Supplemental information for Learning probabilistic protein-DNA recognition codes from DNA-binding specificities using structural mappings

Joshua L. Wetzel, Kaiqian Zhang, and Mona Singh*†

*Department of Computer Science and Lewis-Sigler Institute for Integrative Genomics, Princeton University
†Email mona@cs.princeton.edu

1 Supplemental Methods

1.1 Gathering PWM-protein pairs and splitting datasets for testing rCLAMPS

DNA-binding specificities for homeodomain (HD) and Cys2-His2 zinc finger (C2H2-ZF) transcription factors (TFs) were first extracted as position weight matrices (PWMs) from the Cis-BP database (build 2.00) [1]. For HDs, we used all specificities from wildtype TFs containing only a single HD and no alternative DBD. For C2H2-ZFs, we kept only PWMs for which the corresponding protein contained only a short array of two to four C2H2-ZF domains linked by short loops of at most eight amino acids and no alternative DNA-binding domains. C2H2-ZF protein sequences were scanned for matches to PFAM PF00096 using HMMer v.2.3.2 with the default gathering threshold, since it returns more reliable matches for short domains than does HMMer v.3. An additional filter was then applied to remove C2H2-ZF PWM-protein pairs for which the PWM has fewer than 3**n*+1 columns, where *n* is the number of C2H2-ZF domains in the corresponding array; this further helps to avoid examples wherein potentially not all domains in these shorter arrays were engaged with the DNA. In all cases, if there exists more than one specificity for the same TF, the specificity from the most recent publication was used. This ultimately resulted in a single PWM for each of 627 distinct HD TFs and 227 distinct C2H2-ZF TFs.

We further augmented the HD dataset with 30 mutant HDs from the Barrera et al. dataset [2] (restricting only to "bound" mutants wherein a specificity based on at least 10 enriched PBM 8-mers was reported), and 151 characterized synthetic HDs with validated binding activity from Chu et al. by aligning their selected binding sites (Chu et al. [3], Supplemental Table 6) in the same manner described in their publication. The C2H2-ZF dataset was further augmented with PWMs for 36 distinct fly proteins for which the specific portions of PWMs mapping to particular C2H2-ZF domains within these proteins was previously determined experimentally (Enuameh et al. [4], Supplemental_Dataset_1.txt), restricting using the same criteria as above except allowing contiguous arrays of more than four adjacent C2H2-ZF domains as long as for an array of *n* domains a full length corresponding PWM of at least 3**n*+1 columns was available from the Fly Factor Survey website (Dataset A (public)). Specifically, we used the full length PWM versions based on Solexa sequencing. These processing steps ultimately resulted in datasets of 763 HD proteins and 263 C2H2-ZF proteins, each paired with a single corresponding PWM.

When training and testing using the HD protein dataset, rCLAMPS was first run on all 763 of these PWM-protein pairs to establish holdout validation performance. Afterwards, to allow a completely external test set for fair comparison to other state-of-the-art *de novo* HD specificity predictors [3, 5], and for testing our hybrid approach to transfer specificity information from wildtypes to "mutants", we reran rCLAMPS while withholding a random subset of half the specificities from Chu et al., the 30 mutant HDs kept from Barrera et al., and any other Cis-BP HDs whose combinations of DNA-binding residues were identical to any HD from either of these two sets. This resulting training set contained 593 PWM-protein pairs. When testing our hybrid transfer approach, each protein in this training set was considered as a "wildtype" protein and each remaining (held out) protein *y* was considered as a "mutant" if for at least one protein *x* in the training set, *x* and *y* were identical in at least 80% if their HMM match states and *x* and *y* differed in exactly one DNA-contacting residue. When comparing predicted C2H2-ZF specificities to their experimental counterparts, experimental PWMs were first trimmed of columns containing information content lower than 0.25 bits to avoid comparisons for potentially unengaged flanking finger residues.

We note that all input PWMs, protein sequences (or HMM match-state residues), contact maps, and output/predicted PWMs used in our analysis (along with all code needed to generate the recognition codes, mappings, and analyses), are available in the rCLAMPS code repository linked via GitHub and in the Supplemental Materials directly from the *Genome Research* website.

References

[1] Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJ, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, Hughes TR. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443, 2014. ISSN 00928674.

