Published in final edited form as:

J Chem Inf Model. 2019 February 25; 59(2): 931–943. doi:10.1021/acs.jcim.8b00950.

Cys.sqlite: A structured-information approach to the comprehensive analysis of cysteine disulfide bonds in the Protein Databank

Theodore L. Fobe[†,‡], **Andrei Kazakov**[¶], **Demian Riccardi**[¶]

[†]University of Maryland, Department of Chemical and Biomolecular Engineering, College Park, Maryland, 20742

[‡]Summer Undergraduate Research Fellowship, National Institute of Standards and Technology, Boulder, CO

[¶]National Institute of Standards and Technology, Thermodynamic Research Center (647), 325 Broadway, Boulder, CO 80305

Abstract

Cysteine is a multifaceted amino acid that is central to the structure and function of many proteins. A disulfide bond formed between two cysteines restrains protein conformations through the strong covalent bond and torsions about the bond that prefer, energetically, ±90 degrees. In this study, we transform over 30k Protein Databank files (PDBx/mmCIFs) into a single file, SQLite database (Cys.sqlite). The database schema is designed to accommodate the structural information of both oxidized and reduced cysteines and to retain essential protein metadata to establish informational and biological provenance. Cys.sqlite contains over 95k peptide chains and 500k cysteines (700k structural conformers); there are over 265k cysteine disulfide bond conformations from structures solved with all available experimental methods. The structural information is analyzed with respect to sequence identity cutoff, the experimental method, and energetics of the disulfide. We find that as the experimental information becomes limiting and the influence of modeling becomes more pronounced, the observed average strain increases artificially. The database and analyses presented here can be used to improve the refinement of biological structures from experiments that are known to contain one or more disulfide bonds.

Introduction

The presence of cysteine residues in a protein sequence implies a diverse set of chemical and structural activities unique to the other encoded amino acids.[1] The cysteine sidechain extends a sulfhydryl group ($-C_\beta SH$) that has a pK_a (~ 8–9) slightly above the pH of

demian.riccardi@nist.gov.

Supporting Information Available

The library and scripts used to construct and analyze Cys.sqlite is available at https://github.com/usnistgov/Cys.sqlite.

biological conditions (7.4). The cysteine thiolate forms strong bonds with metal ion cofactors that are critical to enzyme catalysis and metal ion homeostatis.[2] Under oxidizing conditions and sufficiently high pH, a disulfide bond is readily formed between two proximate cysteine residues. Disulfide bonds join different regions of proteins, which can be separated in sequence space within a polypeptide chain or between two different chains. These covalent interactions provide stability to the protein native state by constraining the number of configurations available to the unfolded state.[3]

Organisms have evolved redox proteins to maintain specific structural or functional activities of cysteine residues that depend on the subcellular environment; cysteines are actively reduced in the cytoplasm and oxidized/isomerized in the periplasm.[4–7] Such active intervention is required because disulfide bonds are strong (~250 kJ·mol^{-1}) yet susceptible to exchange with other thiolates at room temperature and sufficiently high pH (exchange barrier ~ 60 kJ·mol^{-1}).[8–11] The dihedral angle about the disulfide bond (C_β-S-S-C_β, Fig.1) favors ± 90 degrees energetically, providing additional structural restraints on the folded state of the protein. The folding free energy can be used to twist and strain the conformation of the disulfide to directly increase its reactivity to thiolate exchange or reduction. These attributes have led to many experimental and computational studies, spanning over 30 years, into how disulfide bonds (natural or engineered) influence protein structure[12–22] and allostery,[23–27] perturb biochemical activity,[28] facilitate cellular redox signaling,[29] react under mechanical stress,[30,31] enable engineered peptide therapeutics[32] and provide targets for cancer therapy.[33]

Recent structural surveys have used the concept of disulfide strain energy to identify and predict the presence of allosteric switches[24,25] and between-strand disulfide (BSD) redox switches found in β sheets.[21,22,34,35] Surveys by Hogg and coworkers[24,25] used a simple, empirical model of disulfide strain energy (DSE, in units of kilojoules per mole of disulfide) that was introduced in an earlier study of disulfides engineered into subtilisin.[15] NMR structures were found to have significantly higher populations of disulfide bonds with high strain energy.[25] However, it is unclear whether the populations of highly strained disulfide bonds observed in NMR result from the limitations of the structural models.

All Protein Databank structures are models, which use theoretical insights and complex computational algorithms,[38–41] based on experimental information. For example, structures determined from X-ray crystallography represent diffracted intensities from large lattices of repeating units. Configurations that vary with lattice displacements reduce the resolution and the amount of information available for modeling. Within a Protein Databank file, the effects of uncorrelated atomic motion or configurational variability are present as modeling parameters, *i.e.*, temperature factors and occupancies. In fact, disulfide bonds are susceptible to reduction when exposed to X-rays during experiments; a process that generally increases B factors and reduces occupancies of the associated cysteines.[42–46] The structures from NMR and cyro electron microscopy model information are typically gathered from solution phase configurations. Without the restraints of the crystal environment, the structure is difficult to define uniquely. For all methods, as the information degrades, the structural representation of the model becomes less accurate. The present study seeks to quantify how

cysteine disulfide structural information depends on different experimental methods and structural resolutions.

At this time, the Protein Databank has over 138k entries;[47] around 22 % of them contain one or more disulfide bonded cysteines. Removing entries with over 50 % sequence identity, the number of entries drops to around 40k, of which around 16 % contain disulfide bonds. This number grows each week as new structures are released. Typical structural surveys involve parsing thousands of files (in PDB or PDBx/mmCIF formats) to acquire the information needed; parsing efforts are complicated by edge cases, and the datasets used are quickly out of date. The present study uses the unique characteristics of cysteine to structure the information within a single-file database (Cys.sqlite) that retains biological and structural provenance, resolves problematic edge cases. Cys.sqlite is simple to acquire, update, and use.

Beyond aiding future efforts to derive new biological information, the structural information provided in Cys.sqlite may be used to improve structural models where disulfide moieties are present. Similar to the “allowed” and “forbidden” regions of a Ramanchandran plot[48] consequential to sidechain packing and secondary structure, the strain energy of a disulfide should influence the native structure containing them. Whether a given biological structure or complex contains one or more disulfide bonds can be established with reasonable certainty experimentally. Thus, extensive structural information specific to disulfide conformations can directly impact structural models of experimental information and structure prediction as has been established with applications of backbone-dependent libraries of side chain rotamers.[49–55]

Methods

All Protein Databank entries containing one or more disulfide bonds are identified using XML queries, and the corresponding PDBx/mmCIF files downloaded from rcsb.org using FTP downloads.[47,56] Compared to the previous standard PDB format, PDBx/mmCIF files contain more well-structured information and support for large structures.[57] Such queries depend on the accumulation of metadata by the RCSB. Erroneous information has the potential to generate entries with no cysteine disulfide bonds and miss entries with cysteine disulfide bonds. The former is discussed in more detail below; for the latter, entries are added as they are identified. For example, several entries associated with a study of radiation damage on disulfide bonds[44] do not report the presence of disulfide bonds in the PDBx/mmCIF files and are not resolved by RCSB XML queries.

The relevant information is extracted from each file and entered into an SQLite database using HackaMol[58] and DBIx::Class; both libraries are open source and available from the Comprehensive Perl Archive Network.[59] SQL is an acronym for Structured Query Language, and SQLite is a SQL database engine.[60] SQLite was chosen because it is in the public domain and can be used directly from most programming languages; additionally, an SQLite database is a single file that can be downloaded and then accessed from freely available database management tools. To facilitate future use of the database, we describe the schema in detail to clarify its structure and the logical decisions made during

construction. Overall, the schema was developed to accommodate all Protein Databank structures. Some columns are not shared by all experimental methods; for example the resolution (*PDB.resolution*) is a null value for NMR structures whereas there is only one model number (*Cys_Conf.model num = 1*) for most X-ray structures. Water molecules and hydrogen atoms, which are typically included in X-ray and NMR structures, respectively, are irrelevant to the current study and were ignored throughout the construction of Cys.sqlite.

General issues of constructing such a database are the continual growth and evolution of the Protein Databank. New structures are released each week. A structure may be deposited and then replaced by a subsequent, improved structure. In the present study, all available structural information as of December 12, 2018 was processed and entered into the database. With each update, obsolete structures are not removed; rather, SQL queries can be used to curate the lists of PDB Ids used in analysis. Furthermore, the schema may be expected to evolve with the needs of each application. The code used to generate the database, along with example queries and scripts used in the present analysis, are included in the GitHub repository (https://github.com/usnistgov/Cys.sqlite) to enable other users to interact with the database and/or modify the database schema, which is described next.

Cys.sqlite

Overall, the database contains six tables, two being major parent tables, PDB and Entity Cys, and the four remaining are hierarchically related (Fig. 2). The PDB table contains information specific to a given Protein Databank entry while the Entity_Cys table contains all of the unique amino acid sequences containing cysteine residues. The Chain_Cys table is the child table of both the PDB and Entity_Cys tables and parent to the Cys table; the Cys table is parent to the Cys_Conf table, which is parent to the Cys_Cys table. Table relationships are established with foreign key columns of the child table that contain the primary key of the parent table. The design of the schema makes some tradeoffs of file size for convenience. For example, the Cys_Conf table contains all information required to rebuild the original coordinates of the corresponding cysteine residue; thus, the Cys_Cys table contains information, such as the S-S bond length, that is essentially redundant. To simplify SQL queries, the schema cascades many foreign keys to the child tables below (Fig. 2); for example, the Cys_Cys table contains foreign keys for the PDB and two Entity_Cys, one for each cysteine, entries that may exist in different entities. Combining the two examples, the added foreign keys trivializes the query of all S-S bond lengths for a given entity at the expense of a larger Cys.sqlite file. Relevant details of the schema are described below, see SI for detailed tables and columns.

Each row of the parent table, *PDB*, corresponds to a single Protein Databank file with the PDB ID as the primary key (*id*). The RCSB REST API is used to determine the status (*status* of “CURRENT” or “OBSOLETE”) and populate the sequence identity cutoff (*exp_method_identity_cutoff*) separately for each experimental method for all current entries; all values of status and identity cutoff are updated after new entries are added. The present study includes cutoffs of 50 % and 100 % identity for X-ray diffraction, solution NMR, and electron microscopy entries. All other current entries will have a NULL entry denoting no applied identity cutoff. The 50 % group is contained within the 100 % group.

Thus, the 100% group can be selected from all current entries matching 50% or 100%, or those that are not null; the overall set with no identity cutoffs corresponds to all "CURRENT" entries. The overall number of chains, residues, and cysteines are determined from the sequence (*chain_count*, *res_count*, and *cys_count* respectively) for both the PDB and Entity_Cys tables.

The use of sequence identity values generated by RCSB to reduce redundant structural information is routinely reported in the literature; however, the sequence identity of a given PDB entry is somewhat obfuscated when more than one biological entity are present in the reported structure. For example, there are solution NMR structures of the Lac repressor DNA-binding domain complexed with different DNA sequences (PDB IDs: 2KEI, 2KEJ, 2KEK, 1L1M, 1OSL). The sequence of the protein is exactly the same, but the entries are all within the RCSB 50% cutoff due to differences in the DNA sequence. Clearly, the treatment of such an entry depends on the context of the analysis. Such considerations motivated the design of the Entity_Cys table. Each entry of the Entity_Cys table corresponds to a unique sequence that contains at least one cysteine. As with all tables except the *PDB* table, the primary key *id* is an integer that is incremented with each entry. Entities without cysteines are ignored. Entities are added only if the entity is present in Cys structural coordinates; this sidesteps the need to use heuristics to distinguish DNA/RNA sequences from protein sequences as both may shared characters, such as C. The Entity_Cys table facilitates the exploration of how a given entity varies in different structures. The table can also be used to launch more elaborate sequence analyses.

