bbcontacts – Supplementary information

Jessica Andreani and Johannes Söding

S1 Supplementary methods

S1.1 Datasets

BetaSheet916 (Cheng and Baldi, 2005) consists of 916 protein chains with an available X-ray structure of resolution below 2.5 Å. These chains contain 31,638 β-residue contacts distributed into 4519 antiparallel β-strand contacts, 2214 parallel β-strand contacts and 1429 isolated β-bridges.

BetaSheet1452 (Savojardo *et al.*, 2013) was built from the structures deposited in the Protein Data Bank after 2004 but using a procedure similar to the BetaSheet916 building procedure. BetaSheet1452 involves 56,552 β-residue contacts distributed into 3937 antiparallel β-strand contacts, 7892 parallel β-strand contacts and 2412 isolated β-bridges.

To build our training dataset, we extracted all CATH domains that did not belong to any of the fold groups identified in the test datasets in CATH v3.5. This set of 22,563 domains belonging to 864 fold groups was then filtered to reduce redundancy. For this purpose, we used the `pdbfilter.pl` script from the HH-suite (Remmert *et al.*, 2011) with parameters -cov 0 -e 0.01 -id 0 (no sequence identity restriction for filtering, but the minimum E-value between any two representative sequences is 0.01 and no minimum coverage was applied when discarding redundant sequences). Among the 1482 PDB domains in this redundancy-filtered dataset, 943 domains containing β-contacts form our training dataset (867 X-ray structures with resolution below 3.5 Å and 76 NMR structures). These 943 domains contain 19,339 β-contacts: 2511 parallel β-contacts, 16,041 antiparallel β-contacts and 787 β-bridges.

Because not all chains in BetaSheet916 and BetaSheet1452 were fully annotated in CATH v3.5, there might remain some redundancy between the training dataset and the test dataset. We verified that the results for BetaSheet916 and BetaSheet1452 did not deteriorate when the dataset was restricted to the subset of each dataset containing all chains fully annotated in CATH v3.5 (and thus non-redundant with the training dataset) (see section S2.1 below and Figure S1). In Figures S21, S22, S23, S24 and S25, we also show results for the training dataset and the test dataset BetaSheet1452.

Because bbcontacts relies on correlated mutations and thus predicts side-chain and not backbone contacts, the positions involved in β-bulges were adjusted to reflect the expected pattern: for a β-bulge between res1 and res2 (in one strand) and resX (in the other strand), all three side-chains must point in the same direction with respect to the plane formed by the β-sheet.

S1.2 Data used for HMM training

To build the multiple sequence alignments, we started from sequences based on the ATOM records of the PDB files: this allowed us to have perfect matching between the structure,

the DSSP assignment and the first sequence of the MSA for each protein in the training dataset, while the two test datasets were unaffected by this choice as by construction they contain only proteins with no backbone interruption (Cheng and Baldi, 2005; Savojardo *et al.*, 2013).

We first ran HHblits v2.0.15 (Remmert *et al.*, 2011) against the uniprot20 database (dated March 2013), with options -all -maxfilt 100000 -realign_max 100000 -B 100000 -Z 100000, thus avoiding any filtering in order to retrieve as many homologous sequences as possible. We then performed a filtering step using HHfilter with options -id 90 -neff 15 -qsc -30 (each alignment is filtered down to 90% sequence identity).

The distribution of the number of sequences in the MSAs for the training dataset and the two test datasets is provided in Figure S2.

The secondary structure predictions were obtained with PSIPRED (Jones, 1999), as implemented in the `addss.pl` script from the HH-suite (Remmert *et al.*, 2011). This means that the MSAs were first filtered down to Neff ≤ 7 and that the procedure included fine-tuning of the secondary structure predictions with psipass2.

Direct coupling predictions were obtained with CCMpred (Seemayer *et al.*, 2014) run with the default options, including initial sequence reweighting and final post-processing using the average-product correction (Dunn *et al.*, 2008). The MSAs were not filtered to remove columns or rows with many gaps.

When building MSAs of reduced diversity for the training dataset, rather than sampling from the alignment at random, we ran HHfilter (Remmert *et al.*, 2011) with different values of the qsc parameter, describing the entropy per column in the MSA. We tried different qsc values through a dichotomic search, until the filtered alignment contained the number of sequences expected for a given $\eta \in \{0.05, 0.1, 0.2, ..., 1.0, 1.2\}$. For an initial MSA diversity value of η_0, diversity-filtered alignments can be obtained for each $\eta < \eta_0$.

The respective numbers of domains and numbers of β-contacts in each diversity-filtered dataset are given in Table S1.

Table S1: Number of domains (#domains), number of parallel residue-residue β-contacts (#parallel) and number of antiparallel residue-residue β-contacts (#antiparallel) in each diversity-filtered dataset

η	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.2
#domains	805	621	384	235	155	103	68	48	35	23	18	10
#parallel	2144	1606	1148	588	287	186	102	57	49	33	23	9
#antiparallel	11888	8470	4549	2601	1675	962	656	409	281	208	160	84

S1.3 HMM parameters

The HMM emission probability $e_z(i,j)$ for a given HMM state z and a given position (i,j) contains a product of two terms: one term based on couplings $ec_z(i,j)$ and one term based on secondary structure $ess_z(i,j)$.

S1.3.1 Coupling-based emissions.

The coupling-based part of the emission probability at position (i,j) was expressed as the product of three odds-ratios relative to the background: one for the central coupling at position (i,j) and one for each of the two couplings at the positions adjacent to (i,j) that belong to the secondary diagonals of the pattern. This can be written as follows for the case of parallel β-strands, for any HMM state z apart from the "start" and "end" states:

$$ec_z(i,j) = \frac{p(x_{i,j}|z,i,j)}{p(x_{i,j}|bg,i,j)} \frac{p(x_{i+1,j-1}|z,i,j)}{p(x_{i+1,j-1}|bg,i,j)} \frac{p(x_{i-1,j+1}|z,i,j)}{p(x_{i-1,j+1}|bg,i,j)} \tag{1}$$

where $x_{i,j}$ denotes the coupling value at position (i,j) and bg denotes the background.

Still for the case of parallel β-strands, we fitted only 3 distributions:

- $p(x_{i,j}|z,i,j)$ for the main diagonal of the pattern,
- $p(x_{i+1,j-1}|z,i,j) = p(x_{i-1,j+1}|z,i,j)$ for the secondary diagonals of the pattern,
- $p(x_{i,j}|bg,i,j) = p(x_{i+1,j-1}|bg,i,j) = p(x_{i-1,j+1}|bg,i,j)$ for the background.

The case of antiparallel β-strands was treated similarly, except that the positions on the secondary diagonals are $(i+1,j+1)$ and $(i-1,j-1)$.

To fit the distributions, the diversity-filtered alignments were used. After centering each coupling distribution at zero by subtracting a shift parameter x_0, we fitted it using a combination of two transformed Gamma functions, one for positive coupling values x_+ and one for negative coupling values x_-. For instance, in the case of the main diagonal fit:

$$p(x_{i,j}|z,i,j) = \begin{cases} f_+(x_{i,j} - x_0) & \text{if } x_{i,j} \geq x_0 \\ f_-(-x_{i,j} + x_0) & \text{if } x_{i,j} < x_0 \end{cases} \tag{2}$$

where

$$f_+ : x_+ \mapsto w_+ \frac{{b_+}^{-\frac{1}{\alpha_+}}}{\Gamma\left(\frac{1}{\alpha_+}\right)} \frac{\alpha_+}{\beta_+} \exp\left(-\frac{\left(\frac{|x_+|}{\beta_+}\right)^{\alpha_+}}{b_+\left[1+\left(\frac{|x_+|}{\beta_+}\right)^{\alpha_+}\right]}\right) \tag{3}$$

$$f_- : x_- \mapsto (1-w_+) \frac{{b_-}^{-\frac{1}{\alpha_-}}}{\Gamma\left(\frac{1}{\alpha_-}\right)} \alpha_- \exp\left(-\frac{|x_-|^{\alpha_-}}{b_-}\right) \tag{4}$$

To describe each of the transformed Gamma fits for a given value of η, we thus need 7 parameters: the shift x_0 needed to center the coupling distribution, the relative weight of the positive and negative sides w_+, plus b_- and α_- for negative couplings and b_+, α_+ and β_+ for positive couplings.

To describe the dependency of the coupling distributions on η, we expressed each parameter as a function of the alignment diversity η: first, the shift was fitted as a quadratic function of η by linear regression and the fitted shift was subtracted from all coupling values. All remaining parameters were expressed as linear functions of η.

The optimization was performed using a global maximum likelihood estimation with the BFGS algorithm, using multiple initialization values for the parameters (32 for the background fit and over 1,000 for the signal fits). We performed each round of optimization using the training data (predicted coupling matrices) generated for all 12 values of η. The training was performed after local background correction of the coupling matrices with $S = 10$, but the resulting parameters were found to be very similar when the matrices without local background correction were used for training. We picked the fit with the maximum likelihood over all runs. The final fits are illustrated in Figure S4.

The final number of parameters for the coupling-based emission probabilities is thus 90: 2 (parallel/antiparallel) times 3 (background, main-diagonal signal, secondary-diagonal signal) times 15 (6 parameters with a linear dependency on η and one parameter with a quadratic dependency).

Because the coupling densities are fitted on cases with alignments from a limited range of η values, and because the parameters are expressed as linear or quadratic functions of η, the formulae obtained for each parameter as a function of η do not give acceptable results when η is either too small or too large (for instance, the b parameters can become negative, which is unacceptable). On the other hand, we observed that for very low values of η, the coupling distributions derived from true β-contact patterns are almost superimposed with the background distributions (the noise overcomes the signal) and for high values of η, the distributions do not change much. Therefore, we set boundaries on the values of η as follows:

- if $\eta < 0.022$, only the secondary structure part of the emission probabilities is used (we assume that there is no coupling signal so that the odds-ratios in the coupling-based emissions are always equal to 1),
- if $\eta > 1.98$, we set $\eta = 1.98$ for calculating the coupling-based emissions.

In addition, because the fits are not perfect representations of the coupling densities, and because the range of coupling values observed in the training dataset is limited, the fits can display unexpected behaviors for relatively low or relatively high coupling values. Typically, for low couplings, $p(x_{i,j}|z)/p(x_{i,j}|bg)$ can become larger than 1, which is unexpected. For high couplings, $p(x_{i,j}|z)/p(x_{i,j}|bg)$ can become smaller than 1, which is also unexpected. To avoid such problems, we set (for both the main-diagonal and secondary-diagonal odds-ratios):

$$\begin{cases} \frac{p(x_{i,j}|z)}{p(x_{i,j}|bg)} = 1 \text{ if } \frac{p(x_{i,j}|z)}{p(x_{i,j}|bg)} > 1 \text{ and } x_{i,j} < x_0(bg) \\ \frac{p(x_{i,j}|z)}{p(x_{i,j}|bg)} = 1 \text{ if } \frac{p(x_{i,j}|z)}{p(x_{i,j}|bg)} < 1 \text{ and } x_{i,j} > x_0(maindiag) + 0.1 \end{cases}$$

where $x_0(bg)$ is the fitted shift for the background distribution and $x_0(maindiag)$ is the fitted shift for the main diagonal positions. This has a very minor effect on the vast majority of our results but makes bbcontacts more robust.

S1.3.2 Secondary-structure-based emissions.

Here, we distinguished between different HMM states (denoted by z) because they exhibit a different behaviour with respect to observed secondary structure predictions.

We tested two different versions of the secondary-structure-based emissions $ess_z(i,j)$.

The first version (hereafter called "non-conditional" for simplicity) was based on the probability $p(\sigma_i, \sigma_j|z)$ of observing a pair of secondary structure states (σ_i, σ_j) in state z at position (i,j):

$$ess_z(i,j) = \frac{p(\sigma_i, \sigma_j|z)}{p(\sigma_i, \sigma_j|bg)} \tag{5}$$

The second version (hereafter called "conditional") was based on the probability $p(\sigma_i, \sigma_j|\sigma_{i_{\text{prev}}}, \sigma_{j_{\text{prev}}}, z)$ of observing a pair of secondary structure states (σ_i, σ_j) in state z at position (i, j) given that we additionally observed secondary structure states $(\sigma_{i_{\text{prev}}}, \sigma_{j_{\text{prev}}})$ at the previous position $(i_{\text{prev}}, j_{\text{prev}})$, where

$$(i_{\text{prev}}, j_{\text{prev}}) = \begin{cases} (i-1, j-1) & \text{for the parallel case,} \\ (i-1, j+1) & \text{for the antiparallel case.} \end{cases} \tag{6}$$

In order to reduce the number of parameters necessary to describe the secondary-structure-based "conditional" emission probabilities, we used the following factorization:

$$p(\sigma_i, \sigma_j|\sigma_{i_{\text{prev}}}, \sigma_{j_{\text{prev}}}, z) = p(\sigma_i|\sigma_{i_{\text{prev}}}, z)p(\sigma_j|\sigma_{j_{\text{prev}}}, z) \tag{7}$$

We verified that the numerical difference between the initial and the factorized form was small.

The "conditional" secondary-structure-based emissions can thus be expressed as:

$$ess_z(i, j) = \frac{p(\sigma_i|\sigma_{i_{\text{prev}}}, z)}{p(\sigma_i|\sigma_{i_{\text{prev}}}, bg)} \frac{p(\sigma_j|\sigma_{j_{\text{prev}}}, z)}{p(\sigma_j|\sigma_{j_{\text{prev}}}, bg)} \tag{8}$$

We also added pseudocounts derived from the non-conditional probability distribution to the conditional probabilities, i.e. we replaced

$$p(\sigma_i|\sigma_{i_{\text{prev}}}, z) = \frac{N_{\sigma_i, \sigma_{i_{\text{prev}}}, z}}{N_{\sigma_{i_{\text{prev}}}, z}} \tag{9}$$

by a term including N_0 counts from the non-conditional distribution. The resulting emission probabilities are:

$$ess_z(i, j) = \frac{\frac{N_{\sigma_i, \sigma_{i_{\text{prev}}}, z} + N_0 \frac{N_{\sigma_i, z}}{N_z}}{N_0 + N_{\sigma_{i_{\text{prev}}}, z}}}{\frac{N_{\sigma_i, \sigma_{i_{\text{prev}}}, bg} + N_0 \frac{N_{\sigma_i, bg}}{N_{bg}}}{N_0 + N_{\sigma_{i_{\text{prev}}}, bg}}} \frac{\frac{N_{\sigma_j, \sigma_{j_{\text{prev}}}, z} + N_0 \frac{N_{\sigma_j, z}}{N_z}}{N_0 + N_{\sigma_{j_{\text{prev}}}, z}}}{\frac{N_{\sigma_j, \sigma_{j_{\text{prev}}}, bg} + N_0 \frac{N_{\sigma_j, bg}}{N_{bg}}}{N_0 + N_{\sigma_{j_{\text{prev}}}, bg}}} \tag{10}$$

where the $N_{\ldots}$ terms represent counts observed in the training dataset (e.g. $N_{\sigma_i, z}$ is the number of counts observed for secondary structure state σ_i in state z and $N_{\sigma_i, \sigma_{i_{\text{prev}}}, z}$ is the number of counts observed for secondary structure state σ_i and previous secondary state $\sigma_{i_{\text{prev}}}$ in state z).

The number of pseudocounts N_0 was optimized on the training dataset, as illustrated in Figure S7.

Because we found that the secondary structure states for the coupling matrix cells situated immediately before and immediately after a β-strand interaction contain information about the likelihood to start and end this interaction, there is also a secondary-structure-based emission term in the Viterbi initialization step ("start" state) and in the Viterbi termination step ("end" state). The term in the initialization step is always a non-conditional probability, for a clean termination of the chain-rule based product of emissions:

$$p(\sigma_{i_{\text{end}}}|\sigma_{i_{\text{end}}-1}, z) * \ldots * p(\sigma_i|\sigma_{i_{\text{prev}}}, z) * \ldots * p(\sigma_{i_{\text{start}}+1}|\sigma_{i_{\text{start}}}, z) * p(\sigma_{i_{\text{start}}}|z)$$

This initialization term can also be seen as a prior based on secondary structure.

For each situation (DSSP-based emissions or PSIPRED-based emissions), the number of parameters is therefore 415: the number of pseudocounts N_0; 216 parameters for the

non-conditional probabilities (2 directions times 12 states (11 HMM states plus the background) times 9 possible combinations of (σ_i, σ_j)); and 198 parameters for the conditional probabilities (2 directions times 11 states (11 HMM states minus the start state plus the background) times 9 possible combinations of $(\sigma_i, \sigma_{i_{\text{prev}}})$). Note that for the DSSP-based case, many of these emission parameters are 0 or 1 since β-β contacts can only be detected between residues assigned as "E" by DSSP.

S1.3.3 Prior probability distribution depending on sequence separation.

We introduced a prior for starting a β-strand interaction depending on the sequence separation between the first pair of interacting residues. The prior contains explicit probabilities to have a contact starting at a sequence separation of up to 12; the probability is then modeled as a linear function of the sequence separation between 13 and 20 and as an exponentially decreasing function starting from a sequence separation of 21:

$$prior(i-j) = \begin{cases} p_{i-j} & \text{if } |i-j| \leq 12 \\ l_1 + |i-j| * l_2 & \text{if } 13 \leq |i-j| \leq 20 \\ e_1 + e_2 * e^{\frac{-|i-j|}{e_3}} & \text{if } |i-j| \geq 21 \end{cases} \tag{11}$$

The corresponding 17 parameters (p_1, p_2, ..., p_{12}, l_1, l_2, e_1, e_2 and e_3) were trained independently for parallel and antiparallel β-contacts. In addition, for the DSSP-based predictions, the training was limited to regions of the coupling matrix where both residues belong to a β-strand, because no HMM path can be detected outside of these regions. The fitted parameters are shown in Figure S5.

We also introduced constraints to prevent decoding in regions of the coupling matrix too close to the diagonal. For this purpose, we always "mask" a region around the diagonal, i.e. we set the emission probabilities to 0 for all positions (i, j) in this region and for all states. This region contains all pairs of positions with a sequence separation of up to (and including) 1 for antiparallel contacts and 6 for parallel contacts.

S1.4 HMM decoding

The local Viterbi algorithm consists of four major steps: initialization, recursion, termination and back-tracing.

In the initialization step, the Viterbi variables $V[i, j, \text{start}]$ are initialized for all positions (i, j) in the coupling matrix. Because the coupling matrices are symmetric, only positions where $i > j$ can receive non-zero Viterbi scores $V[i, j, z]$ for $z \notin \{\text{start}, \text{end}\}$. To make the implementation easier, when going from the start state to the first state in a β-contact, we take a step in the coupling matrix:

$$\begin{cases} (i,j) \to (i+1, j+1) & \text{for the parallel case} \\ (i,j) \to (i+1, j-1) & \text{for the antiparallel case} \end{cases}$$

This means that $V[i, j, \text{start}]$ also has to be initialized for $i \in \{-1, 0, \ldots, L\}$ and $j = -1$ for the parallel case and for $i = j$ for the antiparallel case (where L is the protein length).

In principle, all positions should receive an initial probability of 1, but the priors described above (secondary structure prior and prior depending on sequence separation) are also applied during this initialization step:

$$V[i, j, \text{start}] = \begin{cases} prior(i-j) * p(\sigma_{i_{\text{start}}}|\text{start}) * p(\sigma_{j_{\text{start}}}|\text{start}) & \text{for the parallel case} \\ prior(i-j+2) * p(\sigma_{i_{\text{start}}}|\text{start}) * p(\sigma_{j_{\text{start}}}|\text{start}) & \text{for the antiparallel case} \end{cases} \tag{12}$$

The reason for having $prior(i-j+2)$ in the antiparallel case is that we take a step in the coupling matrix when going from the start state to the first state in a β-contact, so the sequence separation (measuring the distance to the diagonal in the coupling matrix) is unchanged for the parallel case and increased by 2 for the antiparallel case.

