Supporting Information

King et al. 10.1073/pnas.1321126111

SI Methods

Epitope Support Vector Machine Construction Details. One support vector machine (SVM) model was trained for each major histocompatibility complex (MHC) using publicly available peptide–MHC binding constants for 29 human and mouse MHC alleles. 15mer Peptide sequences were first aligned and then encoded into numerical feature vectors. Amino acid sequence encoding relied on two sources: BLOSUM62 amino acid transition probabilities and position-specific scoring matrix (PSSM) values derived from the initial sequence alignment. After encoded, support vector regression (nu-SVR) models were trained using libSVM to recapitulate log-transformed IC_{50} values by minimizing mean-squared error in fivefold cross-validation tests.

Alignment of SVM Training Data Sequences. MHC peptide-binding training data for 29 MHC alleles were downloaded from the Immune Epitope Database, non-15mers were discarded, and sequences were split into binder and nonbinder groups for each allele with a 1,000-nM cutoff. Peptides in each group were aligned using NetMHCII version 2.2 to find the highest scoring peptide core frame in each 15mer. Aligned 15mers were then extracted by including the 9mer core and three residues upstream and downstream, substituting X for gapped termini positions.

Calculation of MHC Allele-Specific PSSMs. For each allele, a position–frequency matrix was calculated by counting residue frequencies from all aligned 9mer cores from peptides with $IC_{50} < 1{,}000$ nM. Values less than 0.001 in the matrix were given a pseudovalue of 0.001, and columns were renormalized. PSSM values were calculated as the log odds ratio of each amino acid position frequency to the baseline frequency of that amino acid in the host organism.

Peptide Sequence SVM Encoding. Each 15mer sequence was encoded as a 240-element feature vector for SVM training as follows. Each amino acid of the 9mer core sequence is represented by a 21-element vector from the corresponding row of the probability-transformed BLOSUM62 matrix. Additionally, the upstream and downstream 3mers, the peptide-flanking residues, were similarly encoded as a weighted average over three positions with N-terminal weights of (1/6, 2/6, 3/6) and C-terminal weights of (3/6, 2/6, 1/6). Another nine features are added by including the score for each 9mer core position from the PSSM described above.

SVM Training. SVM models were created using libSVM. A regression model was trained (nu-SVR) to recapitulate log-transformed IC_{50} values from the feature vectors described above. IC_{50} values were transformed into scores according the same manner as in the work by Nielsen et al. (1), where each score $S = 1 - \log(IC_{50})/\log(50{,}000)$, such that the strongest binder receives a score of 1.0 and the weakest binder receives a score of 0.0. libSVM models were generated using the radial basis function kernel with the shrinking heuristic, and c, f parameters were chosen for each MHC allele model to minimize the average mean-squared error in fivefold cross-validation tests.

Deimmunization Design Simulation Details. ***Rosetta greedy optimization design.*** All design simulations were implemented using Rosetta Scripts (2). Greedy sequence design and rotamer optimization were carried out using the Rosetta greedy descent optimization algorithm as previously described (3). First, every amino acid point mutant and rotamer state at every position is sampled independently, and after rotamer optimization and gradient minimization of all neighbor side chains within an 8-Å sphere, the total energy is stored. After every position's point mutants have been evaluated, substitutions at each position are sorted by energy, and positions are rank-ordered by the value of the optimal substitution at each position. Substitutions are combined by first attempting placement of the optimal substitution at the optimal position, evaluating the total energy, and accepting if the total score improves. The substitution at the second ranked position is then attempted and so on until substitutions have been attempted. This approach converges reliably to identical solutions, although multiple diverse solutions can be generated by optionally attempting combination of near-optimal substitutions at each position and only considering substitutions with scores that remain within a certain threshold from the position's optimal substitution.

