Controlling Uncertainty in Aptamer Selection

Supporting Information, Appendix

Fabian Spill, Zohar B. Weinstein, Atena Irani Shemirani, Nga Ho, Darash Desai, Muhammad H. Zaman

1 Derivation of the Model

The model introduced in the main text essentially describes the following set of chemical reactions:

$$\begin{aligned} A_i + T &\underset{k_i^-}{\overset{k_i^+}{\rightleftharpoons}} [A_iT], \\ A_i + S &\underset{k_S^-}{\overset{k_S^+}{\rightleftharpoons}} [A_iS], \end{aligned} \tag{S1}$$

Here, $k_i^{\pm}$ are the forward and backward rates for aptamer A_i and target T complex association and dissociation, and $k_S^{\pm}$ the corresponding rates for binding of aptamers with the substrate S. Furthermore, as in the main text, $i = 1, \ldots, M^B$ denotes a label enumerating the different aptamers (unique aptamers, or aptamers with similar rate constants binned together). There are several mathematical techniques which can be used to describe such reactions. The most common is to use ordinary differential equations describing the time-evolution of the concentrations A_i, T and S of aptamers, target, and substrate, respectively, using the law of mass-action [1]. The differential equations corresponding to reactions (S1) are given by

$$\begin{aligned} \frac{dA_i}{dt} &= -k_i^+ A_i T - k_S^+ A_i S + k_i^- [A_iT] + k_S^- [A_iS], \\ \frac{d[A_iT]}{dt} &= k_i^+ A_i T - k_i^- [A_iT], \\ \frac{dT}{dt} &= -k_i^+ A_i T + k_i^- [A_iT], \\ \frac{d[A_iS]}{dt} &= k_S^+ A_i S - k_S^- [A_iS], \\ \frac{dS}{dt} &= -k_S^+ A_i S + k_S^- [A_iS]. \end{aligned} \tag{S2}$$

While these equations describe a dynamic system, SELEX experiments are typically performed at timescales where chemical equilibrium is reached. The timescale to reach equilibrium can be on the order of hours [2, 3, 4], but this depends strongly on ligand pool design, substrate and target used, and the involved concentrations. For instance, nucleic acid hybridization, which forms the basis of substrate-ligand binding in [2], is dependent on conditions such as temperature and salt concentration. Accordingly, equilibrium may require up to a day to reach [5]. Assuming experiments are performed under equilibrium conditions, Eqs. (S2) can be approximated by its steady state solution. If $T^{tot} = \sum_{i=1}^{M^B} [TA_i] + T^{free}$ denotes the total concentration of target molecules, $S^{tot} = \sum_{i=1}^{M^B} [SA_i] + S^{free}$ denotes the total concentration of substrate, and A_i^I is the total concentration of aptamer i present before the reactions (S1) occur, then the steady-state equations are given by

$$\begin{aligned} [SA_i] &= \frac{1}{K_S}(A_i^I - [SA_i] - [TA_i])S^{free}, \\ [TA_i] &= \frac{1}{K_{D,i}}(A_i^I - [SA_i] - [TA_i])T^{free}, \\ S^{tot} &= \sum_{i=1}^{M^B} [SA_i] + S^{free}, \\ T^{tot} &= \sum_{i=1}^{M^B} [TA_i] + T^{free}, \\ i &= 1, \ldots, M^B. \end{aligned} \tag{S3}$$

These are the same equations as given in Eq. (1) in the main text. The dissociation constants are obtained from the forward and backward rates through

$$K_{D,i} = \frac{k_i^-}{k_i^+} \tag{S4}$$

$$K_S = \frac{k_S^-}{k_S^+} \tag{S5}$$

Notice that while Eqs. (S2) are structurally simple equations, they are quadratic in the $2M^B$ unknowns $[A_iS]$ and $[A_iT]$, since S^{free} and T^{free} also depend on $[A_iS]$ and $[A_iT]$, respectively. Now, as we argued in the main text, neither the steady-state Eqs. (S3) nor their dynamic analogues Eqs. (S2) are expected to be valid when the number of molecules is very small. Certainly, if there is only one aptamer of type i present, then a deterministic model does not apply. We thus need to describe reactions involving molecules in low copy number with a stochastic approach. We will employ the chemical master equation [6, 7] to describe those reactions. First of all, we note that from a theoretical perspective, it will be straight-forward to derive a master equation for the whole system of reactions given by (S1). However, such a system will not be solvable analytically, nor can it be simulated with standard techniques, such as the Gillespie algorithm [8], due to the large number of involved molecules. We thus treat only those aptamers present in low copy number stochastically, whereas those aptamers present in large copy numbers as well as the target and the substrate concentrations are treated deterministically; their concentrations are obtained by Eqs. (S3). Moreover, since we only treat those aptamers present in low copy numbers stochastically, as long as the free target and substrate concentrations predicted by the deterministic model are sufficiently larger than the sum of the stochastic aptamers, the outcome of the stochastic binding of those aptamers will not significantly alter the free target and substrate concentrations. However, this implies that all of the stochastic reactions decouple and can be treated independently. Thus, we consider, without loss of generality, that only one kind of aptamer is present in low copy number, and let this be aptamer $i = 1$.

As in the main text, let $\tilde{A}_1$ denote the number of aptamers of type $i = 1$. Now, since we model an experimental approach in which S and T are present in concentrations of nM or more, this translates into billions of molecules of T and S. Likewise, we assumed that the other aptamers for $i = 2, \ldots, M^B$ are present in much higher copy numbers, so that the deterministic steady-state Eqs. (S3) remain valid for $i = 2, \ldots, M^B$. We can thus consider the reaction of aptamer $i = 1$ with either a free target molecule T^{free} or a free substrate molecule S^{free}, where those free concentrations are given by Eq. (S3) for $i = 2, \ldots, M^B$. We thus only need to model the numbers $\tilde{A}_1$, $[\tilde{A}_1\tilde{T}]$ and $[\tilde{A}_1\tilde{S}]$ of free aptamer, aptamer-target and aptamer-substrate complexes, respectively. The joint probability $p(\tilde{A}_1, [\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}])$ to have $\tilde{A}_1$, $[\tilde{A}_1\tilde{T}]$ and $[\tilde{A}_1\tilde{S}]$ present in the system is given by the master equation

$$\begin{aligned}
\frac{dp(\tilde{A}_1, [\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}])}{dt} &= k_1^+ \frac{1}{N_A V}\left((\tilde{A}_1+1)(\tilde{T}^{free}+1)p(\tilde{A}_1+1, [\tilde{A}_1\tilde{T}]-1, [\tilde{A}_1\tilde{S}]) - \tilde{A}_1\tilde{T}^{free}p(\tilde{A}_1, [\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}])\right) \\
&+ k_1^-\left([\tilde{A}_1\tilde{T}]+1)p(\tilde{A}_1-1, [\tilde{A}_1\tilde{T}]+1, [\tilde{A}_1\tilde{S}]) - [\tilde{A}_1\tilde{T}]p(\tilde{A}_1, [\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}])\right) \\
&+ k_S^+ \frac{1}{N_A V}\left((\tilde{A}_1+1)(\tilde{S}^{free}+1)p(\tilde{A}_1+1, [\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}]-1) - \tilde{A}_1\tilde{S}^{free}p(\tilde{A}_1, [\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}])\right) \\
&+ k_S^-\left([\tilde{A}_1\tilde{S}]+1)p(\tilde{A}_1-1, [\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}]+1) - [\tilde{A}_1\tilde{S}])p(\tilde{A}_1, [\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}])\right).
\end{aligned} \tag{S6}$$