In the recursion step, all probabilities $V[i,j,z]$ for $z \notin \{\text{start}, \text{end}\}$ are calculated using the transition and emission probabilities:

$$V[i,j,z] = e_z(i,j) \max_k (V[i_{\text{prev}}, j_{\text{prev}}, k] * t[k][z]) \tag{13}$$

where k is any of the HMM states apart from the end state, $t[k][z]$ is the transition probability from state k to state z, and the previous Viterbi score $V[i_{\text{prev}}, j_{\text{prev}}, k]$ is taken from position

$$(i_{\text{prev}}, j_{\text{prev}}) = \begin{cases} (i-1, j-1) & \text{in the parallel case if } z \notin \{bulgei_1, bulgei_2, bulgej_1, bulgej_2\} \\ (i, j-1) & \text{in the parallel case if } z \in \{bulgei_1, bulgei_2\} \\ (i-1, j) & \text{in the parallel case if } z \in \{bulgej_1, bulgej_2\} \\ (i-1, j+1) & \text{in the antiparallel case if } z \notin \{bulgei_1, bulgei_2, bulgej_1, bulgej_2\} \\ (i, j+1) & \text{in the antiparallel case if } z \in \{bulgei_1, bulgei_2\} \\ (i-1, j) & \text{in the antiparallel case if } z \in \{bulgej_1, bulgej_2\} \end{cases}$$

In the termination step, the $V[i,j,\text{end}]$ probabilities are calculated for all positions where $i > j$. Like with the start state, to make the implementation easier, we take a step in the coupling matrix between the last β-contact and the end state, so that the (i,j) position corresponding to the end state is not part of the final path. The formula for $V[i,j,end]$ includes the classical maximum over states where the path can end and a secondary-structure-based emission term, as mentioned above (section S1.3.2):

$$V[i,j,\text{end}] = ess_{\text{end}}(i,j) \max_k (V[i_{\text{prev}}, j_{\text{prev}}, k] * t[k][\text{end}]) \tag{14}$$

In the recursion and termination steps, pointers are used to keep track of the most likely sequences of states.

The initialization, recursion and termination steps of the Viterbi decoding are performed separately for the parallel and antiparallel directions, but all Viterbi scores are then merged before the fourth and final back-tracing step. In the back-tracing step, all $V[i,j,\text{end}]$ probabilities (Viterbi scores) are sorted in decreasing order. The first (most likely) path, corresponding to the highest $V[i,j,\text{end}]$ probability, is retrieved by back-tracing through the saved pointers and saved. Then, we cross out a region corresponding to a "corridor" around this path (we cross out all residue pairs belonging to the path, plus all residue pairs within $\pm$ 3 residues of those belonging to the path) in the Viterbi matrix corresponding to the direction of the path (parallel or antiparallel), i.e. we do not take into account any more Viterbi probabilities for this region and this direction. This avoids retrieving many suboptimal versions of a contact between the same β-strands. The next path that does not contain any crossed-out residues is then saved and a region around this path is crossed-out.

We proceed iteratively in this manner until we reach a given Viterbi score threshold. This threshold is chosen to be low enough that the precision-recall curve shows only a precision drop after this threshold and no more gain in recall, but not too low for computational efficiency. It is adjusted depending on the parameters used to run bbcontacts (DSSP or PSIPRED-based predictions, PSM triggering).

For numerical stability, all probabilities are expressed in logarithmic space.

S1.4.1 Prediction-shortening mode (PSM)

In practice, we apply a decrease by 0.3 per PSM iteration in all log-scale transition probabilities, except for the transitions to the end state that are used to maintain the sum of transition probabilities leaving any state equal to 1. If paths exceeding a length of 50 are predicted, then the decrease in all log-scale transition probabilities is 0.6 per iteration to speed-up the PSM process. A maximum of 20 iterations is also set to limit the runtime of bbcontacts when PSM gets triggered.

S1.5 Evaluation

Residue-level evaluation is straightforward in all cases: a pair of residues predicted as a β-contact (i.e. belonging to one of the accepted paths) is counted as a true positive if it is actually a β-contact (defined by DSSP) and as a false positive if it is not. False negatives are all the true β-contacts which have not been predicted above a given Viterbi threshold. If bbcontacts predicts a contact that actually corresponds to a β-bridge as part of a parallel or antiparallel path, then the contact is counted as a true positive at the residue level and at the orientation-independent strand level.

Strand-level evaluation is only straightforward for DSSP-based results, because in this case a given predicted path will contain residues belonging to exactly one strand on each side. For PSIPRED-based results, strand-level evaluation is performed in the following manner. If a predicted path contains interactions between residues belonging to more than one pair of β-strands, then each pair of strands predicted to be in contact is counted in the strand-level evaluation. If a predicted path contains (on one or both sides) only residues that are not part of a β-strand, then this path is counted as a false positive in the strand-level evaluation. Finally, if a predicted path contains on both sides a mixture of residues contained in β-strands and other residues, the interactions between residues that are not part of a strand are ignored in the strand-level evaluation. Because PSIPRED-based strand-level evaluation is based on these additional criteria, it is provided only in an indicative manner and the residue-level evaluation forms the solid basis for comparison between different versions of our method.

S2 Supplementary results

S2.1 Verification that the results are not affected by any potential redundancy between training and test datasets

Because not all domains from all protein chains contained in the test datasets BetaSheet916 and BetaSheet1452 were annotated in CATH v3.5, we need to make sure that the bbcontacts performance is not over-estimated due to over-training. For this purpose, we evaluated bbcontacts (DSSP-based predictions and PSIPRED-based predictions without and with PSM) on the subsets of the test datasets which are fully annotated in CATH v3.5 (and thus non-redundant with the training dataset), i.e. all protein chains in each test dataset for which all domains are annotated in CATH v3.5.

These subsets contain 873 out of 916 chains for BetaSheet916 and 403 out of 1452 chains for BetaSheet1452. The difference in the proportion of annotated chains between the two test datasets comes from the fact that BetaSheet916 was published in 2005, while BetaSheet1452 was built from PDB structures deposited later than May 2004, so that many structures in BetaSheet1452 are too recent to have been annotated in CATH v3.5.

Because we built the training dataset by taking domains not belonging to any of the folds (CATH Topologies) observed in the annotated protein chains of the test datasets, we are sure that these subsets do not have any redundancy with the training dataset.

In Figure S1, we see that the bbcontacts performance on the subset of BetaSheet916 is almost identical to the performance on the full test dataset; the final recall for PSIPRED-based predictions is even slightly higher. The bbcontacts performance on the subset of BetaSheet1452 is slightly better than the performance on the full dataset for both DSSP-based and PSIPRED-based predictions.

Therefore, we can be confident that the method is not over-trained, because when we remove all chains that are potentially redundant with the training dataset from the evaluation, the performance of bbcontacts is maintained or even slightly increased.

S2.2 Results for the training dataset

It must be noted that the training dataset is rather different in composition from the test datasets: because it is built from all CATH v3.5 annotated domains not contained in the BetaSheet916 and BetaSheet1452 datasets, it contains many protein domains with few β-residues, low resolution or missing residues.

S2.2.1 Influence of the number of pseudocounts in the secondary-structure-based emissions

We looked at the influence of the number of pseudocounts from the non-conditional distribution added to the conditional probabilities for the definition of the secondary-structure-based emissions.

The results are displayed in Figure S7. A number of pseudocounts of 10,000 was chosen as it gives the best precision-recall compromise on the training dataset. A number of pseudocounts of 100,000 gives similar results, with slightly higher initial precision and slightly lower final recall.

S2.2.2 Choice of Viterbi score threshold for F1-score evaluation

The threshold for calculating F1-scores on the test datasets was chosen as the Viterbi score giving the maximum residue-level F1-score on the training dataset. The evolution of the

F1-score when including predictions with decreasing Viterbi score is shown in Figure S16 for the DSSP-based predictions and the PSIPRED-based predictions without and with PSM. A vertical line marks the chosen threshold: 1.7 for DSSP-based predictions and -1.6 for PSIPRED-based predictions.

S2.2.3 Contribution of the different terms in bbcontacts and final results

The precision-recall plots shown in the main text for BetaSheet916 (main Figures 3, 4a and 4b) are shown for the training dataset in Figures S21 and S22.

These figures show that the trends, choices and conclusions discussed in the main text for BetaSheet916 results also hold for results obtained on the training dataset. They also show that the performance of bbcontacts on the training dataset is not higher than on the test datasets, which is a sign that our method is not overfitted.

S2.3 Additional results for BetaSheet916

Comparison between Figure 3a in the main text and Figure S9(a) shows that the effect of changing the secondary-structure-based emissions is very different for DSSP and PSIPRED-based results. For the DSSP case, the probabilities for HMM (non-background) states are essentially unaffected by any of the changes, as they simply reflect the fact that β-contacts can only occur between two β-residues. On the other hand, background probabilities are strongly affected by the change from non-conditional to conditional. Adding pseudocounts from the non-conditional distribution to the conditional probabilities almost does not affect the background, as discussed above, and this explains why the blue and purple lines are superimposed in Figure S9(a). For PSIPRED-based results, adding pseudocounts from the non-conditional distribution to the conditional probabilities is a good compromise that improves the performance of bbcontacts.

Comparison between Figure 3b in the main text and Figure S9(b) also shows a difference in the impact of local background correction on DSSP-based results compared to PSIPRED-based results. The DSSP-based predictions are only impacted by darker regions if the corresponding couplings occur between two β-strands, while in the PSIPRED-based case, strong couplings cause the coupling-based emission probabilities to overtake the secondary-structure-based emissions, so that β-contacts can be predicted even in a region where the secondary structure composition is highly unfavorable.

For an easier comparison with the results from previous papers, Tables S2 and S3 provide recall, precision and F1-scores at the residue level and at the strand level, on the BetaSheet916 dataset. The "SS source" column specifies whether true or predicted secondary structure was used as an input. In these tables, the results for bbcontacts are given for a Viterbi score threshold corresponding to the threshold giving the maximum residue-level F1-score on the training dataset (see above, section S2.2.2). The results for all methods except bbcontacts are taken from Savojardo *et al.* (2013). In Table S3, the column "F1 $\geq$ 70%" shows the percentage of chains in the BetaSheet916 dataset that have an F1-score higher than 70% at the strand level (correct β-strand pairing).

S2.3.1 Comparison of bbcontacts with BCov* and CMM*

BCov (Savojardo *et al.*, 2013) uses PSICOV (Jones *et al.*, 2012) to generate direct coupling matrices, but it has been shown that pseudo-likelihood-based methods give better precision (Kamisetty *et al.*, 2013). CMM (Burkoff *et al.*, 2013) uses a different correlated mutation measure which has not been assessed in terms of general contact prediction performance.

Table S2: Residue-level performance on the BetaSheet916 dataset
(the largest value in each column is highlighted in bold)

Method	SS source	Recall (%)	Precision (%)	F1-score (%)
bbcontacts	PSIPRED	47.3	54.8	50.7
bbcontacts+PSM	PSIPRED	47.2	55.0	50.8
bbcontacts	DSSP	**60.9**	**69.4**	**64.8**
BCov6	DSSP	43.9	42.4	43.1
BCov	DSSP	42.4	40.9	41.6
CMM	DSSP	44.0	44.0	44.0
MLN-2S	DSSP	42.7	47.3	44.9
MLN	DSSP	39.3	46.1	42.4
BetaPro	DSSP	44.1	38.0	40.8

Table S3: Strand-level performance (correct β-strand pairing) on BetaSheet916
(the largest value in each column is highlighted in bold)

Method	SS source	Recall (%)	Precision (%)	F1-score (%)	F1 $\geq$ 70%
bbcontacts	PSIPRED	48.2	81.1	60.5	39.7
bbcontacts+PSM	PSIPRED	48.3	79.8	60.1	39.5
bbcontacts	DSSP	57.6	**83.7**	**68.3**	**55.9**
BCov	DSSP	**62.0**	59.5	60.7	44.2
CMM	DSSP	55.0	61.0	58.0	35.0
MLN-2S	DSSP	59.8	58.4	59.1	36.2
MLN	DSSP	55.5	59.8	57.6	33.7
BetaPro	DSSP	59.7	53.1	56.2	31.7

However, almost 80% of the original CMM alignments for BetaSheet916 have less than 1000 sequences, as opposed to 27% for the original BCov alignments and 34% for the alignments used in the present paper (compare Figure 1 in Burkoff *et al.* (2013) with Figure 3 in Savojardo *et al.* (2013) and Figure S2 in the present paper). Therefore, we can expect than by using better contact predictions as an input, the performance of BCov and CMM should improve.

A comparison of the DSSP-based residue-level performance of bbcontacts, BCov, CMM, BCov* and CMM* is shown in Figure S13. BCov* and CMM* correspond to using coupling matrices predicted with CCMpred as well as DSSP assignments as inputs to the β-topology prediction algorithms of BCov and CMM.

For BCov* and CMM*, we do not apply local background correction to the coupling matrices, because local background correction does not contribute a lot to DSSP-based results (see Figure S9). The default BCov parameters are used, in particular the minimum sequence separation of 6 for parallel strand pairing. For CMM*, the recommended parameters that were used in the original publication (Burkoff *et al.*, 2013) are used for sampling (50 resets, 1 million samples for each reset).

Figure S13 shows that BCov* displays intermediate results between BCov and bbcontacts. The precision-recall curve for bbcontacts displays a better robustness for high-confidence contacts than CMM*, as the bbcontacts precision remains above 80 % for recall up to almost 50 %. However, the CMM* precision-recall curve displays a higher initial precision than the bbcontacts curve. This can be explained by several factors. First,

CMM was developed to make use of strong topological constraints for β-strand interactions, but some of these constraints rely heavily on the availability of the exact β-strand positions and we thus decided not to include them in bbcontacts. This is the case for instance of the specific treatment of DSSP E (or B) assignments of length 1, which always correspond to β-bridges, and of the explicit modelling of the number of residues with no β-partners at the end of β-strands (Burkoff *et al.*, 2013). In addition, the CMM output contains a probability for each pair of β-residues to be in contact, while in bbcontacts, the final score is given to a path containing several β-contacts. Thus, contrary to CMM, bbcontacts cannot distinguish between central pairs of β-residues in a strand-strand contact and (less confident) pairs of β-residues close to strand extremities.

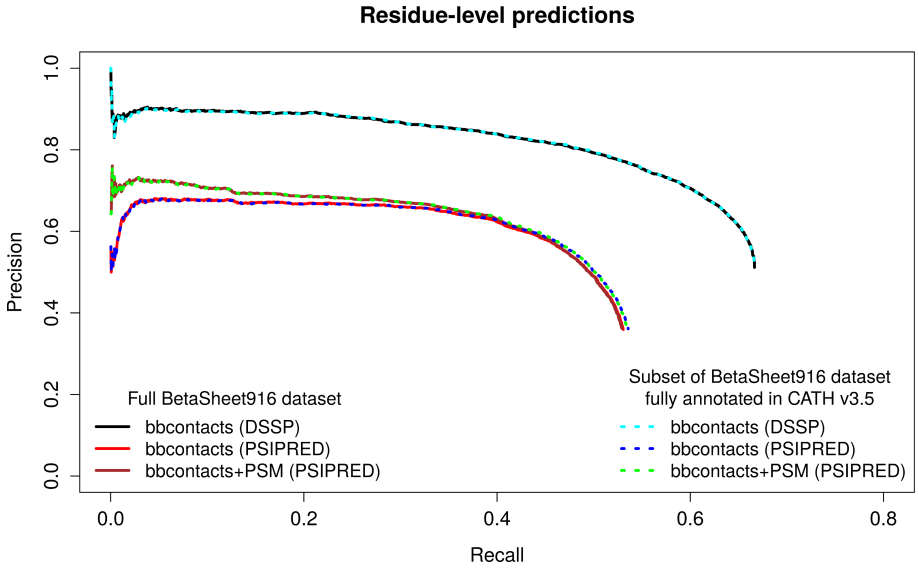
S2.4 Results for BetaSheet1452

The precision-recall plots shown in the main text for BetaSheet916 (main Figures 3, 4a and 4b) are shown for the BetaSheet1452 test dataset in Figures S23 and S24. In addition, the F1-scores for individual test cases depending on the number of MSA sequences are shown in Figure S25.

These figures show that the trends, choices and conclusions discussed in the main text for BetaSheet916 results also hold for results obtained on BetaSheet1452. In particular, comparison with the results from previous methods BCov and CMM (obtained from Savojardo *et al.* (2013)) also shows that bbcontacts performs much better than these previous methods when using the DSSP assignment, and the residue-level precision and recall reached by bbcontacts when using PSIPRED predictions are higher than the precision and recall of BCov and CMM when these methods use DSSP assignments.

S3 Supplementary figures

(a)



(b)

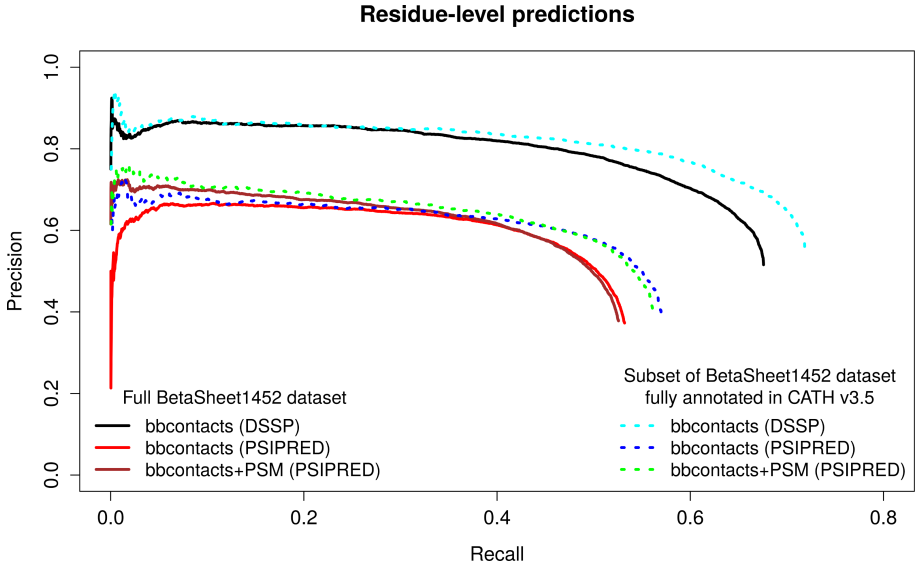
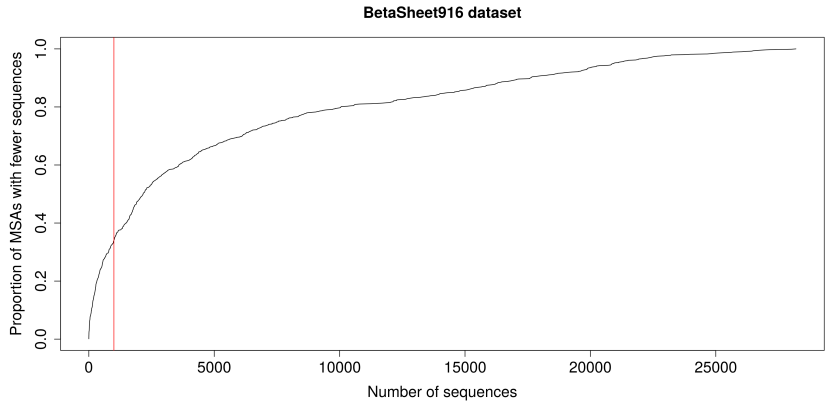
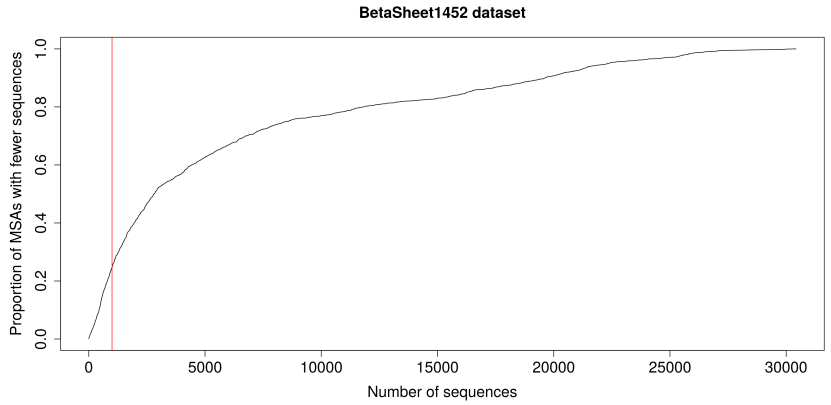


Figure S1: Comparison between residue-level performance evaluated on the full test dataset or on a subset of the test dataset for which all domains are annotated in CATH v3.5, thus making sure that there is no redundancy between this subset and our training dataset. (a) Test dataset BetaSheet916. The fully annotated subset contains 873 out of 916 chains (95%). (b) Test dataset BetaSheet1452. The fully annotated subset contains 403 out of 1452 chains (28%).
Further discussion of these results is provided in section S2.1.

