**Methods for Hydrogen Parameter Revision**

Quantum calculations

Quantum mechanical methods for calculating electron density offer an independent, theory-based method to estimate how much the electron cloud is perturbed around each hydrogen. We have followed out that direction, to relate calculated electron density with observed difference density and with possible atomic interactions where electron clouds meet between atoms. Density for each amino acid was calculated with several quantum-chemical methods and basis sets (HF/3-21G, HF/6-31G(d,p), B3LYP/6-31G(d,p), and MP2/6-31G(d,p)) using the 64-bit Intel version of GAMESS [1], with grid spacings of 0.2A or 0.1Å for electron density output. An alanine dipeptide was also included, to obtain data for the backbone NH. Each molecule was geometry-minimized by each of the QM methods and basis sets. Accuracy and consistency of the QM method and basis set combinations was assessed by consistency and reasonableness in fitting spherical patches to the calculated density contours in the procedure described below. It was determined that the closer grid spacing of 0.1Å was required to lower noise in the results. Furthermore, neither of the HF calculations provided a consistency of results when compared to the B3LYP and MP2 calculations even with the same basis set. The B3LYP and MP2 results were similar, and B3LYP was chosen for the major work.

The effect of the Polarizable Continuum Model (PCM) on the bond lengths and electron density was determined to be negligible. However, because the protein is always interacting with itself and other molecules, the influence of polar and nonpolar molecules on the x-H bond lengths was investigated at the MP2 level. The geometries of a water and of a methane molecule were optimized while fixing the O or C atom at a range of distances from the Hδ22 hydrogen on the side-chain N of asparagine or Hγ1 on the side-chain O of threonine. Geometrical analysis of the electron density is hampered by the fact that H and solvent densities merge at most contour levels, but it can be determined that the electron-cloud center is further out when a solvent atom is nearby, compared with vacuum, or compared with the Hδ21 (away from the solvent molecule) in the same calculation.

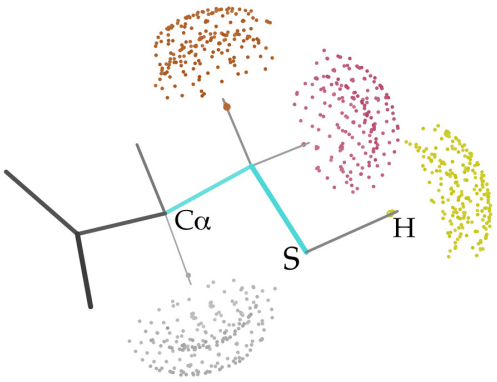
Sphere-fitting calculations

The resulting electron density was analyzed using a geometrical algorithm to fit partial spheres to a range of contour levels, looking for consistency of those sphere centers to identify the effective center of the H electron cloud. Treating the electron-cloud hydrogens as spherical on their external surfaces is of course an approximation, but a fairly good one, and more complex models are not currently feasible. A second approximation, which proves quite accurate for isolated amino acids, is that the sphere center lies along the line defined by the x-H covalent bond between the nuclei. Therefore, we only need to determine the best estimate of the sphere-center distance out along that line.

The surface of the electron density for an amino acid has a complex, multi-lobed shape. We need to fit one spherical patch (at each contour level) to represent the external shape of each hydrogen atom, a problem that differs from

those usually treated: for instance, standard sphere fitting techniques like RANSAC [2] and coresets [3] work reliably when one needs to fit just one sphere to an entire shape. Therefore, we tackle the problem of fitting multiple spheres by dividing it into two parts: 1) segmenting the contour surface into regions each of which individually should contain just one spherical patch; and 2) fitting one sphere to each such patch.

