SUPPLEMENTARY DATA

Statistical Significance Testing

Estimation of Model Accuracy

Table S1 shows Bonferroni-corrected p-values for difference in GDT-TS score correlation for models trained on Mixed and X-ray only training sets. P-values are computed by applying Fisher's transformation (see below) to convert the raw Pearson correlations to z-scores. We use Pearson for hypothesis testing because this formula is not robust for Spearman correlation. Bold indicates significance at $\alpha = 0.05$.

Fisher's transformation of a Pearson correlation ρ is computed by

$$z = \frac{1}{2}\log(\frac{1+\rho}{1-\rho})$$

where z is approximately normally distributed with constant standard deviation which depends on the number of datapoints N:

$$SD = \sqrt{\frac{1}{N-3}}$$

We can use the calculated z and SD to compute a two-sided p-value using a standard z-score hypothesis test.

Protein Sequence Design

Table S2 shows pairwise Bonferroni-corrected p-values for the protein sequence design task. Significance is assessed by Mann-Whitney U-test, a non-parametric test of whether two samples are drawn from the same distribution. In other words, we are testing whether there is a statistically significant shift in the distribution of protein design perplexities (1) between X-ray, NMR, and EM structures for each training set and (2) between each training set for each structure determination method. Bold indicates significance at $\alpha = 0.05$.

Catalytic Residue Prediction

Table S4 shows Bonferroni-corrected p-values for the enzyme class prediction task. As above, significance is assessed by Mann-Whitney U-test. We test whether there is a statistically significant shift in the distribution of AUPRCs between the model trained on only X-ray data and the model trained on mixed data for both X-ray (row 1) and NMR (row 2) test structures. We also test whether there is a statistically significant shift in the distribution of AUPRCs between the X-ray and NMR test structures for both the model trained on only X-ray data (row 3) and the model trained on mixed data (row 4). Bold indicates significance at $\alpha = 0.05$ after correcting for multiple hypothesis testing ($n = 4$).

Table S5 shows Bonferroni-corrected p-values for the enzyme class prediction task when the test structures are partitioned by enzyme class. As above, significance is assessed by Mann-Whitney U-test. No significant differences were found at $\alpha = 0.05$ after correcting for multiple hypothesis testing ($n = 24$).

Table S1.

Target	p-value
T0769	6.0516
T0773	2.4567
T0853	23.268
T0855	19.9525
T0857	3.1693
T0865	**0.041**
T0902	0.1559
T0918	3.6242
T0950	19.3462
T0951	0.6405
T0953s1	1.8857
T0953s2	3.5739
T0954	7.5116
T0955	**0.0041**
T0957s1	15.4739
T0957s2	0.914
T0958	**0.0424**
T0960	**0.0006**
T0963	0.1484
T0966	**0.0000**
T0968s1	12.3549
T0968s2	11.3608
T1003	**0.0043**
T1005	0.6508
T1008	11.1116
T1009	**0.0348**
T1011	**0.0009**
T1016	20.6921

Table S2.

		Train: Mixed			Train: X-ray		
		X-ray	NMR	EM	X-ray	NMR	EM
Train: Mixed	X-ray		$\mathbf{1.42 \times 10^{-30}}$	$\mathbf{1.83 \times 10^{-5}}$			
	NMR			1.6080			
	EM						
Train: X-ray	X-ray	3.8199				$\mathbf{1.20 \times 10^{-123}}$	$\mathbf{2.92 \times 10^{-9}}$
	NMR		$\mathbf{2.19 \times 10^{-91}}$				1.7898
	EM			0.0832			

Table S3.

