Supplementary Information

Sequence Design for a Test Tube of Interacting Nucleic Acid Strands

Brian R. Wolfe[1] and Niles A. Pierce[1,2,*]

[1]Division of Biology & Biological Engineering, [2]Division of Engineering & Applied Science,
California Institute of Technology, Pasadena, CA 91125, USA

[*]Email: niles@caltech.edu

Contents

S1 Additional Algorithm Details

S1.1 Pseudocode

```
OptimizeTube($\Psi^{\rm on}, \Psi^{\rm off}, \Psi, s_\Psi, y_\Psi$)
    $\Psi^{\rm active}, \Psi^{\rm passive} \leftarrow \Psi^{\rm on}, \Psi^{\rm off}$
    $\phi_{\Psi^{\rm active}} \leftarrow$ InitSeq($s_{\Psi^{\rm active}}$)
    $\Lambda, D \leftarrow$ MakeForest($s_{\Psi^{\rm active}}$)
    $\phi_\Lambda, \tilde{C}_1 \leftarrow$ OptimizeForest($\phi_\Lambda, D$)
    $C \leftarrow$ EvaluateDefect($\phi_\Psi$)
    $\hat{\phi}_\Psi, \hat{C} \leftarrow \phi_\Psi, C$
    while $\hat{C} > \max(C_{\rm stop}, \tilde{C}_1)$
        $\Psi^{\rm active}, \Psi^{\rm passive} \leftarrow$ RefocusTube($\Psi^{\rm active}, \Psi^{\rm passive}, \hat{x}_{\Psi^{\rm passive}}$)
        $\Lambda, D \leftarrow$ AugmentForest($\Lambda, D, \hat{P}_{\Psi^{\rm active}}$)
        $\hat{\phi}_\Lambda, \tilde{C}_1 \leftarrow$ OptimizeForest($\hat{\phi}_\Lambda, D$)
        $\hat{C} \leftarrow$ EvaluateDefect($\hat{\phi}_\Psi$)
        if $\hat{C} < C$
            $\phi_\Psi, C \leftarrow \hat{\phi}_\Psi, \hat{C}$
    return $\phi_\Psi$

OptimizeForest($\phi_\Lambda, D$)
    $\tilde{C}_d \leftarrow \infty \ \forall d \in \{1, \ldots, D\}$
    $\beta_{\rm merge} \leftarrow$ false
    while $\neg\beta_{\rm merge}$
        $\phi_{\Lambda_D}, \tilde{C}_D \leftarrow$ OptimizeLeaves($\phi_{\Lambda_D}, D$)
        $d \leftarrow D - 1$
        $\beta_{\rm merge} \leftarrow$ true
        while $d \geq 1$ and $\beta_{\rm merge}$
            $\hat{\phi}_{\Lambda_d} \leftarrow$ MergeSeq($\phi_{\Lambda_{d+1}}$)
            $\hat{C}_d \leftarrow$ EstimateDefect($\hat{\phi}_{\Lambda_d}$)
            if $\hat{C}_d < \tilde{C}_d$
                $\phi_{\Lambda_d}, \tilde{C}_d \leftarrow \hat{\phi}_{\Lambda_d}, \hat{C}_d$
            if $\hat{C}_d > \max(C_d^{\rm stop}, \tilde{C}_{d+1}/f_{\rm stringent})$
                $\beta_{\rm merge} \leftarrow$false
                $\Lambda, D \leftarrow$ RedecomposeForest($\Lambda, D, s_{\Lambda_d}, \hat{P}_{\Lambda_d}$)
                $\phi_{\Lambda_D} \leftarrow$ SplitSeq($\hat{\phi}_{\Lambda_d}$)
                $\tilde{C}_{d'} \leftarrow \infty \ \forall d' \in \{d+1, \ldots, D\}$
            $d \leftarrow d - 1$
    return $\phi_{\Lambda_1}, \tilde{C}_1$

OptimizeLeaves($\phi_{\Lambda_D}, D$)
    $\phi_{\Lambda_D}, \tilde{C}_D \leftarrow$ MutateLeaves($\phi_{\Lambda_D}, D$)
    $m_{\rm reopt} \leftarrow 0$
    while $\tilde{C}_D > C_D^{\rm stop}$ and $m_{\rm reopt} < M_{\rm reopt}$
        $\hat{\phi}_{\Lambda_D} \leftarrow$ ReseedSeq($\phi_{\Lambda_D}, \{\tilde{C}_D^a\}$)
        $\hat{\phi}_{\Lambda_D}, \hat{C}_D \leftarrow$ MutateLeaves($\hat{\phi}_{\Lambda_D}, D$)
        if $\hat{C}_D < \tilde{C}_D$
            $\phi_{\Lambda_D}, \tilde{C}_D \leftarrow \hat{\phi}_{\Lambda_D}, \hat{C}_D$
            $m_{\rm reopt} \leftarrow 0$
        else
            $m_{\rm reopt} \leftarrow m_{\rm reopt} + 1$
    return $\phi_{\Lambda_D}, \tilde{C}_D$

MutateLeaves($\phi_{\Lambda_D}, D$)
    $\tilde{C}_D \leftarrow$ EstimateDefect($\phi_{\Lambda_D}$)
    $\gamma_{\rm bad} \leftarrow \emptyset, \quad m_{\rm bad} \leftarrow 0$
    while $\tilde{C}_D > C_D^{\rm stop}$ and $m_{\rm bad} < M_{\rm bad}$
        $\xi, \hat{\phi}_{\Lambda_D} \leftarrow$ SampleMutation($\phi_{\Lambda_D}, \{\tilde{C}_D^a\}$)
        if $\xi \in \gamma_{\rm bad}$
            $m_{\rm bad} \leftarrow m_{\rm bad} + 1$
        else
            $\hat{C}_D \leftarrow$ EstimateDefect($\hat{\phi}_{\Lambda_D}$)
            if $\hat{C}_D < \tilde{C}_D$
                $\phi_{\Lambda_D}, \tilde{C}_D \leftarrow \hat{\phi}_{\Lambda_D}, \hat{C}_D$
                $\gamma_{\rm bad} \leftarrow \emptyset, \quad m_{\rm bad} \leftarrow 0$
            else
                $\gamma_{\rm bad} \leftarrow \gamma_{\rm bad} \cup \xi, \quad m_{\rm bad} \leftarrow m_{\rm bad} + 1$
    return $\phi_{\Lambda_D}, \tilde{C}_D$

EstimateDefect($\phi_{\Lambda_d}$)
    $\tilde{Q}_{\Lambda_d}, \tilde{P}_{\Lambda_d} \leftarrow$ ConditionalNodalProperties($\phi_{\Lambda_d}$)
    $\tilde{Q}_{\Psi^{\rm active}} \leftarrow$ EstimateComplexPfuncs($\tilde{Q}_{\Lambda_d}$)
    $\tilde{P}_{\Psi^{\rm active}} \leftarrow$ EstimateComplexPairProbs($\tilde{P}_{\Lambda_d}$)
    $\tilde{x}^0_{\Psi^0} \leftarrow$ DeflateMassConstraints($x^0_{\Psi^0}$)
    $\tilde{x}_{\Psi^{\rm active}} \leftarrow$ EstimateComplexConcentrations($\tilde{Q}_{\Psi^{\rm active}}, \tilde{x}^0_{\Psi^0}$)
    $\tilde{n}_{\Psi^{\rm on}} \leftarrow$ EstimateComplexDefects($\tilde{P}_{\Psi^{\rm on}}, s_{\Psi^{\rm on}}$)
    $\tilde{c}_{\Psi^{\rm on}} \leftarrow$ EstimateTubeDefectTerms($\tilde{n}_{\Psi^{\rm on}}, \tilde{x}_{\Psi^{\rm on}}, y_{\Psi^{\rm on}}$)
    $\tilde{C}_d = \sum_{j \in \Psi^{\rm on}} \tilde{c}_j$
    return $\tilde{C}_d$
```