One variant of the first problem of segmenting molecular surfaces has been considered by Natarajan et al. [4], where they segment the surface into protrusions and cavities using a mean-curvature-like function. However, two of our atoms (such as an Hβ1-Hβ2 pair) may be joined as a single ellipsoidal protrusion by that method, since curvatures in orthogonal directions can cancel each other out across the shallow saddle point between the atoms. Therefore, instead of mean-curvature-like functions, we do the segmentation step using a function we call *sphericity*, related to curvature. Each vertex $v$ on the contour surface (which is computed using the marching cubes algorithm [5]) has a sphericity value of 1/radius of the sphere best fit to all vertices within 0.3Å of $v$; the value is negative if the sphere is outside the surface. Finding the thinnest fitting spherical shell is a non-convex problem [6] and hence known to be computationally expensive. However, the computationally cheaper alternatives do not perform well on the flat regions of the contour surface that occur at boundaries between spherical patches. Thus our definition of sphericity, although more expensive computationally, is suitable for the purpose of atomic segmentation because spherical patches are regions of high sphericity separated from each other in all directions by regions of low or negative sphericity. Thresholding on the value of sphericity divides the surface into the desired patches, but the threshold differs for different boundaries. This can be handled by the hierarchical process of constructing a join tree [7] whose nodes correspond to patches on the contour surface. Effectively a rising "water level" of sphericity value separates out new join-tree nodes where their patches are first divided into separate islands. At a level of the join tree with a number of nodes approximating the number of atoms that are end nodes in the covalent connectivity graph of the molecule, a user can easily identify the patches that correspond to the desired set of atoms, including one for each hydrogen. Figure S1 shows cysteine with fitted spherical patches (dot surfaces) at two of the contour levels for each of the atoms, with the sphere centers marked along the bond vector.

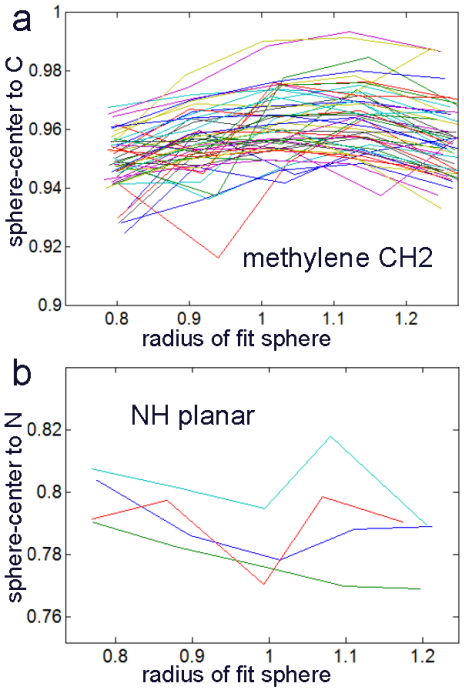


Cys: H atom sphere-fit to QM electron density

Having completed step 1, we are now left with multiple instances of fitting a sphere to a chosen patch at a given contour level: that is, computing a thin spherical shell that contains a large number of the vertices in that patch. We use several properties of this specific molecular system to help choose the right compromise between shell thinness and vertex coverage. a) For isolated amino acids, the convex portions of the electron density contours around each

protruding atom do not have points or dimples, but they are distorted to lower sphericity levels around the edges toward their boundaries with surrounding atoms; we use that fact to discard the lowest-sphericity half of the vertices in each potential patch. b) The patch corresponding to a hydrogen should have a fairly smooth boundary, so we discard vertices where the local boundary has an internal angle less than 120°. c) The patch should be connected, so if not, we choose the largest connected component. These rules allow us to shrink down to a well-behaved patch, and then fit a spherical shell that contains its vertices, using the lifting map technique [8]. We found our overall procedure insensitive to the parameter values involved (such as the 0.3Å radius for vertex neighbors).

We know approximately the radius (which translates here to an electron-density level) at which atom-atom contact effectively occurs: generously, the possible range is no wider than 0.7 to 1.3Å for hydrogens. Therefore, we analyze the sphere centers found for multiple contour levels across that range, by plotting the distance of those sphere-patch centers from the parent heavy atom as a function of sphere radius, expecting reasonable consensus across contour levels (sphere radii) and across examples of a specified atom type. Experimental scatter of both sorts was considerably reduced for quantum electron density values calculated on a 0.1Å grid spacing rather than 0.2Å, so 0.1Å was used in this work.