CATH Class	**Train set 1**	**Test set 1**	**Train set 2**	**Test set 2**	**p-value**
Mainly Alpha	X-ray	X-ray	Mixed	X-ray	10.4517
	X-ray	NMR	Mixed	NMR	$\mathbf{1.11 \times 10^{-9}}$
	X-ray	X-ray	X-ray	NMR	$\mathbf{1.92 \times 10^{-17}}$
	Mixed	X-ray	Mixed	NMR	$\mathbf{1.71 \times 10^{-3}}$
Mainly Beta	X-ray	X-ray	Mixed	X-ray	13.1935
	X-ray	NMR	Mixed	NMR	$\mathbf{1.77 \times 10^{-13}}$
	X-ray	X-ray	X-ray	NMR	$\mathbf{2.24 \times 10^{-22}}$
	Mixed	X-ray	Mixed	NMR	1.9779
Alpha Beta	X-ray	X-ray	Mixed	X-ray	9.7910
	X-ray	NMR	Mixed	NMR	$\mathbf{8.00 \times 10^{-20}}$
	X-ray	X-ray	X-ray	NMR	$\mathbf{4.36 \times 10^{-82}}$
	Mixed	X-ray	Mixed	NMR	$\mathbf{9.03 \times 10^{-18}}$
Few Secondary Structures	X-ray	X-ray	Mixed	X-ray	16.4823
	X-ray	NMR	Mixed	NMR	$\mathbf{3.90 \times 10^{-3}}$
	X-ray	X-ray	X-ray	NMR	$\mathbf{1.77 \times 10^{-18}}$
	Mixed	X-ray	Mixed	NMR	$\mathbf{2.85 \times 10^{-4}}$

Table S4.

Train set 1	**Test set 1**	**Train set 2**	**Test set 2**	**p-value**
X-ray	X-ray	Mixed	X-ray	1.2588
X-ray	NMR	Mixed	NMR	0.6264
X-ray	X-ray	X-ray	NMR	0.1880
Mixed	X-ray	Mixed	NMR	**0.0404**

Table S5.

Enzyme Class	**Train set 1**	**Test set 1**	**Train set 2**	**Test set 2**	**p-value**
Oxidoreductases	X-ray	X-ray	Mixed	X-ray	11.4456
	X-ray	NMR	Mixed	NMR	6.3624
	X-ray	X-ray	X-ray	NMR	3.9312
	Mixed	X-ray	Mixed	NMR	11.3136
Transferases	X-ray	X-ray	Mixed	X-ray	10.0272
	X-ray	NMR	Mixed	NMR	12.0000
	X-ray	X-ray	X-ray	NMR	2.7624
	Mixed	X-ray	Mixed	NMR	3.2688
Hydrolases	X-ray	X-ray	Mixed	X-ray	10.5024
	X-ray	NMR	Mixed	NMR	8.6088
	X-ray	X-ray	X-ray	NMR	0.9696
	Mixed	X-ray	Mixed	NMR	0.3456
Lyases	X-ray	X-ray	Mixed	X-ray	3.9840
	X-ray	NMR	Mixed	NMR	3.3408
	X-ray	X-ray	X-ray	NMR	7.5936
	Mixed	X-ray	Mixed	NMR	11.0880
Isomerases	X-ray	X-ray	Mixed	X-ray	8.9688
	X-ray	NMR	Mixed	NMR	12.0000
	X-ray	X-ray	X-ray	NMR	1.9896
	Mixed	X-ray	Mixed	NMR	2.892
Translocases	X-ray	X-ray	Mixed	X-ray	12.0000
	X-ray	NMR	Mixed	NMR	12.0000
	X-ray	X-ray	X-ray	NMR	4.4520
	Mixed	X-ray	Mixed	NMR	4.4520

Relative performance on NMR structures by CASP year in test set

The only difference between our test set and the standard CASP13 dataset is that we add 6 additional NMR structures from CASP 11-12 due to the lack of NMR structures in recent CASP editions. However, the characteristics of decoys in CASP have changed over the years, especially with the advent of deep learning–based structure prediction models, resulting in potential bias when performance structures from different years are compared directly.[43] To demonstrate that our results are consistent despite the inclusion of prior CASP data, we separate out the per-target Spearman correlations for NMR structures in each CASP in Figure S1. It is clear from this that performance decreased when training only on X-ray data regardless of which year the test structures came from.

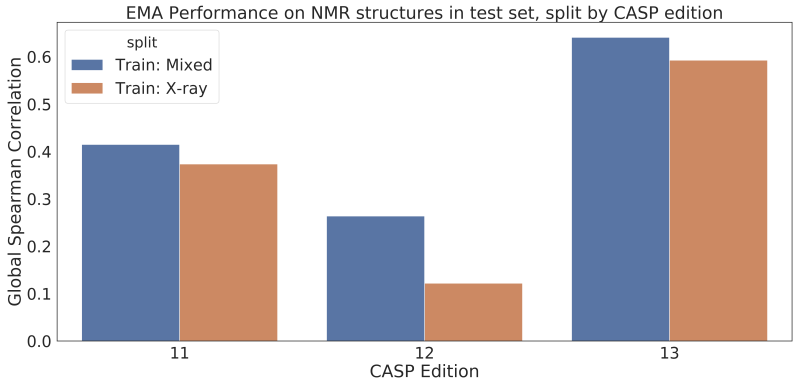

Fig. S1. Per-target Spearman correlations for NMR structures in the test set, split out by CASP edition. Regardless of CASP edition, the correlations decrease when trained on X-ray structures only, with even greater decreases for the most recent editions.

