Supporting Information (SI)

Text S1:

Structural complexes retained in each data set after filtering can be found in SI Table S1. Data filtering resulted in an expected but minimal decrease in the size of both the protein receptor and its ligand, primarily in the protein-DNA/RNA and protein-protein data sets (SI Fig. S1). In the protein-DNA/RNA data set, filtering reduced the size of the protein receptor at most 1.36-fold (t-test p=0.04, u-test p=0.40), and we observed a 1.77-fold decrease in ligand size (t-test p=0.01, u-test p=7.6x10^{-5}), primarily due to the exclusion of histone-like complexes with extremely large DNA strands. The size of the larger protein receptor in the protein-protein data set decreased at most 1.75-fold after filtering (t-test p=0.02, u-test p=0.55), while the size of the smaller protein ligand decreased 1.60-fold (t-test p=1.0x10^{-12}, u-test p=5.0x10^{-11}). The distribution of protein size did not change significantly in the small-molecule data set after filtering (t-test p=0.94, u-test p=0.29).

The filtered protein-DNA/RNA data set primarily consisted of proteins bound to double-stranded DNA; 80% of the complexes had a protein bound to DNA, with 63% of these being double-stranded. Of the 15% of complexes having proteins bound to RNA, the majority (65%) was single-stranded.

We observed a difference in the molecular-functional repertoire and functional diversity of proteins that bind DNA/RNA, compared to those involved in small-molecule and protein-protein interactions (SI Fig. S2A). Over 60% of the protein-DNA/RNA complexes in our filtered data set were classified by the enzyme commission (EC) as having either transferase or kinase functions, while protein-small molecule and protein-protein data sets each contained < 30% transferase/kinase enzymes and > 60% lipase/amylase/peptidase proteins. Additionally, > 10% of protein-small molecule and protein-protein complexes were classified into functional categories other than transferase/kinase and lipase/amylase/peptidase, whereas < 6% of protein-DNA/RNA complexes were classified as other than these main functional categories.

Protein-small molecule, protein-DNA/RNA and protein-protein data sets also differed in a number of atomic interaction features extracted from structural information (SI Figs. S2B,S3,S4). As expected, given the acidity of DNA and RNA, DNA/RNA-binding proteins had ~2-fold more basic residues in their ligand-binding surfaces than proteins from small-molecule and protein binding data sets (t-test p<1.6x10^{-42}, u-test p<4.8x10^{-44}, SI Fig. S3). Protein-DNA/RNA complexes had more hydrogen bonds to the ligand than the other data sets (15.10 for protein-DNA/RNA, vs. 9.84 and 6.60 for small-molecule and protein-protein data sets, respectively, t-test p<1.9x10^{-23}, u-test p<1.3x10^{-25}). Protein-small molecule complexes had higher van der Waals interaction forces, when compared to protein-DNA/RNA and protein-protein data sets (-1073.53 vs. -1982.94 and -1991.58, respectively, t-test p<1.6x10^{-40}, u-test p<6.6x10^{-57}). Deformation effect was reduced in the protein-small molecule data set (9.67 for small-molecule, 25.41 for DNA/RNA, 17.74 for protein-protein, t-test p<2.9x10^{-15}, u-test p<1.3x10^{-14}), and the accessible to solvent area of the ligand was greatest in the protein-DNA/RNA data set (625.09 for small-molecule, 10,786.40 for DNA/RNA, 6071.85 for protein-protein, t-test p<3.4x10^{-3}, u-test p<1.4x10^{-6}). These differences are largely expected, given our current general understanding of how proteins are likely to interact with different ligand types.

Although binding affinity (pKd) was marginally different among all three data sets (t-test p<0.04, u-test p<3.5x10^{-3}), the differences were small (mean pKd=6.26 for small-

molecule, 7.20 for DNA/RNA, and 6.97 for protein-protein), suggesting that the range of biologically-meaningful binding affinity values is at least generally comparable for different types of molecular interactions.

Text S2:

We measured the correlation between each atomic interaction feature and experimentally-determined binding affinity (SI Fig. S11A). The protein-small molecule data set showed the strongest correlations between single features and ligand binding. For example, this data set had two single features with r^2>0.65. In contrast, none of the other data sets had any single atomic interaction correlated more than 40% with binding affinity. The protein-small molecule data set also had the largest average correlation between single atomic interaction features and binding affinity (43%); the protein-DNA/RNA data set had 25% average correlation between single atomic interaction features and ligand binding, and the protein-protein data set had 21% average correlation. Hydrophobic contacts and van der Waals interactions were the atomic interactions most strongly correlated with binding affinity in the small-molecule data set (r^2=0.70 and 0.69, respectively), with hydrogen bonds showing the least correlation (r^2=0.07). In contrast, hydrogen bonding was among the atomic interactions most correlated with binding affinity in the protein-protein data set (r^2=0.35), with hydrophobic contacts also being relatively highly-correlated (r^2=0.34). Protein-DNA/RNA binding affinity was most correlated with the solvent-accessible area of the protein receptor (r^2=0.36), deformation effect (r^2=0.30), and van der Waals interactions (r^2=0.36).

