# iCFN: an efficient exact algorithm for multistate protein design
(Supplementary Data)

Mostafa Karimi and Yang Shen
Department of Electrical and Computer Engineering and
TEES-AgriLife Center for Bioinformatics and Genomic Systems
Engineering, Texas A&M University, College Station, 77843, USA.

## 1 Sequential reading and pruning

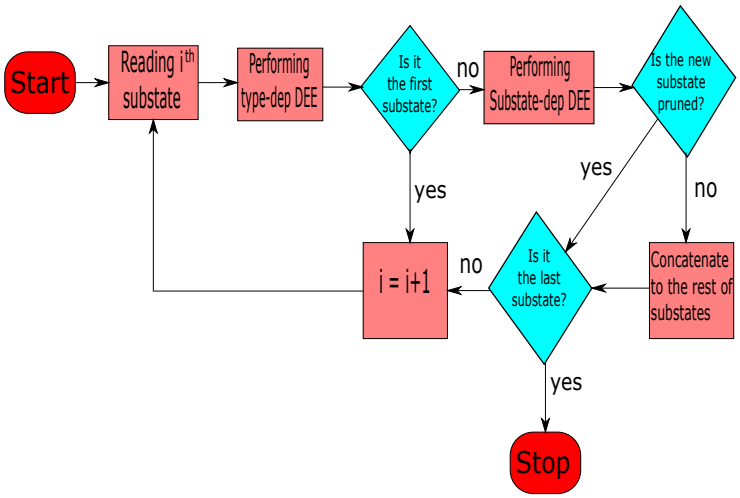


Figure S1: Flowchart of the pre-processing for iCFN: sequential reading and pruning

**Algorithm 1** Pseudocode for sequential reading and pruning.

```
for i = 1: N do
  Read_substate(i)
  Type_Dep_DEE(i,δ)
  if Is_it_first_substate() then
    Create_main_substate()
  else
    Substate_Dep_DEE(i,δ)
    if Is_new_substate_pruned() == 0 then
      Concat_to_main_substate()
    end if
  end if
end for
```

## 2 Across-substate type-dependent DEE

**Theorem 1.** *Rotamer $i_a$ of substate 1 provably pruned by rotamer $i_b$ of substate 2, is not part of the optimal solution if both substates belong to the same state (positive or negative), both rotamers are of the same amino acid type, and the following criterion holds:*

$$\begin{aligned} & c_1 + E_1(i_a) + \sum_{j,j\neq i} \min_{s_1} \left(E_1(j_{s_1}) + E_1(i_a, j_{s_1})\right) \\ & + \sum_{j>k,k\neq i,j\neq i} \min_{s_1,u_1} E_1(j_{s_1}, k_{u_1}) \\ & > c_2 + E_2(i_b) + \sum_{j,j\neq i} \max_{s_2} \left(E_2(j_{s_2}) + E_2(i_b, j_{s_2})\right) \\ & + \sum_{j>k,k\neq i,j\neq i} \max_{s_2,u_2} E_2(j_{s_2}, k_{u_2}) \end{aligned} \tag{1}$$

*Proof.* Following Eq. 1 in the main text, the energy of substate 1 (used as a subscript) with rotamer $i_a$ at residue $i$ and its upper bound can be written as:

$$\begin{aligned} & c_1 + \sum_m E_1(m_r) + \sum_{m<j} E_1(m_r, j_s) \\ & = c_1 + E_1(i_a) + \sum_{m\neq i} E_1(m_r) + \sum_{j\neq i} E_1(i_a, j_s) + \\ & \sum_{m<j,m\neq i,j\neq i} E_1(m_r, j_s) \\ & > c_1 + E_1(i_a) + \sum_{j,j\neq i} \min_s \left(E_1(j_s) + E_1(i_a, j_s)\right) + \\ & \sum_{j>m,m,j\neq i} \min_{s,r} E_1(j_s, m_r) \triangleq L_1(i_a) \end{aligned} \tag{2}$$

By doing the same for rotamer $i_b$ in substate 2:

$$
\begin{aligned}
& c_2 + \sum_{m} E_2(m_r) + \sum_{m<j} E_2(m_r, j_s) \\
& = c_2 + E_2(i_b) + \sum_{m \neq i} E_2(m_r) + \sum_{j \neq i} E_2(i_b, j_s) + \\
& \sum_{m<j, m \neq i, j \neq i} E_2(m_r, j_s) \\
& < c_2 + E_2(i_b) + \sum_{j, j \neq i} \max_{s} \big( E_2(j_s) + E_2(i_b, j_s) \big) + \\
& \sum_{j>m, m \neq i, j \neq i} \max_{s,r} E_2(j_s, m_r) \triangleq U_2(i_b)
\end{aligned}
\tag{3}
$$

Therefore, if $L_1(i_a) > U_2(i_b)$, then $i_a$ is pruned by $i_b$ and cannot be part of the global optimum.

