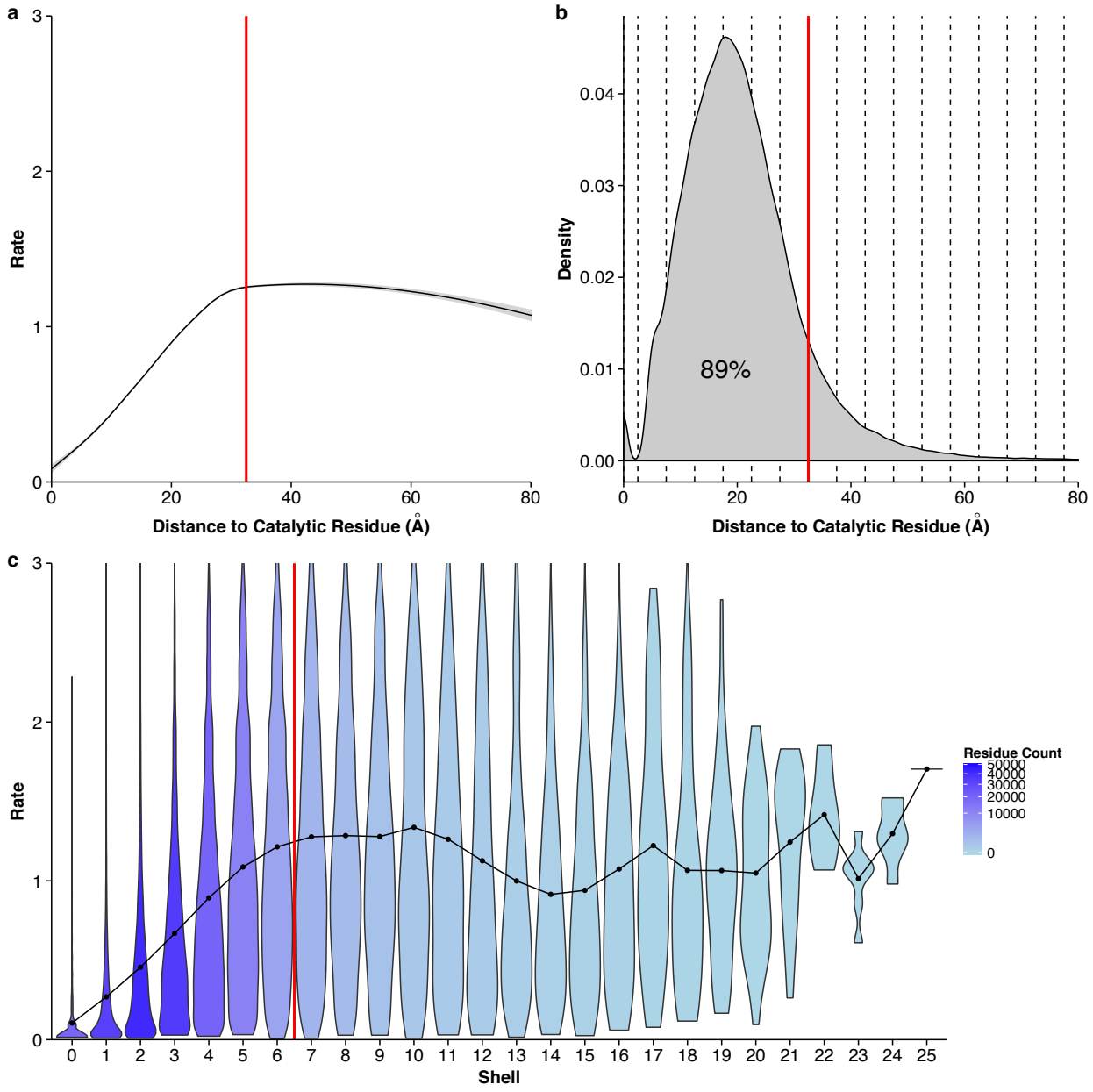
Supplementary Text 1

The following figures were generated using single subunits with interface residues included.

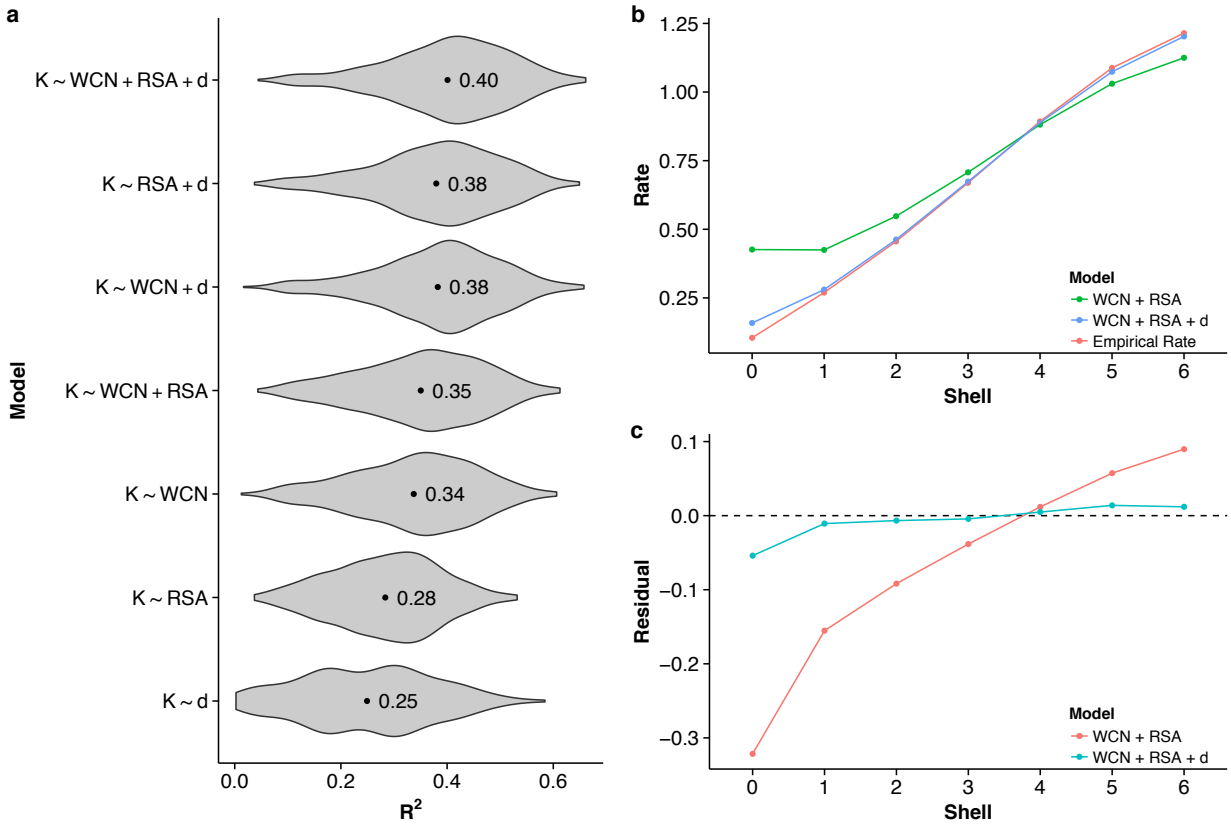
Figure 1



Site-specific evolutionary rates increase with distance to a catalytic residue. Only data from single subunits with interface residues included are considered. (a) Site-specific evolutionary rate versus distance to the nearest catalytic residue. The curve represents the mean rate for all residues, smoothed with a generalized additive model. Standard error is shown in light gray. Individual residues are not shown. The vertical line indicates a distance of 32.5 Å. (b) Density plot of residues with a given distance to a catalytic residue. Dashed lines indicate the boundaries of shells with a thickness of 5 Å each. The vertical red line again shows

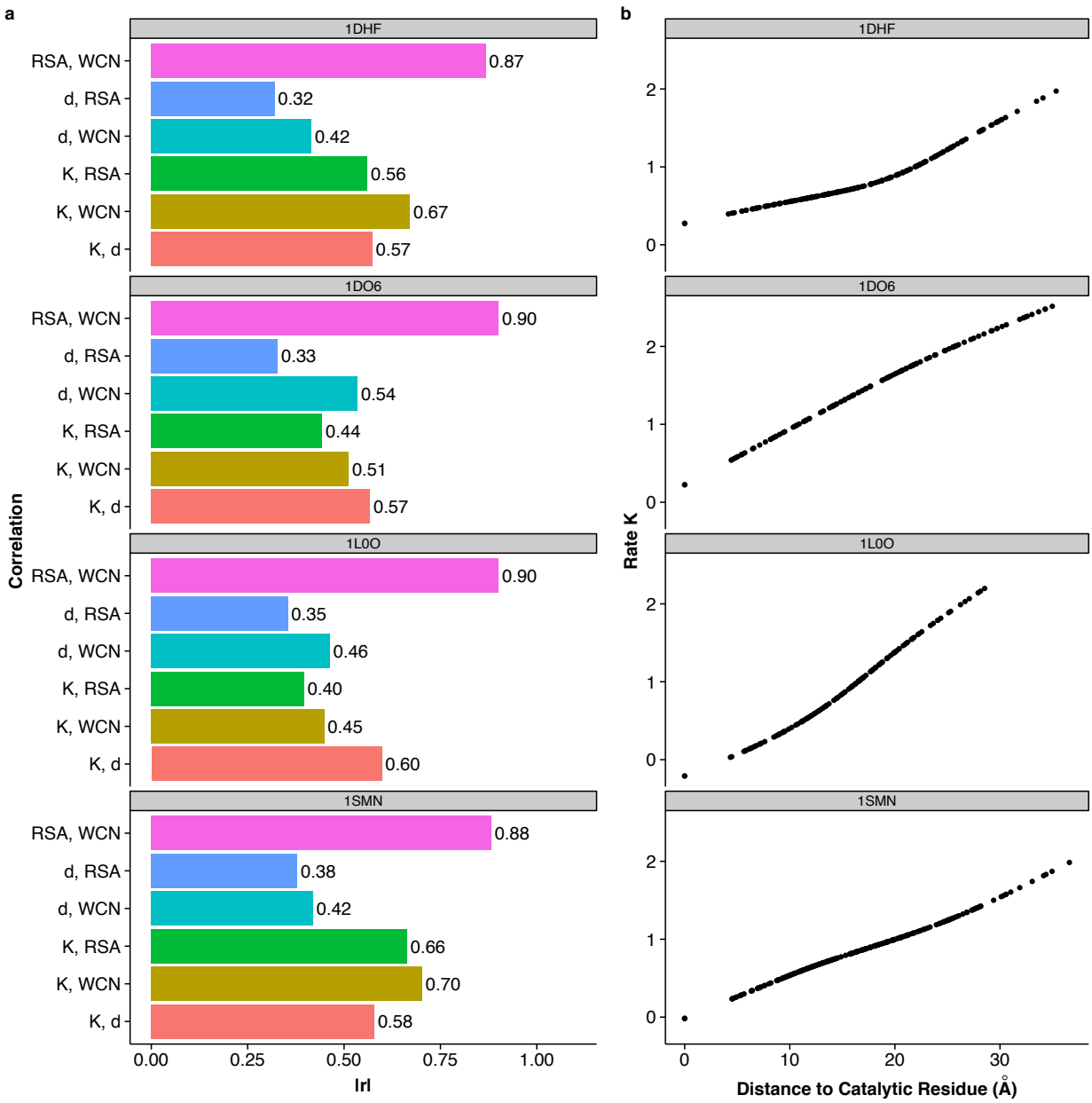
a distance of 32.5, and the percentage shows the proportion of residues within this distnace. (c) Distribution of site-specific evolutionary rates within each shell. Violins show the distribution of rates in each shell, points represent the mean, and the shading represents the number of residues found in that shell. The red line indicates a distance of 32.5 Å, corresponding to the boundary of shell 6.

