Supplementary Information

Nucleic Acid Sequence Design via Efficient Ensemble Defect Optimization

Joseph N. Zadeh[1], Brian R. Wolfe[1], Niles A. Pierce[1,2]

[1]*Department of Bioengineering, California Institute of Technology, Pasadena, CA 91125*
[2]*Department of Applied & Computational Mathematics, California Institute of Technology, Pasadena, CA 91125*
Correspondence to: niles@caltech.edu

DESIGNSEQ(s)
  $m_{\text{leafopt}} \leftarrow 0$
  $\phi, n \leftarrow$ OPTIMIZELEAF(s)
  while $n > f_{\text{stop}}|s|$ **and** $m_{\text{leafopt}} < M_{\text{leafopt}}$
    $\hat{\phi}, \hat{n} \leftarrow$ OPTIMIZELEAF(s)
    if $\hat{n} < n$
      $\phi, n \leftarrow \hat{\phi}, \hat{n}$
    $m_{\text{leafopt}} \leftarrow m_{\text{leafopt}} + 1$
  return ϕ

OPTIMIZELEAF(s)
  $m_{\text{unfavorable}} \leftarrow 0$
  $\gamma_{\text{unfavorable}} \leftarrow \emptyset$
  $\phi \leftarrow$ INITSEQ(s)
  $n \leftarrow$ ENSEMBLEDEFECT(ϕ, s)
  while $n > f_{\text{stop}}|s|$ **and** $m_{\text{unfavorable}} < M_{\text{unfavorable}}|s|$
    $\xi, \hat{\phi} \leftarrow$ UNIFORMMUTATIONSAMPLING(ϕ, s)
    if $\xi \in \gamma_{\text{unfavorable}}$
      $m_{\text{unfavorable}} \leftarrow m_{\text{unfavorable}} + 1$
    else
      $\hat{n} \leftarrow$ ENSEMBLEDEFECT($\hat{\phi}, s$)
      if $\hat{n} < n$
        $\phi, n \leftarrow \hat{\phi}, \hat{n}$
        $m_{\text{unfavorable}} \leftarrow 0$
        $\gamma_{\text{unfavorable}} \leftarrow \emptyset$
      else
        $m_{\text{unfavorable}} \leftarrow m_{\text{unfavorable}} + 1$
        $\gamma_{\text{unfavorable}} \leftarrow \gamma_{\text{unfavorable}} \bigcup \xi$
  return ϕ, n

Algorithm S1: Single-scale ensemble defect optimization with uniform mutation sampling.

DESIGNSEQ(s)
  $m_{\text{leafopt}} \leftarrow 0$
  $\phi, n \leftarrow$ OPTIMIZELEAF(s)
  while $n > f_{\text{stop}}|s|$ **and** $m_{\text{leafopt}} < M_{\text{leafopt}}$
    $\hat{\phi}, \hat{n} \leftarrow$ OPTIMIZELEAF($\hat{\phi}, s$)
    if $\hat{n} < n$
      $\phi, n \leftarrow \hat{\phi}, \hat{n}$
    $m_{\text{leafopt}} \leftarrow m_{\text{leafopt}} + 1$
  return ϕ

OPTIMIZELEAF(s)
  $m_{\text{unfavorable}} \leftarrow 0$
  $\gamma_{\text{unfavorable}} \leftarrow \emptyset$
  $\phi \leftarrow$ INITSEQ(s)
  $n \leftarrow$ ENSEMBLEDEFECT(ϕ, s)
  while $n > f_{\text{stop}}|s|$ **and** $m_{\text{unfavorable}} < M_{\text{unfavorable}}|s|$
    $\xi, \hat{\phi} \leftarrow$ WEIGHTEDMUTATIONSAMPLING($\phi, s, n_1, \ldots, n_{|s|}$)
    if $\xi \in \gamma_{\text{unfavorable}}$
      $m_{\text{unfavorable}} \leftarrow m_{\text{unfavorable}} + 1$
    else
      $\hat{n} \leftarrow$ ENSEMBLEDEFECT($\hat{\phi}, s$)
      if $\hat{n} < n$
        $\phi, n \leftarrow \hat{\phi}, \hat{n}$
        $m_{\text{unfavorable}} \leftarrow 0$
        $\gamma_{\text{unfavorable}} \leftarrow \emptyset$
      else
        $m_{\text{unfavorable}} \leftarrow m_{\text{unfavorable}} + 1$
        $\gamma_{\text{unfavorable}} \leftarrow \gamma_{\text{unfavorable}} \bigcup \xi$
  return ϕ, n

Algorithm S2: Single-scale ensemble defect optimization with defect-weighted mutation sampling.