Algorithm S1. Pseudocode for test tube ensemble defect optimization. For a target test tube containing the set of complexes, Ψ (comprising on-target complexes, $\Psi^{\rm on}$, and off-target complexes, $\Psi^{\rm off}$), with target secondary structures, s_Ψ, and target concentrations, y_Ψ, a set of designed sequences, ϕ_Ψ, is returned by the function call OptimizeTube($\Psi^{\rm on}, \Psi^{\rm off}, \Psi, s_\Psi, y_\Psi$).

S1.2 Calculation of Conditional Partition Functions and Base-Pairing Probabilities

Let E_k denote the set of base pairs that are enforced in node k, and are hence adjacent to a split-point in some ancestor. We seek to calculate the conditional partition function (6):

$$\tilde{Q}_k \equiv Q(\phi_k|E_k)$$

and the conditional equilibrium base-pairing probabilities (8):

$$\tilde{P}_k \equiv P(\phi_k|E_k)$$

over the conditional ensemble, $\tilde{\Gamma}_k$, comprising all structures in Γ_k that contain all base pairs in E_k. Our approach is to apply standard dynamic programs[1] that operate over ensemble Γ_k, but to modify the free energy model so that contributions from structures that do not contain all the base pairs in E_k are effectively neglected.

The partition function is calculated in a forward sweep moving from short subsequences to the full nodal sequence. Equilibrium base-pairing probabilities are then calculated in a backward sweep moving from the full nodal sequence back to short subsequences. During the forward sweep, each time a base pair in E_k is encountered, the standard free energy is augmented with a bonus energy, ΔG^{clamp}, thus increasing the partition function contribution of every structure containing that base pair by a factor of $\exp(-\Delta G^{\text{clamp}}/k_B T)$. By choosing ΔG^{clamp} to be sufficiently negative, contributions from structures lacking any base pair in E_k are effectively neglected. After the forward sweep over ensemble Γ_k using the modified free energy model:

- the backward sweep returns conditional pair probabilities $P(\phi_k|E_k)$ over the conditional ensemble $\tilde{\Gamma}_k$ for the standard free energy model,
- the computed partition function value for the full nodal sequence is divided by $\exp(-|E_k|\Delta G^{\text{clamp}}/k_B T)$ to recover the conditional partition function, $Q(\phi_k|E_k)$, over the conditional ensemble $\tilde{\Gamma}_k$ for the standard free energy model.

These calculations are performed using dynamic programs suitable for complexes containing arbitrary numbers of strands.[1] Energetic bonuses (and/or penalties) have previously been used to enforce experimentally determined structural constraints as a means of improving structure prediction accuracy.[2,3]

After calculating the left-child and right-child conditional partition functions, $\tilde{Q}_{k_l}$ and $\tilde{Q}_{k_r}$, the partition function estimate for parent k with split-point F is then (7):

$$\tilde{Q}_k = \tilde{Q}_{k_l} \tilde{Q}_{k_r} \exp\bigl(-\Delta G_F^{\text{interior}}/k_B T\bigr),$$

where $\Delta G_F^{\text{interior}}$ is the free energy for the interior loop formed by the base-pairs sandwiching F. In each child, the base pair adjacent to F in the parent is enforced in the child using an energetic bonus as described above. By supposition, this base pair terminates a duplex in every structure in the conditional child ensemble, but borders an interior loop in the conditional parent ensemble. Using nearest neighbor free energy models that include a free energy penalty, $\Delta G^{\text{terminal}}$, for terminating a duplex with a non-G·C base pair,[2,4] if a child has a non-G·C base pair as the terminal clamped pair, the parental partition function must be compensated to remove the penalty present in the child partition function. This is achieved by multiplying the child partition function by a factor of $\exp(\Delta G^{\text{terminal}}/k_B T)$ at the time the parental partition function is reconstructed using (7).

S1.3 Distinguishability Issues

If a complex contains multiple copies of the same strand species, subtleties arise in the definition of the structural ensemble and in the calculation of experimental observables (see Reference 1 for details). Consider L strands that form complex $j \in \Psi$ with strand ordering π_j. The structural ensemble of complex j contains all connected polymer graphs with no crossing lines. Let Γ_j denote the ensemble in which each strand is treated as distinct (i.e., each strand has a unique identifier in $\{1, \ldots, L\}$) and let Γ'_j denote the ensemble in which strands of the same species are treated as indistinguishable. Two secondary structures are indistinguishable if their polymer graphs can be rotated so that all strands are mapped onto indistinguishable strands, all base pairs are mapped onto base pairs, and all unpaired bases are mapped onto unpaired bases; otherwise the structures are distinct.[1] The ensemble $\Gamma'_j \subseteq \Gamma_j$ is a maximal subset of distinct secondary structures for strand ordering π_j.

Hierarchical ensemble decomposition treats each strand within complex j as distinct and is performed with respect to ensemble Γ_j. Likewise, the equilibrium base-pairing probabilities, P_j, and the complex ensemble defect, n_j, characterize equilibrium base-pairing over the ensemble Γ_j treating each strand within complex j as distinct.

When measuring complex concentrations in the laboratory, strands of the same species are indistinguishable, so it is desirable to calculate the complex partition function, Q_j, over ensemble Γ'_j in order to obtain a complex concentration, x_j, that is directly comparable to experimental measurement. The dynamic programs that are used to calculate the complex partition function and equilibrium base-pairing probabilities treat each strand within complex j as distinct and operate over ensemble Γ_j.[1] The periodic strand repeat v_j of complex j with strand ordering π_j is defined as the number of different rotations of the polymer graph that map indistinguishable strands onto each other (e.g., $v_j = 4$ for complex AAAA, $v_j = 3$ for complex ABABAB, $v_j = 2$ for ABAABA). For complexes in which all strands are distinct, $v_j = 1$. Complexes containing multiple copies of the same strand species may have $v_j > 1$, in which case the calculated partition function will be incorrect due to symmetry and overcounting errors that are different for different structures in Γ_j.[1] Fortunately, these errors interact in such a way that they can be exactly and simultaneously corrected by dividing the calculated partition function by the integer v_j:

$$Q_j = Q_j^{\text{calc}}/v_j,$$

yielding a partition function over ensemble Γ'_j using a symmetry-corrected physical model.[1] In the context of hierarchical ensemble decomposition, the calculated complex partition function estimate, $\tilde{Q}_j^{\text{calc}}$, based on conditional partition functions calculated at any level $d \in \{2, \ldots, D\}$ within the decomposition forest Λ (see Section S1.2), is corrected by dividing by v_j.

