Appendix A

Naming Convention

Within in the whole thesis, we use the following naming convention:

- PDB [22] ids are giving as four letter codes, for example, $2uag$.
- The chain id is giving in capital letters after the PDB id, for example, $2uagA$. In proteins containing only a single chain, the chain is denoted with '_', e.g., $1ars\_$.
- SCOP [169] domain identifiers are denoted numerical by a single number after the PDB id and the chain id: $2uagA1$. Single chains containing only a single domain are denoted with '_': $1dhs\_\_$.
- CATH [176] domain identifiers are denoted by two numbers after the PDB id and the chain id: $2uagA01$. Single chains are denoted with '0' as well as single domains: $1dhs000$.
- PTGL [157] folding graphs are denoted in capital letters in alphabetic order from the N-to the C-terminus: $1timAA$ or $1ars\_A$.

Appendix B

Statistical Potentials

A pairwise *statistical interaction potential* measures the preference of pairs of amino acids to be in spatial proximity in the 3D protein structure. These *statistical potentials of mean force* (or *knowledge-based potentials*) are derived from the inverse Boltzmann relation to a given set of known 3D protein structures [163,212] setting the energy of a certain state in relation to the probability that the state is observed. The advantage of knowledge-based potentials is their small computational expense compared to more complex potential expressions used in today's force fields derived from first principle laws. The disadvantage of knowledge-based potentials is the fact that they are derived from a specific dataset of proteins. Therefore they depend on the quality of the set and might vary considerably using different datasets. Here, we derive the statistical potential in a similar way then used before by Lu *et al.* [150]. The method can be used for all types of contact definitions between residues of different residue subsets R. These subsets could include all residues (*total*), all SSE residues (SS), all helix residues (H), or all strand residues (E). The potential energy between two types of amino acids i and j is then given by

$$P_R(i,j) = -\log\left(\frac{N_R(i,j)}{\frac{n_R(i)n_R(j)}{n_R^2}N_R}\right) \quad with \tag{B.1}$$

$$N_R = \sum_{\substack{i,j \\ i<j}} N_R(i,j) \;\; and \tag{B.2}$$

$$n_R = \sum_i n_R(i) \;\; . \tag{B.3}$$

$N_R(i,j)$ is the number of contacts between residue types i and j within the residue subset R, and $n_R(i)$ the total number of residues of type i that have contact to any other residue. The number of contacts also depends on the contact definition used. For example, when looking at all contacts between SSEs the total number of contacts is much bigger than when looking only at contacts between helices and strands. For both potentials the total number of residues would be the same. Statistical potentials are represented as a 20×20 interaction potential matrix giving the propensities of pairwise interactions of the 20 standard amino acids for a certain subset R.

Appendix C

Graph-theoretical Definitions

In the following all graphs are simple undirected labeled graphs $G = (V, E)$ as defined in Definition 2. If two vertices $u, v \in V$ of G are connected by an edge $e \in E$ this is denoted by $e = (u, v)$ and the two vertices are said to be *adjacent*. The edge e is said to be *incident* to both vertices u and v.

Definition 42 (Complete Graph)**.** *A graph $G = (V, E)$ is called a* complete *graph, if all its vertices are adjacent, i.e., $\forall i, j \in V, i \neq j : (i, j) \in E$.*

Definition 43 (Subgraph)**.** *A* subgraph *of a graph $G = (V, E)$ is a subset $S \subseteq V$ of vertices of G together with a subset of edges connecting pairs of vertices in S.*

Definition 44 (Clique)**.** *A complete subgraph $H \subseteq G$ is called a* clique *or* maximal complete subgraph *if there is no clique J such that $H \subseteq J \subseteq G$ with $|H| < |J|$, i.e., a clique is a complete subgraph that is not contained in any other complete subgraph.*

Definition 45 (Graph Isomorphism)**.** *Two graphs G_1 and G_2 are said to be* isomorphic*, denoted by $G_1 \equiv G_2$, if there exists a bijection $f : V_{G_1} \rightarrow V_{G_2}$ between their vertices preserving all adjacencies. This means if $(u, v) \in E_1 \Rightarrow (f(u), f(v)) \in E_2$.*

Definition 46 (Subgraph Isomorphism)**.** *The subgraphs $G'_1 \subseteq G_1$ and $G'_2 \subseteq G_2$ are said to be* isomorphic *if $G'_1 \equiv G'_2$. Then, G'_1 and G'_2 are called* common subgraphs *of G_1 and G_2.*

Appendix D

Datasets

D.1 ASTRAL SCOP40 Dataset

The ASTRAL SCOP40 dataset [41] is a non-redundant dataset consisting of protein domains defined from SCOP [169] version 1.69. The domain structures have at most 40% sequence identity. The SCOP40 dataset can be downloaded from the ASTRAL webpage[1]. Since most of the used methods rely on SSEs, we excluded all domain structures with less than two SSEs and less than 30% of their residues within SSEs, and all NMR structures. Additionally, only domains from the four main SCOP classes were included (*all* α, *all* β, α/β, $\alpha+\beta$) resulting in 5,397 protein domain structures. This dataset is used for all database searches with the web version of GANGSTA.

From the SCOP40 dataset we generated two additional datasets that we used to calculate the statistical significance of GANGSTA alignments:

- *SAME40* consists of 4,982 random pairs of domains from the SCOP40 dataset where the two domains are from the same SCOP superfamily. The protein pairs involve 672 different SCOP domains taken from 113 different SCOP superfamilies belonging to 99 different SCOP folds.
- *DIFF40* consists of 88,909 random pairs of domains from the SCOP40 dataset where for each pair the protein domains are from different SCOP superfamilies. This dataset of protein pairs involves 500 different SCOP domains from 317 different SCOP superfamilies belonging to 243 different SCOP folds.