DESIGNSEQ(ϕ, s, n, k)
  $a \leftarrow$ DEPTH(k)
  if HASCHILDREN(k)
    $m_{\text{reopt}} \leftarrow 0$
    if $n = \emptyset$
      $\phi_l \leftarrow$ DESIGNSEQ$(\emptyset, s_{l+}, \emptyset, k_l)$
      $\phi_r \leftarrow$ DESIGNSEQ$(\emptyset, s_{r+}, \emptyset, k_r)$
    else
      UPDATECHILDREN$(k, a, a-1)$
      child, $\phi \leftarrow$ UNIFORMCHILDSAMPLING(ϕ, s, n_l, n_r)
      $\phi_{\text{child}} \leftarrow$ DESIGNSEQ$(\phi_{\text{child}+}, s_{\text{child}+}, n_{\text{child}+}, k_{\text{child}})$
    $n^{k,a} \leftarrow$ ENSEMBLEDEFECT(ϕ, s)
    UPDATECHILDREN$(k, a, a+1)$
    while $n^{k,a} > \max(f_{\text{stop}}|s_l|, n^{k_l,a}_{\text{native}}) + \max(f_{\text{stop}}|s_r|, n^{k_r,a}_{\text{native}})$
        and $m_{\text{reopt}} < M_{\text{reopt}}$
      child, $\hat{\phi} \leftarrow$ UNIFORMCHILDSAMPLING$(\phi, s, n^{k,a}_l, n^{k,a}_r)$
      $\hat{\phi}_{\text{child}} \leftarrow$ DESIGNSEQ$(\phi_{\text{child}+}, s_{\text{child}+}, n^{k,a}_{\text{child}+}, k_{\text{child}})$
      $\hat{n} \leftarrow$ ENSEMBLEDEFECT$(\hat{\phi}, s)$
      if $\hat{n} < n^{k,a}$
        $\phi, n^{k,a} \leftarrow \hat{\phi}, \hat{n}$
        UPDATECHILDREN$(k, a, a+1)$
      $m_{\text{reopt}} \leftarrow m_{\text{reopt}} + 1$
  else
    $m_{\text{leafopt}} \leftarrow 0$
    $\phi, n^{k,a} \leftarrow$ OPTIMIZELEAF(s)
    while $n^{k,a} > f_{\text{stop}}|s|$ **and** $m_{\text{leafopt}} < M_{\text{leafopt}}$
      $\hat{\phi}, \hat{n} \leftarrow$ OPTIMIZELEAF(s)
      if $\hat{n} < n^{k,a}$
        $\phi, n^{k,a} \leftarrow \hat{\phi}, \hat{n}$
      $m_{\text{leafopt}} \leftarrow m_{\text{leafopt}} + 1$
  return ϕ_{native}

UPDATECHILDREN(k, a, b)
  if HASCHILDREN(k)
    $n^{k_l,a} \leftarrow n^{k_l,b}$
    $n^{k_r,a} \leftarrow n^{k_r,b}$
    UPDATECHILDREN(k_l, a, b)
    UPDATECHILDREN(k_r, a, b)

OPTIMIZELEAF(s)
  $m_{\text{unfavorable}} \leftarrow 0$
  $\gamma_{\text{unfavorable}} \leftarrow \emptyset$
  $\phi \leftarrow$ INITSEQ(s)
  $n \leftarrow$ ENSEMBLEDEFECT(ϕ, s)
  while $n > f_{\text{stop}}|s|$ **and** $m_{\text{unfavorable}} < M_{\text{unfavorable}}|s|$
    $\xi, \hat{\phi} \leftarrow$ UNIFORMMUTATIONSAMPLING(ϕ, s)
    if $\xi \in \gamma_{\text{unfavorable}}$
      $m_{\text{unfavorable}} \leftarrow m_{\text{unfavorable}} + 1$
    else
      $\hat{n} \leftarrow$ ENSEMBLEDEFECT$(\hat{\phi}, s)$
      if $\hat{n} < n$
        $\phi, n \leftarrow \hat{\phi}, \hat{n}$
        $m_{\text{unfavorable}} \leftarrow 0$
        $\gamma_{\text{unfavorable}} \leftarrow \emptyset$
      else
        $m_{\text{unfavorable}} \leftarrow m_{\text{unfavorable}} + 1$
        $\gamma_{\text{unfavorable}} \leftarrow \gamma_{\text{unfavorable}} \bigcup \xi$
  return ϕ, n

Algorithm S3: Hierarchical ensemble defect optimization with uniform sampling. Pseudocode conventions follow those of Algorithm 1.