The statistical models we employed apply coefficients to each atomic interaction term as a way of weighting the contribution of each term when predicting binding affinity. We plotted the distribution of each atomic interaction term's coefficients obtained from the 100 best-fit models for each data set (SI Fig. S11B). We observed a number of atomic interactions for which the best-fit coefficients were large—in absolute value—but differed in sign between different data sets. Hydrophobic contacts were an example of this pattern, having relatively large positive coefficients in the small-molecule and protein-protein data sets ($5.12\text{x}10^{-3}$ and $3.03\text{x}10^{-3}$, respectively) but a relatively large negative coefficient in the DNA/RNA data set ($-1.59\text{x}10^{-3}$). The coefficients weighting deformation effect also differed in sign between the small-molecule and DNA/RNA data sets. In this case, the small-molecule models had negative weights ($-5.93\text{x}10^{-2}$, on average), whereas the DNA/RNA models had coefficients that were of similar absolute value but positive in direction ($1.37\text{x}10^{-2}$).

Finally, we observed cases in which the magnitude of the coefficients applied to a particular atomic interaction term differed among data sets, suggesting that the interaction term may play the same role in determining binding affinity across data sets, but may be more or less important, depending on the ligand type. As expected, hydrogen bonds had positive coefficients in all three data sets, but the coefficient was larger for the DNA/RNA data set (mean=$2.17\text{x}10^{-1}$) than for the other data sets ($1.18\text{x}10^{-1}$ and $1.79\text{x}10^{-1}$ for small-molecule and protein-protein, respectively, t-test $p<4.5\text{x}10^{-41}$, u-test $p<1.3\text{x}10^{-25}$). Similarly, the coefficients for repulsive interactions were significantly larger for the DNA/RNA data set than for the others (>1.92-fold difference, t-test $p<2.0\text{x}10^{-53}$, u-test $p<1.2\text{x}10^{-32}$). Interestingly, we found that the intercept—which defines the theoretical binding affinity when all other terms are zero—was significantly greater in the DNA/RNA data set than in the others (intercept=6.20 for DNA/RNA, 1.61 for small-molecule, 3.46 for protein-protein, t-test $p<3.5\text{x}10^{-165}$, u-test $p<2.6\text{x}10^{-34}$), possibly

suggesting that interactions involving only a small number of structurally favorable features may be stronger in the case of protein-DNA/RNA binding than for the other types of ligands.

For each data set, we identified all statistical interaction terms combining two atom-atom interaction types that were present in at least 95 of the top 100 models selected by AIC (SI Fig. S12). We first measured the simple correlation between each statistical interaction term and binding affinity (SI Fig. S13A). We found that single statistical interaction terms combining two types of atom-atom interactions were more closely correlated with binding affinity in the small-molecule data set than in the other data sets. Not only was the average correlation between statistical interaction terms and binding affinity higher (mean r^2=0.46 for small-molecule, 0.29 for DNA/RNA and 0.11 for protein-protein, William's test $p<5.8\text{x}10^{-3}$), but the small-molecule data set had 5 statistical interaction terms that were > 60% correlated with binding affinity, whereas the other data sets had no statistical interaction terms with > 50% correlation.

We found that the most-correlated statistical interaction terms in the DNA/RNA data set (deformation effect : hydrogen bonds, r^2=-0.41; van der Waals : deformation effect, r^2=0.44, and van der Waals : repulsive interactions, r^2=0.44) were more correlated with DNA/RNA affinity than any of the simple atom-atom interaction features (max r^2=0.36, $p<6.3\text{x}10^{-3}$, see SI Fig. S11A). In contrast, binding affinity was highly correlated with both complex statistical interaction terms and simple atom-atom interaction features in the small-molecule data set (see SI Figs. S11A,S13). The simple hydrophobic contacts and van der Waals features were correlated with small-molecule affinity with r^2=0.70 and -0.69, respectively, whereas the most highly-correlated statistical interaction term had r^2=0.66.

Table S1. Number of complexes in each data set before and after filtering.

Type of Interaction	Complexes before filtering	Complexes after filtering
Protein-Small Molecule	2897	2342
Protein-DNA/RNA	510	300
Protein-Protein	1743	784
General	4755	3426

Data sets each contain < 1% transmembrane proteins. The filtered protein-DNA/RNA data set consists of 80% protein-DNA complexes, 15% protein-RNA complexes and 5% protein-nucleotide complexes.

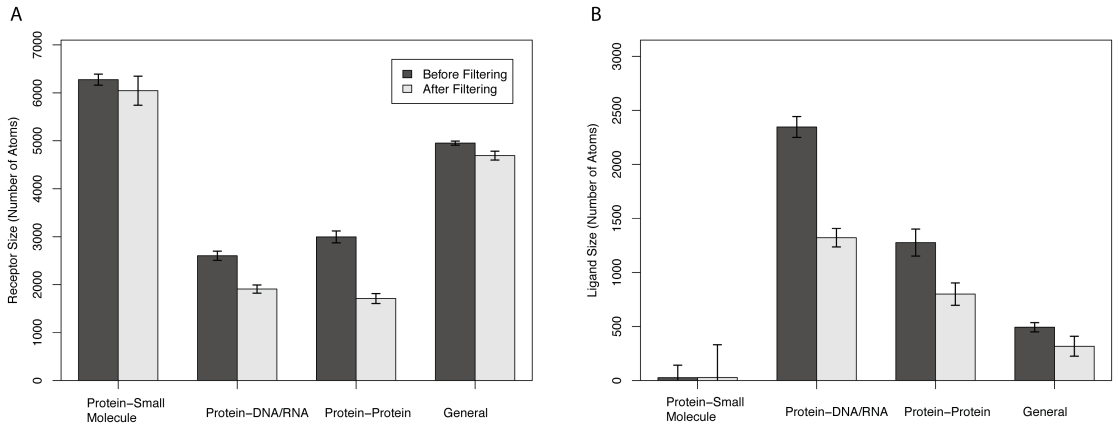

Figure S1. The distribution of receptor and ligand size decreases moderately after filtering. We plot the size (number of atoms) of the protein receptor (A) and the bound ligand (B) of each structural complex, before (dark) and after (light) data filtering. For the protein-protein data set, the larger protein in the complex was considered the 'receptor,' and the smaller protein the 'ligand' of each complex. Bars indicate standard error.