Each row of the *Chain_Cys* table (Fig. 2) is related to a single *Entity_Cys* and *PDB* entry through foreign keys *entity_id* and *pdb_id*, respectively. Thus, *Chain_Cys* provides a pivot between the two tables. For example, a 100% set of X-ray diffraction structures, reduced with respect to the RCSB generated set described above, can be constructed from queries on the Entity_Cys table pivoted to the highest resolution structure from the PDB table.

Each row of the *Cys* table is connected to the *Chain_Cys*, *PDB*, and *Entity_Cys* tables via the *chain_id*, *pdb_id*, and *entity_id* foreign keys, respectively (Fig. 2). The table contains metadata specifying the sequence id (*seq_id*) and insertion code (*insert_code*), which is most often NULL. Each *Cys* entry may have one or more cysteine conformers resulting from multiple models, often in solution NMR structures, or coordinate occupancies, as is often in X-ray diffraction structures.

The *Cys_Conf* table is the widest table of the schema. Each row contains the information needed to rebuild the original Cartesian coordinates of each complete cysteine residue corresponding to a *Cys* table entry. A conformer is skipped only if a complete cysteine cannot be built, uniquely; this arises for missing coordinates or logical ambiguities arising when more than two alternative conformers are present. The backbone dihedrals (ω,ϕ, and ψ) are stored if the cysteine has a neighboring residue before (ω and ϕ) or after (ψ). For example, a terminal cysteine will have a null entry for the ψ angle. However, null entries for backbone dihedrals should not be used to imply a terminal location, because neighboring residues may not be structurally resolved. The cysteine neighbors are determined geometrically rather than by using the sequence due to the ambiguity introduced by residue

insertion codes in some proteins. Identifying neighbors for cysteines containing alternative locations requires that the alternate id (*alt_id*) be the same or missing for a candidate atom. The backbone dihedrals and the conventional χ_1 angle (*N_CA_CB_SG*, see SI), is consistent with information included for cysteine sidechains in datasets used to compile rotamer libraries.[49,50,52,54] Similarly, the total number of bonds to the sulfur atom, not including hydrogen atoms, are determined for each conformation. The bond counts allow the entries to be separated into sets of reduced and oxidized cysteines that may include other more complex structural scenarios, such as bonds to metal ions or sulfur containing compounds (not included in the *Cys* table).

The *Cys_Cys* table is constructed from two rows of the *Cys_Conf* constituting cysteine disulfide bonds. A cysteine disulfide bond is entered in the table if the bond between cysteines is exclusive, and the bond distance is no more 0.231 nm, i.e. 0.025 nm greater than twice the sulfur covalent radius (0.103 nm).[61] For example, there are no cysteine disulfide bonds corresponding to the entry for PDBID 1F3H, which contains 3 cysteines sharing 2 disulfides and a coordinated Zn^{2+}. Each row of the *Cys_Cys* table contains the internal coordinates for each cysteine disulfide. The *Cys_Cys* table contains columns of internal coordinates that will be analyzed in the present study.

The *Cys_Cys* table also includes the disulfide strain energy calculated from a simple model that was introduced by Katz and Kossiakoff[15] and applied by Hogg and coworkers,[24,25]

$$\begin{aligned}\mathrm{DSE}\left(\mathrm{kJ}\cdot\mathrm{mol}^{-1}\right) &= 8.37\left(1+\cos\left(3\chi_1\right)\right)+8.37\left(1+\cos\left(3\chi_1'\right)\right)\\ &+4.18\left(1+\cos\left(3\chi_2\right)\right)+4.18\left(1+\cos\left(3\chi_2'\right)\right)\\ &+14.64\left(1+\cos\left(2\chi_3\right)\right)+2.51\left(1+\cos\left(3\chi_3\right)\right).\end{aligned} \quad (1)$$

This measure of DSE depends on the five dihedral angles (χ_1, χ_2, χ_3, χ'_2, and χ'_1, see Fig. 1) along the N-CA-CB-SG atoms of both cysteine residues. All parameters, except for the χ_1 parameter, are present in the AMBER forcefield and still used.[62] The model introduced by Katz and Kossiakoff[15] increased the χ_1 parameter from 5.86 to 8.37 kJ/mol. The source of this change was not clearly reported. The *Cys_Cys* table contains values for the DSE model of Eq. 1 to remain consistent with earlier studies. The model is compared to gas-phase quantum chemical calculations of dipropyl disulfide.

Structural and energetic analysis

Gas-phase calculations—Gas-phase quantum chemical calculations carried out for a collection of unique dipropyl disulfide molecule conformations are compared to the model of DSE described above. While neglecting interactions important to the energetics of cysteine disulfides, such as those between polar atoms or those with the surrounding protein, gas-phase dipropyl disulfide does retain key intramolecular interactions present in cysteine disulfides. Collections of similar nonpolar molecules are often used in forcefield development, e.g. Ref. 63. The comparison is expected to provide coarse validation and insights into the limitations of the simple DSE model. Cys.sqlite can be used to frame more detailed investigations, such as modeling the effects of the protein environment on strain,

reduction potentials, or X-ray radiation damage,[46] that are beyond the scope of the present investigation.

The gas-phase conformations are generated from 22.5 degree bins of the the five torsions for all X-ray structures with resolution < 0.2 nm. The set was culled from around 6500 conformations to 573 by eliminating degenerate DSE contributions (calculated in kcal/mol) from combined $\chi_1 : \chi'_1$, combined $\chi_2 : \chi'_2$, and χ_3 computed to 2, 2, and 1 decimal places, respectively. All hydrogens were added 0.1 nm from the sp3 carbons at appropriate angles to generate conventional geometries. All conformers are geometry optimized with the five torsion angles constrained to values of the 22.5 degree bins. Geometry optimization and corresponding conformer energies were determined using density-fitted second-order Møller-Plesset perturbation theory (DF-MP2) as implemented in Psi4 v1.1.[64,65] All DF-MP2 calculations were carried out using triple zeta correlation-consistent Dunning basis sets[66,67] with tight-d functions on sulfur (aug-cc-PV(T+d)Z).[68]

Distributions of internal coordinates and DSE—Beyond comparisons to gas-phase quantum chemical calculations, distributions of DSE (Eq.1) are compared for different experimental methods (X-ray, NMR, or EM). DSE is analyzed with respect to the resolution of the structure (X-ray and EM). The DSE is collected accordingly and compared using the mean, STDEV, and standard error of the mean. The cysteine disulfide torsion angles are analyzed with respect to the components of the DSE model to further validate the model and characterize the influence of experimental approach on the observed populations. For each comparison, the overall set of X-ray structures is included along with high-resolution (≤0.15 nm) and low-resolution (>0.28 nm) subsets.

Probability distributions of the internal coordinates are computed from Cys.sqlite to characterize disulfide structures in the Protein Databank; the distributions are varied by experimental method to characterize the associated effects. Throughout the present study, the set of structures corresponding to an experimental method are denoted by the *exp_method_identity_cutoff* values of “NONE” for the entire database and those filtered to have no more than 50 % or100 % sequence identity. Each configuration within a given set is treated as independent; for example, each member of a solution NMR structure ensemble is treated as if it were independent.

The present study models the DSE dependence of the Ramachandran densities,[48] $\rho(\phi, \psi|$ DSE). The cysteine disulfides from the set of high resolution (≤ 0.15 nm) X-ray structures with 50% sequence identity cutoff is separated into four subsets using standard deviations of the mean (all in units of $kJ \cdot mol^{-1}$: DSE ≤ 5.3 ; 5.3 < DSE < 10.6 ; 10.6 ≤ DSE < 15.9; DSE ≥ 15.9). The regions of DSE are analogous to side-chain rotamers used to construct rotamer libraries. The present study uses two-dimensional adaptive kernel density estimates (AKDE) to provide robust models of the Ramachandran densities in spite of data that is multimodal and may be sparsely sampled in important regions. We follow the approach described by Shapovalov and Dunbrack in Ref. 54 and the associated, detailed, supplemental information. Briefly, two-dimensional adaptive kernel density estimates are calculated for each of the regions of DSE,

$$\rho(\phi,\psi\,|\,\mathrm{DSE}) = \frac{1}{4\pi^2 N_{DSE}} \sum_{i=1}^{N_{DSE}} \frac{1}{\left[I_0\left(\frac{\kappa}{\lambda_i}\right)\right]^2} \exp\left[\frac{\kappa}{\lambda_i}\left(\cos(\phi-\phi_i)+\cos(\psi-\psi_i)\right)\right], \quad (2)$$

using a von Mises distribution for each coordinate (ϕ and ψ). The von Mises distribution,

$$\rho(\theta\,|\,\theta_0,\kappa) = \frac{1}{I_0(\kappa)} \exp(\kappa\cos(\theta-\theta_0)), \quad (3)$$

is an appropriate model for probability densities of a circular parameter (θ), where $I_0(\kappa)$ is the modified Bessel function of first kind, and κ is the von Mises concentration parameter. The scaling parameter, λ_i, in Eq. 2 makes the kernel density estimate adaptive to the data. It is provided by the pilot estimate of the Ramachandran density computed over all DSE regions,

$$\lambda_i = \sqrt{\left(\frac{\left(\Pi_i^N \hat{f}(\phi_i,\psi_i)\right)^{\frac{1}{N}}}{\hat{f}(\phi_i,\psi_i)}\right)}, \quad (4)$$

where N is the total number of data points. The pilot function, $\hat{f}(\phi_i,\psi_i)$, is a nonadaptive KDE (Eq. 2 with a $\lambda_i = 1$) that re-weights the concentration parameter at each point. In the present study, both the AKDE (Eq. 2) and pilot estimate functions are calculated with a κ of 100. The R script used to calculate the ADKE maps is included in the GitHub repository.

Results and Discussion

Overall content summary

The uncompressed PDBx/mmCIF files consist of around 38 gigabytes at the time of writing. The database is a factor of 100 smaller (around 380 megabytes), which is reduced by a factor of 2–3 with compression. Since the initial compilation of the database (2018-09-18), 19 entries have become obsolete. Cys.sqlite contains over 32k PDB, 100k *Chain_Cys*, 545k Cys, 770k *Cys_Conf*, and 290k *Cys_Cys* entries, see Table 1 for a breakdown of entries by the three largest contributing experimental methods (X-ray, NMR, and EM). Other methods include solid-state NMR, powder diffraction, neutron diffraction, electron crystallography, and one entry for infrared spectroscopy (1SSZ).