(a)



(b)



(c)

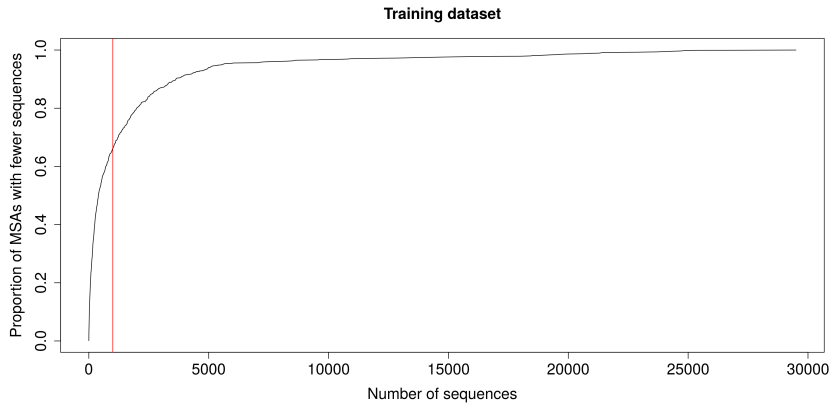


Figure S2: Cumulative distribution of the number of sequences in the MSAs (filtered at 90% sequence identity) for (a) test dataset BetaSheet916, (b) test dataset BetaSheet1452 and (c) training dataset. In each plot, the intersection of the curve with the red vertical line marks the proportion of alignments with less than 1000 sequences: 34% for BetaSheet916, 25% for BetaSheet1452 and 66% in the training dataset.

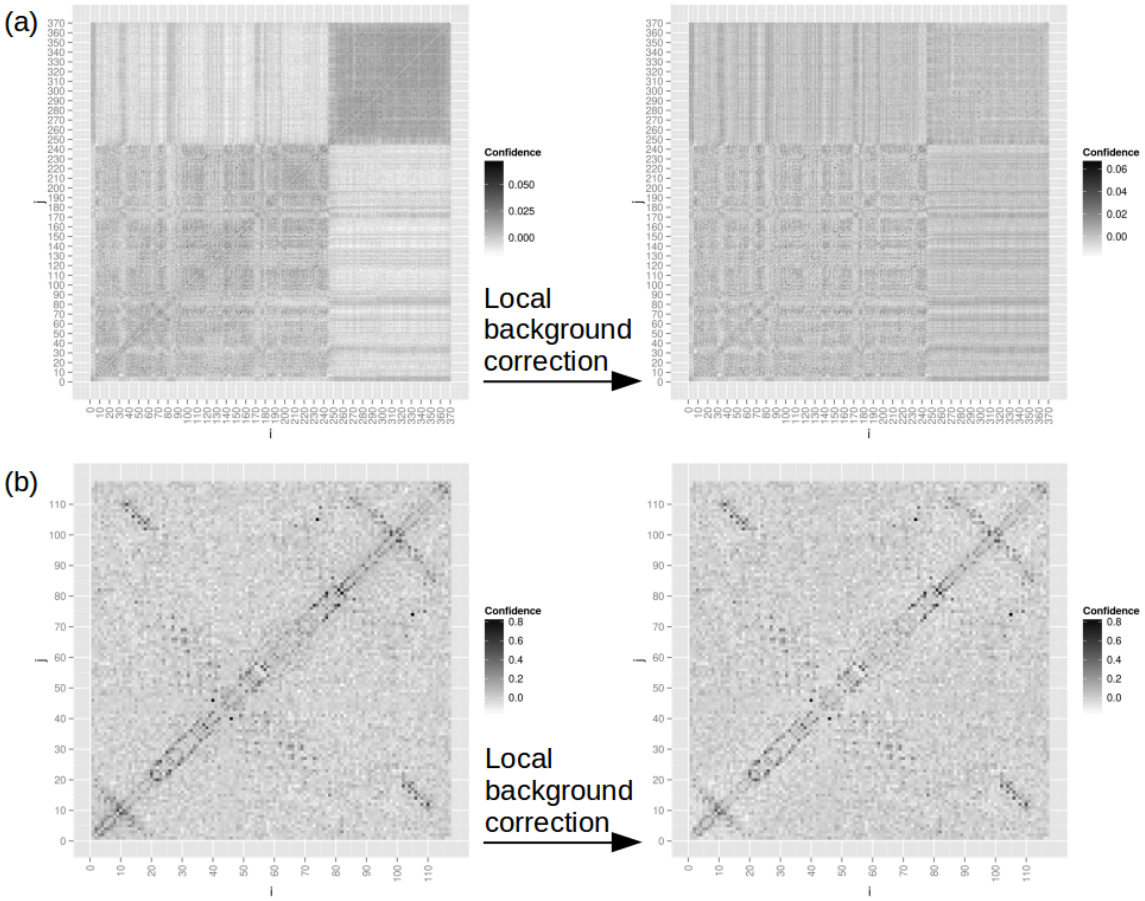


Figure S3: Illustration of the effect of the local background correction procedure with $S = 10$ on two cases, both belonging to the BetaSheet916 dataset.
(a) Protein chain 1gygB (370 residues): the initial coupling matrix shows a darker region in the top-right corner; the local background correction has a strong effect on the predicted couplings; in particular, the coupling values for the top-right corner get strongly reduced.
(b) Protein chain 1p9yA (117 residues): the local background correction has a very mild effect on the values of the predicted couplings and does not change the overall appearance of the contact map; the visible patterns are not affected by the local background correction.

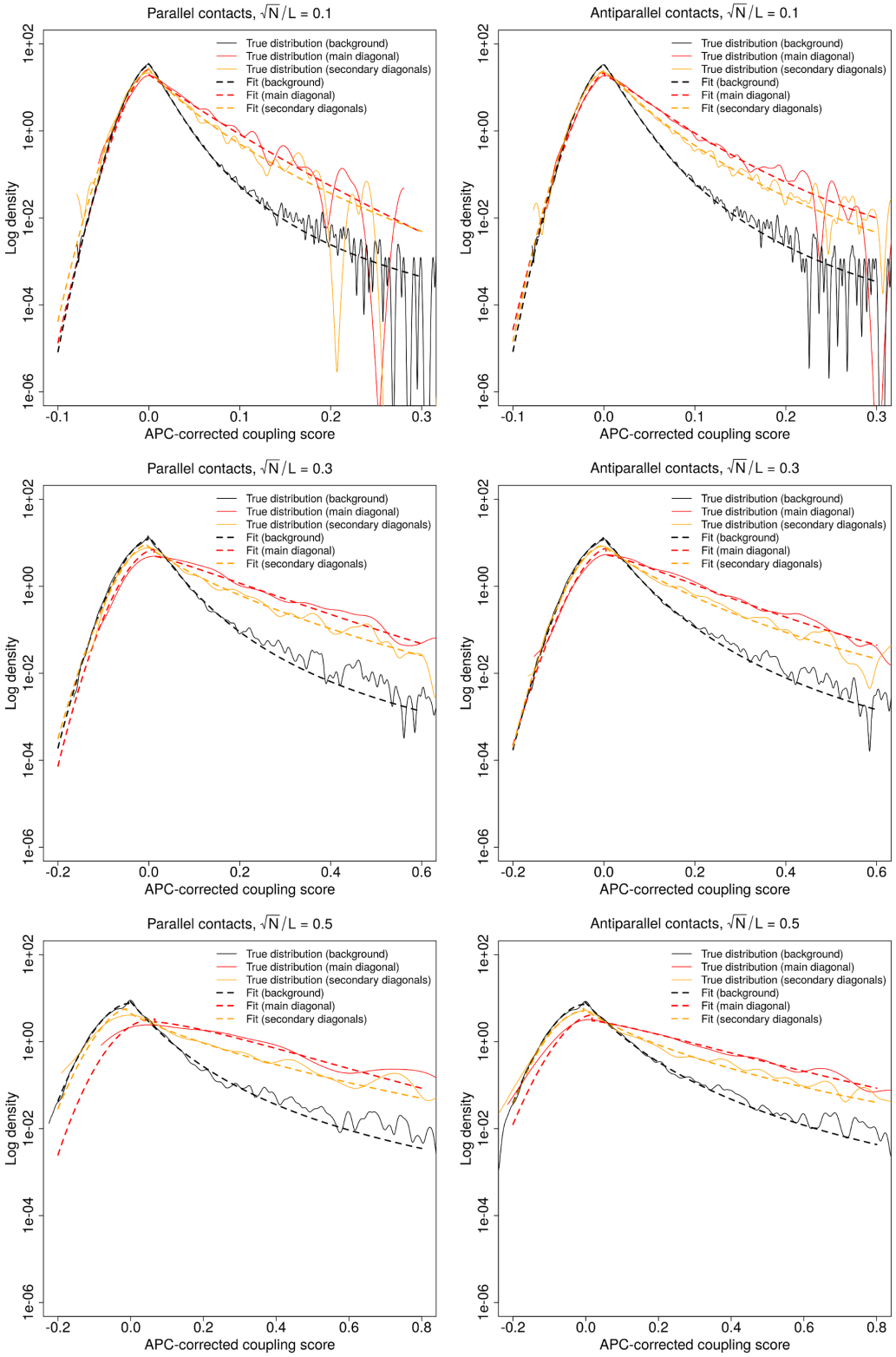


Figure S4: Coupling distribution densities and corresponding fits, for parallel (left) and antiparallel (right) β-contacts, for $\sqrt{N}/L \in \{0.1, 0.3, 0.5\}$

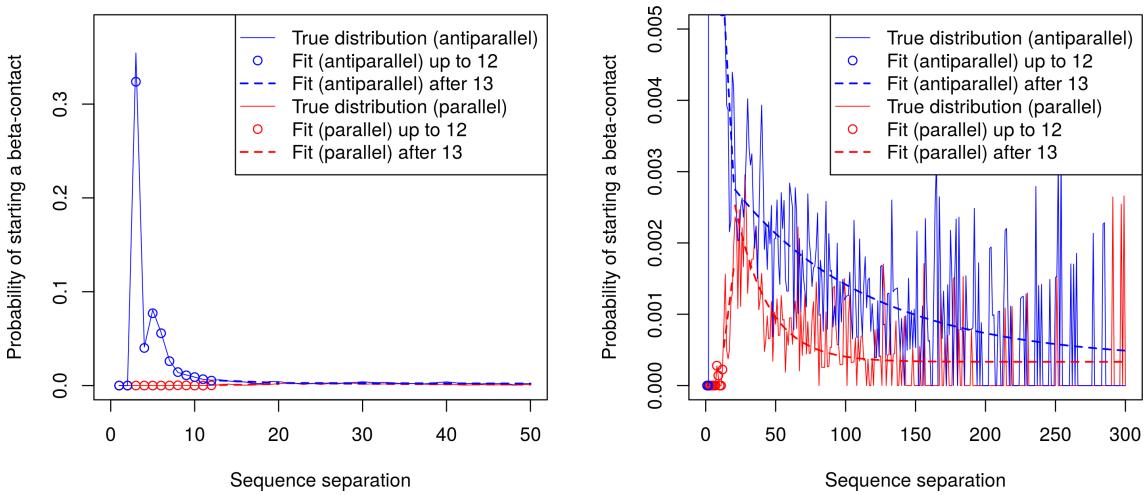


Figure S5: True distributions of the probability for starting an interaction between two β-strands, depending on sequence separation, and fits used for the prior depending on sequence separation. The two panels use two different scales, with the left panel focused on the large probabilities associated with antiparallel β-contacts at short sequence separation and the right panel focused on larger sequence separations. The fits contain three regions: up to a sequence separation of 12, there is an explicit probability for each sequence separation; between 13 and 20, the fit is a linear function of sequence separation; starting at a sequence separation of 21, the fit is an exponentially decreasing function of sequence separation.
This plot corresponds to the DSSP case (where the probabilities are calculated only for regions containing exclusively β-residues). In the PSIPRED case, the distributions have a similar shape but the probabilities are much lower because they are normalized over all possible pairs of residues within a protein.

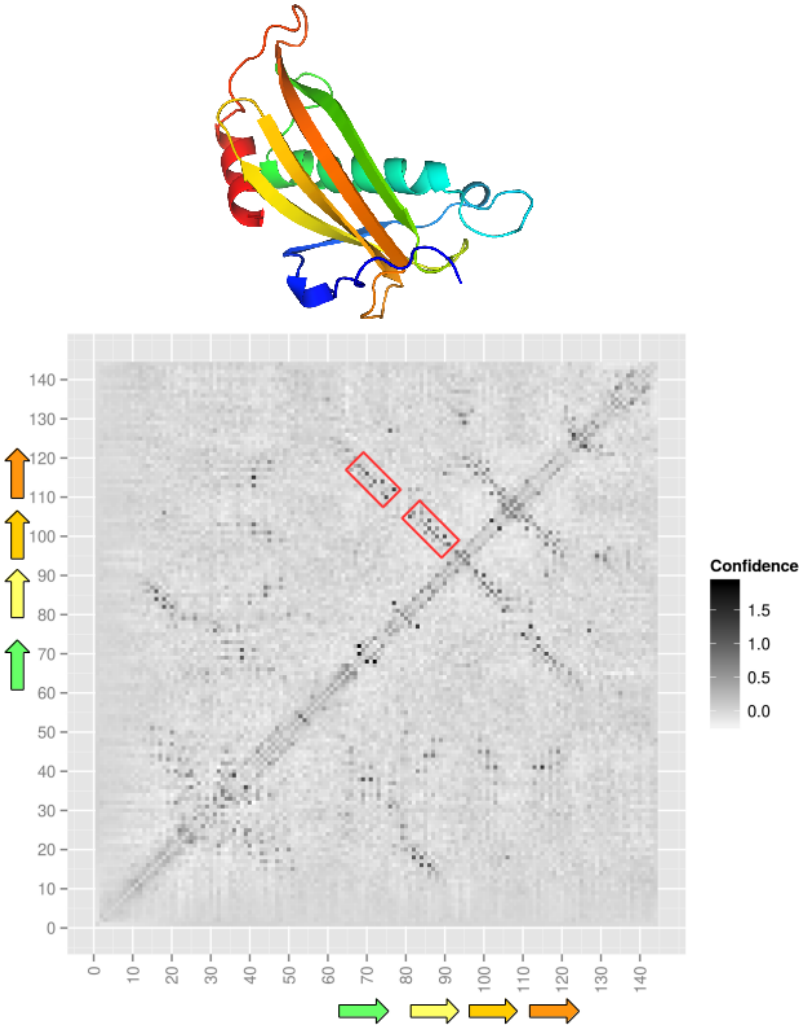


Figure S6: An example where prediction-shortening mode (PSM) is necessary. The two contacts between the yellow and light-orange β-strands and between the green and dark-orange β-strands form patterns (highlighted in red boxes) that are very close in the coupling matrix and almost aligned. Therefore, without PSM, only one path is detected, which leads to the prediction of several false positive residue-residue contacts between the green-to-yellow linker (around positions 75-80) and the light-orange-to-dark-orange linker (around positions 105-110). When PSM is triggered, it shortens the predictions until two separate paths are detected.

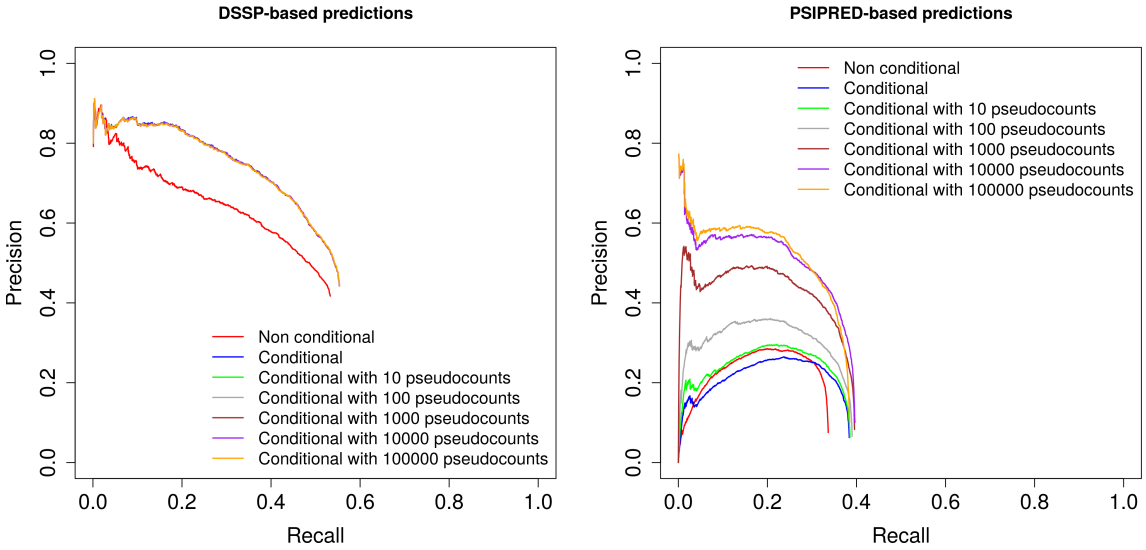
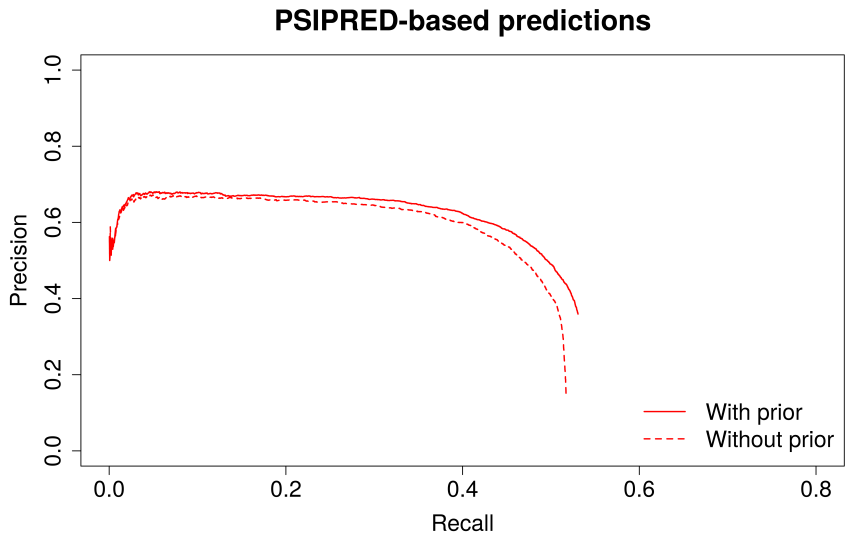


Figure S7: Influence of the number of pseudocounts added from the non-conditional distribution to the conditional distribution, evaluated on the training dataset, using no prior depending on sequence separation, no local background correction of the coupling matrices and with PSM turned off. (left) DSSP-based predictions: all conditional distributions without pseudocounts or with any number of pseudocounts are superimposed. (right) PSIPRED-based predictions.

(a)



(b)

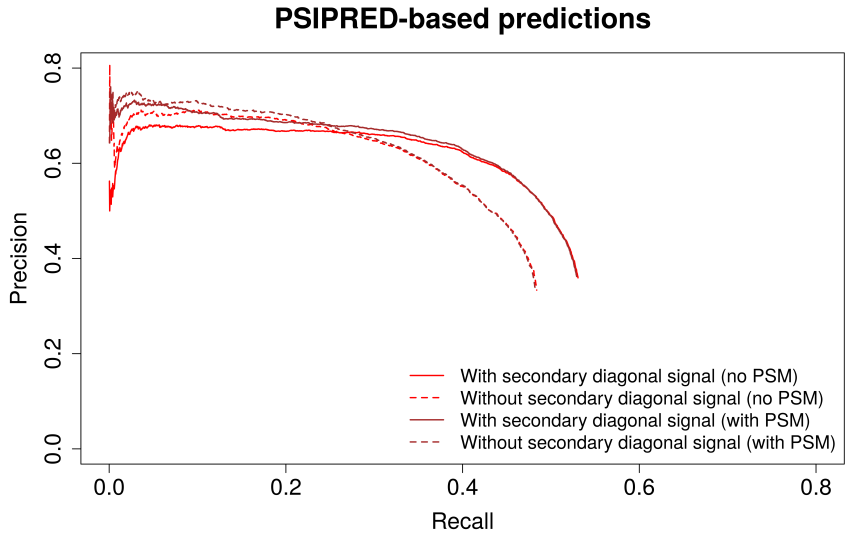


Figure S8: Influence of different model parameters on the residue-level performance of bbcontacts on the BetaSheet916 dataset, using PSIPRED predictions as an input and using the reference version of bbcontacts mentioned in the main text (local background correction of the coupling matrices with $S = 10$, conditional secondary-structure-based emissions with 10,000 pseudocounts).
(a) Influence of the prior depending on sequence separation.
(b) Influence of the signal coming from the secondary diagonals of the patterns on test dataset BetaSheet916: runs "without secondary diagonal signal" (dashed lines) contain only signal from the main diagonal of the pattern, while runs "with secondary diagonal signal" (solid lines) contain signal from both the main and the secondary diagonals. We also test the influence of PSM (predictions without PSM in red, with PSM in brown).

