Determining the minimum number of types necessary to represent the sizes of protein atoms

Jerry Tsai[1], Neil Voss[2], and Mark Gerstein[2†]

[1]Department of Biochemistry & Biophysics
103 Biochemistry / Biophysics Building
Texas A&M University, 2128 TAMU
College Station, Texas 77843-2128

[2]Department of Molecular Biophysics and Biochemistry
Yale University, Bass Center, 266 Whitney Avenue
P.O. Box 208114, New Haven, CT 06520-8114

[†]corresponding author: phone - (203) 432-6105
fax - (360) 838-7861

Running Title: Minimum Set of Atom Types

Total Number of Pages - Cover & Manuscript 18
Tables 2
Figures 4

Abstract

Motivation: Traditionally, for packing calculations people have collected atoms together into a number of distinct "types". These, in fact, often represent a heavy atom and its associated hydrogens (i.e. a united atom). Also, atom typing usually is done according to basic chemistry, giving rise to 20 to 30 protein atom types, such as carbonyl carbons, methyl groups, and hydroxyl groups. No one has yet investigated how similar in packing these chemically derived types are. Here we address this question in detail, using Voronoi volume calculations on a set of high-resolution crystal structures.

Results: We perform a rigorous clustering analysis with cross-validation on tens of thousands of atom volumes and attempt to compile them into types based purely on packing. From our analysis, we are able to determine a "minimal" set of 18 atom types that most efficiently represent the spectrum of packing in proteins. Furthermore, we are able to uncover a number of inconsistencies in traditional chemical typing schemes, where differently typed atoms have almost the same effective size. In particular, we find that tetrahedral carbons with two hydrogens are almost identical in size to many aromatic carbons with a single hydrogen.

Availability: Programs available from http://bioinfo.mbb.yale.edu/geometry and http://molmovdb.org/geometry.

Contact: Jerry Tsai, jwtsai@u.washington.edu
Neil Voss, neil.voss@yale.edu
Mark Gerstein, Mark.Gerstein@yale.edu

Supplementary Information: Available at http://bioinfo.mbb.yale.edu/geometry and http://molmovdb.org/geometry .

Keywords : protein atom volumes, protein atom typing, packing, Voronoi polyhedra

Introduction

Numerous methods have been developed to determine protein atom radii and volumes (Bondi 1964; Chothia 1974; Richards 1974; Finney 1975; Harpaz, Gerstein et al. 1994; Li and Nussinov 1998; Liang, Edelsbrunner et al. 1998; Tsai, Taylor et al. 1999). These radii and volumes have been necessary in understanding protein structure and particularly, in uncovering the relationship between packing and stability. A more accurate protein radii and volume set helps make these calculations more accurate. Examples of studies requiring an accurate radii and volumes have characterized a number of protein properties, such as protein energies (Chothia 1975), protein-protein interactions (Janin and Chothia 1990), standard residue volumes (Harpaz, Gerstein et al. 1994), internal core packing (Janin 1979; Richards 1985), packing at the water interface (Gerstein, Tsai et al. 1995; Gerstein and Chothia 1996), protein cavities (Richards 1979; Hubbard and Argos 1995; Liang, Edelsbrunner et al. 1998; Liang, Edelsbrunner et al. 1998), the quality of crystal structures (Pontius, Richelle et al. 1996), and even measurement of the fit between an enzyme and its substrate (Finney 1978; David 1988). All methods that calculate volumes of atoms use some sort of atom typing scheme with an associated radii set assigned to these types. As shown in Table I, the most common approach for typing protein atoms uses united atoms and a chemical typing scheme. United atoms are necessary because most protein structures solved by X-ray diffraction do not have resolved hydrogen atoms. This united model convention produces atomic groups instead of individual atoms for many atom types, where an atom type includes a heavy atom and its associated hydrogen atoms. However, for the sake of simplicity, we sometimes refer to atom types as atoms, even if some are actually a group of atoms. A chemical typing scheme results from atom types derived from the radii used. Since

the chemistry of an atom determines its radii, atom typing schemes based on atom radii are also tied to an atom's chemistry. The work described in this paper investigates this assumption that the best atom typing for protein volume calculations should be solely dependent on the chemistry of atoms and attempts to find the minimum number of atom types necessary to describe protein packing. Because atom typing has always been associated with the radii used, the separate issue of an appropriate atom typing scheme has not been well addressed. In order to isolate the effects of atom typing from the effects of a radii set, we calculated atom volumes without radii and compare them based on the atom typing scheme used.

To calculate atom volumes, we turned to Voronoi polyhedra (Voronoi 1908). Bernal and Finney (Bernal and Finney 1967) first applied this method to molecular systems and Richards (Richards 1974) first used it with proteins. The method used in this work has been previously described (Tsai, Taylor et al. 1999). As with our earlier work, we only include atoms whose volumes are well defined e.g., non-surface atoms and atoms not next to ligands. Figure 1 illustrates how a Voronoi polyhedron is built. Because this construct partitions space such that all points within a polyhedron are closer to that atom than any other, the Voronoi method provides a good estimate of the true volume of an atom and in turn, reliable values for the comparison of atom volumes. As a point of departure, we show the ProtOr radii set in Table I (Tsai, Taylor et al. 1999). The typing follows standard united atom conventions and chemical atom typing. Using Voronoi polyhedra, clustering, and cross validation analysis, we compare chemical and numerical typing schemes to find which typing scheme and the minimum number of atom types that can provide the best general description of protein volumes. Along the way, we show that protein atoms do not always group themselves according to their chemical type or possess volumes that correlate well with the size of

their radii. In other words, the different chemistry of atoms does not automatically imply that atoms will occupy distinct volumes when packed in proteins structures. As expected, in terms of radii and volumes, a chemical typing scheme is somewhat degenerate. In the end, we decided on a compromise between the numerical and chemical typing schemes, and the atom typing used for our radii set does not deviate far from the aforementioned united atom, chemically based approaches (see Table II).