As expected from the overall growth in the Protein Databank,[47] structure deposits with one or more disulfide bonds has been growing rapidly in recent years for both X-ray and EM. The lag between when the structure is deposited and when it is released can be observed in the drop from 2016 to 2017 (Fig. 3) for X-ray. Structures deposited in 2016 and 2017 continue to be released each week in 2018. The number of EM structures deposited has grown from around 1–10 per year from 1999 to 2012 to 248 (and climbing) for 2018 (as released by 2018-12-12). The number of NMR deposits has remained consistent at ~100 since 1995. Between 1990 and 2005, the number of disulfide bond configurations deposited

from NMR was larger than X-ray (Fig. 3) due to the number of models included with each NMR structure. The large peak for EM in the number of cysteines and disulfide bonds in 2013 corresponds to the deposition of two HIV-1 capsid structures (3J3Y and 3J3Q) that contributed 1176 and 1356 disulfide entries in the *Cys_Cys* table, respectively. Along with 3J4F and 3J34, these entries share a single entity and are associated with an EM study that used molecular dynamics for structural refinement.[69]

EM structures are composed of multiple peptide chains while most NMR entries include multiple configurations. As a result the overall number of chains and cysteines is much lower for NMR compared to X-ray or EM, but the number of cysteine conformations and disulfide bond entries is much larger. There are slightly more than one conformation per cysteine for both X-ray and EM due to multiple occupancies. There are ~18 conformations per cysteine for NMR, which agrees well with the typical number of models reported. At 50% cutoff, around one half, one fourth, and one twelfth of cysteines in the *Cys* table are involved in disulfide bond entries of the *Cys_Cys* table for NMR, X-ray, and EM, respectively (Table 1).

There is a large number of structures in the PDB that are degenerate with respect to the entity. There are ~32k entries in the *PDB* table but only around 25k in the *Entity_Cys* table. The top five most frequently solved entities are found in 2.2k *PDB* entries, the top entity being that of lysozyme (*Entity_Cys.id*=7) with 699 PDB entries and around 4.8 cysteine disulfides per entity. Using identity cutoffs, the degeneracy is largest for X-ray, which is reduced by a factor of I~5 comparing the entire set (NONE) to 50 % sequence identity; NMR and EM are reduced by around half for the same comparison (Table 1, PDB row). On average for the 100% cutoff, there are approximately 1.2 and 5.3 *Entity_Cys* entries per *PDB* entry for X-ray diffraction and EM, respectively. The larger number of entities for EM reflects its application to larger biomolecular complexes. The same comparison for NMR is slightly less than one *Entity_Cys* entries per *PDB* entry, which is consistent with identical entities being present after the identity cutoffs are applied using the RCSB REST API, as mentioned above for the case of Lac Repressor.

The schema support for partial occupancy atoms substantially increases the number of cysteine disulfide bonds. In an earlier version of the schema, which supported only full occupancy atoms, there are around 3k *PDB* entries with no cysteine disulfide bonds. The current Cys.sqlite has around 2k such entries, which may be attributed to the distance between sulfur atoms being beyond the bond length cutoff, missing cysteine atoms, or errors in the information available from RCSB queries. Overall, there are ~6.6k *Cys_Cys* entries with either an *SG_occ* < 1.0 or an *alt_id* that is not null.

Quantum chemical comparisons for sulfur-sulfur bond lengths and DSE

The quantum chemical calculations of dipropyl disulfide reveal a clear dependence of the sulfur-sulfur bond distance on the χ_3 torsion angle (Fig. 4.A). The plot of sulfur-sulfur distance follows the functional form of the χ_3 component of Eq. 1. As described in the methods, configurations from X-ray diffraction are generated for 22.5 degree bins of the torsional angles, which are constrained during geometry optimization. The χ_3 angle and bond distance of the configuration with lowest energy on this constrained potential energy

surface are −90 degrees and 0.204 nm, respectively. Bond distances between 0.202 to 0.205 nm are observed in the low-energy regions of χ_3 (columns of purple filled circles in Fig. 4.A); the bond distance lengthens to around 0.208 to 0.211 nm in high-energy regions (around $\chi_3 = 0$ and $\chi_3 = \pm 180$ degrees).

A sulfur-sulfur dependence on χ_3 is observed, only, in high-resolution X-ray structures (Fig. 4.A) in the regions of $|\chi_3| = 90 \pm 22.5$ that contain six, well-sampled, bins. The gas-phase DF-MP2 conformations correspond to 0 K; thus, the longer bond-length observed for X-ray is reasonable. In both low energy regions ($\chi_3 = \pm 90$), the mean bond-length is around 0.205 nm and increases with χ_3 deviations. The X-ray points with large errors contain too few PDBs to draw robust conclusions; for example, there are only two PDB entries, 5P4G and 5P4N, from a single high-throughput crystallography study[70] in the point at −135 degrees (Fig. 4.A).

The χ_3 dependence of the sulfur-sulfur distance will not be successfully modeled with a simple bond-stretching molecular mechanics energy function, particularly in the regions of ±180 degrees where nonbonded repulsions will be reduced. Structural models that depend more strongly on such molecular mechanics potentials, such as those from solution NMR, are expected to be somewhat flat around the parameterized sulfur-sulfur bond length (0.2029 in CHARMM[71] and 0.2038 in AMBER[62]). Flattened bond lengths are observed clearly for NMR, EM, and low-resolution X-ray structures (Fig. 4A.)

The agreement between the DSE model (Eq. 1) and the quantum chemical calculations of dipropyl disulfide energetics is promising (Fig. 4.B). The majority of DSE values fall along the diagonal with a spread of 5–10 kJ/mol (See SI for comparisons of additional models). Of the 573 configurations, there are around 30 outliers that are all predicted to be more highly strained by DF-MP2. Generally, these conformers contain repulsive interactions that are ignored by Eq. 1, such as eclipsed 1–4 groups or other nonbonded repulsions, see inline figure in Fig. 4B. Such repulsions must be compensated within the context of the given protein, i.e. transforming the gas-phase dipropyl disulfide back into cysteine disulfide within the protein environment. The DSE modeled with Eq. 1 is sufficiently accurate to identify general trends between DSE and experimental methods in the present study.

Disulfide strain energy increases as X-ray resolution decreases

The resolution (reported for X-ray and EM) is a measure of the quality of the experimental data. Lower numerical values of resolution correspond to more distinguishable structural features, such as sulfur-sulfur bonds. The disulfide strain energy (DSE, Eq. 1) increases significantly as the quality of the structure decreases (Fig. 5). For X-ray cysteine disulfides, the average DSE increases from around 10.0 $kJ \cdot mol^{-1}$ ±0.2 $kJ \cdot mol^{-1}$ for the ultra high-resolution structures (≤0.10 nm) up to around 11.4 $kJ \cdot mol^{-1}$ ±0.4 $kJ \cdot mol^{-1}$ for structures with good resolution ($0.18 < Res \leq 0.20$ nm). The overall trend is similar for that calculated with a sequence identity cutoff of 50 %. However, for the lowest resolution structures (>0.35 nm), the 50 % set has a lower numerical mean resolution (0.4 nm) and higher mean DSE (22.7 $kJ \cdot mol^{-1}$ ±0.6 $kJ \cdot mol^{-1}$) compared to overall (18.9 $kJ \cdot mol^{-1}$ ±0.1 $kJ \cdot mol^{-1}$)(Fig. 5); This observation is consistent with reduced DSE due to the utilization of redundant information during structural refinement. The EM mean and standard deviation agree well

with the trend extrapolated from the X-ray structures for the NONE set. The mean resolution and DSE are 0.57 nm and 20.6 kJ·mol^{-1} ±0.1 kJ·mol^{-1}, respectively.

The increase in DSE with decreasing structural quality is associated with modeling limitations. Decreasing experimental information effectively increases the parameter space available to the molecular model; disulfide strain increases, on average, as a consequence of reducing error elsewhere within the model. Following this logic, the NMR structures have the largest influence of modeling compared to EM or X-ray. The mean DSE for the NMR structures is significantly higher than that of EM or X-ray (Fig. 5). However, the difference in the experimental temperature must be considered. X-ray and EM structures are usually solved at cryogenic temperatures (~100 K). Cryogenic cooling affects side-chain packing and eliminates packing defects.[72] NMR data is collected for biological molecules in solution at room temperature, which may be expected to increase the average DSE, as is observed here, and broaden coordinate distributions as is not observed in the present study; this will be discussed further below with respect the distributions of DSE and internal coordinates.

In the analysis above, there was no filter used to separate the results by type of disulfide (see Fig. 1). For X-ray structures with resolution ≤ 0.2 nm, the DSE increases significantly as the disulfide type goes from interchain to intrachain to being vicinal. There are 10478 high resolution structures (≤ 0.15 nm), the average DSE being 10.6 kJ·mol^{-1} ± 0.1 kJ·mol^{-1}. In that set, there are 29 vicinal, 10036 intrachain, and 413 interchain disulfides with average DSE being 13.2 kJ·mol^{-1} ± 0.7 kJ·mol^{-1}, 10.7 kJ·mol^{-1} ± 0.1 kJ·mol^{-1}, and 7.6 kJ·mol^{-1} ± 0.2 kJ·mol^{-1}, respectively. As the resolution is reduced (> 0.2 nm), the interchain disulfide DSE grows to be larger than the intrachain DSE. Between 0.20 nm and 0.22 nm there are 70038 and 465 intrachain and interchain disulfides with the average DSE being 11.4 kJ·mol^{-1} ± 0.0 kJ·mol^{-1} and 14.7 kJ·mol^{-1} ± 0.3 kJ·mol^{-1}, respectively, and the interchain average DSE remains larger when all resolutions > 0.20 nm are included.

As described in detail above, the DSE increases significantly as the quality of experimental information degrades. It may follow that partial occupancy cysteine disulfide bonds would have higher values DSE. However, partial occupancy cysteine disulfides may be associated with the role of X-ray induced radiation damage of cysteine disulfides.[44–46] Adding a single electron to the disulfide will change the potential energy surface of the disulfide and the accuracy of the DSE model Eq. 1 would no longer be sufficient. Modeling the effects of photoelectron reduction would be required and is beyond the scope of the present study. The following are presented with these limitations in mind. For X-ray structures with resolution ≤ 0.20 nm, there are 12 vicinal, 3901 intrachain, and 189 interchain disulfides modeled with partial occupancies; the average DSE are 13.1 kJ·mol^{-1} ± 0.5kJ·mol^{-1}, 16.0 kJ·mol^{-1} ± 0.2kJ·mol^{-1}, and 19.5 kJ·mol^{-1} ± 0.9kJ·mol^{-1}, respectively. The partial occupancy vicinal disulfides appear less strained, although the number is small while the other two types are more strained according to Eq. 1.

Distribution of disulfide strain energy and dihedral angles from Cys.Sqlite

The probability distributions of DSE skew to higher values with longer tails for EM and NMR compared to X-ray (Fig. 6). Within the set of configurations for X-ray, the distributions skew to higher energy configurations as the value of the experimental

resolution increases (Fig. 6). For high-resolution structures (≤ 0.15 nm) the distribution peaks at around 7 kJ·mol^{-1} and falls to near zero by 30 kJ·mol^{-1}. In contrast, the distribution for low-resolution structures (> 0.28 nm) peaks around 13 kJ·mol^{-1} and continues to have nonzero values to 50 kJ·mol^{-1}. While the number of EM configurations is limited when compared to X-ray or NMR, the distribution is more aligned with NMR with a peak around 15 kJ·mol^{-1} and a relatively large population of configurations with DSE > 30 kJ·mol^{-1}.