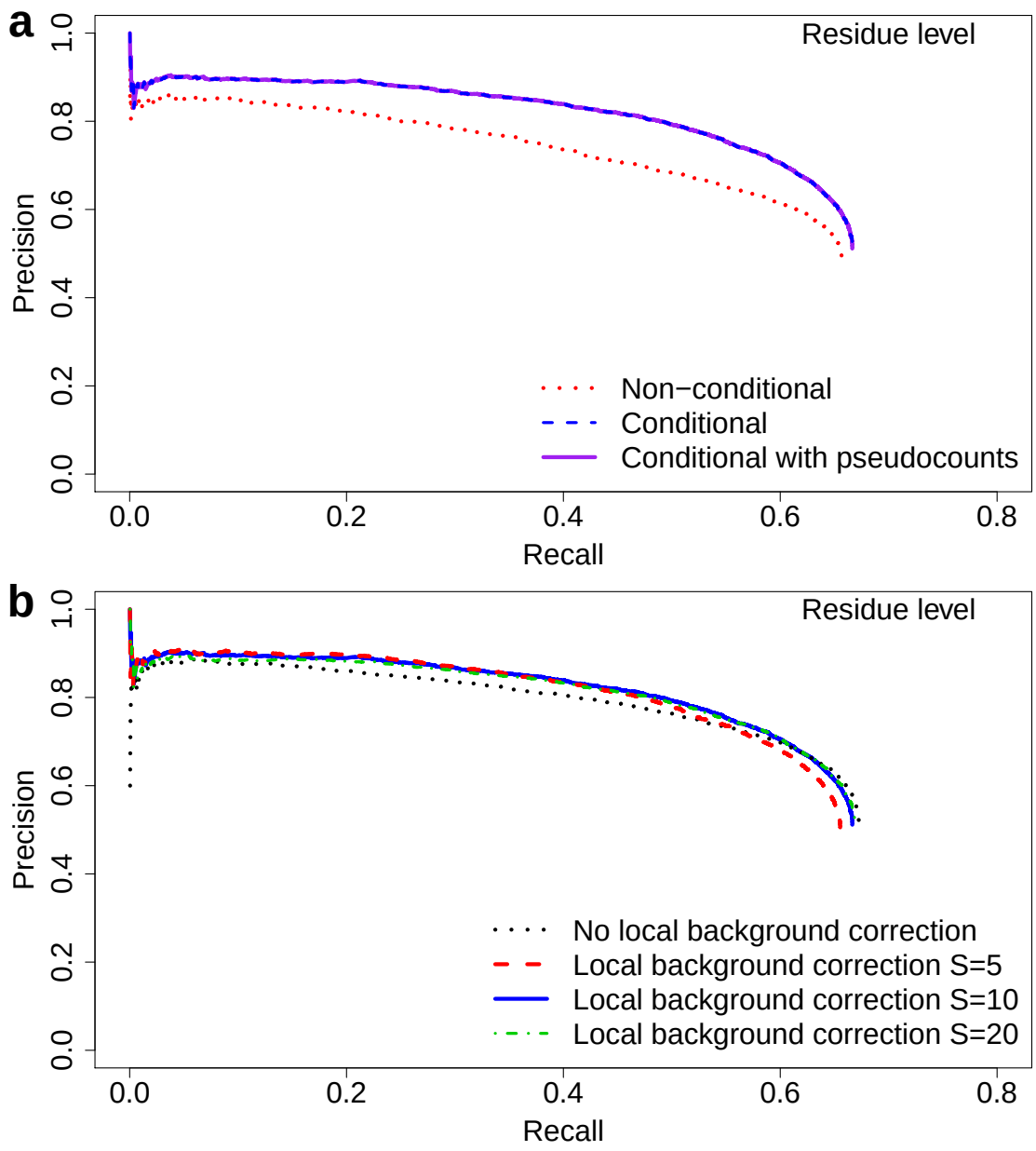
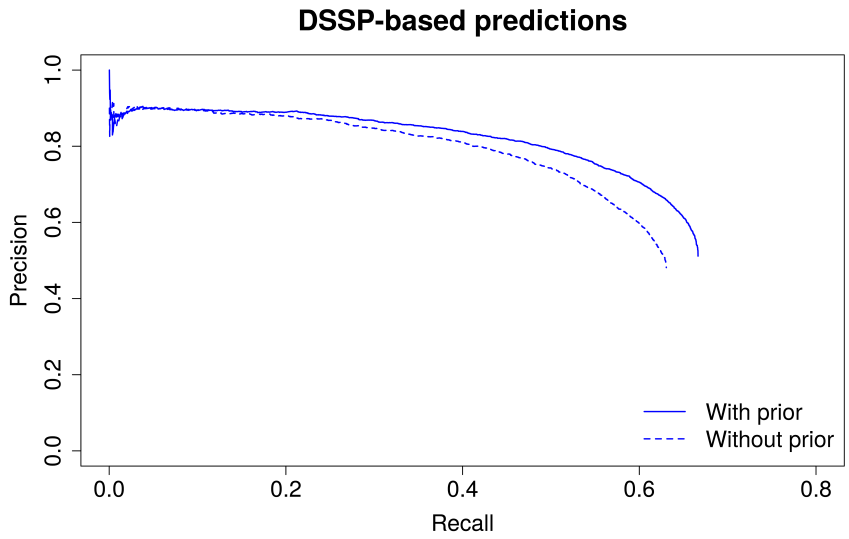


Figure S9: Influence of different model parameters on the residue-level performance of bbcontacts on BetaSheet916, using DSSP assignments as input for secondary structure. (a) Influence of the type of secondary-structure-based emission probabilities on the residue-level performance of bbcontacts: non-conditional (red), conditional (blue), conditional with 10,000 pseudocounts (purple). The blue and purple lines are superimposed. (b) Effect of local background correction applied to coupling matrices, for different values of S.

(a)



(b)

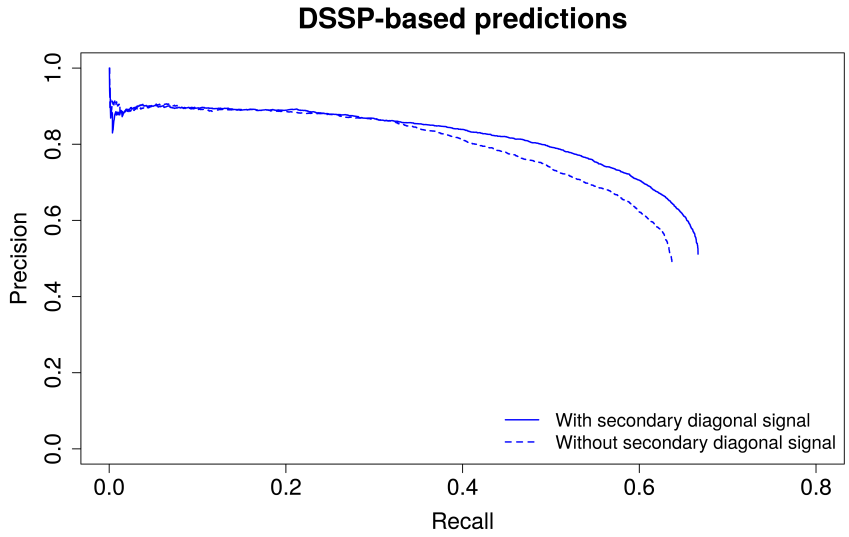


Figure S10: Influence of different model parameters on the residue-level performance of bbcontacts on the BetaSheet916 dataset, using DSSP assignments as an input and using the reference version of bbcontacts mentioned in the main text (local background correction of the coupling matrices with $S = 10$, conditional secondary-structure-based emissions with 10,000 pseudocounts).
(a) Influence of the prior depending on sequence separation.
(b) Influence of the signal coming from the secondary diagonals of the patterns on test dataset BetaSheet916: runs "without secondary diagonal signal" (dashed lines) contain only signal from the main diagonal of the pattern, while runs "with secondary diagonal signal" (solid lines) contain signal from both the main and the secondary diagonals. We also test the influence of PSM (predictions without PSM in red, with PSM in brown).

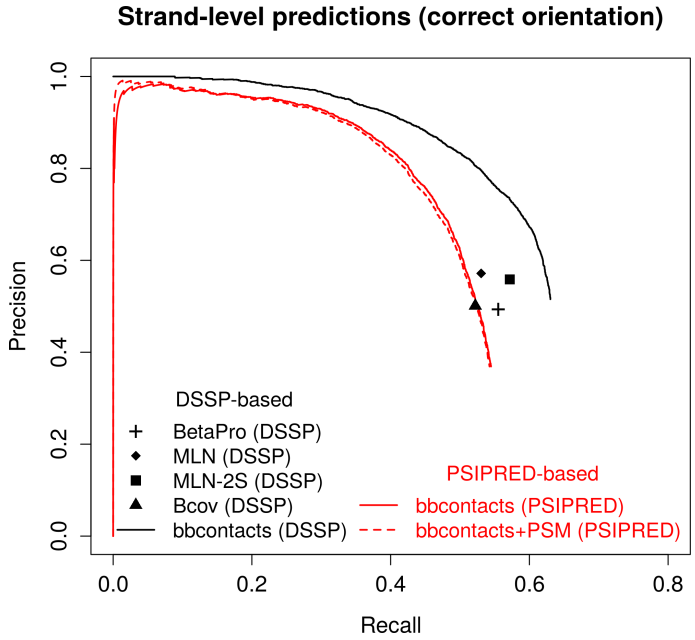


Figure S11: Final strand-level performance of bbcontacts on the BetaSheet916 dataset, compared to previous methods, when testing not only for correct pairing of β-strands but also for correct orientation. For all methods apart from bbcontacts, the results are obtained by multiplying the strand-level precision and recall by the percentage of correct directions provided in Savojardo *et al.* (2013). This result is not available for CMM.

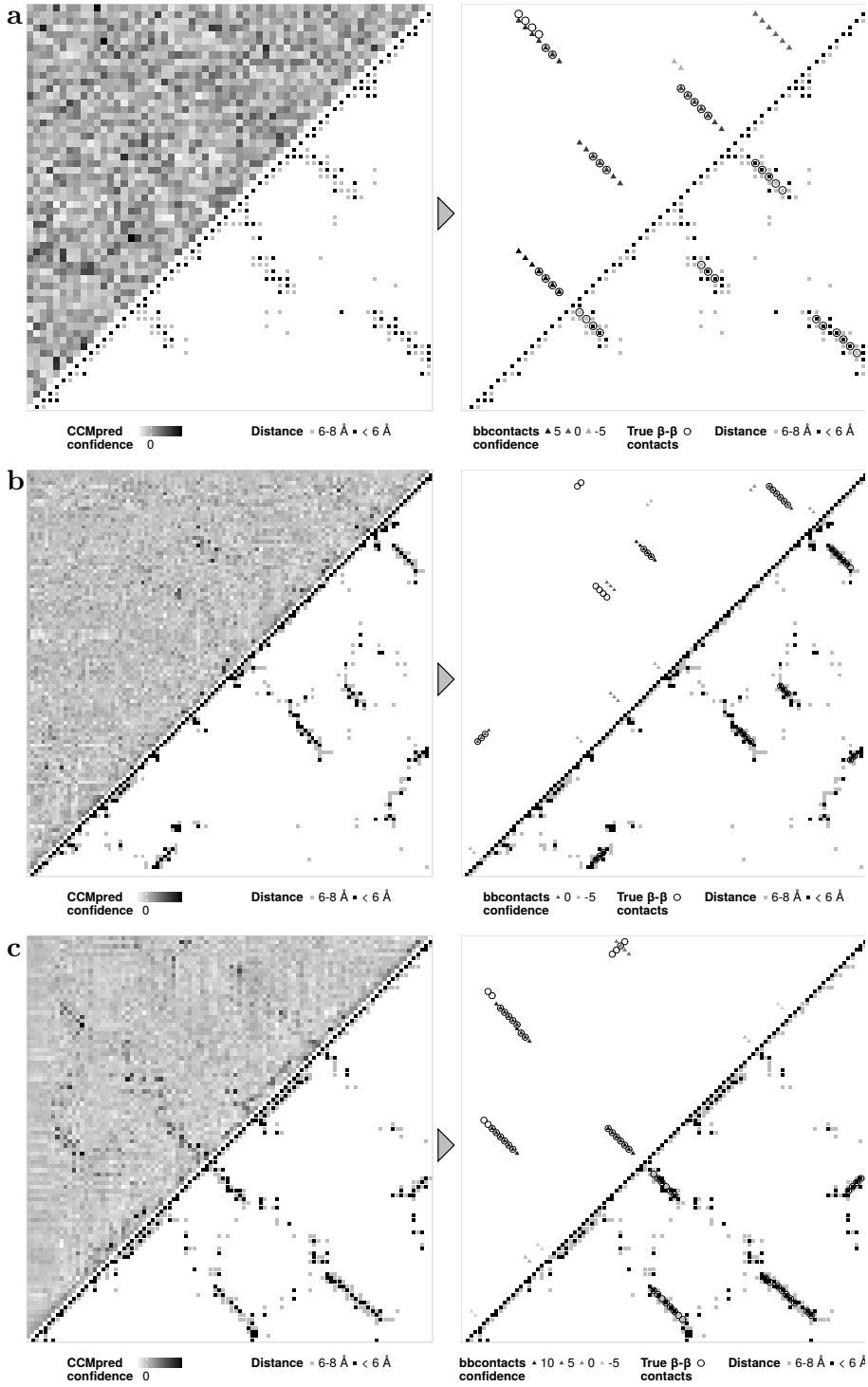


Figure S12: Examples of predicted contact maps for three PDB chains from the BetaSheet916 dataset. Each of the panels is built like main Figure 1, panels (a-b). On the left: CCMpred coupling matrix (upper-left) and coarse distance matrix (lower-right). On the right: β-β contacts predicted by bbcontacts using predicted secondary structure (upper-left) and coarse distance matrix (lower-right). The Viterbi score of the local alignment is the confidence value. The true β-β contacts (annotated by DSSP) are shown as open circles. (a) 1iguB (η=0.09). (b) 1jerA (η=0.29). (c) 2acyA (η=0.49).

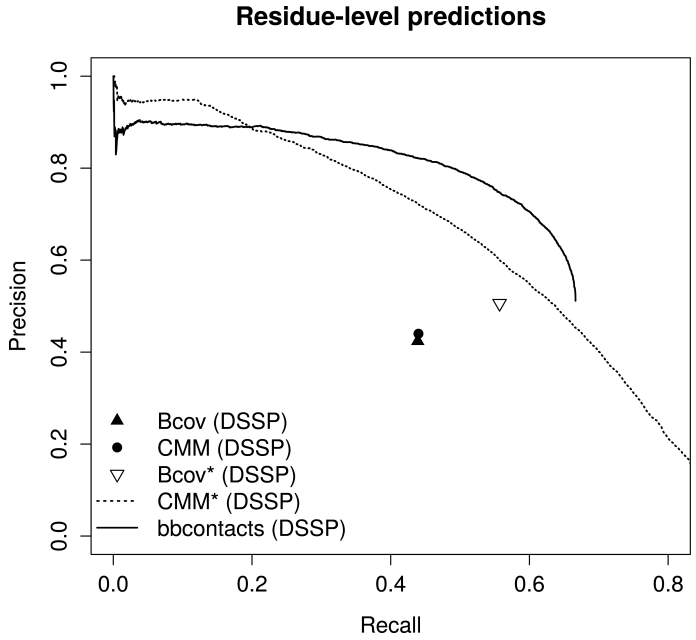


Figure S13: Residue-level performance of bbcontacts on the BetaSheet916 dataset, when using DSSP assignments, compared to the original BCov and CMM results and two new reference points BCov* and CMM*. BCov* and CMM* correspond to a situation where DSSP assignments and couplings predicted with CCMpred are used as an input to the β-contact prediction algorithms from BCov and CMM. Note that both algorithms require the DSSP-assigned secondary structure. Additional discussion of these results is provided in section S2.3.1.

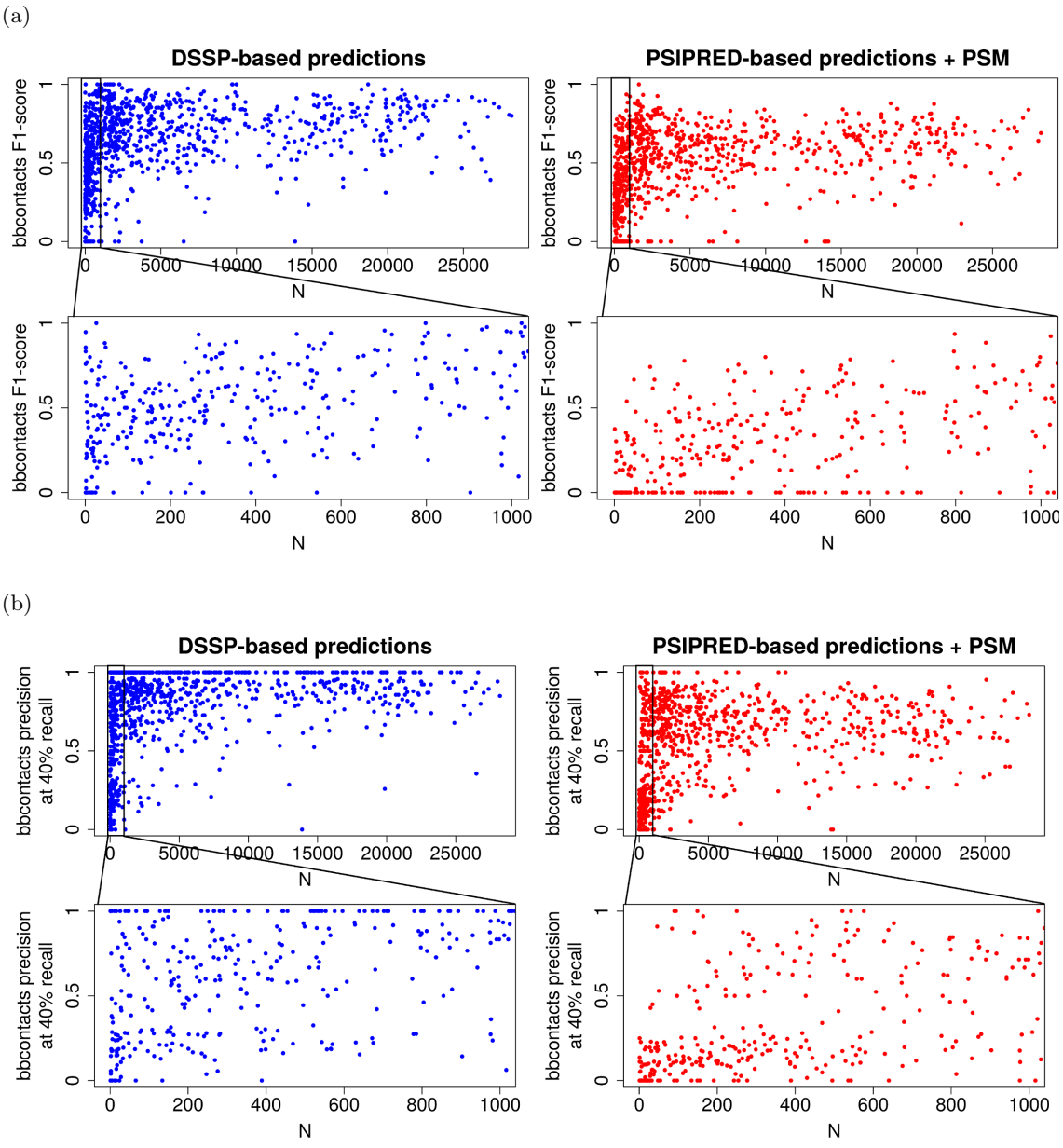
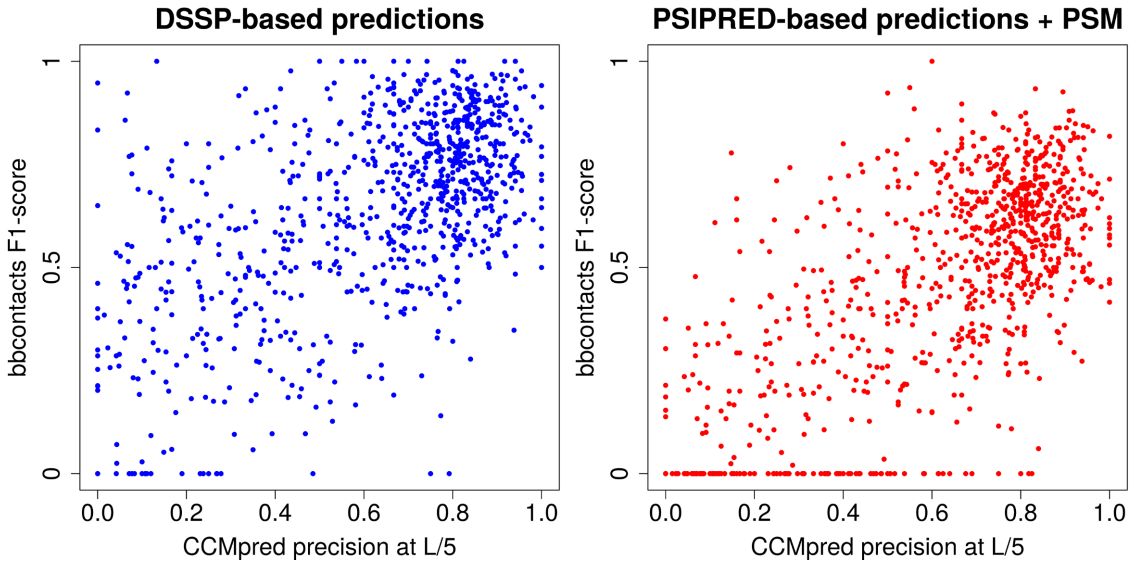


Figure S14: Residue-level performance for individual test cases in BetaSheet916 as a function of the number of sequences N in the alignment (filtered at 90% sequence identity). (a) Performance expressed as the bbcontacts F1-score (calculated on results above a threshold chosen as the Viterbi score giving the maximum residue-level F1-score on the training dataset). (b) Performance expressed as the bbcontacts precision at 40% recall (i.e. for each test case, all predictions up to 40% recall are taken into account when calculating precision). For each panel: (left) DSSP-based predictions, (right) PSIPRED-based predictions.