A natural extension for the top $\delta$ kcal/mol ensemble is that rotamer $i_a$ of substate 1 is pruned by rotamer $i_b$ of substate 2 if $L_1(i_a) > U_2(i_b) + \delta$. □

# 3 Global sequence search

**Algorithm 2** Main algorithm

```
best_score = Max_Value
best_score = Global_Search_GMEC()
Global_Search_Ensemble()
```

**Algorithm 3** Global_Search_GMEC()

```
if LDS_constraint() then
  return
end if
if Is_fully_defined() then
  if Lower_bound_fully_defined() == 0 then
    Backbone_pruning()
    Seq_defined_GMEC()
    if best_score > Lowest_energy_pos - Lowest_energy_neg then
      best_score = Lowest_energy_pos - Lowest_energy_neg
    end if
  end if
else
  i = Variable_ordering()
  a = Amino_ordering()
  Assign_amino(i,a)
  if Lower_bound_Not_fully_defined() == 0 then
    Global_Search_GMEC()
  else
    Remove_amino(i,a)
    Global_Search_GMEC()
  end if
end if
```

**Algorithm 4** Seq_defined_GMEC()

```
Update_pos = 0
Lowest_energy_pos = Max_Value
for i = 1 : N do
  if Stability_condition(τ) then
    temp_best = SCP_GMEC(i)
    if temp_best < Lowest_energy_pos then
      Lowest_energy_pos = temp_best
      Update_pos = 1
    end if
  end if
end for
if Update_pos == 0 then
  return
end if
Lowest_energy_neg = Max_Value
for j = 1 : M do
  temp_best = SCP_GMEC(j)
  if temp_best < Lowest_energy_neg then
    Lowest_energy_neg = temp_best
  end if
  if best_score < Lowest_energy_pos - Lowest_energy_neg then
    return
  end if
end for
```

**Algorithm 5** Global_Search_ensemble()

```
if LDS_constraint() then
  return
end if
if Is_fully_defined() then
  if Lower_bound_fully_defined() == 0 then
    Backbone_pruning()
    Seq_defined_GMEC()
    Seq_defined_ensemble()
    if best_score + ε > Lowest_energy_pos - Lowest_energy_neg then
      print conformation for this sequence
    end if
  end if
else
  i = Variable_ordering()
  a = Amino_ordering()
  Assign_amino(i,a)
  if Lower_bound_Not_fully_defined() == 0 then
    Global_Search_ensemble()
  else
    Remove_amino(i,a)
    Global_Search_ensemble()
  end if
end if
```

**Algorithm 6** Seq_defined_ensemble()

```
for i = 1 : N do
  if backbone_pruned(i) == 0 then
    SCP_ensemble(i)
  end if
end for
for j = 1 : M do
  if backbone_pruned(j) == 0 then
    SCP_ensemble(j)
  end if
end for
```

# 4 Bounding in global sequence search

**Theorem 2.** *For any sequence space $S$, a lower bound of the objective function for multistate protein design with substate ensembles (Formulation in Eq. 5 of main text) is given by*

$$\min_{(k,l)\in\mathcal{P}\times\mathcal{Q}}\Big(\Delta c_{kl}+\sum_i \min_{\mathbf{a}\in S(i)}\min_{(r,r')}\big(\Delta E_{kl}(i_{r,r'})+ \sum_{j>i}\min_{\mathbf{a}'\in S(j)}\min_{(s,s')}\Delta E_{kl}(i_{r,r'},j_{s,s'})\big)\Big),\ \textit{where} \tag{4}$$

$$\begin{aligned}\Delta c_{kl} &= c_k^+ - c_l^-,\\ \Delta E_{kl}(i_{r,r'}) &= E_k^+(i_r) - E_l^-(i_{r'}),\\ \Delta E_{kl}(i_{r,r'},j_{s,s'}) &= E_k^+(i_r,j_s) - E_l^-(i_{r'},j_{s'}),\end{aligned} \tag{5}$$