Generally, the minima and maxima of the disulfide dihedral angle distributions are in good agreement with those expected from the corresponding energy functions for both NMR and X-ray structures (Fig. 6). The positive and negative dihedrals are not symmetrically sampled. Most importantly, the NMR and EM distributions for rotations about the S-S bond (χ_3) have significant populations in the region that reflects the trans configuration of the disulfide (100° to 180° ; −180° to −100°); NMR also has larger populations of disulfides 50 to 100 compared to X-ray and EM. For χ_1, the cysteine sidechain dihedral, NMR has significant population in the −100° to −150° region that is forbidden similarly for X-ray and EM. For EM, most structures have been deposited much more recently, as described above, and the agreement for the χ_1 angle may reflect the influence of side-chain rotamer libraries on the EM models.

Distributions of internal coordinates

There are significant differences between the probability distributions of some internal coordinates of the cysteine and cysteine disulfide for NMR compared to X-ray or EM (Fig. 7). The ω torsion angle of the peptide backbone is strongly peaked around 180° for all methods. Resonance with the carbonyl introduces double-bond character to the peptide bond between the N and C atoms; the barrier for a cis/trans rotation is around 84 kJ·mol^{-1}, and the cis conformer, being around 17 kJ·mol^{-1} higher in energy, is rarely observed.[73] The NMR distribution is significantly sharper; taking the distribution from the high X-ray as the true distribution, the NMR modeling restraints for that coordinate may be too strong. The same holds true for the angle between the backbone to sidechain angle (CA CB SG) and the S-S bond length (Fig. 7), which both have multiple, sharp peaks. The multiple peaks for the S-S bond agree well with the bond-length parameters for the AMBER (0.2038 nm) and CHARMM (0.2029 nm) forcefields.

The distribution of CA_CA distances is a measure of the stretch of the cysteine disulfide. There is a small peak in the region between 0.35 and 0.45 nm for all methods. For NMR and EM, this peak has higher populations near 0.375 nm than X-ray (Fig. 7) reflecting a larger number of disulfides that are more compressed. At the higher end of the distribution (~ 0.7 nm), there is a larger population of stretched disulfides for NMR, EM, and low resolution X-ray. For high-resolution X-ray structures, the largest peaks in the distribution are around 0.50 and 0.575 nm. These peaks shift out to larger values for low resolution structures. If the thermal motion and solution conditions reflected in NMR experiments were able to lead to the broader distributions of CA distances, the internal coordinates described above would also be expected to have broader distributions.

Ramachandran density maps

Ramachandran densities characterize the distribution of observed backbone conformations. The present study uses ADKE to model the true distributions for reduced cysteine and cysteine disulfide (Fig. 8) for high-resolution (<0.15 nm) X-ray structures with 50% identity cutoff. All cysteine Ramachandran densities have a strong peak around ϕ, ψ of −50,−50 in the α-helix region ($-180° \leq \phi \leq -30°$; $-120° \leq \psi \leq 30°$). Compared to reduced cysteine, the disulfide bond distribution has a larger population in the β_P$_{II}$ region ($-180° \leq \phi \leq -30°$; $30° \leq \psi \leq 180°$);the P$_{II}$ region is an oval oriented to the upper right of the β_P$_{II}$ region and slanted downward with increasing ϕ.[74] The ratio of integrated density (α-helix / β_P$_{II}$) decreases from 0.77 for the reduced cysteine distribution to 0.60 for cysteine disulfide distribution.

The Ramachandran density for cysteine disulfide is a mixture of low and high DSE conformations (see, Fig. 6). There is a clear dependence of the distribution on the DSE; four gradations of DSE are plotted in Fig. 9. For the lowest energy gradation (0 to 5.3 kJ·mol^{-1}) there is a significant population in the β_P$_{II}$ region with a ratio of integrated density (α-helix / β_P$_{II}$) of 0.45. That ratio increases dramatically to 0.83 for the next gradation (5.3 to 10.6 kJ·mol^{-1}) before dropping again to 0.52 (10.6 to 15.9) and finally to 0.40 for the most strained disulfides (highest DSE region). For strained disulfides (bottom two panels of Fig. 9) a strong peak in the density forms around [ϕ, ψ] of [−125,150] in the β_P$_{II}$ with associated decreases in the P$_{II}$ region. Considering these population shifts, with disulfide strain, for high resolution X-ray structures, there is a clear correlation of strained disulfides with the β-sheet secondary structure element. The population shifts from both the P$_{II}$ and α-helix regions to the β-sheet region reflect the central importance of the strong nonbonded interactions within β sheets in facilitating cross-strand disulfide redox switches.[21,22,34,35]

The shifts in Ramachandran densities is largely absent for NMR structures. Overall, the maps for NMR cysteine disulfide bonds contain the same strong peak in the α-helix region but are much more di use (see SI). There is more density in the β_P$_{II}$ region than the α-helix region; the ratios of integrated density (α-helix / β_P$_{II}$) for the four increasing DSE gradations are 0.46, 0.47, 0.42, and 0.46. There are no distinguishable changes in the features, as are observed for the corresponding X-ray distributions (see SI).

Concluding Remarks

In the present study, we used the structure and energetics of the cysteine disulfide bonds from the entire Protein Databank to compare structures within and across experimental methods. The simple approximation for disulfide strain energy (Eq. 1) reveals that cysteine disulfides appear to become more strained as the structural information degrades for X-ray structures. The strained populations were much larger for EM and NMR than X-ray. The effects of modeling in NMR are much clearer than in X-ray structures where some internal coordinates (e.g. the ω and CA_CB_S angles) appear overly restrained. These observations are independent of sequence identity cutoffs.

Gas-phase quantum chemical calculations reveal a clear dependence of the sulfur-sulfur bond length on the χ_3 torsion angle. This dependence is only observed in high-resolution X-

ray structures in low DSE regions. For high-resolution X-ray structures, there is a striking DSE dependence of the cysteine disulfide Ramachandran density that is analogous to dependencies associated with side-chain rotamer regions used in the development of rotamer libraries. There is a clear opportunity to improve the modeling of cysteine disulfides in low-resolution and NMR structures.

To our knowledge this is the first attempt to use a highly structured single file to make a slice of the vast information contained in the Protein Databank. The unique structural characteristics of cysteine are well-represented by the database schema presented here. Cys.sqlite provides cysteine-centric information with comprehensive coverage for proteins with one or more disulfide bonds in the Protein Databank. The presence of cysteine disulfides is not a requirement; the hierarchy of the schema establishes a parent-child relationship between the cysteine and the cysteine disulfide, respectively. The schema developed here may provide guidance to developing schema for cofactors such as metal ions, hydrophobic packing, and salt bridges.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Kenneth Kroenlein and Chris Muzny for helpful suggestions that improved the Cys.sqlite schema.

References

(1). Jones DP; Go YM Mapping the cysteine proteome: Analysis of redox-sensing thiols. Curr. Opin. Chem. Biol 2011, 15, 103–112. [PubMed: 21216657]

(2). Waldron KJ; Rutherford JC; Ford D; Robinson NJ Metalloproteins and metal sensing. Nature 2009, 460, 823–830. [PubMed: 19675642]

(3). Wedemeyer WJ; Welker E; Narayan M; Scheraga HA Disulfide Bonds and Protein Folding. Biochemistry 2000, 39, 4207–4216. [PubMed: 10757967]

(4). Derman AI; Prinz WA; Belin D; Beckwith J Mutations that allow disulfide bond formation in the cytoplasm of Escherichia coli. Science 1993, 262, 1744–1748. [PubMed: 8259521]

(5). Loferer H; Hennecke H Protein disulphide oxidoreductases in bacteria. Trends Biochem. Sci 1994, 19, 169–171. [PubMed: 8016867]

(6). Raina S; Missiakas D Making and breaking disulfide bonds. Annu. Rev. Microbiol 1997, 51, 179–202. [PubMed: 9343348]

(7). Stewart EJ; Åslund F; Beckwith J Disulfide bond formation in the Escherichia coli cytoplasm: an in vivo role reversal for the thioredoxins. EMBO J. 1998, 17, 5543–5550. [PubMed: 9755155]

(8). Szajewski RP; Whitesides GM Rate constants and equilibrium constants for thiol-disulfide interchange reactions involving oxidized glutathione. J. Am. Chem. Soc 1980, 102, 2011–2026.

(9). Rothwarf DM; Scheraga HA Equilibrium and kinetic constants for the thiol-disulfide interchange reaction between glutathione and dithiothreitol. Proc. Natl. Acad. Sci. U.S.A 1992, 89, 7944–7948. [PubMed: 1518818]

(10). Lees WJ; Whitesides GM Equilibrium constants for thiol-disulfide interchange reactions: a coherent, corrected set. J. Org. Chem 1993, 58, 642–647.

(11). Fernandes PA; Ramos MJ Theoretical insights into the mechanism for thiol/disulfide exchange. Chem. Eur. J 2004, 10, 257–266. [PubMed: 14695571]

(12). Thornton J Disulphide bridges in globular proteins. J. Mol. Biol 1981, 151, 261–287. [PubMed: 7338898]

(13). Richardson JS In The Anatomy and Taxonomy of Protein Structure; Anfinsen C, Edsall JT, Richards FM, Eds.; Advances in Protein Chemistry; Academic Press, 1981; Vol. 34; pp 167–339.

(14). Perry LJ; Wetzel R Disulfide bond engineered into T4 lysozyme: stabilization of the protein toward thermal inactivation. Science 1984, 226, 555–558. [PubMed: 6387910]

(15). Katz BA; Kossiakoff A The crystallographically determined structures of atypical strained disulfides engineered into subtilisin. J. Biol. Chem 1986, 261, 15480–5. [PubMed: 3096989]

(16). Hazes B; Dijkstra BW Model building of disulfide bonds in proteins with known three-dimensional structure. Prot. Eng. Des. Sel 1988, 2, 119–125.

(17). Pjura PE; Matsumura M; Wozniak JA; Matthews BW Structure of a thermostable disulfide-bridge mutant of phage T4 lysozyme shows that an engineered cross-link in a flexible region does not increase the rigidity of the folded protein. Biochemistry 1990, 29, 2592–2598. [PubMed: 2334683]

(18). Wetzel R Harnessing disulfide bonds using protein engineering. Trends Biochem. Sci 1987, 12, 478–482.

(19). Sowdhamini R; Srinivasan N; Shoichet B; Santi DV; Ramakrishnan C; Balaram P Stereochemical modeling of disulfide bridges. Criteria for introduction into proteins by site-directed mutagenesis. Prot. Eng. Des. Sel 1989, 3, 95–103.

(20). Almeida AM; Li R; Gellman SH Parallel β-sheet secondary structure is stabilized and terminated by interstrand disulfide cross-linking. J. Am. Chem. Soc 2011, 134, 75–78. [PubMed: 22148521]

(21). Haworth NL; Wouters MA Between-strand disulfides: forbidden disulfides linking adjacent β-strands. RSC Adv. 2013, 3, 24680–24705.

(22). Haworth NL; Wouters MA Cross-strand disulfides in the non-hydrogen bonding site of antiparallel β-sheet (aCSDns): poised for biological switching. RSC Adv. 2015, 5, 86303–86321.