(a)



(b)

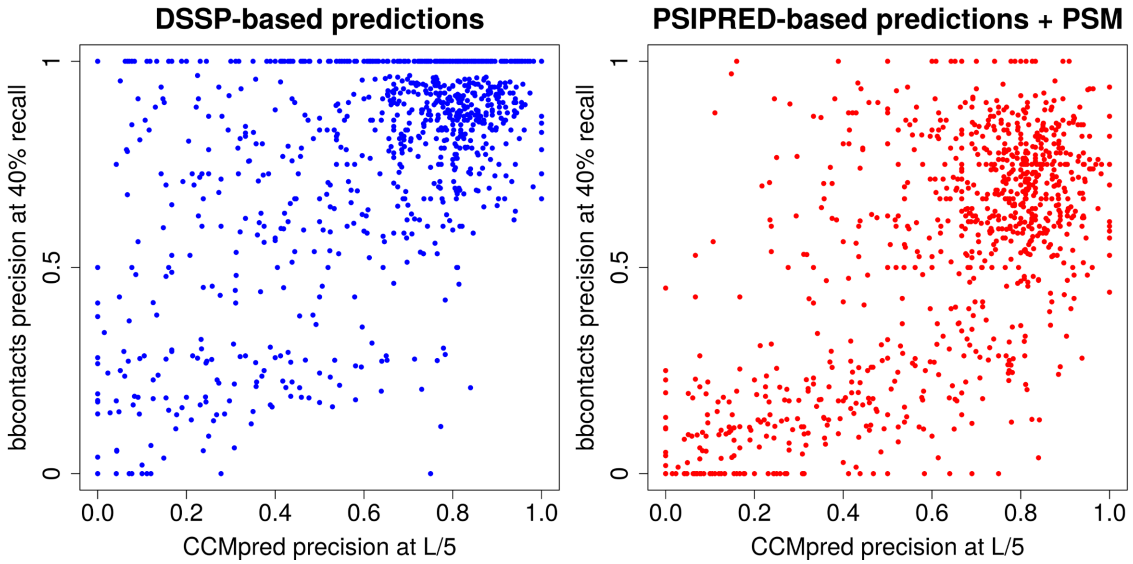


Figure S15: Residue-level performance for individual test cases in BetaSheet916 as a function of the CCMpred precision for $L/5$ predictions (L being the length of the protein). (a) Performance expressed as the bbcontacts F1-score (calculated on results above a threshold chosen as the Viterbi score giving the maximum residue-level F1-score on the training dataset). (b) Performance expressed as the bbcontacts precision at 40% recall (i.e. for each test case, all predictions up to 40% recall are taken into account when calculating precision). For each panel: (left) DSSP-based predictions, (right) PSIPRED-based predictions.

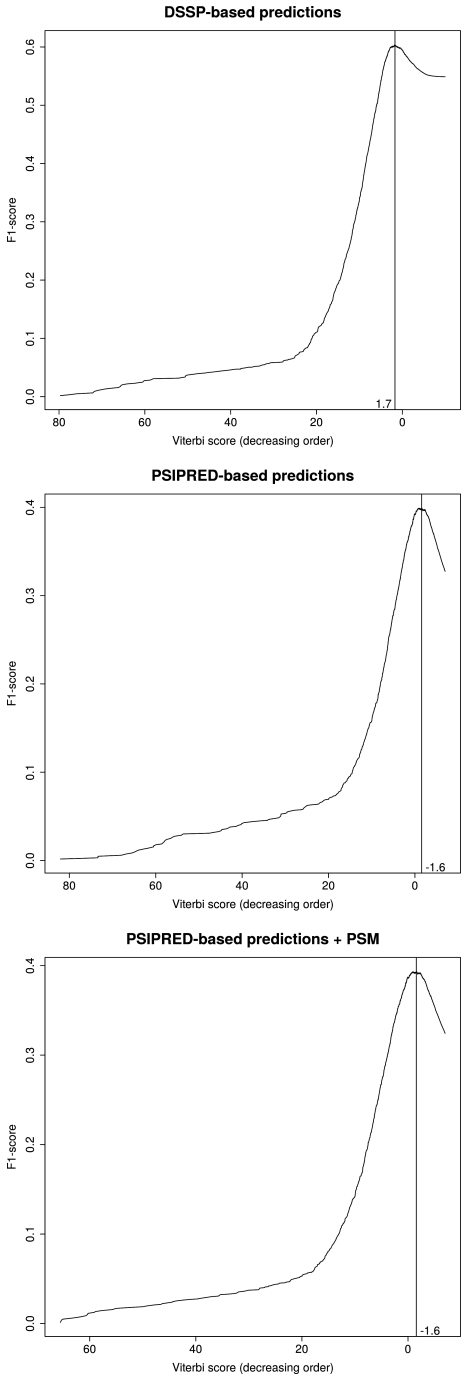


Figure S16: Evolution of the F1-score on the training dataset when predictions with decreasing Viterbi score are progressively added to the evaluation.
The Viterbi score giving the maximum F1-value is marked by a vertical line and was chosen as a threshold to calculate F1-values on the test datasets.

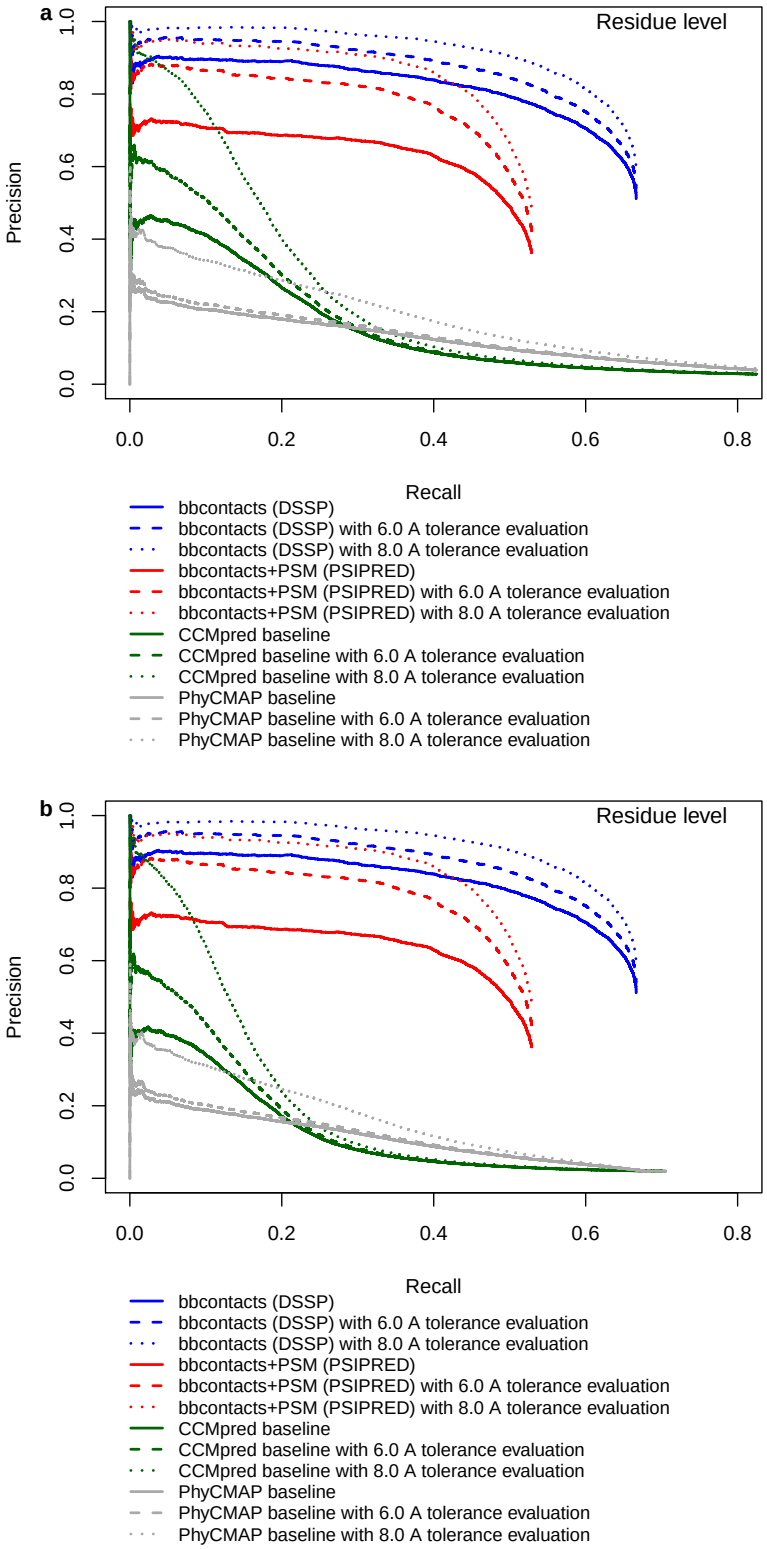


Figure S17: Residue-level performance of bbcontacts on the BetaSheet916 dataset, compared to CCMpred (Seemayer *et al.*, 2014) and PhyCMAP (Wang and Xu, 2013) baselines obtained by restricting the predictions to (a) DSSP-assigned β-strand regions and (b) β-strand regions predicted by PSIPRED. As in main Figure 4c, the false positive predictions with sequence separation smaller than 6 are excluded for CCMpred and PhyCMAP. Three types of evaluation are used: standard and evaluation with 6 Å tolerance (as in main Figure 4c) and evaluation with 8 Å tolerance (i.e. all false positives that have a Cβ distance lower than 8 Å are excluded from the set of false positives) (dotted lines).

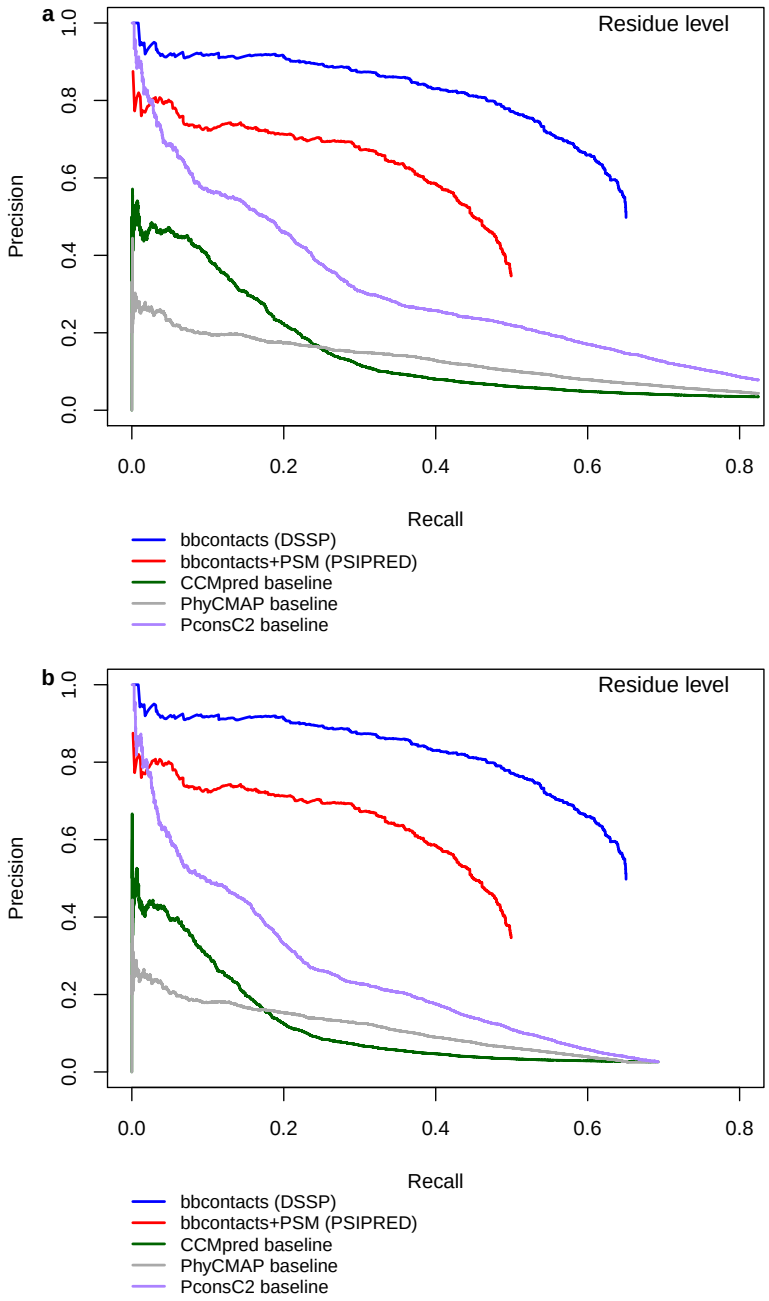


Figure S18: Residue-level performance of bbcontacts compared to CCMpred (Seemayer *et al.*, 2014), PhyCMAP (Wang and Xu, 2013) and PconsC2 (Skwark *et al.*, 2014) baselines obtained by restricting the predictions to (a) DSSP-assigned β-strand regions and (b) β-strand regions predicted by PSIPRED. This plot contains only results for the subset of the BetaSheet916 dataset for which PconsC2 predictions were obtained (Supplementary Dataset S2). As in main Figure 4c, the false positive predictions with sequence separation smaller than 6 are excluded for CCMpred, PhyCMAP and PconsC2. For clarity, in this plot only the standard evaluation is used.

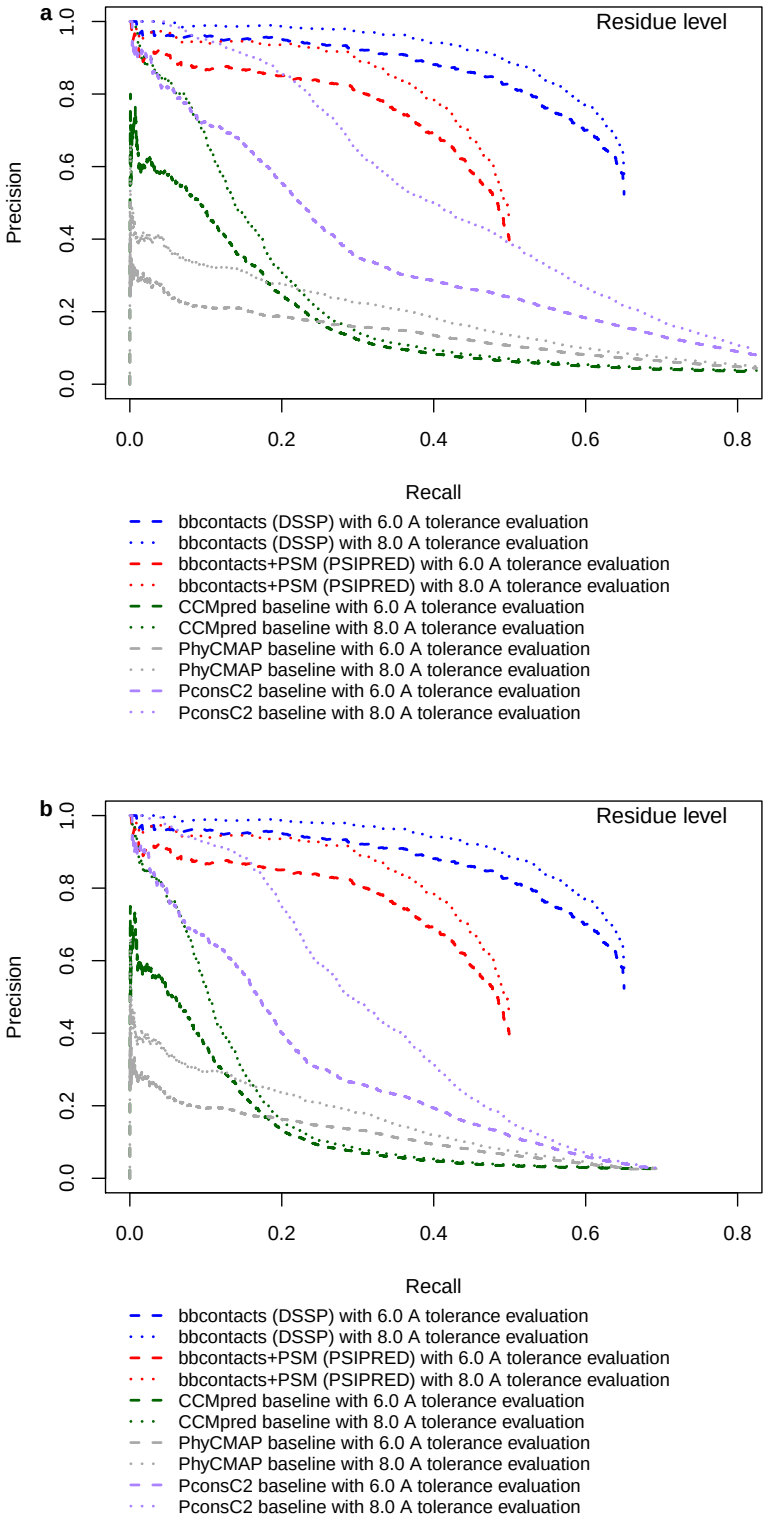


Figure S19: Residue-level performance of bbcontacts compared to CCMpred (Seemayer *et al.*, 2014), PhyCMAP (Wang and Xu, 2013) and PconsC2 (Skwark *et al.*, 2014) baselines obtained by restricting the predictions to (a) DSSP-assigned β-strand regions and (b) β-strand regions predicted by PSIPRED. This plot contains only results for the subset of the BetaSheet916 dataset for which PconsC2 predictions were obtained (Supplementary Dataset S2). As in main Figure 4c, the false positive predictions with sequence separation smaller than 6 are excluded for CCMpred, PhyCMAP and PconsC2. For clarity, in this plot only the 6 Å tolerance and 8 Å tolerance evaluations are used.