Figure 2



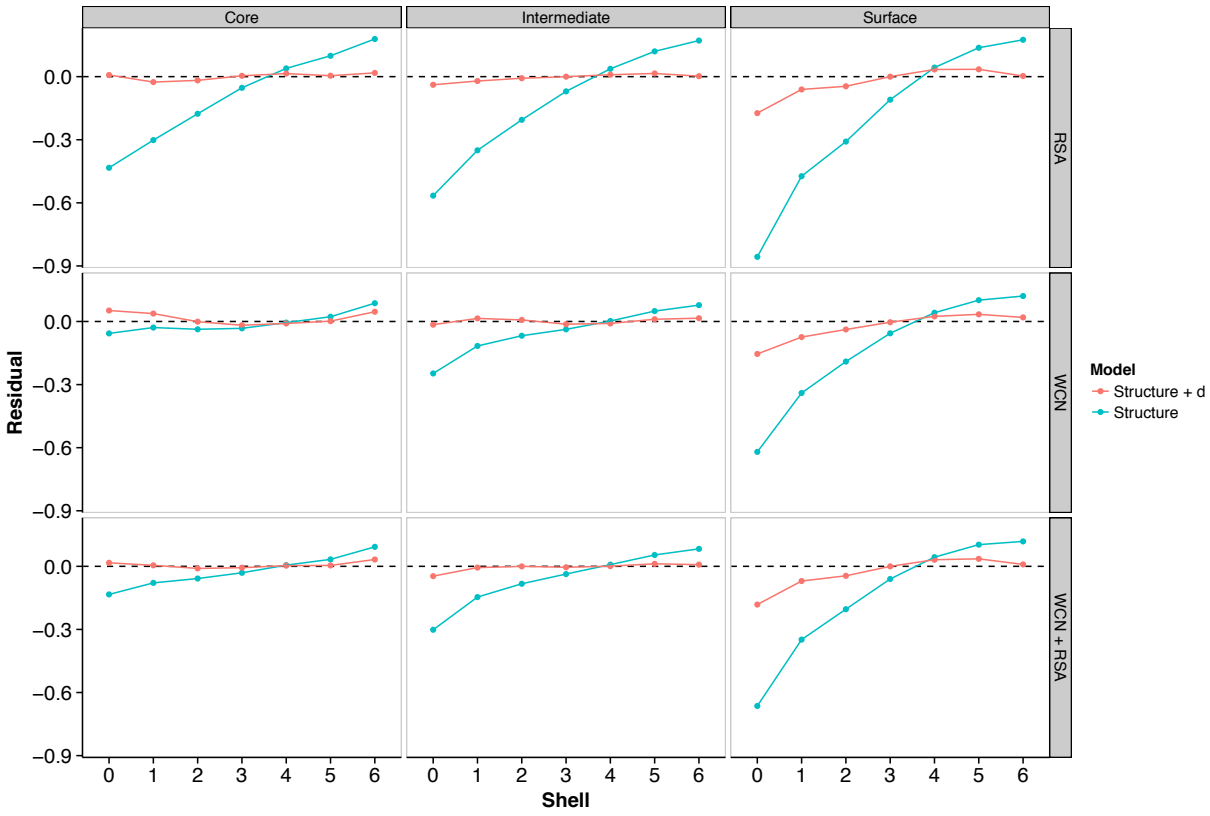
Relative performance of functional and structural predictors of rate. Only data from single subunits with interface residues included are considered. (a) Distribution of R^2 values for various linear models explaining rate, fitted individually for each protein. Labeled points within each distribution indicate the mean R^2 value across proteins. Here, K denotes the site-specific evolutionary rate, d the distance to the nearest catalytic residue, WCN the weighted contact number, and RSA the relative solvent accessibility. On average, the addition of distance to the structural constraints WCN and RSA increases the percent variance explained by the linear models by at least 5 percentage points. (b) Mean empirical and predicted rates, separated by shell. A model containing only WCN and RSA overestimates rates near the active site of the enzyme and the addition of distance corrects this behavior. (c) Mean residuals for linear models with and without distance as a parameter, separated by shell.}

Figure 3



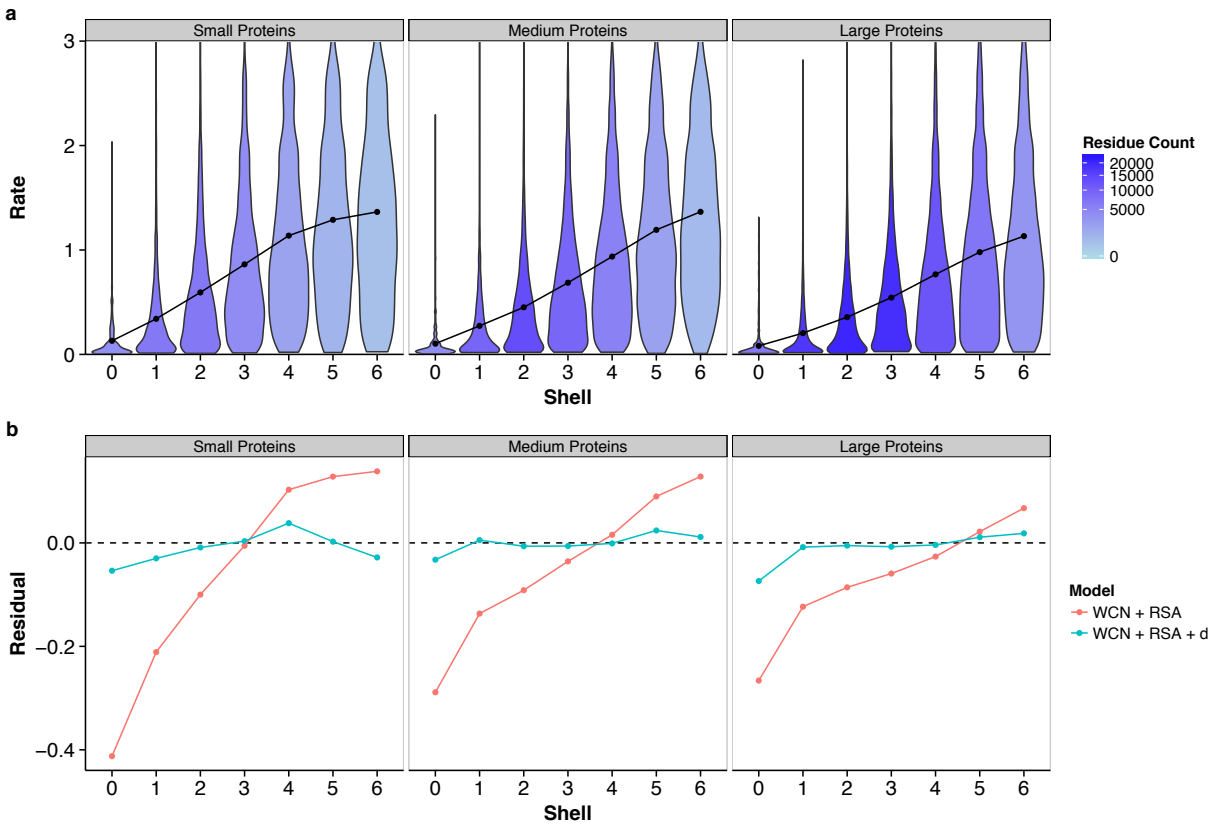
Example enzymes from main text. From top to bottom, the enzymes shown are 1DHF, 1DO6, 1L0O, and 1SMN. Only data from single subunits with interface residues included are considered. (a) Absolute Pearson correlations (r) between combinations of distance d, weighted contact number (WCN), relative solvent accessibility (RSA), and rate K. Distance correlates weakly with WCN or RSA, and strongly with rate, indicating that distance is a strong predictor independent of structural constraints RSA and WCN. (b) Rates smoothed with a GAM (generalized additive model) as a function of d, shown for every residue in each enzyme structure. Smoothed rates increase with distance throughout the entire enzyme structure.

Figure 4



Effect of active-site location on the relationship between site-specific evolutionary rates and distance to the nearest catalytic residue. Only data from single subunits with interface residues included are considered. Enzymes are grouped into three categories, based on the mean solvent exposure of their catalytic residues: core (RSA < 0.05), intermediate ($0.05 <$ RSA < 0.25), and surface (RSA > 0.25). Lines represent mean residuals for different structural linear models (vertical labels, right) and those same models with distance added as a parameter.

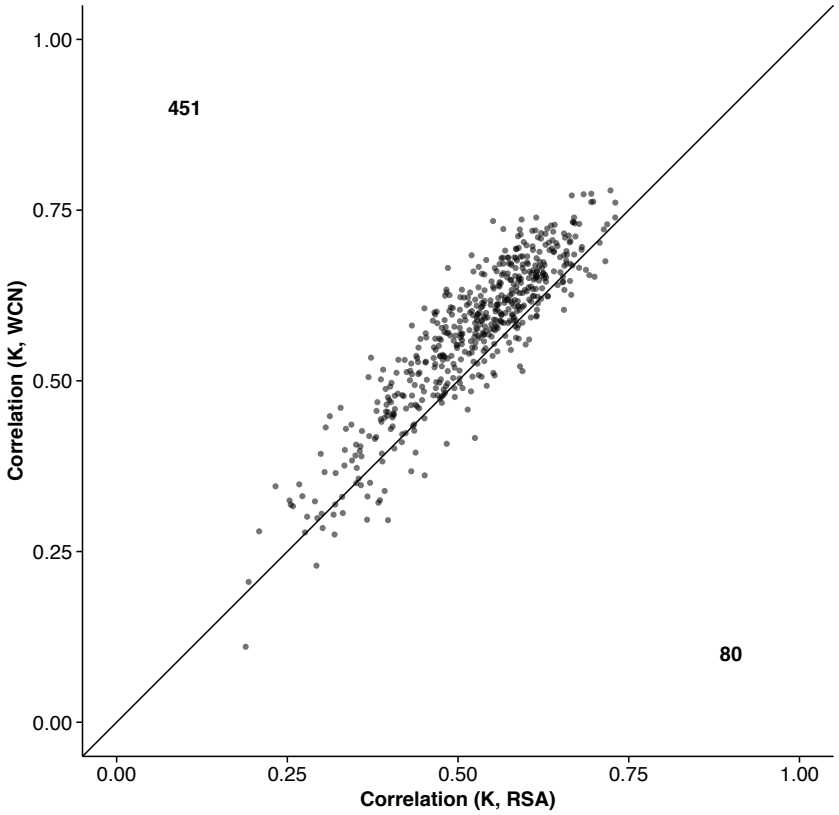
Figure 5



Effect of protein size on the relationship between site-specific evolutionary rates and distance to the nearest catalytic residue. Only data from single subunits with interface residues included are considered. (a) Distribution of rates within each shell, where proteins have been separated into small (95–326 sites), medium (327–484 sites), and large (485–1287 sites) based on amino-acid sequence length. Points represent the mean rate in each shell. As the size of the protein increases, the distance–rate slope decreases and distance effects extend further from the active site. (b) Mean residuals in each shell for models with and without distance, again separated by protein size. The constraining effects of catalytic residues depend on protein size, with stronger, more local effects in small proteins, and weaker, longer-range effects in large proteins.