The test tube ensemble contains the set of complexes

$$\Psi = \Psi^{\text{on}} \cup \Psi^{\text{off}} \quad \text{with} \quad \Psi^{\text{on}} \cap \Psi^{\text{off}} = \emptyset,$$

where $|\Psi^{\text{on}}|$ and $|\Psi^{\text{off}}|$ are arbitrary. In practice, it is convenient to define Ψ^{off} to contain all complexes of up to size L_{max} (excluding those that are in Ψ^{on}). Consider a test tube containing the set of strand species Ψ^0 that interact to form all possible complexes of size L, for $L \in \{1, \ldots, L_{\text{max}}\}$. Each complex j is specified as a distinct strand ordering π_j. L distinct strands can be ordered around a circle in $(L-1)!$ distinct ways (e.g., strands A, B, and C can be ordered ABC and ACB). If some of the L strands are of the same species, there will be fewer than $(L-1)!$ distinct strand orderings (e.g., strands A, A, and B can only be ordered AAB). For a given set of L strands, each unpseudoknotted connected secondary structure is found in the structural ensemble, Γ_j, corresponding to exactly one strand ordering, π_j (i.e., in exactly one complex $j \in \Psi$).[1] For the set of strands Ψ^0, the total number of complexes of up to size L_{max} is given by (2):[1]

$$\sum_{L=1}^{L_{\text{max}}} \sum_{l=1}^{L} \frac{|\Psi^0|^{\gcd(l,L)}}{L}.$$

The set Ψ^{off} is then defined to include all of these complexes except those that are in Ψ^{on}.

S1.4 Selection and Use of Multiple Exclusive Split-Points

During probability-guided decomposition, the optimal set of exclusive split points, $\{F\}^*$, is chosen to minimize the *cost*:

$$\text{cost}(\{F\}) \equiv \sum_{F_i \in \{F\}} \left(|\phi_{k_{l_i}}|^3 + |\phi_{k_{r_i}}|^3 \right),$$

subject to constraints (16) on the *collective probability*:

$$P(\{F\}) \equiv \sum_{F_i \in \{F\}} \min_{a \cdot b \in F_i^{\pm}} P_k^{a,b},$$

the minimum child size, and split-point exclusivity. We solve this optimization problem using a depth-first branch and bound procedure, where each branch in the optimization tree corresponds to a set of split-points satisfying the exclusivity and child-size constraints. Starting at the root with no split-points, at each branching step, a split-point is added so as to greedily maximize the collective probability, $P(\{F\})$, of the current branch, $\{F\}$, while satisfying the exclusivity and child-size constraints. The cost of the current branch, cost$(\{F\})$, serves as a lower bound on the cost of all unexplored subtrees of this branch. Branching continues until a branch is encountered that satisfies the collective probability constraint, $P(\{F\}) \geq f_{\text{split}}$. This branch defines the current minimal-cost solution, $\{\bar{F}\}$, with $\text{cost}(\{\bar{F}\})$ providing an upper bound on the cost of the optimal split-set. Backtracking commences by removing the final split-point from $\{\bar{F}\}$ and exploring rival branches by greedily maximizing collective probability. During backtracking, branches are explored only if their cost is less than $\text{cost}(\{\bar{F}\})$, otherwise they are trimmed. If an untrimmed branch is encountered that satisfies the collective probability constraint, it becomes the current minimal-cost solution and $\{\bar{F}\}$ and $\text{cost}(\{\bar{F}\})$ are updated. After all branches have been trimmed, the current $\{\bar{F}\}$ is the optimal split set $\{F\}^*$. The computational cost of selecting the optimal set of split-points is typically negligible compared to the cost of evaluating partition functions and equilibrium pair probabilities.

For structure- and probability-guided decomposition using multiple exclusive split-points, the same approach is used except that the first branch is selected from $B(S_k)$ (15). Note that for a child node, k, generated using multiple exclusive split-points, the target structure matrix, S_k, may have row sums not equal to unity; if a nucleotide a in node k is intended to form an on-target base pair with a nucleotide not contained in node k, this will cause S_k to be zero for all entries in row a.

Figure S1 shows the extent to which the algorithm exploits multiple exclusive split-points for the engineered test set. Only a small fraction of parents use multiple exclusive split-points (panel a), but for test tubes with 400-nt on-targets, approximately 40% of complexes (panel b) and 100% of design trials (panel c) contain a parent node with at least four children (and thus at least two exclusive split-points). The cost savings using multiple exclusive split-points become substantial as the size of the decomposed target structures increases (Figure 7; *cf.* solid and dotted lines).

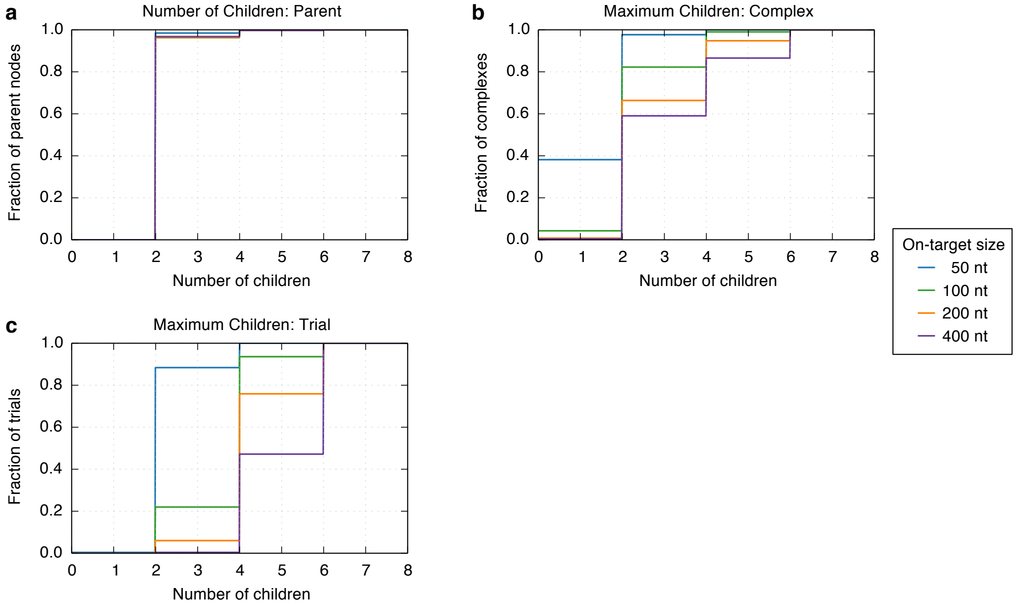

Figure S1. Usage of multiple exclusive split-points in the final decomposition forest. a) Children per parent. b) Maximum number of children for any parent node in each complex. c) Maximum number of children for any parent node in each design trial. RNA design at 37 °C for the engineered test set.

S2 Additional Design Studies

S2.1 Sensitivity of Performance to Algorithm Parameters

Here, we examine the sensitivity of algorithm performance to each algorithm parameter. Sensitivity studies were carried out for RNA design at 37 °C for the subsets of the engineered and random test sets with 200-nt on-targets. Two design trials were carried out for each target test tube, yielding a total of 200 design trials per parameter value. The sensitivity study for each parameter was carried out with all other parameters at their default values (Table 1).

