Predicting DNA Hybridization Kinetics from Sequence

Jinny X. Zhang*,[1,2] John Z. Fang*,[1] Wei Duan,[1] Lucia R. Wu,[1] Angela W. Zhang,[1] Neil Dauchau,[3] Boyan Yordanov,[3] Rasmus Petersen,[3] Andrew Phillips,[3] and David Yu Zhang[1,2]

[1]*Department of Bioengineering, Rice University, Houston, TX*
[2]*Systems, Synthetic, and Physical Biology, Rice University, Houston, TX*
[3]*Microsoft Research, Cambridge, UK*

0. Experimental Methods

Here, we describe the experimental methods used to perform the fluorescence characterization of hybridization kinetics. The data fitting and modeling methods are described in later sections.

Fluorescence Studies:

Oligonucleotide Synthesis and Formulation. All DNA oligonucleotides were ordered from Integrated DNA Technologies in 100 μM LabReady format, pre-suspended in Tris EDTA buffer. Target T and probe P oligonucleotides were ordered as standard desalted oligos at the 25 nanomole scale. Fluorophore F and quencher Q oligonucleotides were ordered as HPLC purified oligos at the 250 nanomole scale. All oligonucleotide stock solutions were quantitated by Nanodrop to determine concentration.

Working secondary stocks of the oligonucleotides were prepared at the following concentrations: 5 μM for P, 10 μM for T, 5 μM for F, and 25 μM for Q. Stocks of QT were prepared for each target T by mixing 10 μL Q, 15 μL T, and 75 μL 5x PBS. Similarly, stocks of FP were prepared for each probe by mixing 10 μL F, 15 μL P, and 75 μL 5x PBS. These secondary stocks were then thermally annealed, cooling from 95 °C to 20 °C over the course of 75 minutes. Unless otherwise specified, all annealing process mentioned below occurred over 75 min, cooling from 95 °C to 20 °C.

Fluorescence Observation of Hybridization Kinetics. Fluorescence experiments were performed using two Horiba Fluoromax-4 instruments, and a 4-sample changer. The slit sizes used for the excitation and emission monochromators were 8 nm and 8 nm. For each 50 s time point, each cuvette's fluorescence was measured for 9 s.

An appropriate amount of FP stock was pipetted into each cuvette at the beginning of the experiment, and the fluorescence was allowed to stabilize over the course of 5 to 30 minutes (fluorescence was observed during this time). Subsequently, the cuvettes were removed from the instruments, and an appropriate amount of the QT stock was added to the cuvette, following which the cuvette was placed back into the instrument. Experimental concentrations of FP and QT are listed in Supplementary Excel spreadsheet, which is further described in Supplementary Section 6.

For each hybridization experiment, positive and negative control experiments were performed to allow mathematical conversion of observed fluorescence values into instantaneous hybridization yields. Negative control experiments included only the FP species, and show the high fluorescence corresponding to 0% yield. Positive control experiments included the FP and QT species thermally annealed (at the hybridization experiment concentrations), and show the low fluorescence corresponding to 100% yield.

Nupack Thermodynamics Calculations:

Nupack (http://www.nupack.org) was used for calculation of a number of features, such as nGp. Where Nupack was used, the experimental hybridization reaction conditions (temperature, buffer, DNA concentration) were used as inputs. Nupack's "dangles" parameter was set to the default option of "some," and other parameters were set to default values. Note that although Nupack reference concentration is indeed 55 M rather than 1 M, this pecularity does not affect our WNV model predictions because all two-stranded species' energies are equally affected.

NGS Studies:

Oligonucleotide Preparation. All DNA oligonucleotides were ordered from Integrated DNA Technologies in 100 μM LabReady format, pre-suspended in Tris EDTA Buffer. ERCC spike-in oligos, Probe oligos and Adaptor-primers were purchased as standard desalted oligonucleotides at the 100 nanomole scale. The Biotinylated universal oligo was purchased as HPLC purified oligonucleotides at 250 nanomole scale. The Displacer oligo used for removing captured Targets/Probes from the magnetic bead surface was purchased as a standard desalted oligos at the 25 nanomole scale. The Beads-Blocker oligos used for suppressing nonspecific binding of to the magnetic bead surface were purchased as standard desalted oligos at the 25 nanomole scale. The Adaptor-Blocker oligo used for preventing ligated genomic DNA fragments from forming hairpin structure was purchased as a standard desalted oligo at the 25 nanomole scale. The Adaptor-primers that were purchased as standard desalted oligos at the 25 nanomole scale used to amplify the released product for a second round of capture and amplification.

The ERCC spike-in oligo and 36nt complementary oligo were mixed with the universal Biotinylated oligo as so called ERCC mixture in 5×PBS Buffer with 0.1% Tween 20 at concentrations of 10 nM, 20 nM, and 40 nM, respectively. The ERCC mixture was thermally annealed, and then combined with the Target mixture for further use. Two separate pools were made by each mixing 65 probes together in 1×TE Buffer to a concentration of

1.3 μM per probe. After diluting both probe pools to 500 nM per probe using 5×PBS Buffer with 0.1% Tween 20, probe pools were separately combined with Biotinylated universal oligo in 5×PBS Buffer with 0.1% Tween 20, and diluted to achieve final concentrations of 64 nM per Probe and 8.32 μM Biotinylated universal oligo. These two mixtures were then thermally annealed. Capture Probe Pool, as mentioned in the following text, was the mixture of equal amount of two separate pools and the concentration of Capture Probe Pool means the concentration of probe of one target.

The Blocker mix was created by mixing the 5 blocker oligos with a concentrations of 16 μM each, in 1× Tris EDTA Buffer, 0.75 M NaCl (purchased as Ambion 5 M NaCl stock solution from Thermo Fisher Scientific). 4 μ L Dynabeads MyOne Streptavidin T1 (Thermo Fisher Scientific) were pre-incubated with 20 μL of 16 μM Beads Blocker Mix at 55 °C for at least 30 minutes before use in capture reactions, noticed as Pre-Capture Beads mixture in the following text.