*i.e., differences in constant, singleton, and pairwise energies between a positive substate $k$ (position $i$ and $j$ taking rotamer $r$ and $s$) and a negative substate $l$ (position $i$ and $j$ taking rotamer $r'$ and $s'$).*

*Proof.* For an arbitrary sequence $\mathbf{a}$ in the space $S$, its rotamer vector $\mathbf{r}$ is in the space of $\mathcal{R}_k(\mathbf{a})$ for substate $k$. The highest specificity is thus

$$\begin{aligned}
&\min_{\mathbf{a}\in S}\big(\min_{k\in\mathcal{P}}\min_{\mathbf{r}\in\mathcal{R}_k(\mathbf{a})}E_k^+(\mathbf{r})-\min_{l\in\mathcal{Q}}\min_{\mathbf{r}'\in\mathcal{R}_l(\mathbf{a})}E_l^-(\mathbf{r}')\big)\\
&\geqslant \min_{\mathbf{a}\in S}\min_{(k,l)\in\mathcal{P}\times\mathcal{Q}}\min_{(\mathbf{r},\mathbf{r}')\in\mathcal{R}_k(\mathbf{a})\times\mathcal{R}_l(\mathbf{a})}\big(E_k^+(\mathbf{r})-E_l^-(\mathbf{r}')\big)\\
&\geqslant \min_{\mathbf{a}\in S}\min_{(k,l)\in\mathcal{P}\times\mathcal{Q}}\Big(\Delta c_{kl}+\min_{(r,r')}\big(\sum_i\Delta E_{kl}(i_{r,r'})\\
&+\sum_{j>i}\Delta E_{kl}(i_{r,r'},j_{s,s'})\big)\Big)\\
&\geqslant \min_{\mathbf{a}\in S}\min_{(k,l)\in\mathcal{P}\times\mathcal{Q}}\Big(\Delta c_{kl}+\sum_i\min_{(r,r')}\big(\Delta E_{kl}(i_{r,r'})\\
&+\sum_{j>i}\min_{(s,s')}\Delta E_{kl}(i_{r,r'},j_{s,s'})\big)\Big)\\
&= \min_{(k,l)\in\mathcal{P}\times\mathcal{Q}}\Big(\Delta c_{kl}+\min_{\mathbf{a}}\sum_i\min_{(r,r')}\big(\Delta E_{kl}(i_{r,r'})\\
&+\sum_{j>i}\min_{(s,s')}\Delta E_{kl}(i_{r,r'},j_{s,s'})\big)\Big)\\
&\geqslant \min_{(k,l)\in\mathcal{P}\times\mathcal{Q}}\Big(\Delta c_{kl}+\sum_i\min_{\mathbf{a}\in S(i)}\min_{(r,r')}\big(\Delta E_{kl}(i_{r,r'})\\
&+\sum_{j>i}\min_{\mathbf{a}'\in S(j)}\min_{(s,s')}\Delta E_{kl}(i_{r,r'},j_{s,s'})\big)\Big)
\end{aligned} \tag{6}$$

□

The complexity of evaluating the lower bound for undefined sequences is given as follows:

**Theorem 3.** *The lower bound in Theorem 2 can be computed in $O\big((nRa)^2 r\big)$, where $n$ is the number of positions, $R$ the average number of rotamers per position, $a$ the average number of substates per state, and $r$ the average number of rotamers per amino acid.*

*Proof.* we prove the complexity by starting with the most inner minimization:

$$\begin{aligned}
&\min_{(s,s')} \Delta E_{kl}(i_{r,r'}, j_{s,s'}) \\
&= \min_{(s,s')} \Big( E_k^+(i_r, j_s) - E_l^-(i_{r'}, j_{s'}) \Big) \\
&= \min_{s} E_k^+(i_r, j_s) + \min_{s'} \Big( - E_l^-(i_{r'}, j_{s'}) \Big) \\
&= \min_{s} E_k^+(i_r, j_s) - \max_{s'} E_l^-(i_{r'}, j_{s'})
\end{aligned} \tag{7}$$