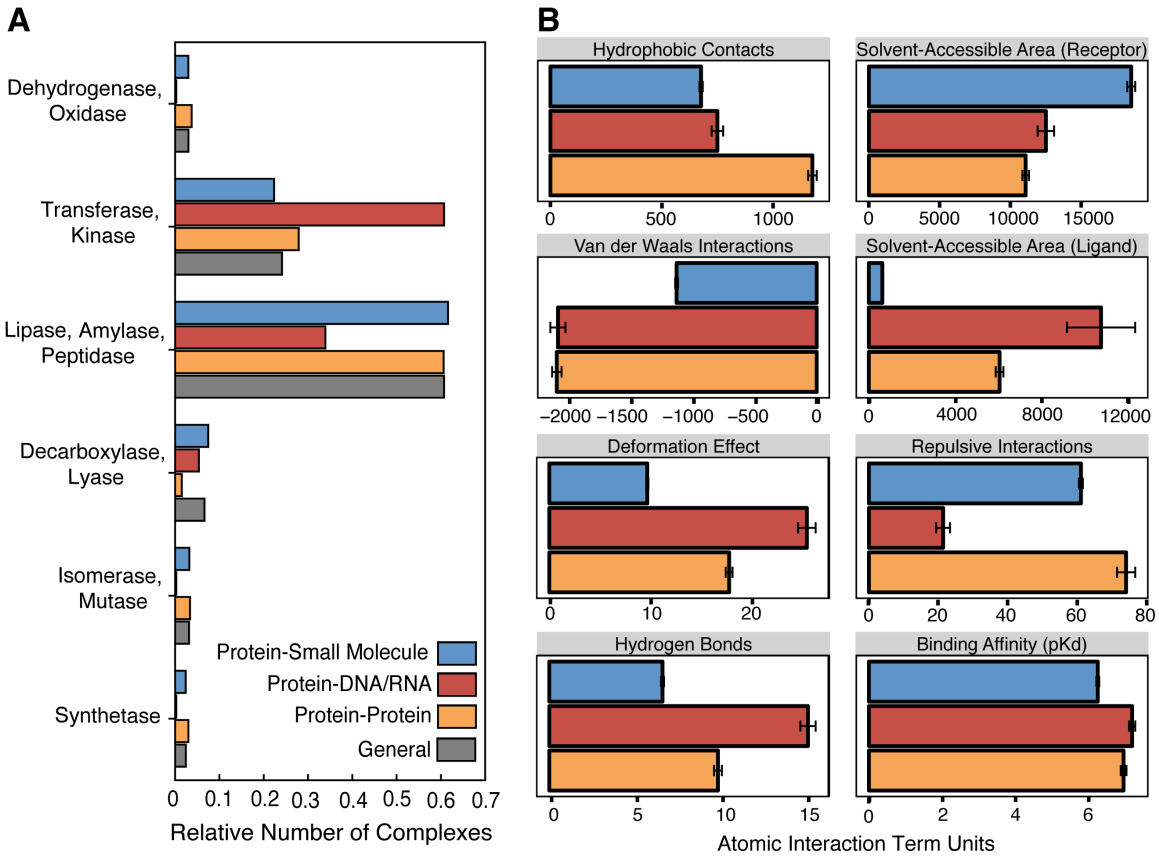

Figure S2. Proteins interacting with small molecules, DNA/RNA and other proteins have different molecular functions and quantitatively different structural features. A. We plot the proportion of protein-small molecule (blue), protein-DNA/RNA (red) and protein-protein (orange) complexes performing each category of molecular function, based on enzyme commission (EC) classification. The general data set includes all complexes (gray). B. For complexes of each interaction type, we plot the mean and standard error of each extracted structural feature (see Methods, Fig. 1).

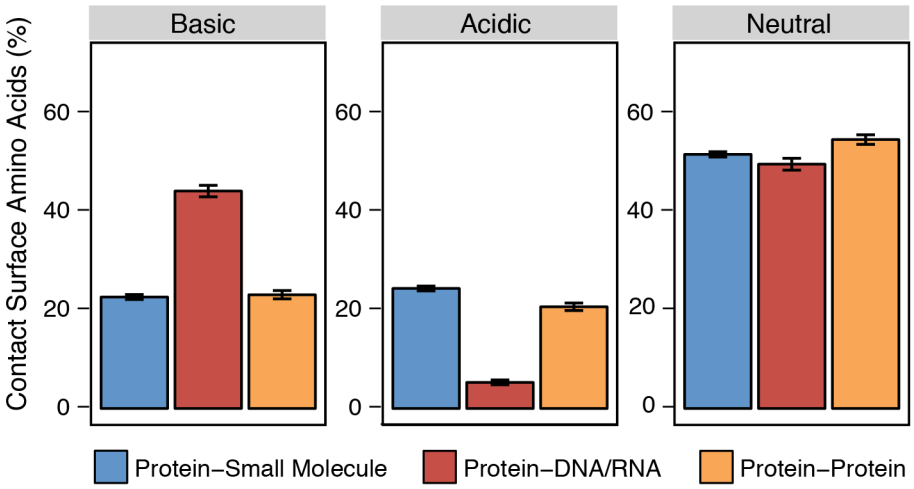

Figure S3. Proteins binding DNA/RNA have ~2-fold more basic residues on their binding surfaces than proteins involved in interactions with small molecules or other proteins. We identified the binding surface of each protein receptor as consisting of those amino acids with at least one atom < 3.5 Å distance from the ligand. We plot the mean percentage of amino acids in the binding surface that are basic, acidic or neutral, divided into protein-small molecule (blue), protein-DNA/RNA (red) and protein-protein (orange) data sets. Bars indicate standard error.