(23). Hogg PJ Disulfide bonds as switches for protein function. Trends Biochem. Sci 2003, 28, 210–4. [PubMed: 12713905]

(24). Schmidt B; Ho L; Hogg PJ Allosteric disulfide bonds. Biochemistry 2006, 45, 7429–33. [PubMed: 16768438]

(25). Schmidt B; Hogg PJ Search for allosteric disulfide bonds in NMR structures. BMC Struct. Biol 2007, 7, 49. [PubMed: 17640393]

(26). Zhou B; Baldus IB; Li W; Edwards SA; Gräter F Identification of allosteric disulfides from prestress analysis. Biophys. J 2014, 107, 672–681. [PubMed: 25099806]

(27). Pijning AE; Chiu J; Yeo RX; Wong JWH; Hogg PJ Identification of allosteric disulfides from labile bonds in X-ray structures. Royal Soc. Open Sci 2018, 5, 171058.

(28). Østergaard H; Henriksen A; Hansen FG; Winther JR Shedding light on disulfide bond formation: engineering a redox switch in green fluorescent protein. EMBO J. 2001, 20, 5853–5862. [PubMed: 11689426]

(29). Zheng M; Åslund F; Storz G Activation of the OxyR transcription factor by reversible disulfide bond formation. Science 1998, 279, 1718–1722. [PubMed: 9497290]

(30). Wiita AP; Ainavarapu SRK; Huang HH; Fernandez JM Force-dependent chemical kinetics of disulfide bond reduction observed with single-molecule techniques. Proc. Natl. Acad. Sci. U.S.A 2006, 103, 7222–7227. [PubMed: 16645035]

(31). Baldus IB; Gräter F Mechanical force can fine-tune redox potentials of disulfide bonds. Biophys. J 2012, 102, 622–629. [PubMed: 22325286]

(32). Góngora-Benítez M; Tulla-Puche J; Albericio F Multifaceted roles of disulfide bonds. Peptides as therapeutics. Chem. Rev 2013, 114, 901–926. [PubMed: 24446748]

(33). Hogg PJ Targeting allosteric disulphide bonds in cancer. Nat. Rev. Cancer 2013, 13, 425–431. [PubMed: 23660784]

(34). Wouters MA; Lau KK; Hogg PJ Cross-strand disulphides in cell entry proteins: poised to act. Bioessays 2004, 26, 73–79. [PubMed: 14696043]

(35). Wouters MA; Fan SW; Haworth NL Disulfides as redox switches: from molecular mechanisms to functional significance. Antioxid. Redox. Sign 2010, 12, 53–91.

(36). Humphrey W; Dalke A; Schulten K VMD: Visual molecular dynamics. J. Mol. Graph 1996, 14, 33–38. [PubMed: 8744570]

(37). Richardson JS; Videau LL; Williams CJ; Richardson DC Broad analysis of vicinal disulfides: Occurrences, conformations with cis or with trans peptides, and functional roles including sugar binding. J. Mol. Biol 2017, 429, 1321–1335. [PubMed: 28336403]

(38). Brünger AT; Adams PD; Clore GM; DeLano WL; Gros P; Grosse-Kunstleve RW; Jiang JS; Kuszewski J; Nilges M; Pannu NS; Read RJ; Rice LM; Simonson T; Warren GL Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D 1998, 54, 905–921. [PubMed: 9757107]

(39). Herrmann T; Güntert P; Wüthrich K Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J. Mol. Biol 2002, 319, 209–227. [PubMed: 12051947]

(40). Trabuco LG; Villa E; Mitra K; Frank J; Schulten K Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. Structure 2008, 16, 673–683. [PubMed: 18462672]

(41). Scheres SH A Bayesian view on cryo-EM structure determination. J. Mol. Biol 2012, 415, 406–418. [PubMed: 22100448]

(42). Ravelli RB; McSweeney SM The 'fingerprint'that X-rays can leave on structures. Structure 2000, 8, 315–328. [PubMed: 10745008]

(43). Weik M; Ravelli RB; Kryger G; McSweeney S; Raves ML; Harel M; Gros P; Silman I; Kroon J; Sussman JL Specific chemical and structural damage to proteins produced by synchrotron radiation. Proc. Natl. Acad. Sci. U.S.A 2000, 97, 623–628. [PubMed: 10639129]

(44). Petrova T; Ginell S; Mitschler A; Kim Y; Lunin VY; Joachimiak G; Cousido-Siah A; Hazemann I; Podjarny A; Lazarski K; Joachimiak A X-ray-induced deterioration of disulfide bridges at atomic resolution. Acta Crystallogr. D 2010, 66, 1075–1091. [PubMed: 20944241]

(45). Sutton KA; Black PJ; Mercer KR; Garman EF; Owen RL; Snell EH; Bernhard WA Insights into the mechanism of X-ray-induced disulfide-bond cleavage in lysozyme crystals based on EPR, optical absorption and X-ray diffraction studies. Acta Crystallogr. D 2013, 69, 2381–2394. [PubMed: 24311579]

(46). Gerstel M; Deane CM; Garman EF Identifying and quantifying radiation damage at the atomic level. J. Synchrotron Radiat. 2015, 22, 201–212. [PubMed: 25723922]

(47). Berman HM; Westbrook J; Feng Z; Gilliland G; Bhat TN; Weissig H; Shindyalov IN; Bourne PE The Protein Data Bank. Nucleic Acids Res. 2000, 28, 235–242. [PubMed: 10592235]

(48). Ramachandran G. t.; Sasisekharan V. Adv. Prot. Chem; Elsevier, 1968; Vol. 23; pp 283–437.

(49). Ponder JW; Richards FM Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol 1987, 193, 775–791. [PubMed: 2441069]

(50). Dunbrack RL; Karplus M Backbone-dependent rotamer library for proteins application to side-chain prediction. J. Mol. Biol 1993, 230, 543–574. [PubMed: 8464064]

(51). Dunbrack RL Jr; Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997, 6, 1661–1681. [PubMed: 9260279]

(52). Lovell SC; Word JM; Richardson JS; Richardson DC The penultimate rotamer library. Proteins 2000, 40, 389–408. [PubMed: 10861930]

(53). Dunbrack RL Jr Rotamer libraries in the 21st century. Current opinion in structural biology 2002, 12, 431–440. [PubMed: 12163064]

(54). Shapovalov MV; Dunbrack RL A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 2011, 19, 844–858. [PubMed: 21645855]

(55). Hintze BJ; Lewis SM; Richardson JS; Richardson DC Molprobity's ultimate rotamer-library distributions for model validation. Proteins 2016, 84, 1177–1189. [PubMed: 27018641]

(56). PDBx/mmCIF Dictionary Resources. http://mmcif.wwpdb.org, Last visited: January 29, 2019.

(57). wwPDB consortium, Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2019, 47, D520–D528. [PubMed: 30357364]

(58). Riccardi D; Parks JM; Johs A; Smith JC HackaMol: An Object-Oriented Modern Perl Library for Molecular Hacking on Multiple Scales. J. Chem. Inf. Model 2015, 55, 721–726. [PubMed: 25793330]

(59). CPAN https://metacpan.org. Last visited: January 29, 2019.

(60). SQLite https://www.sqlite.org. Last visited: January 29, 2019.

(61). Pyykkö P; Atsumi M Molecular Single-Bond Covalent Radii for Elements 1–118. Chem. Euro. J 2009, 15, 186–197.

(62). Weiner SJ; Kollman PA; Case DA; Singh UC; Ghio C; Alagona G; Profeta S; Weiner P A new force field for molecular mechanical simulation of nucleic acids and proteins. J. Am. Chem. Soc 1984, 106, 765–784.

(63). Zhu X; MacKerell AD Jr Polarizable empirical force field for sulfur-containing compounds based on the classical Drude oscillator model. J. Comput. Chem 2010, 31, 2330–2341. [PubMed: 20575015]

(64). Turney JM; Simmonett AC; Parrish RM; Hohenstein EG; Evangelista FA; Fermann JT; Mintz BJ; Burns LA; Wilke JJ; Abrams ML; Russ NJ; Leininger ML; Janssen CL; Seidl ET; Allen WD; Schaefer HF; King RA; Valeev EF; Sherrill CD; Crawford TD Psi4: an open-source ab initio electronic structure program. Wiley Interdiscip. Rev. Comput. Mol. Sci 2012, 2, 556–565.

(65). Parrish RM; Burns LA; Smith DGA; Simmonett AC; DePrince AE; Hohenstein EG; Bozkaya U; Sokolov AY; Di Remigio R; Richard RM; Gonthier JF; James AM; McAlexander HR; Kumar A; Saitow M; Wang X; Pritchard BP; Verma P; Schaefer HF; Patkowski K; King RA; Valeev EF; Evangelista FA; Turney JM; Crawford TD; Sherrill CD Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. J. Chem. Theory Comput 2017, 13, 3185–3197, [PubMed: 28489372]

(66). Dunning Thom H., J. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. J. Chem. Phys 1989, 90, 1007–1023.

(67). Kendall RA; Thom H Dunning J; Harrison R. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. J. Chem. Phys 1992, 96, 6796–6806.

(68). Dunning T Jr.; Peterson K; Wilson A Gaussian basis sets for use in correlated molecular calculations. X. The atoms aluminum through argon revisited. J Chem Phys 2001, 114, 9244–9253.

(69). Zhao G; Perilla JR; Yufenyuy EL; Meng X; Chen B; Ning J; Ahn J; Gronenborn AM; Schulten K; Aiken C; Zhang P Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. Nature 2013, 497, 643–646. [PubMed: 23719463]

(70). Schiebel J; Krimmer SG; Röwer K; Knörlein A; Wang X; Park AY; Stieler M; Ehrmann FR; Fu K; Radeva N; Krug M; Huschmann FU; Glöckner S; Weiss MS; Mueller U; Klebe G; Heine A High-throughput crystallography: reliable and efficient identification of fragment hits. Structure 2016, 24, 1398–1409. [PubMed: 27452405]

(71). Brooks BR; Brooks III CL; Mackerell AD; Nilsson L; Petrella RJ; Roux B; Won Y; Archontis G; Bartels C; Caflisch SBA; Caves L; Cui Q; Dinner AR; Feig M; Fischer S; Gao J; Hodoscek M; Im W; Kuczera K; Lazaridis T; Ma J; Ovchinnikov V; Paci E; Pastor RW; Post CB; Pu JZ; Schaefer M; Tidor B; Venable RM; Woodcock HL; Wu X; Yang W; York DM; Karplus M CHARMM: The biomolecular simulation program. J Comput Chem 2009, 30, 1545–1615. [PubMed: 19444816]

(72). Fraser JS; van den Bedem H; Samelson AJ; Lang PT; Holton JM; Echols N; Alber T Accessing protein conformational ensembles using room-temperature X-ray crystallography. Proc. Natl. Acad. Sci. U.S.A 2011, 108, 16247–16252. [PubMed: 21918110]

(73). Yonezawa Y; Nakata K; Sakakura K; Takada T; Nakamura H Intra-and inter-molecular interaction inducing pyramidalization on both sides of a proline dipeptide during isomerization: an ab initio QM/MM molecular dynamics simulation study in explicit water. J. Am. Chem. Soc 2009, 131, 4535–4540. [PubMed: 19267429]

(74). Hollingsworth SA; Karplus PA A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. Biomol. Concepts 2010, 1, 271–283. [PubMed: 21436958]