[2] Barrera LA, Vedenko A, Kurland JV, Rogers JM, Gisselbrecht SS, Rossin EJ, Woodard J, Mariani L, Kock KH, Inukai S, Siggers T, Shokri L, Gordân R, Sahni N, Cotsapas C, Hao T, Yi S, Kellis M, Daly MJ, Vidal M, Hill DE, Bulyk ML. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*, 351(6280):1450–1454, 2016. ISSN 10959203.

[3] Chu SW, Noyes MB, Christensen RG, Pierce BG, Zhu LJ, Weng Z, Stormo GD, Wolfe SA. Exploring the DNA-recognition potential of homeodomains. *Genome Research*, 22(10):1889–1898, 2012. ISSN 10889051.

[4] Enuameh MS, Asriyan Y, Richards A, Christensen RG, Hall VL, Kazemian M, Zhu C, Pham H, Cheng Q, Blatti C, Brasefield JA, Basciotta MD, Ou J, McNulty JC, Zhu LJ, Celniker SE, Sinha S, Stormo GD, Brodsky MH, Wolfe SA. Global analysis of Drosophila Cys2-His2 zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. *Genome Research*, 23(6):928–940, 2013. ISSN 10889051.

[5] Christensen RG, Enuameh MS, Noyes MB, Brodsky MH, Wolfe SA, Stormo GD. Recognition models to predict DNA-binding specificities of homeodomain proteins. *Bioinformatics*, 28(12):i84–i89, 2012. ISSN 13674803.