Sheared gDNA Preparation. 10 μg Genomic DNA was sheared for 20min using the Covaris M220 Focused-Ultrasonicator and Holder XTU Insert microTUBE 130L to achieve the product with basepair-peak at 120bp. This shearing was performed by Genomic and RNA Profiling Core Lab from Baylor College of Medicine. 3.6 μg of sheared genomic DNA was ligated and size-selected using NEBNext Ultra II DNA Library Prep Kit for Illumina. Final product from ligation was approximately 100 μL in 0.1×Tris EDTA Buffer and was mixed with 100 μL of 8 μM Adaptor-Blocker in 10×PBS buffer with 0.1% Tween 20 for thermally annealing. 5 μL of 1 pM ERCC mixture was spike in afterwards as standard to the annealed Adaptor-Blocked Target pool together as the Target Mixture for next capture step.

Hybrid-Capture Protocol. For better result in suppressing background, Dual-Capture-PCR protocol was used. First capture aimed at capture different amount of targets depending on different kinetics. First PCR was used to amplify captured and released product, which still had lots of background sequences, for second capture use. Second capture was performed to capture all remaining targets in a fast process using high concentration of capture probes. Second PCR was used to attach NGS adaptor and index to 2nd-capture released product.

The 10 μL 0.25 nM Capture Probe mixture was mixed with 17 μL 5×PBS Buffer with 0.1% Tween 20, then pre-incubated at 55 °C for at least 30 minutes prior to hybridization reaction to allow temperature stabilization. 23 μL of Target mixture was added to the Probe mixture to achieve final concentrations of 0.05 nM per probe in total volume of 50 μL. Hybridization reactions were incubated at 55°C in a Eppendorf Mastercycler Personal for 20 minutes or 24 hours.

After hybridization, the hybridization mixture was transferred to the tube containing Pre-Capture Beads mixture(4 μL Dynabeads MyOne Streptavidin T1 pre-incubated with 20 μL of 16 μM Blocker oligos mix) , then fully mixed and incubated at 55°C for an additional 15 minutes, while shaking at 430 rpm in a Multi-Therm shaker (Benchmark Scientific). Subsequently, the bead solution was washed 3 times using 100 μL Hybridization Washing Buffer1(1×Tris EDTA Buffer with 0.75M NaCl), and then additional 3 times using Hybridization Washing Buffer2(1×Tris EDTA Buffer with 0.15M NaCl). Every washing step of both buffers was incubated at 55 °C for 5min, to remove unbound Probe and background sequences from beads surface. Additionally, the beads were transferred to a new tube after every wash to minimize nonspecific retention of tube walls.

Release of Targets and Probes from the beads was performed using strand displacement. 18 μL of 500 nM of the Displacer oligo (in 1×PBS Buffer) was added to wash the bead sample, and allowed to react for 5 minutes at 55° while shaking at 430 rpm in the Multi-Therm shaker. The displaced oligos in the supernatant was then collected for the subsequent first PCR amplification step, which is used to amplify released product for second capture. Second Capture was performed to lower background.

10 μL Amplicon of first PCR was mixed with 30 μL 80nM Adaptor-Blocker in 10×PBS buffer with 0.1% Tween 20 and 40 μL of water for thermally annealing. Second Capture was performed with the Capture Probe Pool and the amplicon from first PCR after clean up and annealing. 10 μL of annealed Amplicon was mixed with Capture Probe Pool to achieve a final concentration of roughly 1nM per probe.

PCR Amplification. The first PCR was using Adaptor-Primers to increase the concentration of released product. The second PCR was performed using NEBNext Multiplex Oligos for Illumina to append Illumina P5/P7 and index sequences, while simultaneously increasing the concentration of released product after the second capture. Both PCR amplifications were performed in a Eppendorf Mastercycler Personal with the following protocol: an initial 3 min incubation at 95°C, followed by 16 cycles for the first PCR and 14 cycles for the second PCR of 10 s at 95°C and 30 s at 60°C, followed by a final 45 s extention at 60°C. The final concentrations of the first and second PCR primers were 400 nM each. 15 μL of the displacement released oligo sample served as the template in each 60 μL PCR amplification reaction. Purification of the amplicon was performed using DNA Clean & Concentrator-25 kit (Zymo Research).

Quantification, Library Pooling and NGS. Quantification of the purified libraries was performed with

Qubit 3.0 Fluorometer (Thermo Fisher Scientific) using Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific). All libraries were pooled evenly together for a total concentration of 4 nM in water. Next Generation Sequencing was performed using the Illumina Miseq using a MiSeq v2 300-cycle Reagent Kit (Illumina).

1. Experimental Controls and Reproducibility

Instantaneous yield determination. For each fluorescence hybridization experiment, we performed positive and negative controls to determine yield. Fig. 1-1 shows a representative fluorescence trace, wherein the FP probe species is introduced at time $t = 0$, and the QT target species is introduced at $t \approx 900$ s. The stable fluorescence of the FP species in the cuvette between 15 minutes are averaged to calculate the negative control of 0% hybridization yield. The hybridization of FP and QT proceeded until $t \approx 6900$ s. Subsequently, the cuvette was removed from the fluorimeter, heated to 95°C, and then cooled to the hybridization temperature. The solution fluorescence after this thermal anneal is averaged and considered as the positive control corresponding to 100% hybridization yield. The instantaneous hybridization yield at every single timepoint is calculated as the linear interpolation between the positive and negative controls.

