*Structural bioinformatics*

# Gaia: Automated Quality Assessment of Protein Structure Models

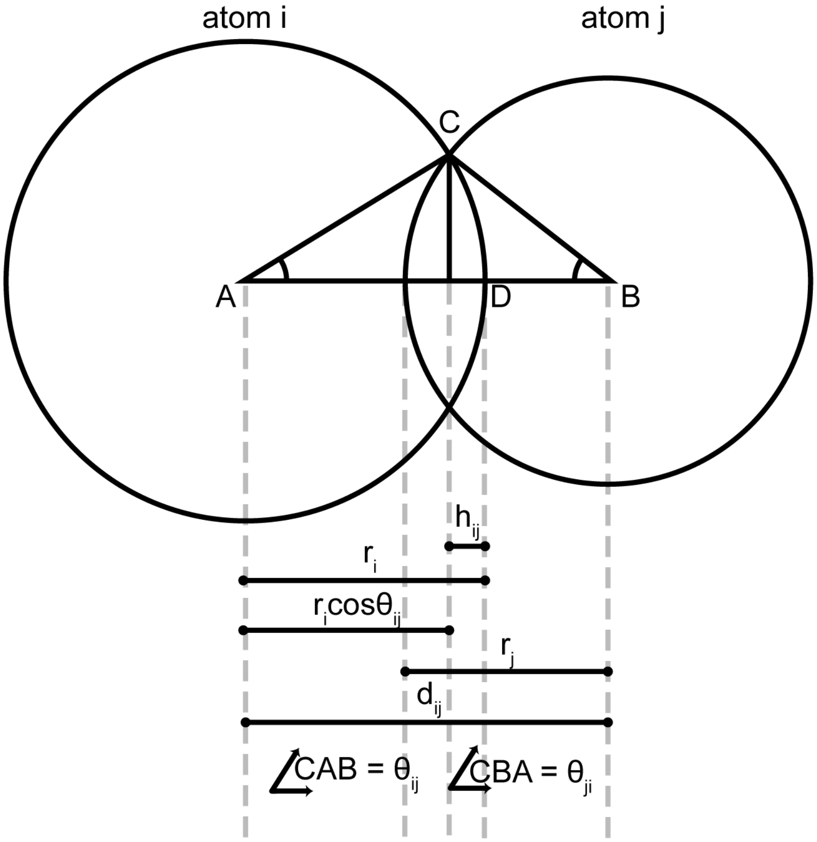
Pradeep Kota[1,2,3,†], Feng Ding[1,3, †], Srinivas Ramachandran[1,2,3 †], Nikolay V. Dokholyan[1,2,3*]