- f_{stop} (permitted normalized test tube ensemble defect, default: 0.01, Figure S2): For values between 0.01 and 0.1, the desired design quality is typically achieved and there is little affect on design cost. Using 0.001 and 0.003, the desired design quality is typically not achieved and design cost increases.
- f_{passive} (fraction of the permitted test tube ensemble defect that is allocated to off-target complexes, default: 0.01, Figure S3): For values between 0 and 0.1, there is little affect on design quality or design cost.
- H_{split} (minimum number of stable base pairs on either side of a split-point during hierarchical ensemble decomposition, default: 2, Figure S4): For values between 1 and 4, there is little affect on typical design quality, but using 1, design quality degrades for the hardest cases. For the engineered test set, typical design cost is comparable using 2 or 3 but increases slightly using 1 or 4. For the random test set, typical design cost is comparable using 1 or 2 but increases using 3 or 4.
- N_{split} (minimum number of nucleotides in a node within the decomposition forest, default: 12, Figure S5): For values between 8 and 60, there is little affect on design quality. Design cost is similar for values between 8 and 40 and typically increases slightly using 60.
- f_{split} (minimum probability accounted for by exclusive split-points, default: 0.99, Figure S6): For values between 0.900 and 0.999, there is little affect on design quality for either test set, and there is little affect on design cost for the engineered test set. For the random test set, design cost is similar for values between 0.900 and 0.997 and typically increases slightly using 0.999.
- $f_{\text{stringent}}$ (factor by which the stop condition is made more stringent at each level moving down the decomposition forest, default: 0.99, Figure S7): For values between 0.970 and 1.000, there is little affect on design quality or design cost. Using 0.900, design quality typically overshoots the stop condition and design cost increases for the harder cases.
- ΔG^{clamp} (bonus free energy used to enforce base pairs in conditional child ensembles within the decomposition forest, default = -25 kcal/mol, Figure S8): For values between -5 and -50 kcal/mol, there is little affect on design quality or design cost.
- M_{bad} (maximum number of consecutive failed mutation attempts before leaf mutation terminates unsuccessfully, default: 300, Figure S9): For the engineered test set, there is little affect on design quality or design cost for values between 10 and 300. For the random test set, design quality and design cost are both comparable using 100 or 300, but design quality sometimes decreases and design cost typically increases using 10 or 30.
- M_{reseed} (number of nucleotides reseeded prior to leaf reoptimization, default = 50, Figure S10): For values between 10 and 250, there is little affect on design quality or design cost.
- M_{reopt} (maximum number of consecutive failed leaf reoptimization attempts before leaf optimization terminates unsuccessfully, default: 3, Figure S11): For values between 1 and 30, there is little affect on design quality or design cost.
- f_{redecomp} (fraction of the decomposition defect that must be accounted for during parental redecomposition, default: 0.03, Figure S12): For values between 0 and 0.1, there is little affect on design quality or design cost.
- f_{refocus} (fraction of the focusing defect that must be accounted for during test tube ensemble refocusing, default: 0.03, Figure S13): For values between 0 and 0.1, there is little affect on design quality for either test set, and there is little affect on design cost for the engineered test set. For the random test set, design cost typically decreases as the value increases from 0 to 0.1.

It remains an open question whether some parameters can be removed from the algorithm without decreasing design quality or increasing design cost for target test tubes that arise in practice.

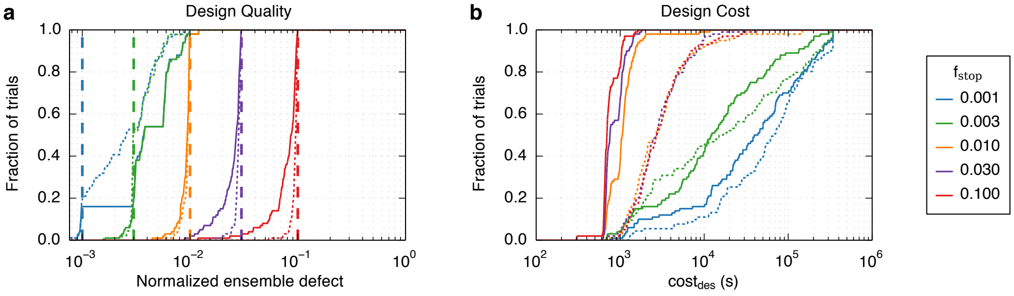

Figure S2. Sensitivity of algorithm performance to f_{stop} (default: 0.01). a) Design quality. Each stop condition is depicted as a dashed colored line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

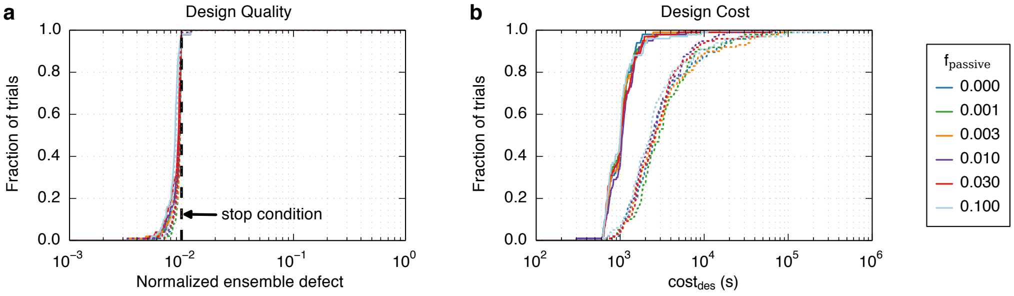

Figure S3. Sensitivity of algorithm performance to f_{passive} (default: 0.01). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

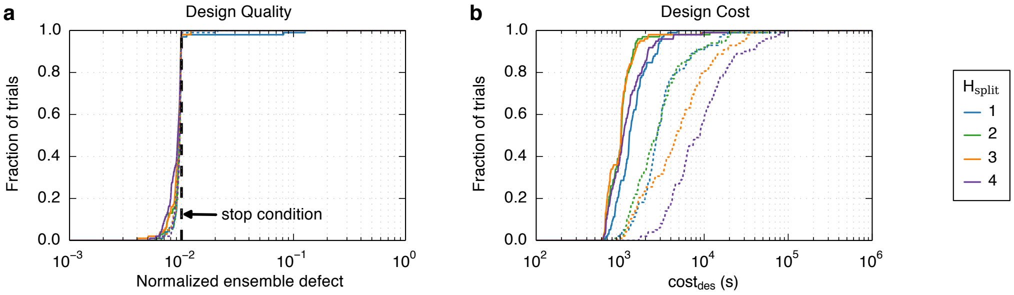

Figure S4. Sensitivity of algorithm performance to H_{split} (default: 2). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