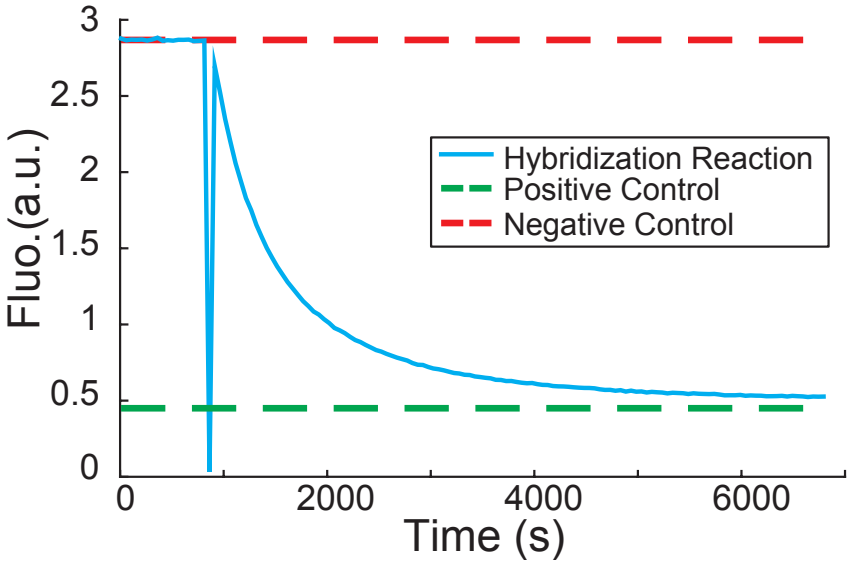

FIG. 1-1: Determining instantaneous hybridization yield using positive and negative controls. The displayed reaction uses Target/Probe pair #85. The fluorophore oligo (F) and the probe oligo (P) were pre-annealed into FP, and introduced into the cuvette before $t = 0$. At $t \approx 900$s, the cuvette was removed from the fluorimeter, and the pre-annealed QT (quencher oligo Q and target oligo T) was introduced. The average of the solution fluorescence in the timepoints where only FP was present is taken as the negative control (red dashed line). After the end of the reaction at $t \approx 6900$ s, the solution was thermally annealed, and the fluorescence was measured to generate the positive control (green dashed line).

Vendor oligonucleotide synthesis reproducibility. Oligonucleotides purchased commercially from vendors such as Integrated DNA Technologies (IDT) or Sigma show significant synthesis-to-synthesis variability. To ensure that the hybridization rate constants k_{Hyb} reflected primarily DNA sequence variability rather than vendor synthesis variability, we selected 3 different target/probe systems (37, 70, and 85) for detailed study.

Each of T and P oligonucleotides were ordered as 3 separate syntheses (2 from IDT and 1 from Sigma). Additionally, for each target/probe pair, we performed 3 distinct dilutions from primary stock provided by the vendor, in order to characterize dilution and reaction preparation variability. These results are shown in Fig. 1-2. There is insignificant difference between the 3 different dilutions and reaction preparations, indicating that our experimental reproducibility is very high. There is marginally higher difference between the 3 different oligonucleotide syntheses, indicating that synthesis reproducibility is high.

Fig. 1-3 summarized the k_{Hyb} and Bad Fraction observed for the different oligo syntheses. The hybridization rate constant k_{Hyb} generally showed low variability — the largest different observed was in Target/Probe pair #70, which exhibited a difference in $\log_{10}(k_{\mathrm{Hyb}})$ of about 0.2. As this difference is small compared to the range of k_{Hyb} values observed experimentally, we believe that k_{Hyb} are primarily affected by sequence.

In contrast, the Bad Fraction varied significantly across different syntheses, with Target/Probe pair #70 varying between 0.14 and 0.01. Thus, the variability across syntheses is large compared to the variability across sequences. For this reason, we did not apply the WNV model to predicting the Bad Fraction parameter. As an aside, we were surprised that Bad Fraction observed for IDT was in general higher than for Sigma, given IDT's reputation in the field as the technologically leading supplier of oligonucleotides.