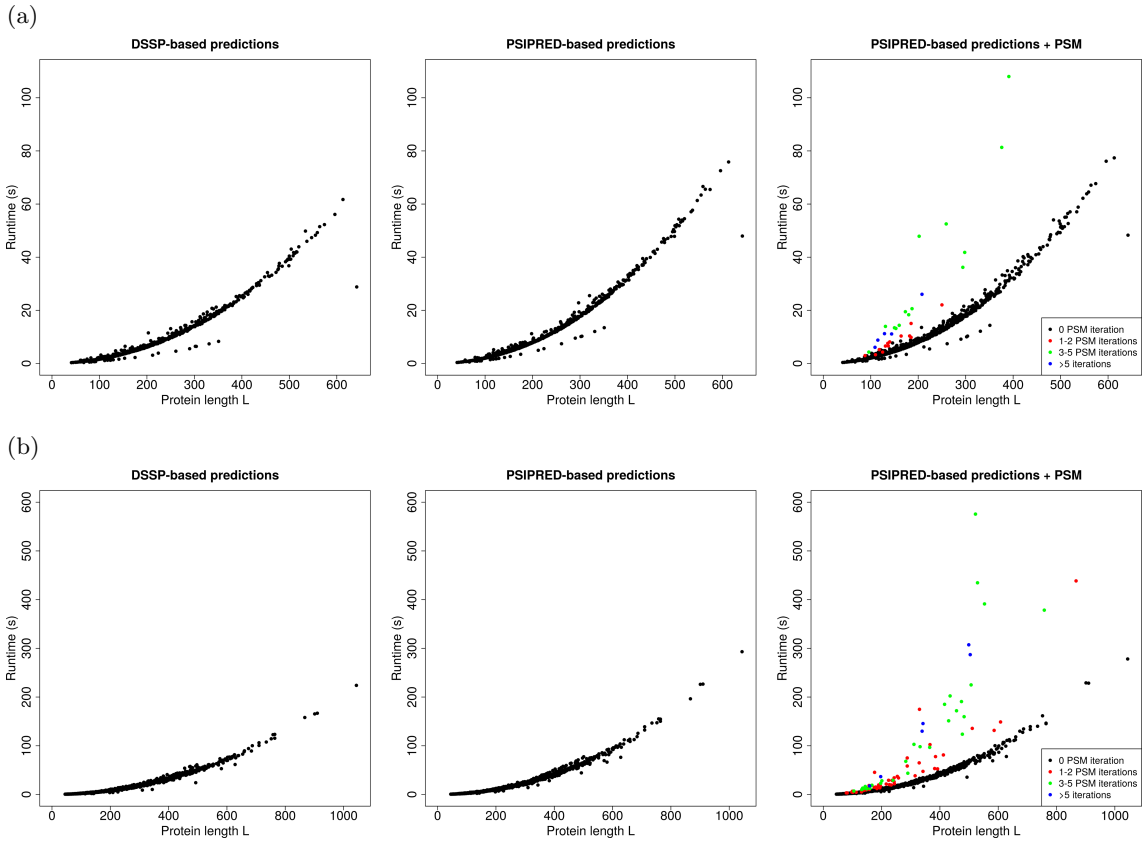


Figure S20: Runtimes depending on protein length for all cases in test datasets (a) BetaSheet916 and (b) BetaSheet1452. Because BetaSheet1452 contains much larger protein chains, the scales are different between (a) and (b).
When PSM is enabled, points are colored according to the number of PSM iterations effectively done while running bbcontacts.
The few points that have a runtime lower than the general trend in all plots correspond to cases where $\eta < 0.022$, in which case no coupling-based emissions are calculated or used for the predictions.

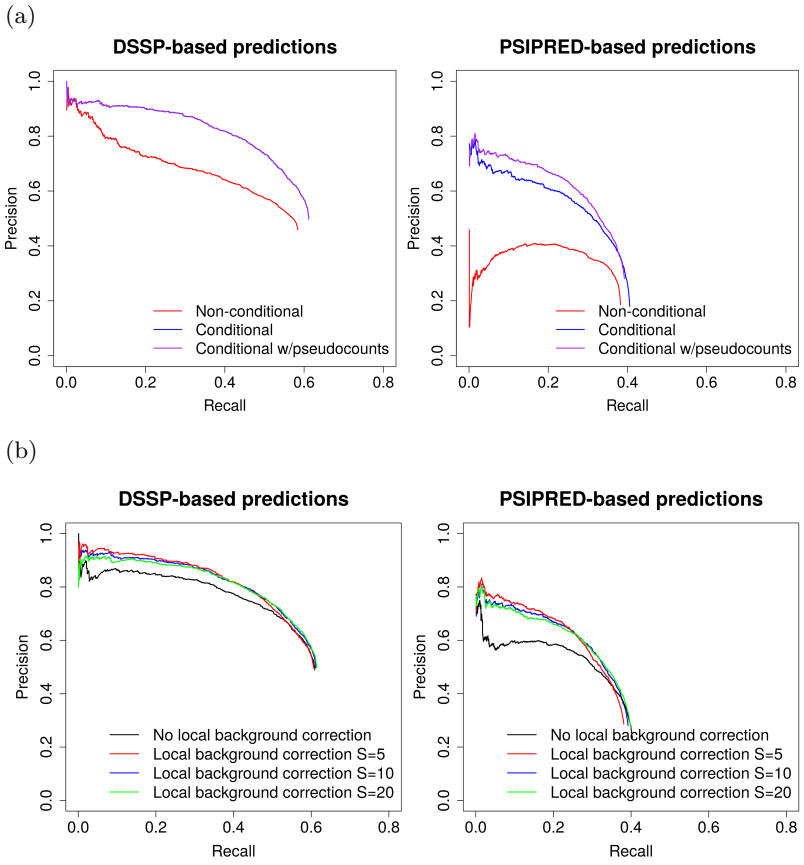
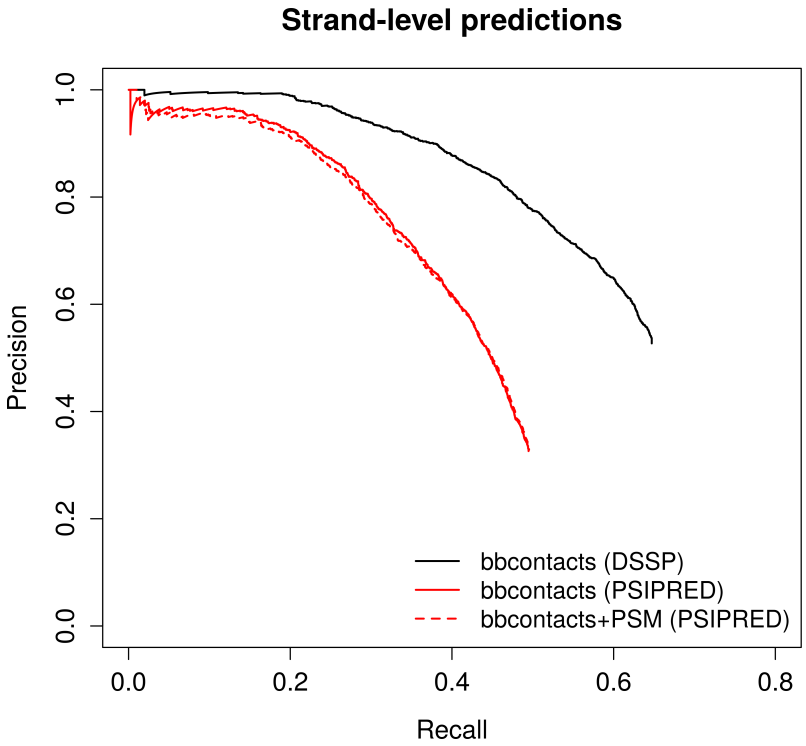


Figure S21: Influence of different model parameters on the residue-level results for the training dataset.
(a) Influence of the type of secondary-structure-based emission probabilities on the residue-level performance of bbcontacts: non-conditional (red), conditional (blue), conditional with 10,000 pseudocounts (purple). (left) DSSP-based predictions: the blue and purple lines are superimposed. (right) PSIPRED-based predictions.
(b) Effect of local background correction applied to coupling matrices, for different values of S. (left) DSSP-based predictions. (right) PSIPRED-based predictions.

(a)



(b)

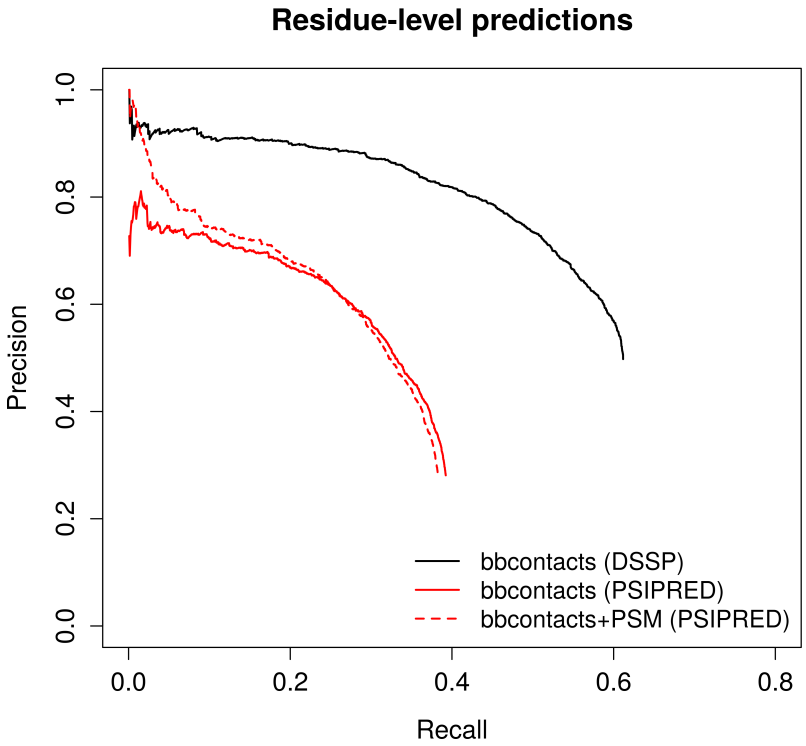


Figure S22: Performance of bbcontacts on the training dataset.
(a) Strand-level performance (correct pairing of β-strands).
(b) Residue-level performance.

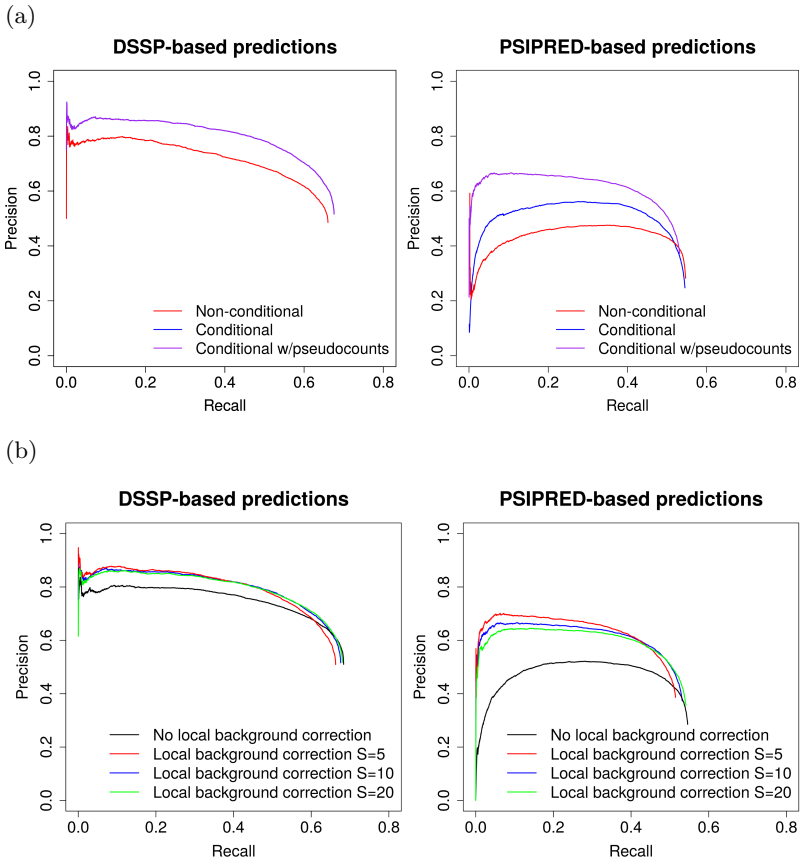
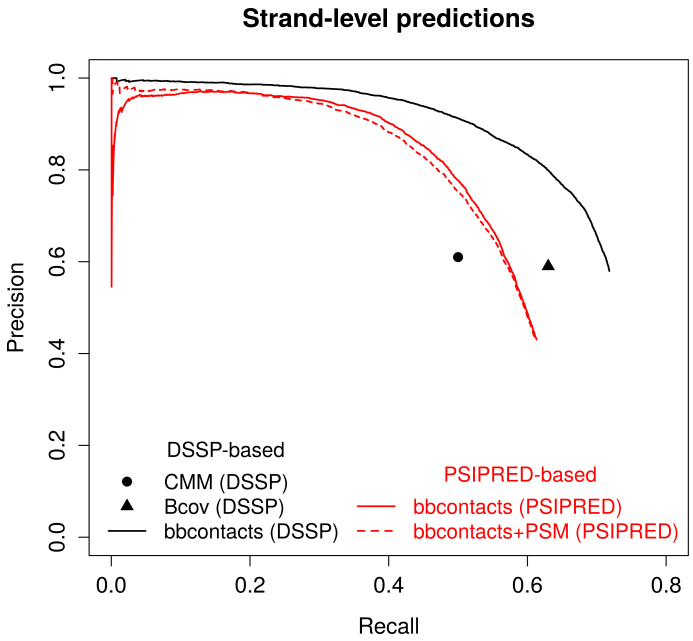


Figure S23: Influence of different model parameters on the residue-level results for test dataset BetaSheet1452.
(a) Influence of the type of secondary-structure-based emission probabilities on the residue-level performance of bbcontacts: non-conditional (red), conditional (blue), conditional with 10,000 pseudocounts (purple). (left) DSSP-based predictions: the blue and purple lines are superimposed. (right) PSIPRED-based predictions.
(b) Effect of local background correction applied to coupling matrices, for different values of S. (left) DSSP-based predictions. (right) PSIPRED-based predictions.

(a)



(b)

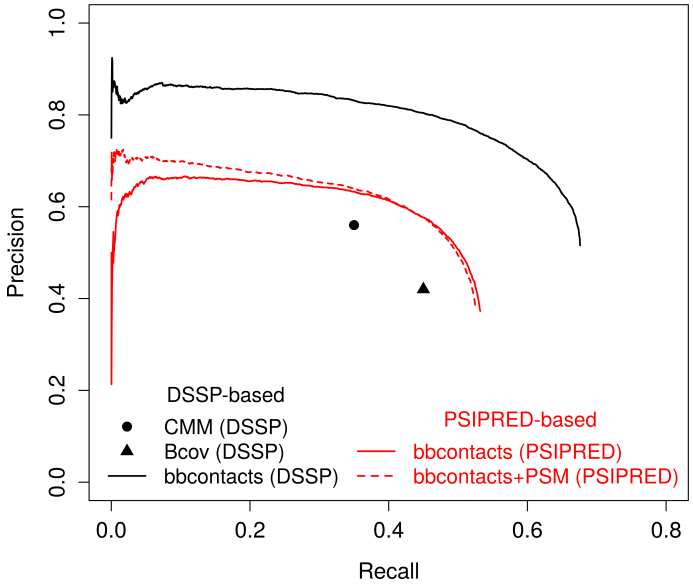


Figure S24: Performance of bbcontacts on dataset BetaSheet1452 and comparison with previous methods.
(a) Strand-level performance (correct pairing of β-strands).
(b) Residue-level performance.
Results for CMM and BCov on the test dataset BetaSheet1452 are taken from Savojardo *et al.* (2013).

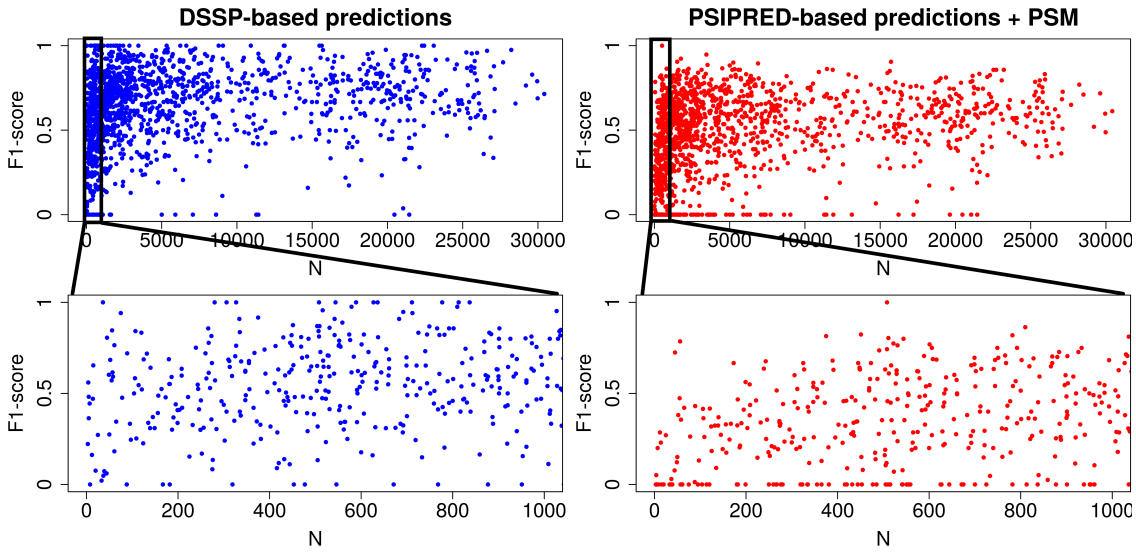


Figure S25: Residue-level performance for individual test cases in BetaSheet1452, expressed as the F1-score (calculated on results above a threshold chosen as the Viterbi score giving the maximum residue-level F1-score on the training dataset) as a function of the number of sequences N in the alignment (filtered at 90% sequence identity). (left) DSSP-based predictions, (right) PSIPRED-based predictions.

S4 Supplementary datasets

S4.1 Supplementary dataset S1: training dataset (943 domains)

The four columns are the CATH domain identifier, the length of the domain L, the resolution of the PDB structure and the number of sequences N in the HHblits alignment.