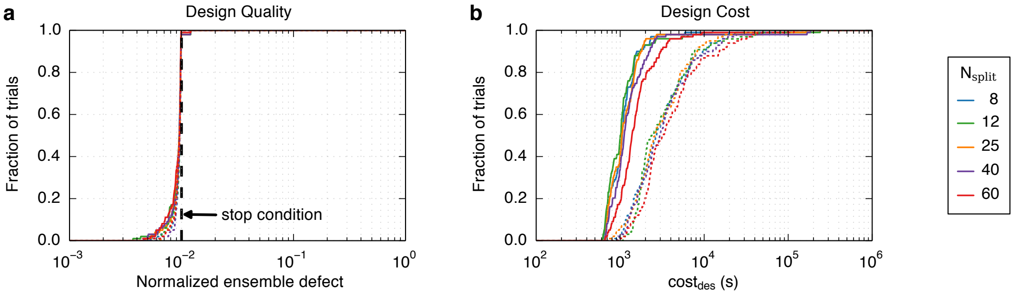

Figure S5. Sensitivity of algorithm performance to N_{split} (default: 12). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

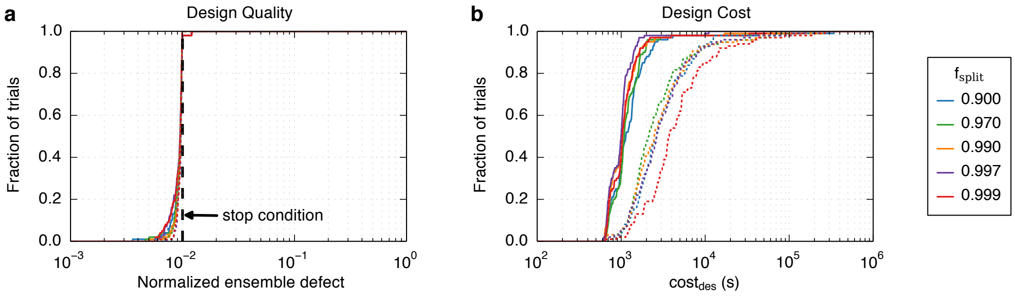

Figure S6. Sensitivity of algorithm performance to f_{split} (default: 0.99). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

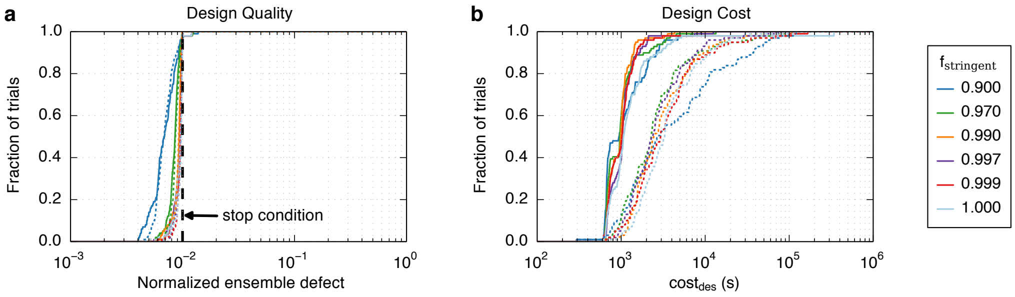

Figure S7. Sensitivity of algorithm performance to $f_{\mathrm{stringent}}$ (default: 0.99). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

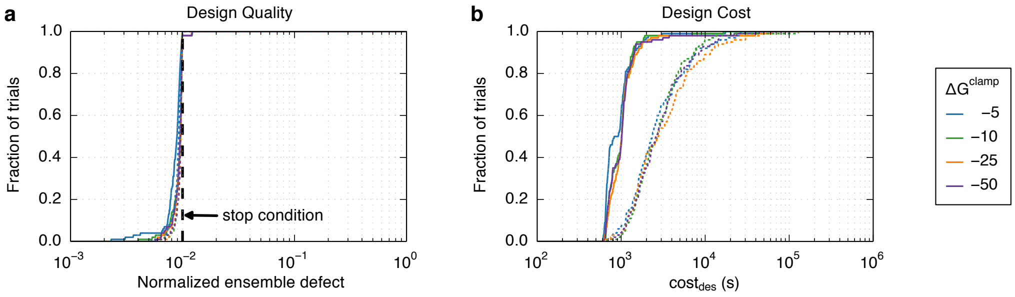

Figure S8. Sensitivity of algorithm performance to $\Delta G^{\mathrm{clamp}}$ (default: -25 kcal/mol). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

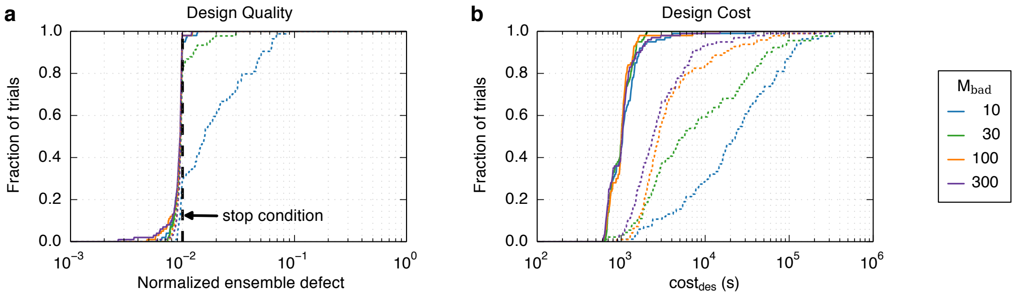

Figure S9. Sensitivity of algorithm performance to M_{bad} (default: 300). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

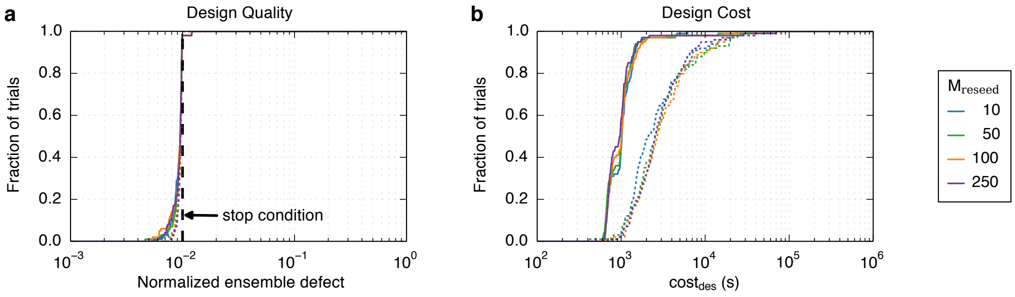

Figure S10. Sensitivity of algorithm performance to M_{reseed} (default: 50). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

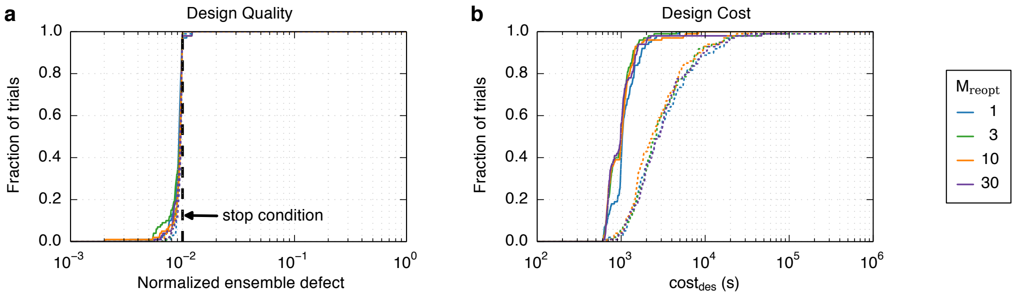

Figure S11. Sensitivity of algorithm performance to M_{reopt} (default: 3). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