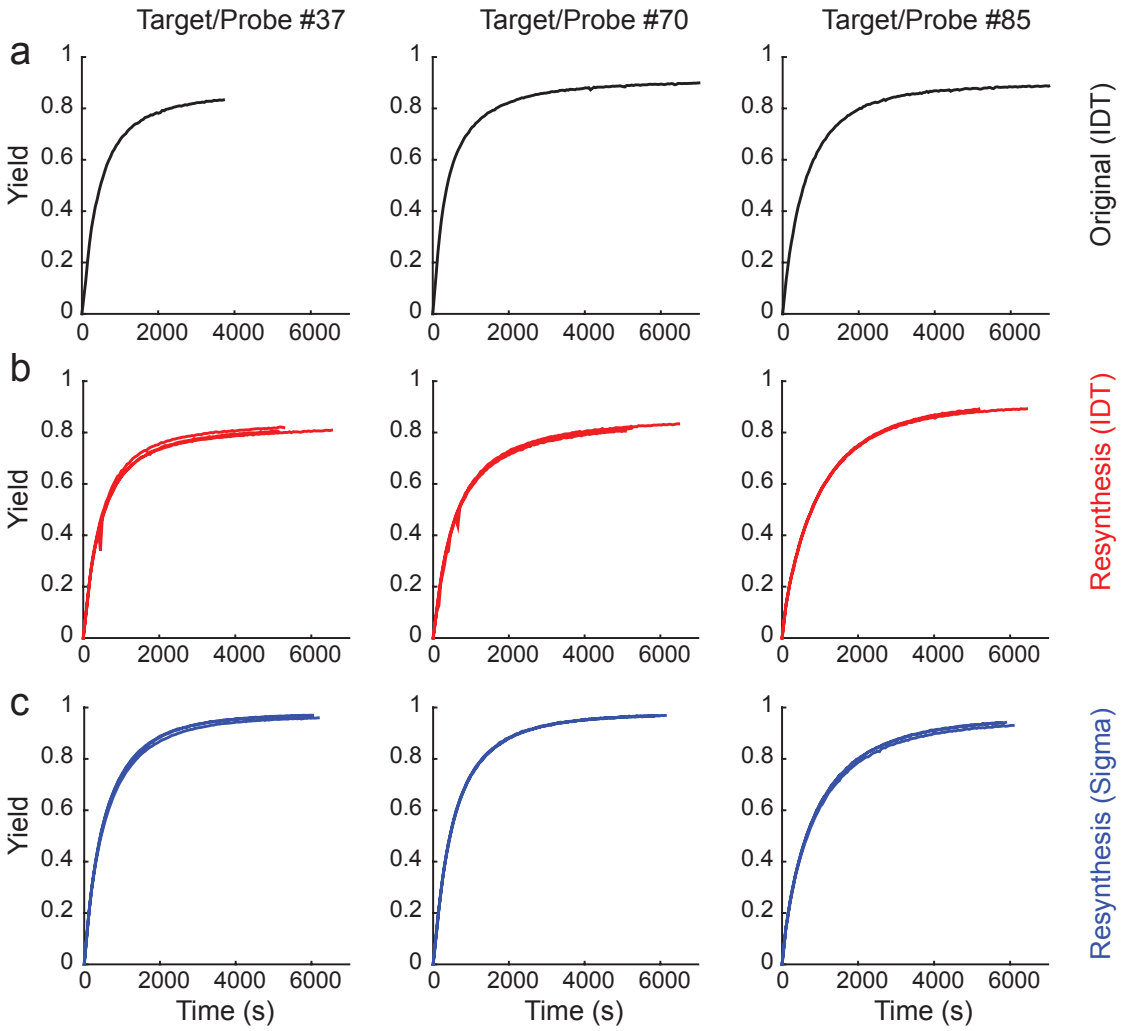

FIG. 1-2: Experimental results on re-synthesized and re-diluted target and probe oligonucleotides. All shown hybridization reactions were performed at 55 °C. The resynthesized oligo experiments in panels (b) and (c) show triplicate experiments.

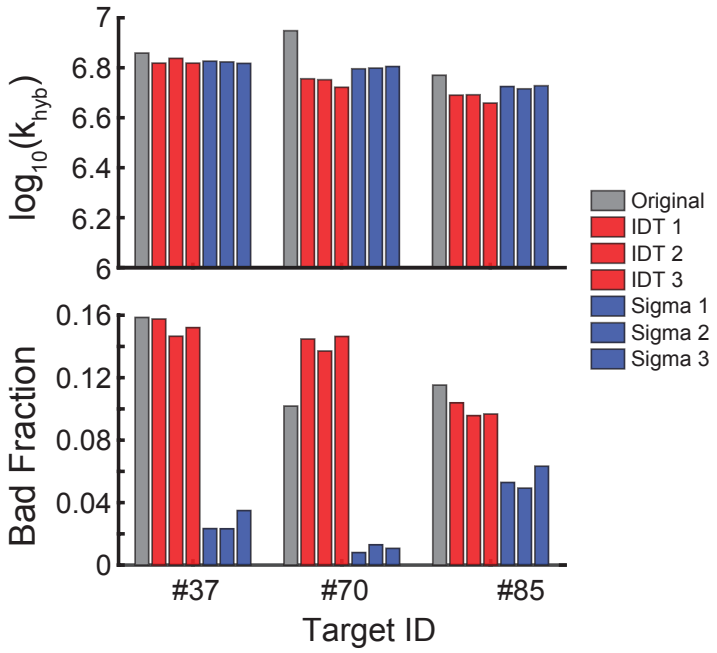

FIG. 1-3: Summary of k_{Hyb} and Bad Fraction for re-synthesized and re-diluted oligos.

2. Experimental Data and Reaction Model Fitting

As a reminder to the reader, we considered three hybridization reaction models H1, H2, and H3, with the following modeled reactions:

Model H1, with parameters $k_{\mathbf{Hyb}}$ and $f_{\mathbf{good}}$:

$$\begin{aligned} T + P_{\text{good}} &\overset{k_{\text{Hyb}}}{\rightarrow} TP \\ \frac{[P]_0 - [P_{\text{good}}]_0}{[P]_0} &= \text{Bad Fraction} \end{aligned}$$

Model H2, with parameters $k_{\mathbf{Hyb}}$, k_1, and k_2:

$$\begin{aligned} T + P &\overset{k_{\text{Hyb}}}{\rightarrow} TP \\ T + P &\underset{k_2}{\overset{k_1}{\rightleftharpoons}} TP_{\text{bad}} \end{aligned}$$

Model H3, with parameters $k_{\mathbf{Hyb}}$, k_1, k_2, and $f_{\mathbf{good}}$:

$$\begin{aligned} T + P_{\text{good}} &\overset{k_{\text{Hyb}}}{\rightarrow} TP \\ T + P_{\text{good}} &\underset{k_2}{\overset{k_1}{\rightleftharpoons}} TP_{\text{bad}} \\ \frac{[P]_0 - [P_{\text{good}}]_0}{[P]_0} &= \text{Bad Fraction} \end{aligned}$$

For a fair comparison of the three models against one another, we obtain the best-fit parameter values of each model for each hybridization experiment.

To provide a formal view of the inference problem being solved, let $\mathcal{H}$ denote a hypothesized dynamic model, which is parameterized by values $\vec{w}$, and $\vec{y}$ be a vector of observation data. We denote by $p(\vec{w}|\vec{y},\mathcal{H})$ the *posterior* density of the parameter values for a given hypothesis, given the observation data. Under Bayes' formula, this is given by

$$p(\vec{w}|\vec{y},\mathcal{H}) = \frac{p(\vec{y}|\vec{w},\mathcal{H})p(\vec{w}|\mathcal{H})}{p(\vec{y}|\mathcal{H})} \tag{1}$$

where $p(\vec{y}|\vec{w},\mathcal{H})$ is the *likelihood* of observing $\vec{y}$ for a given hypothesis and parameter set. The parameter *priors* are given by $p(\vec{w}|\mathcal{H})$, and encode any prior belief of the distribution of the parameters. The denominator is independent of the parameters, and is called the *marginal likelihood*.

The posterior can be approximated using Markov chain Monte Carlo (MCMC) methods. Here, we make use of the Filzbach software, which uses a variant of the classic Metropolis-Hastings scheme. The Filzbach software has been applied to a variety of biological and ecological model inference problems, and is available from the Microsoft Research webpage (`http://research.microsoft.com/filzbach`). Metropolis-Hastings MCMC creates Markov chains of parameter sets, by making multidimensional perturbations to a current parameter set, creating a *proposal*. The likelihood function is compared between the current and proposal sets, and the chain moves to the proposal with probability 1 if the likelihood has increased, and with some probability less than 1 if the likelihood is lower. As this process is repeated, the chain tends towards regions of parameter space that have higher likelihood scores. In the limit, as the number of samples (steps in the Markov chain) is increased, the posterior distribution will be described by the parameter values in the samples. The marginal distribution of each parameter can then be evaluated by simply creating a histogram of the values of each parameter.