Here, $\tilde{T}^{free}$ and $\tilde{S}^{free}$ denote the number of free target and substrate molecules. Since we consider the other aptamers for $i > 2$ to be in deterministic steady state, $\tilde{T}^{free}$ and $\tilde{S}^{free}$ are fixed. Furthermore, since we assume that target and substrate concentration are in the nM range, it is safe to assume that $\tilde{T}^{free} \gg 1$ and $\tilde{S}^{free} \gg 1$. Thus, we have $\tilde{T}^{free}+1 \approx \tilde{T}^{free}$ and $\tilde{S}^{free}+1 \approx \tilde{S}^{free}$. Then, we can absorb the factor $\frac{1}{N_A V}$ into the two forward reactions, so we will introduce the concentrations $T^{free} = \frac{\tilde{T}^{free}}{N_A V}$ and $S^{free} = \frac{\tilde{S}^{free}}{N_A V}$, respectively. Next, we notice that since $\tilde{A}_1 + [\tilde{A}_1\tilde{T}] + [\tilde{A}_1\tilde{S}] = \tilde{A}_1^I$, which is a constant in each round, only two of the three variable in the master equation are independent. Furthermore, the number of aptamers which are selected into the next round is the sum of the free ones and the aptamer-target complexes,

$$\tilde{A}_1^{S,D} = [T\tilde{A}_1] + \tilde{A}_1 = \tilde{A}_1^I - [S\tilde{A}_1]. \tag{S7}$$

This is the same as Eq. (2) for concentrations given in the main text. Thus, it is sufficient if we can calculate the probability $p([\tilde{A}_1\tilde{S}])$, which is obtained as a marginal distribution from $p(\tilde{A}_1, [\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}])$. We can also identify $\tilde{A}_1$ with the number of free, unbound aptamers $\tilde{A}_1^{free}$. We thus have

$$\begin{aligned}
p\left([\tilde{A}_1\tilde{S}]\right) &= \sum_{[\tilde{A}_1\tilde{T}]=0}^{\tilde{A}_1^I} p\left(\tilde{A}_1 = \tilde{A}_1^I - [\tilde{A}_1\tilde{T}] - [\tilde{A}_1\tilde{S}], [\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}]\right), \\
&= \sum_{[\tilde{A}_1\tilde{T}]=0}^{\tilde{A}_1^I} p\left([\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}]\right).
\end{aligned} \tag{S8}$$

In the last step, we simply eliminated $\tilde{A}_1$ since it is fully determined by Eq. (S7). We thus plug Eq. (S8) into Eq. (S6) and

obtain the master equation for $p\left([\tilde{A}_1\tilde{S}]\right)$:

$$\frac{dp\left([\tilde{A}_1\tilde{S}]\right)}{dt} = \sum_{[\tilde{A}_1\tilde{T}]=0}^{\tilde{A}_1^I} \left(k_1^+\left((\tilde{A}_1+1)T^{free}p\left([\tilde{A}_1\tilde{T}]-1,[\tilde{A}_1\tilde{S}]\right)-\tilde{A}_1T^{free}p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]\right)\right)\right. \tag{S9}$$

$$+k_1^-\left(([\tilde{A}_1\tilde{T}]+1)p\left([\tilde{A}_1\tilde{T}]+1,[\tilde{A}_1\tilde{S}]\right)-[\tilde{A}_1\tilde{T}]p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]\right)\right) \tag{S10}$$

$$+k_S^+\left((\tilde{A}_1+1)S^{free}p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]-1\right)-\tilde{A}_1S^{free}p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]\right)\right) \tag{S11}$$

$$\left.+\ k_S^-\left(([\tilde{A}_1\tilde{S}]+1)p([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]+1)-[\tilde{A}_1\tilde{S}]p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]\right)\right)\right) \tag{S12}$$

We now calculate the four contributions to the equation above. Eq. (S9) is given by:

$$\begin{aligned}
&k_1^+T^{free}\sum_{[\tilde{A}_1\tilde{T}]=0}^{\tilde{A}_1^I}\left((\tilde{A}_1+1)p\left([\tilde{A}_1\tilde{T}]-1,[\tilde{A}_1\tilde{S}]\right)-\tilde{A}_1p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]\right)\right)\\
=&k_1^+T^{free}\sum_{[\tilde{A}_1\tilde{T}]=0}^{\tilde{A}_1^I}\left((\tilde{A}_1^I-[\tilde{A}_1\tilde{T}]-[\tilde{A}_1\tilde{S}]+1)p\left([\tilde{A}_1\tilde{T}]-1,[\tilde{A}_1\tilde{S}]\right)-(\tilde{A}_1^I-[\tilde{A}_1\tilde{T}]-[\tilde{A}_1\tilde{S}])p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]\right)\right)\\
=&k_1^+T^{free}\left(\tilde{A}_1^I-\tilde{A}_1^I-[\tilde{A}_1\tilde{S}]\right)p\left([\tilde{A}_1\tilde{T}]=\tilde{A}_1^I,[\tilde{A}_1\tilde{S}]\right)\\
=&0
\end{aligned} \tag{S13}$$

In the first step, we find that this is a telescoping sum, so only the term with $[\tilde{A}_1\tilde{T}]=\tilde{A}_1^I$ survives. Finally, if $[\tilde{A}_1\tilde{T}]=\tilde{A}_1^I$, this necessitates $[\tilde{A}_1\tilde{S}]=0$ and the last term vanishes. Next, Eq. (S10) gives:

$$\begin{aligned}
&k_1^-\sum_{[\tilde{A}_1\tilde{T}]=0}^{\tilde{A}_1^I}\left(([\tilde{A}_1\tilde{T}]+1)p\left([\tilde{A}_1\tilde{T}]+1,[\tilde{A}_1\tilde{S}]\right)-[\tilde{A}_1\tilde{T}]p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]\right)\right)\\
&=-k_1^-\times 0\times p\left([\tilde{A}_1\tilde{T}]=0,[\tilde{A}_1\tilde{S}]\right)\\
&=0
\end{aligned} \tag{S14}$$

This is again a telescoping sum with the term $[\tilde{A}_1\tilde{T}]=0$ surviving. However, this term is multiplied by $[\tilde{A}_1\tilde{T}]$ and thus vanishes. Next, Eq. (S11) gives:

$$\begin{aligned}
&k_S^+S^{free}\sum_{[\tilde{A}_1\tilde{T}]=0}^{\tilde{A}_1^I}\left((\tilde{A}_1+1)p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]-1\right)-\tilde{A}_1p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]\right)\right)\\
=&k_S^+S^{free}\sum_{[\tilde{A}_1\tilde{T}]=0}^{\tilde{A}_1^I}\left((\tilde{A}_1^I-[\tilde{A}_1\tilde{T}]-[\tilde{A}_1\tilde{S}]+1)p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]-1\right)-(\tilde{A}_1^I-[\tilde{A}_1\tilde{T}]-[\tilde{A}_1\tilde{S}])p\left([\tilde{A}_1\tilde{T}],[\tilde{A}_1\tilde{S}]\right)\right)\\
=&k_S^+S^{free}\left((\tilde{A}_1^I-E\left([\tilde{A}_1\tilde{T}]\mid[\tilde{A}_1\tilde{S}]-1\right)-[\tilde{A}_1\tilde{S}]+1)p\left([\tilde{A}_1\tilde{S}]-1\right)-(\tilde{A}_1^I-E\left([\tilde{A}_1\tilde{T}]\mid[\tilde{A}_1\tilde{S}]\right)-[\tilde{A}_1\tilde{S}])p\left([\tilde{A}_1\tilde{S}]\right)\right)
\end{aligned} \tag{S15}$$