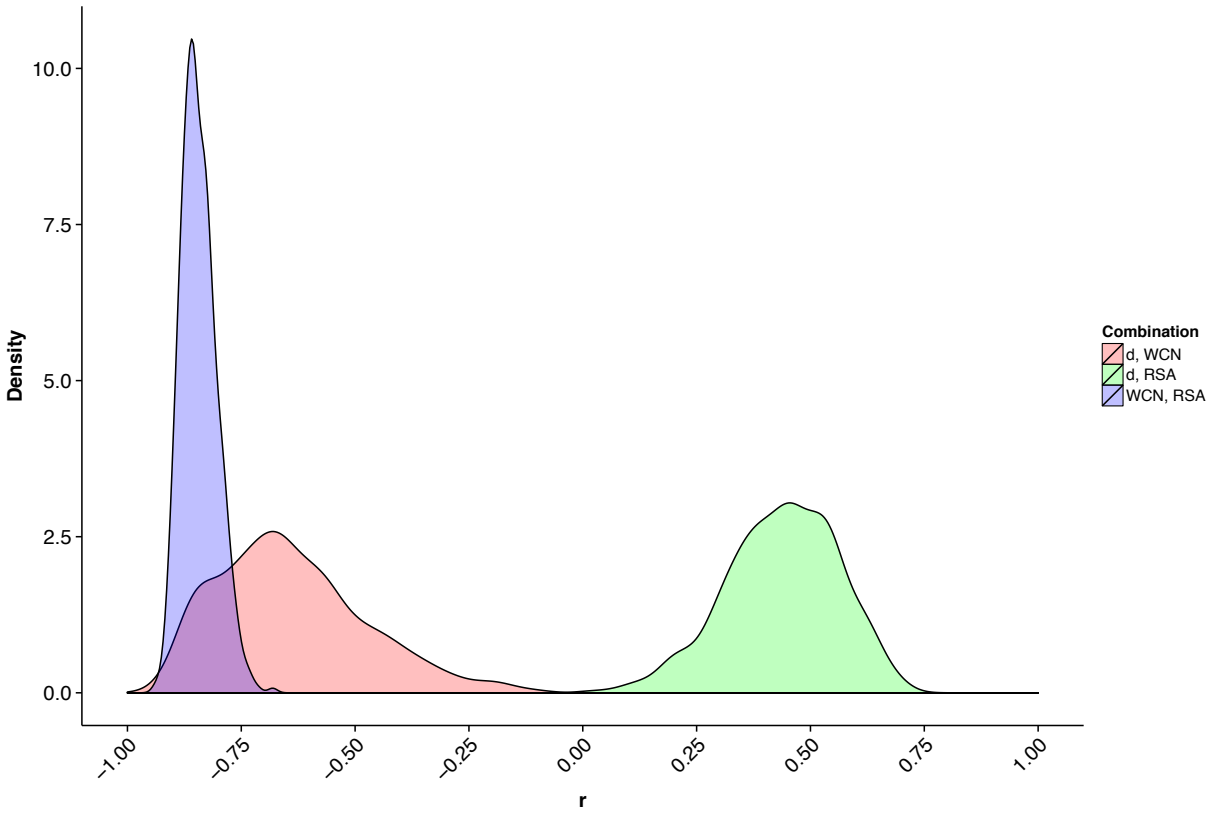
Supporting Figures

Figure S1



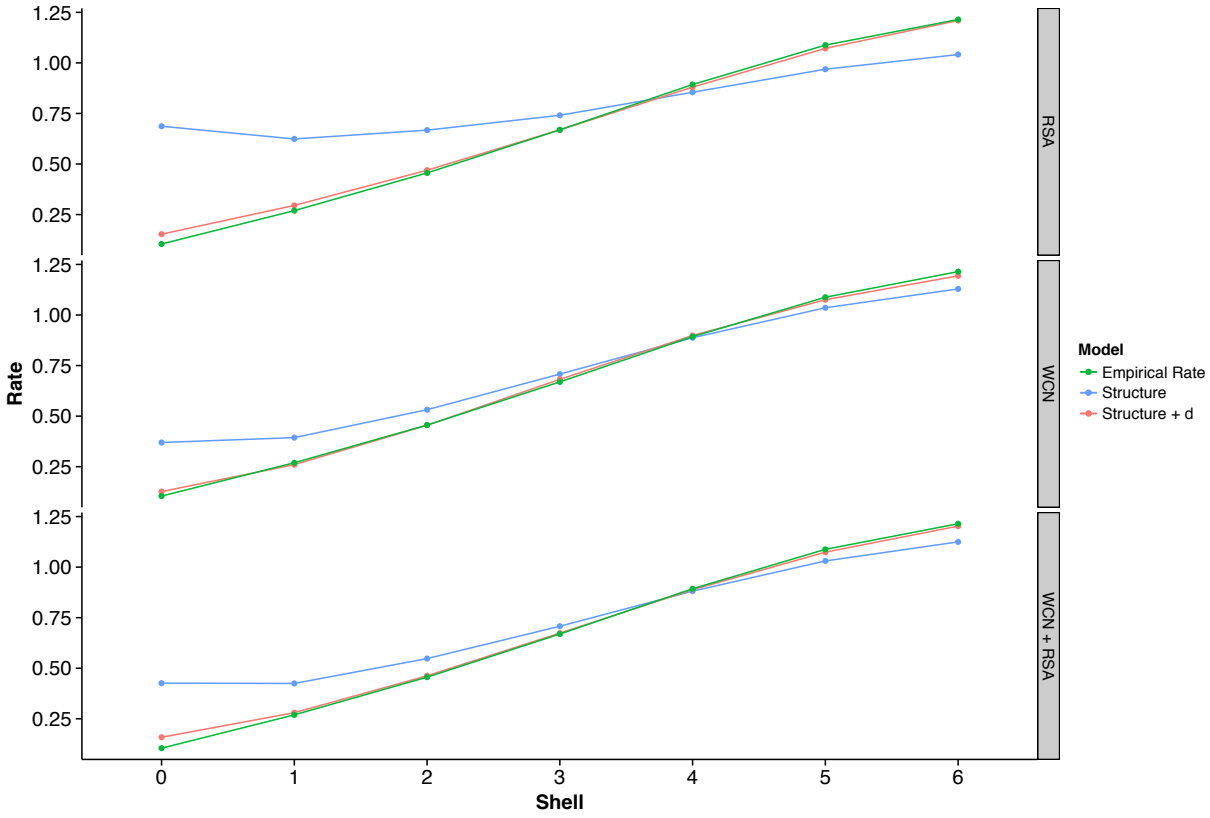
Pearson r values of side-chain WCN and RSA correlated with site-specific evolutionary rates. Only data from single subunits with interface residues included are considered. The sign of the correlation coefficients for WCN are switched from negative to positive for a simpler comparison with RSA. Each point corresponds to an individual protein, and the numbers refer to the number of proteins above or below the $y = x$ line. In aggregate, WCN is a better predictor of site-specific evolutionary rate than RSA.

Figure S2



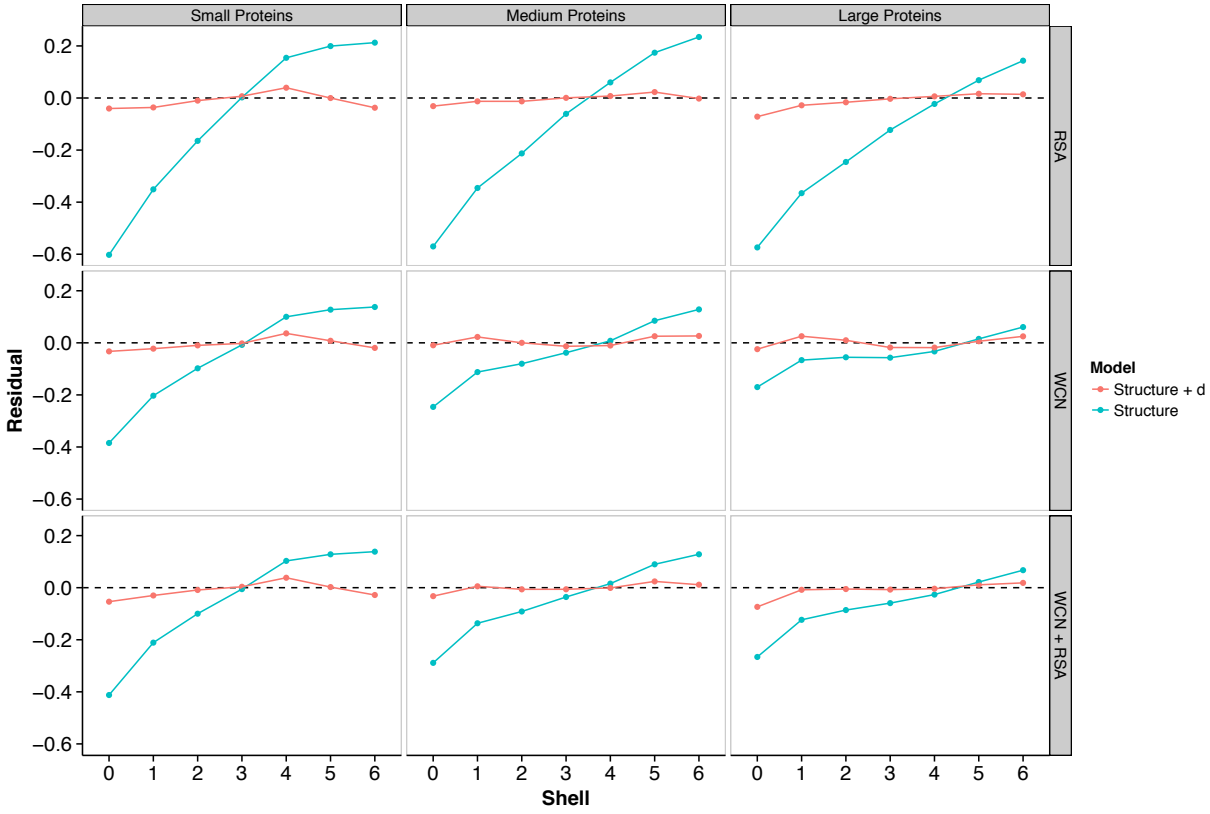
Distributions of Pearson r correlations between pairs of site-specific evolutionary rate predictors on a per protein basis. Only data from single subunits with interface residues included are considered. Predictor pairs include distance and WCN, distance and RSA, and WCN and RSA. WCN and RSA correlate more strongly with each other than either does with distance.

Figure S3



Empirical and predicted site-specific evolutionary rates from different combined linear models. Only data from single subunits with interface residues included are considered. "Structure" refers to one of three structural models used to predict rate: RSA, WCN, or RSA + WCN. "Structure + d" refers to these same three structural models, with the addition of distance to the nearest catalytic residue as a variable. Each point is the mean predicted rate for a given shell across all residues in the data set. In all cases, models that include distance as a parameter predict rate more accurately than models containing only structural parameters, especially near the active site.

Figure S4

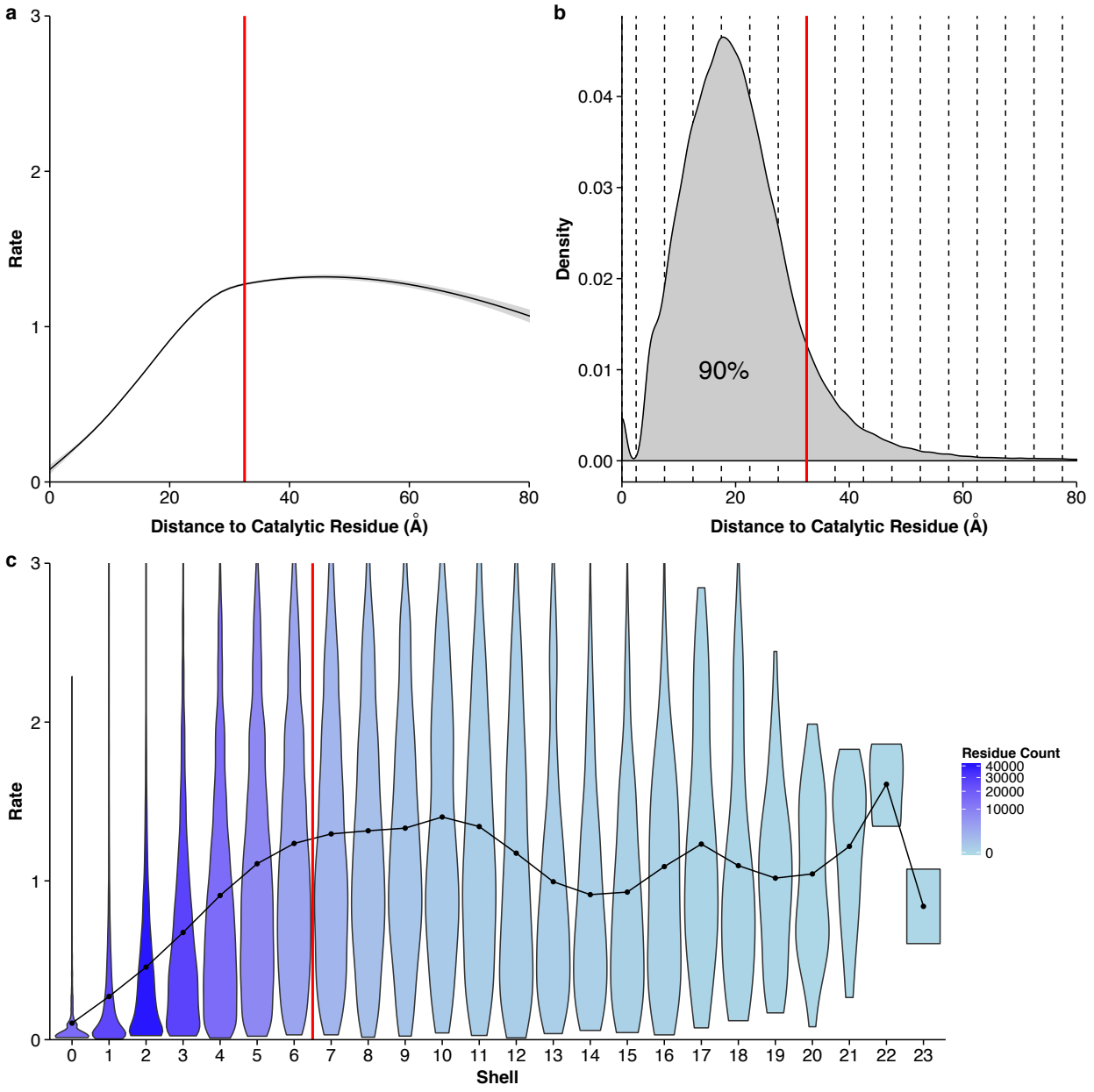


Mean residuals for different combined linear models used to predict site-specific evolutionary rate. Only data from single subunits with interface residues included are considered. The data set is divided into small (95–326 sites), medium (327–484 sites), and large (485–1287 sites) proteins. Each point represents the mean predicted rate for all residues in a given shell.

Supplementary Text 2

The following figures were generated using single subunits with interface residues removed.