Per-protein protein design results for paired dataset

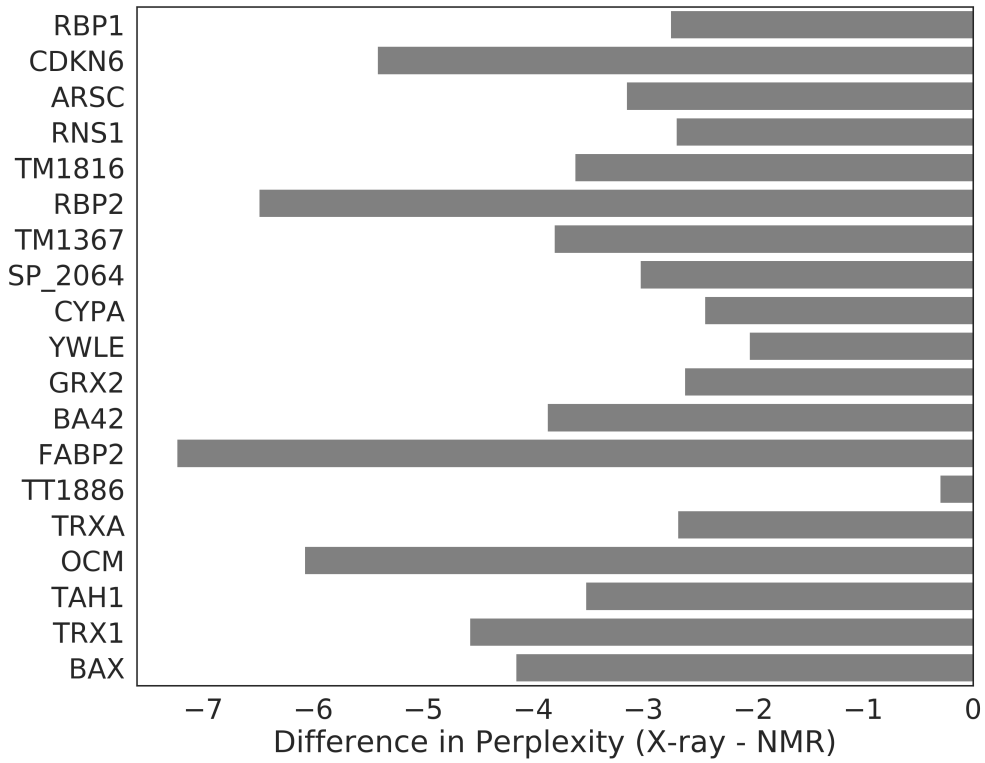

Fig. S2. Difference in perplexity between X-ray and NMR structures for protein each protein in paired dataset from Mei et al..[10] All proteins have lower perplexity for X-ray than for NMR structures.

Error analysis for protein design task

Figure S3 shows the full confusion matrix for protein design on X-ray data in order to help interpret the drop in performance for certain residues. Rows are true amino acids, columns are predicted amino acids. Amino acids are ordered by biochemical properties: positively charged (H, K, R), negatively charged (D, E), small polar (S, T, N, Q), small hydrophobic (A, V, L, I, M), large hydrophobic (F, Y, W), and unique (P, G, C). Table S6 outlines most-commonly predicted residues (incorrect predictions with $> 5\%$ frequency) for the cases where performance decreased upon inclusion of NMR data in training.