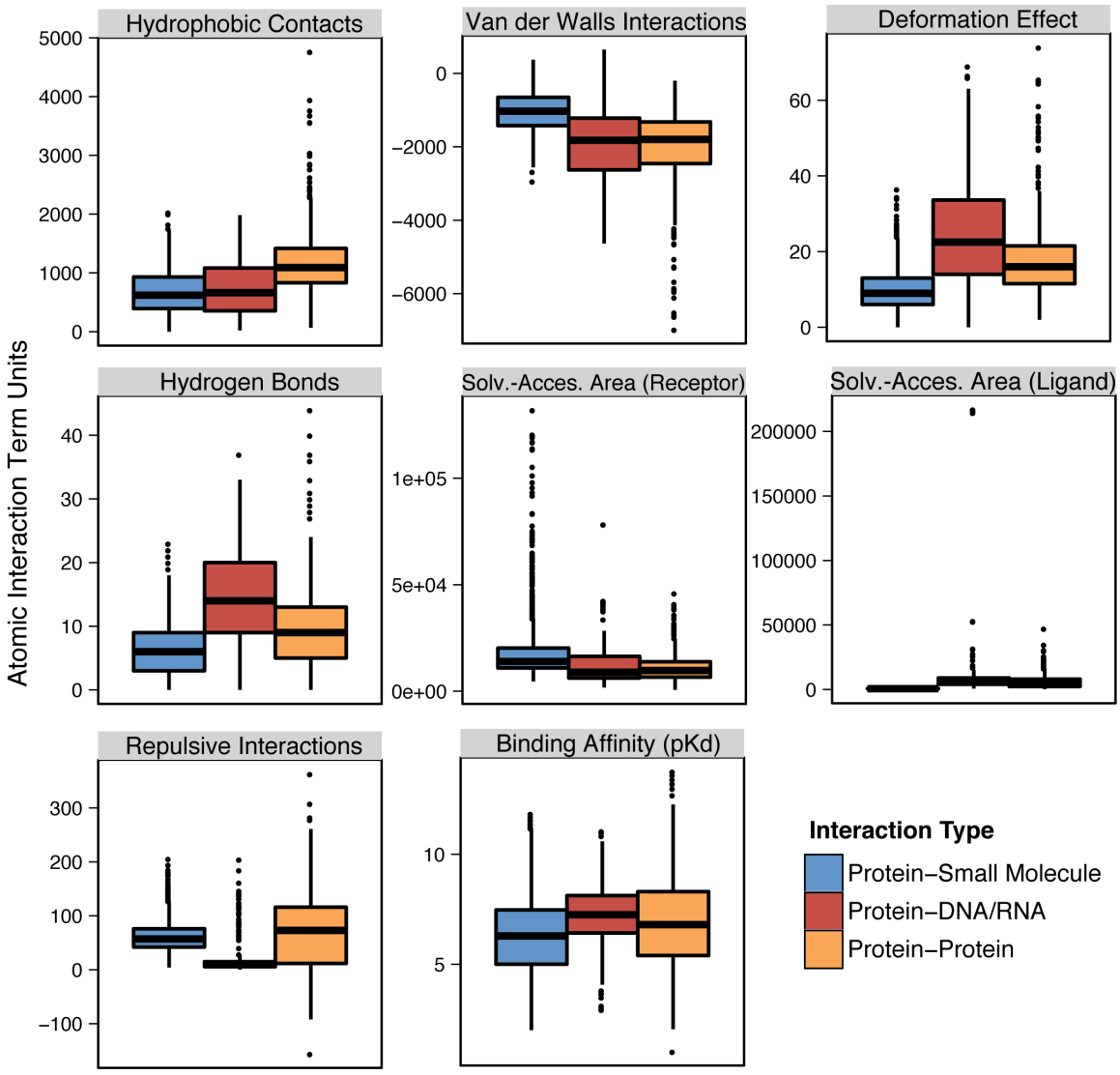

Figure S4. The distributions of structural features differ among proteins that bind small molecules, those that bind DNA/RNA and those interacting with other proteins. From each protein-ligand complex in our filtered data set, we calculated a number of structural features thought to correlate with ligand affinity (see Methods). For each structural feature, we plot the median, 1st and 3rd quartiles, and 1.5 interquartile range as a "box plot." Blue boxes indicate complexes with protein bound to a small molecule, red indicates protein-DNA/RNA complexes, and orange indicates protein-protein complexes.