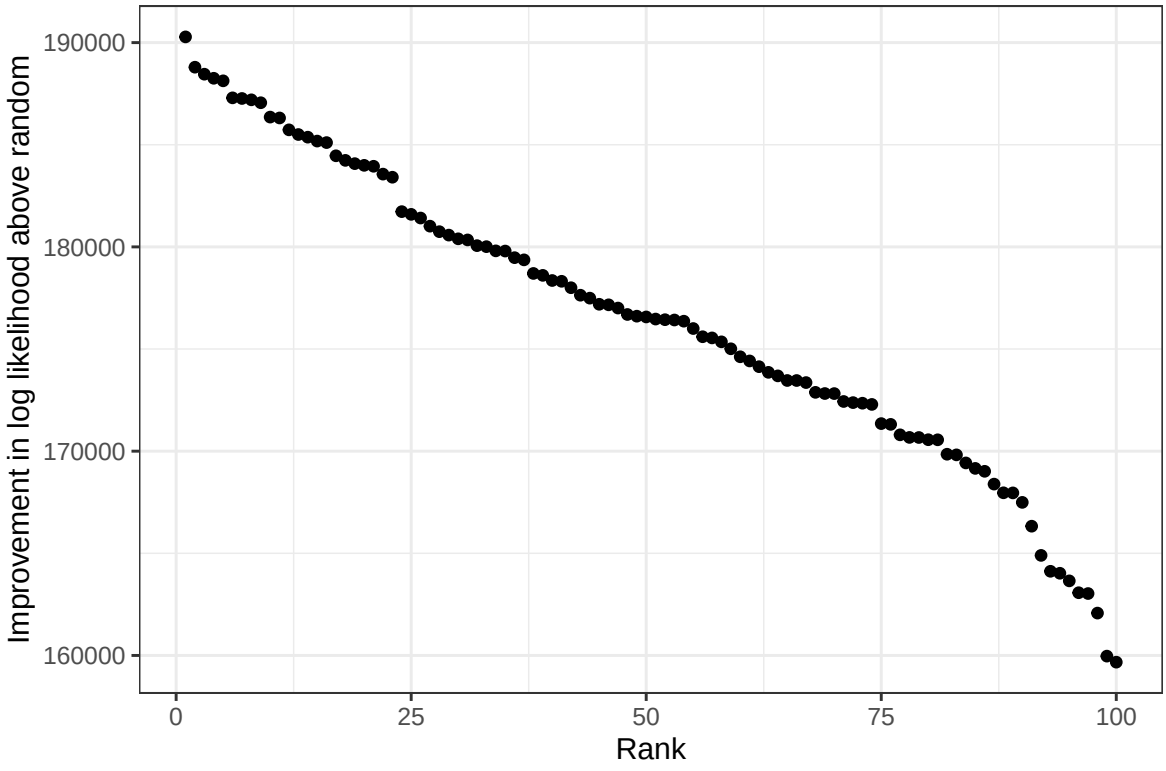

Figure S1: **Performance of rCLAMPS across sampling runs.** We ran our Gibbs sampling routine 100 times on the 763 homeodomain PWM-protein pairs, providing different random starting registrations for the PWMs each time. For each of these 100 samplings, we show the improvement in likelihood (y-axis) of each mapping output after iterating the Gibbs sampler, relative to the expected likelihood of a random mapping. In particular, the expected likelihood of a random mapping was computed by training 100 models using the same 763 homeodomain PWM-protein pairs, each with a different random starting registration and without iterating the Gibbs sampler. We note that the top 10% of the mappings (ranked by likelihood, x-axis) show at most 2% lower improvement above random than the best mapping, indicating that although the precise mapping found is dependent on the starting point, high quality models can be found from many random starting points, indicating robustness of our framework.

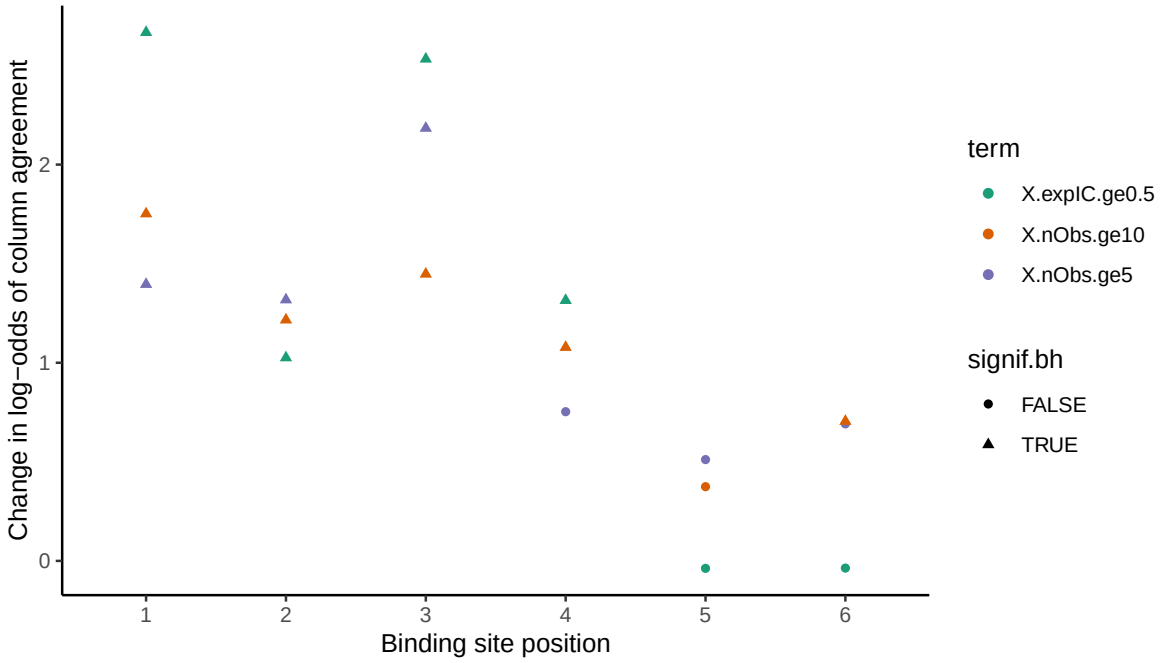

Figure S2: **Higher experimental column IC and more amino acid contact observations increase odds of accurate predictions.** For each binding site position, we evaluated the impact of several factors on the probability that rCLAMPS would accurately predict experimental PWM columns in that position. In particular, for each binding site position (x-axis), we partitioned predictions for that position according to several factors and ran independent logistic regressions for each factor to estimate the impact of that factor on whether the prediction was correct (i.e., PCC ≥ 0.5 between predicted and actual specificities), as measured by the coefficient learned for that factor, which corresponds to the log-odds (y-axis) of rCLAMPS producing an accurate prediction. In particular, we partitioned predictions according to whether the information content of the corresponding experimental PWM column was at least 0.5 ('X.expIC.ge0.5', green), whether the amino acid residues contacting the binding site position being predicted had been observed at least 5 times across our dataset ('X.nObs.ge5', purple), or at least 10 times across our dataset ('X.nObs.ge10', red). In nearly all cases, the odds of accurate prediction increases (i.e., the learned coefficient is > 0) if the condition is met. For each case, the estimated impact on log-odds was tested for significance (Wald test), then multiple hypothesis testing correction was performed using the Benjamini Hochberg procedure to control the false discovery rate (FDR). Significant effects (FDR < 0.05) are depicted as triangles.