Methods, System, and Implementation

Protein Data Set

As was used in a previous study (Tsai, Taylor et al. 1999), a set of 87 structures was used to calculate protein volumes. Based on a 1.75 Å resolution cutoff, these structures were chosen from a larger list of structures that contained the best example of a SCOP (Structural Classification of Proteins) classified domain (Murzin, Brenner et al. 1995). The primary goal of this set of proteins is to contain the broadest representation of protein environments. The 4 letter Protein Data Bank (PDB) (Bernstein, Koetzle et al. 1977; Abola, Sussman et al. 1997) codes are as follows: 1cbn, 1lkk, 2erl, 8rxn, 1bpi, 1ctj, 1igd, 1rge, 1amm, 1arb, 1cse, 1jbc, 2sn3, 1cus, 7rsa, 1rro, 1aac, 193l, 1utg, 5p21, 1hms, 1xyz, 256b, 2olb, 2phy, 3ebx, 3sdh, 2end, 1xso, 1cka, 1cyo, 1edm, 1ezm, 1isu, 1mla, 1poa, 1rie, 1whi, 2ctb, 2eng, 2ovo, 2cba, 3grs, 1lit, 1ra9, 1tca, 1csh, 1epn, 1mrj, 1phc, 1ptf, 1smd, 1vcc, 2dri, 2ilk, 2sil, 3pte, 4fgf, 2cpl, 1kap, 1lcp, 1php, 1snc, 1sri, 2wrp, 1krn, 2trx, 1ctf, 1fnb, 1gai, 1gof, 1knb, 1llp, 1mol, 1pdo, 1rop, 1tad, 1tfe, 1vhh, 1vsd, 2act, 1fkd, 1chd, 1kpt, 1thw, 2bbk, 3cla.

Analysis of the Effect of Atom Typing

For protein atoms, we followed the conventions used in the PDB with one exception. We made a distinction between the two oxidation states of a cysteine residue

by using the usual code CYS for reduced cysteine and the special code CSS for disulfide bonded cystine. Altogether, this adds 1 extra residue for a total of 21 and 6 more protein atoms for a total of 173.

For our atom type notation (see Table I), the uppercase letter in the first register identifies the heavy atom: C, N, O, and S (carbon, nitrogen, oxygen, and sulfur, respectively). The number in the second register indicates the number of covalent bonds the atom makes. The third register is always an H for hydrogen. The fourth register shows the number of hydrogen atoms connected to the heavy atom. Therefore, in our notation an sp^3 carbon with two hydrogens is C4H2, an sp2 carbon with one hydrogen atom would be classified as C3H1, and a hydroxyl group is O2H1. Also, as explained below and used in Table II, an additional lowercase letter is used in the fifth register to describe the atom type: s, b, or u (small, big, or unique, respectively).

Our procedure to test atom typing schemes requires the creation of 1) a reference set of protein atom volumes and 2) a procedure for testing the parameterization of atom types. One important feature of the reference volumes is that it must allow us to separate the effects of atom typing from the influence of atom radii. Therefore, we decided to calculate the reference set of 173 protein atom volumes with Voronoi polyhedra using the bisection plane-positioning method (Voronoi 1908), which is a method that does not use atom radii. These "raw" volumes allow us to compare different typing schemes without any bias from a radii set. To test the parameterization of atom types, we first begin with a scheme with n types and incrementally decrease the number of types to one.

Chemical Typing

Because the atom volumes naturally cluster by chemical type, we initially chose a typing scheme based on one used in previous studies.(Chothia 1974; Harpaz, Gerstein et al. 1994; Gerstein, Tsai et al. 1995) Figure 2A shows this scheme. At the level with most types, we used standard chemical types with the proviso that mainchain atoms are separated from sidechain ones. From this initial set, each derivative set was created by successively collapsing logical types together. The first derivative level collapses the mainchain and sidechain types together. From this level, a third groups atoms according to heavy atom type and number of possible covalent bonds. The fourth level combines atoms solely based on heavy atom type, and the fifth distinguishes atoms by hydrophobicity.

Numerical Typing

To contrast the chemical typing scheme, we created two typing schemes based purely on numerical criteria. As shown in Figures 2B and C, these numerical typing schemes do not follow the conventions of chemistry. In deriving both schemes, we successively added one atom to a cluster or brought two clusters together, such that the overall number of clusters started at the 173 individual protein atom volumes (the "raw" 173 volumes) and decreased incrementally to 1. In the first, single-linkage clustering scheme (Figure 2B), atom volumes were grouped based on nearest neighbors i.e., the next cluster was formed by joining the two clusters that were closest in distance. The distance between clusters was defined simply as the smallest volume difference between them. In a second scheme (Figure 2C), we applied multi-linkage clustering to the raw 173 volumes. A new cluster was generated by minimizing on its width. By width, we mean the difference between the largest and smallest volume within that cluster. In Figures 2B and C, the atoms are arranged from smallest to largest volume.

Comparing the two shows that multi-linkage clustering produces a more symmetrical pattern, indicating that multi-linkage clustering groups volumes more evenly by size than the single-linkage clustering.

Residual Calculation and Cross-Validation

To compare the three typing schemes with each other, we calculated a residual for each set in a typing scheme, which we called E_{stat}. The raw 173 volumes were collapsed into n derived atom type volumes, depending upon atom typing (determined from the clustering). These n derived volumes were then expanded back out into a predicted 173 volumes, which were used to see how well the n derived volumes estimated the raw 173 volumes. For each of the 173 protein atoms, the difference between the predicted and raw values were summed.

$$E_{stat}=\Sigma_{(i=1,173)}(V(i)-V_c(i))^2, \quad (1)$$

where V_i is the volume of type i in the original 173 type set and $V_c(i)$ is its predicted volume based on the clustering. The predicted value is the mean volume for cluster c. The number of elements in a given cluster c is N_c and the cluster variance is $\sigma_c^2 = <(V(i)-<V(i)>)^2>$ (where the averaging is over all types i in the cluster). Given these definitions, formula 1 is mathematically equivalent to:

$$E_{stat} = \Sigma_c N_c \sigma_c^2, \quad (2)$$

where the sum is over all clusters c.

Obviously, a clustering with more types essentially allows for more parameters in the calculation, leading to the possibility of over-fitting (Efron and Tibshirani 1991). Consequently, we cross-validated the single- and multi-linkage typing schemes. To do this, we randomly excluded 10 values from the unclustered 173 volumes and used the

remaining 163 to calculate values for cluster centers. In terms of notation, we denote one of the 10 excluded volumes from the 173 as $^{173}V_{ex}(i)$ and the cluster volumes derived from only the 163 types as $^{163}V_c$. Finally, we calculated a cross-validated residual difference E_{cv} between the excluded volumes and their predicted volumes based on the 163 volume clustering:

$$E_{cv}(j) = \Sigma_{(i=1,10)} (^{173}V_{ex}(i) - ^{163}V_c(i))^2. \quad (3)$$

We subscript the residual difference by j to indicate that it is for one distinct set of ten volumes excluded. For each distinct number of atom types from 1 to 25, we repeated this exclusion procedure 100 times, excluding a different set of 10 atoms each time, to generate an averaged residual.