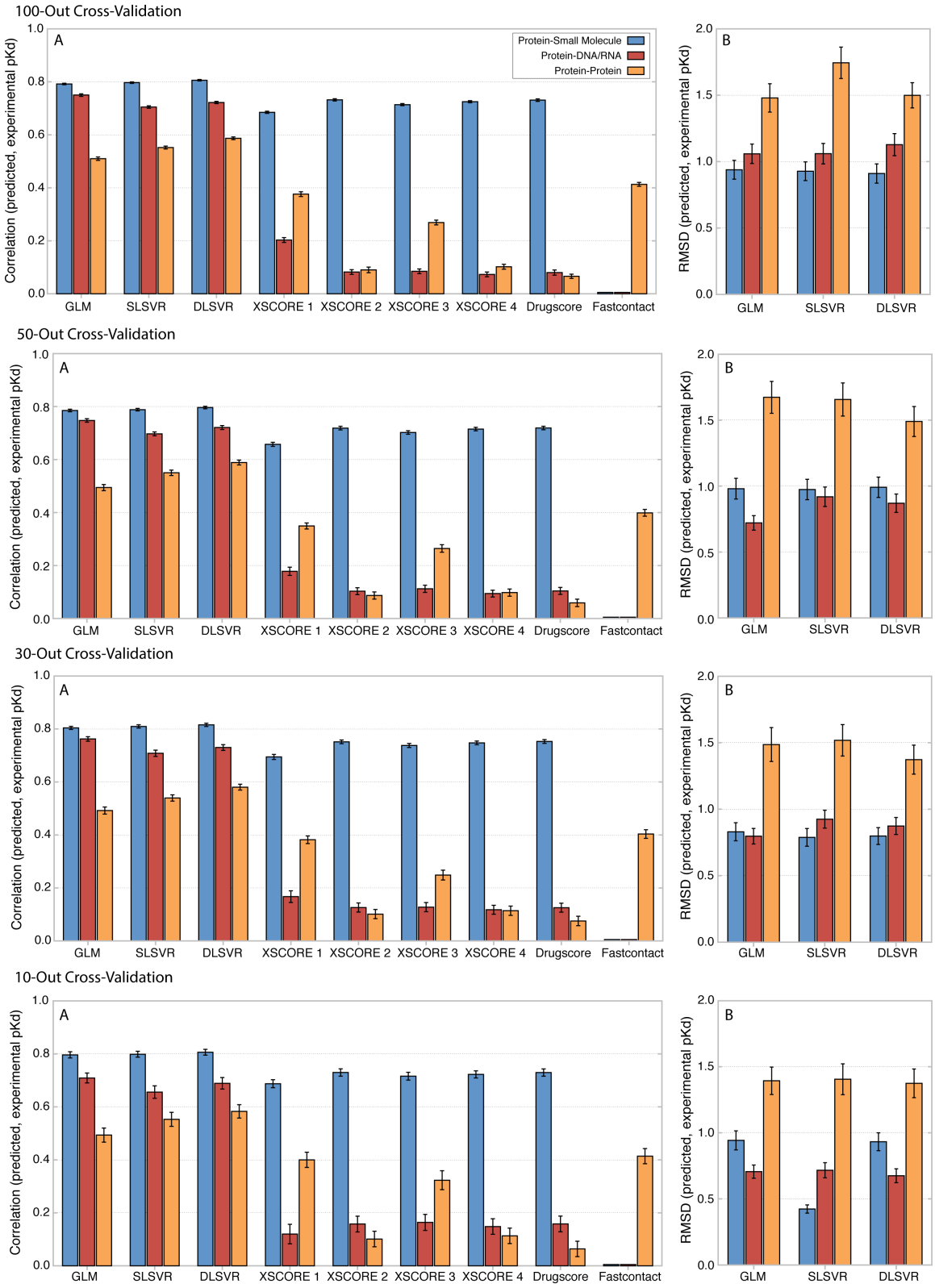

Figure S5. Different cross-validation strategies have minimal effect on inferred accuracy. We used replicated cross-validation to evaluate the accuracy with which novel statistical models and existing prediction tools can predict ligand affinity (pKd) from

structural information (see Methods, Figs. 1,2). In each labeled panel, we plot results using a different number of complexes set aside as testing data for each replication (from 10 to 100), with A showing the correlation (r^2) between predicted and experimental binding affinity (pKd), and B showing root mean squared deviation (RMSD) between predicted and experimental affinity. Bars indicate standard error, and colors indicate different interaction data sets (blue=protein-small molecule, red=protein-DNA/RNA, orange=protein-protein). Results are shown for a generalized linear model (GLM), single-layer (SLSVR) and dual-layer (DLSVR) support vector regression, and a number of existing prediction tools (see Methods).