D.2 Four-Helix-Bundle Dataset

The four-helix-bundle dataset comprises ten proteins belonging to four different folds and six different superfamilies in the SCOP [169] classification scheme. Table D.1 in the Appendix shows the dataset of the ten proteins with their SCOP annotations. This dataset was used before in [62, 63].

[1] `http://astral.berkeley.edu/scopseq-os-1.67.html`

Table D.1: **Four-Helix-Bundle dataset.** SCOP [169] fold and superfamily identifiers.

protein	SCOP fold	SCOP superfamily
$2hmzA$	47161	47188
$2ccyA$	47161	47175
$256bA$	47161	47175
$1bbhA$	47161	47175
$1le2$	47161	47162
$3inkC$	47265	47266
$1bgeB$	47265	47266
$1rcb$	47265	47266
$1aep$	47856	47857
$1flx$	*designed*[a]	*designed*[a]

[a] SCOP category for a selection of artificial protein structures.

D.3 TRAF Dataset

The TRAF dataset consists of eight proteins that belong to two different folds in the all-β class of the SCOP [169] database. Four proteins (PDB-IDs: $1czyA2$, $1kzzA1$, $1lb4$, $1k2fA$) belong to the 'TRAF (TNF Receptor Associated Factor) domain-like' fold but are members of two different families: $1czyA$, $1kzzA1$, and $1lb4$ are taken from the 'TRAF domain' family; $1k2fA$ belongs to the 'SIAH' family. Four proteins (PDB-IDs: $1bmg$, $1frtB$, $1igtA2$, $1k8iA1$) of the TRAF dataset belong to the 'C1 set domains' family of the 'Immunoglobulin-like beta-sandwich' fold. This dataset was used before in [62, 63].

D.4 C2 Dataset

The C2 dataset consists of ten proteins taken from two families of the 'C2 domain' superfamily in SCOP [169]. The proteins $1a25A$, $1rsy$, $3rpbA$, and $1dsyA$ are from the 'Synaptotagmin-like' family. $1rlw$, $1gmiA$, $1bdyA$, $1e8yA2$, $1qasA2$, and $1d5rA1$ are from the 'PLC-like' family.

D.5 Rossmann-Fold Dataset

The Rossmann-fold dataset consists of seven protein domains that contain motifs that are classified as 'Rossmann'-fold or 'Rossmann-like'-fold according to the CATH [176] or SCOP [169] classification schemes. The proteins (target structures) are listed in Table D.2. All proteins have pairwise less than 40% sequence similarity.

Table D.2: **Rossmann dataset.** Correspondence between SCOP- and CATH-ids and the CATH hierarchy identifiers.

SCOP id	*CATH id*	CAT^a	$CATH^b$
1*cjcA*2	1*cjcA*01	3.40.50	3.40.50.720
1*rqlA*_	1*rqlA*01	3.40.50	3.40.50.1000
	1*rqlA*02	1.10.164	1.10.164.10
1*geeA*_	1*geeA*00	3.40.50	3.40.50.720
1*dhs*__	1*dhs*000	3.40.910	3.40.910.10
1*dih*_1	1*dih*001	3.40.50	3.40.50.720
1*f*0*kA*_	1*f*0*kA*01	3.40.50	3.40.50.2000
	1*f*0*kA*02	3.40.50	3.40.50.2000
1*f*8*yA*_	1*f*8*yA*00	3.40.50	3.40.50.1810

[a] CAT means Class-Architecture-Topology code according to CATH [176]
[b] CATH means Class-Architecture-Topology-Homologous Superfamily code according to CATH.

D.6 CP Dataset

The CP dataset consist of seven protein pairs that are known as circular permuted proteins from literature [97,117]: $1rin-2cna$, $1nkl-1qdm$, $1rsy-1qas$, $1aqi-1boo$, $1onr-1fba$, $1gbg-1ajk$, and $1avd-1swg$.

D.7 DIFFAL Dataset

The DIFFAL dataset consist of ten protein-structure pairs introduced by Fischer *et al.* [70] and used by Novotny *et al.* [172] that are known representing difficult pairwise alignments. Novotny added the last pair ($1g61/1jdw$). The PDB ids of the protein pairs are:
$1bgeB/2gmfA$, $1cewI/1molA$, $1cid/2rhe$, $1crl/1ede$, $1fxiA/1ubq$, $1ten/1hhrB$, $1tie/4fgf$, $2azaA/1paz$, $2sim/1nsbA$, $3hlaB/2rhe$, and $1g61/1jdw$.

D.8 Novotny Dataset

The Novotny dataset consists of representative proteins from four different CATH [176] classes (classes: mainly-α, mainly-β, mixed-*alpha*-β, few SSEs) and was applied in a recent structure alignment performance test by Novotny *et al.* [172]. For all protein domains their corresponding CATH class, the CATH topology classification, the number of different superfamilies per topology level in the dataset, and the CATH domain identifier are given in Table D.3. The whole Novotny dataset and the benchmark results are available on `http://xray.bmc.uu.se/~marian/servers/index.htm`.

Table D.3: **Novotny dataset [172].** CATH [176] classification for the Novotny dataset.