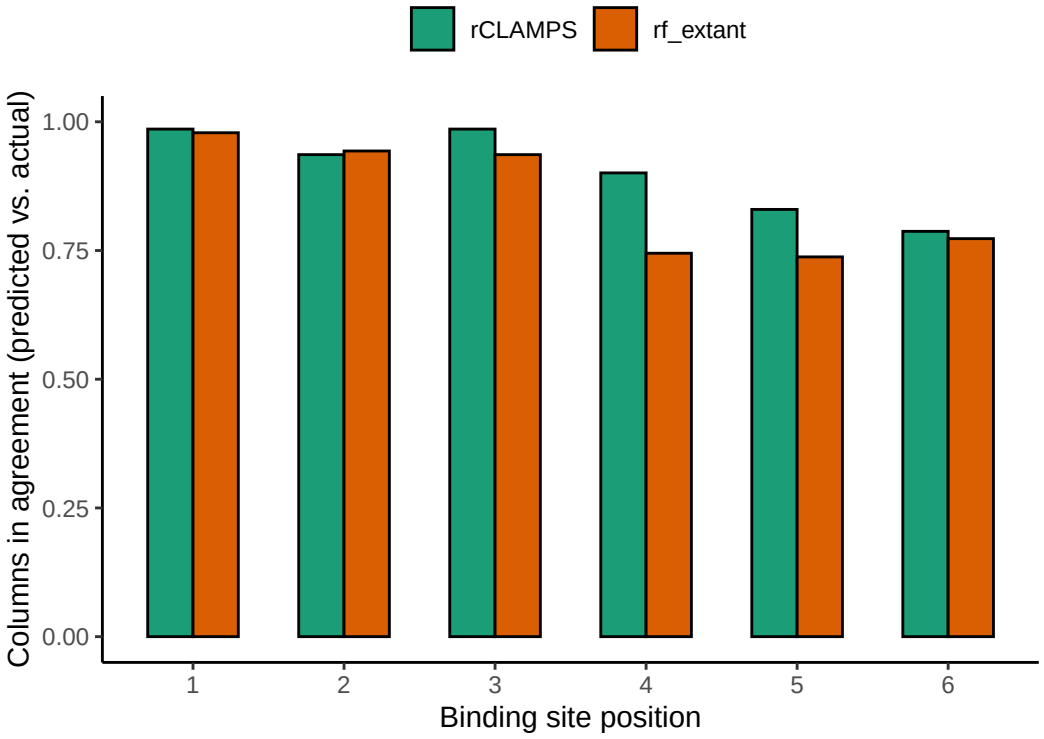

Figure S3: **Per binding site position comparison of rCLAMPS to `rf_extant`.**We compare the performance of rCLAMPS (green) and the random forest method `rf_extant` (red) when predicting HD DNA-binding specificities *de novo*, considering 141 HD proteins unseen in training by either method. Agreement for all paired (predicted vs. experimental) PWM columns are plotted per binding site position (x-axis). We compare the fraction of columns in agreement (y-axis) at a 0.5 PCC threshold for each binding site position.

