteins*. **17**, 412–425 (1993).

253. P. H. Hünenberger, A. E. Mark, W. F. van Gunsteren, Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *J. Mol. Biol.* **252**, 492–503 (1995).

254. R. Abseher, M. Nilges, Are there non-trivial dynamic cross-correlations in proteins? *J. Mol. Biol.* **279**, 911–920 (1998).

255. T. Noguti, N. Gō, Dynamics of Native Globular Proteins in Terms of Dihedral Angles. *J. Phys. Soc. Jpn.* **52**, 3283–3288 (1983).

256. J. K. Bray, D. R. Weiss, M. Levitt, Optimized torsion-angle normal modes reproduce conformational changes more accurately than cartesian modes. *Biophys. J.* **101**, 2966–2969 (2011).

257. M. Levitt, C. Sander, P. S. Stern, Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* **181**, 423–447 (1985).

258. S. Hayward, N. Go, Collective variable description of native protein dynamics. *Annu. Rev. Phys. Chem.* **46**, 223–250 (1995).

259. G. G. Maisuradze, D. M. Leitner, Free energy landscape of a biomolecule in dihedral principal component space: sampling convergence and correspondence between structures and minima. *Proteins*. **67**, 569–578 (2007).

260. Y. Mu, P. H. Nguyen, G. Stock, Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins*. **58**, 45–52 (2005).

261. J. N. Brønsted, K. J. Pedersen, Zeitschrift für Phys. Chemie. *Stöchiometrie und Verwandtschaftslehre*. **108**, 185–235 (1924).

262. G. Robillard, R. G. Shulman, High resolution nuclear magnetic resonance study of the histidine—Aspartate hydrogen bond in chymotrypsin and chymotrypsinogen. *J. Mol. Biol.* **71**, 507–511 (1972).

263. M. I. Page, W. P. Jencks, Intramolecular general base catalysis in the aminolysis of acetylimidazole and methyl formate by diamines. *J. Am. Chem. Soc.* **94**, 8818–8827 (1972).

264. S. C. L. Kamerlin, Z. T. Chu, A. Warshel, On Catalytic Preorganization in Oxyanion Holes: Highlighting the Problems with the Gas-Phase Modeling of Oxyanion Holes and Illustrating the Need for Complete Enzyme Models. *The Journal of Organic Chemistry*. **75** (2010), pp. 6391–6401.

265. M. M. Pinney, A. Natarajan, F. Yabukarski, D. M. Sanchez, F. Liu, R. Liang, T. Doukov, J. P. Schwans, T. J. Martinez, D. Herschlag, Structural Coupling Throughout the Active Site Hydrogen Bond Networks of Ketosteroid Isomerase and Photoactive Yellow Protein. *J. Am. Chem. Soc.* **140**, 9827–9843 (2018).

266. D. A. Kraut, P. A. Sigala, B. Pybus, C. W. Liu, D. Ringe, G. A. Petsko, D. Herschlag, Testing Electrostatic Complementarity in Enzyme Catalysis: Hydrogen Bonding in the Ketosteroid Isomerase Oxyanion Hole. *PLoS Biology*. **4** (2006), p. e99.

267. A. Natarajan, J. P. Schwans, D. Herschlag, Using unnatural amino acids to probe the energetics of oxyanion hole hydrogen bonds in the ketosteroid isomerase active site. *J. Am. Chem. Soc.* **136**, 7643–7654 (2014).

268. J. L. Markley, M. A. Porubcan, The charge-relay system of serine proteinases: proton magnetic resonance titration studies of the four histidines of porcine trypsin. *J. Mol. Biol.* **102**, 487–509 (1976).

269. J. L. Markley, I. B. Ibañez, Zymogen activation in serine proteinases. Proton magnetic resonance pH titration studies of the two histidines of bovine chymotrypsinogen A and chymotrypsin Aalpha. *Biochemistry*. **17**, 4627–4640 (1978).

270. T. K. Harris, A. S. Mildvan, High-precision measurement of hydrogen bond lengths in proteins by nuclear magnetic resonance methods. *Proteins*. **35**, 275–282 (1999).

271. N. D. Rawlings, M. Waller, A. J. Barrett, A. Bateman, MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **42**, D503-9 (2014).

Acknowledgments: We thank members of the Herschlag lab for thoughtful discussions and manuscript feedback.

Funding:

National Science Foundation grant MCB2322069 (DH, TJM)

Author contributions:

Conceptualization: SD, RCK, JPG, MMP, DH

Methodology (pseudo-ensembles): SD, RCK, JPG, FY

Data curation (pseudo-ensembles and additional crystallographic data): SD, RCK, JPG

Data curation (QM and MD): EP, VWDC, MZ

Investigation: SD, RCK, JPG, JPS

Visualization: SD

Funding acquisition: DH, TJM

Writing – original draft: SD, DH

Writing – review & editing: all authors

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

Data and materials availability: All data and computer code used for the analyses are available in the supplementary materials or at doi: 10.5281/zenodo.10724909. The *Ensemble*PDB package is available at: https://github.com/Herschlag-Lab/EnsemblePDB.

Supplementary Materials

Methods

Supplementary Text

Figs. S1 to S70

Tables S1 to S25

References (*184–271*)