Class	CAT^a	NoH^b	CATH entries
mainly-α	1.10.40	2	1*rlr*001 1*yfm*003 1*furA*03 1*auwA*03 1*jswB*03 1*hylC*03 1*i*0*aA*03
	1.10.164	3	1*aq*6*A*02 1*c*3*uA*02*c*[c] 1*fezA*02 1*jud*002 1*zrn*001
	1.25.10[d]	3	1*b*3*uA*00 1*bk*6*A*00 1*gcjA*00 1*ialA*00 1*ibrB*00 1*qbkB*00 2*bct*000
mainly-β	2.30.110	2	1*ci*0*A*00 1*dnlA*00 1*ejeA*00 1*i*0*rA*00
	2.40.100	1	1*a*33000 1*awgA*00 1*cynA*00 1*dywA*00 1*ihgA*01 1*lopA*00 1*qngA*00 1*qoiA*00 2*rmcA*00
	2.100.10	3	1*c*3*kA*00 1*ciy*002 1*jacA*00 1*jotA*00 1*dlc*003 1*vmoA*00
mixed-α-β	3.10.50[e]	2	1*bkf*000 1*grj*002 1*pbk*000 1*rot*000 1*yat*000
	3.40.91	3	1*bhmA*00 1*cfr*000 1*d*2*iA*00 1*fokA*03
	3.70.10	3	1*axcA*00 1*b*77[f] 1*czd*[f] 1*dmlA*00 1*ge*8*A*00 1*plq*000
few SSEs	2.40.20	1	1*b*2*iA*00 1*ceaA*00 1*kdu*000 1*kiv*000 1*krn*000 1*pk*4000 1*pmlA*00 5*hgpg*[g]

[a] Class-Architecture-Topology code according to CATH.
[b] Number of homologous superfamilies (H level in CATH) of the topology level.
[c] Has changed CATH topology from L-2-haloacid Dehalogenase, domain 2, to Fumarase C Chain A, domain 2.
[d] 1.25.30 in the original Novotny dataset. All domains moved to 1.25.10.
[e] 3.10.70 in the original Novotny dataset. All domains moved to 3.10.50.
[f] Not classified anymore.
[g] Not in the PDB [22] and therefore not classified anymore.

D.9 Fischer Dataset

The Fischer dataset [70], as used, for example, by Novotny *et* al. [172], consists originally of 68 *reference* sequences and 301 *target* structures. For every *reference* sequence there is at least one *target* structure with similar fold type in the dataset. All pairs were hand selected showing high structure similarity but low sequence similarity. Since two of the *reference* structures and 32 of the *target* structures contain more than one domain according to SCOP [169] classification scheme we extended the number of *reference-target* pairs too 70 and the total number of *target* structures to 333 structures, respectively.

Table D.4: **Fischer dataset [70].** For every *reference* structure the matching *target* structures are given according SCOP [169] fold classification.

reference	targets
Mostly alpha	
1dxtb_	1hbg__ 1f99b_ 1mbc__ 1cpca_
1cpcl_	1cpca_ 1hbg__ 1f99b_ 1mbc__
1c2ra_	1ycc__ 451c__ 2mtac_
2mtac_	1ycc__ 451c__
1bbha_	2ccya_ 256ba_
1bgeb_	2gmfa_ 1huw__ 1rcb__ 1rfba_
1rcb__	2gmfa_ 1bgeb_ 1rfba_ 1huw__
1aep__	256ba_ 2ccya_
1osa__	4cpv__ 2scpa_ 2bbma_ 1rec__ 1scmb_ 1scmc_ 1top__
2sas__	2scpa_ 4cpv__ 2bbma_ 1rec__ 1scmb_ 1scmc_ 1top__
1enh__	1lfb__
1lgaa_	2cyp__ 1mypa1 1mypc1
2hpda_	2cpp__
Mostly Beta	
1fc1a1	2fb4h2 8fabb1 8fabb2 2rhe__ 3hlab_ 3cd4_1 3cd4_2 3hlaa_ 1cid_1 1cid_2 8faba1 8faba2 1mcoh1 1mcoh2 1mcoh3 1mcoh4
1fc1a2	2fb4h2 8fabb1 8fabb2 2rhe__ 3hlab_ 3cd4_1 3cd4_2 3hlaa_ 1cid_1 1cid_2 8faba1 8faba2 1mcoh1 1mcoh2 1mcoh3 1mcoh4
2fbjh1	8fabb1 2fb4h2 8fabb2 2rhe__ 3hlab_ 3cd4_1 3cd4_2 3hlaa_ 1cid_1 1cid_2 8faba1 8faba2 1mcoh1 1mcoh2 1mcoh3 1mcoh4
2fbjh2	8fabb2 8fabb1 2fb4h2 2rhe__ 3hlab_ 3cd4_1 3cd4_2 3hlaa_ 1cid_1 1cid_2 8faba1 8faba2 1mcoh1 1mcoh2 1mcoh3 1mcoh4
1cid_1	2rhe__ 2fb4h2 8fabb1 8fabb2 3hlab_ 3cd4_1 3cd4_2 3hlaa_ 8faba1 8faba2 1mcoh1 1mcoh2 1mcoh3 1mcoh4
1cid_2	2rhe__ 2fb4h2 8fabb1 8fabb2 3hlab_ 3cd4_1 3cd4_2 3hlaa_ 8faba1 8faba2 1mcoh1 1mcoh2 1mcoh3 1mcoh4
1pfc__	3hlab_ 2fb4h2 8fabb1 8fabb2 2rhe__ 3cd4_1 3cd4_2 3hlaa_ 1cid_1 1cid_2 8faba1 8faba2 1mcoh1 1mcoh2 1mcoh3 1mcoh4
1ten__	3hhrb1
1tlk__	2rhe__ 2fb4h2 8fabb1 8fabb2 3hlab_ 3cd4_1 3cd4_2 3hlaa_ 1cid_1 1cid_2 8faba1 8faba2 1mcoh1 1mcoh2 1mcoh3 1mcoh4
3cd4_1	2rhe__ 2fb4h2 8fabb1 8fabb2 3hlab_ 3hlaa_ 1cid_1 1cid_2 8faba1 8faba2 1mcoh1 1mcoh2 1mcoh3 1mcoh4
3cd4_2	2rhe__ 2fb4h2 8fabb1 8fabb2 3hlab_ 3hlaa_ 1cid_1 1cid_2 8faba1 8faba2 1mcoh1 1mcoh2 1mcoh3 1mcoh4
3hlab_	2rhe__ 2fb4h2 8fabb1 8fabb2 3cd4_1 3cd4_2 1cid_1 1cid_2 8faba1 8faba2 1mcoh1 1mcoh2 1mcoh3 1mcoh4
1aaj__	1paz__ 1aoza1 2pcy__ 2azaa_ 2afna1

```
reference   targets

Mostly beta
2afna1      1aoza1 1paz__ 2pcy__ 2azaa_
2azaa_      1paz__ 1aoza1 2pcy__ 2afna1
4sbva_      2tbva_ 2plv1_ 1tmv11 1tmv21 2mev3_
1bbt1_      2plv1_ 2tbva_ 1tmv11 1tmv21 2mev3_
1saca_      2ayh__ 2cna__ 1slta_
1ltsd_      1bova_
1tie__      4fgf__
8i1b__      4fgf__
1arb__      5ptp__ 1bbre1 2pkaa1
2sga__      5ptp__ 1bbre1 2pkaa1
2snv__      5ptp__ 1bbre1 2pkaa1
1ftpa_      1ifc__ 1rbp__
1mup__      1rbp__ 1ifc__
2sim__      1nsba_
1caub_      1caua_
Alpha/Beta
1chra1      2mnr_1 4enl_1
2mnr_1      4enl_1
3rubl1      6xia__ 5ruba1
1crl__      1ede__ 1tca__ 3tgl__ 3sc2b1 1taha_ 2ace__
1taha_      1tca__ 1ede__ 3tgl__ 3sc2b1 2ace__
1aba__      1kte__ 2trxa_ 1gp1a_ 1dsba2 6gsta2
1dsba2      2trxa_ 1kte__ 1gp1a_ 6gsta2
1gp1a_      2trxa_ 1kte__ 1dsba2 6gsta2
1atna1      1atr_1 1glag1
1hrha1      1rnh__
3chy__      2fox__
2ak3a1      1gky__ 3adk__ 5p21__ 2reb_1 1nipb_
1gky__      3adk__ 5p21__ 2reb_1 1nipb_
2cmd_1      6ldh_1 8adh_2 1gd1o1 1dhr__ 1gdha1 2pgd_2
1eaf__      4cla__
2gbp__      2liv__
1mioc_      2minb_
2pia_2      1fnb_2
1gal_1      3cox_1 3grs_1 1trb_1 1phh_1 2tmda2
1npx_1      3grs_1 3cox_1 1trb_1 1phh_1 2tmda2
Alpha+Beta
1fxia_      1ubq__ 2pia_3
1cewi_      1mola_
1stfi_      1mola_
2pnb__      1shaa_
2sara_      9rnt__
1onc__      7rsa__
5fd1__      1iqza_
Other
2hhma_      1fbpa_
1hip__      2hipa_
1isua_      2hipa_
```