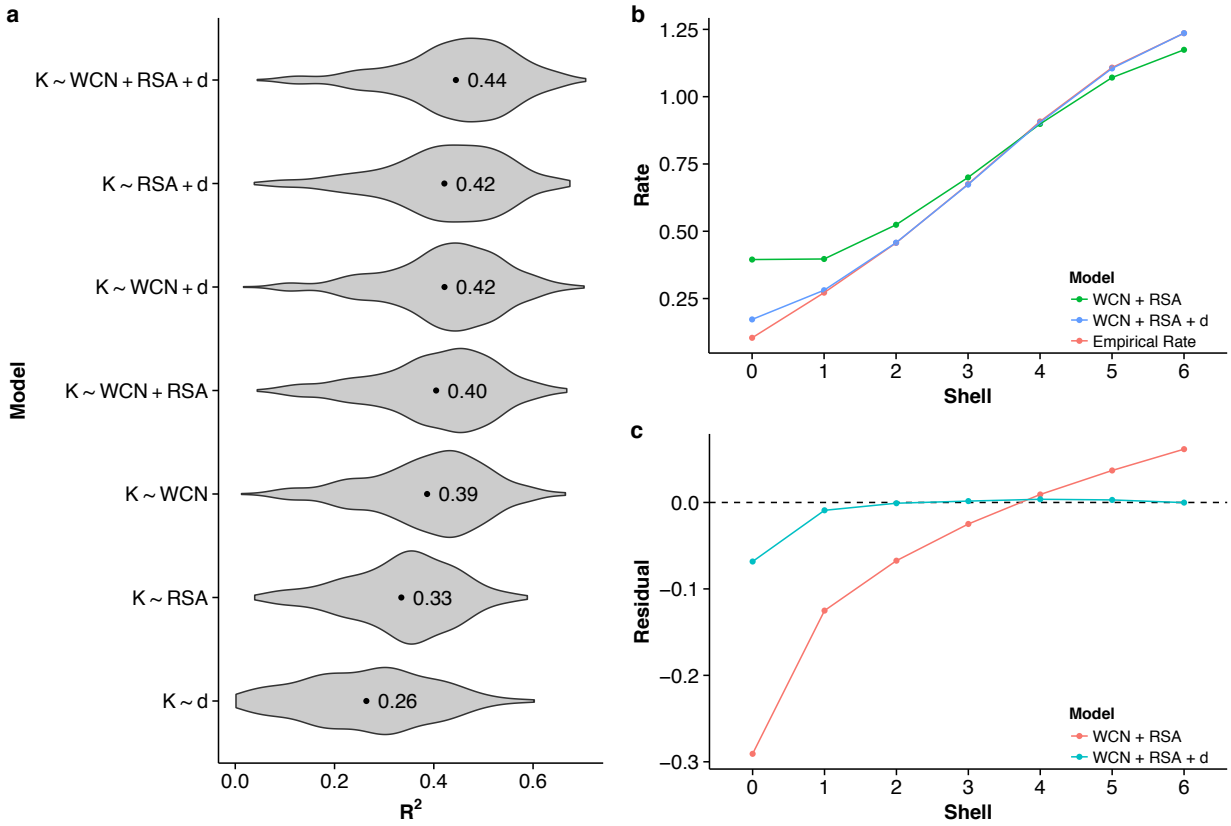
Figure 1



Site-specific evolutionary rates increase with distance to a catalytic residue. Only data from single subunits with interface residues removed are considered. (a) Site-specific evolutionary rate versus distance to the nearest catalytic residue. The curve represents the mean rate for all residues, smoothed with a generalized additive model. Standard error is shown in light gray. Individual residues are not shown. The vertical line indicates a distance of 32.5 Å. (b) Density plot of residues with a given distance to a catalytic residue. Dashed lines indicate the boundaries of shells with a thickness of 5 Å each. The vertical red line again shows

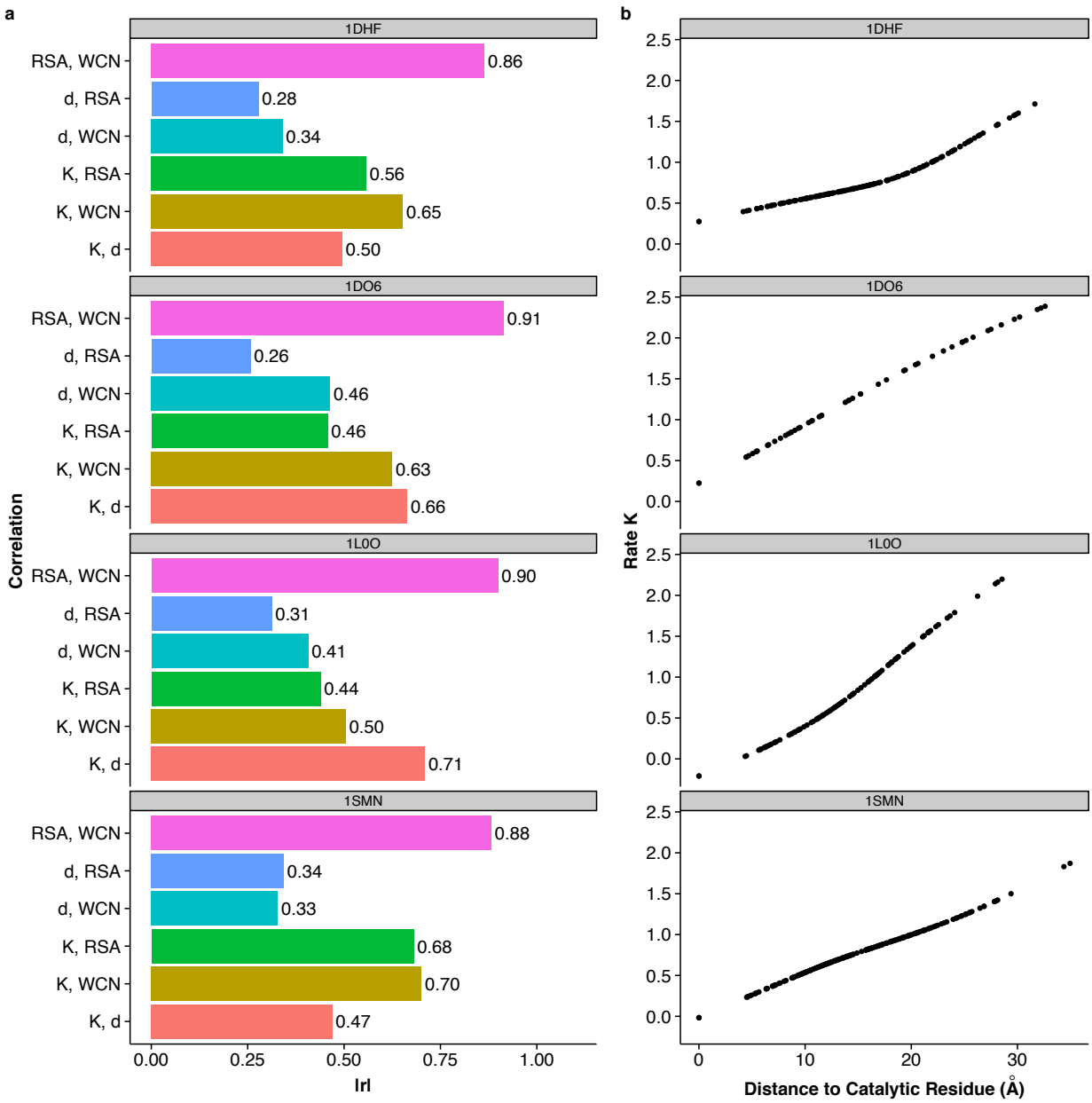
a distance of 32.5, and the percentage shows the proportion of residues within this distnace. (c) Distribution of site-specific evolutionary rates within each shell. Violins show the distribution of rates in each shell, points represent the mean, and the shading represents the number of residues found in that shell. The red line indicates a distance of 32.5 Å, corresponding to the boundary of shell 6.

Figure 2



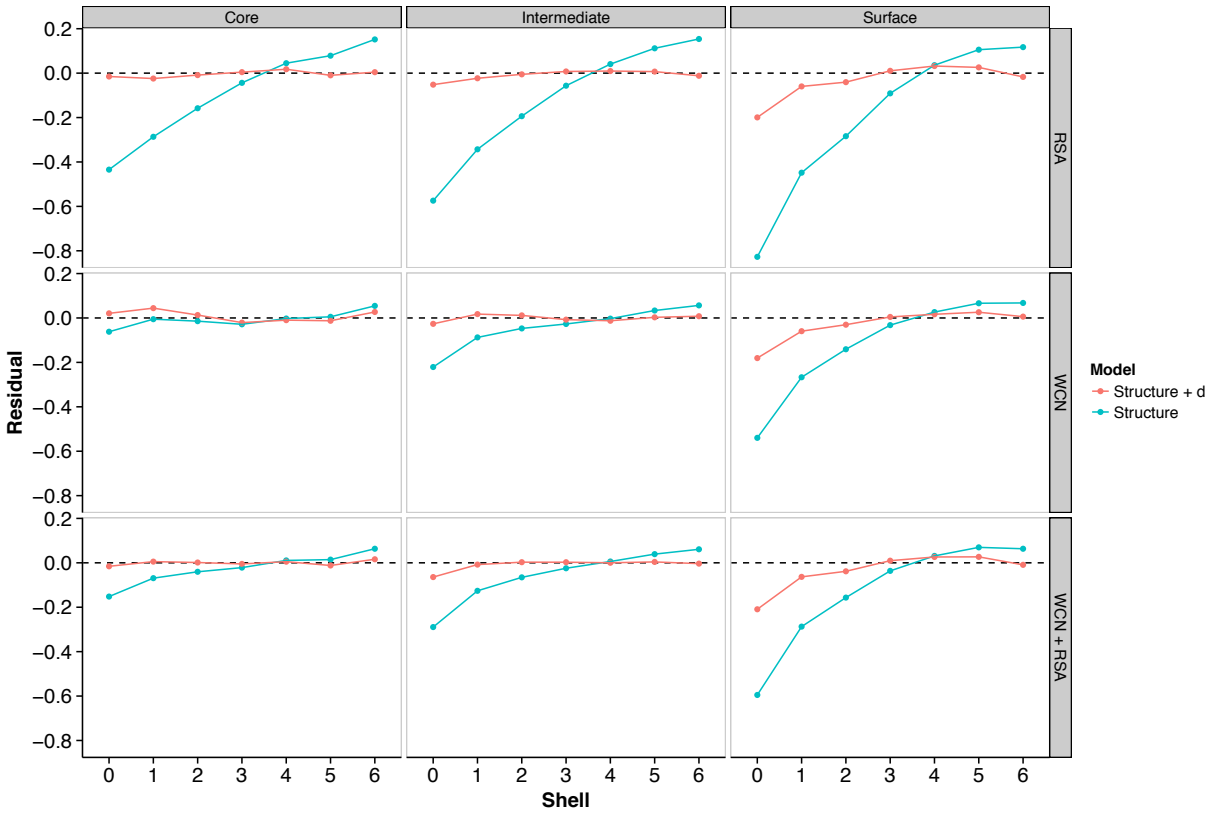
Relative performance of functional and structural predictors of rate. Only data from single subunits with interface residues removed are considered. (a) Distribution of R^2 values for various linear models explaining rate, fitted individually for each protein. Labeled points within each distribution indicate the mean R^2 value across proteins. Here, K denotes the site-specific evolutionary rate, d the distance to the nearest catalytic residue, WCN the weighted contact number, and RSA the relative solvent accessibility. On average, the addition of distance to the structural constraints WCN and RSA increases the percent variance explained by the linear models by at least 5 percentage points. (b) Mean empirical and predicted rates, separated by shell. A model containing only WCN and RSA overestimates rates near the active site of the enzyme and the addition of distance corrects this behavior. (c) Mean residuals for linear models with and without distance as a parameter, separated by shell.}