Large-scale design benchmarking. The crystal structures of eight proteins isolated from human pathogens, all containing known T-cell epitopes, were downloaded from the Protein Data Bank. Target protein sequences were scanned for MHC-binding sequences using Rosetta as described above. Each sequence was scanned for eight human leukocyte antigen DR beta (HLA-DRB) alleles (HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*11:01, HLA-DRB1*13:02, and HLA-DRB1*15:01). Epitopes were identified as those with a log-averaged predicted IC_{50} less than or equal to 1,500 nM. The predicted 15mer cores of all these epitopes were targeted for design. For the SVM score term, design simulations were carried out with varying score weights using the greedy optimization scheme described above. For host genome 9mer score term, three design simulations were carried out using Monte Carlo with 150 steps per designable position. Human 9mer content as a fraction of total epitopes subject to design and Rosetta energy values were averaged from these simulations. Final predicted epitope count was calculated as those with a log-averaged predicted IC_{50} less than or equal to 500 nM.

L-asparaginase II design simulations. The crystal structure of L-asparaginase II [Protein Data Bank (PDB) ID code 1NNS] was downloaded from the Protein Data Bank as the homotetrameric biological assembly. Design positions were restricted to those 9mer epitope regions identified by Cantor et al. (4). The design simulation was carried out with greedy sequence optimization as mentioned above, with a subsequence SVM score weight of 1.0. Epitopes were designed using the SVM for HLA-DRB1*04:01. [L-asparaginase simulations only used one allele predictor to better match the experiment in the work by Cantor et al. (4). This single allele has lower average binding affinity and thus, higher average scores than the eight-allele set used in the erythropoietin simulations, and therefore, its weight was decreased to compensate.] Predicted IC_{50} values are reported as the strongest binding epitope that overlaps the design target 9mer. Rank is reported as the highest-ranking epitope frame that encompasses all design positions. Rosetta energies for epitope regions are calculated by summing intra- and interresidue energies over the target segment.

Erythropoietin design simulations. The crystal structure of erythropoietin complexed to the binding domain of the erythropoietin receptor (PDB ID code 1EER) was downloaded from the Protein Data Bank. Design positions were restricted to those chosen by Tangri et al. (5) for mutation (residues 102, 103, 104, 107, 141, 143, 144, 146, and 147) using a subsequence SVM score weight of 3.5. Epitopes were redesigned using SVMs corresponding to eight HLA-DR alleles (HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*11:01, HLA-DRB1*13:02, and HLA-DRB1*15:01).

Predicted IC_{50} values are reported as the strongest binding epitope that overlaps the design target 9mer. Rank is reported as the highest-ranking epitope frame that encompasses all design positions. Predicted allele binders were calculated using an IC_{50} cutoff of 500 nM.

Superfolder GFP design simulations. Superfolder GFP (sfGFP) design simulations used the available sfGFP crystal structure (PDB ID code 2B3P). Eight designs were calculated using Rosetta for predicting both epitopes and the mutations' effects on MHC binding. To generate multiple design sequences, candidate mutations at each design position included all amino acids within 1.5 Rosetta energy units of the lowest-energy mutation. Candidate mutations at each position were chosen randomly during the greedy optimization design stage as described above.

Exotoxin A design simulations. Exotoxin A design simulations used the available exotoxin-eEF2 cocrystal structure (PDB ID code 1ZM4). Epitopes were redesigned using SVMs corresponding to 14 HLA alleles (HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB3*01:01, HLA-DRB4*01:01, HLA-DRB5*01:01, HLA-DQA1*05:01-DQB1*03:01, and HLA-DQA1*03:01-DQB1*03:02). Three designs were calculated using Rosetta for predicting the mutations' effects on MHC binding of the known residues 466–480 and 547–564 epitope region. Candidate mutations at each position were chosen randomly during the greedy optimization design stage as described above.