Examples of the resulting distance-vs-radius plots are shown in Figure S2a for methylene CH2 hydrogens (22 H atoms from 8 different amino acids) and Figure S2b for planar NH groups (4 H atom examples). The electron-cloud centers inferred from the near-flat regions in these plots are 0.96 ±0.01Å for tetrahedral CH2 and 0.79 ±0.01Å for planar NH.

Database analyses: nuclear x-H distances

To document nuclear x-H distances, we first revisited the classic neutron crystallography structures from the 1970's that cover 12 of the amino acids (e.g., [9-11]. Those included 50-80 measurements for N-H and C-H but only 4-5 values for O-H and S-H.

We then tabulated and plotted x-H distances for 78 relevant neutron structures in the Cambridge Structural Database (CSD; version 5.33 plus 4 updates; [12]), chosen as normal amino acids or nucleotides, or found by search with a fragment drawing such as C-S-H (see list of CSD codes below). Distances were measured in the Conquest viewer with atoms labeled and packing turned on, and were grouped into planar sp2 versus tetrahedral sp3 geometries, with the latter divided into CH1, CH2, or CH3 examples. For O-H, the neutron data included a good representation of water, phosphate, and carboxylate as well as Ser/Thr/Tyr/ribose examples, with quite distinct values, so those were each

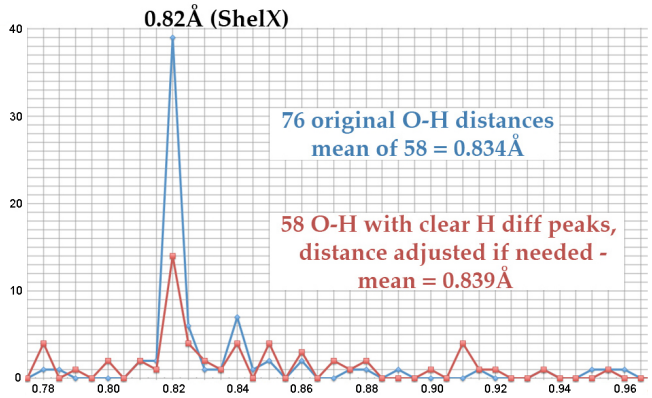
tabulated separately. In contrast, we found for a given geometry and parent atom that x-D and x-H were indistinguishable at the level of accuracy achievable (also found for electron diffraction) [13], so those were combined. However, we could identify only one additional S-H value even with an open search just requiring an S and an H atom.

Checking neutron structures in the wwPDB (worldwide Protein Data Bank; [14]), we found only three with a free cysteine (3KKX, 3KMF, 4G0C); all 8 S-H distances were 0.96 +/- .004Å; that reflects an unlikely target value in CNS presumably copied from C-H in the absence of a ShelX value for S-H, and we have not used those datapoints. We also compared the few but precise measurements from electron diffraction [13] and nuclear magnetic resonance [15], and methane C-H from *ab initio* calculation [16].

Database analyses: electron-cloud-center x-H distances

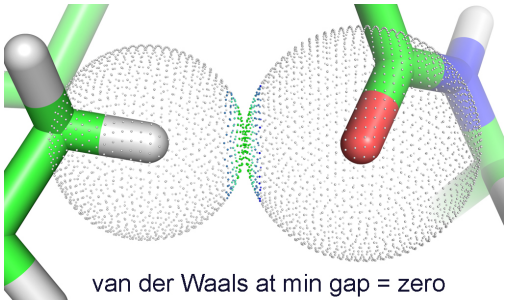
For electron-cloud C-H distances, the QM/sphere-fit calculated values closely matched the ShelX [17] values, but for polar x-H they were substantially shorter than in ShelX, thus requiring further study. N-H, O-H, and S-H distances were tabulated and plotted from 162 relevant X-ray structures in the CSD (see file list below). Examining distributions for these distance values showed many problems, such as apparent use of incorrect values (as above), overuse of exact "ideal" values rather than derived from the data, searching on 'x-ray' also returning neutron structures from joint refinement or just from the same paper, and so on. We have trimmed really extreme outliers on the distribution tails, but also undertook a new study where we could readily examine difference peaks for the H atoms.