Filzbach requires the user to specify the log likelihood function, and the parameter priors. Here, our observations are fluorescence measurements $y_1, y_2, \ldots, y_n$ at successive time-points $t_1, t_2, \ldots, t_n$. We assume that the fluorescence measurements have Gaussian-distributed deviations from the underlying concentration of fluorophore-containing single-stranded DNA molecules, and thus use the likelihood function

$$p(\vec{y}|\vec{w},\mathcal{H}) = \prod_{i=1}^{n} p(y_i|\vec{w},\mathcal{H}), \qquad \text{where} \quad p(y_i|\vec{w},\mathcal{H}) \sim \mathcal{N}(s_i^{(\vec{w},\mathcal{H})}, \sigma^2) \tag{2}$$

where the $s_i^{(\vec{w},\mathcal{H})}$ are simulated fluorescence from a model $\mathcal{H}$ with parameters $\vec{w}$. Here, σ represents the standard deviation of the data, and is also inferred during the parameter inference procedure. For numerical stability, the logarithm of the likelihood function is used in the implementation throughout. As such, the product in (**??**) gets converted into a sum.

For all kinetics traces, we evaluated 5 independent MCMC chains with 10,000 burn-in iterations and 30,000 sample iterations (see Filzbach documentation for more details). Only the parameter set with the highest observed likelihood score was retained for further analysis.

The raw fluorescence data and the best-fit traces for each model are shown in Fig. 1-1 through 1-11. Because the simulation traces for the three models appear very similar for many hybridization experiments, Fig. 1-12 through 1-22 show the relative error RE = $\mathrm{Abs}\left(\frac{s_i - y_i}{y_i}\right)$ of the each model for each hybridization experiment. Note that some figure subpanels are empty (e.g. Target #30 at 55 °C in Fig. 1-9) because the hybridization reaction was not thermodynamically favorable for the target and probe sequence pair at the listed temperature. For these experiments, no significant decrease of fluorescence was observed upon addition of the QT complex.

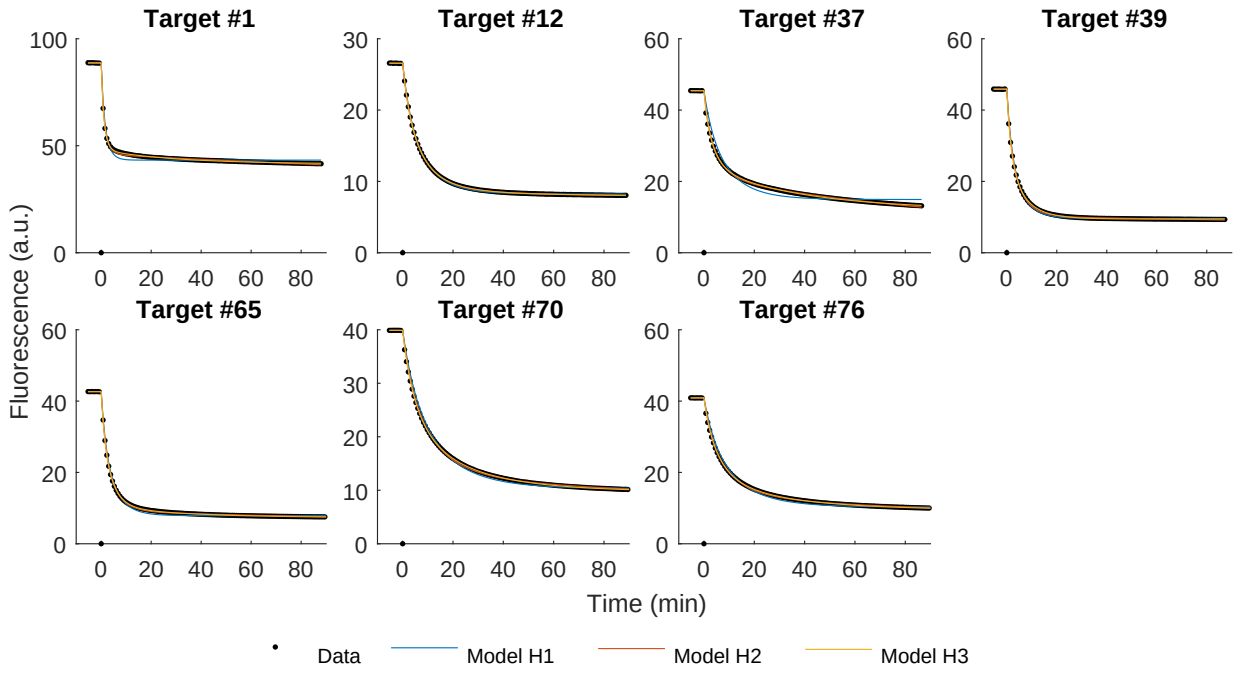

FIG. 2-1: Fluorescence data and best-fit traces for hybridization experiments performed at 28 °C.

