Figure S1. Binding preferences learned by each network in the NNAlign model trained on DR15 Ph, with and without P1 burn-in implementation (panels “B” and “A”, respectively). Each logo represents a binding core learned by the network architecture out of a total of fifty (corresponding to ten seeds, and five different hidden neuron values). Logo plots were constructed from the predicted cores in the top 1% scoring predictions of 1 million random natural peptides.
Figure S2. Pearson correlation coefficient (PCC) heatmaps between single EL and BA dataset core motifs. To calculate the correlation between sequence motifs, we extracted the amino acid frequencies from the motifs displayed in figure 2, and next performed a bootstrapped correlation analysis comparing the amino acid frequency values at the four anchor positions (P1, P4, P6 and P9) of the binding core between all pairs of motifs. Heatmaps report the bootstrapped PCC values; standard deviations are reported in Table S1.

Table S1. Average Pearson correlation coefficient and standard deviations for Figure S2.
**Figure S3.** Binding preferences learned by the combined NNAlign models, trained on both Binding Affinity (BA) and Eluted Ligands (EL) data. In the top row, motifs for the DRB1*01:01 allele are shown, with overlined logoplots (right) corresponding to EL output neuron of the combined model, and the non-overlined logo (left) corresponding to the BA output neuron. Similarly, binding motifs for DRB1*15:01 are displayed in the bottom row, with overlined logos (right) also indicating the EL output neuron preference, and the non-overlined logoplots (left) indicating the BA output neuron predilection. Logos were constructed from the predicted binding cores in the top 1% scoring predictions of 900,000 random natural peptides.
Figure S4. Processing signals located at N and C terminal regions for all six datasets employed in this work. For each terminus region in every dataset, we first filtered out all ligands with PFR length lower than 3. Then, we constructed the logos by selecting the closest three PFR and context residues neighboring the N and C terminus. For additional details on this processing signal construction, please refer to Figure 5.
Figure S5. Heatmap representation of the processing signal’s Position-Specific Scoring Matrix (PSSM), for all six datasets employed in this work. For each dataset, the closest three PFR and Context residues neighboring the N and C terminus are shown. Color code correspond to the information content present in each position of the PSSM, for upstream and downstream. Colorbar values indicate the power of two to which information is increased or decreased (i.e. a value of -3 indicates a 23 fold position-specific information depletion, meaning eight times more probability to find such particular depletion than the natural background frequency). For more information on how this processing signal is constructed, refer to Figure 5.
Figure S6. Correlation between processing signals found in all six datasets employed in this work, for Upstream and Downstream regions, removing the MS-derived Cysteine enrichment artifact. Each matrix entry displays the Pearson Correlation Coefficient (PCC) value of two datasets under study. A PCC value of one corresponds to a maximum correlation, while a PCC value of zero means no correlation at all. Processing signals used in this figure were generated as explained in Figure 5.
Figure S7. Processing signal learned by the NNAlign model trained on DR15 Pm dataset with context information. For each logo, three amino acids of the PFR and context residues neighbouring the N and C terminus are shown. Logos were constructed from the predicted binding cores in the top 1% scoring predictions of 1 million random natural peptides.