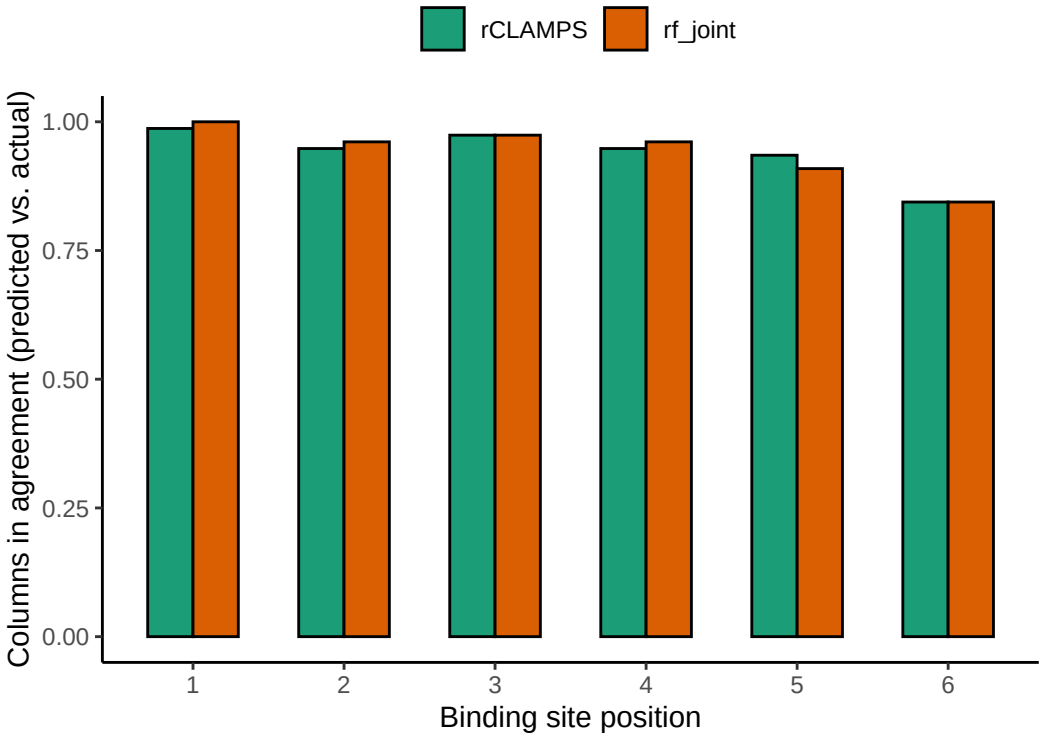

Figure S4: **Per binding site position comparison of rCLAMPS to `rf_joint`.** We compare the performance of rCLAMPS (green) and the random forest method `rf_joint` (red) when predicting HD DNA-binding specificities *de novo*, considering 77 HD proteins unseen in training by either method. Agreement for all paired (predicted vs. experimental) PWM columns are plotted per binding site position (x-axis). We compare the fraction of columns in agreement (y-axis) at a 0.5 PCC threshold for each binding site position.