Following the E_{stat} analysis described above, Figure 3 shows how closely the type sets are able to predict the raw 173 volumes. As expected, the E_{stat} falls to zero when a type for each atom is used (i.e. 173 atom types). Although not entirely surprising, the chemical typing set does not generally do as well as the numerically generated sets, since it produces a much more scattered distribution. At higher numbers of types, the chemical typing performs almost as well as the single-linkage set with the same number of atom types (inset to Figure 3). The multi-linkage typing scheme does better than both the chemical and the single-linkage typing ones. It also possesses a much smoother distribution. Focusing on the region up to 25 types (inset to Figure 3), we see that the results from the multi-linkage cluster are not perfectly smooth. To make sure that we were doing the calculations correctly, we also show the residual for both the single and multi-linkage clustering schemes. The residual for both numerical clustering schemes follows their respective E_{stat}, which gives us confidence in our calculations.

Discussion

Differences between Chemical and Numerical Typing

The results in Figure 3 show that typing protein atoms using numeric criteria produces a better fit than strict chemical typing. To find an explanation for this difference, we looked at the distribution of the raw 173 protein atom volumes separated by purely chemical attributes (Figure 4). As expected, the raw 173 volumes clustered loosely according to chemical type, but surprisingly the distribution also shows significant overlap in atom types between 19 and 24 $Å^3$. The types within this range are also surprising because their sizes do not reflect their chemical type.

The most striking feature is the similar sizes of sp^3 carbons with two hydrogens (C4H2) and sp^2 carbons with one hydrogen (C3H1). We believe the reason for the similar sizes of these two chemically different carbon atoms is that all of the sp^2 carbons with one hydrogen (C3H1) seem to have an increased volume. These C3H1 atoms belong to the ring systems of aromatic residues, and packing around these planar residues within the core must not be as tight as for the aliphatics. We do not believe that this effect is an artifact or due to the lack of radii set. If it were, we would expect the effect to be general in nature and affect all atom volumes. Looking at both types of atoms with one less hydrogen, we do not find the same phenomenon. The sp^2 carbons with no hydrogens (C3H0) are smaller than sp^3 carbons with one hydrogen (C4H1), which makes sense chemically. Since we do not see a similar effect with the C3H0 and C4H1 groups, the effect must be real.

There are some other chemically dissimilar types that also overlap in volume. In particular, the similarity in size of the two types of oxygen atoms (O1H0 and O2H1) is most likely due to a decrease in the hydroxyl atom (O2H1) size from electroconstriction. As stated earlier, this analysis only considered buried protein atoms. For O2H1 atoms

to exist in the interior of a protein, they must take part in hydrogen bonds. Otherwise, the energetically unfavorable situation arises where an unsatisfied dipole would exist in a primarily nonpolar environment. Hydrogen bonded neighbors are closer, which reduces the size of the O2H1 atom.

Other interesting, yet somewhat intuitive differences can be found between protein mainchain versus sidechain atoms. As illustrated in Figure 4, the densest clusters within a chemical type belong to mainchain atoms i.e., the first group of C4H1 atoms are all mainchain C_α. These tight clusters indicate that packing is more regular around mainchain atoms and less so around sidechain atoms.

An Optimal Set of 18 Types: A Hybrid Chemical and Numerical Typing

The primary goal of this work was to develop a typing scheme that would be generally useful and more accurate in calculations of protein properties than currently used chemical typing schemes. The previous analysis clearly favors a numerical typing scheme over a chemical one, since a chemical typing scheme does not account for the clustering of the raw 173 volumes (Figure 4). However, a completely numeric atom typing scheme would be unwieldy and confusing outside of the calculation of protein volumes, since many atoms of the same chemical type would end up in different groups. Considering these points, we have decided upon a compromise between the two and have named it the ProtOr typing scheme. We retain the original chemical typing scheme as shown in Table I and add a few minor adjustments influenced by the numerical typing. Our proposed typing scheme is summarized in Table II. It is basically 18 different atom types: the 13 basic chemical types found in proteins and the expansion of 5 of these types into groups of two. For the 8 chemical types that are not expanded, there is only one or a unique type, so the letter “u” is added to the atom type

name. Of the 5 types that are expanded, the C3H0, C3H1, C4H1, C4H2 and N3H1 chemical types are divided into small or big ("s" or "b", respectively) groups based loosely on the multi-linkage, numerical clustering of the volumes within an atom type. It is these 5 that we discuss in detail:

C3H0. Possessing the smallest volumes, atoms with the C3H0s type are carbonyl carbons with branching bonds at one of their covalently bonded neighbors. This describes mainchain carbonyl carbons of residues possessing a C_β atom: all residues except glycine. The larger C3H0b atom type contains glycine's mainchain carbonyl carbon and all sidechain carboxyl and carbonyl carbons.

C3H1. Adding a hydrogen, the C3H1s and C3H1b atom types consist of aromatic carbons and were typed according to numerical, multi-linkage clustering criteria.

C4H1. For aliphatic carbons, the atoms of the C4H1s type have covalent neighbors with branched bonds and are not part of a ring structure. So, the C4H1s type includes all mainchain C_α atoms except those from glycine, alanine, and proline. The C4H1b type contains all the rest: alanine and proline's C_α and all of the sidechain aliphatic carbons with one hydrogen.

C4H2. As with the aromatic carbons, the aliphatic carbons with two hydrogens were split into C4H2s and C4H2b types using multi-linkage clustering criteria.

N3H1. Lastly, nitrogen atoms with one hydrogen separate into two groups. All mainchain nitrogens except for proline's are in the N3H1s group (proline's nitrogen has no hydrogen or is an imide and is its own group, see Table II). All sidechain nitrogens with one hydrogen are included in N3H1b group. As mentioned above, the difference within this atom type is most likely due to the packing environment.

Conclusion

After comparing chemical versus numerical typing schemes, the ProtOr atom typing scheme that best represents the packing environment of proteins is a compromise. The analysis points out that the overlap in volumes between atoms that are generally considered chemically different i.e., the C4H2 and C3H1 atoms. As a result, the final atom typing scheme consists of primarily a chemical typing scheme with subgroups based on numerical criteria. This allows for an increased accuracy of atom type volumes while retaining the intuitive nature of a chemical scheme. While the change has not been drastic, we expect that the increased accuracy of the atom typing scheme will increase the accuracy and understanding of measurements using a protein atom radii set and protein volumes such as those mentioned in the introduction.

Source Code and Parameter Database Available on the Web

We make available a general code base for geometric calculations on macromolecular structures. This includes: (1) code and executables for calculating Voronoi polyhedra and its dual, Delaunay triangulations, as well as (2) programs to calculate related geometric quantities -- such as accessible surfaces, helix axes, least-squares fits, H-bonds, VDW contacts, and crystal symmetry. We also make available the ProtOr atom typing and radii set as well as an extensive collection (i.e. database) of geometric parameters associated with the calculations. These items can be retrieved by sending e-mail to Mark.Gerstein@yale.edu or by using the World Wide Web to access the following URL:

http://bioinfo.mbb.yale.edu/geometry.