Here, we have introduced the conditional expectation

$$E\left([\tilde{A}_1\tilde{T}]\mid[\tilde{A}_1\tilde{S}]\right)=\sum_{[\tilde{A}_1\tilde{T}]=0}^{\tilde{A}_1^I}[\tilde{A}_1\tilde{T}]p\left([\tilde{A}_1\tilde{T}]\mid[\tilde{A}_1\tilde{S}]\right). \tag{S16}$$

We thus need to find the conditional probability $p([\tilde{A}_1\tilde{T}]\mid[\tilde{A}_1\tilde{S}])$. However, this probability can be simply found by noticing that when $[\tilde{A}_1\tilde{S}]$ is given, then only the reaction $A_1+T\underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}}[A_1T]$ matters. The total number of aptamer molecules for this reaction is now not $\tilde{A}_1^I$, but $\tilde{A}_1^I-[\tilde{A}_1\tilde{S}]$, since $[\tilde{A}_1\tilde{S}]$ aptamers are bound in a complex with the substrate. The master

equation for $p([\tilde{A}_1\tilde{T}] \mid [\tilde{A}_1\tilde{S}])$ is given by

$$\begin{aligned}\frac{\partial p\left([\tilde{A}_1\tilde{T}] \mid [\tilde{A}_1\tilde{S}]\right)}{\partial t} &= k_1^+ T^{free}\left((\tilde{A}_1+1)p\left([\tilde{A}_1\tilde{T}]-1 \mid [\tilde{A}_1\tilde{S}]\right) - \tilde{A}_1 p\left([\tilde{A}_1\tilde{T}] \mid [\tilde{A}_1\tilde{S}]\right)\right)\\ &+ k_1^-\left(([\tilde{A}_1\tilde{T}]+1)p\left([\tilde{A}_1\tilde{T}]+1 \mid [\tilde{A}_1\tilde{S}]\right) - [\tilde{A}_1\tilde{T}]p\left([\tilde{A}_1\tilde{T}] \mid [\tilde{A}_1\tilde{S}]\right)\right)\\ &= k_1^+ T^{free}\left((\tilde{A}_1^I - [\tilde{A}_1\tilde{S}] - [\tilde{A}_1\tilde{T}]+1)p\left([\tilde{A}_1\tilde{T}]-1 \mid [\tilde{A}_1\tilde{S}]\right) - (\tilde{A}_1^I - [\tilde{A}_1\tilde{S}] - [\tilde{A}_1\tilde{T}])p\left([\tilde{A}_1\tilde{T}] \mid [\tilde{A}_1\tilde{S}]\right)\right)\\ &+ k_1^-\left(([\tilde{A}_1\tilde{T}]+1)p\left([\tilde{A}_1\tilde{T}]+1 \mid [\tilde{A}_1\tilde{S}]\right) - [\tilde{A}_1\tilde{T}]p\left([\tilde{A}_1\tilde{T}] \mid [\tilde{A}_1\tilde{S}]\right)\right).\end{aligned} \tag{S17}$$

This is a simple master equation with all coefficients linear in the unknown $[\tilde{A}_1\tilde{T}]$. It follows that in the steady state distribution $\frac{\partial p([\tilde{A}_1\tilde{T}]|[\tilde{A}_1\tilde{S}])}{\partial t} = 0$, the distribution is binomial:

$$\begin{aligned}p\left([\tilde{A}_1\tilde{T}] \mid [\tilde{A}_1\tilde{S}]\right) &\sim B\left(\tilde{A}_1^I - [\tilde{A}_1\tilde{S}], \frac{k_1^+ T^{free}}{k_1^+ T^{free} + k_1^-}\right)\\ &= B\left(\tilde{A}_1^I - [\tilde{A}_1\tilde{S}], \frac{T^{free}}{T^{free} + K_{D1}}\right)\end{aligned} \tag{S18}$$

Thus, we get for the conditional expectation value

$$E\left([\tilde{A}_1\tilde{T}] \mid [\tilde{A}_1\tilde{S}]\right) = \left(\tilde{A}_1^I - [\tilde{A}_1\tilde{S}]\right)\frac{T^{free}}{T^{free} + K_{D1}}. \tag{S19}$$

Finally, we get for Eq. (S12):

$$\begin{aligned}&k_S^- \sum_{[\tilde{A}_1\tilde{T}]=0}^{\tilde{A}_1^I}\left(([\tilde{A}_1\tilde{S}]+1)p\left([\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}]+1\right) - [\tilde{A}_1\tilde{S}]p\left([\tilde{A}_1\tilde{T}], [\tilde{A}_1\tilde{S}]\right)\right)\\ =&k_S^-\left(([\tilde{A}_1\tilde{S}]+1)p\left([\tilde{A}_1\tilde{S}]+1\right) - [\tilde{A}_1\tilde{S}]p\left([\tilde{A}_1\tilde{S}]\right)\right)\end{aligned} \tag{S20}$$

Collecting all terms together, we get the master equation for the marginal distribution $p\left([\tilde{A}_1\tilde{S}]\right)$

$$\begin{aligned}\frac{dp\left([\tilde{A}_1\tilde{S}]\right)}{dt} &= k_S^+ S^{free}\left((\tilde{A}_1^I - \left(\tilde{A}_1^I - [\tilde{A}_1\tilde{S}]+1\right)\frac{T^{free}}{T^{free}+K_{D1}} - [\tilde{A}_1\tilde{S}]+1)p\left([\tilde{A}_1\tilde{S}]-1\right)\right.\\ &\left. - \left(\tilde{A}_1^I - \left(\tilde{A}_1^I - [\tilde{A}_1\tilde{S}]\right)\frac{T^{free}}{T^{free}+K_{D1}} - [\tilde{A}_1\tilde{S}]\right)p\left([\tilde{A}_1\tilde{S}]\right)\right)\\ &+ k_S^-\left(\left([\tilde{A}_1\tilde{S}]+1\right)p\left([\tilde{A}_1\tilde{S}]+1\right) - [\tilde{A}_1\tilde{S}]p\left([\tilde{A}_1\tilde{S}]\right)\right).\end{aligned} \tag{S21}$$

We observe that while Eq. (S21) has more complicated coefficients, it is structurally of the same type as Eq. (S17). Thus, the solution to the steady state problem is a Binomial distribution

$$\begin{aligned}p\left([\tilde{A}_1\tilde{S}]\right) &\sim B\left(\tilde{A}_1^I, \frac{k_S^+ k_1^- S^{free}}{k_S^+ k_1^- S^{free} + k_S^-(k_1^+ T^{free} + k_1^-)}\right)\\ &= B\left(\tilde{A}_1^I, \frac{\frac{K_{D1}}{T^{free}}}{\frac{K_{D1}}{T^{free}} + \frac{K_S}{S^{free}} + \frac{K_{D1}}{T^{free}}\frac{K_S}{S^{free}}}\right)\end{aligned} \tag{S22}$$