sfGFP Deimmunization Design Sequences Alignment.

```
sfgfp SKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v2.2 SKGEELFKGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v1.2 SKGEELFLGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v4.2 SKGEELFLGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v4.1 SKGEELFLGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v1.1 SKGEELFLGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v3.1 SKGEELFLGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v3.2 SKGEELFKGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v2.1 SKGEELFTGVVQILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
******* * * * ********************************************
sfgfp TTLGYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v2.2 TTLGYGVQCFSRYPDHMKRHDFFKSSMPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v1.2 TTLGYGVQCFSRYPDHMKRHDFFKSAQPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v4.2 TTLGYGVQCFSRYPDHMKRHDFFKSAMSDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v4.1 TTLGYGVQCFSRYPDHMKRHDFFKSAMSDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v1.1 TTLGYGVQCFSRYPDHMKRHDFFKSAQPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v3.1 TTLGYGVQCFSRYPDHMKRHDFMKSAMPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v3.2 TTLGYGVQCFSRYPDHMKRHDFMKSAMPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v2.1 TTLGYGVQCFSRYPDHMKRHDFFKSSMPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
********************:**:::*********************************
sfgfp RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v2.2 RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v1.2 RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v4.2 RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v4.1 RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v1.1 RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v3.1 RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v3.2 RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v2.1 RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
************************************************************
sfgfp YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVTAAGITHG
sfgfp.di.v2.2 YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVTAAGIDDG
sfgfp.di.v1.2 YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVRAAGITDQ
sfgfp.di.v4.2 YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVRAAGITDG
sfgfp.di.v4.1 YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVEAAGIQEQ
sfgfp.di.v1.1 YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVTAGGIQEE
sfgfp.di.v3.1 YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVRAAGIQEE
sfgfp.di.v3.2 YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVRAAGIQEE
sfgfp.di.v2.1 YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVRAAGIDEG
****************************************** * .** .
```


1. Nielsen M, Justesen S, Lund O, Lundegaard C, Buus S (2010) NetMHCIIpan-2.0—Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome Res* 6(1):9.
2. Fleishman SJ, et al. (2011) RosettaScripts: A scripting language interface to the Rosetta macromolecular modeling suite. *PLoS ONE* 6(6):e20161.
3. Wang P, et al. (2008) Immune epitope database: MHC-II binding dataset. Available at http://tools.iedb.org/mhcii/download/. Accessed February 14, 2013.
4. Cantor JR, et al. (2011) Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. *Proc Natl Acad Sci USA* 108(4):1272–1277.
5. Tangri S, et al. (2005) Rationally engineered therapeutic proteins with reduced immunogenicity. *J Immunol* 174(6):3187–3196.

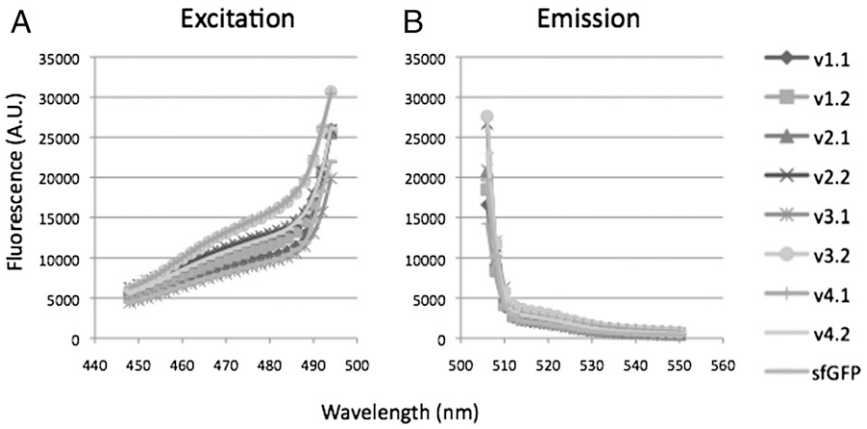

Fig. S1. Native and deimmunized sfGFP excitation and emission spectra for all eight sfGFP designs in arbitrary units (AU). (*A*) Excitation spectrum measured at 510-nm emission. (*B*) Emission spectra measured at 488-nm excitation.