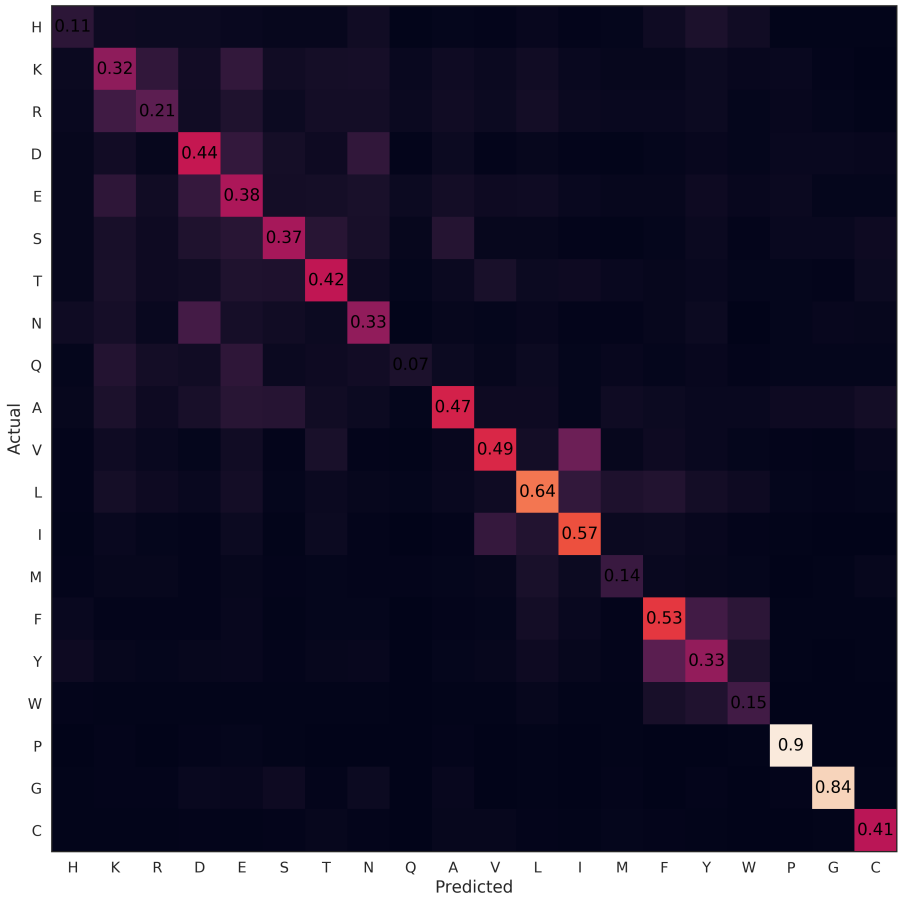

Fig. S3. Confusion matrix for protein design task on X-ray test data. Results show that confusion is generally seen between biochemically similar residues, suggesting that the decrease in performance for certain residues may not be detrimental to model performance.

Table S6.

True Residue	**Predicted Residue**	**Frequency**
Asparagine (N)	ASP	0.199
	GLU	0.079
	LYS	0.067
	SER	0.056
Threonine (T)	GLU	0.084
	VAL	0.076
	SER	0.075
	LYS	0.061
Isoleucine (I)	VAL	0.148
	LEU	0.120
Leucine (L)	ILE	0.082
Phenylalanine (F)	TYR	0.131
	LEU	0.104
Serine (S)	ALA	0.105
	GLU	0.102
	THR	0.093
	ASP	0.067
	LYS	0.055