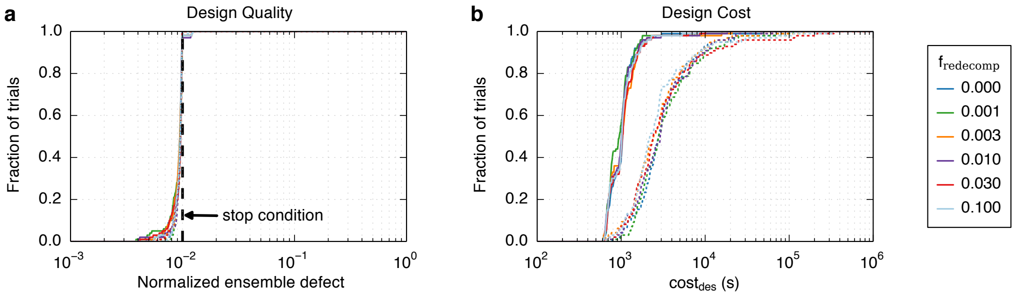

Figure S12. Sensitivity of algorithm performance to f_{redecomp} (default: 0.03). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

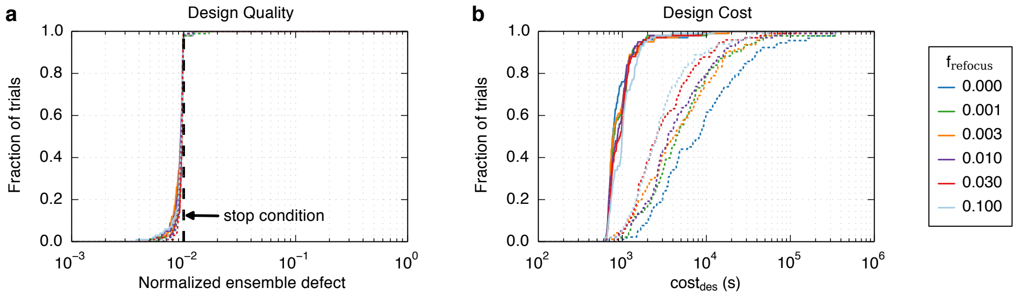

Figure S13. Sensitivity of algorithm performance to f_{refocus} (default: 0.03). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. RNA design at 37 °C for the subsets of the engineered (solid lines) and random (dotted lines) test sets with 200-nt on-targets.

S2.2 Complex Design

Here, we examine algorithm performance for the special case in which test tube design reduces to complex design: a target test tube containing one on-target complex and no off-target complexes. Figures S14 and S15 demonstrate that the performance of the current algorithm and the previously published single-complex design algorithm[5] is similar for the (dimer) on-target structures in the engineered and random test sets, respectively.

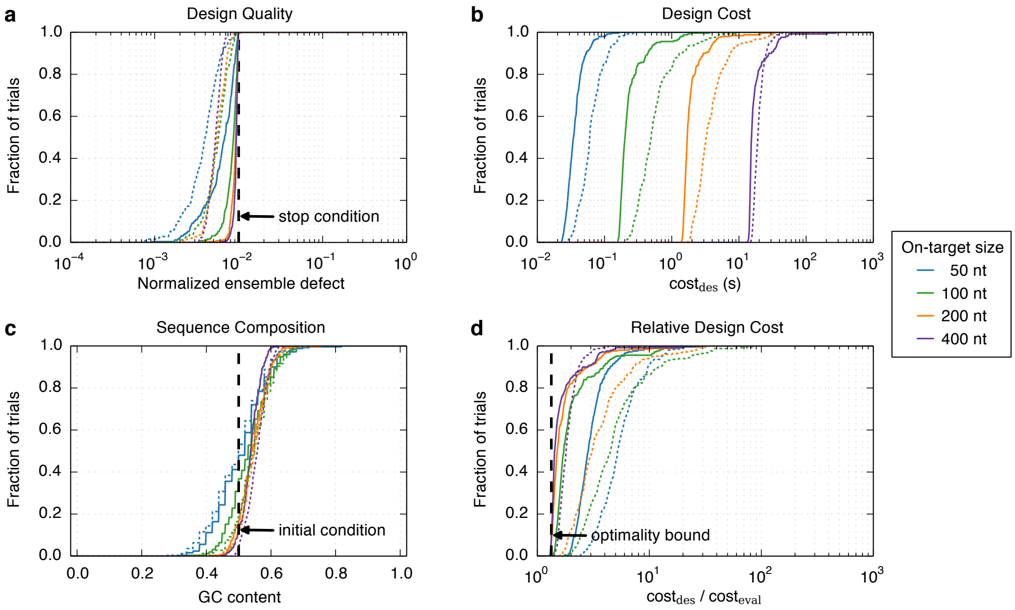

Figure S14. Algorithm performance for complex design using on-target structures from the engineered test set. Comparison of the current test tube design algorithm (solid lines) to the previously published single-complex design algorithm[5] (dotted lines). a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. c) Sequence composition. The initial GC content is depicted as a dashed black line. d) Cost of sequence design relative to a single evaluation of the objective function. The optimality bound[5] is depicted as a dashed black line. RNA design at 37 °C. Each tube contains a single on-target dimer and no off-targets. There are 50 target structures for each on-target size.

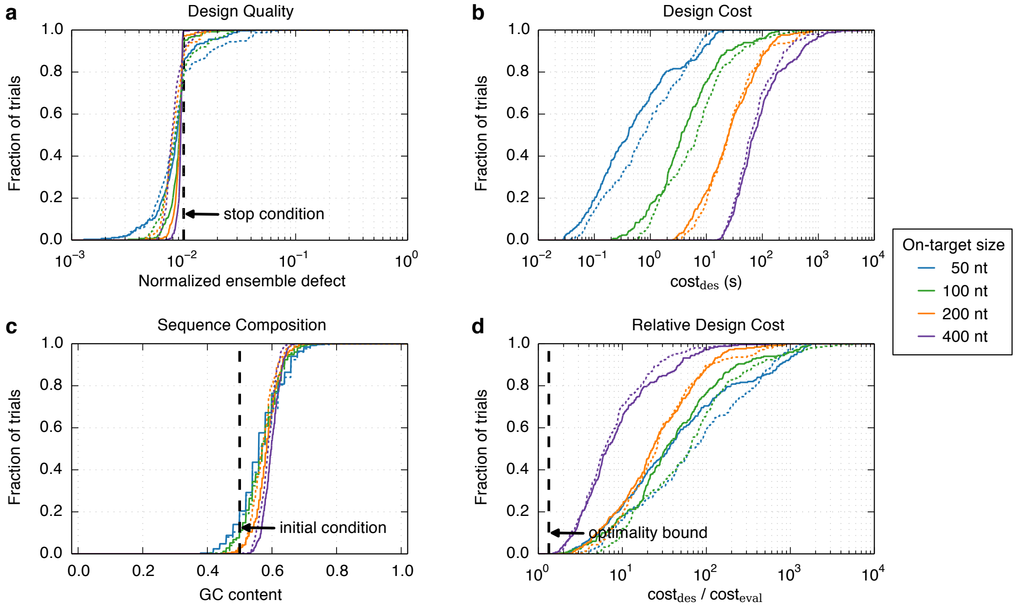

Figure S15. Algorithm performance for complex design using on-target structures from the random test set. Comparison of the current test tube design algorithm (solid lines) to the previously published single-complex design algorithm[5] (dotted lines). a) Design quality. The stop condition is depicted as a dashed line. b) Design cost. c) Sequence composition. The initial GC content is depicted as a dashed black line. d) Cost of sequence design relative to a single evaluation of the objective function. The optimality bound[5] is depicted as a dashed black line. RNA design at 37 °C. Each tube contains a single on-target dimer and no off-targets. There are 50 target structures for each on-target size.