DESIGNSEQ(s)

$m_{\text{leafopt}} \leftarrow 0$
$\phi, \pi \leftarrow$ OPTIMIZELEAF(s)
while $\pi > f_{\text{stop}}$ **and** $m_{\text{leafopt}} < M_{\text{leafopt}}$
  $\hat{\phi}, \hat{\pi} \leftarrow$ OPTIMIZELEAF(s)
  if $\hat{\pi} < \pi$
    $\phi, \pi \leftarrow \hat{\phi}, \hat{\pi}$
  $m_{\text{leafopt}} \leftarrow m_{\text{leafopt}} + 1$
return ϕ

OPTIMIZELEAF(s)

$m_{\text{unfavorable}} \leftarrow 0$
$\gamma_{\text{unfavorable}} \leftarrow \emptyset$
$\phi \leftarrow$ INITSEQ(s)
$\pi \leftarrow$PROBABILITYDEFECT(ϕ, s)
while $\pi > f_{\text{stop}}$ **and** $m_{\text{unfavorable}} < M_{\text{unfavorable}}|s|$
  $\xi, \hat{\phi} \leftarrow$ UNIFORMMUTATIONSAMPLING(ϕ, s)
  if $\xi \in \gamma_{\text{unfavorable}}$
    $m_{\text{unfavorable}} \leftarrow m_{\text{unfavorable}} + 1$
  else
    $\hat{\pi} \leftarrow$PROBABILITYDEFECT($\hat{\phi}, s$)
    if $\hat{\pi} < \pi$
      $\phi, \pi \leftarrow \hat{\phi}, \hat{\pi}$
      $m_{\text{unfavorable}} \leftarrow 0$
      $\gamma_{\text{unfavorable}} \leftarrow \emptyset$
    else
      $m_{\text{unfavorable}} \leftarrow m_{\text{unfavorable}} + 1$
      $\gamma_{\text{unfavorable}} \leftarrow \gamma_{\text{unfavorable}} \bigcup \xi$
return ϕ, π

Algorithm S4: Single-scale probability defect optimization with uniform mutation sampling.

DESIGNSEQ(ϕ, s, μ, k)
  $a \leftarrow$ DEPTH(k)
  if HASCHILDREN(k)
    $m_{\text{reopt}} \leftarrow 0$
    if $\mu = \emptyset$
      $\phi_l \leftarrow$ DESIGNSEQ($\emptyset, s_{l+}, \emptyset, k_l$)
      $\phi_r \leftarrow$ DESIGNSEQ($\emptyset, s_{r+}, \emptyset, k_r$)
    else
      UPDATECHILDREN($k, a, a-1$)
      child, $\phi \leftarrow$ WEIGHTEDCHILDSAMPLING(ϕ, s, μ_l, μ_r)
      $\phi_{\text{child}} \leftarrow$ DESIGNSEQ($\phi_{\text{child}+}, s_{\text{child}+}, \mu_{\text{child}+}, k_{\text{child}}$)
    $\mu^{k,a} \leftarrow$ MFEDEFECT(ϕ, s)
    UPDATECHILDREN($k, a, a+1$)
    while $\mu^{k,a} > \max(f_{\text{stop}}|s_l|, \mu^{k_l,a}_{\text{native}}) + \max(f_{\text{stop}}|s_r|, \mu^{k_r,a}_{\text{native}})$
        and $m_{\text{reopt}} < M_{\text{reopt}}$
      $\hat{\mu}_i \leftarrow \mu_i^{k,a} + \epsilon \quad \forall i \in \{1, \ldots, |s|\}$
      child, $\hat{\phi} \leftarrow$ WEIGHTEDCHILDSAMPLING($\phi, s, \hat{\mu}_l, \hat{\mu}_r$)
      $\hat{\phi}_{\text{child}} \leftarrow$ DESIGNSEQ($\phi_{\text{child}+}, s_{\text{child}+}, \hat{\mu}_{\text{child}+}, k_{\text{child}}$)
      $\hat{\mu} \leftarrow$ MFEDEFECT($\hat{\phi}, s$)
      if $\hat{\mu} < \mu^{k,a}$
        $\phi, \mu^{k,a} \leftarrow \hat{\phi}, \hat{\mu}$
        UPDATECHILDREN($k, a, a+1$)
      $m_{\text{reopt}} \leftarrow m_{\text{reopt}} + 1$
  else
    $m_{\text{leafopt}} \leftarrow 0$
    $\phi, \mu^{k,a} \leftarrow$ OPTIMIZELEAF(s)
    while $\mu^{k,a} > f_{\text{stop}}|s|$ **and** $m_{\text{leafopt}} < M_{\text{leafopt}}$
      $\hat{\phi}, \hat{\mu} \leftarrow$ OPTIMIZELEAF(s)
      if $\hat{\mu} < \mu^{k,a}$
        $\phi, \mu^{k,a} \leftarrow \hat{\phi}, \hat{\mu}$
      $m_{\text{leafopt}} \leftarrow m_{\text{leafopt}} + 1$
  return ϕ_{native}