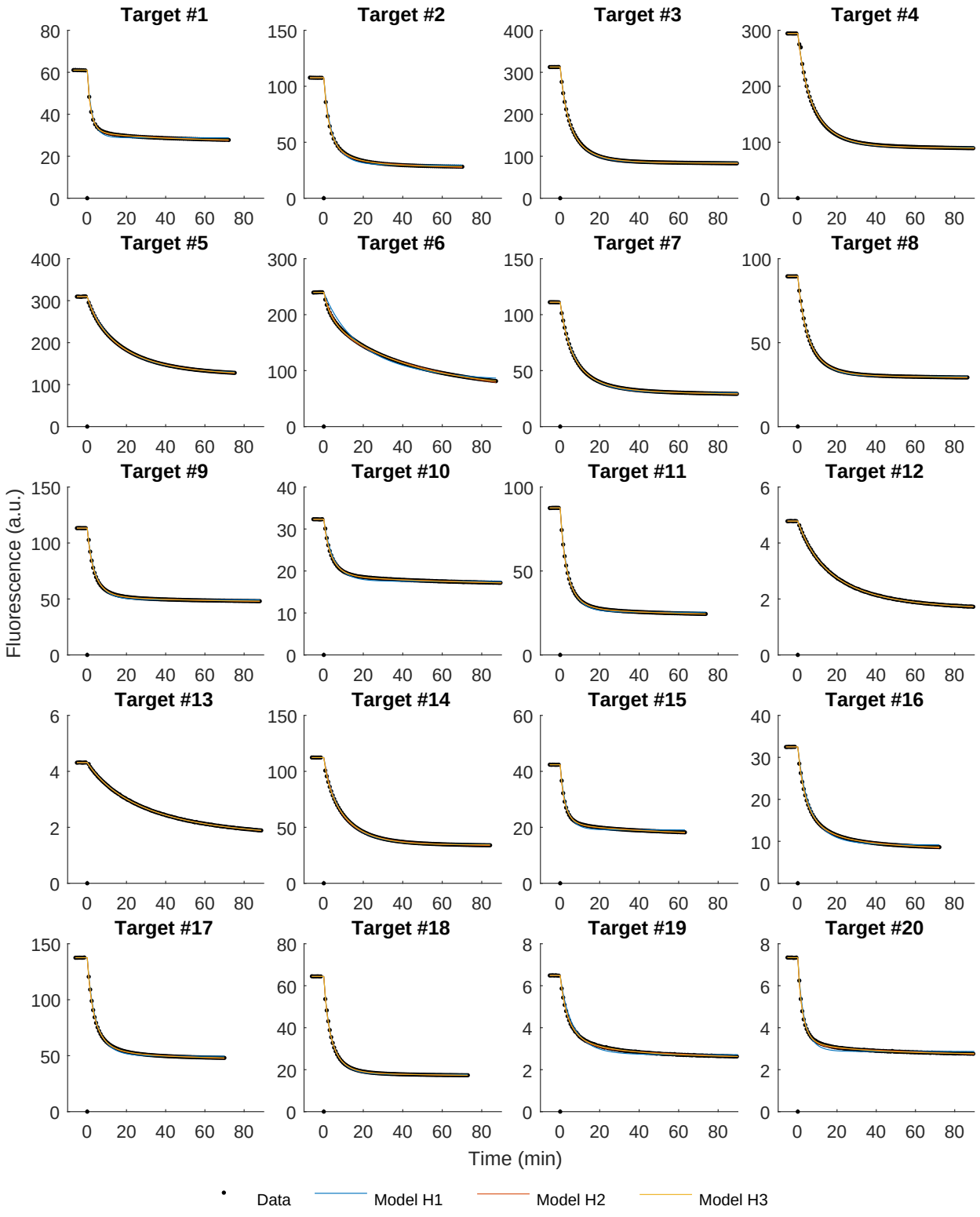

FIG. 2-2: Fluorescence data and best-fit traces for hybridization experiments performed at 37 °C, target sequences 1-20.

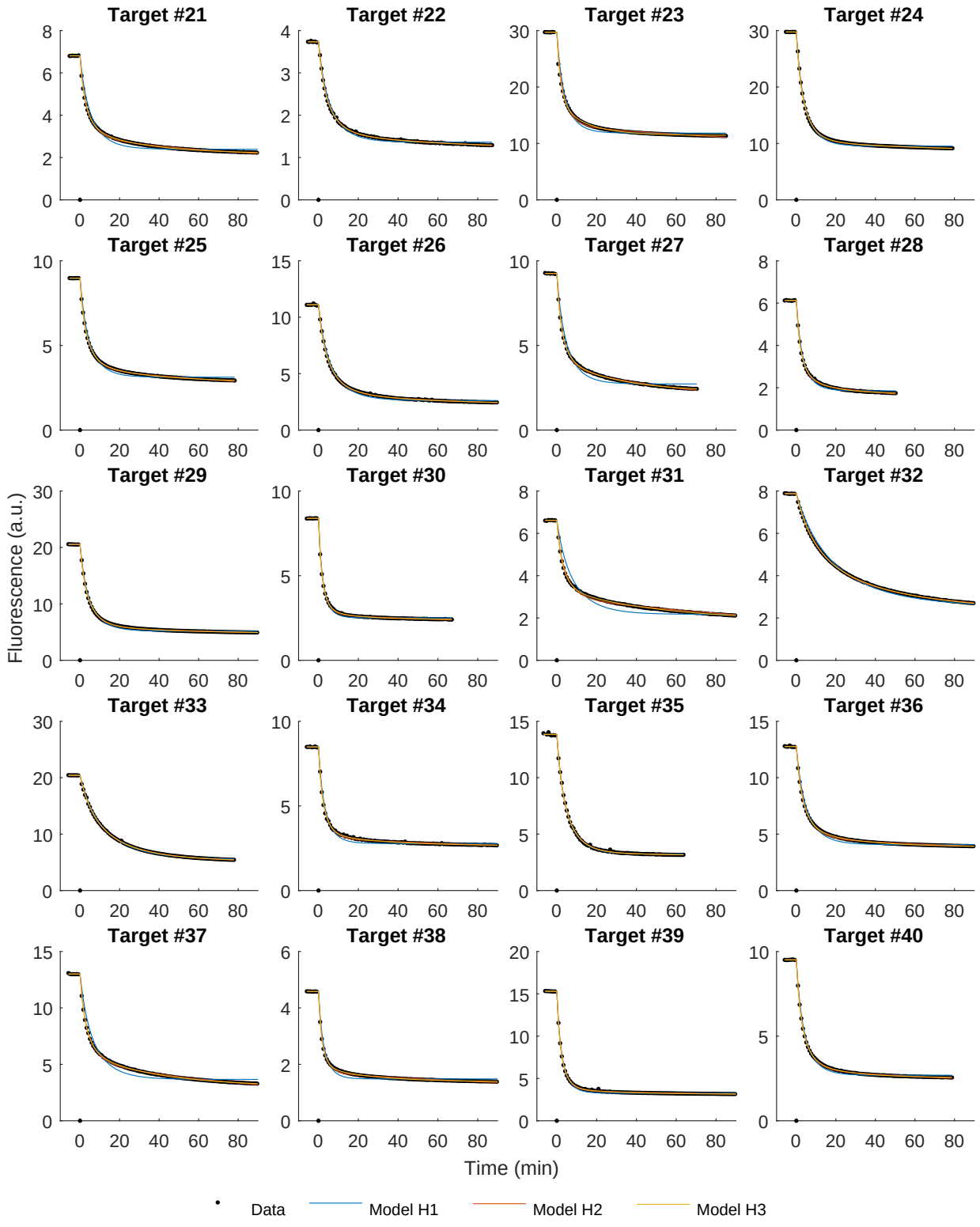

FIG. 2-3: Fluorescence data and best-fit traces for hybridization experiments performed at 37 °C, target sequences 21-40.