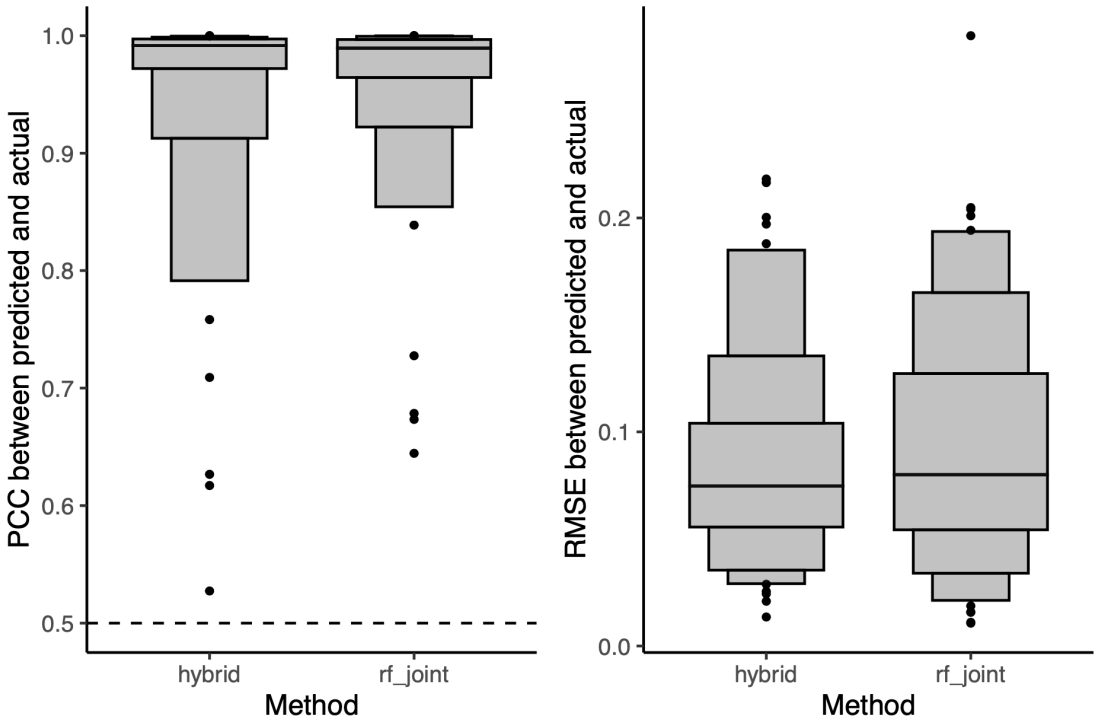

Figure S5: **Comparison of hybrid transfer approach versus `rf_joint`.** For each of the 19 "mutant" homeodomains (defined as above in Supplemental Methods 1.1) that were also not part of the `rf_joint` training set, we compare the actual specificity for a column and that predicted by our hybrid approach ('hybrid') and the `rf_joint` model (rf_joint), considering the 76 columns for which column transfer was possible. We display the PCC agreement scores (left, *y*-axis) and root mean squared error (RMSE) agreement scores (right, *y*-axis) visualized as letter-value (or boxen) plots. In a letter-value plot, the widest box shows the value range spanned by half the data (from the 25th to 75th percentiles), while each successively narrower pair of boxes together show the value range spanned by half the remaining data. **(Left)** All predictions by both methods are accurate at at our nominal agreement threshold (PCC $\geq$ 0.5, dashed line), with hybrid transfer predictions showing slightly higher upper quartile, median, and lower quartile scores than rf_joint, but with lower scores in the lowest quartile. **(Right)** Considering RMSE as the agreement measure, the hybrid approach shows lower (better) median, upper quartile, and maximum scores than `rf_joint`, but slightly higher lower quartile and minimum scores.

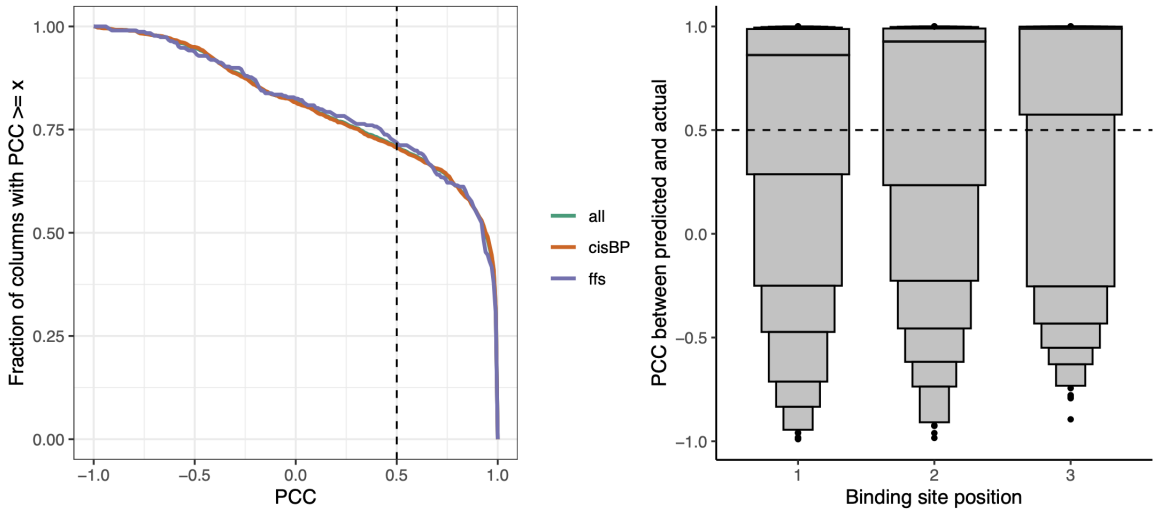

Figure S6: **Probabilistic DNA recognition code for C2H2-ZFs derived from automatically inferred structural mappings has excellent de novo predictive performance.** We compare agreement between predicted PWM columns and corresponding experimental PWM columns for 263 C2H2-ZF proteins in a strict holdout validation setup analogous to that shown for the homeodomains in main Figure 2. **(Left)** Considering all C2H2-ZF binding site positions together, as different thresholds of PCC are considered (x-axis), the fraction of column pairs that have PCC greater than this threshold is plotted (y-axis). Our nominal threshold for agreement (PCC $\geq$ 0.5) is shown as a dashed vertical line. We consider all PWM columns together (all, green), only columns from Fly Factor Survey (ffs, purple), and only columns from Cis-BP (cisBP, orange). **(Right)** For each binding site position within the C2H2-ZF contact map (x-axis, overlapping positions 1 and 4 are combined as position 1), we display the PCC agreement scores (y-axis) for the paired columns at that binding site position, visualized as letter-value (or boxen) plots.