We have thus established the distribution of $[\tilde{A}_1\tilde{S}]$, given in terms of the number of aptamers $\tilde{A}_1^I$, which are present after the non-specific binding steps, the rate constants, and the concentration of free target and substrate T^{free} and S^{free}, which is obtained from solving the deterministic Eqs. (S3). We also note that if we solve the deterministic equations for $i = 1$ also, then those equations predict a steady state for $[SA_1]$ of

$$[SA_1] = \tilde{A}_1^I \frac{\frac{K_{D1}}{T^{free}}}{\frac{K_{D1}}{T^{free}} + \frac{K_S}{S^{free}} + \frac{K_{D1}}{T^{free}}\frac{K_S}{S^{free}}}, \tag{S23}$$

which is exactly the expected value of Eq. (S22). This suggests the following strategy to solve the full model, including the deterministically modeled aptamers for $i > 1$. We first solve the full deterministic Eqs. (S3) for all $1 \le i \le M^B$. Then, for

those aptamers which we model stochastically, i.e. those which are predicted to fall below the threshold condition given in the main text as $A_i^I - [SA_i] < \Theta$ for a given threshold Θ, we can simply simulate the specific binding reactions by drawing a single random number sampled from a binomial Eq. (S22). The parameter for this binomial is then obtained from the steady state solution Eq. (S23). Finally, non-specific losses such as washing contribute an additional probability to the total selection probability for a single aptamer. As such losses constitute independent events, the total selection probability for a given aptamer that incorporates both specific binding and non-specific loss is thus the product of their respective probabilities. This is reflected in Eq. (4) in the main text.

2 Model Implementation

We have implemented our model in Mathematica 10.3, Wolfram Research, Inc., Champaign, IL (2015). The deterministic Eqs. (S3) are solved with the FindRoot function of Mathematica. The impact of stochastic processes within the model are assessed through Monte Carlo simulations based upon the binomial probability distribution described by Eq. (3). For each simulation and each aptamer bin that is below the threshold Θ, a random number is generated using the RandomVariate function of Mathematica, which requires a probability distribution from which this variate should be obtained. We provide Eq. (3) as this distribution, and the parameter is obtained from the solution to the deterministic equation as in Eq. (4).

References

[1] Keener JP, Sneyd J (2009) *Mathematical Physiology: Cellular Physiology. I* (Springer, New York).

[2] Stoltenburg R, Nikolaus N, Strehlitz B (2012) Capture-SELEX: Selection of DNA Aptamers for Aminoglycoside Antibiotics. *Journal of Analytical Methods in Chemistry* 2012.

[3] Park JW, Tatavarty R, Kim DW, Jung HT, Gu MB (2012) Immobilization-free Screening of Aptamers assisted by Graphene Oxide. *Chemical Communications* 48:2071–2073.

[4] Nguyen VT, Kwon YS, Kim JH, Gu MB (2014) Multiple GO-SELEX for Efficient Screening of Flexible Aptamers. *Chemical Communications* 50:10513–10516.

[5] Stevens, PW, Henry, MR, Kelso, DM (1999) DNA hybridization on microparticles: determining capture-probe density and equilibrium dissociation constants. *Nucleic acids research* 27:1719–1727.

[6] Van Kampen NG (1992) *Stochastic Processes in Physics and Chemistry* (North Holland, Amsterdam) Vol. 1.

[7] Gardiner C (2010) *Stochastic Methods* (Springer, Berlin).

[8] Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics* 22:403–434.

Table S1: Default Parameters

Parameter	Symbol	Default Value
Number of Unique Aptamers	M^A	$4^{25} \approx 1.13 \times 10^{15}$
Number of Bins	M^B	100
Number of Cycles	C	20
Stochastic Model Threshold	Θ	100
Substrate Concentration Cycle 1	S^{tot}	$1.66 \times 10^{-6} M$
Substrate Concentration Cycle 2 Onward	S^{tot}	$1.66 \times 10^{-7} M$
Target Concentration	T^{tot}	$10^{-4} M$
Incubation Volume	V	$50 \mu l$
Aptamer-Substrate Dissociation	K_S	$10^{-12} M$
PCR Amplification Factor	α_{PCR}	50

Table S2: The Improved Protocol

Cycle Round	T^{tot}	K_S
1	$10^{-3} M$	$10^{-11} M$
2	$10^{-4} M$	$10^{-12} M$
3	$10^{-4} M$	$10^{-13} M$
4	$10^{-4} M$	$10^{-14} M$
5	$10^{-5} M$	$10^{-15} M$
6-10	$10^{-5} M$	$10^{-15} M$
11-20	$10^{-6} M$	$10^{-18} M$

Table S3: Alternative protocols used in Fig. S8. The fast K_S decrease protocol decreases T^{tot} similarly to the improved protocol, but K_S is decreased faster through the cycles. Likewise, the fast T^{tot} decrease protocol decreases K_S similarly to the improved protocol, but T^{tot} is decreased faster.

Cycle Round	fast K_S decrease	fast T^{tot} decrease
1	$10^{-12} M$	$10^{-4} M$
2	$10^{-14} M$	$10^{-6} M$
3	$10^{-16} M$	$10^{-8} M$
4	$10^{-18} M$	$10^{-8} M$
5	$10^{-20} M$	$10^{-8} M$
6-10	$10^{-20} M$	$10^{-10} M$
11-20	$10^{-22} M$	$10^{-12} M$