Appendix E

List of Structural Alignment Methods

Here, we provide of an alphabetically ordered list of state-of-the-art methods for protein structure alignment that were used in recent evaluation tests for structural alignment methods [37,172] or that were used throughout this thesis:

- **CE** (Combinatorial Extension of the optimal pathway) [210] attempts to find the best possible alignment of two structures by combinatorial extension of the path of aligned fragment pairs that satisfy certain criteria regarding structural similarity. The evaluation of structural similarity is based on inter-residue distances and the RMSD of the matched atoms after rigid body superpositioning. Gaps are allowed, but the maximum size of a gap is restricted. A z-score is used as significance measure, and it is calculated for the alignment of two structures. Therefore, the probability of finding an alignment of the same length when comparing two random structures is evaluated.

- **DALI** [103] calculates residue-residue distance matrices from 3D coordinates of proteins. The distance matrices are first divided into hexapeptide fragments to simplify the alignment task in later stages. DALI attempts to find common local patterns in two fragments of the distance matrices. Such fragments are paired, stored, and combined into larger overlapping segments. The alignment of fragments within the segment is further optimized by a Monte Carlo method, which is not guaranteed to converge to the globally optimal solution. Therefore, several alignments are optimized in parallel, which yields the best, second best, and so on solutions. The method is fully automatic and allows sequence gaps of any length, reversal of chain direction, and free topological connectivity of aligned segments.

- **DEJAVU** [124,153] uses SSEs, represented as vectors, to detect structural similarity between the $reference$ and database structures. The structural similarity is defined with regard to the number of SSEs, their lengths, mutual distances and angles, and, optionally, connectivity and directionality. Results from the SSE-based search are first refined by RMSD minimization (based on Cα atoms) and then by a dynamic programming procedure.

The hits are sorted according to their significance, expressed as a P-value and a z-score as defined by Levitt and Gerstein [146]

- **GRATH** [98] is a graph-based structure comparison program. Protein structures are described as protein graphs composed of nodes and edges, where nodes represent the SSEs and edges correspond to the geometrical relationships (chirality, distances, and angles) between the SSEs. GRATH is intimately coupled to CATH [176].

- **K2** [219, 219] aligns first the SSEs and then extends the alignment to include any equivalent residues found in loops or turns. A genetic algorithm that was later replaced by a simulated annealing procedure determined the initial secondary structure element alignment. After refinement of the SSE alignment, the protein backbones are superposed and a search is performed to identify any additional equivalent residues in a convergent process. Alignments are evaluated using intramolecular distance matrices. Alignments can be performed with or without sequential connectivity constraints.

- **LGA** [247] takes into account local and global structure superpositions. It first generates many local superpositions to detect many different regions where proteins are similar, and additional searches for the largest (not necessary continuous) set of similar residues that deviate not more than a specified distance cutoff. It uses two measures of similarities, the LCS (longest continuous segments) and GDT (global distance test) (see also Equation 4.2).

- **LOCK** [211] is a hierarchical structure-superposition method that tries to minimize the RMSD of two structures at three levels. The starting superposition is obtained by aligning SSEs, represented as vectors, with use of dynamic programming. The RMSD is minimized for corresponding Cα atoms in the second step. In the third step, the core of the structure is defined and an RMSD minimization is once more applied to this core.