The COD (Crystallography Open Database; [18]) was used for evaluating x-ray electron density for hydrogens, by scripting the open-source Olex2 viewer ([19]; http://www.olex2.org) to strip H atoms, calculate and display H difference density, and enable interactive adjustment of x-H distance if unambiguously needed. We examined 124 COD structures of amino acids, nucleic acids, and carbohydrates chosen by suitable Smiles strings, and tabulated N-H, O-H, S-H, and some CH distances. These fell into three cases: a) missing or poor H difference peak – omitted; b) clear H difference peak, fit fairly well by deposited H atom – kept; c) clear H difference peak, but not fit by deposited H atom – the x-H distance was adjusted. There is still some uncertainty in distinguishing a difference peak produced by an H atom from one produced from motion of the parent atom, from solvent issues, or from noise; however, this COD protocol produces many fewer obvious artifacts and is more robustly coupled to the experimental data. Most significantly for OH (Figure S3), the COD adjustments helped correct a clearly evident previous bias in favor of assigning the ShelX distance value. For these reasons, the adjusted COD values are emphasized in our overall compilation.




For another specific set of tests on electron-cloud distance

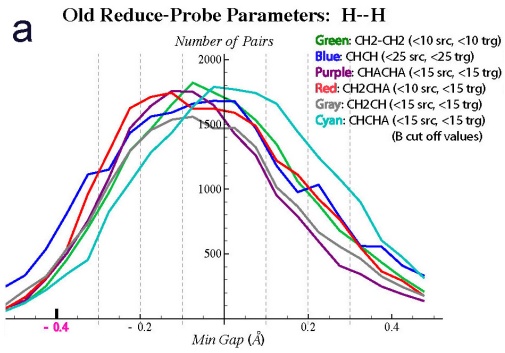
values, PDB x-ray structures at <1Å resolution with >150 residues were checked for occurrence frequency and appearance of H difference peaks. 12 files containing a total of >4000 residues were chosen for examination in Coot [20] and measurement from parent atom to H difference peak center (see file list below), for distinguishing C-H chemical types and for documenting polar x-H distances in H-bonding vs non-polar environments.

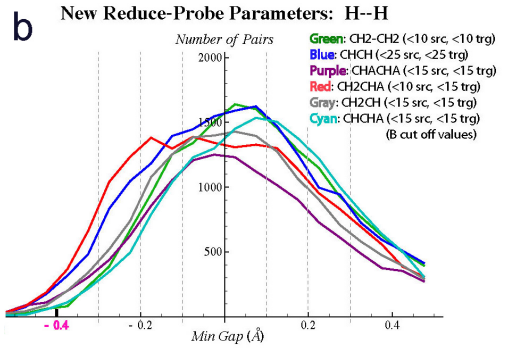

van der Waals at min gap = zero

Database analyses: van der Waals radii

For determination of updated van der Waals radii tuned for each new x-H distance set, we used the Top8000 dataset ([21]; available on GitHub) of non-redundant X-ray protein chains with MolProbity score <2.0 and resolution better than 2Å, and residue-level quality filters. Hydrogens were added and optimized with Reduce [22], including Asn/Gln/His flips where indicated. The one-dot-each modification of Probe [23], which outputs only a single contact dot to the nearest target atom, was used to calculate the smallest distance between the van der Waals spheres for a given pair of atoms – the "min gap". Figure S4 illustrates the just-touching, min-gap zero case. Looking only at the nearest neighbor rather than all neighbors, and using the OneDotEach function, eliminates issues with occluding or very distant atoms and produces much more interpretable distance distributions. Note that these "van der Waals radii" are meant to represent the preferred packing distance, not completely hard spheres, so some min gap values should be negative.