Domain	L	Resol	N	Domain	L	Resol	N
3nirA00	48	0.48	61	1jniA00	62	1.25	293
2b97A00	70	0.75	98	3moeA01	226	1.25	476
1mc2A00	122	0.85	797	3moeA03	272	1.25	534
1j0pA00	108	0.91	358	3fciA00	223	1.27	1619
1vbwA00	68	0.93	414	1qksA02	432	1.28	3381
3judA00	144	0.98	1315	1vlbA06	126	1.28	3245
1gkmA01	193	1.00	1373	1vlbA05	93	1.28	3323
1gkmA02	312	1.00	1192	1gk9A01	148	1.30	920
3nvsA02	205	1.02	4369	2prvB00	150	1.30	643
2v3iA00	433	1.05	1043	1eu1A02	245	1.30	5980
2hbwA02	146	1.05	4753	1gk9B02	73	1.30	895
1n62B03	166	1.09	3636	1rutX01	78	1.30	2635
1n62B02	145	1.09	2556	3eojA00	358	1.30	11
1n62B05	97	1.09	3299	3fegA02	262	1.30	5078
1n62B04	254	1.09	3845	1oqvA00	171	1.30	29
1g8tA00	241	1.10	1293	1gk9B03	161	1.30	467
2aibA00	98	1.10	143	2nr7A00	193	1.30	448
1t8kA00	77	1.10	8302	1eu1A03	87	1.30	683
3bvxA01	382	1.10	1242	3i33A04	84	1.30	4612
3a8gA00	195	1.11	262	1vp8A00	183	1.30	97
3ci3A01	177	1.11	1245	1o9iA02	63	1.33	142
1sauA01	44	1.12	526	2qikA02	154	1.35	1597
2r01A02	42	1.15	111	3pfgA02	59	1.35	60
2awkA00	224	1.15	166	1pinA01	32	1.35	1970
2ciwA00	298	1.15	383	1gppA00	217	1.35	260
1hbnA01	99	1.16	58	1ijyA00	122	1.35	576
1hbnB02	295	1.16	52	2fxuA03	92	1.35	1956
3essA00	199	1.19	5	3bxuA00	71	1.35	544
2bmoA01	308	1.20	1506	2gkpA00	161	1.35	65
3og2A03	89	1.20	166	3p0bA01	407	1.35	1143
3og2A02	184	1.20	187	1ouwA00	148	1.37	658
1w6sA00	595	1.20	3302	1f1eA00	147	1.37	400
1jetA02	124	1.20	10531	1s9uA00	198	1.38	935
1ymtA00	235	1.20	2037	2i3fA00	206	1.38	503
3qvpA03	302	1.20	3542	1v30A00	118	1.40	1104
3molB00	174	1.20	51	1tzpB00	236	1.40	272
1vk1A02	122	1.20	18	1pbjA00	116	1.40	18497
1vr7A00	119	1.20	668	1s2oA02	71	1.40	293
1vk1A01	101	1.20	2910	1l6rA02	64	1.40	120
1jetA03	216	1.20	10832	2fsqA00	224	1.40	82
3qvpA02	64	1.20	2381	1ie9A00	255	1.40	2012
1w6sB00	72	1.20	31	1f8nA03	89	1.40	312
2iayA00	110	1.20	97	1ygeA05	349	1.40	671
2xi8A00	66	1.21	21375	1yc5A02	84	1.40	2763
3moeA02	108	1.25	504	3iisM00	151	1.40	19
2wlvA00	144	1.25	352	2ra9A02	72	1.40	304

Domain	L	Resol	N
1f8nA02	100	1.40	269
2ra9A01	54	1.40	254
1tgrA00	52	1.42	251
1pp0B00	194	1.42	13
1m0kA00	222	1.43	946
1bgfA00	124	1.45	140
1ikpA02	158	1.45	3
2endA00	137	1.45	60
3h0nA00	182	1.45	780
3n2wD00	361	1.45	825
3f6yA01	128	1.45	84
1qz5A02	22	1.45	772
1g3pA01	87	1.46	6
1tkeA02	102	1.46	3144
1g3pA02	104	1.46	1
1tkeA03	58	1.46	2149
2fb6A00	111	1.46	113
3im9A01	239	1.46	7195
3mjfA04	96	1.47	2139
1rkuA02	95	1.47	102
1wpuA00	147	1.48	112
3gvjA04	155	1.48	8
1gqiA03	237	1.48	236
1vykA00	129	1.49	144
2qnlA00	159	1.50	113
1ui0A00	192	1.50	2604
1ocyA01	76	1.50	758
3c9qA00	191	1.50	113
3ckcA03	103	1.50	1
1j77A00	199	1.50	705
1qw2A00	97	1.50	92
1jl0A00	310	1.50	308
3mzfA02	92	1.50	1601
1inlC02	64	1.50	824
3bb0A01	174	1.50	140
3cqlA02	58	1.50	858
3bb0A02	401	1.50	1047
3npkA00	260	1.50	5147
3bhwA00	182	1.50	159
1g6sA01	205	1.50	4296
1ocyA02	122	1.50	15
1ofwA01	155	1.50	304
2y1qA00	137	1.50	4468
1h16A00	759	1.53	1172
1rv9A00	242	1.53	1741
3bhqA00	194	1.54	19638
2jkbA03	98	1.54	32
1sz7A00	159	1.55	513
2hhvA02	115	1.55	1982
4ubpB00	122	1.55	771
1iuqA02	267	1.55	358
1p9hA00	179	1.55	1296
1f0lA01	187	1.55	1
1nc7A00	110	1.55	53
2dxaA00	154	1.58	3855
1wn9A00	123	1.58	13
1k7iA01	234	1.59	3836
1aopA01	166	1.60	2880
1aopA03	143	1.60	3182
2olrA02	79	1.60	654
2o2kA01	243	1.60	1074
1ft5A00	211	1.60	839
1dd9A01	123	1.60	2085
1l5oA01	74	1.60	1121
1wr8A02	69	1.60	122
2pagA00	132	1.60	221
3ku3A01	103	1.60	268
1qgiA01	147	1.60	73
1x7dB01	164	1.60	1563
1ccwB02	66	1.60	67
1s9rA02	78	1.60	492
1p5dX03	119	1.60	4263
1p5dX01	145	1.60	4979
1p5dX02	78	1.60	4964
1kqfC00	216	1.60	2844
2ob5A00	145	1.60	707
1kqfA03	266	1.60	6018
1vkiA00	165	1.60	3789
1k92A02	215	1.60	1497
1hfeL03	143	1.60	1792
3c8wA01	227	1.60	473
2icuA00	207	1.60	1614
1aopA02	144	1.60	2216
3os4A00	387	1.60	1777
1dj7A00	109	1.60	217
1rylA00	157	1.60	186
3ku3A02	221	1.60	223
2hlyA00	205	1.60	18
1nlnA00	203	1.60	45
1vccA00	77	1.60	18
3o79B00	105	1.60	57
3gvoA00	342	1.60	1466
2olrA01	197	1.60	835
2olrA03	251	1.60	4763
2v2gD02	69	1.60	2000
3g9mA00	78	1.61	1600
2p12A01	158	1.63	217
1dw9A02	66	1.65	215
3bs3A00	58	1.65	21444
2d48A00	129	1.65	53
1o54A01	72	1.65	553
2yyvB00	224	1.65	467
1ogoX01	203	1.65	17
2fpqA00	414	1.65	25
1hx6A01	229	1.65	1
2g50A01	176	1.65	2419
1dtdB00	61	1.65	1
3gbyA00	126	1.66	18013
2isbA00	175	1.66	885
2aj7A00	155	1.67	397

Domain	L	Resol	N
1mw9X03	132	1.67	3616
1mw9X02	163	1.67	3230
1mw9X04	117	1.67	3454
2im9A01	116	1.67	208
2r7gA02	158	1.67	161
2im9A02	147	1.67	291
3kgdA01	240	1.68	597
2j8gA02	82	1.69	2526
1rxqD00	166	1.70	1965
1pbyC00	78	1.70	24
1qd1A01	180	1.70	303
2qgmA02	68	1.70	385
1vhhA00	157	1.70	91
1njhA00	105	1.70	73
1ewfA02	276	1.70	454
1vclA03	148	1.70	3
2i71A02	145	1.70	30
1pg6A00	206	1.70	1103
2quyA00	330	1.70	1058
1w27A03	122	1.70	202
3bi7A01	177	1.70	349
1tuoA03	117	1.70	4820
1tuoA02	85	1.70	4843
1ewfA01	180	1.70	346
2id3A02	141	1.70	5626
2qgmA01	200	1.70	430
1wjxA00	112	1.70	1499
1rh6A00	54	1.70	165
1mtyG02	73	1.70	12
2bw0A02	102	1.70	2660
3kqjA02	207	1.70	4326
1kidA00	193	1.70	2364
1hp1A02	187	1.70	2835
3l00A01	62	1.70	2180
2a9iA00	105	1.70	240
1k0rA01	99	1.70	1222
1rwjA00	81	1.70	291
2axqA03	100	1.70	291
2ox6D00	161	1.70	10
2a9dA01	246	1.70	2458
1cpqA00	129	1.72	529
2hy5B00	132	1.72	973
2pbkB00	227	1.73	51
1uehA00	214	1.73	2744
3nt1A02	511	1.73	1299
1px5A02	186	1.74	182
1pcfA00	66	1.74	279
1l1lA02	112	1.75	323
1y0kA00	178	1.75	10
1w99A03	180	1.75	156
1oi2A02	162	1.75	1124
1l1lA03	94	1.75	113
1pv5A00	254	1.75	388
1l1lA01	511	1.75	2106
2wbmA01	81	1.75	385

Domain	L	Resol	N
1u7lA02	183	1.75	221
3claA00	213	1.75	2450
1n93X02	127	1.76	11
2gmqA00	96	1.76	3
3or1C01	40	1.76	533
1n93X01	208	1.76	8
3g0mA00	138	1.76	968
2qhqB00	112	1.76	167
2c42A03	212	1.78	2968
1qkrB00	171	1.80	131
1f1mA00	162	1.80	198
2ou6A00	180	1.80	1553
1fn9A01	140	1.80	13
1lbuA02	129	1.80	873
1mugA00	165	1.80	1197
1lbuA01	84	1.80	4612
1rzhH01	105	1.80	53
2p84A01	60	1.80	49
2q03B00	129	1.80	123
1v54D00	144	1.80	258
1orvA01	470	1.80	4976
1v54A00	513	1.80	4858
1cmbA00	104	1.80	41
1kvdB00	77	1.80	2
1vqqA02	120	1.80	3958
1a9xA04	150	1.80	3064
1l8bB00	190	1.80	716
1jh6A00	181	1.80	1123
2sicI00	107	1.80	158
3canA00	158	1.80	11795
1vdkA01	135	1.80	1843
1mwpA00	96	1.80	71
1fn9A02	225	1.80	4
3c5nA00	231	1.80	399
1v54G00	83	1.80	262
1j09A03	116	1.80	3907
3fdjA03	127	1.80	1864
1vdkA02	266	1.80	5553
1jidA00	114	1.80	371
2q66A01	192	1.80	374
1ja1A03	126	1.80	1675
1lm5B00	193	1.80	245
1ro7C00	240	1.80	235
3bq9A01	110	1.80	124
1dl5A02	116	1.80	2
1v33A01	243	1.80	419
2z0tA00	109	1.80	168
2wjrA00	204	1.80	168
2qh9A00	172	1.80	168
1nxuA01	78	1.80	1000
1rzhH02	132	1.80	229
1yc9A02	83	1.80	3467
2it9A00	115	1.80	79
1t07A00	75	1.80	321
1ow1A00	167	1.80	100

Domain	L	Resol	N
2p1gB01	130	1.80	201
3mhjA00	208	1.80	1432
1k6kA00	142	1.80	4144
2r6zA01	54	1.80	22
1ualA02	81	1.80	1712
2y28B00	176	1.80	2494
1iq4A00	179	1.80	1624
1c96A02	113	1.81	3945
1c96A04	221	1.81	3657
1c96A03	175	1.81	3294
1c96A01	201	1.81	2718
2yvwA01	207	1.81	4477
2oqmB01	166	1.83	367
2w2rA00	177	1.83	13
1hq0A00	295	1.83	22
2pw8I00	60	1.84	10
1u7kA00	129	1.85	68
2auwB02	67	1.85	749
1b25A02	178	1.85	683
2auwA01	82	1.85	425
2qtqB00	192	1.85	23700
2ptrB02	275	1.85	5547
1b25A01	209	1.85	760
3do6A02	113	1.85	1427
1o6aA00	85	1.85	1527
1ssqA01	138	1.85	1743
1upkA01	305	1.85	241
1tolA02	73	1.85	792
2in3A02	123	1.85	3014
1b25A03	213	1.85	623
3cqbA01	85	1.86	4945
1lmlA03	63	1.86	267
2wjnC02	156	1.86	75
2wjnC01	138	1.86	82
1lmlA04	99	1.86	164
1jb7B00	216	1.86	6
1lmlA02	124	1.86	603
2gviA01	62	1.87	16
2ijqA00	145	1.88	455
3q0iA02	104	1.89	2701
7ahlF01	292	1.89	65
1t5oA01	141	1.90	1206
2qe9A01	158	1.90	2088
1svbA02	83	1.90	85
2osoA00	152	1.90	442
1h6wA01	41	1.90	17
2o5hA00	126	1.90	75
1oaoD05	130	1.90	105
1dzfA02	73	1.90	309
1eerA00	166	1.90	43
3l5xA00	101	1.90	27
1mpxA02	67	1.90	135
2dtrA02	66	1.90	1377
1lslA01	56	1.90	2693
1i1qA00	512	1.90	4465

Domain	L	Resol	N
1ae9A00	171	1.90	20746
1u94A02	59	1.90	1731
1musA03	110	1.90	178
2qx2A00	313	1.90	156
1ee8A01	120	1.90	2171
1ux6A01	127	1.90	669
1svbA03	80	1.90	110
1sr8A01	39	1.90	28
1oaoC03	176	1.90	123
1sr8A02	153	1.90	632
1qcsA02	102	1.90	208
1lshA04	251	1.90	231
3cx5C00	385	1.90	5065
1q16B03	77	1.90	333
2q5xA00	151	1.90	294
1rssA00	140	1.90	1598
1iv8A02	114	1.90	411
1nh2D02	48	1.90	149
1nh2C00	50	1.90	129
1vpsB00	289	1.90	73
1at0A00	142	1.90	473
1dzfA01	138	1.90	244
3fn2A00	94	1.90	7
1q16A01	27	1.90	147
2qcvA02	41	1.90	96
1epwA02	331	1.90	24
2ww2A02	166	1.90	1256
2g3wA00	173	1.90	312
1lshA01	263	1.90	506
1lshA03	98	1.90	174
3c9fA02	168	1.90	277
1amuA03	81	1.90	24720
1jdhA00	508	1.90	5373
1svbA01	136	1.90	143
1e4fT02	65	1.90	961
1musA02	276	1.90	316
1wteB01	147	1.90	2
1r4vA00	145	1.90	42
2ck3G02	117	1.90	1686
1wteA02	124	1.90	11
3d7aA01	136	1.90	241
1oisA02	128	1.90	204
2gukA00	107	1.91	82
3bf5A02	42	1.91	1
1lr0A00	123	1.91	1546
3ku8A00	134	1.93	2541
1kblA02	63	1.94	829
1vgjA00	181	1.94	2184
3dt5A00	118	1.94	1
3dnhA02	82	1.94	445
1b12A02	74	1.95	402
1m48B00	126	1.95	31
1p7tA01	85	1.95	435
1g8lA03	92	1.95	2687
1p7tB04	132	1.95	620

Domain	L	Resol	N
2o34A00	241	1.95	867
1qhdA02	227	1.95	25
1g8lA04	74	1.95	2466
1xlyA00	224	1.95	20
2wb0X02	168	1.95	43
1beaA00	116	1.95	160
1p7tB02	156	1.95	463
3c6kA02	56	1.95	828
3kjdA01	136	1.95	335
1g8lA02	57	1.95	2386
2pspA01	65	1.95	414
1ko7A01	129	1.95	1293
2gufA01	118	1.95	21282
1r1hA02	374	1.95	1760
3c8iA00	127	1.95	41
2ra8A01	74	1.95	502
1pucA00	101	1.95	212
1k7wA01	137	1.96	1687
1k7wA02	228	1.96	5606
2oezA01	75	1.97	252
3nx6A00	75	1.97	1872
2ny1A00	305	1.99	15373
2hq4A00	158	1.99	14
3eupB00	200	1.99	24271
1he1A00	135	2.00	18
3cexA00	165	2.00	218
2hqvA00	167	2.00	392
1ex0A02	319	2.00	2459
1kmoA02	519	2.00	24020
1io1A03	95	2.00	16
1oe4A00	245	2.00	485
1qlmA01	118	2.00	173
1io1A01	169	2.00	2906
1e7uA04	158	2.00	1070
1k8kE00	173	2.00	182
1o22A00	146	2.00	2
1lkiA00	172	2.00	53
1i7wD00	57	2.00	317
1xo0A02	200	2.00	19150
1vkyB01	206	2.00	1606
1r7lA00	103	2.00	13
1n7zA02	156	2.00	29
2e2dC01	85	2.00	127
1h3nA03	57	2.00	1
3cngA01	34	2.00	684
1eg3A01	38	2.00	71
2qgqB01	205	2.00	10939
2qv8A00	143	2.00	349
1e8yA05	187	2.00	1487
2hnuA00	81	2.00	106
1fm2B03	66	2.00	818
2hkuB00	182	2.00	22856
1nijA02	116	2.00	2454
1pzxA03	122	2.00	1903
1gl4A00	273	2.00	540

Domain	L	Resol	N
1pocA00	134	2.00	271
1mi8A00	141	2.00	1166
1wpbO02	113	2.00	107
1o0wA01	154	2.00	3174
1qoyA00	303	2.00	29
2ichA01	178	2.00	372
2ichB02	128	2.00	402
1tvfA02	69	2.00	55
2qf7A03	90	2.00	568
2fytA02	170	2.00	927
3reaC00	125	2.00	1047
2a1kA00	215	2.00	46
2ppqA02	212	2.00	7135
1wdjA00	186	2.00	4314
1qlmA02	198	2.00	155
1e8cA01	99	2.00	3980
1m3yA01	188	2.00	139
1m3yA02	215	2.00	121
1xkwA01	100	2.00	19500
1n7zA01	155	2.00	31
1kp8A02	94	2.00	2408
2atzA00	172	2.00	32
1n1bA01	200	2.00	824
1fm2B02	177	2.00	816
1pujA02	93	2.00	1469
1j5uA01	116	2.00	390
3kflA02	121	2.00	3513
1sx3A01	250	2.00	4971
1k8kA02	30	2.00	233
1d2zA00	102	2.00	148
1qakA01	79	2.00	1382
2cvcA03	114	2.00	404
3c2qA01	86	2.00	187
1io1A02	131	2.00	77
1dvoA00	152	2.00	256
1nigA00	146	2.00	3
2arzA02	88	2.00	589
2oyrA01	54	2.00	244
1olzA02	57	2.00	684
1ohtA00	167	2.00	2182
2i5tA00	165	2.01	1585
2h5nC00	124	2.01	841
1ddgA02	105	2.01	1415
2ii0A02	242	2.02	1362
2ii0A01	210	2.02	1037
2nrjA01	323	2.03	148
1em9B00	141	2.05	19
2qnuA00	207	2.05	209
3bjqJ00	292	2.05	258
1y7mA02	115	2.05	3875
1wlfA02	80	2.05	165
1uz5A04	72	2.05	2467
2g7zA02	118	2.05	1884
1i7dA03	141	2.05	3660
2gtqA05	324	2.05	690

Domain	L	Resol	N
1y7mA01	46	2.05	8629
3cjrB01	70	2.05	933
1udxA01	154	2.07	1760
2qs7A00	130	2.09	734
1ppjI00	42	2.10	68
3g4nA02	173	2.10	33
3fnaB00	114	2.10	13986
16vpA00	311	2.10	18
1g3jD00	34	2.10	25
2fiyA00	285	2.10	364
1wruA02	88	2.10	147
1ewnA00	200	2.10	833
3fy6D01	105	2.10	7
1x9mA02	169	2.10	864
2f2gA00	211	2.10	1226
2psbA00	287	2.10	257
3g4nA01	89	2.10	24
1a31A03	150	2.10	550
1h5wB03	45	2.10	6
3c2bA02	146	2.10	2339
3g27A01	65	2.10	84
1dq3A01	177	2.10	755
1mswD01	310	2.10	349
1u8bA01	69	2.10	1065
1qd6C00	240	2.10	350
2fi0A00	79	2.10	559
1okgA03	66	2.10	7
2qzbA00	145	2.10	86
1na6A01	171	2.10	35
3cddF03	60	2.10	149
1jcfA03	76	2.10	1346
1h3iA01	134	2.10	3672
1z1nX02	117	2.10	312
1wruA01	175	2.10	172
1accA01	228	2.10	1124
1fs0G01	130	2.10	2134
3g4nA03	187	2.10	61
2c36A00	274	2.11	29
2raaA00	174	2.12	2822
1n0uA03	107	2.12	834
1dvkA00	149	2.15	205
1rlzA00	344	2.15	556
3csvA02	235	2.15	2443
2bkkC02	173	2.15	5095
2iojA00	117	2.15	1428
2g8yB01	109	2.15	1037
3ar4A01	171	2.15	13464
1h54A03	74	2.15	1199
3ar4A04	244	2.15	7426
2r19A00	135	2.16	1846
2qyaA01	107	2.17	104
2igsA00	211	2.17	3
1k1fD00	63	2.20	16
1p35C00	295	2.20	11
1b4uB00	298	2.20	1636
1scfB00	118	2.20	33
1eteA00	134	2.20	19
1oxwC00	350	2.20	4917
1n7vA02	123	2.20	2
1n7vA03	232	2.20	2
1ba3A05	53	2.20	502
1qhlA00	203	2.20	14576
1vq8A03	78	2.20	500
1p2zA02	152	2.20	85
3hu3A02	93	2.20	360
2i06A01	214	2.20	116
1i5pA03	198	2.20	14
1pfoA01	183	2.20	96
1n7vA01	177	2.20	3
3mudA01	119	2.20	34
1p2zA03	265	2.20	79
2i06A02	77	2.20	71
2rhqB01	69	2.20	1708
2rhqB04	75	2.20	2265
2g03A00	172	2.20	2547
1u19A00	348	2.20	29498
1iq8A02	70	2.20	180
1nmpA01	119	2.20	1902
1p2zA04	208	2.20	40
1up8A00	597	2.20	841
1p2zA01	155	2.20	163
2qziA00	99	2.20	59
1kyqA03	82	2.20	92
1gmlA00	154	2.20	2123
2rhqB03	202	2.20	2365
3bqwA01	210	2.20	168
3n1hA00	161	2.20	60
3ci0J01	97	2.20	232
1zvpA00	128	2.20	361
1vq8P01	55	2.20	281
1pfoA02	53	2.20	43
1sczA00	233	2.20	3598
3bl4A01	69	2.20	14
3mw6F01	82	2.21	249
3cwcA02	230	2.23	1227
1r44A00	202	2.25	1038
1ciyA02	196	2.25	155
1qtqA04	109	2.25	3190
1qtqA03	79	2.25	1008
1p32C00	176	2.25	342
3dclA01	147	2.25	55
3dclA03	39	2.25	49
1xviA02	94	2.26	227
2j58A03	80	2.26	2061
2j58A01	103	2.26	2408
2j58A02	93	2.26	2575
2wvyA04	172	2.26	1197
2pw6A00	234	2.27	1392
3bzcA02	260	2.27	1393
2o3iA02	130	2.30	185

