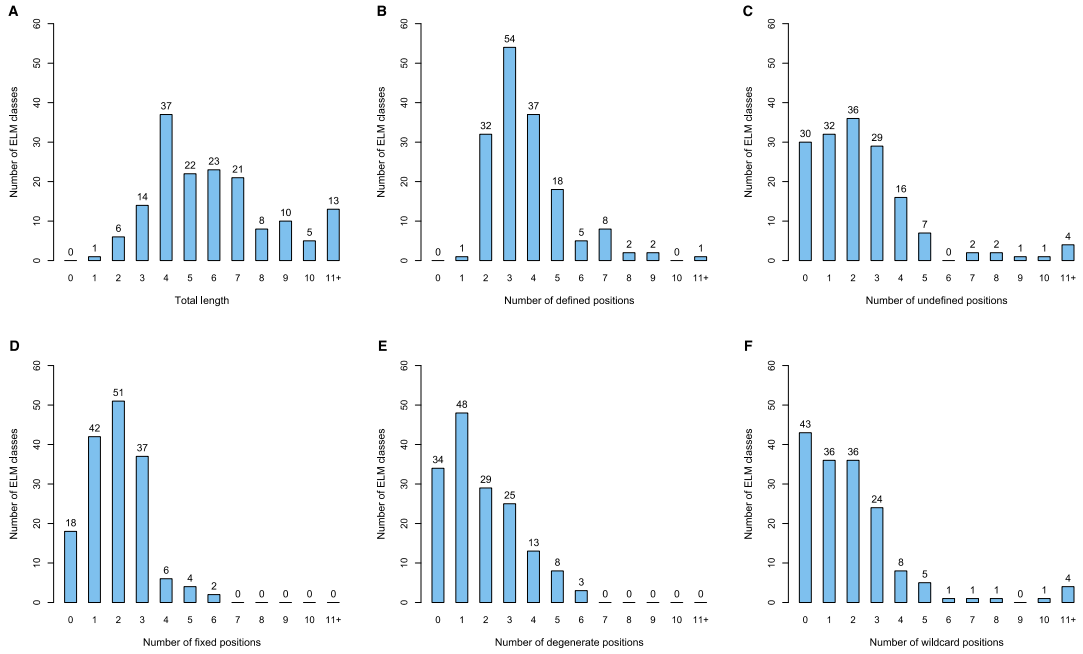
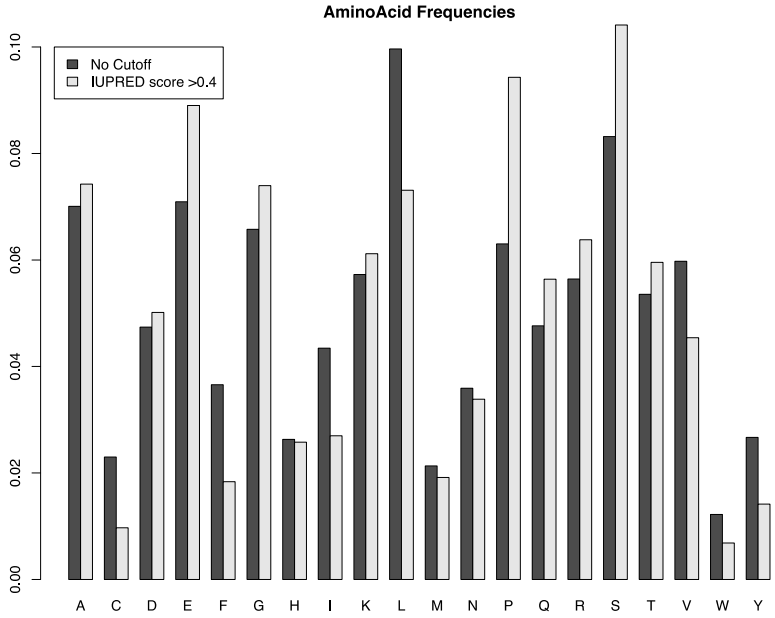
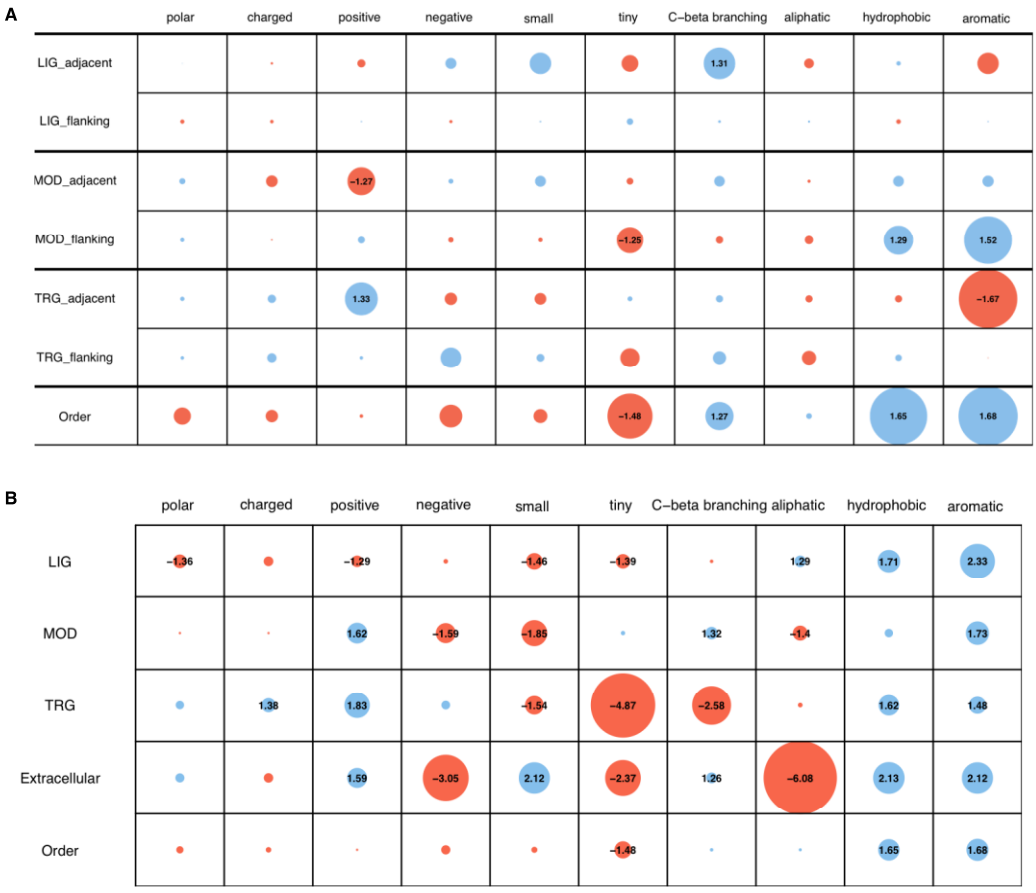
Supplementary Figures