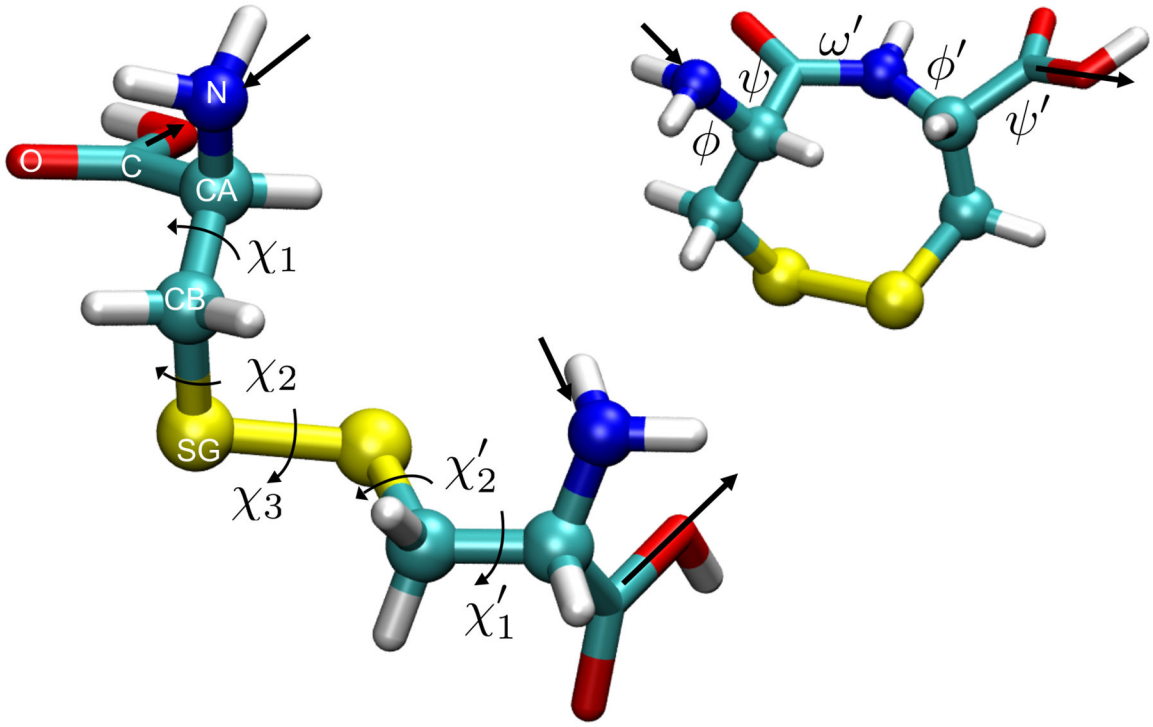

Figure 1:
VMD[36] licorice rendering of a disulfide between two cysteine amino acids. All atoms included in the dihedrals are rendered with overlayed CPK spheres. Cystine (left, SMILES:C(C(C(=O)O)N)SSCC(C(=O)O)N) involves two cysteines that may be on the same or different protein chains; the atom names and disulfide torsion angles are shown in white and black, respectively. Cyclocystine (right, SMILES:C1C(C(=O)NC(CSS1)C(=O)O)N) corresponds to a vicinal disulfide bond[37]; the backbone torsion angles are shown for reference. The arrows point along the direction of the polypeptide chain.

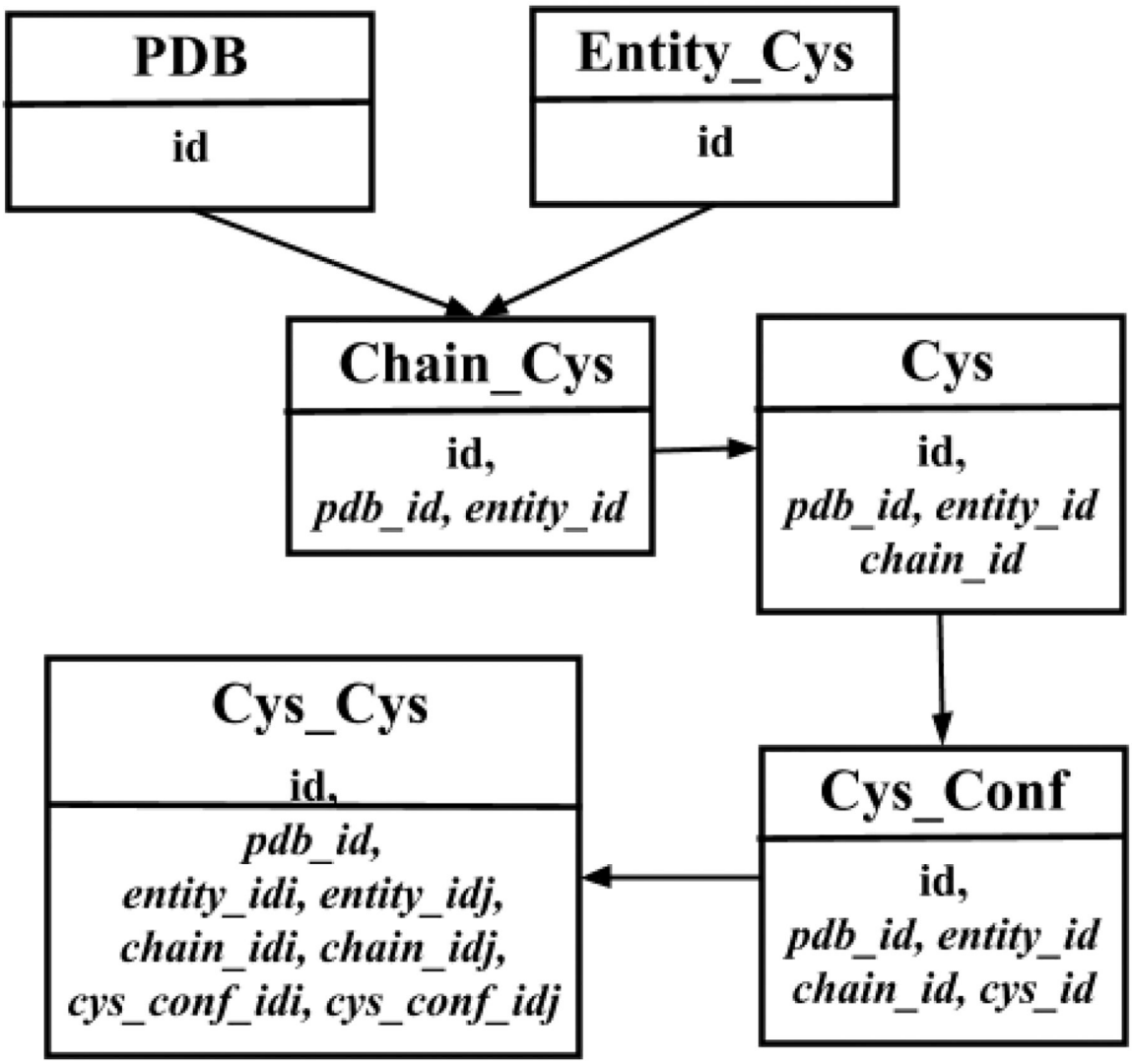

Figure 2:

Relationships between tables for the Cys.sqlite schema. The primary keys and foreign keys (italic) enable connections between tables. All foreign keys cascade down the intuitive hierarchy (arrow) except for those of the Cys table; the *Cys_Cys* table does not contain foreign keys for the Cys table. The *Cys_Cys* table contains a column (*alt_occ_flag*) to facilitate queries of whether the cysteine disulfide is associated partial occupancy residues. See SI for detailed tables and columns.

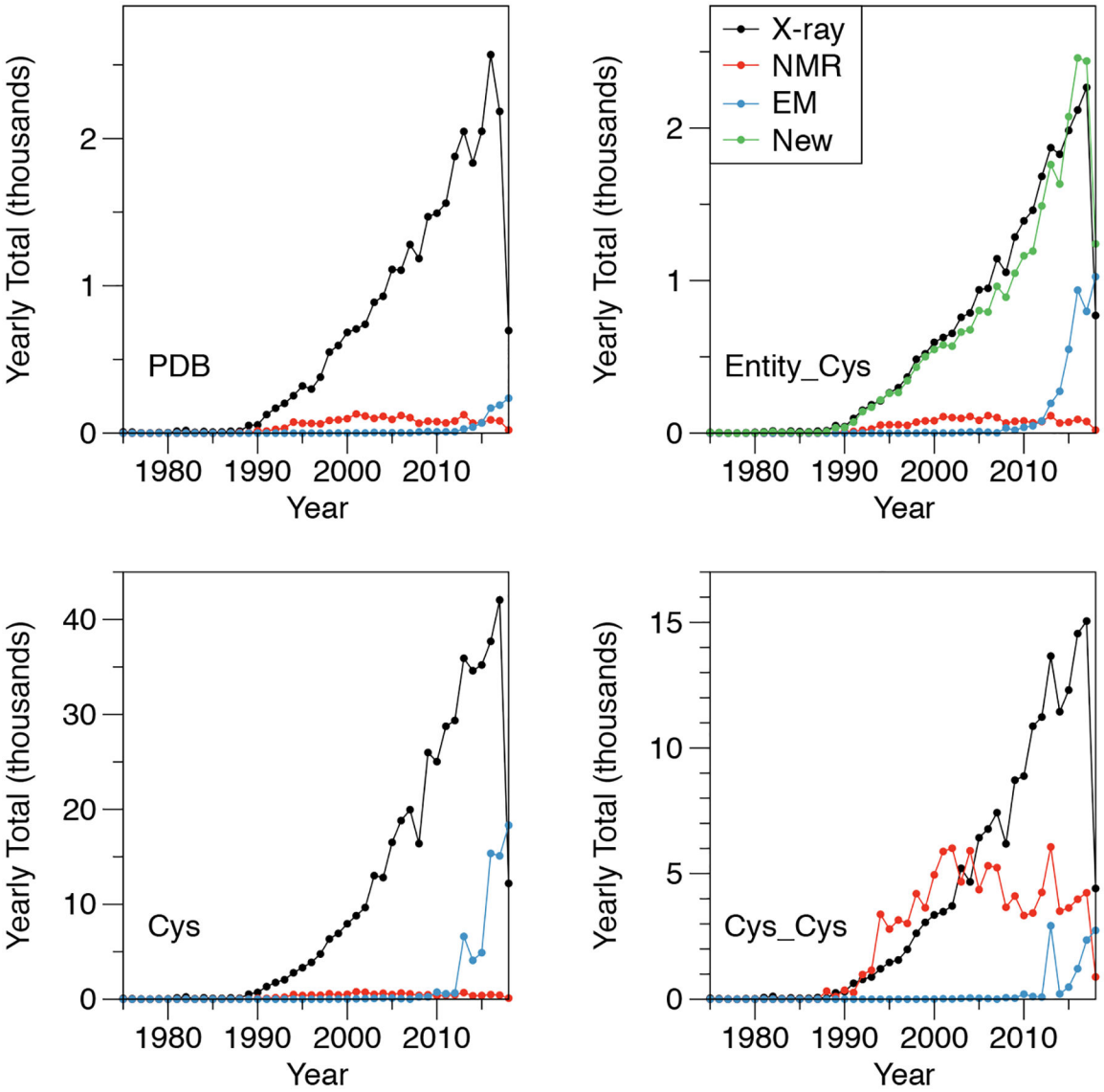

Figure 3:

Characterization of total annual entries, accumulated from the deposit date, for several Cys.sqlite tables. The upper right panel is the only to display the total number of new entries (entities) released each year (the sum of X-ray, NMR, and EM).