UPDATECHILDREN(k, a, b)
  if HASCHILDREN(k)
    $\mu^{k_l,a} \leftarrow \mu^{k_l,b}$
    $\mu^{k_r,a} \leftarrow \mu^{k_r,b}$
    UPDATECHILDREN(k_l, a, b)
    UPDATECHILDREN(k_r, a, b)

OPTIMIZELEAF(s)
  $m_{\text{try}} \leftarrow 0$
  $m_{\text{unfavorable}} \leftarrow 0$
  $\gamma_{\text{unfavorable}} \leftarrow \emptyset$
  $\phi \leftarrow$ INITSEQ(s)
  $\mu \leftarrow$ MFEDEFECT(ϕ, s)
  while $\mu > f_{\text{stop}}|s|$ **and** $m_{\text{unfavorable}} < M_{\text{unfavorable}}|s|$
      and $m_{\text{try}} < M_{\text{try}}$
    $\hat{\mu}_i \leftarrow \mu_i + \epsilon \quad \forall\, i \in \{1, \ldots, |s|\}$
    $\xi, \hat{\phi} \leftarrow$ WEIGHTEDMUTATIONSAMPLING($\phi, s, \hat{\mu}_1, \ldots, \hat{\mu}_{|s|}$)
    if $\xi \in \gamma_{\text{unfavorable}}$
      $m_{\text{unfavorable}} \leftarrow m_{\text{unfavorable}} + 1$
    else
      $\hat{\mu} \leftarrow$ MFEDEFECT($\hat{\phi}, s$)
      if $\hat{\mu} < \mu$ **or** ACCEPTUNFAVORABLE(f_{accept})
        $\phi, \mu \leftarrow \hat{\phi}, \hat{\mu}$
        $m_{\text{unfavorable}} \leftarrow 0$
        $\gamma_{\text{unfavorable}} \leftarrow \emptyset$
      else
        $m_{\text{unfavorable}} \leftarrow m_{\text{unfavorable}} + 1$
        $\gamma_{\text{unfavorable}} \leftarrow \gamma_{\text{unfavorable}} \bigcup \xi$
    $m_{\text{try}} \leftarrow m_{\text{try}} + 1$
  return ϕ, μ

Algorithm S5: Hierarchical MFE defect optimization with defect-weighted sampling. During leaf optimization, we employ defect-weighted mutation sampling, selecting nucleotide i as a mutation candidate with probability $(\mu_i^{k,a} + \epsilon)/(\mu^{k,a} + \epsilon|s|)$. Adding ϵ to each defect contribution ensures that all bases (even those with $\mu_i^{k,a} = 0$) are subject to mutation with a non-zero probability. During leaf optimization, fraction f_{accept} of unfavorable candidate mutations are accepted to assist in escaping from local minima. The leaf stop condition is $\mu^{k,a} < f_{\text{stop}}|s|$; the parental stop condition is $\mu^{k,a} < \max(f_{\text{stop}}|s_l|, \mu^{k_l,a}_{\text{native}}) + \max(f_{\text{stop}}|s_r|, \mu^{k_r,a}_{\text{native}})$. Because some unfavorable mutations are accepted, the total number of mutation attempts during a leaf optimization is limited to M_{try}. Calculations are performed with defaults values: $\epsilon = 0.1$, $f_{\text{accept}} = 0.2$, $M_{\text{try}} = 5000$. Pseudocode conventions follow those of Algorithm 1.