[1]Department of Biochemistry and Biophysics

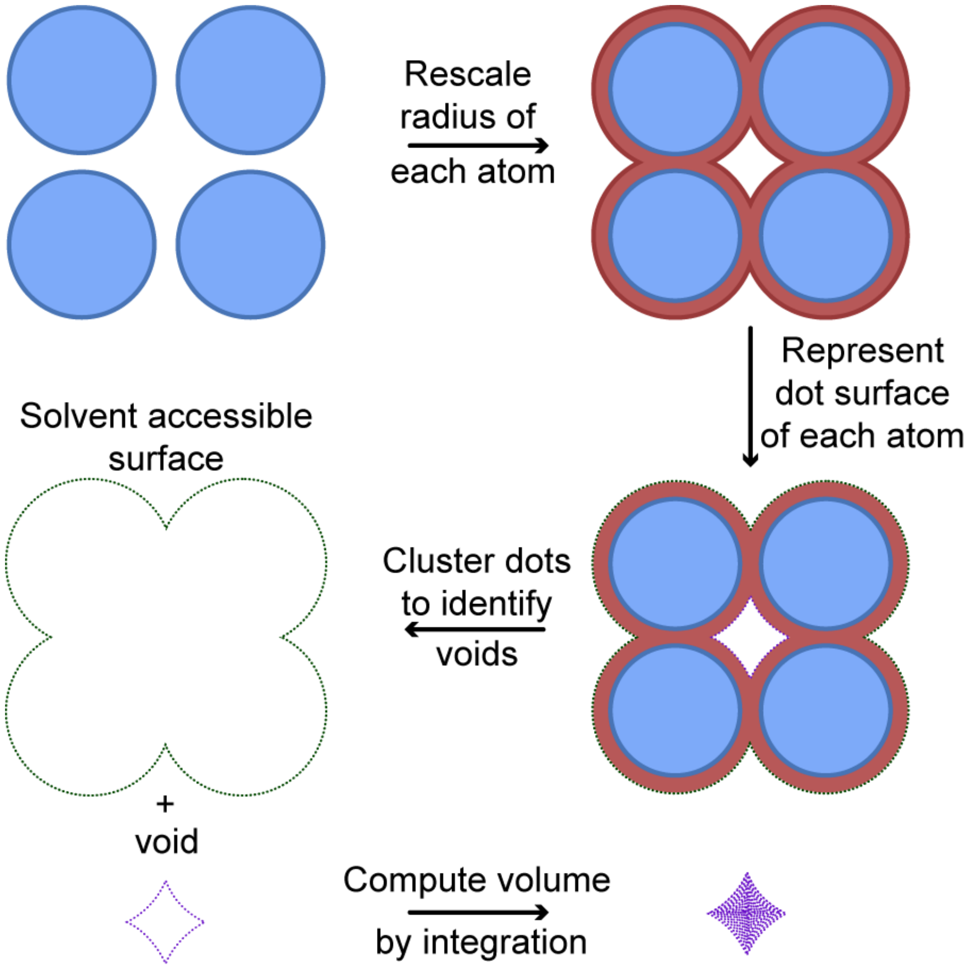
[2]Program in Molecular and Cellular Biophysics

[3]Center for Computational and Systems Biology, University of North Carolina at Chapel Hill, NC 27599-7260 USA