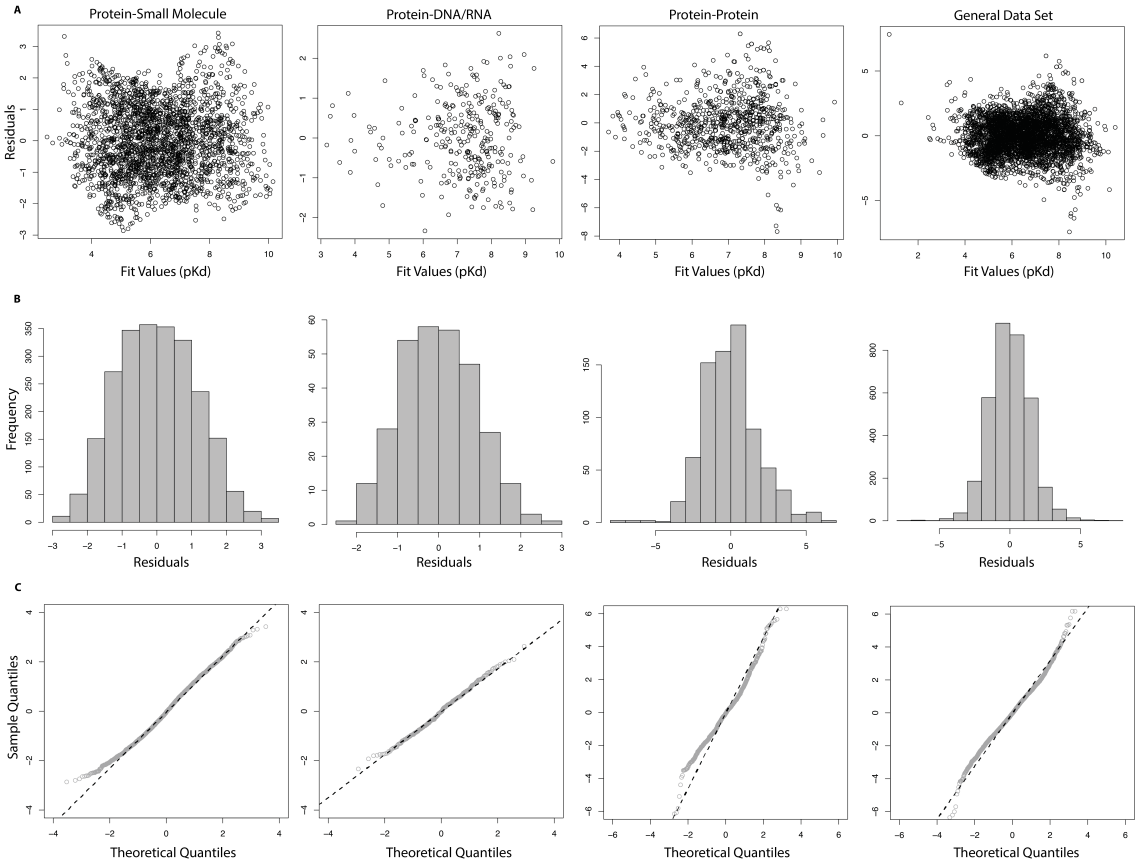

Figure S6. Statistical analysis suggests that affinity-prediction models are not strongly biased. For the best-fit generalized linear model fit to each data set by cross-validation (see Methods), we plot three views of the residuals. A. We plot the predicted pKd (X-axis) vs. the distance between predicted and experimentally-determined pKd (Y-axis) of each complex. B. We plot the frequency distribution of residuals of each size. C. We plot the theoretical (X-axis) vs. observed (Y-axis) quantiles for each data set.

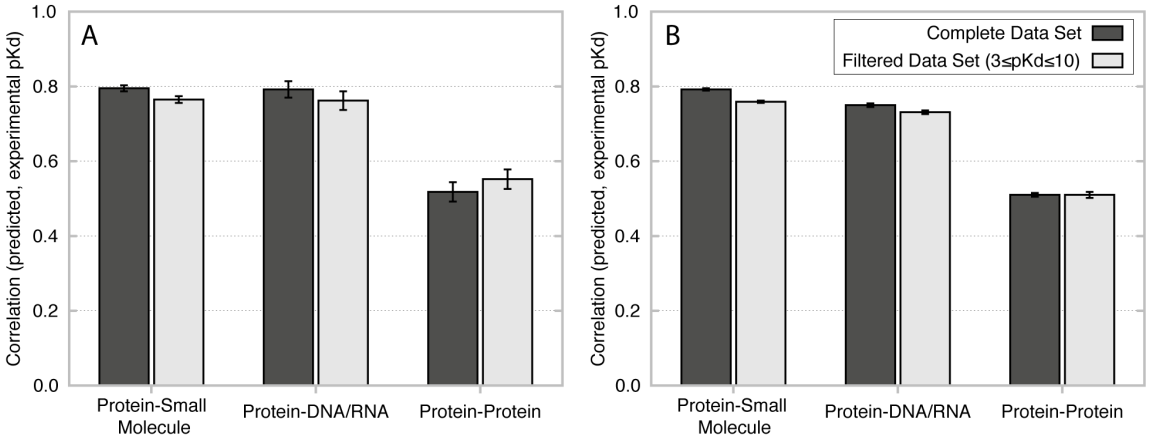

Figure S7. Excluding outlier complexes had minimal effect on predictive accuracy. We removed structural complexes with experimentally-determined binding affinity (pKd) ≤ 3 or ≥ 10 from either the training data (A) or the testing data (B). We plot the mean correlation between predicted and experimental pKd (r^2) for each interaction data set before (dark series) and after (light series) removing outlier structures. Bars indicate standard error.

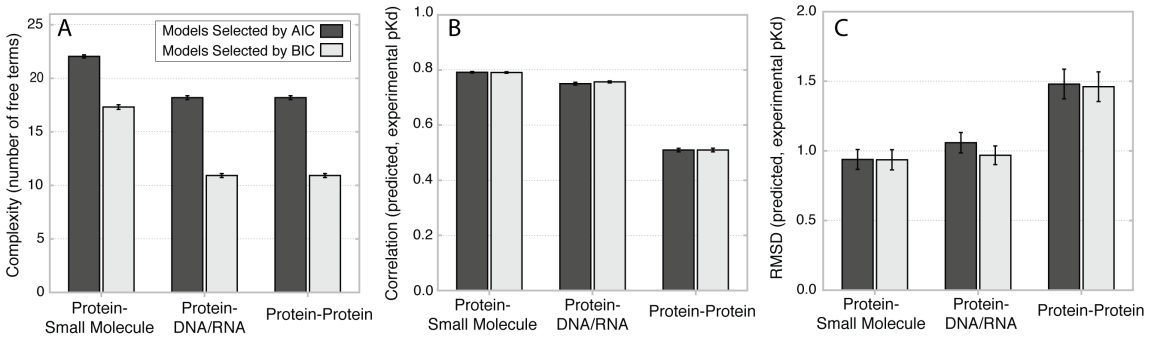

Figure S8. Different model-selection procedures affected model complexity but not predictive accuracy. We used either Akaike or Bayesian information criteria (AIC (dark) and BIC (light), respectively) to identify best-fit statistical models for each interaction data set. A. We plot the mean complexity (number of free parameters in the model) of models selected by each criterion from each data set. B. We plot the correlation between predicted and experimentally-determined binding affinity (pKd) produced by models selected by each criterion for each data set. C. We plot the root mean squared deviation (RMSD) between predicted and experimental pKds. In all panels, bars indicate standard error.