So, we can calculate it in $O(r)$. Since the number of amino acids is known, then

$$\min_{a' \in S(j)} \min_{(s,s')} \Delta E_{kl}(i_{r,r'}, j_{s,s'}) \tag{8}$$

will be again $O(r)$, so by summing over positions it will be $O(nr)$. For calculating:

$$\min_{a \in S(i)} \min_{(r_k, r_l)} \Big( \Delta E_{kl}(i_{r,r'}) + \sum_{j>i} \min_{a' \in S(j)} \min_{(s,s')} \Delta E_{kl}(i_{r,r'}, j_{s,s'}) \Big) \tag{9}$$

similar to previous version, we can compute it in $O(nR^2 r)$ and summing over all positions it will be $O(n^2 R^2 r)$. Finally, since we are calculating (9) for all $a^2$ pairs of substates across the two states, complexity will be $O(a^2 n^2 R^2 r)$. □

When a sequence is specifically defined during search, we have derived a more powerful lower bound as follows:

**Theorem 4.** *For any defined sequence* $\mathbf{s}$ *(*$S = \{\mathbf{s}\}$*), a lower bound can be computed by*

$$\min_{k \in \mathcal{P}} L_k^+(\mathbf{s}) - \min_{l \in \mathcal{Q}} U_l^-(\mathbf{s}) \tag{10}$$

*in which* $L_k^+(\mathbf{s})$ *is lower bound on all rotamer conformation for sequence* $\mathbf{s}$ *and* $k^{th}$ *substate in positive design and* $U_l^-(\mathbf{s})$ *is Upper bound on all rotamer conformation for sequence* $\mathbf{s}$ *and* $l^{th}$ *substate in negative design.*

*Proof.* When sequence is fully defined, a lower bound can be derived by:

$$\begin{aligned} &\min_{k \in \mathcal{P}} \min_{\mathbf{r_k} \in R_k(\mathbf{s})} E_k^+(\mathbf{r_k}) - \min_{l \in \mathcal{Q}} \min_{\mathbf{r_l} \in R_l(\mathbf{s})} E_l^-(\mathbf{r_l}) \\ &\geqslant \min_{k \in \mathcal{P}} L_k^+(\mathbf{s}) - \min_{l \in \mathcal{Q}} U_l^-(\mathbf{s}), \end{aligned} \tag{11}$$

in which $L_k^+(\mathbf{s})$ can be any lower bound from single-state protein design (we use existential directed arc consistency a.k.a. EDAC) and $U_l^-(\mathbf{s})$ can be any upper bound from single state protein design (we use limited discrepancy search a.k.a. LDS). □

# 5 Results

## 5.1 TCR Design: Efficiency

### Additional contributions to performance improvement

Beyond substate pruning enabled by interconnected CFNs, three more improvements we made also contribute to the numerical efficiency. The first affects both reduced iCFN and iCFN: (1) variables (positions) are ordered based on the number of rotamers divided by the median of singleton energies only, which affects the nodes and leaves (and ultimately sequences) visited during tree search. The rest two are both for calculating lower bounds of undefined sequences thus only affect iCFN: (2) a lookup table storing intermediate min/max values for each substate reduces calculations in the order of the number of substates, and (3) an upper bounding when minimizing differences over substate pairs can be accomplished with any feasible solution.

To dissect the contributions of these three additional contributions, we start with none and incrementally introduce them into versions 0 (none), 0.1, and 0.2, where the latter two only apply to iCFN. The latest version in the main text is regarded version 1. By comparing them in the supplemental Tables S1 and S2, we find that the change of position ordering may lead to slightly increased number of nodes expanded or leaves visited but saves run time due to much less time spent on each node for bound estimation. In addition, the lookup tables are created only once and used multiple times in search, which especially speed up large designs (twice for double designs).

| | Reduced iCFN | | iCFN | | | |
|---|---|---|---|---|---|---|
| Position(s) | v0 | v1 | v0 | v0.1 | v0.2 | v1 |
| 26 | 4.06 | 1.46 | 4 | 0.6 | 0.62 | 0.56 |
| 28 | 291 | 24.5 | 289 | 5.88 | 7.32 | 6.29 |
| 98 | 32 | 9.98 | 26 | 3.26 | 4.46 | 3.38 |
| 100 | 33 | 19.85 | 30 | 4.18 | 4 | 4.44 |
| 26,28 | 16 152 | 1335.95 | 12 774 | 676.61 | 248.26 | 228 |
| 26,98 | 3540 | 809.18 | 2627 | 283.63 | 172.89 | 182.10 |
| 26,100 | 3799 | 1510.03 | 3008 | 686.67 | 330.68 | 303.64 |
| 28,98 | 27 252 | 3707.04 | 21 809 | 1522.54 | 717.25 | 745.84 |
| 28,100 | 20 521 | 5603.60 | 16 605 | 1785.23 | 738.04 | 796.96 |
| 98,100 | 19 808 | 4384.48 | 13 726 | 1257 | 534.42 | 526.97 |