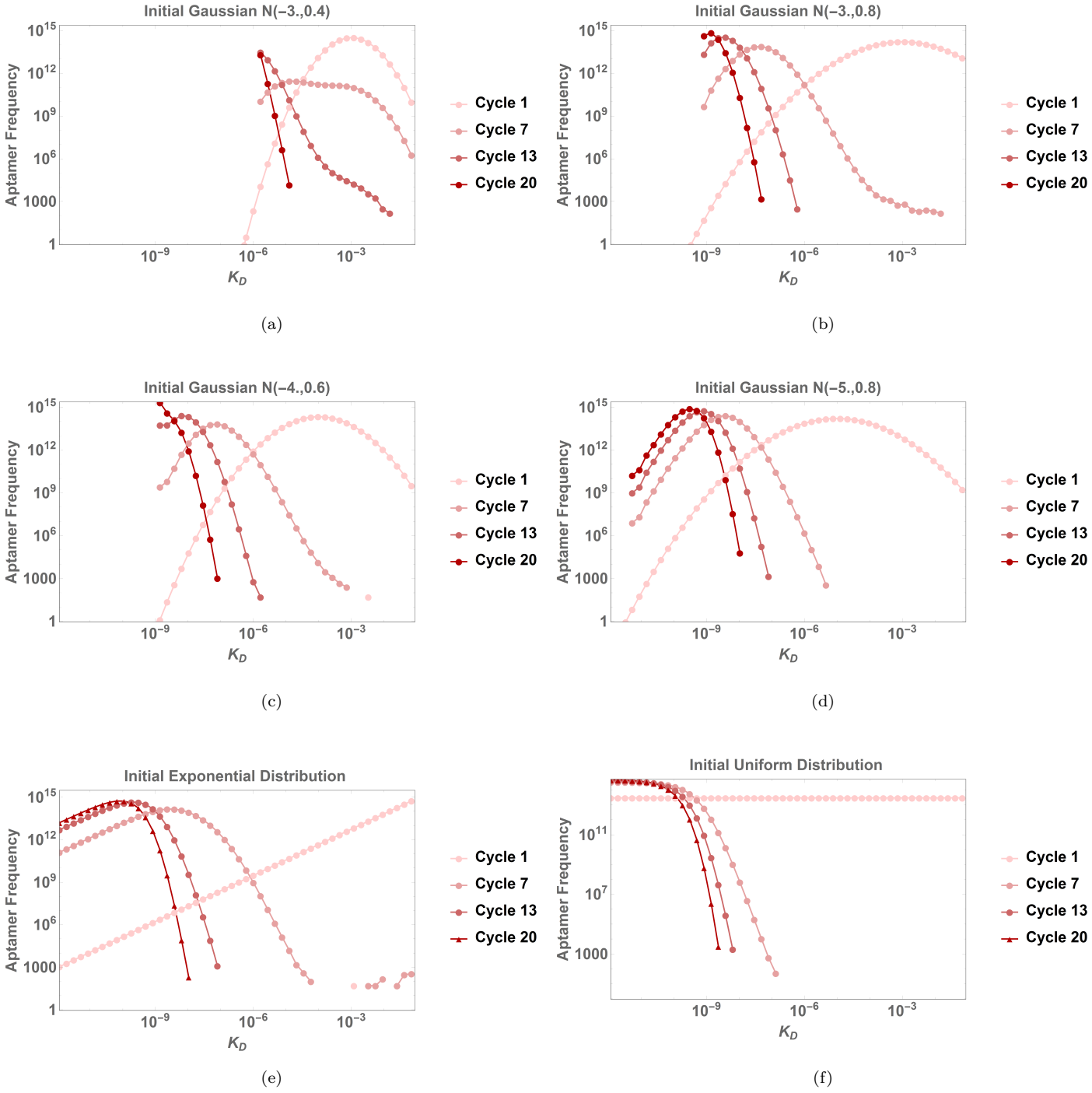

Figure S1: SELEX dynamics over 20 cycles for six initial K_D distributions: four Gaussians with different means and standard deviations, an exponential distribution, and one uniform distribution. While we certainly would not expect the real K_D distribution for targets of interest to be exponential or uniform, it is interesting to see the dynamics of SELEX for such distributions, and illustrates the point that the distribution has an important influence on selection efficiency. We see that the dynamics of evolution is quite different: in the uniform case, due to the large number of good binders, the bad ones are quickly removed. The exponential distribution (e) or the broad Gaussian (d) guarantee a sufficient number of good binders being present, but the protocol is not able to magnify the best binders (here, $K_D = 10^{-12}M$) and slightly worse binders still form the peak of the distribution. On the other hand, for the Gaussians (a)-(c), the protocol quickly selects the best binders present in the initial distribution.

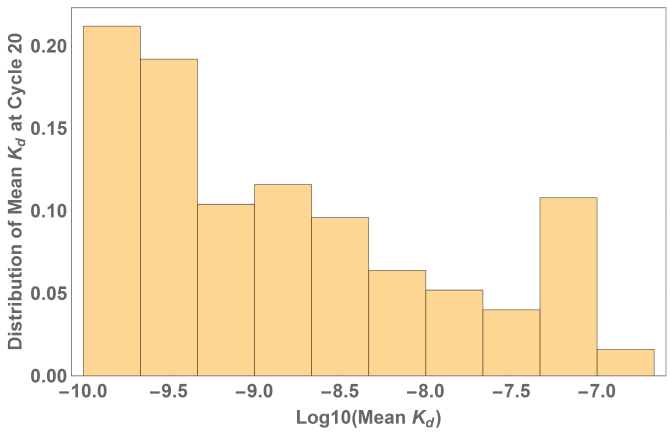

Figure S2: Plot showing the distribution of the mean K_D of aptamers present after cycle 20 for the same experimental condition presented in Fig. 3. The distribution reflects results from 250 identical Monte Carlo simulations. In general, we notice a decreasing distribution with increasing K_D, with a peak at the maximum value of $K_D = 10^{-10}M$. However, there is a second peak near $K_D = 10^{-7}M$. This peak corresponds to results similar to the first simulation in Fig. 3 (blue triangles), where all high-affinity aptamers that were introduced as noise outside the continuous Gaussian distribution are lost by chance.

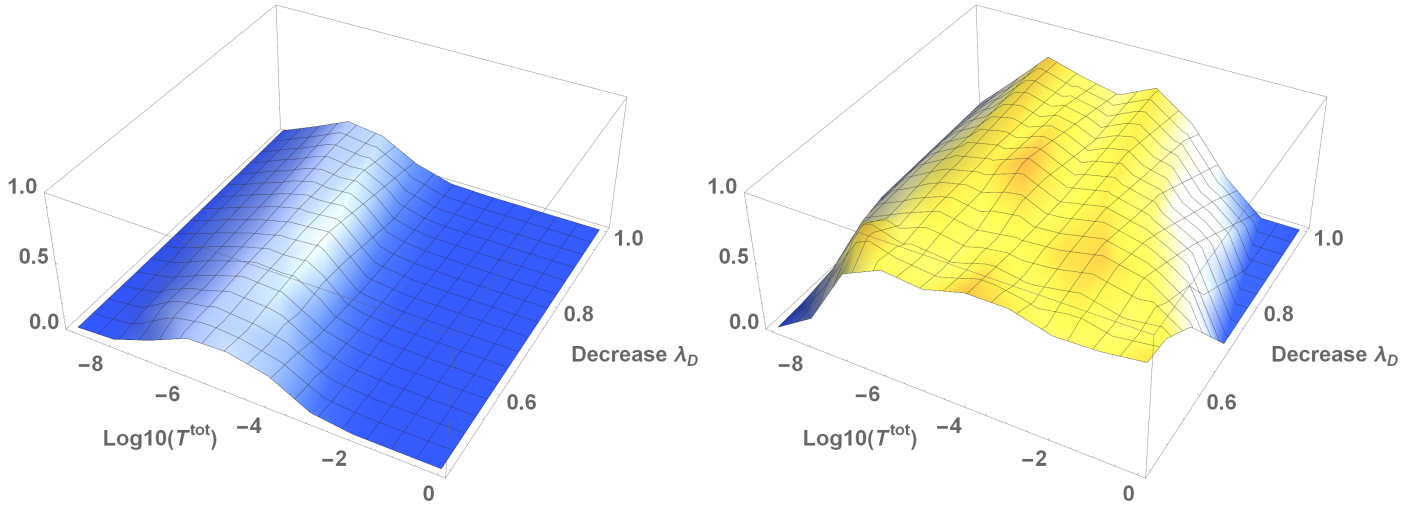

Figure S3: Snapshots of $\phi(c)$, the fraction of high-affinity aptamers with binding affinities stronger than $K_D = 10^{-10}M$, at two different cycles as a function of the initial target concentration T^{tot} and the factor λ_D, by which the target concentration is decreased with each cycle. We used an initial Gaussian distribution $N(-4, 0.8)$ and about 100 additional high-affinity aptamers. (a) At cycle $c = 8$, target concentrations close to $10^{-6}M$ are the first to yield an increase of high affinity aptamers, and faster decreases (lower λ_D) broaden the range of target concentrations which lead to strong binders. (b) At cycle $c = 20$, a wide range of target concentrations leads to strong binders, unless the initial concentration is too low ($T^{tot} < 10^{-7}M$). Very high concentrations ($T^{tot} > 10^{-1}M$) can still lead to success, provided the concentration is decreased sufficiently fast ($\lambda_D < 0.6$).