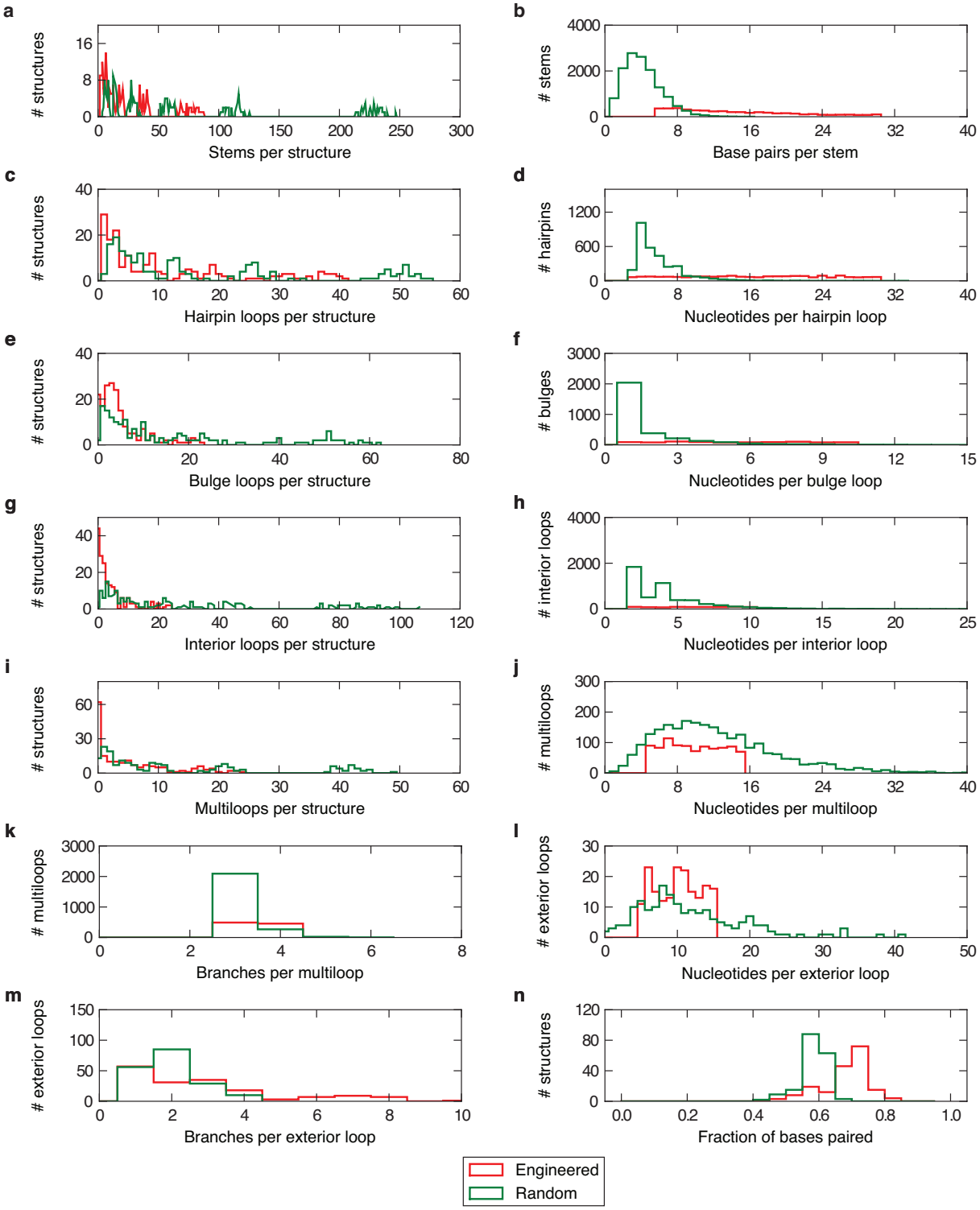

Figure S1: Comparison of the structural features of the engineered and random test sets.