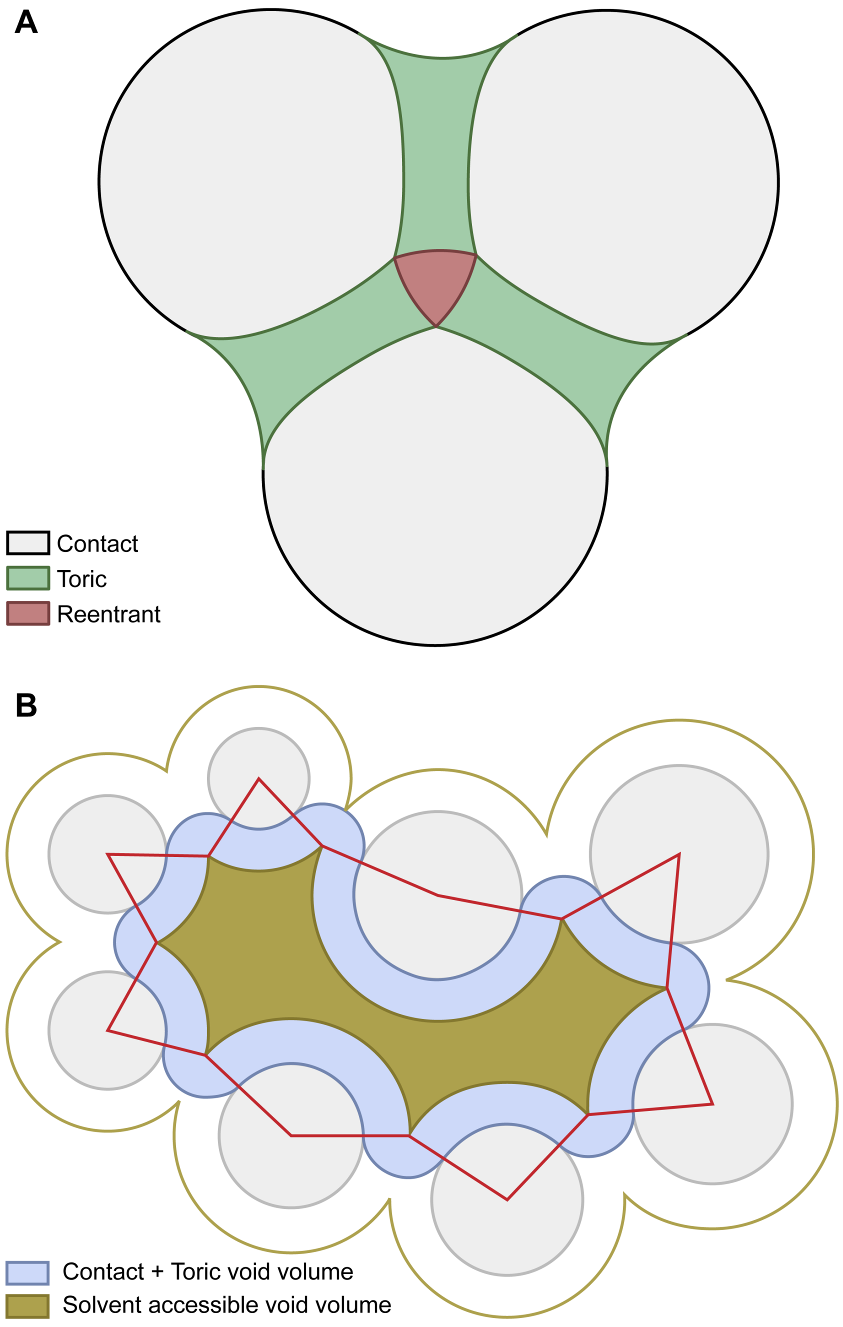
## SUPPLEMENTARY INFORMATION



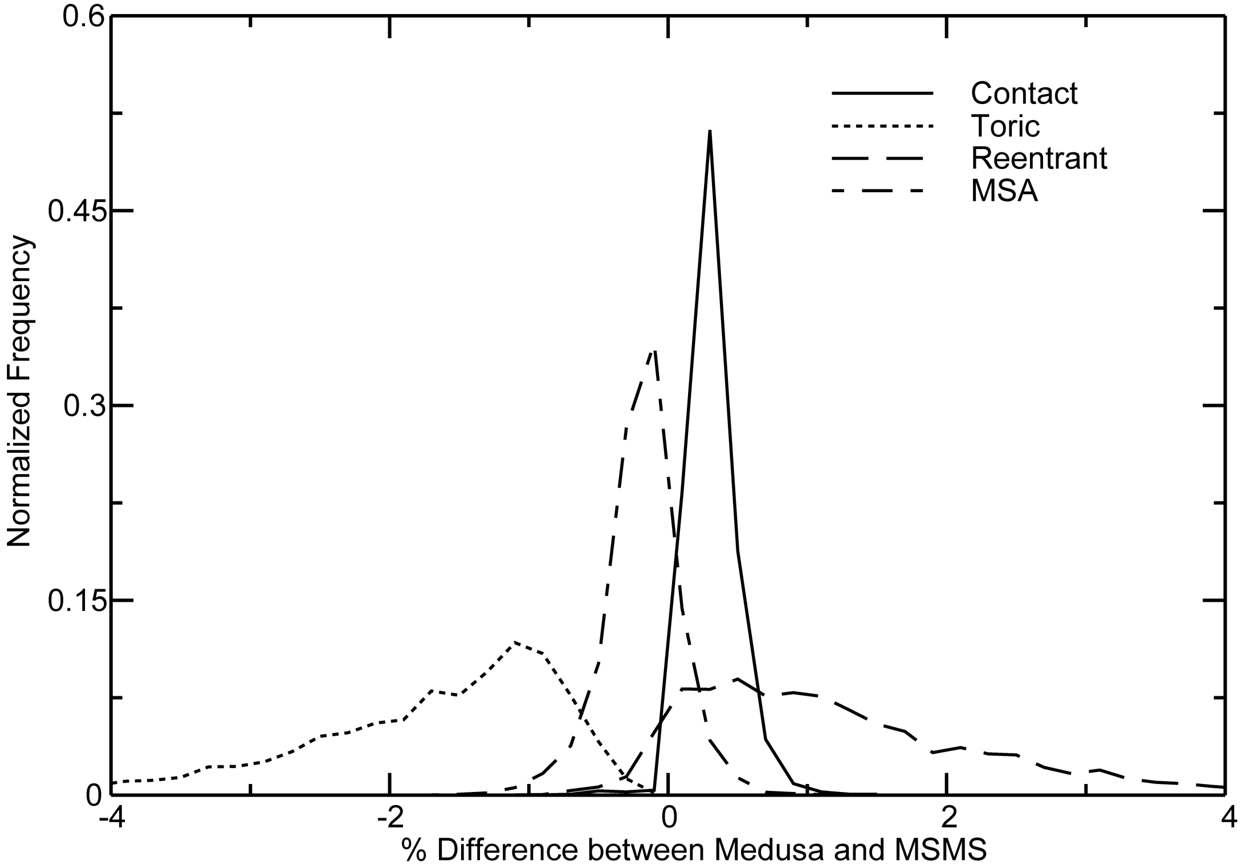
**Supplementary Figure S1. Parameters used for surface area and void calculations.** For any two overlapping atoms *i* and *j*, various distances and angles are represented in the figure. These parameters are used in different equations shown in the main text for computation of MSA/SASA and void volume.