Figure 3



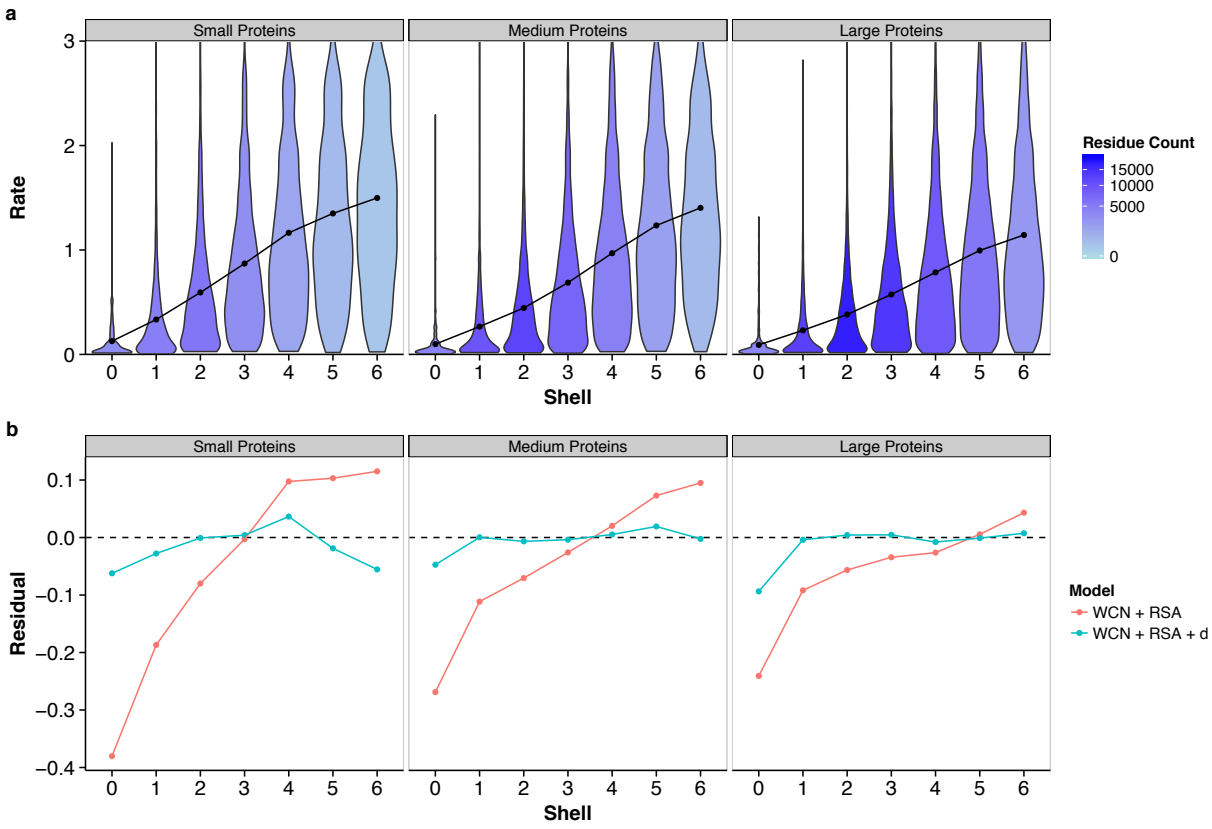
Example enzymes from main text. From top to bottom, the enzymes shown are 1DHF, 1DO6, 1L0O, and 1SMN. Only data from single subunits with interface residues removed are considered. (a) Absolute Pearson correlations (r) between combinations of distance d, weighted contact number (WCN), relative solvent accessibility (RSA), and rate K. Distance correlates weakly with WCN or RSA, and strongly with rate, indicating that distance is a strong predictor independent of structural constraints RSA and WCN. (b) Rates smoothed with a GAM (generalized additive model) as a function of d, shown for every residue in each enzyme structure. Smoothed rates increase with distance throughout the entire enzyme structure.

Figure 4



Effect of active-site location on the relationship between site-specific evolutionary rates and distance to the nearest catalytic residue. Only data from single subunits with interface residues removed are considered. Enzymes are grouped into three categories, based on the mean solvent exposure of their catalytic residues: core (RSA < 0.05), intermediate ($0.05 <$ RSA < 0.25), and surface (RSA > 0.25). Lines represent mean residuals for different structural linear models (vertical labels, right) and those same models with distance added as a parameter.

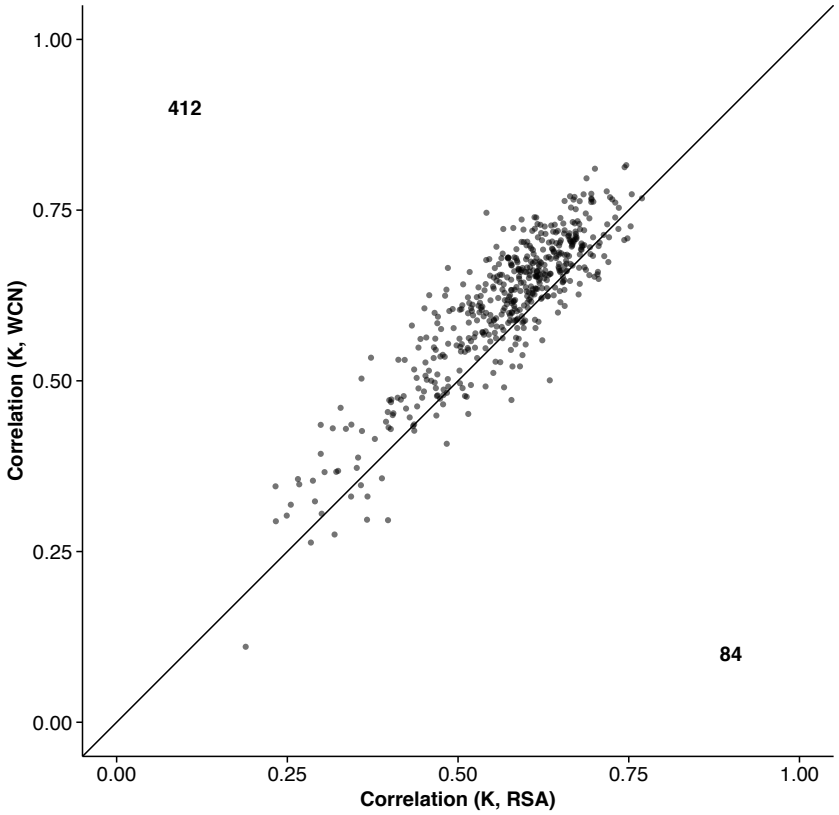
Figure 5



Effect of protein size on the relationship between site-specific evolutionary rates and distance to the nearest catalytic residue. Only data from single subunits with interface residues removed are considered. (a) Distribution of rates within each shell, where proteins have been separated into small (95–326 sites), medium (327–484 sites), and large (485–1287 sites) based on amino-acid sequence length. Points represent the mean rate in each shell. As the size of the protein increases, the distance–rate slope decreases and distance effects extend further from the active site. (b) Mean residuals in each shell for models with and without distance, again separated by protein size. The constraining effects of catalytic residues depend on protein size, with stronger, more local effects in small proteins, and weaker, longer-range effects in large proteins.

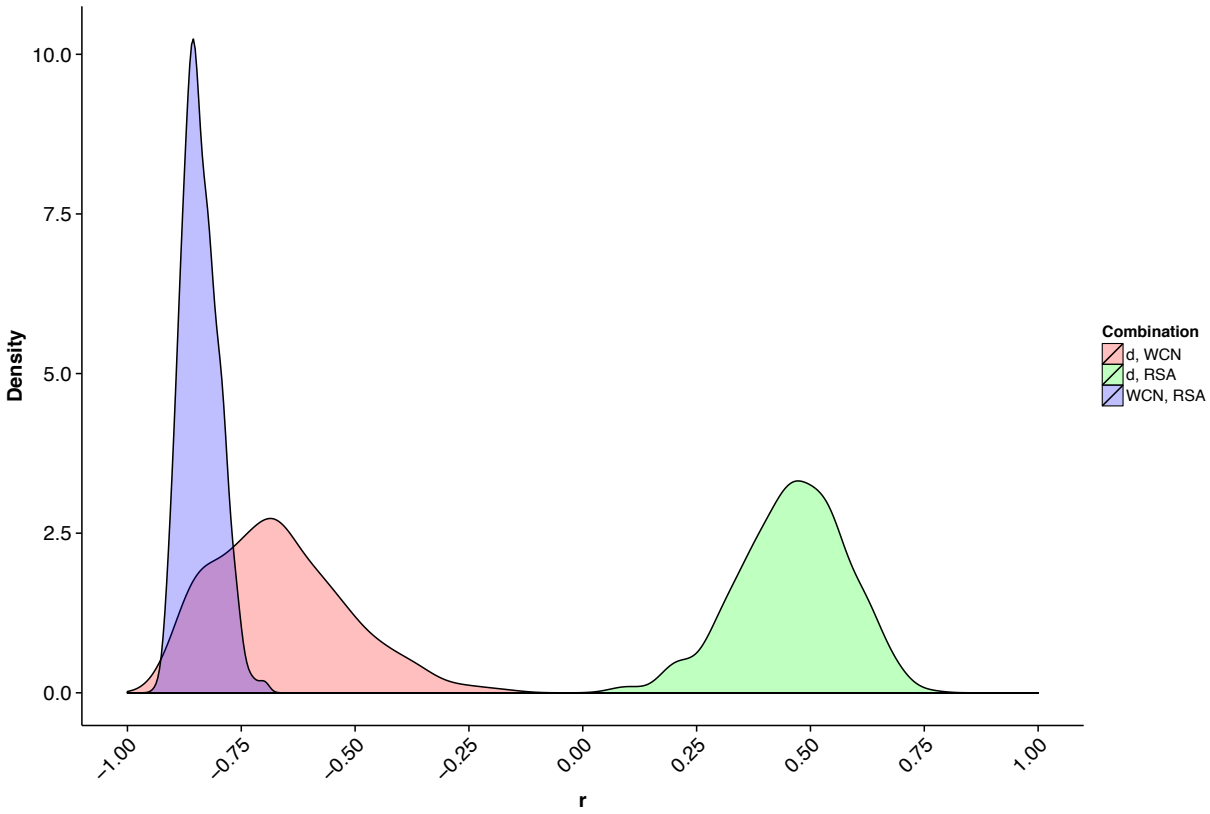
Supporting Figures

Figure S1



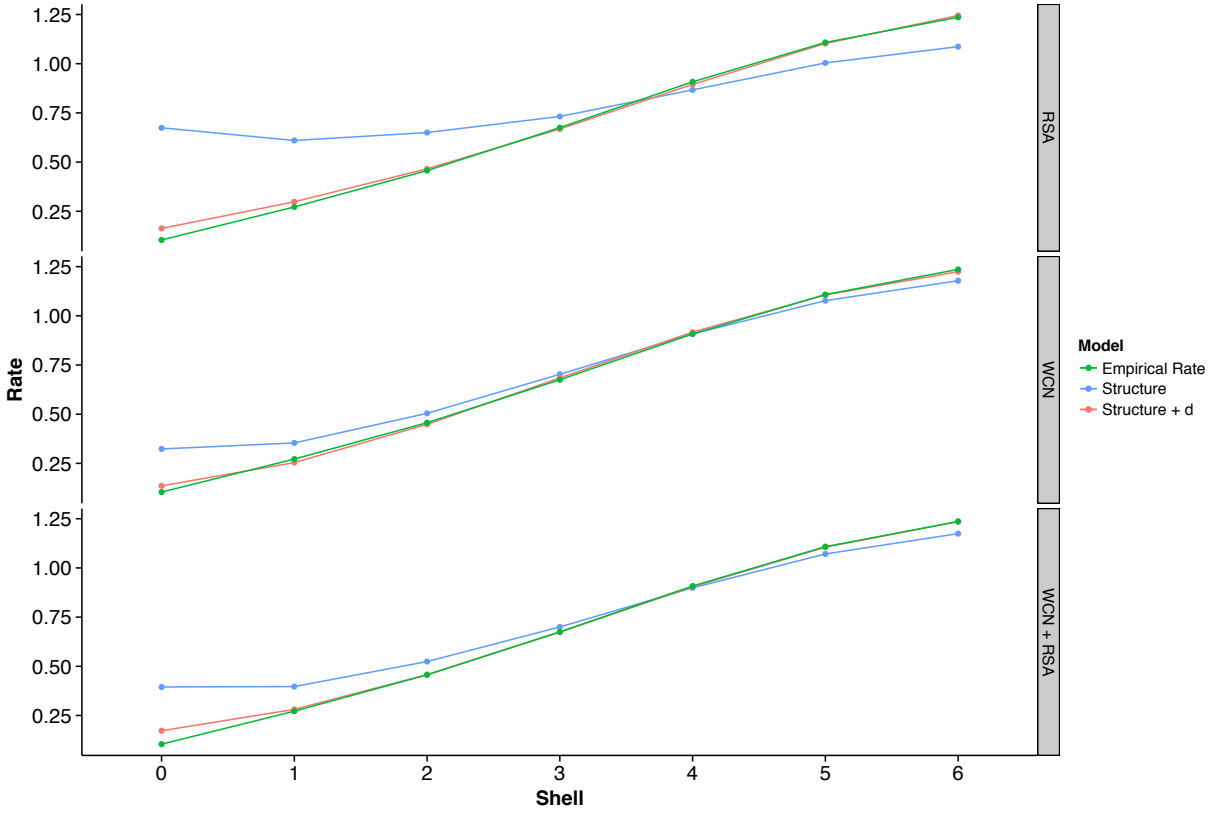
Pearson r values of side-chain WCN and RSA correlated with site-specific evolutionary rates. Only data from single subunits with interface residues removed are considered. The sign of the correlation coefficients for WCN are switched from negative to positive for a simpler comparison with RSA. Each point corresponds to an individual protein, and the numbers refer to the number of proteins above or below the $y = x$ line. In aggregate, WCN is a better predictor of site-specific evolutionary rate than RSA.