Acknowledgements

JT would like to thank the National Institutes of Health (grant number GM41455) and the National Science Foundation (Bioinformatics Fellowship) for support. MG would like to thank the Keck foundation and the National Science Foundation for support (grant number DBI-9723182). We also appreciate the kind support and suggestions from David Baker, Cyrus Chothia, Michael Levitt, and Fred Richards.

References

Abola, E. E., Sussman, J. L., Prilusky, J. and Manning, N. O. (1997). "Protein Data Bank Archives of Three-Dimensional Macromolecular Structures." Meth. in Enzym. **277**: 556-571.

Bernal, J. D. and Finney, J. L. (1967). "Random close-packed hard-sphere model II. Geometry of random packing of hard spheres." Disc. Faraday Soc. **43**: 62-69.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D. j., Rodgers, J. r., Kennard, O., Shimanouchi, O. and Tasumi, M. (1977). "Protein Data Bank: a computer-based archival file for macromolecular structures." J. Mol. Biol. **112**: 535-542.

Bondi, A. (1964). "van der Waals Volumes and Radii." J. Phys. Chem. **68**: 441-451.

Chothia, C. (1974). "Hydrophobic bonding and accessible surface area in proteins." Nature **248**: 338-339.

Chothia, C. (1975). "Structural invariants in protein folding." Nature **254**: 304-308.

David, C. W. (1988). "Voronoi Polyhedra as Structure Probes in Large Molecular Systems." Biopolymers **27**: 339-344.

Efron, B. and Tibshirani, R. (1991). "Statistical Data Analysis in the Computer Age." Science **253**: 390-395.

Finney, J. L. (1975). "Volume Occupation, Environment and Accessibility in Proteins. The Problem of the Protein Surface." J. Mol. Biol. **96**: 721-732.

Finney, J. L. (1978). "Volume Occupation, Environment, and Accessibility in Proteins. Environment and Molecular Area of RNase-S." J. Mol. Biol. **119**: 415-441.

Gerstein, M. and Chothia, C. (1996). "Packing at the Protein-Water Interface." Proc. Natl. Acad. Sci. USA **93**: 10167-10172.

Gerstein, M., Tsai, J. and Levitt, M. (1995). "The volume of atoms on the protein surface: Calculated from simulation, using Voronoi polyhedra." J. Mol. Biol. **249**: 955-966.

Harpaz, Y., Gerstein, M. and Chothia, C. (1994). "Volume Changes on Protein Folding." Structure **2**: 641-649.

Hubbard, S. J. and Argos, P. (1995). "Detection of internal cavities in globular proteins." Protein Engineering **8**(10): 1011-1015.

Janin, J. (1979). "Surface and inside volumes in globular proteins." Nature **277**(5696): 491-492.

Janin, J. and Chothia, C. (1990). "The Structure of Protein-Protein Recognition Sites." J. Biol. Chem. **265**: 16027-16030.

Li, A. J. and Nussinov, R. (1998). "A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking." Proteins **32**(1): 111-27.

Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V. and Subramaniam, S. (1998). "Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape." Proteins **33**(1): 1-17.

Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V. and Subramaniam, S. (1998). "Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins." Proteins **33**(1): 18-29.

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). "Scop: a structural classification of proteins database for the investigation of sequences and structures." J. Mol. Biol. **247**(4): 537-540.

Pontius, J., Richelle, J. and Wodak, S. J. (1996). "Deviations from Standard Atomic Volumes as a Quality Measure of Protein Crystal Structures." J. Mol. Biol. **264**: 121-136.

Richards, F. M. (1974). "The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density." J. Mol. Biol. **82**: 1-14.

Richards, F. M. (1979). "Packing Defects, Cavities, Volume Fluctuations, and Access to the Interior of Proteins. Including Some General Comments on Surface Area and Protein Structure." Carlsberg. Res. Commun. **44**: 47-63.

Richards, F. M. (1985). "Calculation of Molecular Volumes and Areas for Structures of Known Geometry." Meth. in Enzym. **115**: 440-464.

Tsai, J., Taylor, R., Chothia, C. and Gerstein, M. (1999). "The Packing Density in Proteins: Standard Radii and Volumes." J. Mol. Biol. **290**(1): 253-266.

Voronoi, G. F. (1908). "Nouveles applications des paramétres continus á la théorie de formes quadratiques." J. Reine Angew. Math. **134**: 198-287.

Figure Captions

Figure 1. Two-dimensional representation of Voronoi polyhedra construction. A polyhedron is built around the central atom. Points are the centers of atoms. The calculation first finds points within a distance cutoff (the outer circle) to a central atom. For each pair of atoms including the central atom, a face is created perpendicular to the line connecting the two atoms. The intersection of these faces creates vertices, which defines the polyhedron (shown by the shaded area). Those points sharing a polyhedron face are neighbors (circled atoms and bold connecting lines). Faces falling outside of the polyhedron (light connecting lines and broken lines for faces) indicate atoms that are occluded by others and are not direct neighbors to the central atom.

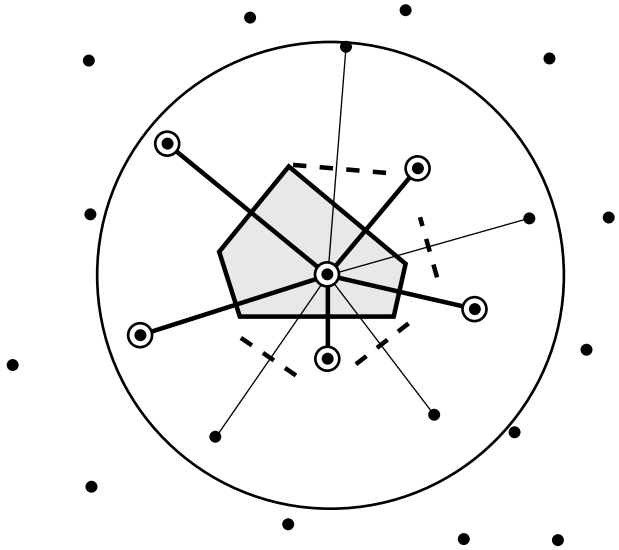
Figure 2. Clustering Trees. A). Tree showing Chemical Typing Scheme. At the left are the atom classes using a chemical typing scheme. The m indicates a group with mainchain atoms only. Moving to the left, groups are subsequently collapsed based loosely on chemical criteria and in the following order: backbone vs. sidechain, number of covalent bonds, heavy atom type, and hydrophobicity B). Single-Linkage Clustering. The leaves of the tree on the left denote the 173 distinct protein atoms types. They are arranged according to their volume from smallest on top to largest on the bottom. At each step, two branches are collapsed if the volumes of the groups are closest than any other. The average volume calculated over the new group is used in the subsequent rounds of clustering. C). Multi-Linkage Clustering. The clustering done here is similar to the single linkage clustering, but the criteria is different. Here, new clusters are formed based on their width. The widths of all possible combinations of clusters are calculated. The cluster with the smallest width is chosen as the next

cluster. For B) and C), length of lines connecting atoms together are proportional to the volumes' size differences.