Rescale radius of each atom
Represent dot surface of each atom
Cluster dots to identify voids
Solvent accessible surface
+
void
Compute volume by integration

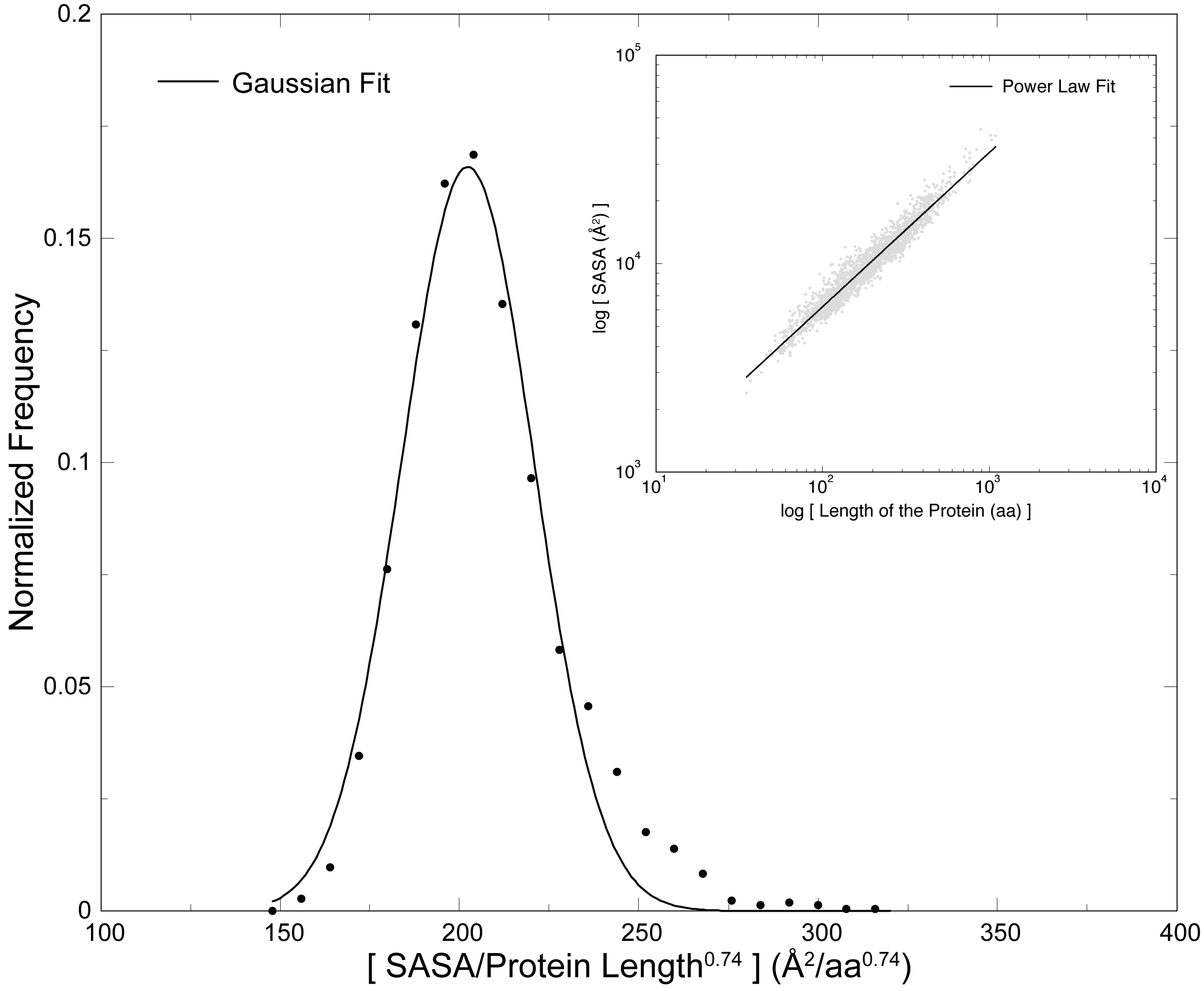
**Supplementary Figure S2. Steps in calculating void volume.** The algorithm used in void-volume calculation in proteins is illustrated. First, the radius of each atom is rescaled by 1.4 Å, the radius of a water molecule. Then, using dot-surface representation and single-linkage clustering, the surface and voids are separated. The volume of each void is then calculated by numerical integration. This volume corresponds to the solvent accessible void volume.