- **MATRAS** [121] uses a Markov transition model of evolution to measure protein structure similarity. Three types of structural similarity scores are used: an environmental score (a combination of local structure and solvent accessibility), a residue distance score, and an SSE score. The program uses a hierarchical alignment algorithm. The first alignment is obtained by comparing SSEs using a branch-and-bound method. Using more detailed environmental and residue distance scores further improve this initial alignment. The significance of the results is expressed as a z-score.

- **MAMMOTH** [179] (MAtching Molecular Models Obtained from THeory) uses a heuristic method to find, in a sequence-independent mode, the maximal structural subset of two proteins with the same backbone and 3D conformation. It provides a score of the significance of the alignment found based on the probability of obtaining the structural superimposition by chance when any two different folds of that length in the database are compared.

- **PRIDE** [40] describes protein structures by a set of distributions of Cα-Cα distances. Structural similarity is evaluated as the similarity of the distance distributions and is expressed as a score that varies between 0 and 1 (where a value of 1 indicates identical structures).

- **SCALI** [245] aligns two protein structures in a three-step process. First a library of gapless local fragments is generated using hidden markov models. The second step is a tree search in alignment space, where each branch point is the addition of a new fragment to the alignment. Finally, the best alignments are pruned and extended.

- **SSAP** [178] uses double dynamic programming to produce a structural alignment based on atom-to-atom vectors in structure space. SSAP constructs vectors using Cβ-atoms for all residues except glycine, a method, which thus takes into account the rotameric state of each residue as well as its location along the backbone. SSAP works by first constructing a series of inter-residue distance vectors between each residue and its nearest non-contiguous neighbors on each protein. A series of matrices are then constructed containing the vector differences between neighbors for each pair of residues for which vectors were constructed. Dynamic programming applied to each resulting matrix determines a series of optimal local alignments which are then summed into a 'summary' matrix to which dynamic programming is applied again to determine the overall structural alignment.

- SSM [134] represents SSEs as vectors that are combined into a protein graph and uses a rapid graph-matching algorithm to match the SSE graphs of query and database structures. Subsequently, the Cα atoms of matched SSEs plus some nearby atoms are superimposed. A target function that depends on the number of matched atoms and their RMSD is minimized. The significance of the hits is evaluated with a P-value and a z-score.

- **TM-align** [248] combines TM-score (Equation 4.3) rotation and dynamic programming. It uses first SSE representation and alignment, and then heuristic iterations of superpositions optimizing the TM-score.

- **TOP** [149] aligns subsets of SSEs in two proteins, and their similarity is measured by the angles between aligned SSEs, the distances between matched SSEs and the RMSD of the superimposed coordinates. If the number of matched SSEs for a structure exceeds a certain fraction of all its SSEs, TOP considers these two structures to be structurally similar. It then proceeds with the second comparison stage, which entails detecting the matching residues.

- **TOPS** [81] compares topology diagrams of protein structures instead of the more conventional approaches involving SSEs and Cα coordinates or distances. A TOPS diagram is a simplified representation of protein structure as a string of SSEs that preserves connectivity and directionality and also contains information about hydrogen bonds and chirality. In pictorial representations, helices are represented as circles and strands as triangles. The query structure is converted into a so-called TOPS pattern, which is basically a TOPS diagram in which gaps for insertion of SSEs are allowed.

The TOPS pattern is compared to a database of TOPS diagrams, which is a fast way to do a symbolic structure comparison.

- **TOPSCAN** [155] is a rapid, but approximate, method for protein structure comparison. It was developed as a prescreen method for the more sophisticated, but slow, SSAP [178] TOPSCAN translates structure information into topology strings. Topology strings are defined at two levels: the primary topology represents the state of the secondary structure for each residue (helix or strand), and the secondary topology contains additional information about length, direction, proximity, and accessibility of SSEs. A global dynamic programming algorithm aligns the topology strings, and a similarity score is calculated.

- **VAST** [79] uses a graph theory-based approach to align SSEs. Pairs of SSEs (one from each structure) are represented as nodes if they are of the same type. These nodes are connected by an edge if the angles and the distances between corresponding SSEs from the two proteins do not violate certain constraints. The graph shows the correspondence of SSE pairs based on type, relative orientation, and connectivity. The significance of the results is indicated by a P-value, which is defined as the probability of obtaining the results by chance alone, multiplied by the number of possible, alternative substructure alignments for the given pair of structures.

- **Vorolign** [26] aligns protein structures using double dynamic programming and measures the similarity between residues on the evolutionary conservation of their Voronoi-contacts (see 2.7.3).

- **Yakusa** [37] is a local structural similarity searching method that works at the residue Cα level and makes use of the backbone internal angles.

Appendix F

Additional Figures

F.1 Protein Structure

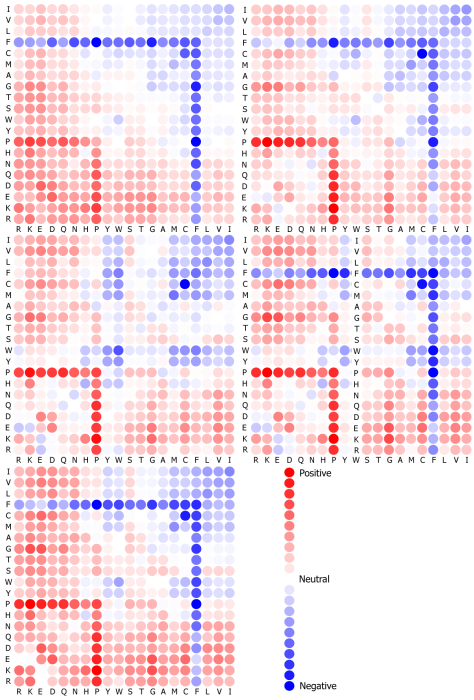

Figure F.1: ***SS* contact potentials using DSSP [119]**. The SSE residue potentials for the five contact definitions using DSSP for SSE assignment. The color spectrum goes from the most negative energy (blue) over neutral (white) to the most positive energy value (red). Top left: *ca*, top right: *cb*, middle left: *all*, middle right: *vdW*, bottom left: *vor*.