Dataset statistics

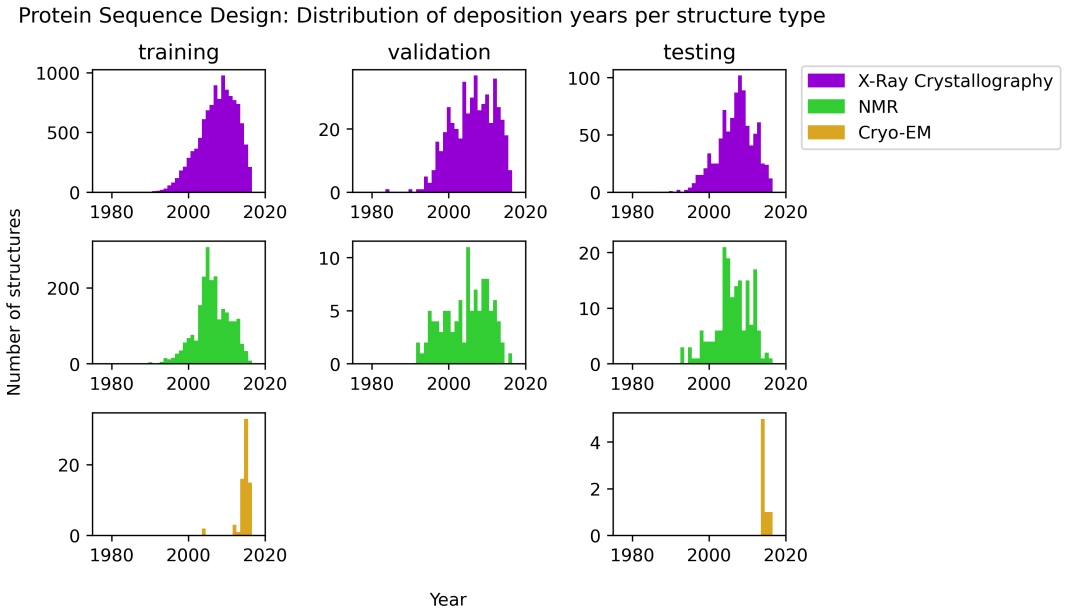

Fig. S4. Distributions of deposition year for each structure in the training, validation, and testing sets for the protein sequence design task. Structures solved in later years tend to be higher quality than those solved in earlier years; the splitting methodology did not introduce any time bias. X-ray structures are dominant in all three splits. No cryo-EM structures were present in the validation set.

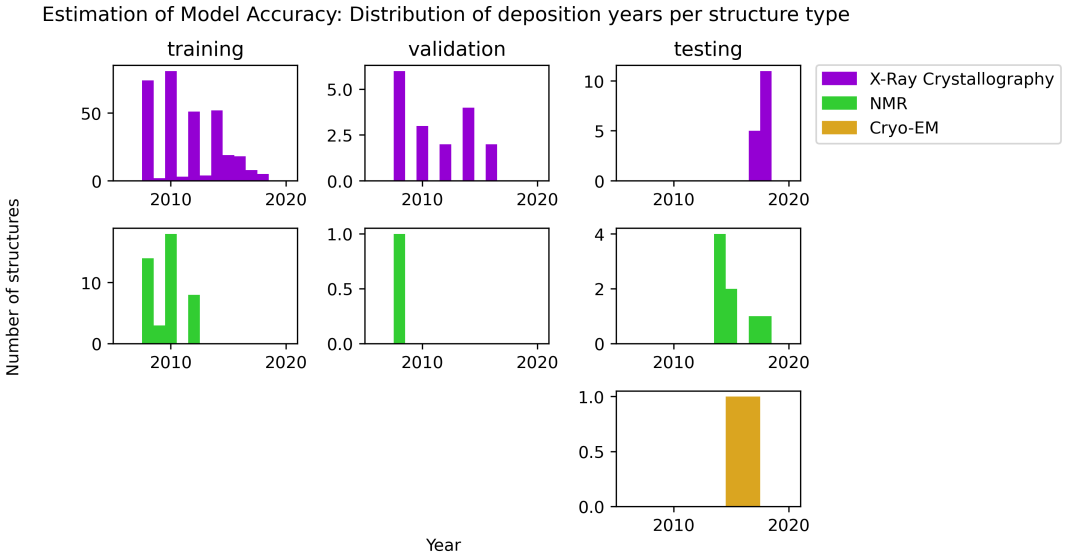

Fig. S5. Distributions of deposition year for each structure in the training, validation, and testing sets for the evaluation of model accuracy task. As per CASP convention, a time-based split was used. X-ray structures are dominant in all three splits. No cryo-EM structures were present in the training or validation sets.