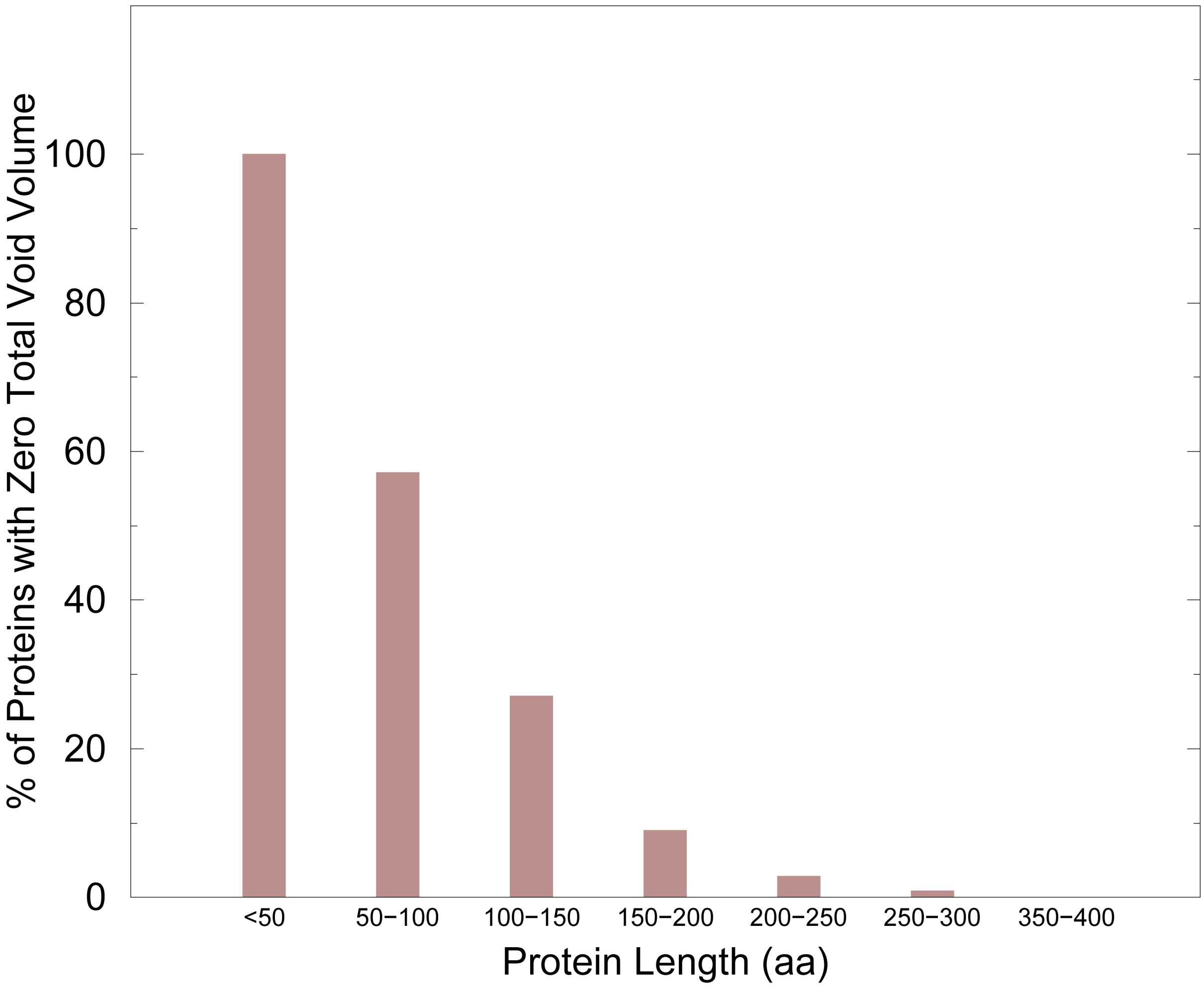
**Supplementary Figure S3. Cartoon representation of different surface components in a protein. A.** A simple case of three atoms overlapping with one another. The area of atoms shown in gray corresponds to the contact surface. Toric and reentrant surfaces are shown in green and red respectively. **B.** A typical void is depicted by the atoms shown in gray. The contact and toric components of void volume are colored blue. Solvent accessible void volume computed by numerical integration is shown in green. The red lines emanating from the center of each atom separate the contact void volume from the toric component. Reentrant void volume is not shown in this figure.