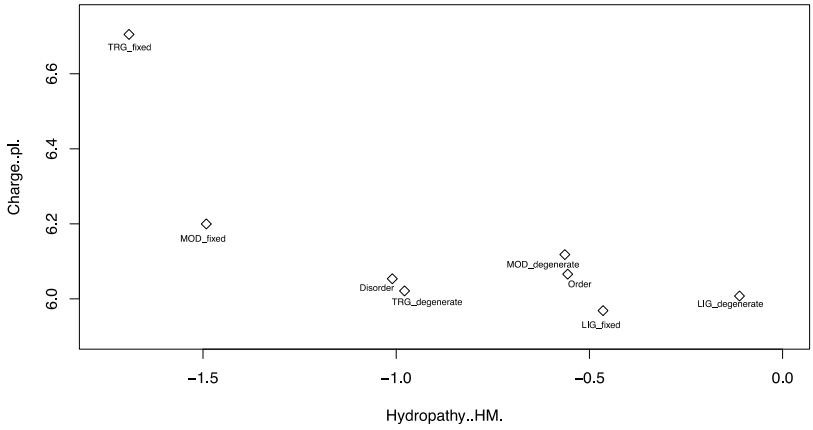
Supplementary Figure 1. The distribution of the number of different types of positions, categorized based on the degree of degeneracy, over all annotated ELM classes. A: Total length; B: Defined positions; C: Undefined positions (wildcard + prohibited); D: Fixed positions; E: Degenerate positions; F: Wildcard positions. To count the total number of wildcard positions in a RegEx, RegExs with a variable length gap were expanded to produce all possible derived RegExs with a fixed number of wildcard positions, the average of which was set as the total number of wildcard positions for that motif (e.g. the LIG_CYCLIN_1 class with the RegEx [RK].L.{0,1} [FYLIVMP] expands into [RK].L[FYLIVMP] and [RK].L.[FYLIVMP], and has a total of (1+2)/2=1.5 wildcard positions).



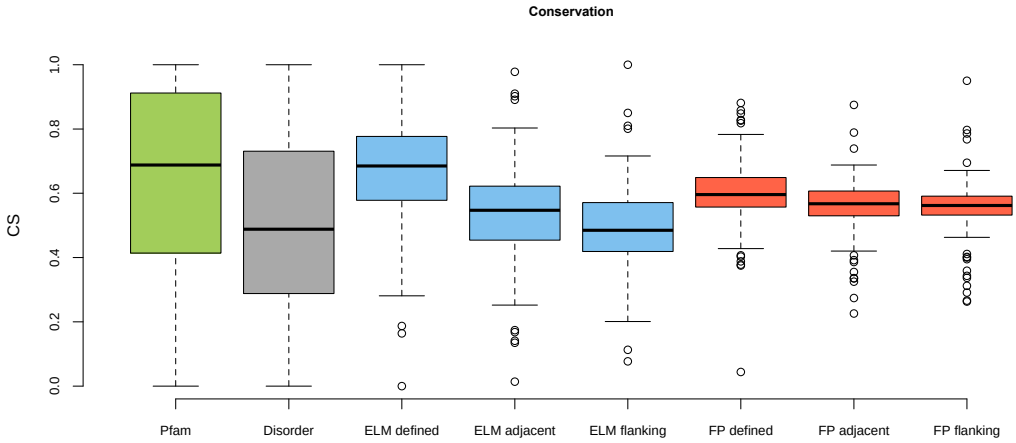
Supplementary Figure 2. Amino acid frequencies derived from 20,225 human sequences (excluding isoforms) from Uniprot taking into account all amino acids (dark grey) or only those that show an IUPRED score above or equal to thethreshold of 0.4 (light grey) (see also supplementary Table 7).