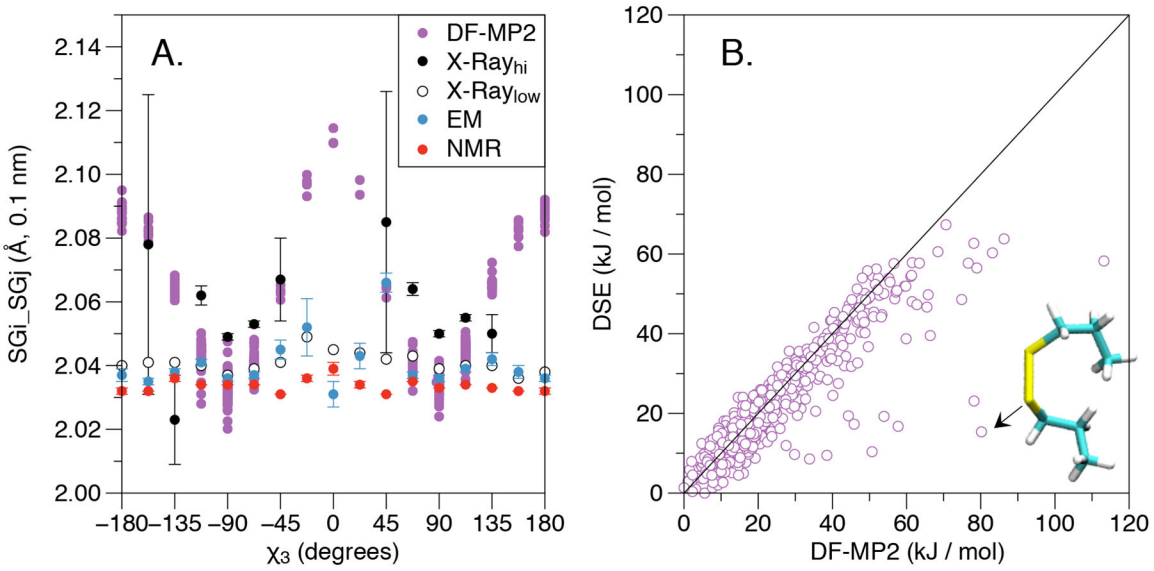

Figure 4:

Comparison with DF-MP2 quantum chemical calculations of dipropyl disulfide. A., The sulfur-sulfur bond distance is plotted as a function of the χ_3 torsion angle for all DF-MP2 geometries. Also plotted are the mean and standard error of the mean for all Cys.sqlite cysteine disulfide bond configurations for X-Ray (high and low resolution ≤ 0.15 and > 0.28 nm, respectively), EM, and NMR. B., The disulfide strain energy (DSE) calculated with the simple model (Eq. 1) is compared to energies determined using DF-MP2 for each conformation (see Fig. S2), both being calculated relative to the minimum energy configuration. A line of unit slope (black line) represents perfect agreement. The configuration of the outlier (black arrow) has χ_1, χ_2, χ_3, χ'_2,χ'_1 torsion angles of 90,90,−90,67.5,−45.0, respectively and a carbon-carbon distance of 0.33 nm between the nearest nonbonded CH_3 and CH_2 groups.

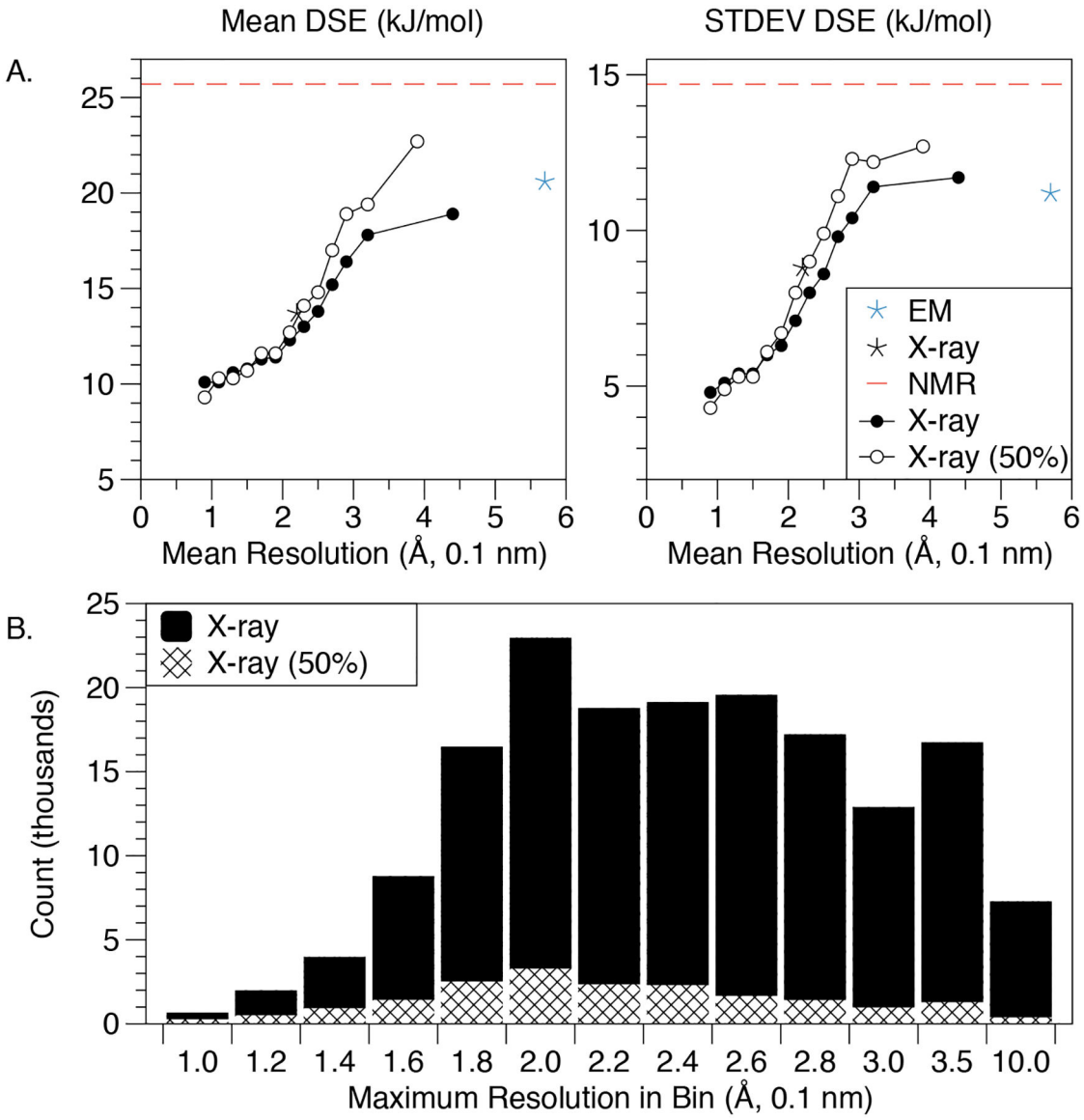

Figure 5:

Dependence of disulfide strain energy (DSE) on resolution. A. The overall mean (left) and standard deviation (right) are shown for NMR (dashed red), X-ray (black star), and EM (blue star) for all cysteine disulfides in Cys.sqlite (Table 1). The resolution (x-axis) of X-ray and EM points correspond to the average resolution for the set; NMR is plotted as a constant because the resolution does not apply. The X-ray DSE measures are plotted overall entries and also for the set with 50 % sequence identity cutoff. In B., The counts of the corresponding X-ray entries are plotted for each bin of resolution. The average (STDEV) resolution for all Cys.sqlite X-ray (black star) and EM (blue star) structures is 0.22 (0.06) and 0.57 (0.48) nm, respectively; the minimum (maximum) resolution is 0.05 (1.0) nm and 0.22 (5.0) nm, respectively.

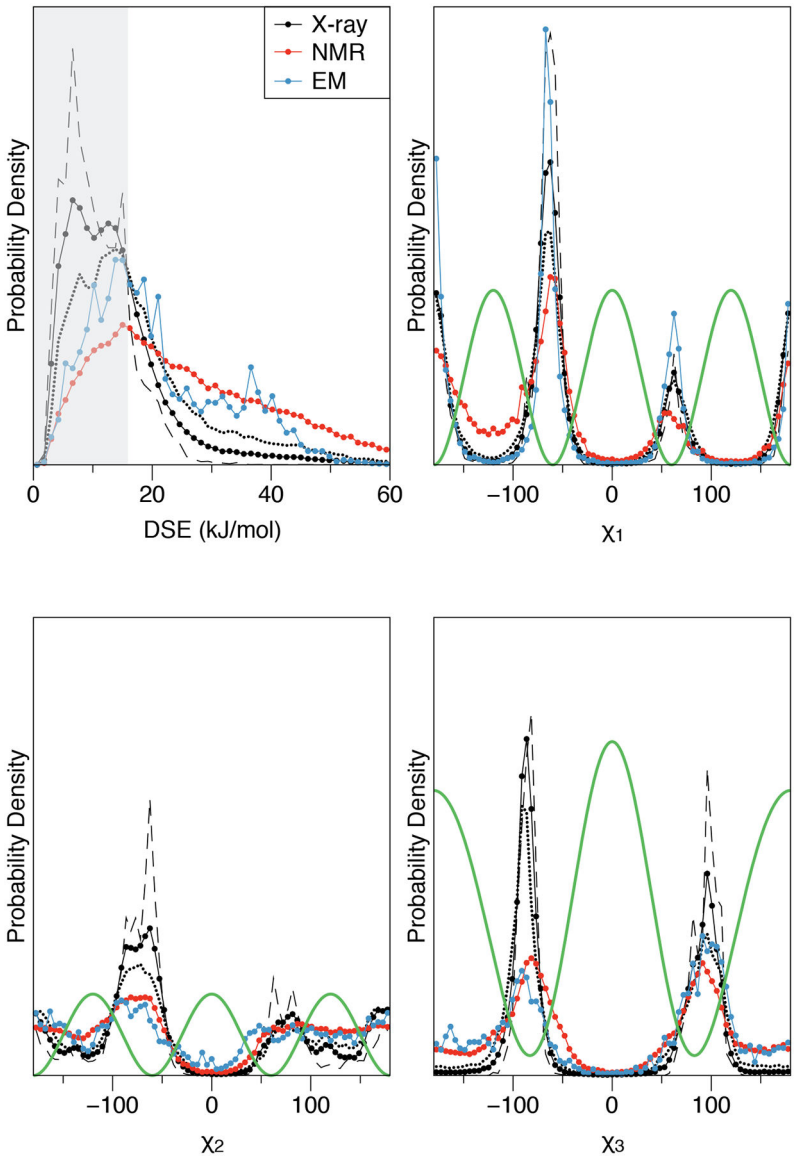

Figure 6:

Probability distributions of the disulfide strain energy (DSE) and disulfide dihedrals plotted for EM, NMR, and X-ray with no sequence identity cutoff. The X-ray is calculated overall (solid), high resolution (≤0.15 nm, dashed), and low resolution (>0.28 nm, dotted). The shaded region of the DSE distribution corresponds to the separation between unstrained and strained disulfides (15.9 kJ·mol^{-1}). The χ_1 and χ_2 angles are combined with the χ'_1 and χ'_2 angles on the same plot. For the three dihedrals (χ_1, χ_2, χ_3), the corresponding contributions to the energy is plotted (green), each scaled by the same amount to be plotted alongside the probability distributions.