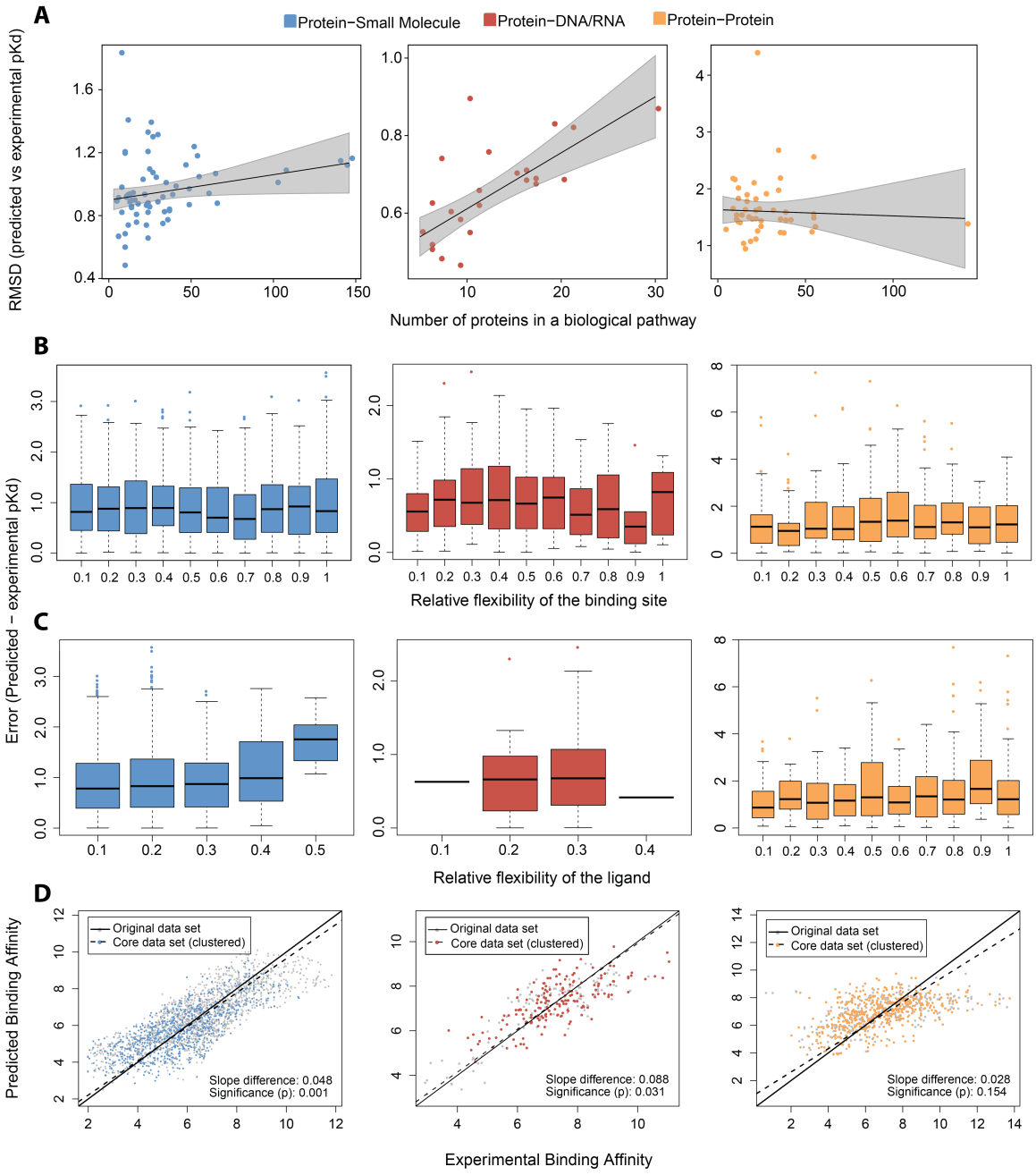

Figure S9. Predictive accuracy is not stongly variable across different data set partitionings. A. We cluster the binding complexes by KEGG metabolic pathway and plot the RMSD across complexes in the cluster (predicted vs. experimental pKd) against the number of complexes in the cluster. B and C. We plot the prediction error vs. protein flexibility and ligand flexibility, respectively. D. We generated a 'nonredundant' data set by clustering proteins of > 90% sequence similarity and selecting one representative complex for each cluster; we plot predicted vs. experimental pKds for the nontredundant data set and the complete data set.