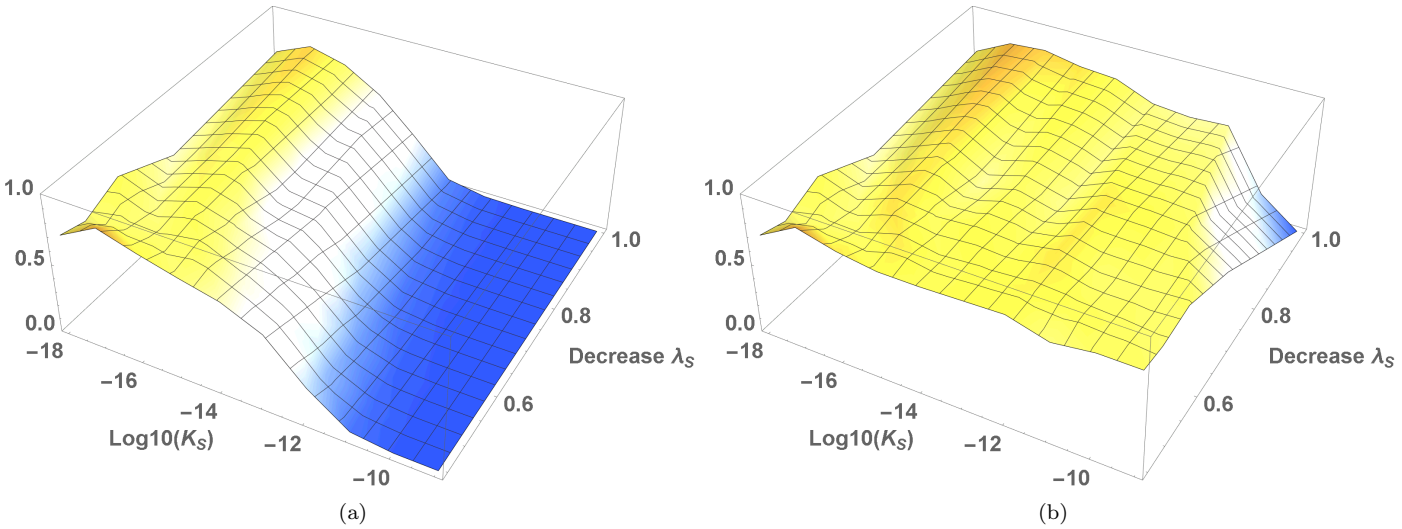

Figure S4: Snapshots of $\phi(c)$, the fraction of high-affinity aptamers with binding affinities stronger than $K_D = 10^{-10}M$, at two different cycles as a function of the initial K_S and the factor λ_S, by which K_S is decreased with each cycle. We used an initial Gaussian distribution $N(-4, 0.8)$ and about 100 additional high-affinity aptamers. (a) At cycle $c = 8$, lower initial values of K_S and faster decreases (lower λ_S) result in faster enrichment of high-affinity aptamers. (b) At cycle $c = 20$, most initial values of K_S eventually lead to enrichment of high-affinity aptamers, provided the affinity is decreased sufficiently fast ($\lambda_S < 0.8$).

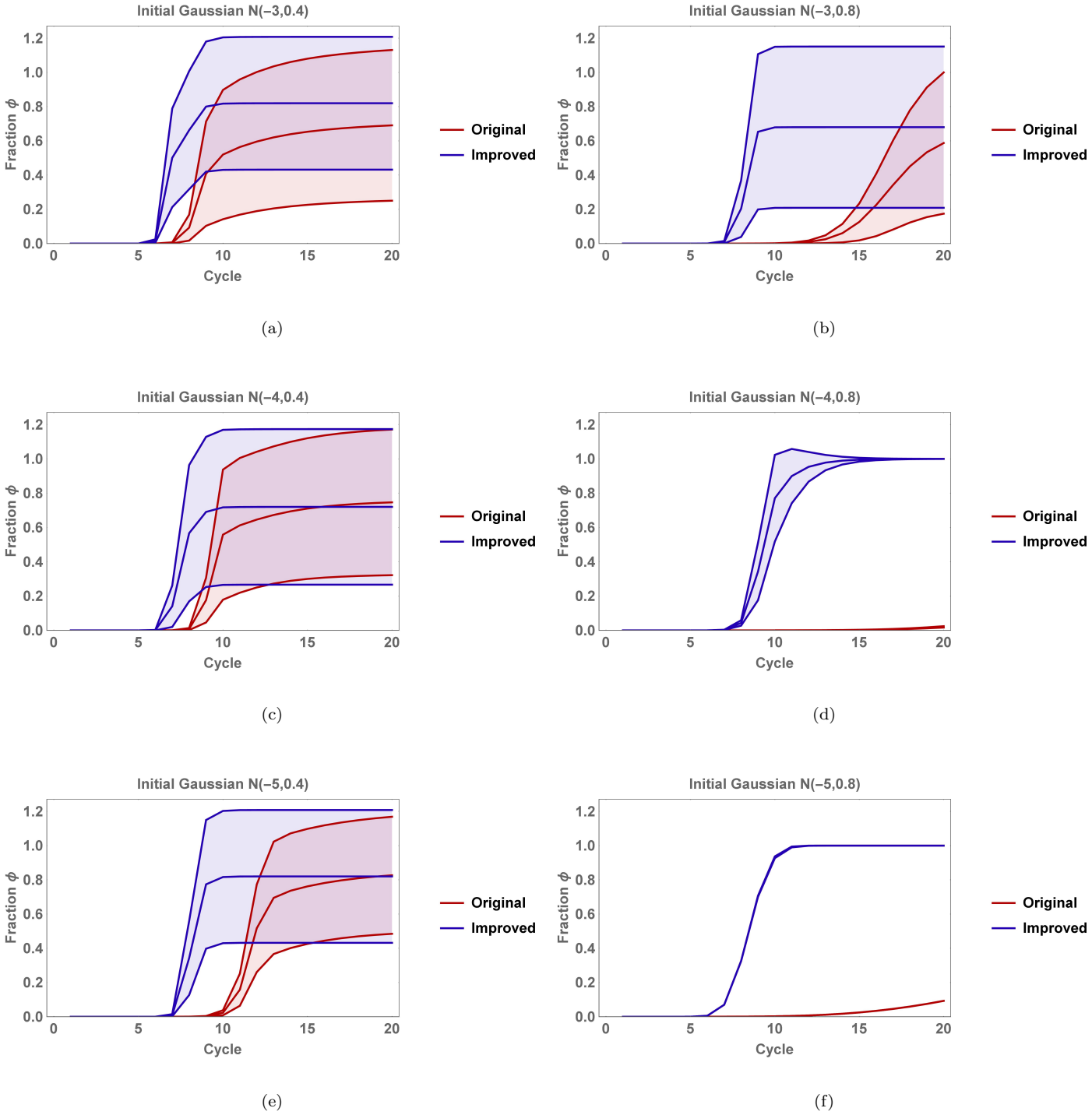

Figure S5: Fraction ϕ of high-affinity binders with cycle. Comparison of SELEX dynamics between the original SELEX protocol (Table S1) with constant T^{tot} and K_S, and the improved protocol with decreasing target concentration and K_S, for six different Gaussian distributions with means $-3, -4, -5$ and standard deviations $0.4, 0.8$. Plot shows fraction of good binders (binding stronger than $K_D = 10^{-10}M$) over SELEX cycles and the standard deviation observed over 50 Monte Carlo simulations. We notice that the improved case reaches higher or equal plateaus as the constant protocol, and it reaches the plateau much faster. There is less variability in the success measures when the initial Gaussian is broader, as in those cases there is initially a large number of good binders present, so stochastic effects play only a small role ((d), (f)). Fig. 8 in the main text shows the final fraction at cycle 20 and the speed with which this fraction is reached.