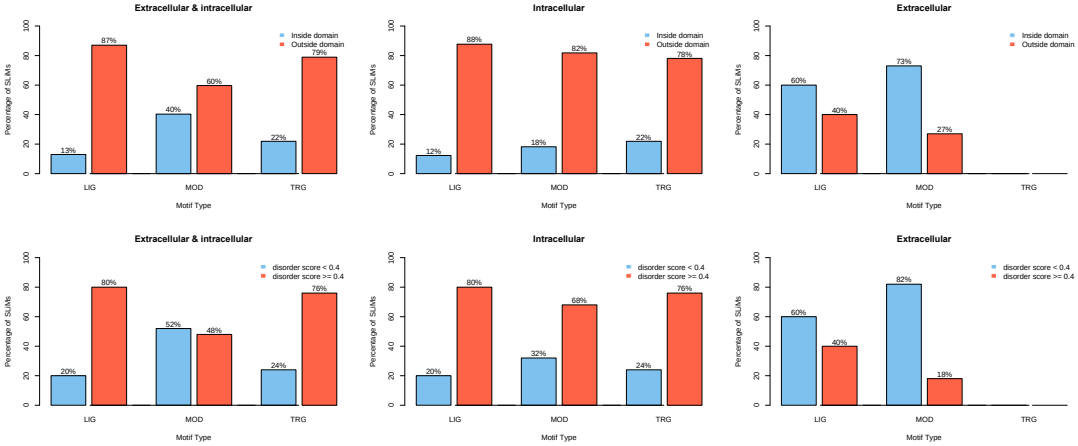
Supplementary Figure 3. (A) Amino acid preferences (grouped by physicochemical properties) of predicted ordered regions (IUPred < 0.4) and the adjacent and flanking regions of ELMs split by motif type (LIG, MOD, TRG). (B) Amino acid preferences (grouped by physicochemical properties) of predicted ordered regions (IUPred < 0.4), extracellular ELM instances and the intracellular ELM instances split by motif type (LIG, MOD, TRG). Circle sizes are proportional to fold change preference compared to disordered regions in general (IUPred >= 0.4). Blue circles indicate amino acids of a given physicochemical property are over-represented compared to disordered regions whilst the red circles indicate depletion of those amino acids. Scores within the circles denote fold change compared to the amino acid preferences of disordered regions.



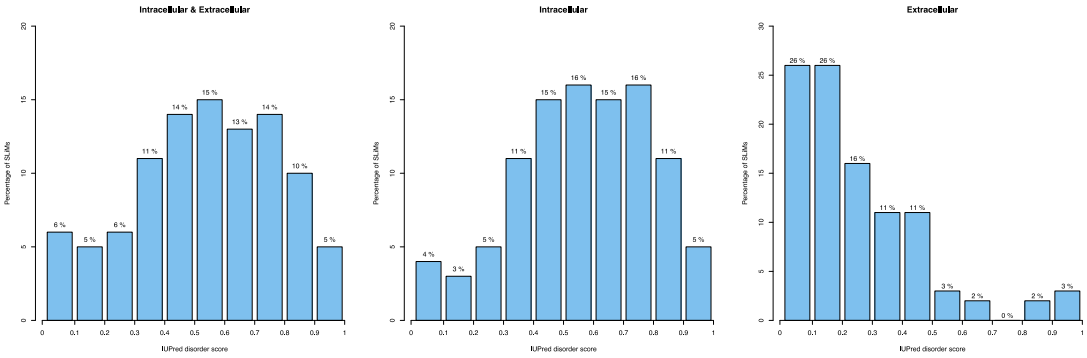
Supplementary Figure 4. A plot comparing mean charge (pI) to mean hydropathy (HM) for the defined positions within the different ELM classes, ordered and disordered regions.



Supplementary Figure 5. Boxplot of CS scores of residues annotated as within Pfam domains (green), CS scores of residues predicted to be within disordered regions (IUPred scores >= 0.4) (grey), mean CS scores for annotated ELM instances, split into defined, adjacent and flanking residues (blue) and mean CS scores for FP instances, split into defined, adjacent and flanking residues (red).



Supplementary Figure 6. (A) Classification of ELM instances split by motif type, and their presence in intracellular or extracellular portions of the protein, according to their position relative to PFAM domains. **(B)** Classification of ELM instances split by motif type, and their presence in intracellular or extracellular portions of the protein, according to mean IUPred score.



Supplementary Figure 7. Binned distribution of mean IUPred disorder scores for defined residues of 1204 annotated motifs, split by their presence in intracellular or extracellular portions of the protein.