Figure S2



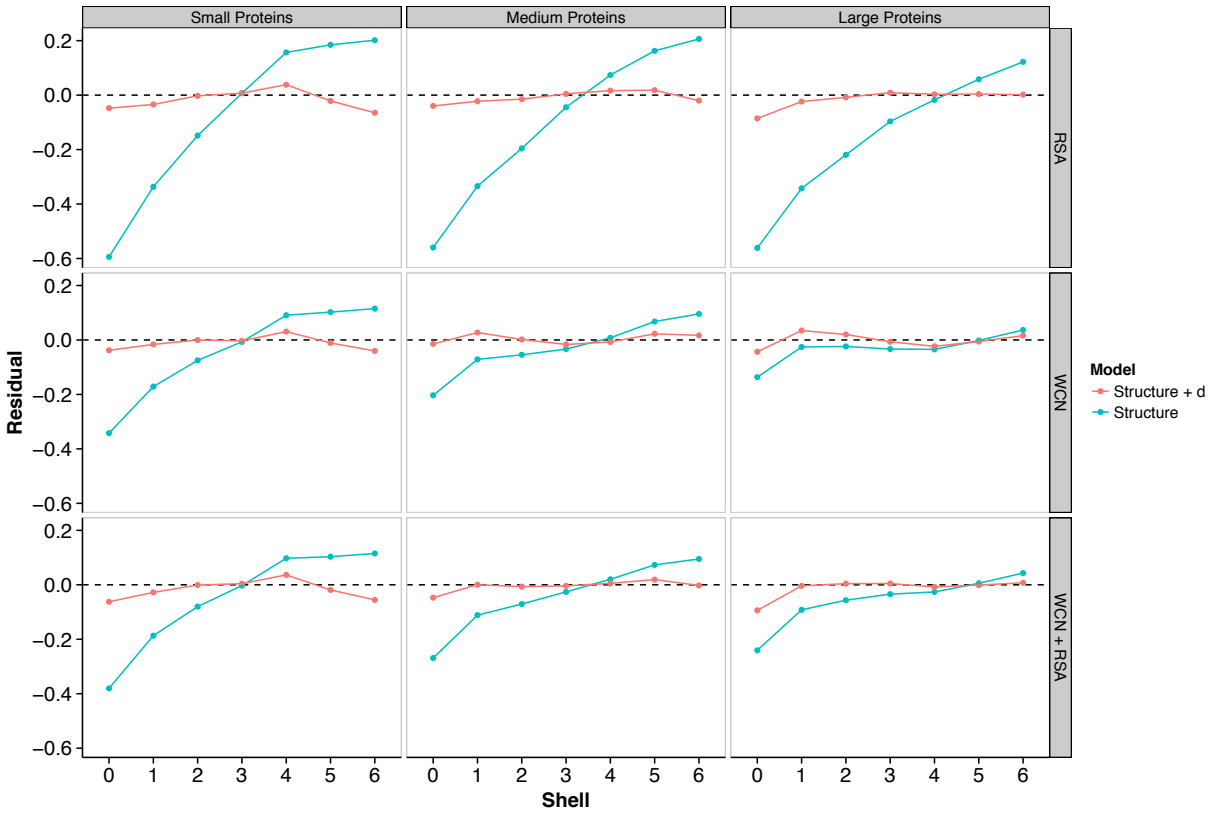
Distributions of Pearson r correlations between pairs of site-specific evolutionary rate predictors on a per protein basis. Only data from single subunits with interface residues removed are considered. Predictor pairs include distance and WCN, distance and RSA, and WCN and RSA. WCN and RSA correlate more strongly with each other than either does with distance.

Figure S3



Empirical and predicted site-specific evolutionary rates from different combined linear models. Only data from single subunits with interface residues removed are considered. "Structure" refers to one of three structural models used to predict rate: RSA, WCN, or RSA + WCN. "Structure + d" refers to these same three structural models, with the addition of distance to the nearest catalytic residue as a variable. Each point is the mean predicted rate for a given shell across all residues in the data set. In all cases, models that include distance as a parameter predict rate more accurately than models containing only structural parameters, especially near the active site.

Figure S4

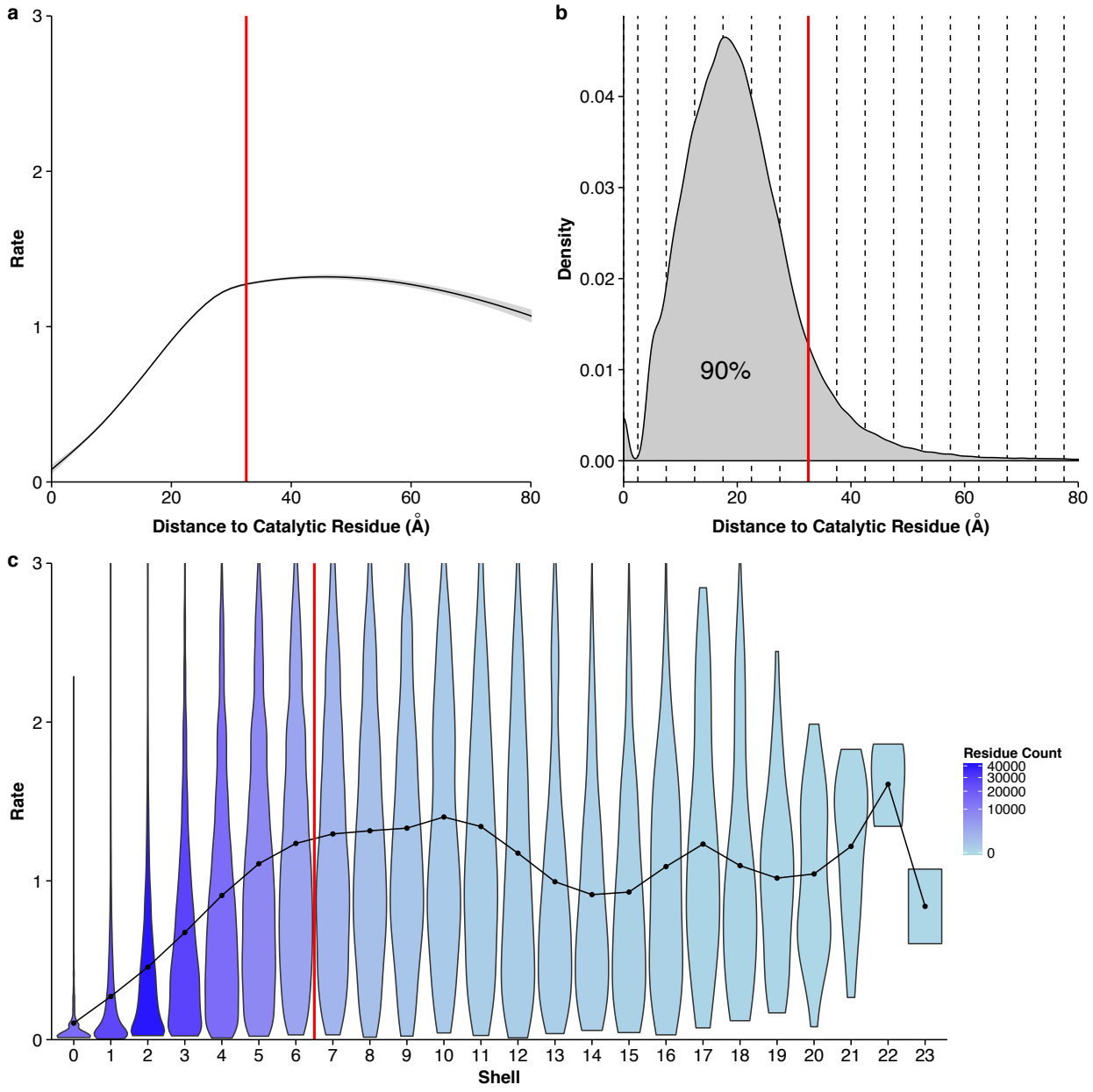


Mean residuals for different combined linear models used to predict site-specific evolutionary rate. Only data from single subunits with interface residues removed are considered. The data set is divided into small (95–326 sites), medium (327–484 sites), and large (485–1287 sites) proteins. Each point represents the mean predicted rate for all residues in a given shell.

Supplementary Text 3

The following figures were generated using biological assemblies with interface residues removed.

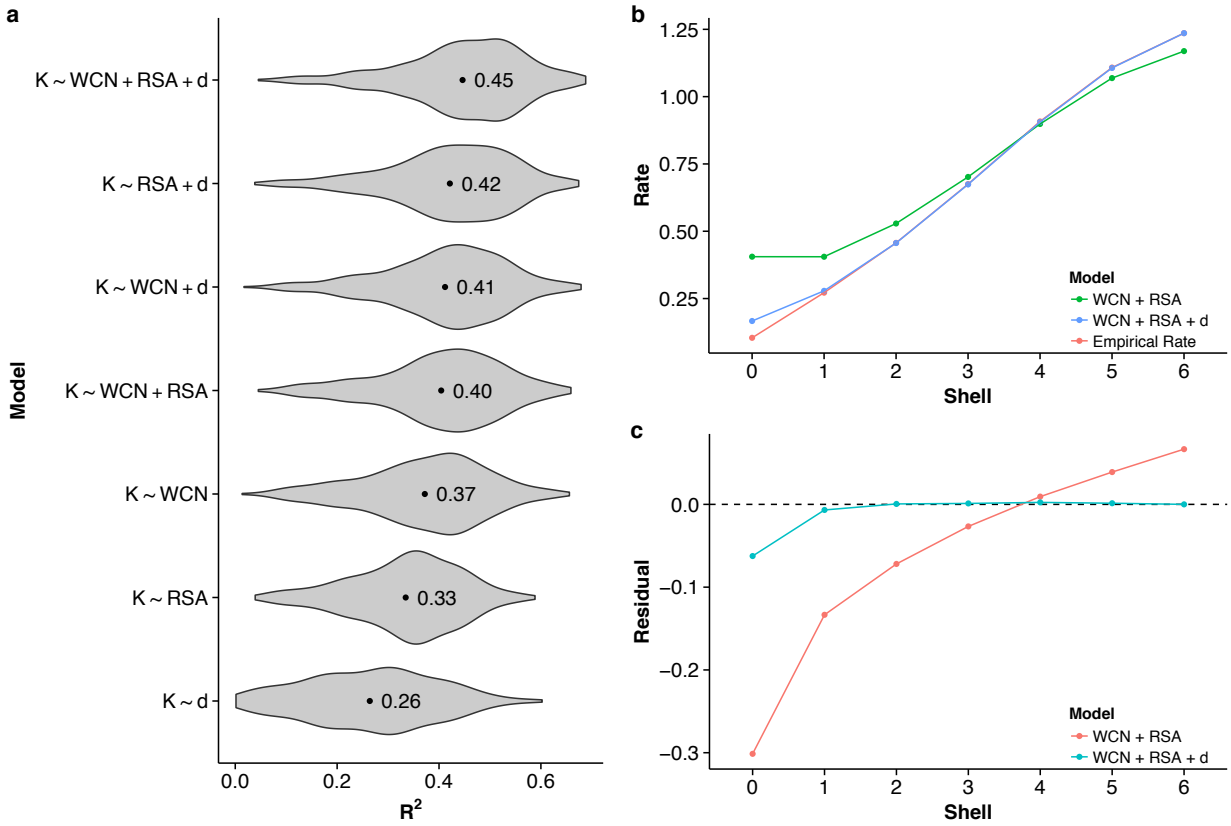
Figure 1



Site-specific evolutionary rates increase with distance to a catalytic residue. Only data from biological assemblies with interface residues removed are considered. (a) Site-specific evolutionary rate versus distance to the nearest catalytic residue. The curve represents the mean rate for all residues, smoothed with a generalized additive model. Standard error is shown in light gray. Individual residues are not shown. The vertical line indicates a distance of 32.5 Å. (b) Density plot of residues with a given distance to a catalytic residue. Dashed lines indicate the boundaries of shells with a thickness of 5 Å each. The vertical red line