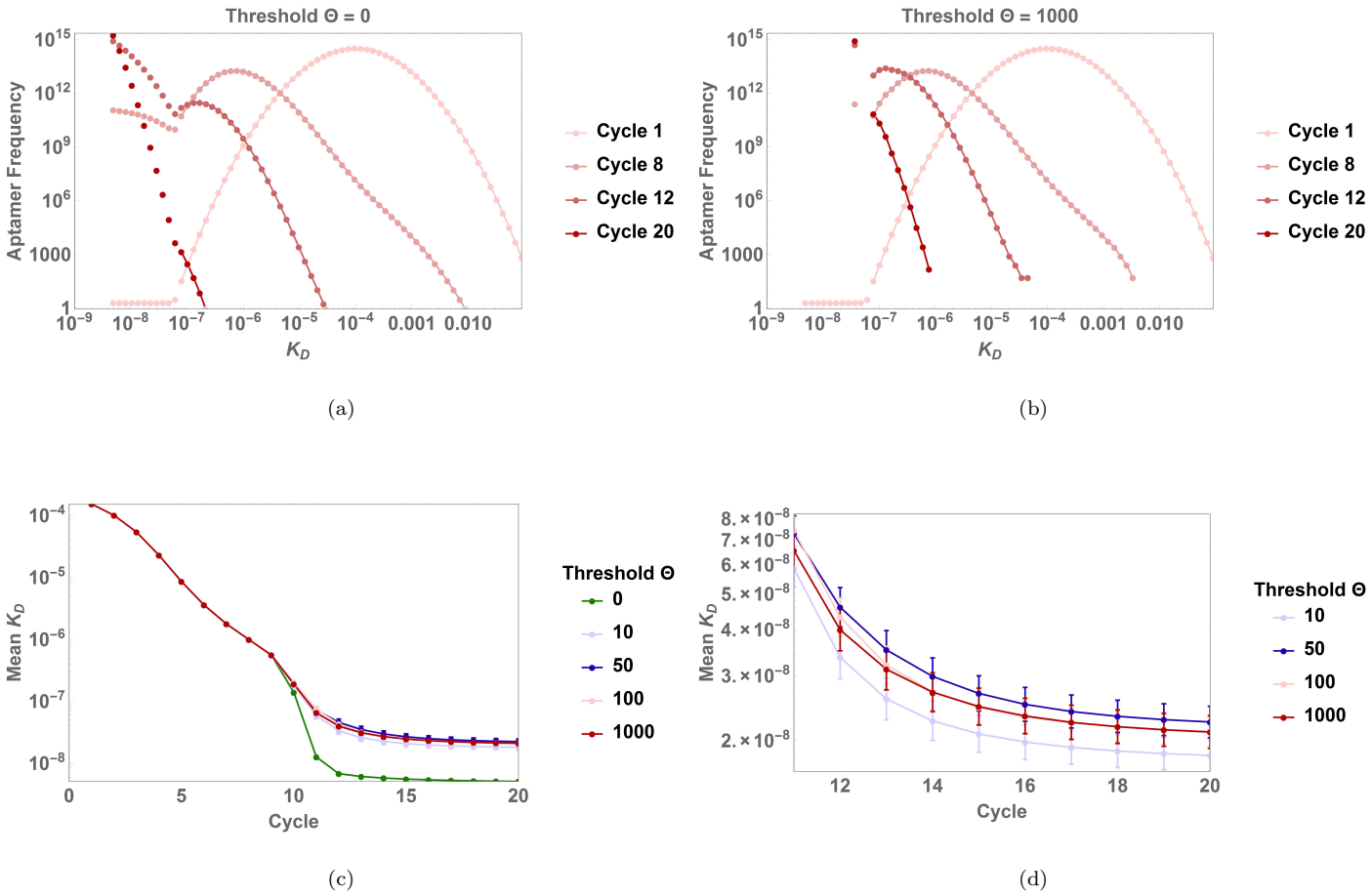

Figure S6: We show the dependence of simulation results on the threshold Θ which defines when we use the stochastic or the deterministic model. In each case the initial distribution is a Gaussian $N(-4, 0.4)$ with added noise. (a) and (b) show sample realization for $\Theta = 0$ and $\Theta = 1000$. The case of $\Theta = 0$ means that we always use the deterministic model. Thus, there is no loss of molecular species, and the handful of aptamers in the low K_D range ($K_D < 10^{-7}M$) always outperform the ones with higher K_D. The case $\Theta = 1000$ does appear qualitatively similar to the case $\Theta = 100$ used in the main text, so repeated runs are required to obtain statistical data. (c) and (d) show the dependence of the mean K_D value, as a measure of protocol performance, for different values of Θ. Each graph is obtained from averaging 150 Monte Carlo simulations and shown with the corresponding error bars. In (c), we notice no visible differences between thresholds between $\Theta = 0$ and $\Theta = 1000$ before cycle 10. This is because in those initial cycles, the randomness seen in low K_D ligands does not affect the bulk of the ligand distribution, and thus does not affect the mean K_D significantly. From cycle 10 on, the results of the non-zero thresholds deviate from the case $\Theta = 0$, which predicts lower mean K_D values. This is because for deterministic dynamics, there is no loss of ligands possible throughout the cycles, and those high affinity ligands always take over the distribution after some time, as seen in (a). (d) shows the same data zoomed in, focusing on the non-zero thresholds from cycle 10 on. We see that from $\Theta = 100$ to $\Theta = 1000$ there is no visible difference, justifying the choice of $\Theta = 100$ in other simulations performed in this work.