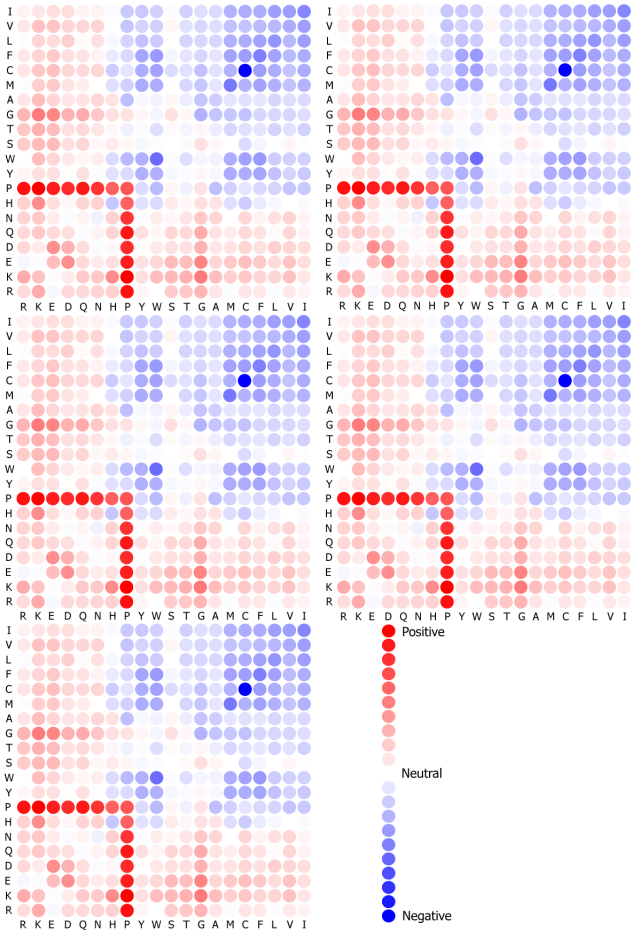

Figure F.2: ***Helix-helix* contact potentials using DSSP [119]**. The *helix-helix* residue potentials for the five contact definitions using DSSP for SSE assignment. The color spectrum goes from the most negative energy (blue) over neutral (white) to the most positive energy value (red). Top left: *ca*, top right: *cb*, middle left: *all*, middle right: *vdW*, bottom left: *vor*.

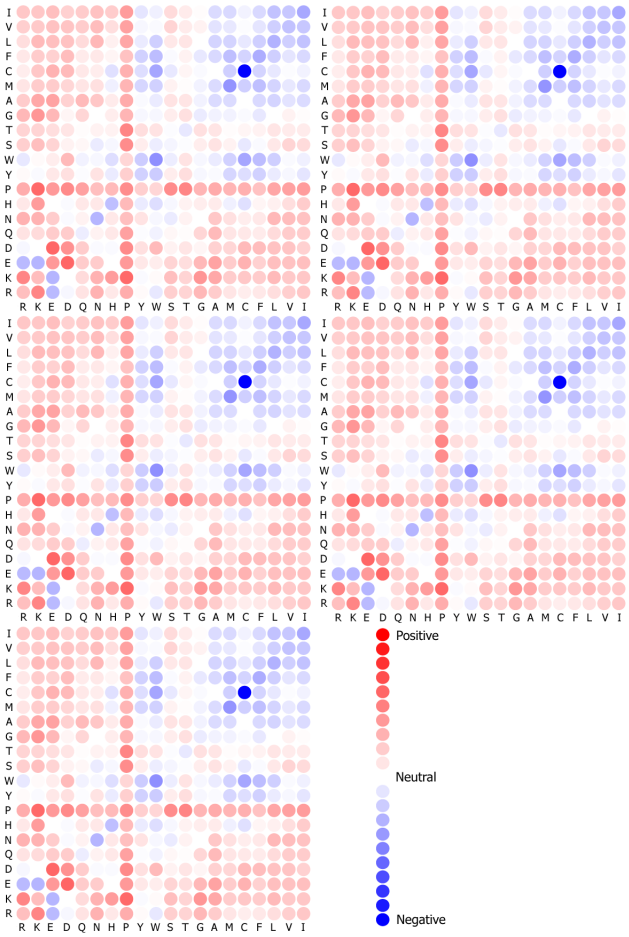

Figure F.3: ***Strand-strand* contact potentials using DSSP [119]**. The *strand-strand* residue potentials for the five contact definitions using DSSP for SSE assignment. The color spectrum goes from the most negative energy (blue) over neutral (white) to the most positive energy value (red). Top left: *ca*, top right: *cb*, middle left: *all*, middle right: *vdW*, bottom left: *vor*.