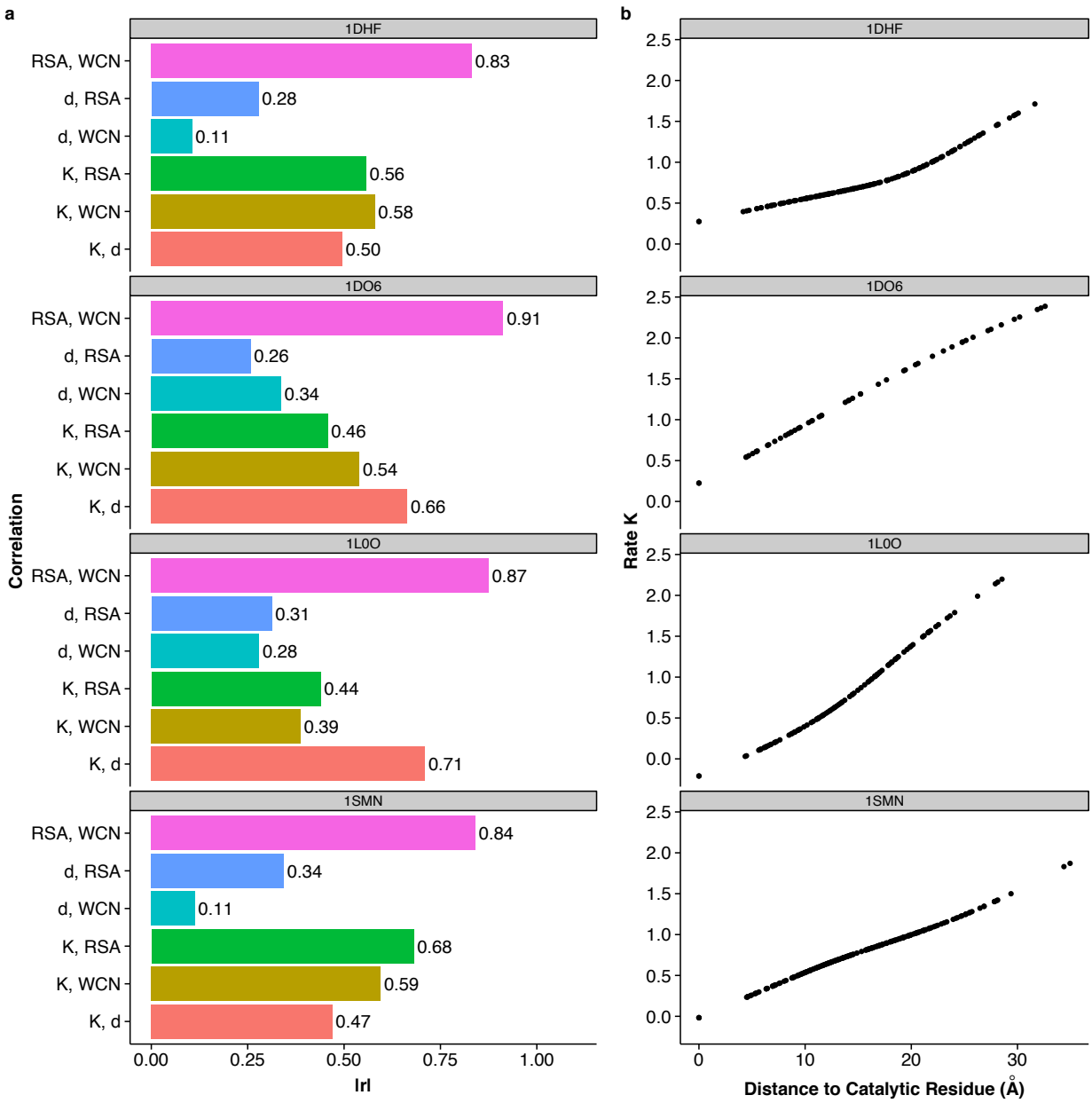
again shows a distance of 32.5, and the percentage shows the proportion of residues within this distnace. (c) Distribution of site-specific evolutionary rates within each shell. Violins show the distribution of rates in each shell, points represent the mean, and the shading represents the number of residues found in that shell. The red line indicates a distance of 32.5 Å, corresponding to the boundary of shell 6.

Figure 2



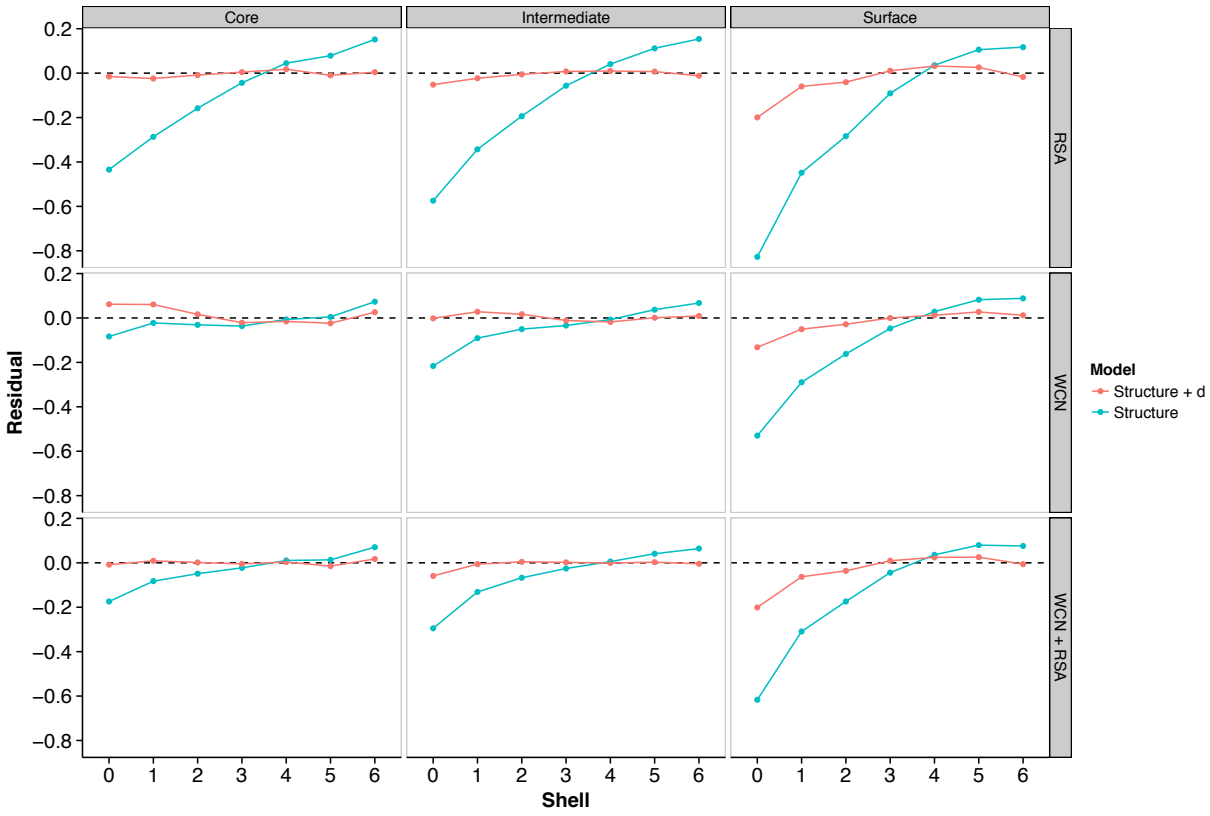
Relative performance of functional and structural predictors of rate. Only data from biological assemblies with interface residues removed are considered. (a) Distribution of R^2 values for various linear models explaining rate, fitted individually for each protein. Labeled points within each distribution indicate the mean R^2 value across proteins. Here, K denotes the site-specific evolutionary rate, d the distance to the nearest catalytic residue, WCN the weighted contact number, and RSA the relative solvent accessibility. On average, the addition of distance to the structural constraints WCN and RSA increases the percent variance explained by the linear models by at least 5 percentage points. (b) Mean empirical and predicted rates, separated by shell. A model containing only WCN and RSA overestimates rates near the active site of the enzyme and the addition of distance corrects this behavior. (c) Mean residuals for linear models with and without distance as a parameter, separated by shell.}

Figure 3



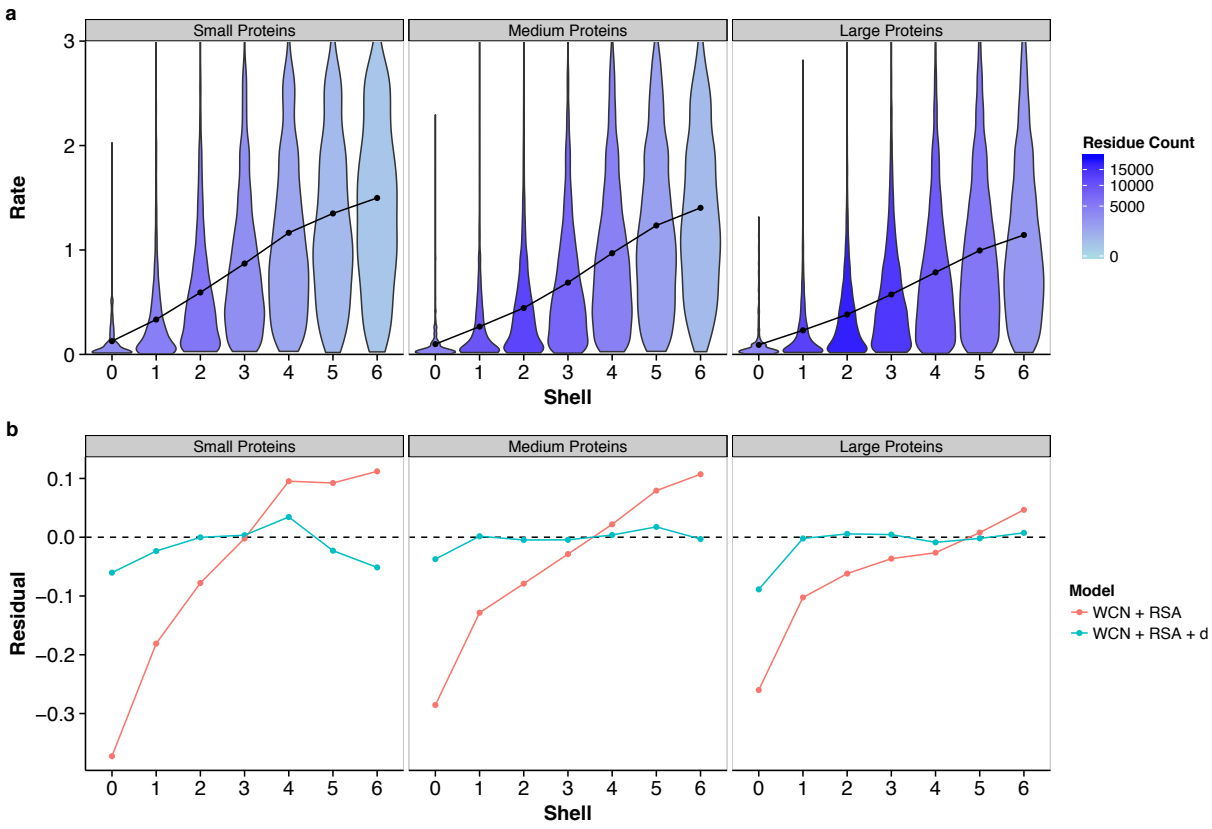
Example enzymes from main text. From top to bottom, the enzymes shown are 1DHF, 1DO6, 1L0O, and 1SMN. Only data from biological assemblies with interface residues removed are considered. (a) Absolute Pearson correlations (r) between combinations of distance d, weighted contact number (WCN), relative solvent accessibility (RSA), and rate K. Distance correlates weakly with WCN or RSA, and strongly with rate, indicating that distance is a strong predictor independent of structural constraints RSA and WCN. (b) Rates smoothed with a GAM (generalized additive model) as a function of d, shown for every residue in each enzyme structure. Smoothed rates increase with distance throughout the entire enzyme structure.

Figure 4



Effect of active-site location on the relationship between site-specific evolutionary rates and distance to the nearest catalytic residue. Only data from biological assemblies with interface residues removed are considered. Enzymes are grouped into three categories, based on the mean solvent exposure of their catalytic residues: core (RSA < 0.05), intermediate ($0.05 <$ RSA < 0.25), and surface (RSA > 0.25). Lines represent mean residuals for different structural linear models (vertical labels, right) and those same models with distance added as a parameter.