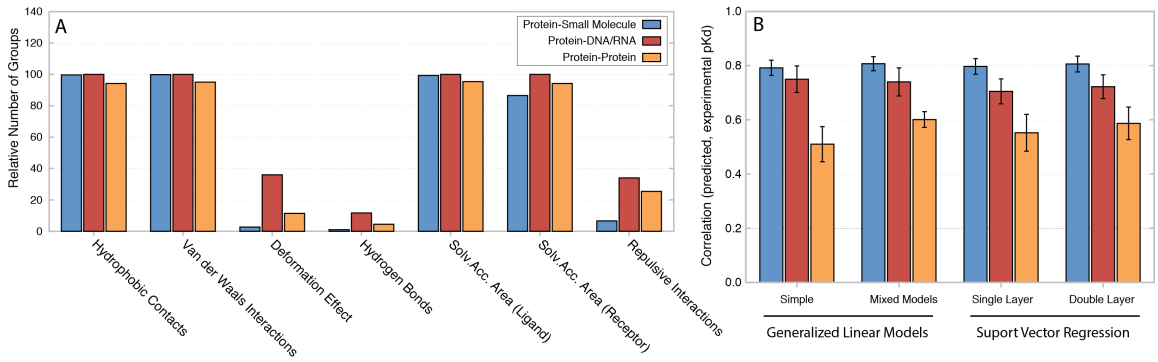

Figure S10: Mixed models identify heterogeneity in atomic interaction features and reduce variation in protein-protein prediction accuracy. We generated mixed models (GLMM) for each data set by adding random effects to the best-fit generalized linear model (GLM) obtained from cross-validation, using the GLMM search algorithm to select the best-fit number of categories for each structural feature (see Methods). A. The best-fit number of categories identified by mixed model analysis for each structural feature is shown for each interaction type. B. We compared the predictive accuracy of the best-fit mixed generalized linear model (GLMM) to the best-fit homogeneous generalized linear model (GLM) and two types of homogeneous support vector regression models (single layer, SLSVR; and double-layer, DLSVR). We plot the mean Pearson correlation (r^2) between predicted and experimentally-determined binding affinity (pKd) for each statistical model and interaction type over 100 replicates of cross-validation (see Methods, Fig. 2). Bars indicate standard deviation.

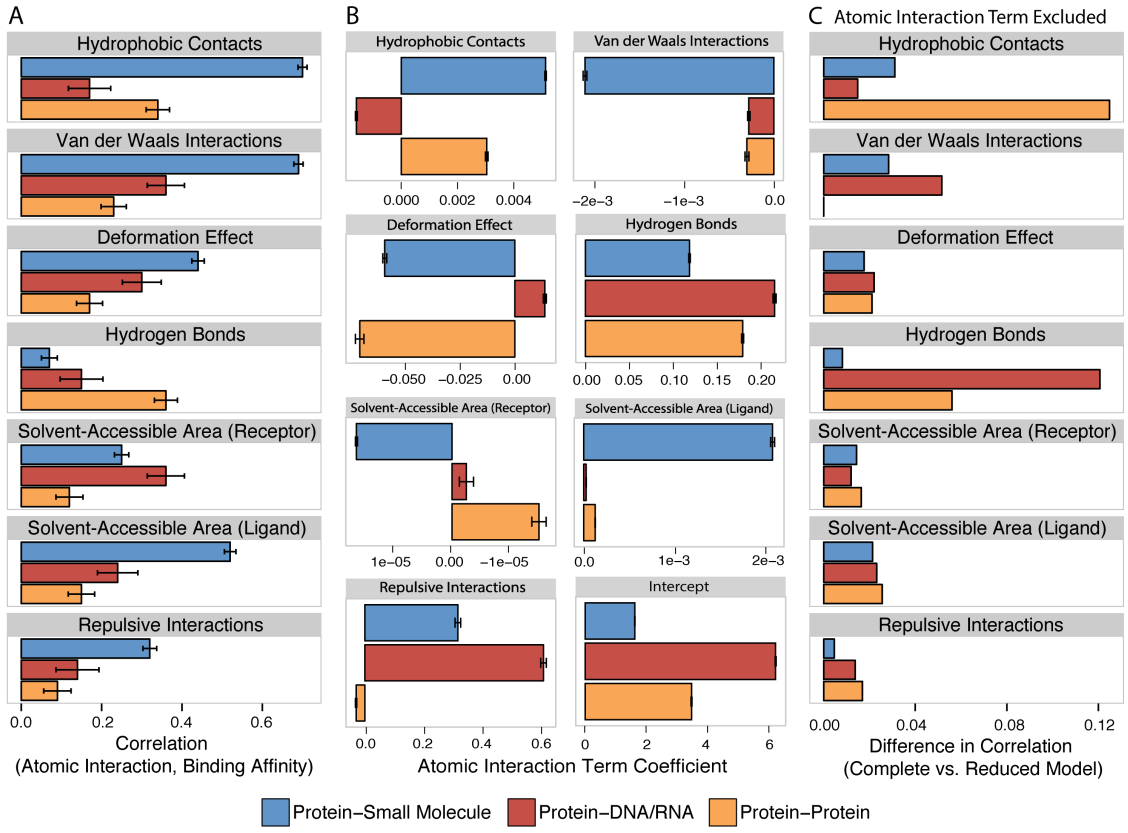

Figure S11. Atomic interactions contribute differently to binding affinity prediction of protein-small molecule, protein-DNA/RNA and protein-protein binding. A. We plot the mean and standard error in spearman correlation between each atomic interaction and experimental binding affinity (see Methods, Fig. 1). B. For each atomic interaction, we plot the mean and standard error of the coefficients applied to that atomic interaction over the 100 best-fit prediction models obtained from each data set (see Methods). Intercept refers to the constant term. C. We generated reduced models by excluding each atomic interaction from the complete statistical model with all atomic interactions included. The plot shows the difference in Pearson correlation (r^2) between predicted and experimental binding affinities, comparing the best-fit complete models for each data set to the best-fit reduced model with the indicated atomic interaction removed.

Figure S12. Different combinations of atomic interactions are selected as statistically important for different data sets. For each data set (protein-small molecule, protein-DNA/RNA, protein-protein, and the 'general' data set containing all complexes) we plot the number of the top 100 best-fit generalized linear models (selected by AIC) having that statistical interaction term. We show only those statistical interaction terms that are present in at least 95 of the top 100 models in at least one data set.