Domain	L	Resol	N
1twfF00	84	2.30	440
2pifA01	138	2.30	373
1gd8A00	105	2.30	1345
1e5rB02	84	2.30	14
1js8B02	105	2.30	56
3rk1B02	87	2.30	604
1cr5B02	95	2.30	199
2pifA02	106	2.30	364
2b5uA03	98	2.30	66
2qm4D01	133	2.30	87
2vutI00	42	2.30	952
1inpA01	47	2.30	56
2o3iA01	227	2.30	324
1bobA01	128	2.30	209
3k3fA00	332	2.30	371
1q3qA02	107	2.30	2044
1twfB04	177	2.30	738
1l1sA00	108	2.30	742
3cniA00	143	2.30	3344
1f3mA00	70	2.30	355
1u2mA00	84	2.30	989
1jsuC00	69	2.30	214
2fokB01	281	2.30	16
3pikA02	77	2.30	3261
1g31A00	107	2.30	106
3fggA00	138	2.30	15
2in5A00	190	2.30	158
2hr7A02	118	2.32	557
2auaA02	89	2.35	16
2auaA01	106	2.35	19
3cdlB02	130	2.36	2572
3jyuB01	125	2.37	504
2gjvA00	136	2.39	43
1l5jA03	173	2.40	2662
1af6A00	421	2.40	311
1fepA02	531	2.40	24813
1ax8A00	130	2.40	54
2gmfA00	121	2.40	28
1l5jA02	200	2.40	4001
1epuA02	91	2.40	552
1t77A02	289	2.40	674
1x87B01	267	2.40	509
1l5jA04	128	2.40	642
3b8oA01	213	2.40	222
1kfqA01	207	2.40	5107
2a6hC02	339	2.40	1899
2ph7A01	119	2.40	3
1yisA02	310	2.40	5706
2a6hC01	211	2.40	2103
2fywA02	122	2.40	1822
2fywC01	126	2.40	1719
2pusA01	334	2.40	13
1r8eA02	73	2.40	9656
1ozjA00	126	2.40	203
2a6hC03	180	2.40	1254
2fdoA00	89	2.40	3
1chkA01	143	2.40	85
2vpzA04	61	2.40	486
3kasA03	142	2.40	607
1n8yC02	119	2.40	426
2ahxB04	138	2.40	1347
2a6hC04	65	2.40	1922
2ftsA04	81	2.41	2454
3g74C00	83	2.43	105
1ya5T01	84	2.44	19
1ei7A00	158	2.45	37
2h21C01	254	2.45	1791
3c6mD01	35	2.45	25
1hk8A00	561	2.45	961
2h21A02	165	2.45	374
2ajrA02	52	2.46	765
1lpbA00	85	2.46	69
3l4jA05	153	2.48	396
3l4jA03	184	2.48	1984
2fpnA01	135	2.49	35
1j3eA00	115	2.50	106
2hdiB00	103	2.50	3
3clqA04	160	2.50	126
1cjyB02	485	2.50	815
1khvA01	59	2.50	3
1ckmA03	54	2.50	5
1a0pA02	180	2.50	20621
3bt3A01	72	2.50	4
2qsdB02	78	2.50	134
1eg7A03	90	2.50	1630
1jb0D00	138	2.50	96
1ztpA01	210	2.50	113
2re3B03	29	2.50	243
3clqA02	150	2.50	134
3c2iA01	62	2.50	261
1jeyB03	102	2.50	181
2vqeR00	73	2.50	1239
1jb0A00	740	2.50	296
1t11A02	166	2.50	1802
1td6A02	92	2.50	2
1uunA02	52	2.50	28
2q83A02	226	2.50	3287
1ibvA00	81	2.50	20
1jeyA03	63	2.50	132
1jb0L00	151	2.50	135
1x9yA01	170	2.50	10
2re3A02	69	2.50	319
2p62A02	93	2.50	10
1jeyB02	201	2.50	977
1jeyA02	126	2.50	394
1fiqC04	156	2.50	3460
3bh1A02	101	2.51	192
3bh1A01	239	2.51	169
1pc6A00	141	2.51	174
1otsB00	441	2.51	3138

Domain	L	Resol	N
1fcdA03	74	2.53	219
1914A00	171	2.53	310
2j8sB05	223	2.54	12852
2j8sA08	89	2.54	8524
2j8sA04	96	2.54	8560
3bf0A02	98	2.55	1753
1w3fA02	165	2.58	207
2r6iA01	94	2.59	402
1zy8K00	44	2.59	3485
2vglB00	579	2.59	2535
1e50B00	130	2.60	48
2p5zX02	97	2.60	1502
1nktA04	177	2.60	1988
1hynR00	300	2.60	1174
2r7jA01	144	2.60	15
2w9jA00	69	2.60	173
1c4zA02	81	2.60	1640
1nktA02	122	2.60	1625
1k2fA02	56	2.60	385
1lktA00	104	2.60	29
3dplC01	117	2.60	593
1fpsA00	348	2.60	5360
1sigA00	305	2.60	5213
1divA02	90	2.60	1742
1jg5A00	83	2.60	44
1lrvA00	233	2.60	757
3kicA00	520	2.60	141
1t8sA01	140	2.60	260
2nrqA00	136	2.60	226
2bghA01	207	2.60	1909
2bghA02	206	2.60	1944
2hwjB01	118	2.61	214
3cygA02	93	2.61	494
1floC02	257	2.65	9
2idgC00	159	2.69	528
1kl7A01	92	2.70	1277
2vsgA02	160	2.70	19
3bu2A02	69	2.70	177
2vsgA01	198	2.70	128
1bcpB01	86	2.70	4
3b8mC01	151	2.70	218
2o8rB02	151	2.70	1417
3hhwK02	200	2.70	72
3hhwK01	183	2.70	84
1nltA02	66	2.70	2956
2ijrA02	173	2.70	41
1c4kA04	129	2.70	1120
3cslA01	128	2.70	19919
1dt9A01	105	2.70	197
2ha9B00	398	2.70	425
1kf6C00	130	2.70	95
1jroB05	94	2.70	3318
3cucA00	257	2.71	3617
2iahA03	543	2.73	25281
1vz6A01	253	2.75	1344

Domain	L	Resol	N
1jmuB03	157	2.80	18
1tljA00	189	2.80	266
1f45B00	133	2.80	40
1fgjA01	240	2.80	523
1dkgA02	60	2.80	2183
2r6gF03	88	2.80	113
2glfA02	133	2.80	1059
2zihB00	281	2.80	324
1vfgA02	217	2.80	3622
2gk9B01	157	2.80	671
2r6gF02	84	2.80	175
3oaeA00	421	2.80	639
2r6gF04	235	2.80	18893
1fgjA02	230	2.80	226
1xp4C02	89	2.80	1225
1yewB00	238	2.80	419
1jx7A00	114	2.80	1739
1jmuB02	124	2.80	17
1jmuB01	120	2.80	11
1tdjA03	161	2.80	1576
3k6eA00	139	2.81	18446
3doaA01	153	2.81	969
1sxjA03	117	2.85	230
1vsgA02	159	2.90	46
1cjaA01	150	2.90	9
3fgxA00	96	2.90	67
1vsgA01	203	2.90	157
1nt2B01	71	2.90	53
1gaxA04	50	2.90	5
1cjaA02	177	2.90	24
1f02T00	66	2.90	11
2idbA02	113	2.90	800
2gy5A03	135	2.90	7080
1jhnA02	146	2.90	556
3bxjA01	34	3.00	2
1mqsA03	120	3.00	430
2ijzA02	107	3.00	1010
1bo1A02	159	3.00	866
1ltlA01	90	3.00	575
2bpa100	426	3.00	21
1ldjA05	74	3.00	811
1x9nA01	270	3.00	809
2alaA01	124	3.00	33
1qvrC01	139	3.00	4088
1tlyA00	251	3.01	276
3fwlA03	85	3.09	277
3fwlA04	172	3.09	4377
3h9vA02	273	3.10	227
3b8pB00	173	3.10	227
1w36B04	146	3.10	736
1e0fI01	33	3.10	1
2r6fA02	154	3.20	2870
2zjsY00	415	3.20	1843
2aj2A01	92	3.21	885
1shyB02	49	3.22	276

Domain	L	Resol	N
1v7nY00	139	3.30	40
1pw4A02	212	3.30	9546
1x9pA02	297	3.30	47
1tx9A00	141	3.31	5
1zcdA00	376	3.45	1126
1g03A00	134	NMR	7
1n91A00	107	NMR	875
1hdlA00	55	NMR	150
1mknA00	59	NMR	32
2jovA01	71	NMR	272
1gccA00	63	NMR	1591
1auuA00	55	NMR	685
1d4uA00	111	NMR	223
1hywA00	58	NMR	60
1szlA01	52	NMR	2198
2jynA01	136	NMR	165
1e8pA00	46	NMR	69
1o8rA00	94	NMR	65
1z60A00	59	NMR	245
1dqcA00	73	NMR	1200
1xn8A00	131	NMR	39
1o6wA02	33	NMR	1241
2ez5W01	36	NMR	1003
1hnrA00	47	NMR	702
1cz4A02	81	NMR	282
1tbaA00	67	NMR	44
1d1rA00	83	NMR	985
2if1A00	126	NMR	1091
1wfeA00	86	NMR	681
1nblA00	46	NMR	63
2jroA01	65	NMR	74
1lv3A00	65	NMR	570
1nd9A00	49	NMR	981
2hg7A00	60	NMR	21
1nr3A00	122	NMR	16132
1imlA00	76	NMR	2854
1x6aA01	62	NMR	2706
1j57A00	143	NMR	86
2hfqA00	85	NMR	47
1n0zA00	45	NMR	533
2hg6A00	106	NMR	20
1d6gA00	47	NMR	37
1t23A00	93	NMR	46
2hh8A00	127	NMR	50
1yuaA01	64	NMR	115
1ev0A00	58	NMR	539
1xu6A00	80	NMR	86
1apjA00	74	NMR	198
2jv8A00	73	NMR	1
1jbiA00	100	NMR	401
1kmxA00	54	NMR	31
1q5fA01	150	NMR	130
1ul4A01	65	NMR	274
1mkcA00	43	NMR	47
1rhxA00	87	NMR	393

Domain	L	Resol	N
1v9vA01	95	NMR	69
1hn6A00	110	NMR	11
1wvkA00	86	NMR	196
2nwtA00	69	NMR	158
2jneA00	71	NMR	125
1ghhA00	81	NMR	208
2jz6A01	50	NMR	1109
1y7jA00	40	NMR	54
1g10A00	102	NMR	142
1widA00	117	NMR	698
1q48A00	134	NMR	1688
1wibA00	92	NMR	1042
1e8rA00	50	NMR	61
1imtA00	80	NMR	166
1nynA00	111	NMR	254
1ngrA00	85	NMR	334
1a1wA00	83	NMR	256
1gh9A00	71	NMR	65
2gpfA01	63	NMR	552
1q60A00	99	NMR	60
1v9xA00	114	NMR	341
1co4A00	42	NMR	112
2joeA01	128	NMR	124
1so9A00	131	NMR	521
2e6iA00	64	NMR	165
1svjA00	136	NMR	7793

S4.2 Supplementary dataset S2: subset of BetaSheet916 for which PconsC2 predictions could be computed

Because of the high computational cost of running PconsC2, the comparison with bbcontacts and other methods was performed on a subset of the main test set BetaSheet916 containing 186 protein chains. This subset was built in the following manner: for each CATH domain present in BetaSheet916, only the shortest chain in BetaSheet916 containing that CATH domain was retained. All chains containing more than 200 residues were excluded to limit the PconsC2 runtime.
In the following table, the four columns are the PDB chain identifier, the length of the chain L, the resolution of the PDB structure and the number of sequences N in the HHblits alignment.

Chain	L	Resol	N	Chain	L	Resol	N
1s5uE	136	1.7	7558	1nvjB	126	2.15	1395
1dhnA	121	1.65	1697	1c9sM	71	1.9	77
1f3zA	150	1.98	2056	1ameA	66	1.65	66
1oo0A	144	1.85	114	1ir2I	140	1.84	547
1q92A	195	1.4	12947	1mogA	67	1.7	399
1qhvA	195	1.51	67	1mkpA	144	2.35	5135
1n2mC	163	1.9	176	1l5bA	101	2	281
1jzaA	66	2.2	267	1nz0D	111	1.2	1741
1fjrA	188	2.3	95	1g2rA	94	1.35	988
1bxyA	60	1.9	1342	1r6jA	82	0.73	6730
1mk0A	97	1.6	553	1e44A	84	2.4	32
1b66A	138	1.9	1702	1n5bA	128	2	17
1mg3F	125	2.4	47	1pugA	94	2.2	1566
1genA	200	2.15	1065	1bjpA	62	2.4	1872
1h59B	45	2.1	134	1k2dB	185	2.2	6262
1h75A	76	1.7	14797	1flmA	122	1.3	1618
1squB	154	2.4	1344	1nycA	111	1.4	3
1svpB	160	2	17	1v5xA	200	2	1897
1ootA	58	1.39	6749	1lshB	174	1.9	111
1j2lA	68	1.7	1030	1ex6A	186	2.3	12668
1is1A	185	2.2	1773	1uslC	158	1.88	1684
1ni0C	158	2.5	14	1d0dA	60	1.62	2
1hufA	123	2	6	1c5eA	95	1.1	33
1ugiA	83	1.55	1	2a8vA	118	2.4	1039
1qmyA	156	1.9	20	1r94B	97	2.3	2850
1btnA	106	2	3405	1mm9A	127	1.66	84
1g13A	162	2	344	1e5kA	188	1.35	16823
1bx7A	51	1.2	11	1iibA	103	1.8	1485
1extA	160	1.85	976	2prdA	174	2	1347
1whiA	122	1.5	1172	1moxC	49	2.5	521
1lqvB	173	1.6	3569	1feuA	185	2.3	1545
1b33N	67	2.3	277	1o7zA	60	1.92	889
1jsgA	111	2.5	47	1jtgB	165	1.73	14
1fqtA	109	1.6	7963	1js2A	89	1.9	155
1d8lB	140	2.5	2523	1b13A	54	1.5	1748
1icfI	65	2	596	1gefA	120	2	672
1jatA	152	1.6	4048	1jj2W	82	2.4	444
1fxkC	133	2.3	796	1gmxA	108	1.1	13191
1hruA	186	2	3867	1ezgB	82	1.4	48

Chain	L	Resol	N
1rlhA	151	1.8	154
1jj2E	172	2.4	2174
1jj2L	194	2.4	296
1jj2T	53	2.4	354
1j85A	156	2	6797
1ufhB	154	2.2	13970
1f39B	101	1.9	4647
1ec6A	87	2.4	2799
1ihnA	113	2.2	428
1ew4A	106	1.4	496
1v54F	98	1.8	234
1lm4B	184	1.45	3002
1d1mB	65	2.05	2237
1ptqA	50	1.95	1299
1hdfB	100	2.35	146
1ihfA	96	2.5	4009
1jj2Y	73	2.4	328
1nrzA	163	1.75	942
1go4C	195	2.05	399
1a5kA	100	2.2	576
1oqjA	90	1.55	149
1qysA	92	2.5	1
1bysA	152	2	7308
1kh8A	125	2	455
1kcqA	103	1.65	798
1i8nA	89	2.2	1
1ltiA	185	2.13	155
1qqhA	144	2.1	270
1jhsA	188	1.9	197
3rhnA	115	2.1	5129
1gpqB	128	1.6	90
1gp0A	133	1.4	434
1e6tA	129	2.20	4
1n07B	155	2.45	2096
1ku6B	61	2.5	287
1h4yA	115	1.61	6272
1i4jB	110	1.8	1917
1iwmA	177	1.9	441
1f47B	144	1.95	321
1o5uA	88	1.83	3723
1j3lB	164	2.3	1748
1oqwB	144	2	6513
1rlkA	116	1.95	684
1i59B	188	1.8	21133
1qgwA	76	1.63	26
1fsjB	134	1.8	128
1pqfA	127	2	836
1fmbA	104	1.8	3089
1dg5A	159	2	3004
1ds6B	179	2.35	282
1nplA	109	2	2022
1o13A	107	1.83	1654
1mi0A	61	1.85	16
1fuxB	164	1.81	1989
1udzA	179	1.8	6684

Chain	L	Resol	N
1jj2N	115	2.4	2230
1l8rA	101	1.65	75
1a73A	162	1.8	18
1m8nA	120	2.45	24
1fx3B	149	2.5	561
1durA	55	2	14150
1b2uD	90	2.1	543
1xxaC	73	2.2	855
1gwyB	175	1.71	51
1e79H	131	2.4	2128
1g3kA	173	1.9	1444
1tulA	102	2.2	28
1k9jA	130	1.9	6505
1agqD	95	1.9	316
1iktA	115	1.75	1483
1ucrB	75	1.2	48
2ablA	163	2.5	4215
1ca9A	191	2.3	1277
1g6gA	127	1.6	5380
1ocuA	134	2.3	2515
1g1bA	164	1.99	499
1ycqA	88	2.3	53
1b78A	184	2.2	2381
1hxrB	115	1.65	130
1d0qA	102	1.71	2742
1vjhA	120	2.1	679
1kptA	105	1.75	46
1jyhA	155	1.8	2695
1q9uB	128	1.8	871
1h8pA	88	1.82	405
1hjzB	192	1.7	2139
1cw0A	155	2.3	1091
1g5cC	169	2.1	2581
1pchA	88	1.8	2258
1c2aA	120	1.9	177
1oapA	108	1.93	8567
1uutA	195	2	27
1a9nA	162	2.38	21317
1rmdA	116	2.1	10578
1no5B	102	1.8	3591
1di2A	69	1.9	2899
1lj0A	89	2	3392
1kuhA	132	1.6	543
1n9nA	108	2.3	4516
1ub4C	75	1.7	1719
1fd4A	41	1.7	191
1nfjA	87	2	285
1ktgA	137	1.8	19521
1nn7A	105	2.1	1428
1bylA	122	2.3	15658
1gmuC	140	1.5	535
1josA	100	1.7	1627
1dqoA	134	2.2	202

Supplementary references

Burkoff,N.S. *et al.* (2013). Predicting protein β-sheet contacts using a maximum entropy-based correlated mutation measure. *Bioinformatics*, **29**, 580–587.

Cheng,J. and Baldi,P. (2005). Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, **21 Suppl 1**, i75–84.

Dunn,S.D. *et al.* (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Jones,D.T. *et al.* (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Kamisetty,H. *et al.* (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 1–6.

Remmert,M. *et al.* (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Savojardo,C. *et al.* (2013). BCov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming. *Bioinformatics*, **29**, 3151–3157.

Seemayer,S. *et al.* (2014). CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, page 10.1093/bioinformatics/btu500.

Skwark,M.J. *et al.* (2014). Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS Comput. Biol.*, **10**(11), e1003889.

Wang,Z. and Xu,J. (2013). Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics*, **29**(13), i266–273.