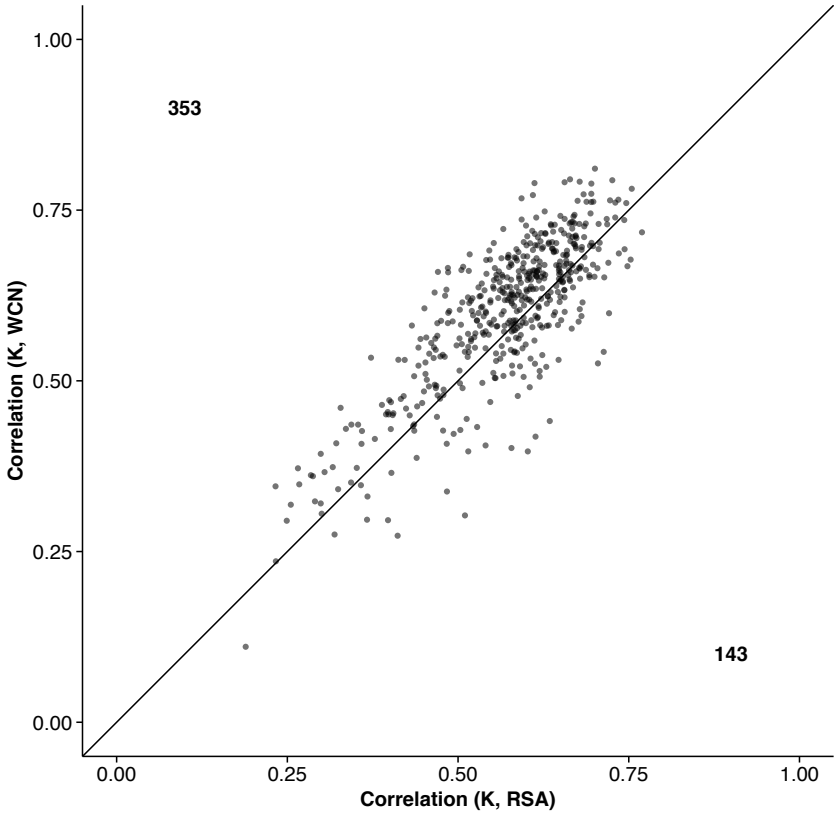
Figure 5



Effect of protein size on the relationship between site-specific evolutionary rates and distance to the nearest catalytic residue. Only data from biological assemblies with interface residues removed are considered. (a) Distribution of rates within each shell, where proteins have been separated into small (95–326 sites), medium (327–484 sites), and large (485–1287 sites) based on amino-acid sequence length. Points represent the mean rate in each shell. As the size of the protein increases, the distance–rate slope decreases and distance effects extend further from the active site. (b) Mean residuals in each shell for models with and without distance, again separated by protein size. The constraining effects of catalytic residues depend on protein size, with stronger, more local effects in small proteins, and weaker, longer-range effects in large proteins.

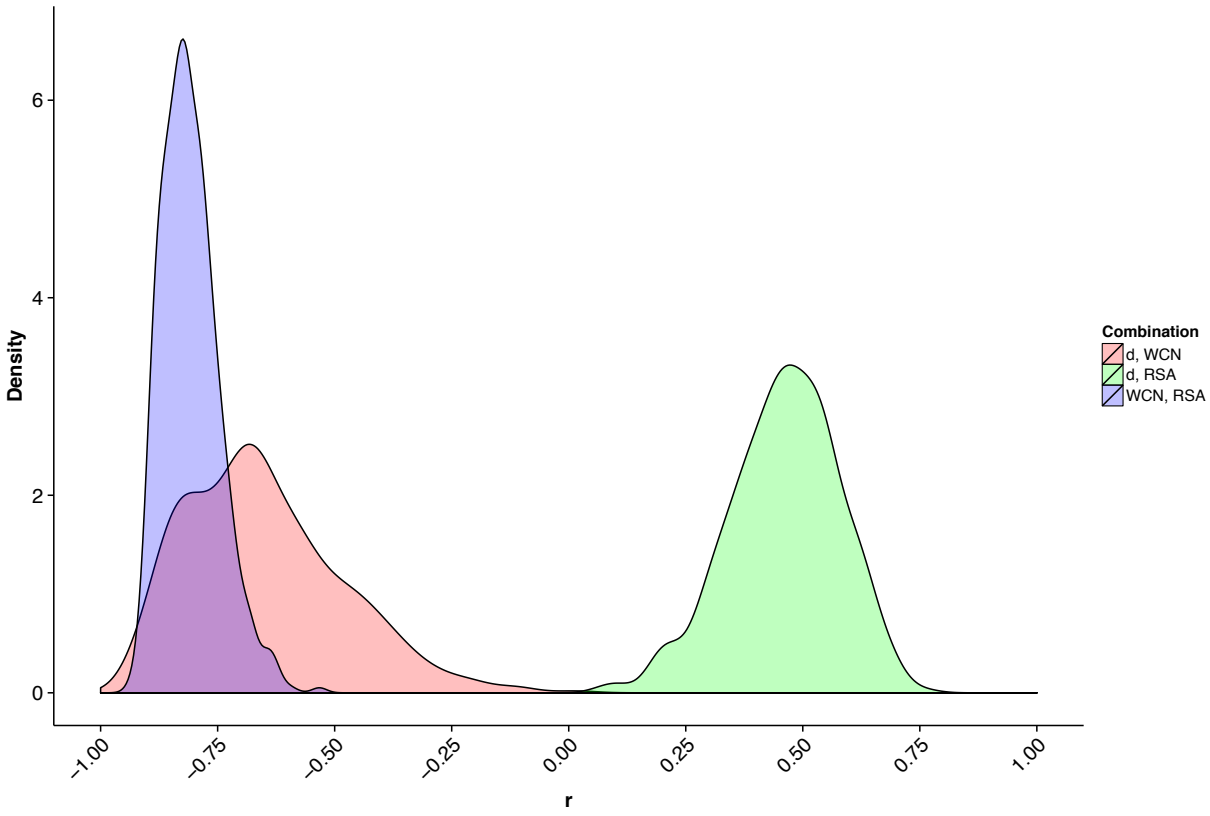
Supporting Figures

Figure S1



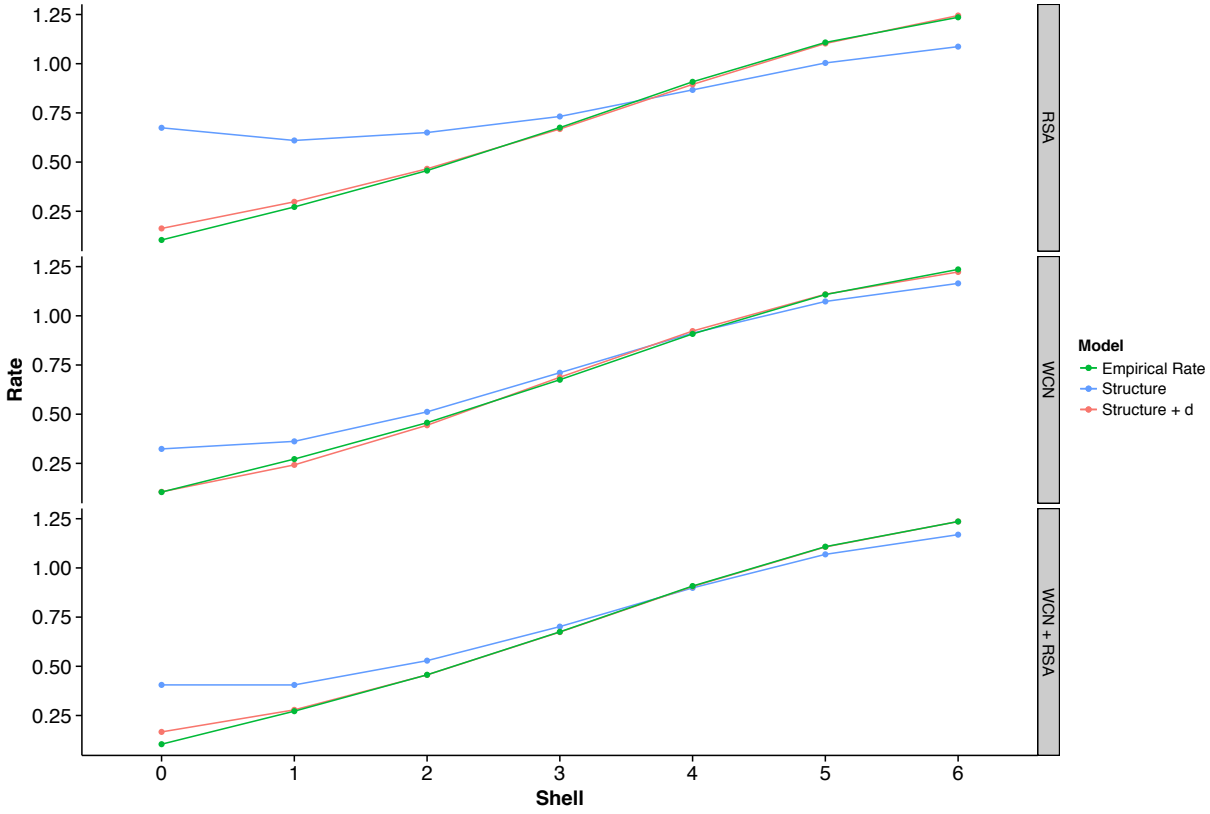
Pearson r values of side-chain WCN and RSA correlated with site-specific evolutionary rates. Only data from biological assemblies with interface residues removed are considered. The sign of the correlation coefficients for WCN are switched from negative to positive for a simpler comparison with RSA. Each point corresponds to an individual protein, and the numbers refer to the number of proteins above or below the $y = x$ line. In aggregate, WCN is a better predictor of site-specific evolutionary rate than RSA.

Figure S2



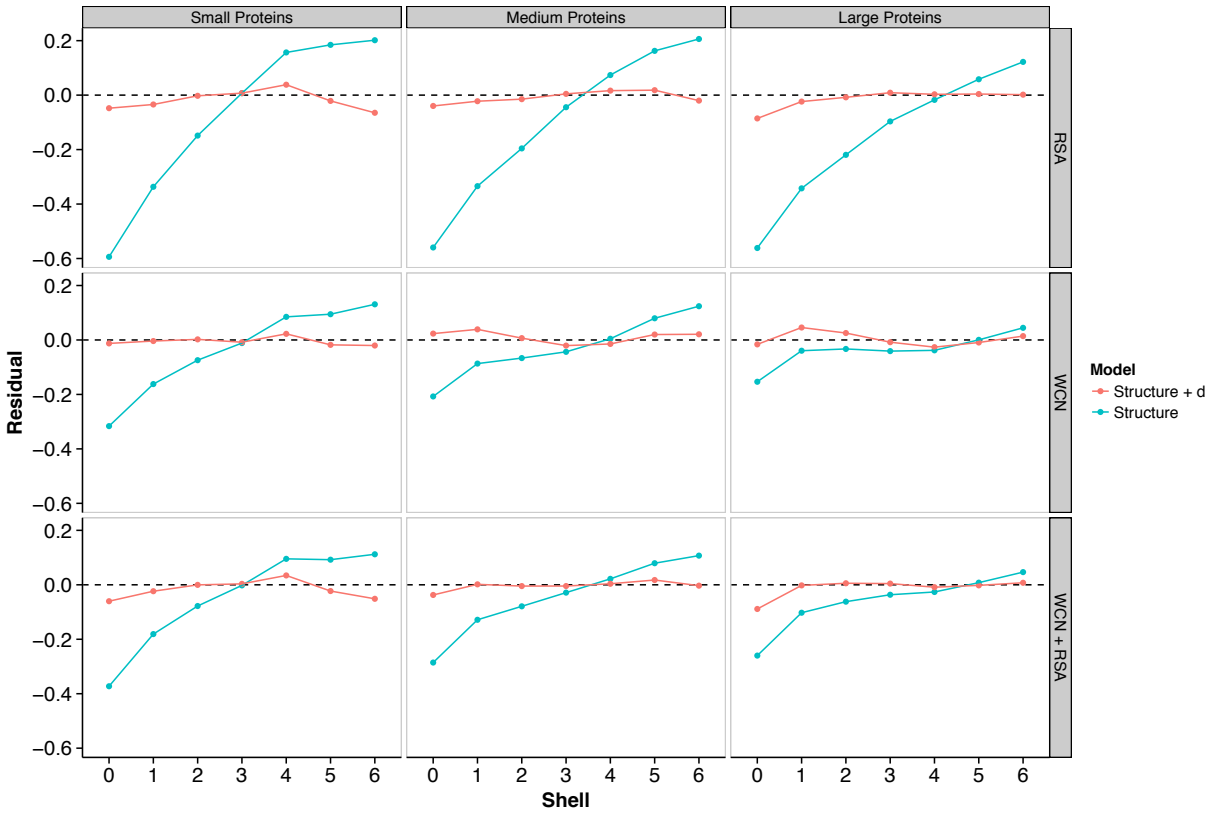
Distributions of Pearson r correlations between pairs of site-specific evolutionary rate predictors on a per protein basis. Only data from biological assemblies with interface residues removed are considered. Predictor pairs include distance and WCN, distance and RSA, and WCN and RSA. WCN and RSA correlate more strongly with each other than either does with distance.

Figure S3



Empirical and predicted site-specific evolutionary rates from different combined linear models. Only data from biological assemblies with interface residues removed are considered. "Structure" refers to one of three structural models used to predict rate: RSA, WCN, or RSA + WCN. "Structure + d" refers to these same three structural models, with the addition of distance to the nearest catalytic residue as a variable. Each point is the mean predicted rate for a given shell across all residues in the data set. In all cases, models that include distance as a parameter predict rate more accurately than models containing only structural parameters, especially near the active site.

Figure S4



Mean residuals for different combined linear models used to predict site-specific evolutionary rate. Only data from biological assemblies with interface residues removed are considered. The data set is divided into small (95–326 sites), medium (327–484 sites), and large (485–1287 sites) proteins. Each point represents the mean predicted rate for all residues in a given shell.