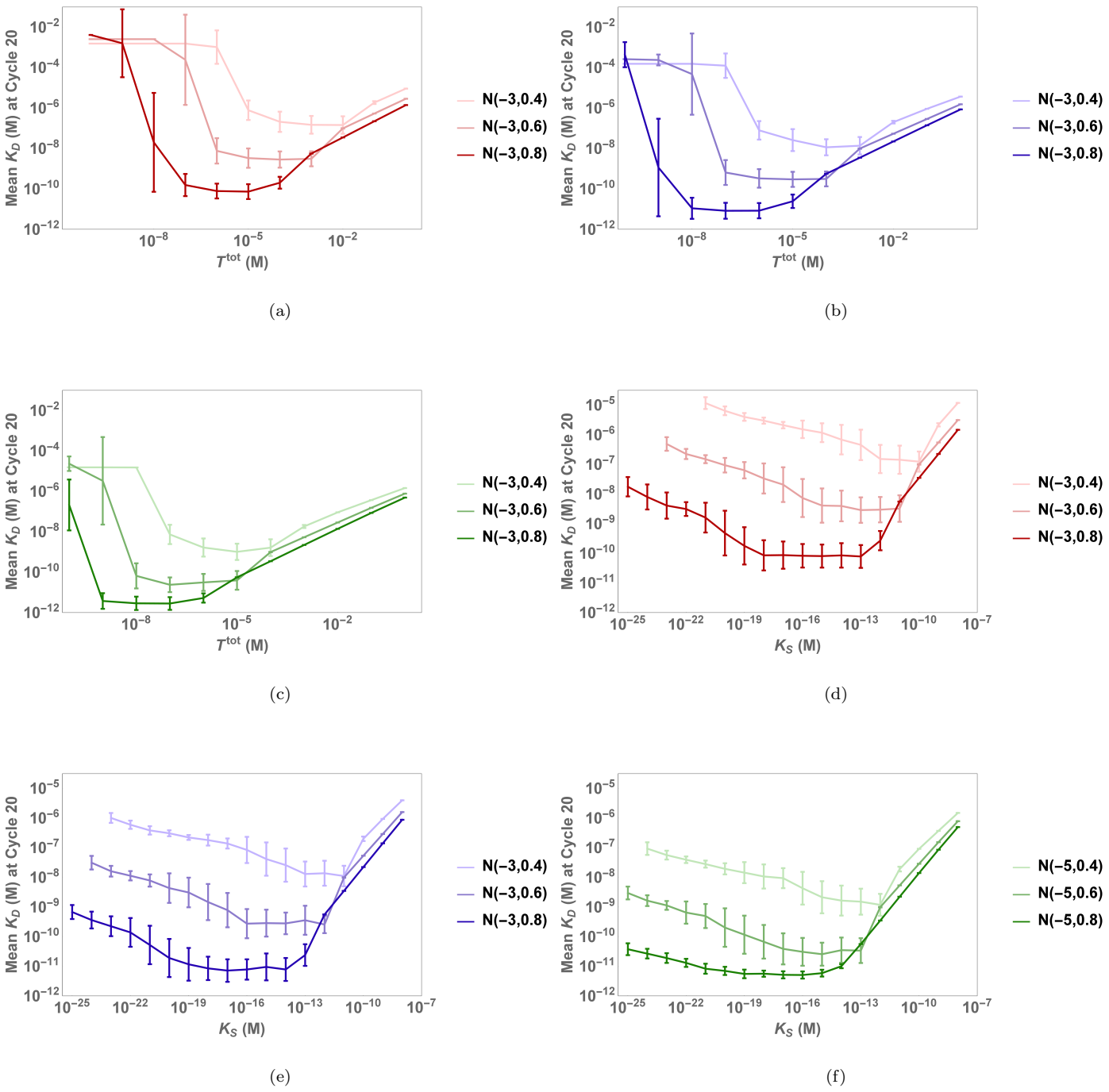

Figure S7: The dependence of the mean K_D value on target concentration T^{tot} and substrate binding affinity K_S is shown at cycle 20 for 9 different initial distributions of ligands, as in Figs. 5, 7, respectively. The difference is that here we added noise to the initial distribution. This noise results in different quantitative, but similar qualitative results. For each data point, 50 Monte Carlo simulations were performed and the mean and standard deviation were computed in log-space. Compared to Figs. 5, 7, we only show three graphs to make the standard deviation, visualized by the added bars, more readable. We see that the noise in the initial distribution results in considerable variability of the obtained mean K_D.

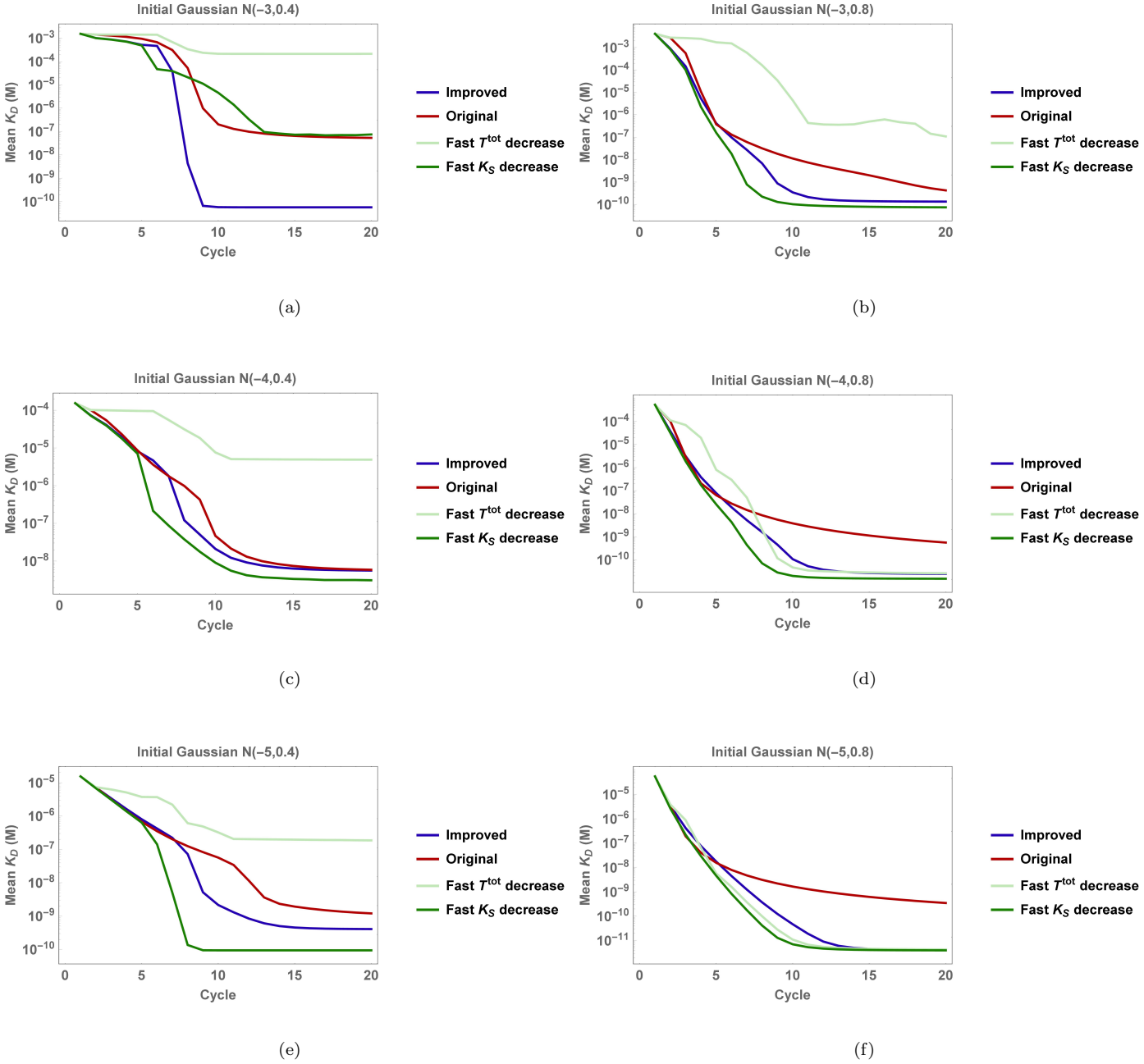

Figure S8: The mean K_D of the distribution as a performance metric for the protocol is shown for 6 different initial Gaussian distributions with added noise, and 4 different protocols. The averages from 100 Monte Carlo simulations are shown in each case. We compare the original protocol where K_S and T^{tot} are constant (Table S1) to the improved protocol discussed in the main text (Table S2), and 2 other protocols where we decrease K_S or T^{tot} faster than in the improved protocol (Table S3). The protocol with the fast T^{tot} decrease does not significantly improve the initial distribution in the cases of the narrow initial Gaussian distribution ((a), (c), (e)), which is expected since we have seen before that decreasing T^{tot} has more of an adverse effect than decreasing K_S (compare Figs. 5 and 7). Deceasing K_S faster than in the improved protocol can often lead to small improvements ((b), (c), (e)) , but can also lead to a much worse outcome ((a)), so such a protocol should only be used if a narrow initial distribution as in (a) can be ruled out.