Runs were done three different ways: 1) using previous MolProbity nuclear values for both x-H and van der Waals, 2) using the previous ShelX/PHENIX x-H distances and MolProbity van der Waals lengthened by 0.05Å for H (but with 1.70Å for non-aromatic C), and 3) using QM/sphere-fit x-H and van der Waals as for run 2. Database tables in MySQL [24] were used to filter the data by crystallographic B-factor, and distributions of min gap value were produced in R [25]. By default, only atom pairs with source and target atom B-factors <10 were retained. If that resulted in <500 pairs, the cutoff was raised, but atom-pair types with <500 examples at a B<30 cutoff were not analyzed. For visual comparison of related distributions, one reference case was chosen whose maximum counts per 0.01Å bin were typical of the group, usually around 2000 counts. A scalar multiplier was applied to the count values of the other, non-reference, distributions in the group, to bring them into comparable range while preserving the relative order of peak heights. This allows comparison of



distribution shapes, without affecting determination of the preferred min gap values. Figure S5 shows the distributions of H-to-H distances for the old x-H plus van der Waals radii (peaking at overlapped min gap) vs for the revised parameters (peaking at the optimal zero min gap). Similarly, Figure 6 in the main text shows that clashscores show a more desirable distribution in the new system.

**References for Supplement**


1 Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, Koseki S, Matsunaga N, Nguyen KA, Su S, Windus TL, Dupuis M, Montgomery JA (1993) General Atomic and Molecular Electronic Structure System. J Comput Chem 14:1347-1363.

2 Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun Assoc Comp Mach 24:381-395.

3 Har-Peled S, Wang Y (2004) Shape fitting with outliers. SIAM J Computing 33:269–285.

4 Natarajan V, Wang Y, Bremer P-T, Pascucci V, Hamann B (2006) Segmenting molecular surfaces. Computer Aided Geom Design 23:495-509.

5 Lorensen WE and Cline HE (1987) Marching cubes: A high resolution 3D surface construction algorithm. Proc Assoc Comp Mach SIGGRAPH 1987:163-169.

6 Agarwal PK, Har-Peled S, Varadarajan KR (2004) Approximating extent measures of points. J Assoc Comp Mach 51:606-635.

7 Carr H, Snoeyink J, Axen U (2000) Computing contour trees in all dimensions. Comp Geom 24:75-94.

8 de Berg M, Cheong O, van Kreveld M (2008) Computational Geometry: Algorithms and Applications, 2nd edition, Springer-Verlag, Berlin.

9 Lehmann MS, Koetzle TF, Hamilton WC (1972) Precision neutron diffraction structure determination of protein and nucleic acid components. I. The crystal and molecular structure of the amino acid L-alanine. J Am Chem Soc 94:2657-60.

10 Coppens P, Vos A (1971) Electron density distribution in cyanuric acid. II. Neutron diffraction study at liquid nitrogen temperature and comparison of X-ray and neutron diffraction results. Acta Cryst B27:146-158.

11 Takusagawa F, Koetzle TF, Kou WWH, Parthasarathy R (1981) Structure of N-acetyl-L-cysteine: X-ray (T=295 K) and neutron (T=16 K) diffraction studies. Acta Cryst B37:1591-1596.

12 Allen FH (2002) The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Cryst B58:380-388.

13 Bartell LS, Kuchitsu K (1978) Representations of molecular force fields. V. On the equilibrium structure of methane. J Chem Phys 68:1213-1215.

14 Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S (2017) Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. Methods Molec Biol (Clifton NJ)1607:627-641.

15 Ottiger M, Bax A (1998) Determination of relative N-$H_N$, N-C', Cα-C', and Cα-Hα effective bond lengths in a protein by NMR in a dilute liquid crystalline phase. J Am Chem Soc 120:12334-12341.