Table S1: Comparing run time (in seconds) between different versions of reduced iCFN and iCFN for the best global optimum conformation in multi-state design problems with ensemble of substates per state for TCR.

| | Reduced iCFN | | iCFN | | | |
|---|---|---|---|---|---|---|
| Position(s) | v0 | v1 | v0 | v0.1 | v0.2 | v1 |
| 26 | 6654 | 66.69 | 3700 | 22.94 | 31.48 | 21.86 |
| 28 | 4283 | 114.22 | 2222 | 22.32 | 29.3 | 23.55 |
| 98 | 10 612 | 103.29 | 6318 | 40.81 | 55.72 | 43.35 |
| 100 | 2109 | 154.51 | 1102 | 21.20 | 20.05 | 23.74 |
| 26,28 | 384 997 | 7454.93 | 205 656 | 16 120 | 1705.68 | 1063.89 |
| 26,98 | — | 15 449.04 | — | 27 596 | 4666.22 | 3872.32 |
| 26,100 | 502 803 | 19 780.68 | 265 185 | 12 162 | 2689.77 | 2226.52 |
| 28,98 | — | 23 378.51 | 487 360 | 20 629 | 3561.14 | 2810.31 |
| 28,100 | 347 872 | 24 631.34 | 177 949 | 11 452 | 2956.08 | 2359.10 |
| 98,100 | — | 17 303.91 | 323 104 | 11 781 | 2700.47 | 2056.47 |

Table S2: Comparing run time (in seconds) between different versions of reduced iCFN and iCFN for the best ensemble conformations in multi-state design problems with ensemble of substates per state for TCR. ("—" indicates an out-of-time error under the 7-day limit.)

## 5.2 TCR Design: Accuracy

Results for Multi-substates are shown in the following table.

| Mutation | AAG ($\Delta\Delta G$) | ELA ($\Delta\Delta G$) | specificity ($\Delta\Delta\Delta G$) |
|---|---|---|---|
| $\alpha D26Y$ | −1.03 | 21.90 | −22.93 |
| αD26F | −1.74 | 9.04 | −10.78 |
| αD26A | −2.56 | 4.41 | −6.97 |
| αD26N | −0.34 | 5.49 | −5.83 |
| αD26P | −3.38 | 2.27 | −5.65 |
| αD26K | −0.39 | 5.14 | −5.53 |
| αD26T | −2.30 | 2.33 | −4.63 |
| αD26C | −2.21 | 2.00 | −4.21 |
| αD26V | −1.53 | 1.48 | −3.01 |
| $\alpha D26W$ | −4.13 | −1.48 | −2.65 |
| αD26M | −3.32 | −0.79 | −2.53 |
| αD26H | −0.38 | 1.96 | −2.34 |
| $\alpha G28L$ | −2.66 | 38.08 | −40.74 |
| αG28E | −5.91 | 22.72 | −28.63 |
| αG28D | −2.39 | 16.00 | −18.39 |
| αG28T | 3.13 | 21.38 | −18.25 |
| $\alpha G28I$ | −5.55 | 8.83 | −14.38 |
| αG28M | −4.32 | 9.52 | −13.84 |
| αG28R | 4.18 | 16.94 | −12.76 |
| αG28V | 0.75 | 11.16 | −10.41 |
| αG28C | −0.04 | 9.45 | −9.49 |
| $\alpha G28Y$ | 4.80 | 13.96 | −9.16 |
| αG28K | −0.20 | 8.94 | −9.14 |
| αG28F | −3.62 | 3.18 | −6.80 |

Table S3: TCR designs considering an ensemble of positive or negative substate (flexible backbone conformation here). Reported for each design is the calculated relative binding affinities $\Delta\Delta G$ (in Kcal/mol) compared to the wild type (WT) for the AAG peptide (MART-1 nonameric epitope) and the ELA peptide (MART-1 decameric epitope), respectively, as well as their difference $\Delta\Delta\Delta G$, or, specificity. Only designs predicted to significantly improve AAG-binding specificity compared to WT ($\Delta\Delta\Delta G \leqslant -2$ Kcal/mol) are reported here. Designs highlighted in red and blue were experimentally validated true or false positives according to a recent study (Pierce *et. al.* 2014) .