S2.3 Sequence Initialization and Reseeding

Figure S16 compares algorithm performance using different GC contents for random sequence initialization and reseeding. Sequences are initialized/reseeded using either: random sequences (default), random sequences using only AU, or random sequences using only GC. The desired design quality is achieved independent of the initial GC content (panel a), with the typical design cost increasing marginally if the initial sequence contains only GC (panel b). Designs initialized using only random AU or only random GC illustrate that the desired design quality can be achieved over a broad range of GC contents (panel c).

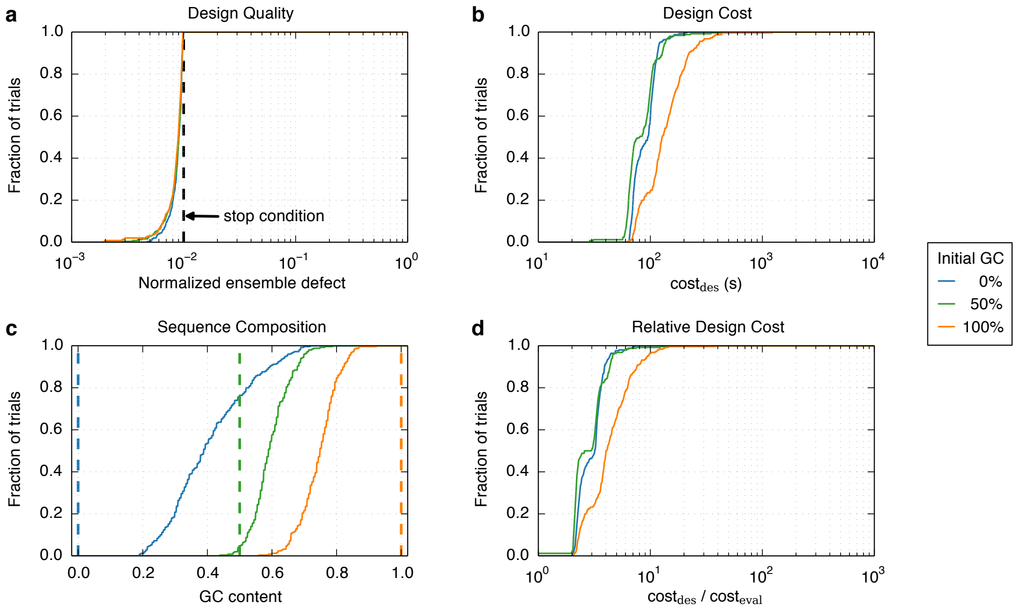

Figure S16. Effect of sequence initialization/reseeding on algorithm performance. a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. c) Sequence composition. The GC contents used for seeding/reseeding are depicted as dashed colored lines. d) Cost of sequence design relative to a single evaluation of the objective function. RNA design at 37 °C for the subset of the engineered test set with 100-nt on-targets.

S2.4 RNA vs DNA Design

Figure S17 compares RNA and DNA design. DNA designs are performed in 1 M Na^+ at 25 °C to reflect that DNA systems are typically engineered for room temperature studies. In comparison to RNA design, DNA design leads to similar design quality (panel a), design cost (panels b and d), and GC content (panel c).

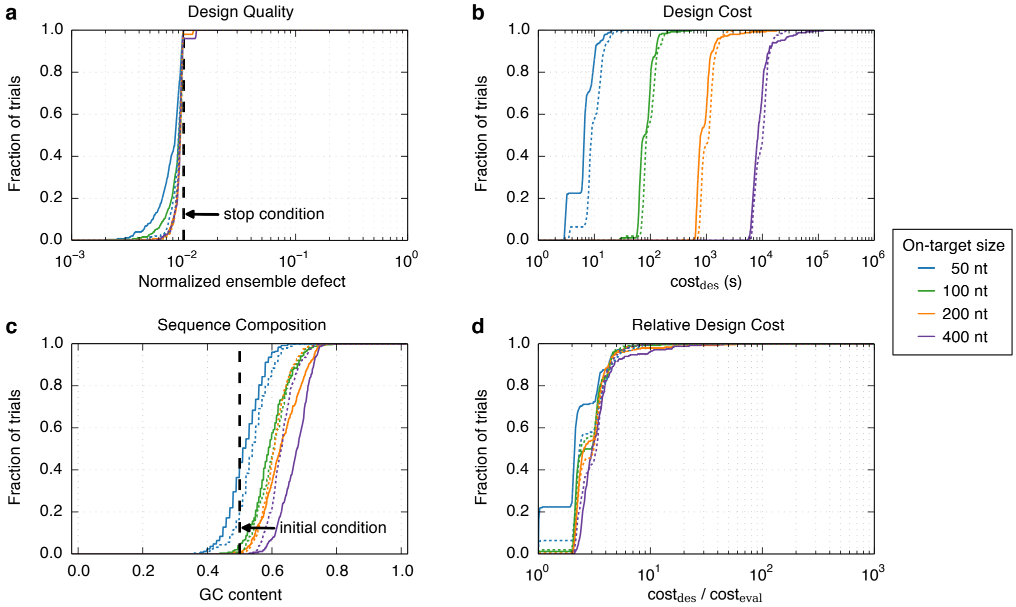

Figure S17. Effect of design material on algorithm performance. a) Design quality. The stop condition is depicted as a dashed black line. b) Design cost. c) Sequence composition. The initial GC content is depicted as a dashed black line. d) Cost of sequence design relative to a single evaluation of the objective function. RNA design at 37 °C (solid lines) and DNA design at 25°C (dotted lines) for the engineered test set.

S3 Structural Features of the Engineered and Random Test Sets

For the engineered test set, each dimer on-target structure was randomly generated with stem and loop sizes randomly selected from a distribution of sizes representative of the nucleic acid engineering literature. For the random test set, each dimer on-target structure was generated by calculating the minimum free energy structure of a different random pair of RNA sequences at 37 °C. Target structures were selected so that the minimum cut (the number of intermolecular base pairs that must be broken to dissociate the dimer strands) was at least 9 bp for the engineered test set and at least 7 bp for the random test set. The structural properties of the on-target structures in the engineered and random test sets are summarized in Figure S18. Typically, the random test set contains on-target structures with a lower fraction of paired nucleotides (panel a), more stems (panel b), shorter stems (panel c), and a higher minimum cut (panel d). The loop composition of each test set is summarized in Figure S19. The on-target structures for the engineered test set, the random test set, and the *big-tube test set* used for Figure 10 are provided as supplementary text files.