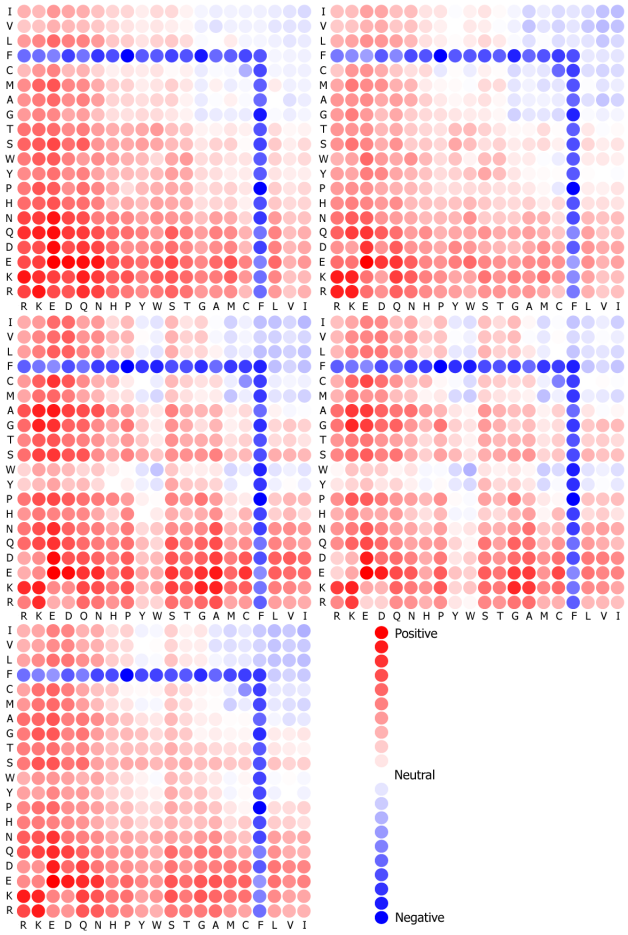

Figure F.4: ***Helix-strand* contact potentials using DSSP [119].** The *helix-strand* residue contact potentials for the five contact definitions using DSSP for SSE assignment. The color spectrum goes from the most negative energy (blue) over neutral (white) to the most positive energy value (red). Top left: *ca*, top right: *cb*, middle left: *all*, middle right: *vdW*, bottom left: *vor*.

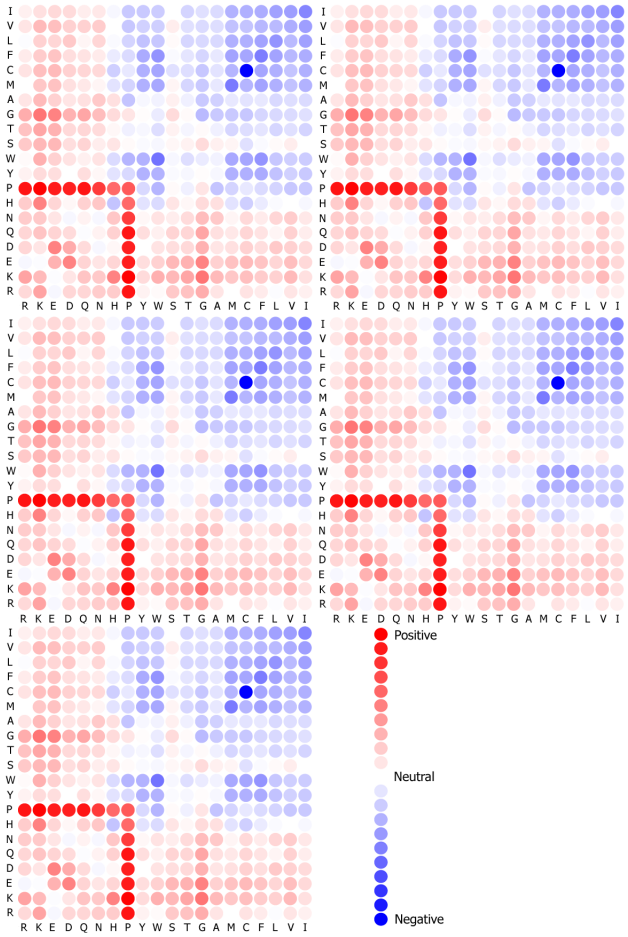

Figure F.5: ***Helix-helix* contact potentials using Stride [74].** The *helix-helix* residue potentials for the five contact definitions using Stride for SSE assignment. The color spectrum goes from the most negative energy (blue) over neutral (white) to the most positive energy value (red). Top left: *ca*, top right: *cb*, middle left: *all*, middle right: *vdW*, bottom left: *vor*.

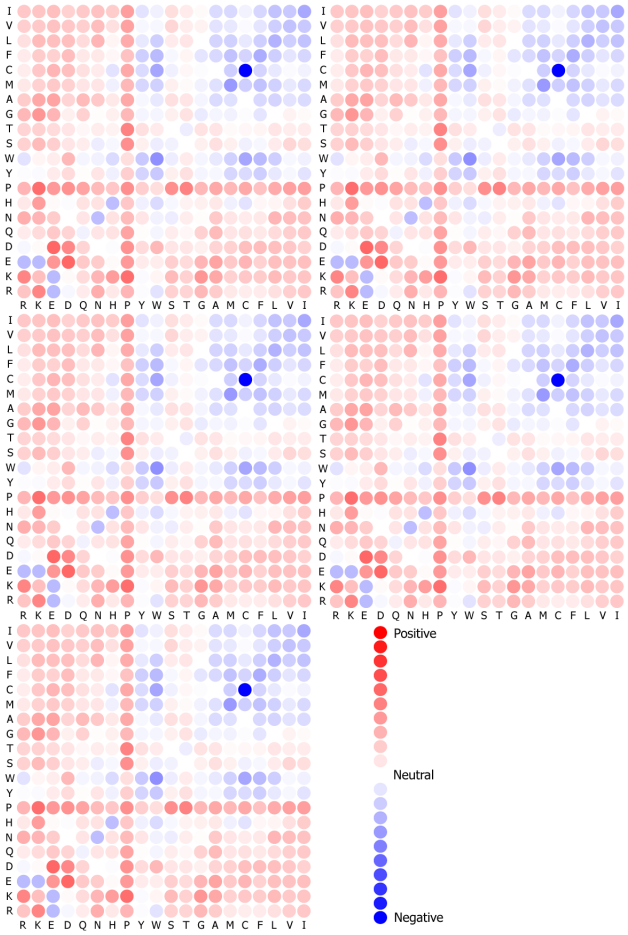

Figure F.6: ***Strand-strand* contact potentials using Stride [74].** The *strand-strand* residue potentials for the five contact definitions using Stride for SSE assignment. The color spectrum goes from the most negative energy (blue) over neutral (white) to the most positive energy value (red). Top left: *ca*, top right: *cb*, middle left: *all*, middle right: *vdW*, bottom left: *vor*.