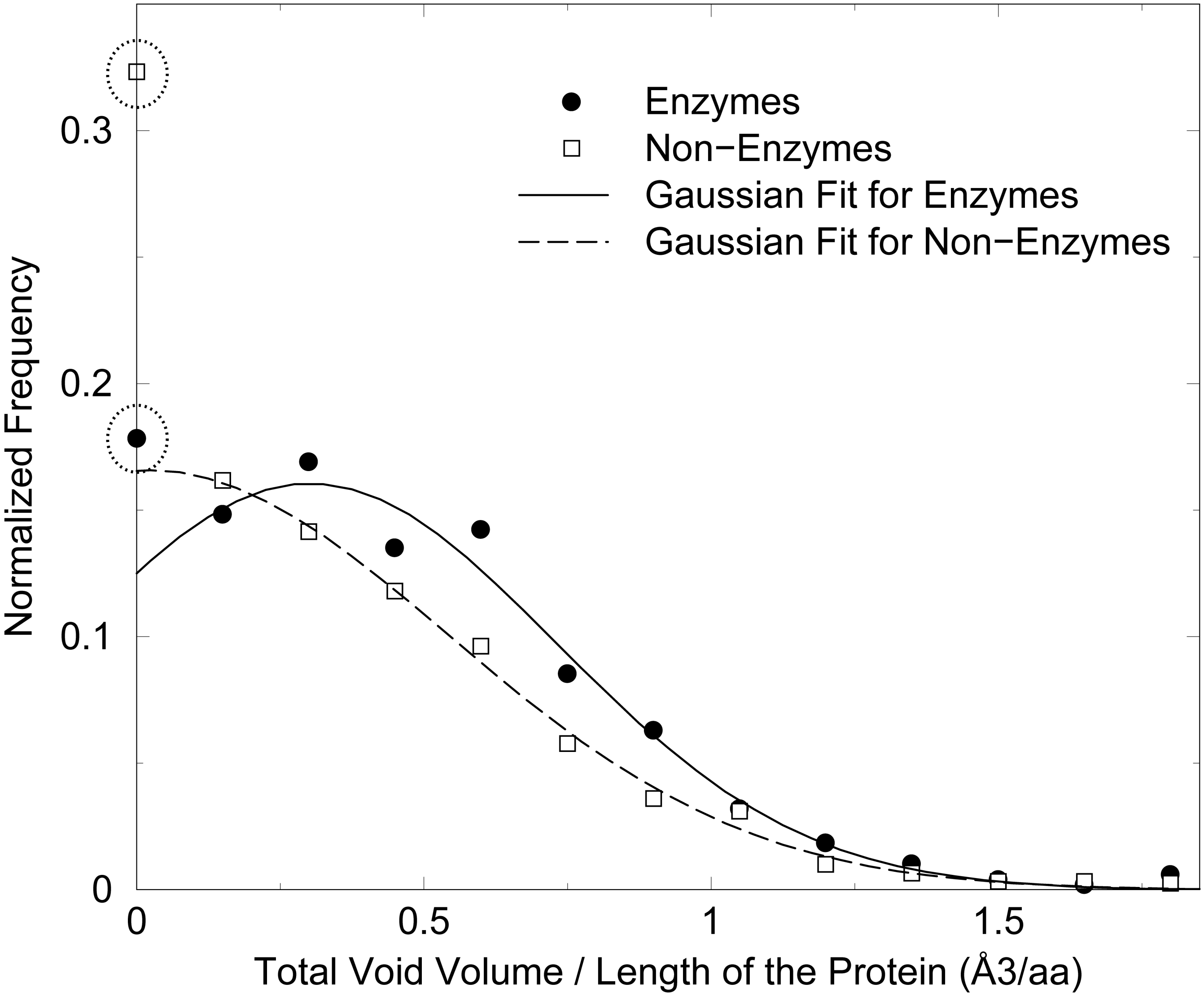
**Supplementary Figure S4. Comparison between MSMS and the current study.** The distribution of the percentage of difference between the areas of different components of molecular surface obtained from all the structures in our high-resolution dataset is plotted.



**Supplementary Figure S5. Scaling of Solvent Accessible Surface Area (SASA) as a function of protein size.** SASA scales with size of the protein as $(\text{length})^{0.74}$. Upon normalization of SASA by $(\text{length})^{0.74}$, histogram of SASA from all structures fits to a Gaussian distribution ($\mu$=202.3, $\sigma$=26.05). The raw plot of SASA vs the protein length is shown as gray points and the power-law fit is shown as a black line (inset).



**Supplementary Figure S6. Percentage of proteins featuring zero void volume, as a function of protein size.** We observe that most of the small proteins feature no voids, while the percentage of proteins possessing voids increases steadily with size.



**Supplementary Figure S7. Distribution of voids in enzymes and non-enzymes.** The distributions of total void volume of enzymes and non-enzymes per residue fit well to Gaussian distributions with the exception of a total void volume corresponding to zero in non-enzymes(dotted circle). The distribution of enzymes (μ=0.38, σ=0.6) is right-shifted compared to that of non-enzymes (μ=0.09, σ=0.74).