Figure 3. Residual and Cross Validation. To measure the predictive power of the different atom typing sets created by clustering, a residual between the volumes predicted by the clustered atom typing scheme and the raw 173 volumes was calculated and is shown versus the number of atom types in that typing scheme cluster. To insure that this analysis is not over parameterized, we cross validated our data by randomly leaving out 10 volumes and recalculating the residual. This cross validated residual is also shown versus number of types in the cluster. The cross validation analysis is scaled to fit the residual values. The inset shows the range of atom types blown up from 0 to 25 with points connected by lines.

Figure 4. Distribution of the raw 173 reference volumes. The 173 raw, reference volumes are shown based on their volume. For clarity, the graph is broken up into three 10 $Å^3$ segments along the x axis. Also, atoms belonging to different chemical atom types are separated along the y axis and a consistent symbol for each heavy atom type is used -- diamonds for carbons, triangles for nitrogens, circles for oxygens, and squares for sulfurs.

Figure 1 - Voronoi Construct



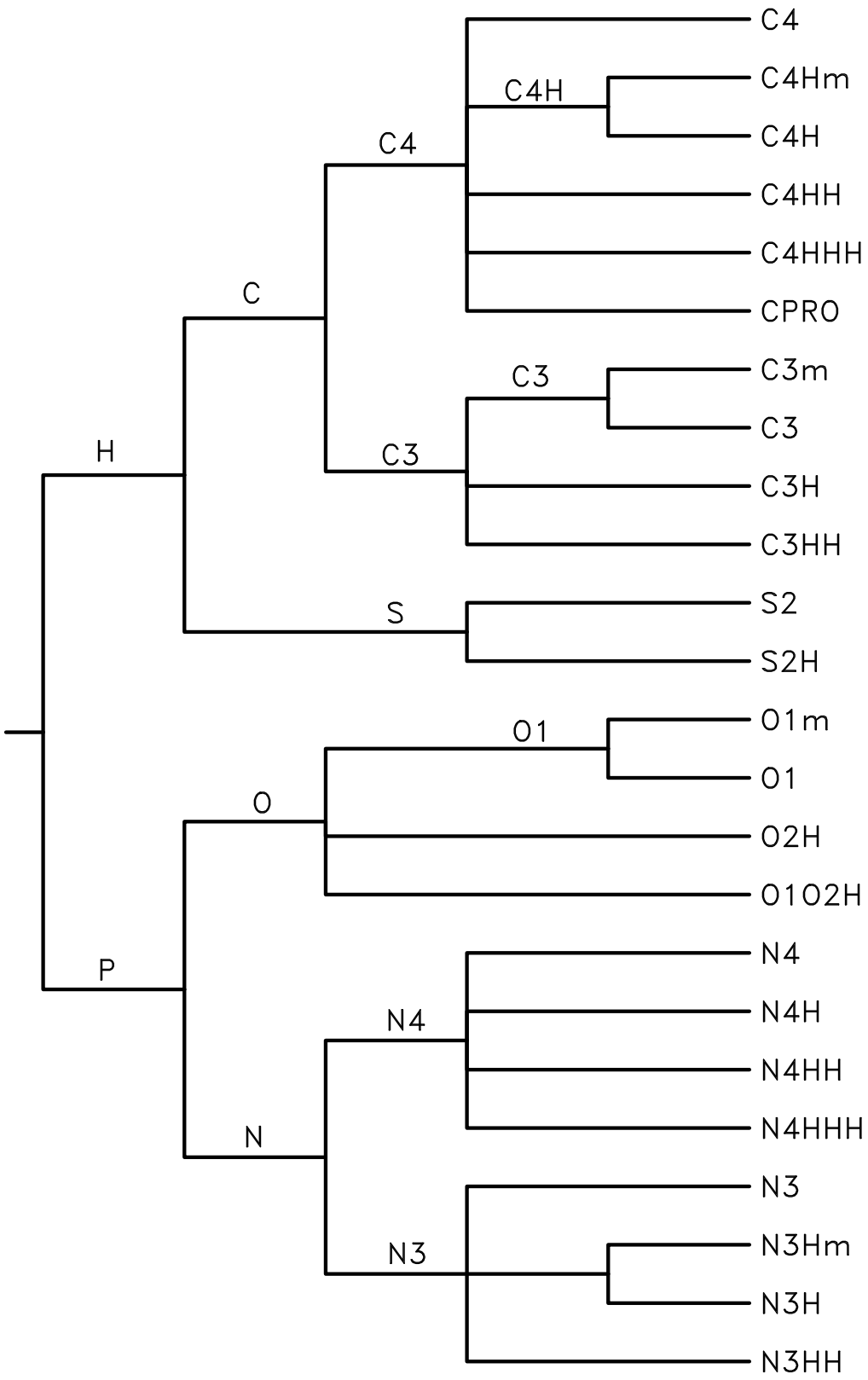


Fig. 2A - Chemical Tree

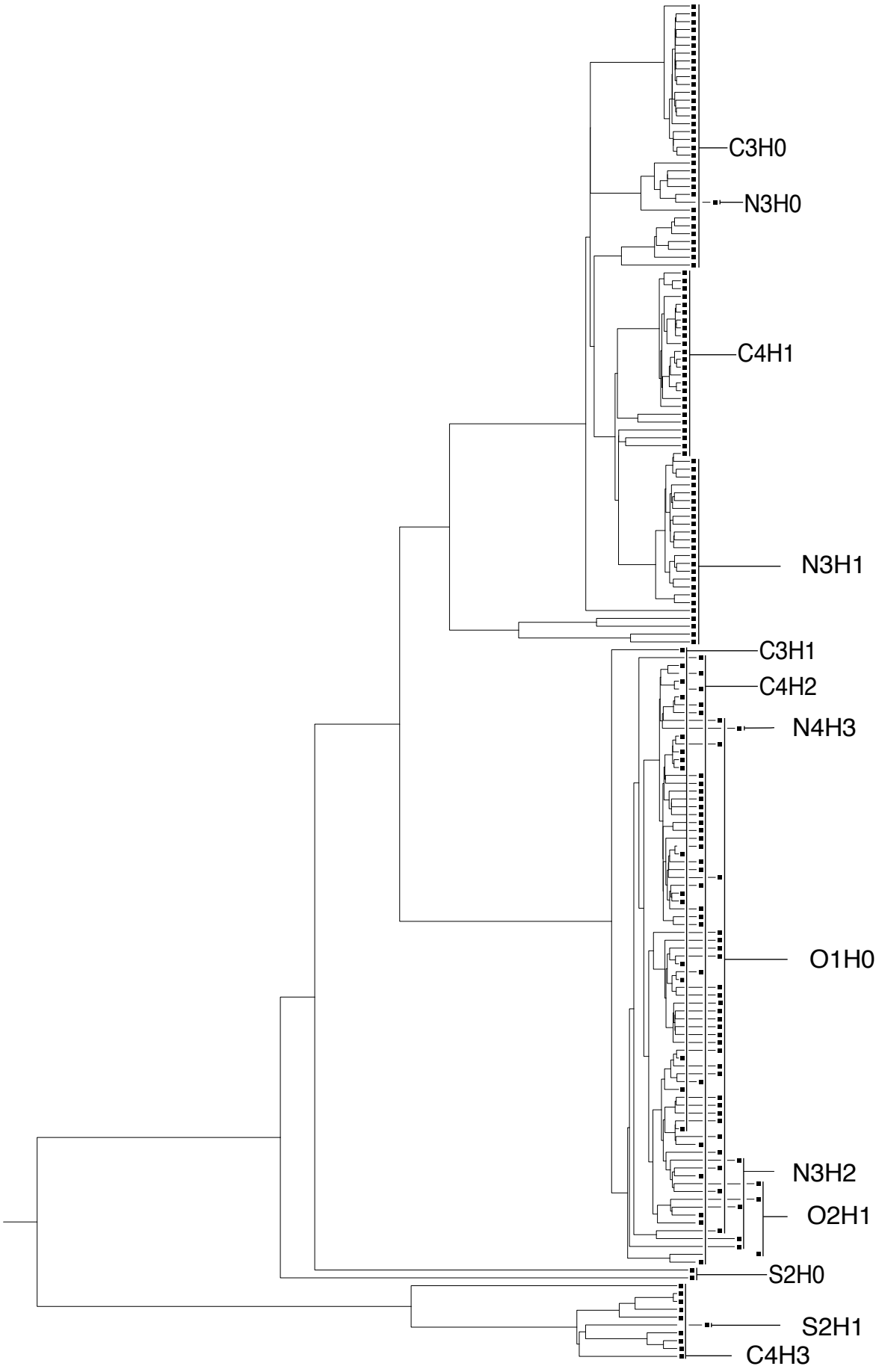


Fig. 2B-Single Linkage Tree

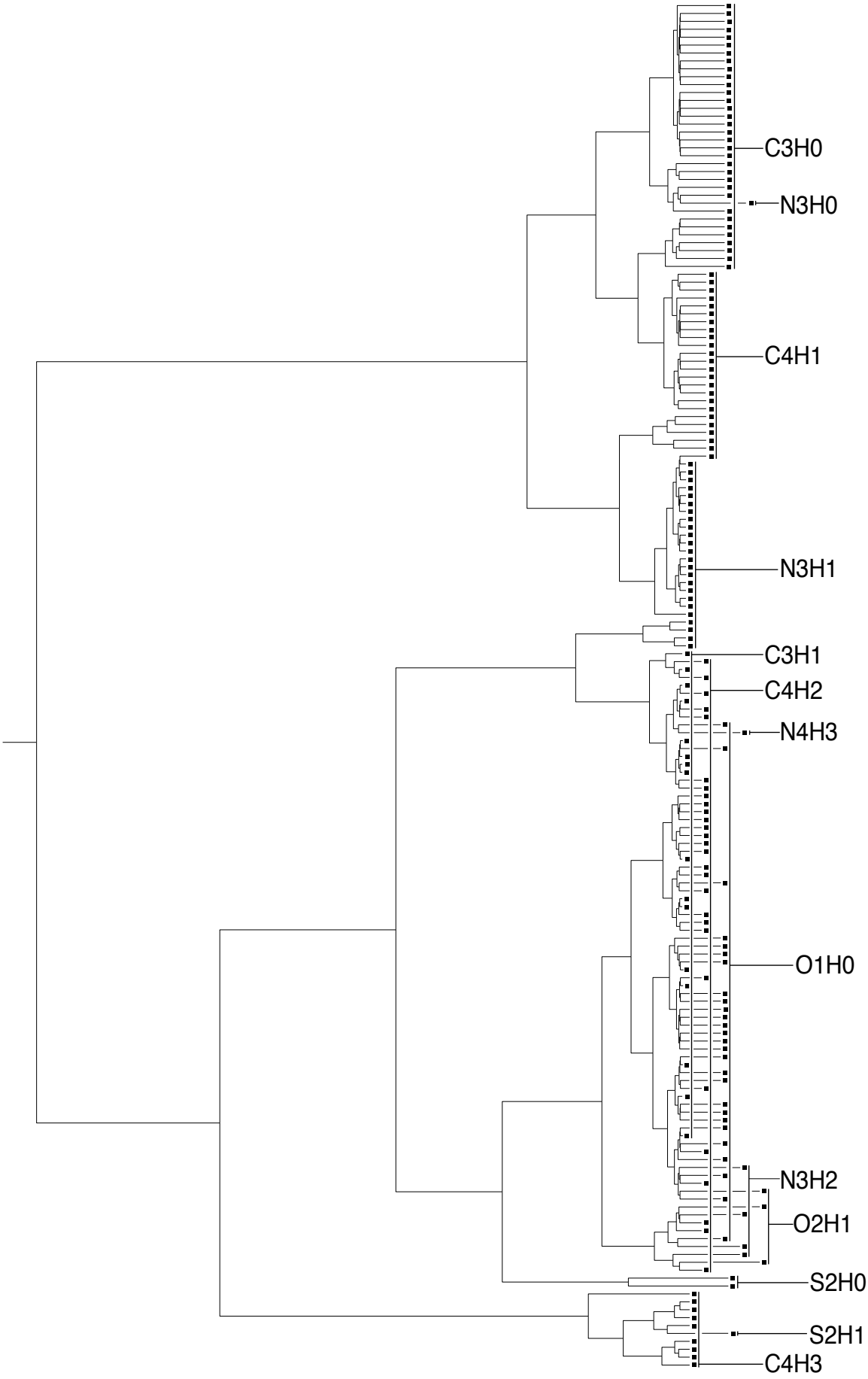


Fig. 2C - Multi-Linkage Tree

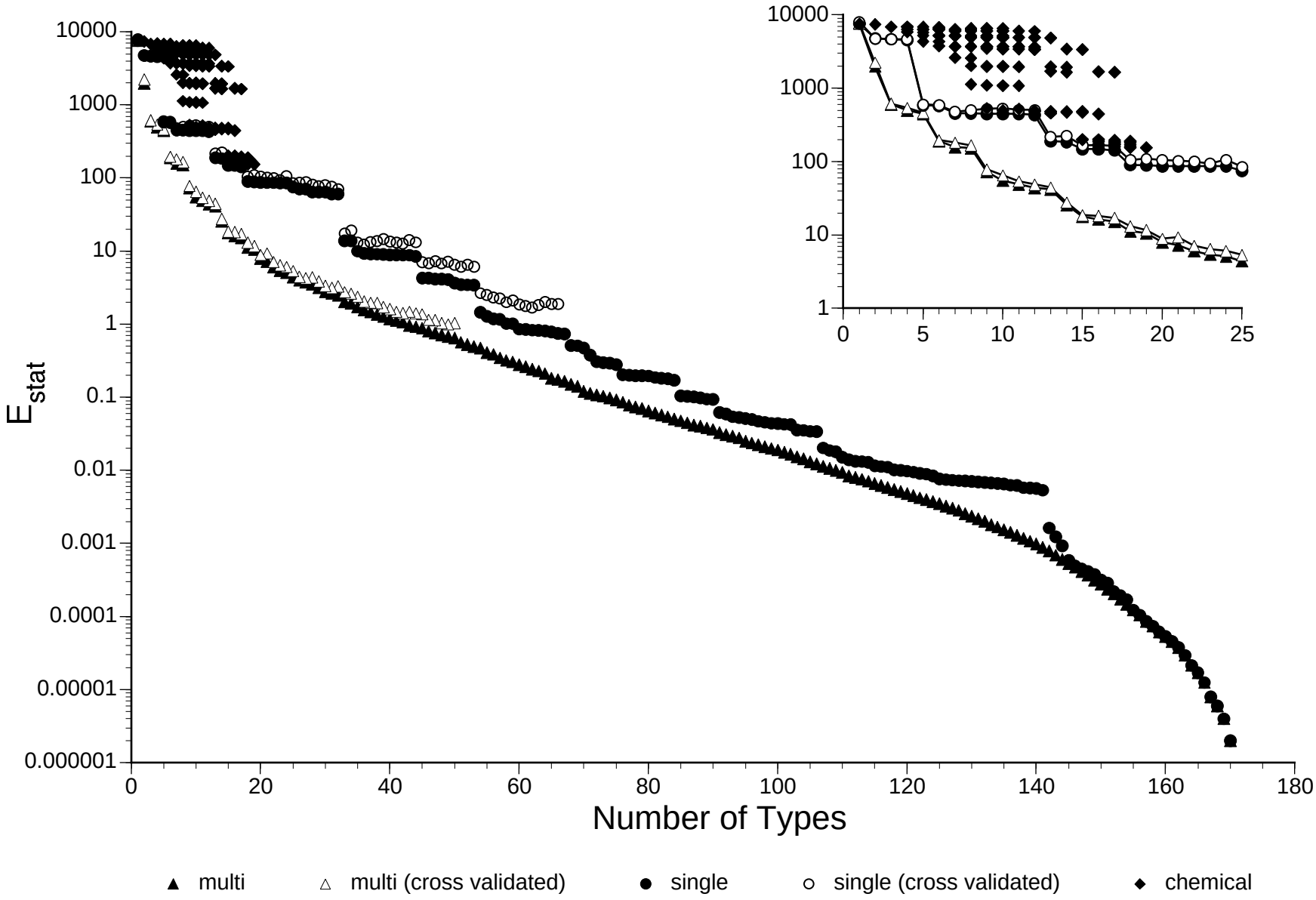

10000
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
0.000001
E stat
0
20
40
60
80
100
120
140
160
180
Number of Types
10000
1000
100
10
1
0
5
10
15
20
25
multi
multi (cross validated)
single
single (cross validated)
chemical