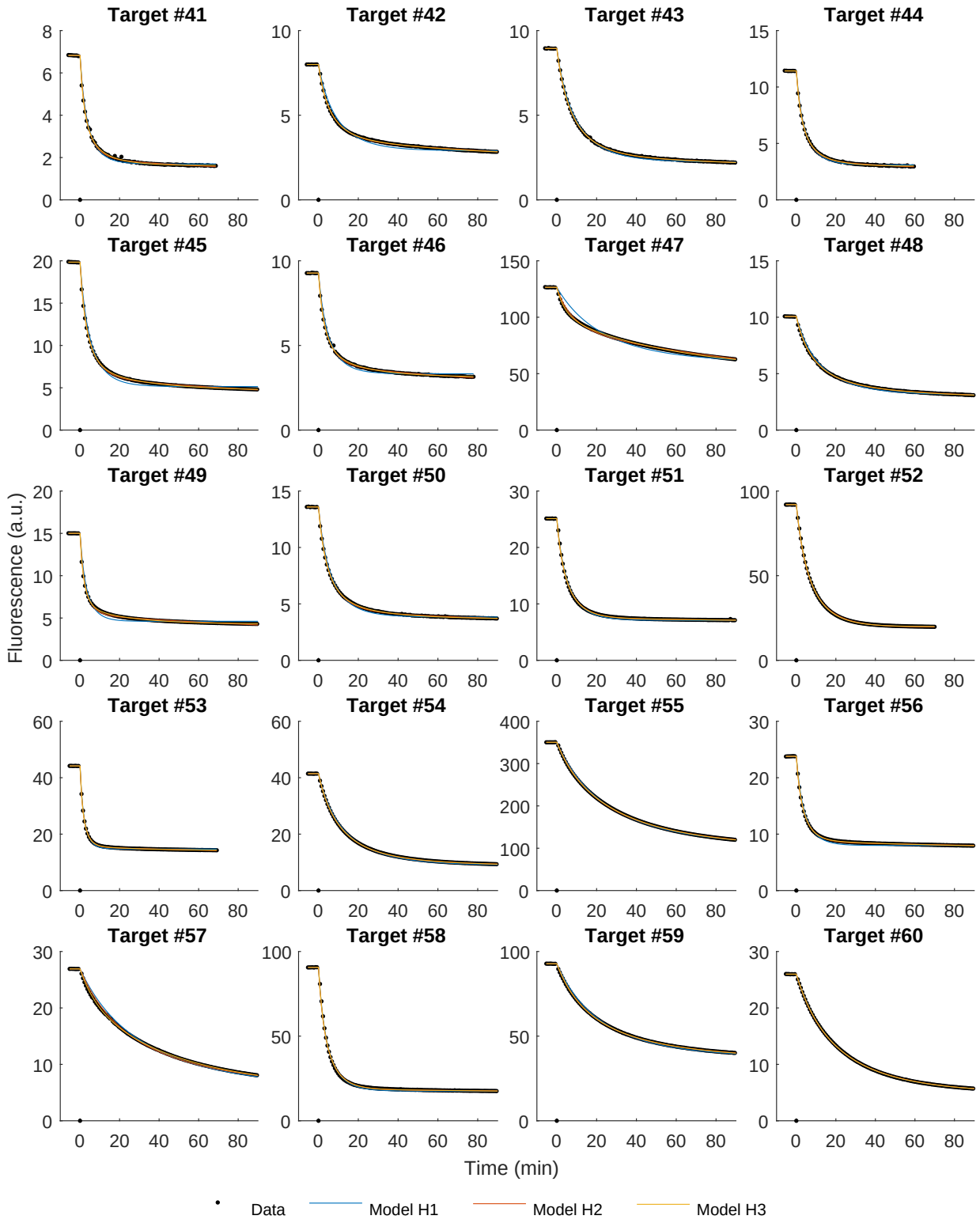

FIG. 2-4: Fluorescence data and best-fit traces for hybridization experiments performed at 37 °C, target sequences 41-60.