16 Meyer W (1973) PNOxCI studies of electron correlation effects. I. Configuration expansion by means of nonorthogonal orbitals, and application to the ground state and ionized states of methane. J Chem Phys 58:1017-1035.

17 Sheldrick GM (2008) A short history of ShelX. Acta Cryst A64:112-122.

18 Grazulis S, Chateigner D, Downs RT, Yokochi AT, Quiros M, Lutterotti L, Manakova E, Butkus J, Moeck P, Le Bail A (2009) Crystallography Open Database – an open-access collection of crystal structures J Appl Cryst 42:726-729.

19 Dolomanov OV, Bourhis LJ, Gildea RJ, Howard JAK, Puschmann H (2009) OLEX2: a complete structure solution, refinement, and analysis program. J Appl Cryst 42:229-341.

20 Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Cryst D66:486-501.

21 Hintze BJ, Lewis SM, Richardson JS, Richardson DC (2016) Molprobity's ultimate rotamer library distributions for model validation. Proteins 84:1177–1189.

22 Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 285:1735–1747.

23 Word JM, Lovell SC, Labean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 285:1711–1733.

24 MySQL AB (2006) MySQL administrator's guide and language reference, 2nd ed, MySQL Press, Indianapolis.

25 Team RDC (2005) R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna.

**Lists of database files:**

CSD neutron entries

ACYGLY11

ADENOS01

AGLYSL01

ALUCAL04

ALUCAL05

ARGIND11

ASPARM02

ASPARM03

ASPARM05

ASPARM07
ASPARM08
ASPARM09
CAXKOB01
CAXKOB11
CREATH04
CREATH05
CYSTAC01
CYSTCL01
CYSTCL02
CYTOSM04
DLASPA02
DLSERN11
GLCICH01
GLUTAM01
GLYCIN03
GLYCIN05
GLYCIN15
GLYCIN16
GLYCIN19
GLYCIN20
GLYCIN21
GLYCIN22
GLYCIN23
GLYCIN24
GLYGLY04
GLYGLY09
GLYGLY11
GLYHCL
HIPPAC02
HISTCM12
HISTPA12
HISTPA14
HOPROL12
IMAZOL04

IMAZOL06
IMAZOL13
IMZMAL11
IMZMAL13
KEPNAU
LALNIN12
LALNIN22
LALNIN23
LARGPH03
LARGPH04
LARGPH07
LCYSTN12
LGLUAC011
LGLUAC03
LHISTD13
LSERMH10
LTHREO02
LTYROS11
LYSCLH02
LYSCLH11
MANMUJ
MEADEN02
METHYM01
NALCYS02
NALCYS10
NRURAM11
PHALNC01
SUXHID01
TGLYSU01
TGLYSU03
TGLYSU11
TGLYSU25
VALEHC11
WEHZAL01

CSD X-ray entries

ABANIC
ABUPUI
ABUPUI01
ABUQAP
ACAGAN
ACMBPN
ACOXUM
ACTYSN
ADEWUC
ADIBOF
ADOTAO
APALTY
ATONAZ
ATYRAN
ATYREE01
ATYREE02
ATYREE03
ATYRMA10
AWONOP
BAZQEZ
BOQCUF
BOQCUF01
BOQCUF010
BOQCUF02
BOQCUF03
BOQCUF04
BOQCUF05
BOQCUF06
BOQCUF07
BOQCUF08
BOQCUF09
BOWKOO
CAWJOA
CEDFAS

CEYCOZ (?)
COQNAX
COSGUM
CYSCLM11
DALREO
DALRIS
DBTYRS
DEPJEO
DEPJOY
DLTYRS
DMTYRS
DTYROS
ETEYOR
EWOVAN
FAGFEZ
FAPKEN
FAPLIS
FAZHET01
FOYTAP
FUQLAE
GIQQUS
GLTLYR10
GLUTAS02
GLUTAS03
GLUTAS04
GLUTAS05
GLUTAS06
GLYTRE02
GUKMUU
GYTRE03
HADFAT
HIDGOQ
HUKJUT
HULGAW
IXETIO