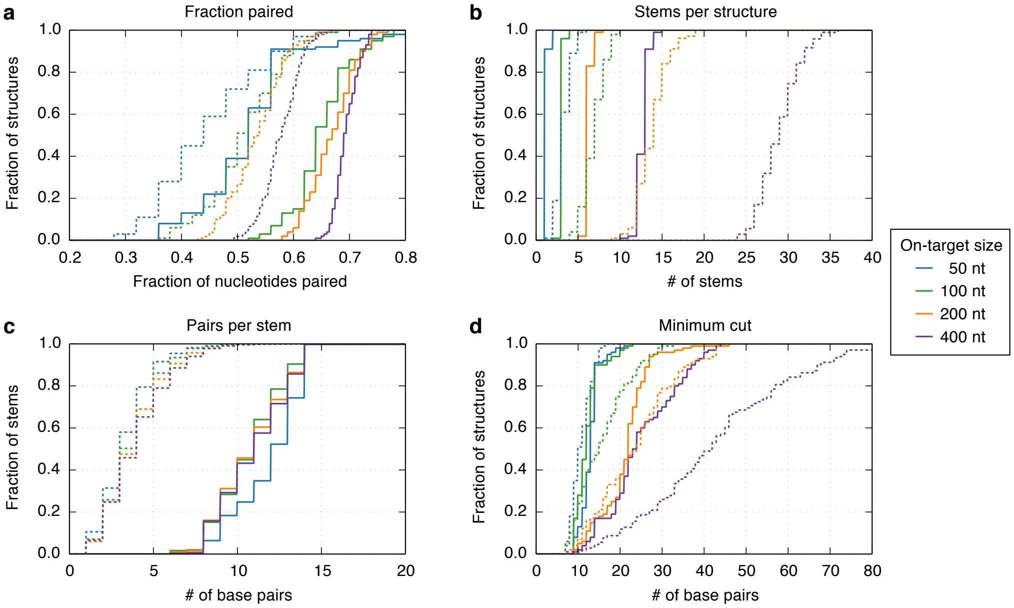

Figure S18. Structural features of the (dimer) on-target structures in the engineered (solid lines) and random (dotted lines) test sets. a) Fraction of nucleotides paired. b) Number of duplex stems per structure. c) Number base pairs per stem. d) Minimum number of base pairs that must be cut to dissociate the dimer strands.

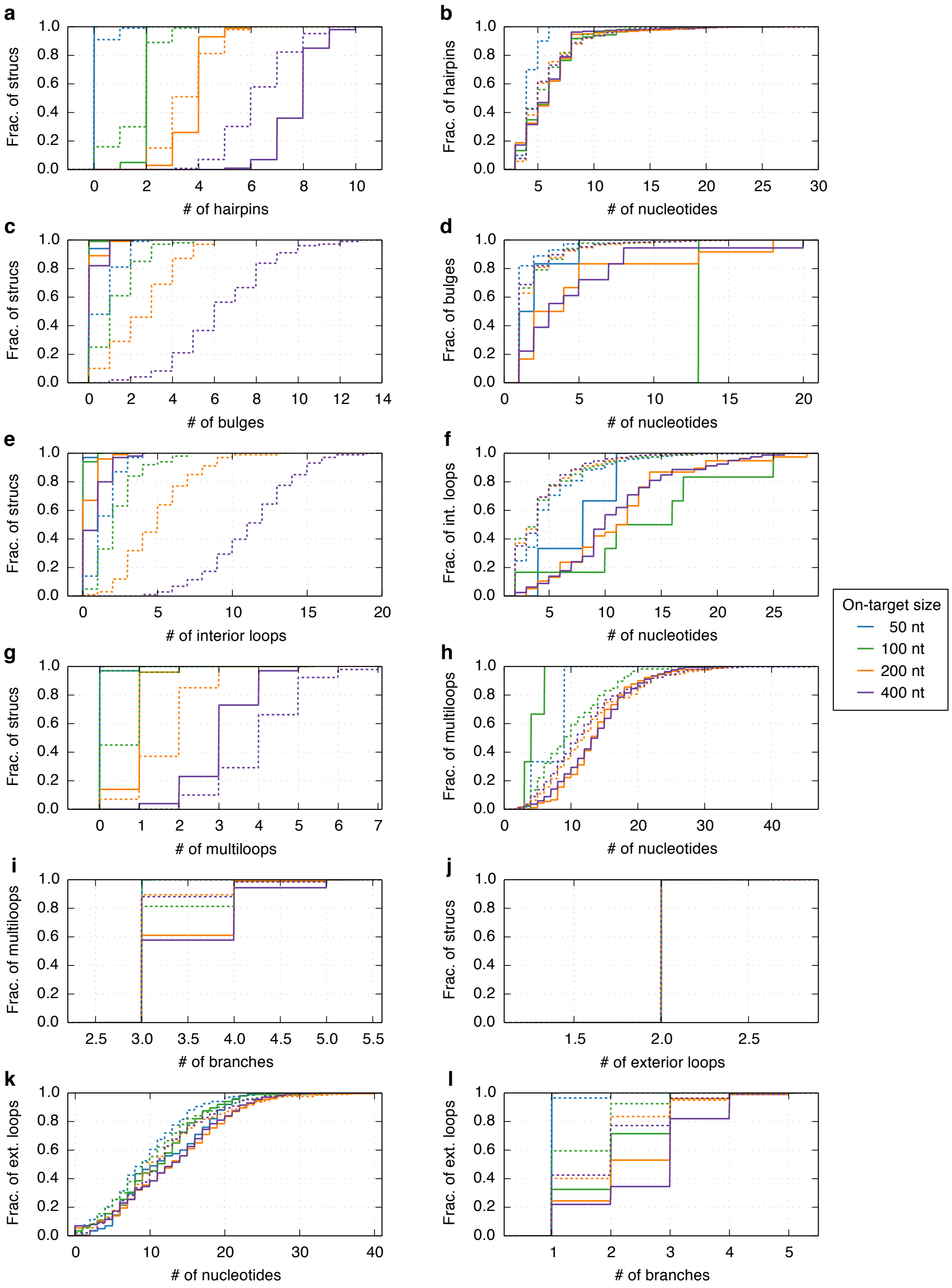

Figure S19. Loop composition of the (dimer) on-target structures in the engineered (solid lines) and random (dotted lines) test sets. a) Number of hairpin loops per structure. b) Number of unpaired nucleotides in each hairpin loop. c) Number of bulge interior loops per structure. d) Number of unpaired nucleotides in each bulge loop. e) Number of non-bulge interior loops per structure. f) Number of unpaired nucleotides in each non-bulge interior loop. g) Number of multiloops per structure. h) Number of unpaired nucleotides in each multiloop. i) Number of branches in each multiloop. j) Number of exterior loops per structure (always 2 since all are dimers). k) Number of unpaired nucleotides in each exterior loop. l) Number of branches in each exterior loop.

References

[1] Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E., and Pierce, N. A. (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.* 49, 65–88.

[2] Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288, 911–940.

[3] Mathews, D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* 10, 1178–1190.

[4] SantaLucia, J., Jr., and Hicks, D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* 33, 415–440.

[5] Zadeh, J. N., Wolfe, B. R., and Pierce, N. A. (2011) Nucleic acid sequence design via efficient ensemble defect optimization. *J. Comput. Chem.* 32, 439–452.