Fig. 3

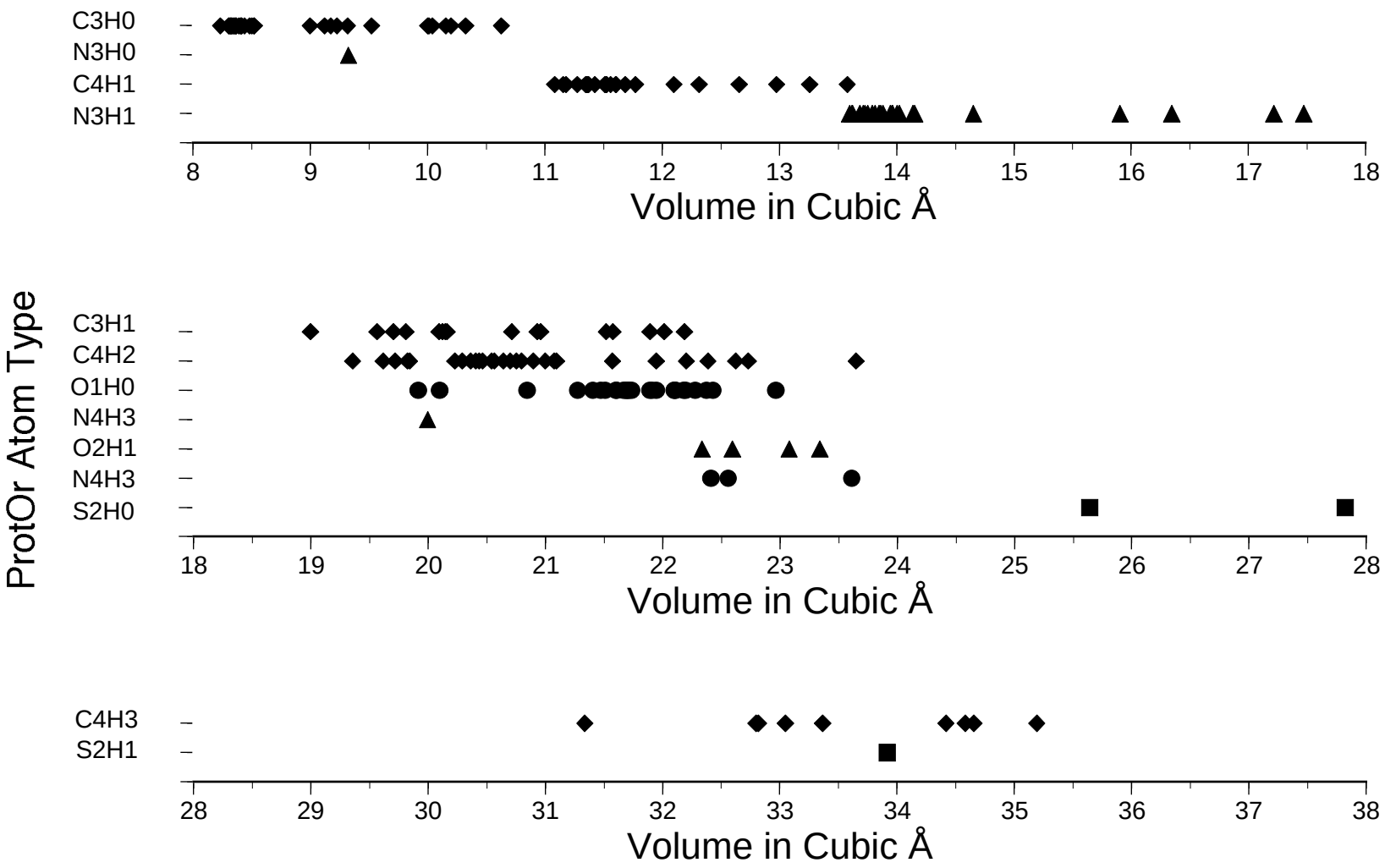

ProtOr Atom Type
C3H0
N3H0
C4H1
N3H1
8
9
10
11
12
13
14
15
16
17
18
Volume in Cubic Å
C3H1
C4H2
O1H0
N4H3
O2H1
N4H3
S2H0
18
19
20
21
22
23
24
25
26
27
28
Volume in Cubic Å
C4H3
S2H1
28
29
30
31
32
33
34
35
36
37
38
Volume in Cubic Å

Fig. 4

Table 1
ProtOr Volumes

ProtOr radii set	
atom	radii
C3H0	1.61
C3H1	1.76
C4H1	1.88
C4H2	1.88
C4H3	1.88
N3H0	1.64
N3H1	1.64
N3H2	1.64
N4H3	1.64
O1H0	1.42
O2H1	1.46
S2H0	1.77
S2H1	1.77

Atom Type	num (173)	Comments	Protein Atoms
C3H0s	20	carbonyl carbons with branching (mainchain carbonyls from residues with a C_β, so no gly carbon)	ALA_C,ARG_C,ASN_C,ASP_C,CSS_C,CYS_C,GLN_C,GLU_C,HIS_C,ILE_C,LEU_C, LYS_C,MET_C,PHE_C,PRO_C,SER_C,THR_C,TRP_C,TYR_C,VAL_C
C3H0b	13	carboxyl and carbonyl carbons w/o branching (side chain and glycine's) and aromatic carbons w/o hydrogen	ARG_CZ,ASN_CG,ASP_CG,GLN_CD,GLU_CD,GLY_C,HIS_CG,PHE_CG,TRP_CD2, TRP_CE2,TRP_CG,TYR_CG,TYR_CZ
C4H1s	18	aliphatic carbons with one hydrogen and branching from all three heavy atom bonds	ARG_CA,ASN_CA,ASP_CA,CSS_CA,CYS_CA,GLN_CA,GLU_CA,HIS_CA,ILE_CA, LEU_CA,LYS_CA,MET_CA,PHE_CA,SER_CA,THR_CA,TRP_CA,TYR_CA,VAL_CA
C4H1b	6	aliphatic carbons with one hydrogen and no branching through at least one heavy atom bond	ALA_CA,ILE_CB,LEU_CG,PRO_CA,THR_CB,VAL_CB
C3H1s	8	small aromatic carbons with one hydrogen	HIS_CD2,HIS_CE1,PHE_CD1,TRP_CD1,TYR_CD1,TYR_CD2,TYR_CE1,TYR_CE2
C3H1b	8	big aromatic carbons with one hydrogen	PHE_CD2,PHE_CE1,PHE_CE2,PHE_CZ,TRP_CE3,TRP_CH2,TRP_CZ2,TRP_CZ3
C4H2s	21	aliphatic carbons with two hydrogens, small	ARG_CB,ARG_CD,ARG_CG,ASN_CB,ASP_CB,GLN_CB,GLN_CG,GLU_CB,GLU_CG, GLY_CA,HIS_CB,LEU_CB,LYS_CB,LYS_CD,LYS_CG,MET_CB,PHE_CB,PRO_CD, SER_CB,TRP_CB,TYR_CB
C4H2b	7	aliphatic carbons with two hydrogens, big	CSS_CB,CYS_CB,ILE_CG1,LYS_CE,MET_CG,PRO_CB,PRO_CG
C4H3u	9	aliphatic carbons with three hydrogens, i.e. methyl groups	ALA_CB,ILE_CD1,ILE_CG2,LEU_CD1,LEU_CD2,MET_CE,THR_CG2,VAL_CG1,VAL_CG2
N3H0u	1	imide nitrogens (only member is Pro N)	PRO_N
N3H1s	20	amide nitrogens with one hydrogen (all other mainchain N's)	ALA_N,ARG_N,ASN_N,ASP_N,CSS_N,CYS_N,GLN_N,GLU_N,GLY_N,HIS_N,ILE_N, LEU_N,LYS_N,MET_N,PHE_N,SER_N,THR_N,TRP_N,TYR_N,VAL_N
N3H1b	4	amide nitrogens with one hydrogen (on sidechains)	ARG_NE,HIS_ND1,HIS_NE2,TRP_NE1
N3H2u	4	all amide nitrogens with 2 hydrogens (only on sidechains)	ARG_NH1,ARG_NH2,ASN_ND2,GLN_NE2
N4H3u	1	amide nitrogen charged, with 3 hydrogens	LYS_NZ
O1H0u	27	all oxygens in carboxyl or carbonyl groups (no distinction made between oxygens in carboxyl group)	ALA_O,ARG_O,ASN_O,ASN_OD1,ASP_O,ASP_OD1,ASP_OD2,CSS_O,CYS_O, GLN_O,GLN_OE1,GLU_O,GLU_OE1,GLU_OE2,GLY_O,HIS_O,ILE_O,LEU_O,LYS_O, MET_O,PHE_O,PRO_O,SER_O,THR_O,TRP_O,TYR_O,VAL_O
O2H1u	3	all hydroxyl atoms	SER_OG,THR_OG1,TYR_OH
S2H0u	2	sulfurs with no hydrogens	CSS_SG,MET_SD
S2H1u	1	sulfurs with one hydrogen	CYS_SG

Table 2. Summary of ProtOr Type Set