| Mutation | AAG ($\Delta\Delta G$) | ELA ($\Delta\Delta G$) | specificity ($\Delta\Delta\Delta G$) |
|---|---|---|---|
| $\alpha$L98K | 1.04 | 6.32 | −5.28 |
| $\alpha$L98R | 0.32 | 4.95 | −4.63 |
| $\alpha F100Y$ | 7.09 | 47.79 | −40.70 |
| $\alpha F100W$ | 15.23 | 40.39 | −25.16 |
| $\alpha$F100R | 13.76 | 21.90 | −8.14 |
| $\alpha$F100Q | 3.34 | 10.28 | −6.94 |
| $\alpha$F100M | 4.08 | 9.77 | −5.69 |
| $\alpha$F100A | 1.73 | 6.40 | −4.67 |
| $\alpha$F100I | 1.71 | 6.21 | −4.50 |
| $\alpha$F100K | 14.61 | 18.32 | −3.71 |
| $\alpha$F100S | 1.70 | 5.41 | −3.71 |
| $\alpha$F100C | 2.39 | 5.90 | −3.51 |
| $\alpha$F100L | 7.07 | 9.47 | −2.40 |
| $\alpha$F100V | 4.36 | 6.64 | −2.28 |
| $\alpha$F100E | 4.97 | 7.22 | −2.25 |

Table S3: (Continued) TCR designs considering an ensemble of positive or negative substate (flexible backbone conformation here). Reported for each design is the calculated relative binding affinities $\Delta\Delta G$ (in Kcal/mol) compared to the wild type (WT) for the AAG peptide (MART-1 nonameric epitope) and the ELA peptide (MART-1 decameric epitope), respectively, as well as their difference $\Delta\Delta\Delta G$, or, specificity. Only designs predicted to significantly improve AAG-binding specificity compared to WT ($\Delta\Delta\Delta G \leqslant -2$ Kcal/mol) are reported here. Designs highlighted in red and blue were experimentally validated true or false positives according to a recent study (Pierce *et. al.* 2014).

| Mutation | AAG ($\Delta\Delta G$) | ELA ($\Delta\Delta G$) | specificity ($\Delta\Delta\Delta G$) |
|---|---|---|---|
| $\alpha$D26N | $-1.98$ | 3.76 | $-5.74$ |
| *$\alpha G28L$* | $-1.65$ | 4.69 | $-6.34$ |
| $\alpha$G28D | $-3.64$ | $-1.60$ | $-2.04$ |
| $\alpha$L98V | 2.54 | 1.32 | 1.22 |
| $\alpha$L98D | 1.92 | 0.62 | 1.3 |
| $\alpha$L98I | 2.20 | 0.87 | 1.33 |
| $\alpha$L98E | 2.29 | 0.56 | 1.73 |
| $\alpha$L98Q | 1.88 | $-0.05$ | 1.93 |
| *$\alpha F100W$* | 39.53 | 100.71 | $-61.18$ |
| *$\alpha F100Y$* | 4.87 | 28.50 | $-23.63$ |

Table S4: TCR designs considering a single positive or negative substate (fixed backbone conformation here). Reported for each design is the calculated relative binding affinities $\Delta\Delta G$ (in Kcal/mol) compared to the wild type (WT) for the AAG peptide (MART-1 nonameric epitope) and the ELA peptide (MART-1 decameric epitope), respectively, as well as their difference $\Delta\Delta\Delta G$, or, specificity. Only designs predicted to significantly improve AAG-binding specificity compared to WT ($\Delta\Delta\Delta G \leqslant -2$ Kcal/mol) are reported here with the exception for position 98 with the top 5 lowest $\Delta\Delta\Delta G$. Designs highlighted in red and blue were experimentally validated true or false positives according to a recent study (Pierce *et. al.* 2014).

| Mutation | AAG ($\Delta\Delta G$) | ELA ($\Delta\Delta G$) |
|---|---|---|
| *$\alpha D26W$* | 4 | 8 |
| *$\alpha G28L$* | 10 | 1 |
| *$\alpha G28I$* | 10 | 3 |
| *$\alpha G28Y$* | 10 | 8 |
| *$\alpha F100Y$* | 7 | 1 |
| *$\alpha F100W$* | 10 | 2 |

Table S5: TCR designs considering an ensemble of positive or negative substates (flexible backbone conformations here). Reported are indices of various backbone conformations that were adopted in iCFN for various successful designs bound to the AAG peptide (MART-1 nonameric epitope) and the ELA peptide (MART-1 decameric epitope).