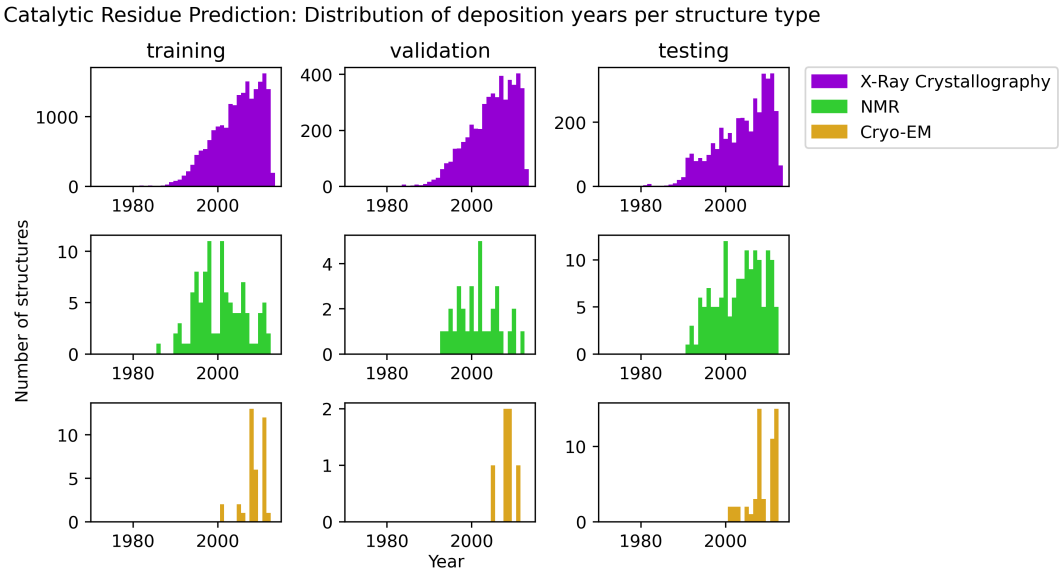

Fig. S6. Distributions of deposition year for each structure in the training, validation, and testing sets for the catalytic residue prediction task. Structures solved in later years tend to be higher quality than those solved in earlier years; the splitting methodology did not introduce any time bias. X-ray structures are dominant in all three splits.

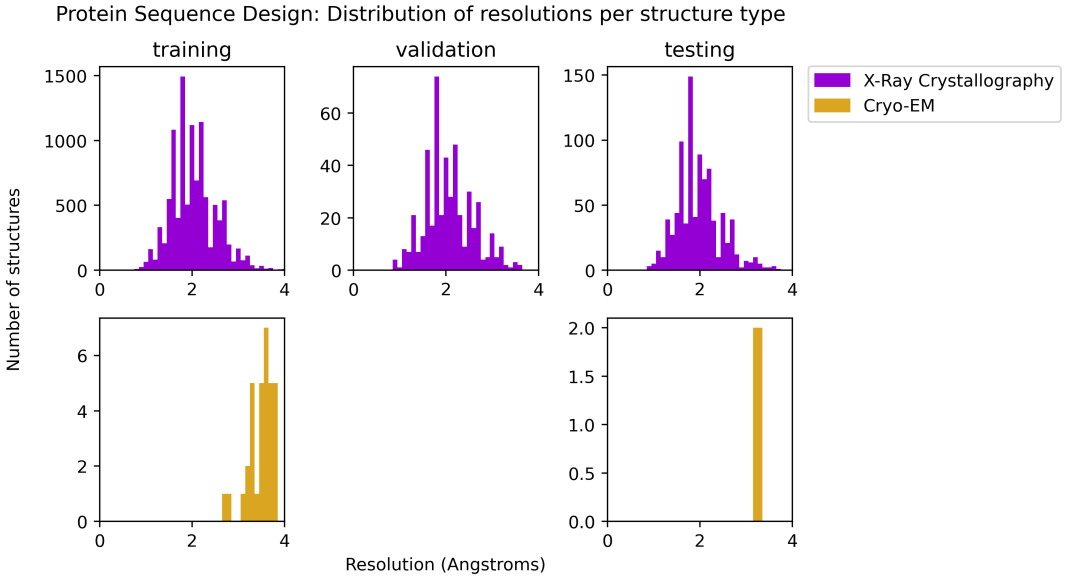

Fig. S7. Distributions of resolution for each X-ray and cryo-EM structure in the training, validation, and testing sets for the protein sequence design task. The splitting methodology did not introduce any resolution bias. X-ray structures are dominant in all three splits. No cryo-EM structures were present in the validation set.