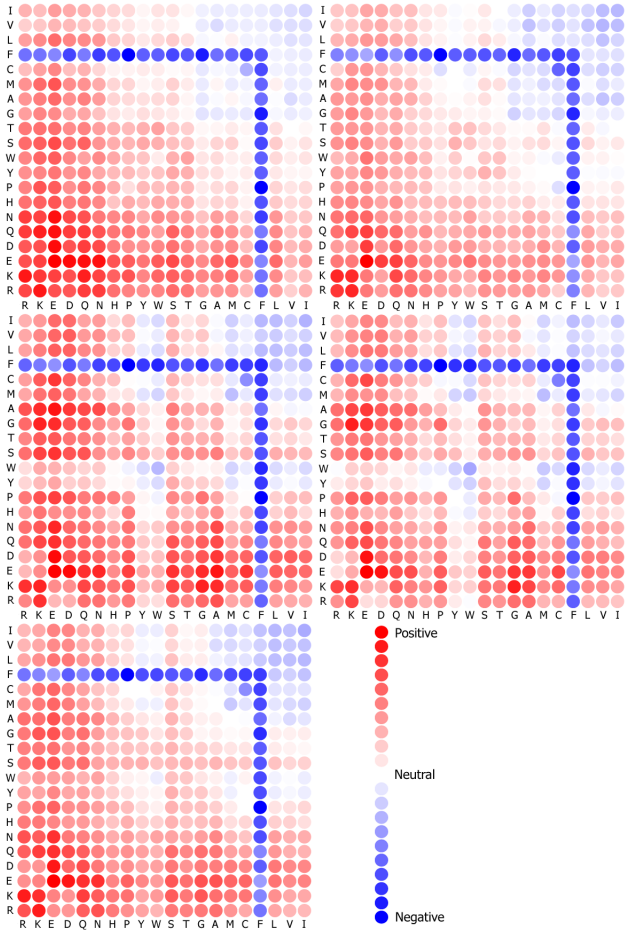

Figure F.7: ***Helix-strand* contact potentials using Stride [74].** The *helix-strand* residue contact potentials for the five contact definitions using Stride for SSE assignment. The color spectrum goes from the most negative energy (blue) over neutral (white) to the most positive energy value (red). Top left: *ca*, top right: *cb*, middle left: *all*, middle right: *vdW*, bottom left: *vor*.

Appendix G

Additional Tables

Table G.1: **Amino acids.** The 20 native amino acids in 1-letter and 3-letter notation.

Amino acid	1-letter	3-letter
Alanine	A	ALA
Arginine	R	ARG
Asparagine	N	ASN
Aspartic acid	D	ASP
Cysteine	C	CYS
Glutamine	Q	GLN
Glutamic acid	E	GLU
Glycine	G	GLY
Histidine	H	HIS
Isoleucine	I	ILE
Leucine	L	LEU
Lysine	K	LYS
Methionine	M	MET
Phenylalanine	F	PHE
Proline	P	PRO
Serine	S	SER
Threonine	T	THR
Tryptophan	W	TRP
Tyrosine	Y	TYR
Valine	V	VAL

Table G.2: **Van-der-Waals radii**. Data from the Structural Biology Glossary (http://www.imb-jena.de/ImgLibDoc/glossary/IMAGE_VDWR.html).

Element	Radius (Å)
H	1.20
C	1.70
N	1.55
O	1.52
F	1.47
P	1.80
S	1.80
Cl	1.89

Table G.3: **Linear notations for common structural motifs.** The first column gives the motif's name, the second column the linear notation according [30], the third the graph type. All linear notations are defined in the RED notation type. For every search we used also the symmetrical variant of the notation, i.e., when searching for '-1a,-1a,3a' we additionally searched also with '1a,1a,-3a'.

Motif	Search strings	Graph type
Four helix bundle	1a, 1a, 1a	Alpha
	1p, 1a, 1p	
Greek key	-1a,-1a,3a	Beta
	3a,-1a,-1a	
Jelly roll	3a,-1a,-1a, 3a,-5a,-1a, 7a	Beta
	7a,-5a, 3a,-1a,-1a, 3a,-5a	
	7a,-1a,-5a, 3a,-1a,-1a, 3a	
	1a, 5a,-3a, 1a, 1a,-3a, 5a	
Immunoglobulin fold	1a, 3a, -1a, -1a, 3a, 1a	Beta
Rossmann Fold	3p, -1p, -1p	Beta
	-1p,-1p,3p,1p,1p	
	-1p,-1p,-3p,1p,1p	
Beta Barrel	1a,1a,1a,1a,1a,1a,1a,-7a	Beta
	7a,-1a,-1a,-1a,-1a,-1a,-1a,-1a	
TIM barrel	1p,1p,1p,1p,1p,1p,1p,1p	Beta
Ubiquitin roll	-1ae, 3pe, -1ae	Alpha-Beta
	3ae, 1ae, -3pe, 1ae	
	-e,-1ae,5pe,-2ae,1ae	
	-1ae,4pe,-2ae,1ae	
	-1me,-1ae,4pe,-1ae	
	-1ae,-1ae,4pe,-1ae	
	1me,3ae,-1ae	

Table G.4: **Objective function parameters.** Parameters of the objective function (Equation 5.11) used for the genetic algorithm. GP: gap penalty, N_{gap}: number of ignored SSE in *source* structure, SB: SSE connectivity parameter

parameter	value
wC	0.6
wO	0.4
GP	0.11 * Ngap
SB	can be set by user, default SB = 0
ϵ	10^{-9}