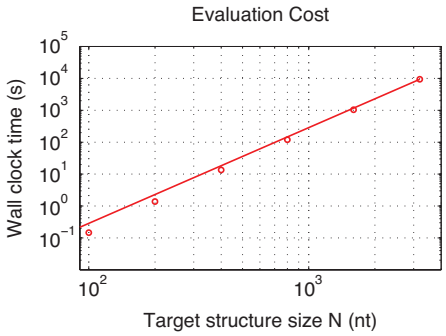

Figure S2: Computational cost, $c_{\text{eval}}(N) = \Theta(N^3)$, of a single evaluation of the ensemble defect, $n(\phi, s)$, for the full sequence and target structure. Each data point represents the median over all sequences for a particular value of N. The line depicts a slope of three, suggesting empirically that the dynamic program is operating approximately within the asymptotic regime for this range of N. RNA design at 37°C on the engineered test set.

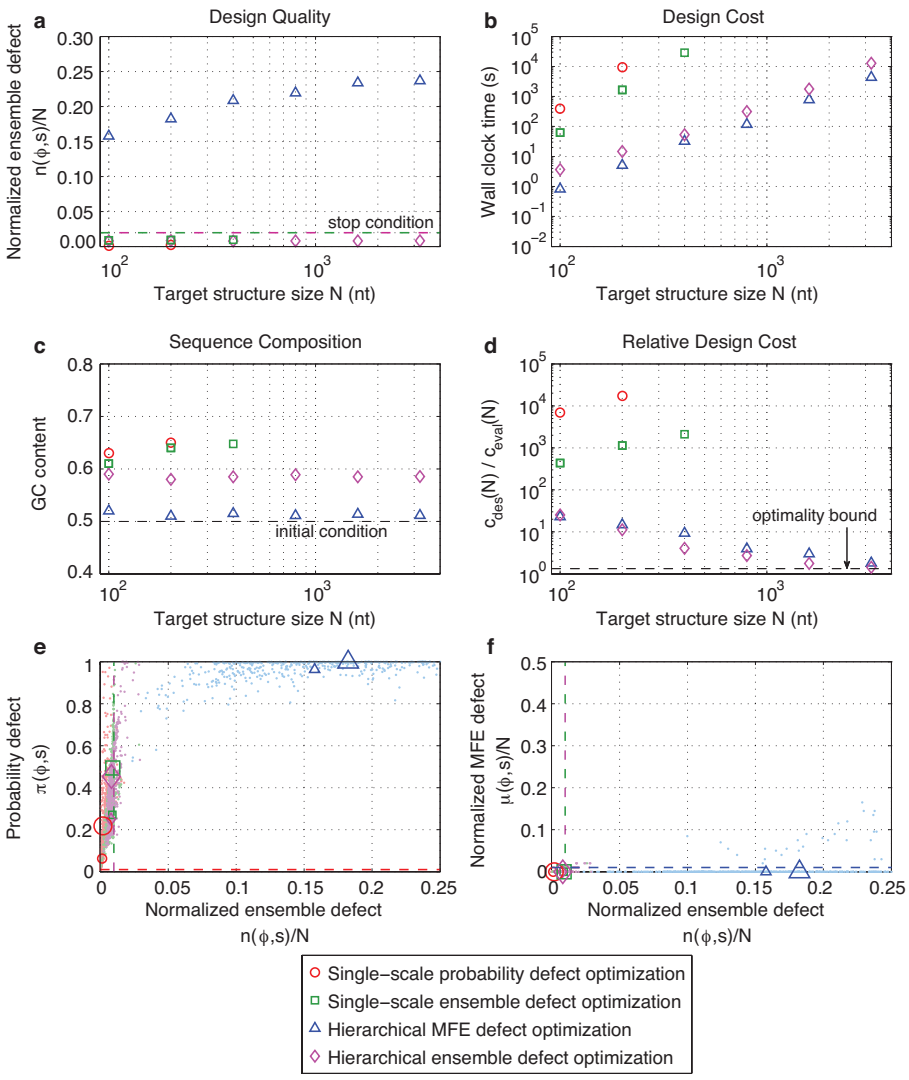

Figure S3: Comparison to algorithms inspired by previous publications. a) Design quality. The ensemble defect stop condition is depicted as a dashed line. b) Design cost. c) Sequence conmposition. The initial GC content is depicted as a dashed line. d) Cost of sequence design relative to a single evaluation of the objective function. The optimality bound is depicted as a dashed line. e,f) Evaluation of each sequence design using three objective functions. Dots represent independent designs. Symbols denote medians for each value of $N \in \{100, 200\}$ (symbol size increases with N). RNA design at 37°C on the random test set.