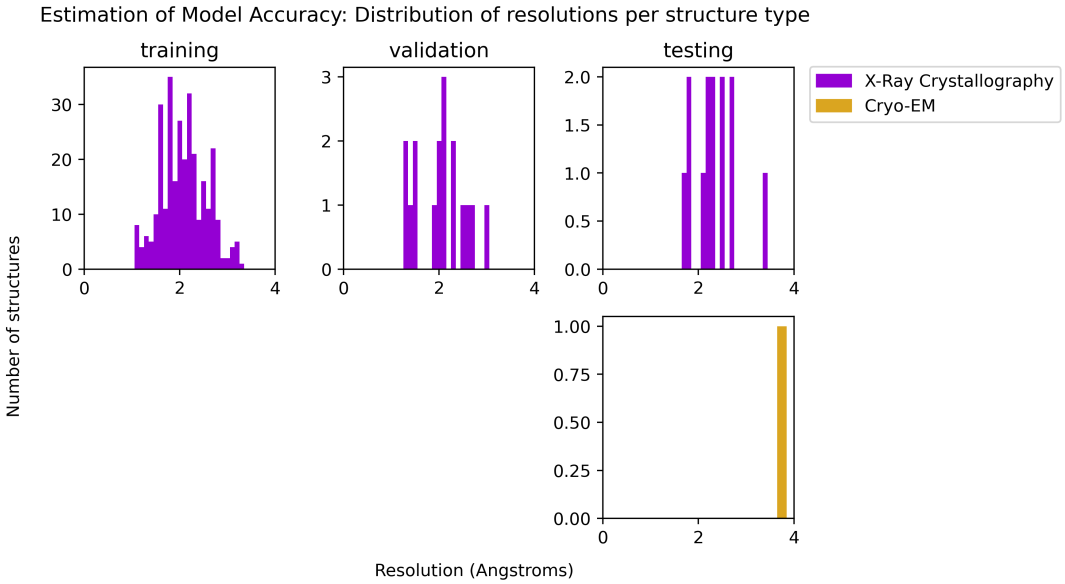

Fig. S8. Distributions of resolution for each X-ray and cryo-EM structure in the training, validation, and testing sets for the estimation of model accuracy task. The splitting methodology did not introduce any resolution bias. X-ray structures are dominant in all three splits. No cryo-EM structures were present in the training or validation sets

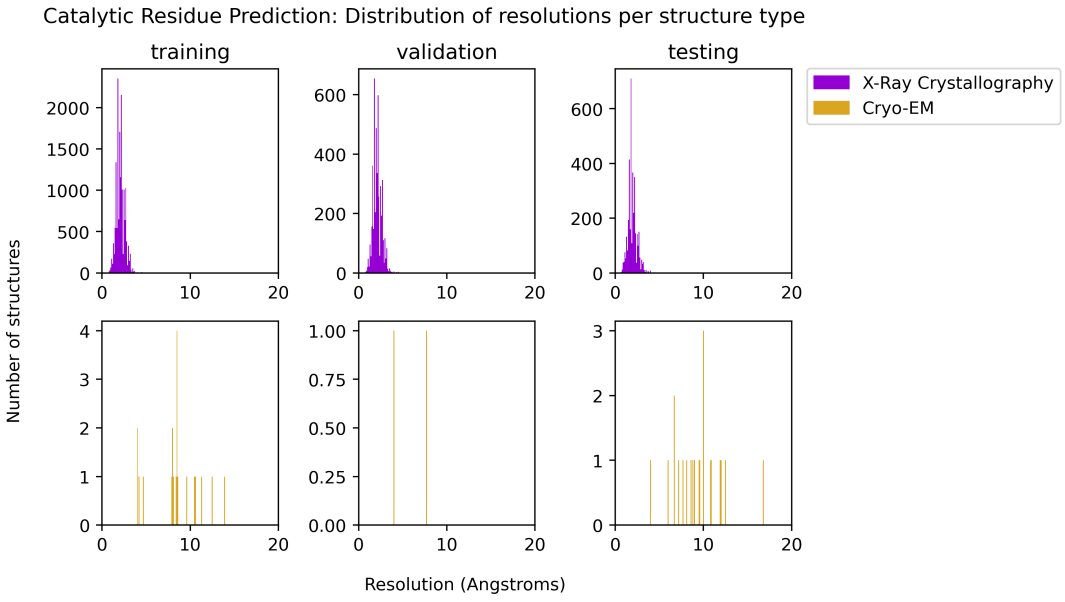

Fig. S9. Distributions of resolution for each X-ray and cryo-EM structure in the training, validation, and testing sets for the catalytic residue prediction task. The splitting methodology did not introduce any resolution bias. X-ray structures are dominant in all three splits.