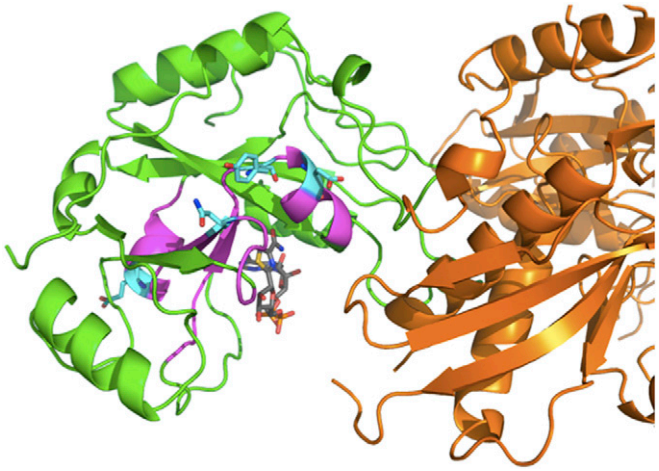

Fig. S2. Rosetta design model for immunotoxin deimmunization. Cyan, design mutations; green, endotoxin A; magenta, T-cell epitopes; orange, eEF-2.

Table S1. Rosetta deimmunization of crystal structures with known MHC epitopes

Source	PDB ID code	Known epitopes (design/native)	Predicted epitopes (design/native)	Human 9mers (design/native)	ΔRosetta energy	Sequence identity
Influenza Matrix M1	1EA3	0/7	40/154	0/0	−33.769	0.68
Tuberculosis MPT63	1LMI	1/17	21/75	0/0	−16.962	0.74
SARS Nucleocapsid	2CJR	0/3	14/72	0/0	−3.281	0.739
SARS ORF9-B	2CME	0/3	2/81	0/0	−29.654	0.436
Malaria AMA1	2Q8A	1/1	58/171	0/0	−43.004	0.778
Tuberculosis esxB	3FAV	0/13	3/28	0/0	−9.108	0.486
Arenavirus L-protein	3JSB	2/21	32/129	0/0	−21.934	0.717
HSV envelope protein D	3U82	2/13	38/165	0/0	−27.104	0.697

The numbers of known and predicted epitopes (from the eight HLA-DR allele set described in the text) after design (design) and in the original native sequence (native) are shown for each target protein along with the change in Rosetta energy (ΔRosetta energy) and the fraction of residues not changed during design (Sequence identity). Known epitope data were excluded from the simulation to provide an unbiased measure of the prediction and design. This table corresponds to Fig. 2*A*.

Table S2. Rosetta native sequence content design of crystal structures with known MHC epitopes

Source	PDB ID code	Known epitopes (design/native)	Predicted epitopes (design/native)	Human 9mers (design/native)	ΔRosetta energy	Sequence identity
Influenza Matrix M1	1EA3	3.33/7	135.33/154	6.33/0	16.58	0.76
Tuberculosis MPT63	1LMI	3.67/17	81.67/75	3.33/0	13.48	0.79
SARS Nucleocapsid	2CJR	0/3	65.67/72	4/0	2.74	0.8
SARS ORF9-B	2CME	0/3	68.33/81	15/0	10.65	0.32
Malaria AMA1	2Q8A	1/1	170.67/171	0/0	−18.59	0.98
Tuberculosis esxB	3FAV	0/13	55.33/28	10/0	11.92	0.32
Arenavirus L-protein	3JSB	1.33/21	120.33/129	19.67/0	66.21	0.45
HSV envelope protein D	3U82	8/13	168.33/165	0/0	−7.98	0.98

Values are averaged from three simulations. The numbers of known and predicted epitopes (from the eight HLA-DR allele set described in the text) after design (design) and in the original native sequence (native) are shown for each target protein along with the change in Rosetta energy (ΔRosetta energy) and the fraction of residues not changed during design (Sequence identity). Known epitope data were excluded from the simulation to provide an unbiased measure of the prediction and design. This table corresponds to Fig. 2*B*.