IYEBIX
JECYUL
JUKMEH
KAHSOB
KIXBOJ
KIYFED
KIYFIH
KIYPUD
LAWKIE
LCYSTN04
LCYSTN22
LCYSTN23
LCYSTN24
LCYSTN25
LIPYIT
LIPYUF
LIXJIL
LOCJET
LOCLOF
LOCLOF01
LODJOD
LODJUJ
LODKAQ
LTHREO01
LTHREO03
LTYRHC10
LTYROS10
LTYROS11
MAPKOE
MAWGUM
MEMTYR10
MOQLOU
MOVLOZ
MOVLOZ01
MTYROS

MTYROS01
NALCYS02
NALCYS10
NANYIK
NEPMIE
NEPMOK
NIZGEJ
OTROSC
PCTRIB10
QAGBIK
OJOMOP
QANGER
QAQPOP
QAQPUV
QOZNAN
RAZPUE
REPFEX
SEMQUK
SEZLOZ
SOJPAI
TALZUC
TANCOC
TANCUI
TICFIV
TUSMOJ
TYRPXL10
UCUXEW
UPUVOR
UPUWAE
UPUWEI
UZUKUW
VAGHIV
VAWTAQ
VEDCEM
VEDCOW

VEGHEV
VIFFEW
VINDIF
WASVAN
WOVTOQ
XAWBUU
XIJKIK01
XIMJAF
XIMJAF01
YASKEJ
YEBMEX
YEDGOD
YEFTUZ
YEFVAH
YEJTIQ
YIPWEZ
YIPWID
YIPWOJ
YIPWUP
ZAMZES
ZEFZAL10
ZOPZOT
ZULWEI

PDB high-resolution entries with good H difference peaks

1byi 0.97 224 dethiobiotin synthase
1gwe 0.88 503 catalase
1ix9 0.90 205x2 Mn superoxide dismutase
1m40 0.85 263 TEM-1 beta lactamase TS-complex
2ddx 0.86 333 xylanase
2e4t 0.96 519 cellulase Cel44A
2p74 0.88 263x2 CTX-M-9 apo beta lactamase
2xtt 0.93 223+36 trypsin
2z6w 0.96 165+11 cyclophilin/cyclosporin

3f7l 0.99 152 Cu,Zn superoxide dismutase

3g63 0.88 381 PfluDING (now obsoleted by 4f1v)

3lz5 0.95 316 aldose reductase

COD X-ray entries

2007362

2008732

2009193

2010697

2010698

2010756

2010852

2010913

2011015

2011462

2011463

2011619

2011627

2011663

2011719

2011951

2011995

2012166

2012283

2012378

2012875

2012949

2012974

2013180

2013332

2013353

2013821

2013893

2014209
2014266
2014370
2014935
2014954
2014976
2015074
2015152
2015153
2015219
2015336
2015363
2015476
2015749
2016423
2016490
2016491
2016492
2016493
2016644
2017100
2018039
2103180
2103453
2103480
2103481
2103647
2103648
2103746
2104650
2200057
2200281
2200453

2200500
2200501
2201007
2202215
2202470
2202913
2202947
2203711
2204520
2205292
2205705
2206130
2206391
2206612
2206918
2207108
2207177
2207595
2207905
2208394
2209426
2209573
2211269
2211552
2211856
2212022
2212920
2213032
2213115
2213204
2213333
2213757
2214252

2214410
2215349
2217293
2217301
2217410
2217653
2218369
2218405
2218435
2218503
2219071
2220934
2221397
2221894
2223740
2224387
2224476
2224690
2225028
2225082
2225187
2225832
2225977
2226014
2227303
2227519
2227617
2228836
2229608
2230555