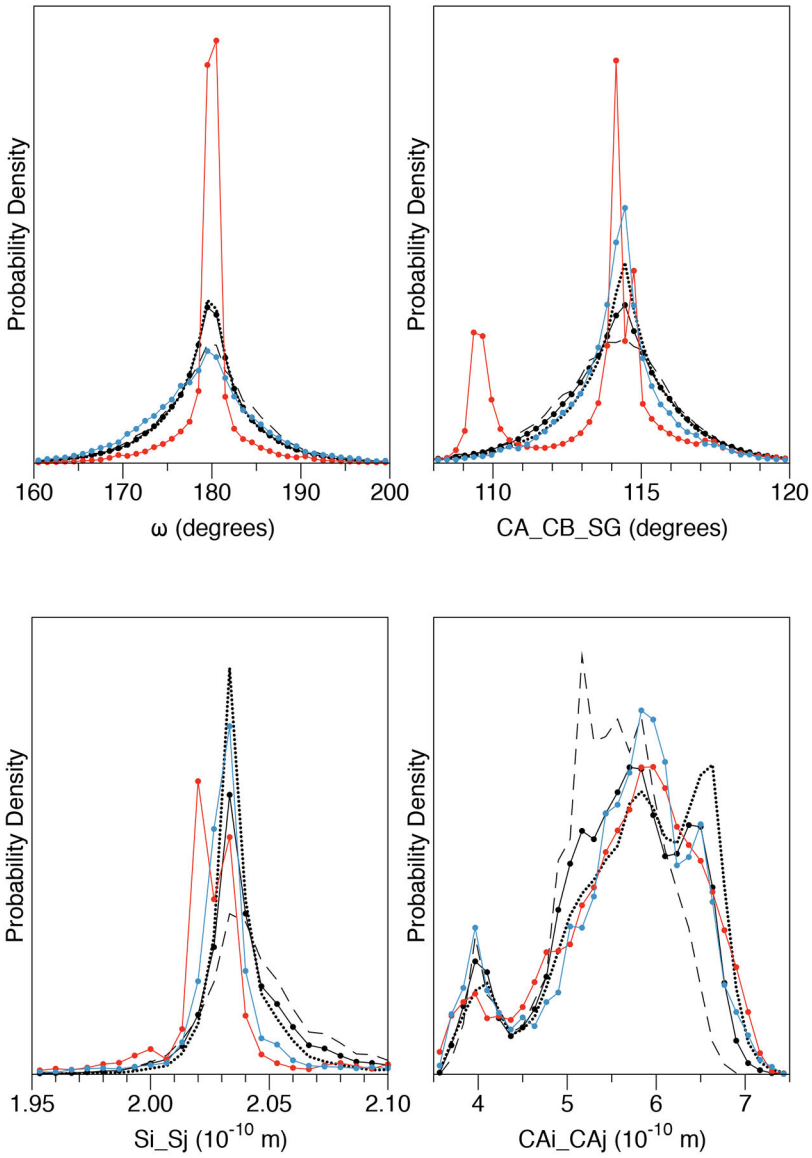

Figure 7:

Probability distributions of selected cysteine and cysteine disulfide internal coordinates for X-ray, NMR, and EM with no sequence identity cutoff. The X-ray is calculated overall (solid), high resolution (≤0.15 nm, dashed), and low resolution (>0.28 nm, dotted).

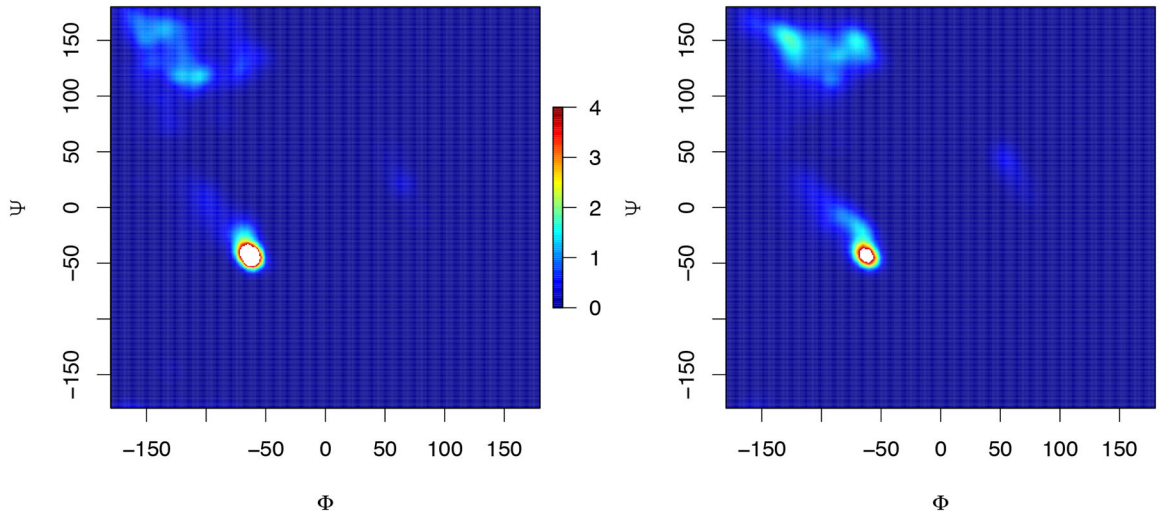

Figure 8:

AKDE models of the Ramachandran densities reduced cysteine (left) and cysteine disulfide (right) in high-resolution X-ray structures with 50% sequence identity cutoff. The ϕ, ψ region around −50,−50 (white) is higher than the ceiling of 4×10^{-4}. There are 884 and 4427 data-points associated with the left and right panels, respectively.

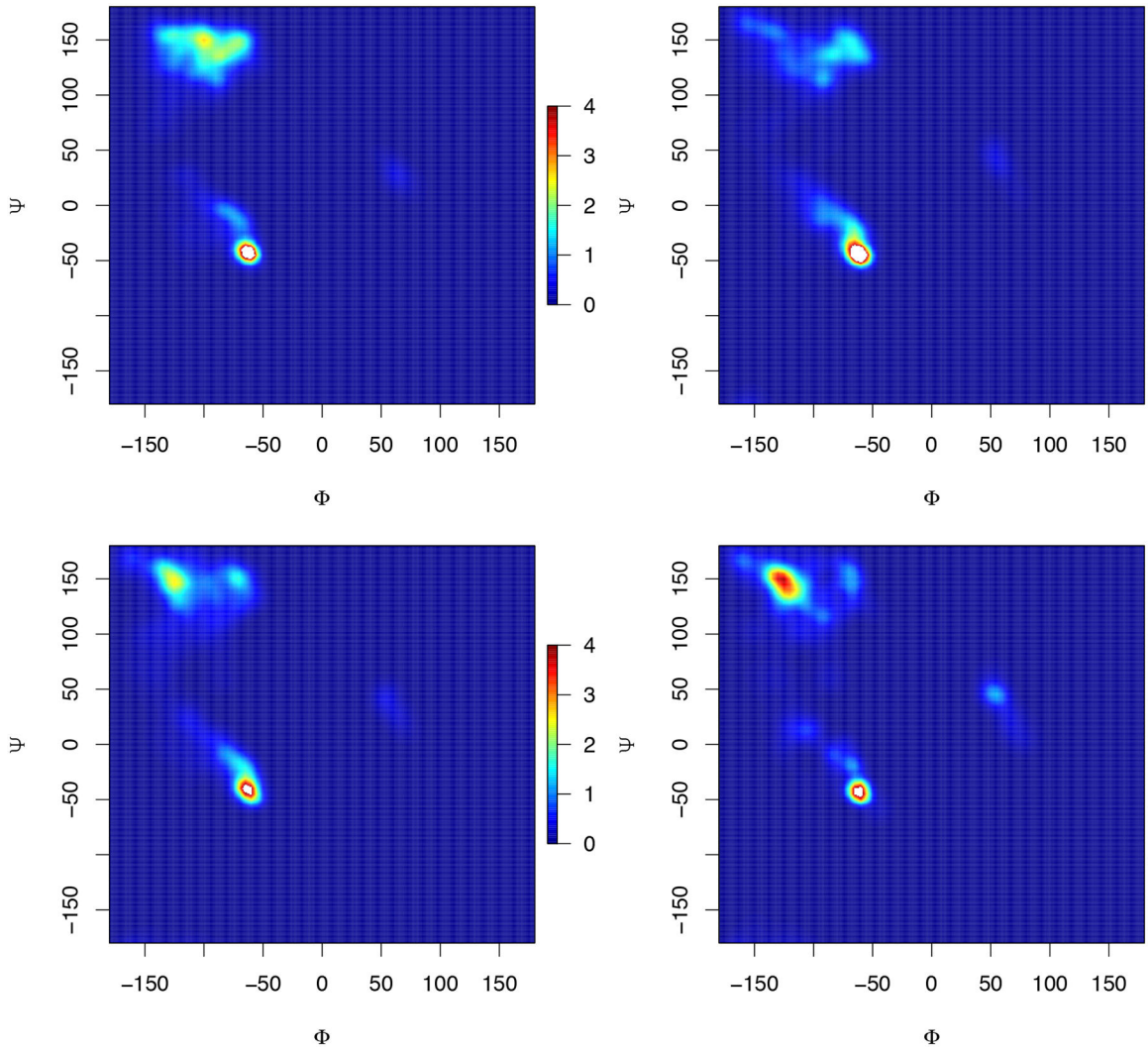

Figure 9:

AKDE models of the Ramachandran densities for four regions of DSE for high-resolution X-ray structures with 50% sequence identity cutoff. The DSE of the region increases from left to right and from top to bottom; the upper right panel corresponds to a lower energy region than the lower left panel. The overall number of data points for all plots is 4427, which consists of 716, 1865, 1238, and 608 data points in order of increase DSE region. See Methods for the definition of the regions. For all plots the ϕ, ψregion around −50,−50 (white) is higher than the ceiling of 4×10^{-4}. All plots are on the same scale ($\times10^{-4}$).

Table 1:

Summary of Cys.sqlite contents corresponding to current, released PDB entries (query run on 2018-12-12). The counts are determined by method and sequence identity cutoff. Obsolete entries (20) are not included; the two *Entity_Cys* entries with no current PDB entries are also ignored. For each value of the sequence identity cutoff (50 100 NONE), the count represents the total in that set. The *Cys_Cys* table includes both the total and unique disulfides, in parenthesis, for NMR. The *Entity_Cys* values are determined using a join on the *Chain_Cys* and *PDB* tables.

Table	exp_method	50	100	NONE
PDB	X-ray	5543	10454	29561
	NMR	1207	1631	2254
	EM	361	468	811
	All			32717
EntityCys	X-ray	6089	12896	24990
	NMR	1189	1624	2088
	EM	2022	2460	3342
	All			25566
ChainCys	X-ray	19467	32535	85489
	NMR	1287	1740	2558
	EM	7020	8125	14556
	All			102748
Cys	X-ray	83041	157561	466685
	NMR	6332	8852	12820
	EM	28234	34186	68004
	All			548513
CysConf	X-ray	85048	161100	476796
	NMR	116926	160491	224266
	EM	28313	34284	68102
	All			771161
CysCys	X-ray	20449	49959	173724
	NMR	55060(2987)	76303(4215)	106758(6107)
	EM	2445	3711	10712
	All			291847