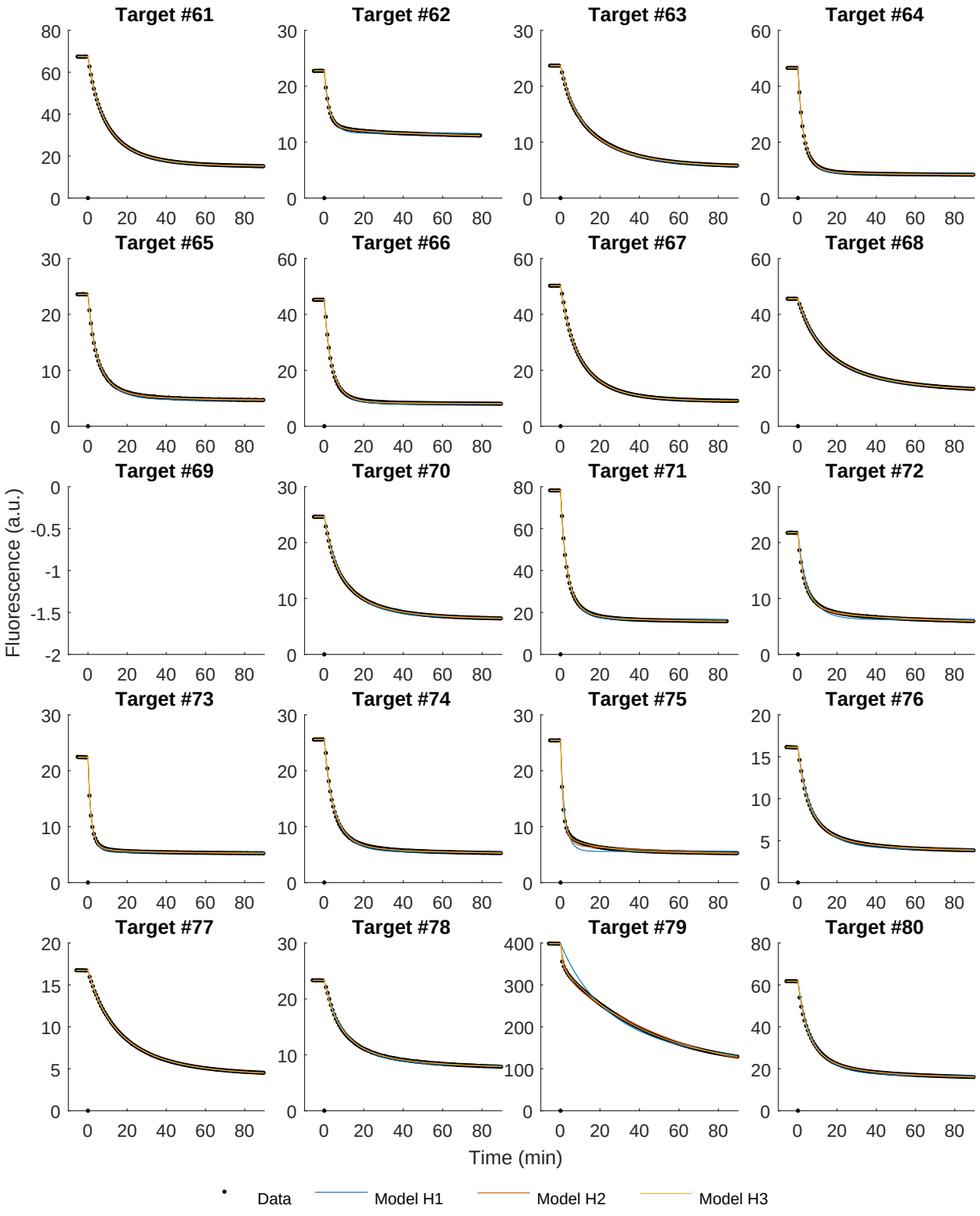

FIG. 2-5: Fluorescence data and best-fit traces for hybridization experiments performed at 37 °C, target sequences 61-80.

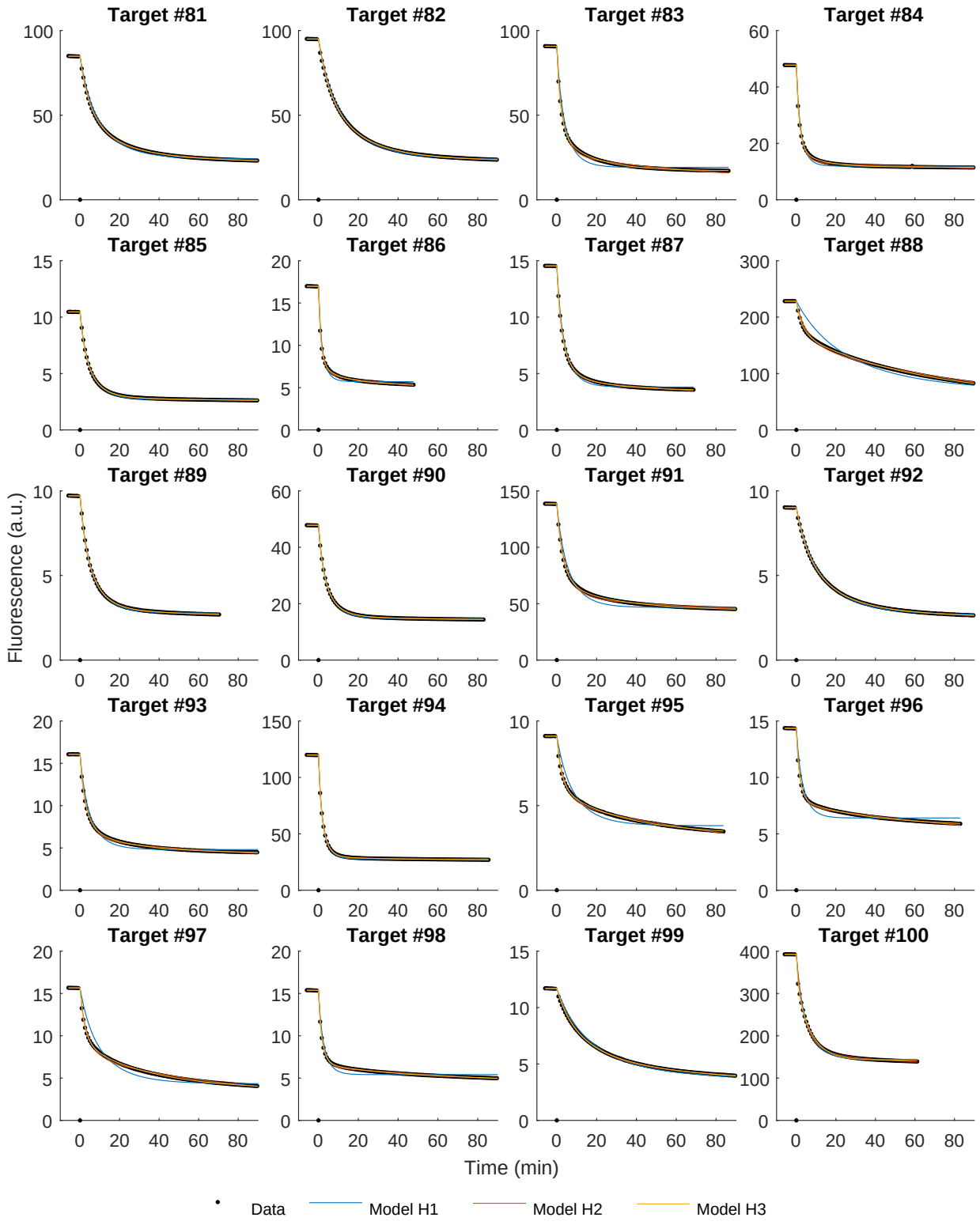

FIG. 2-6: Fluorescence data and best-fit traces for hybridization experiments performed at 37 °C, target sequences 81-100.

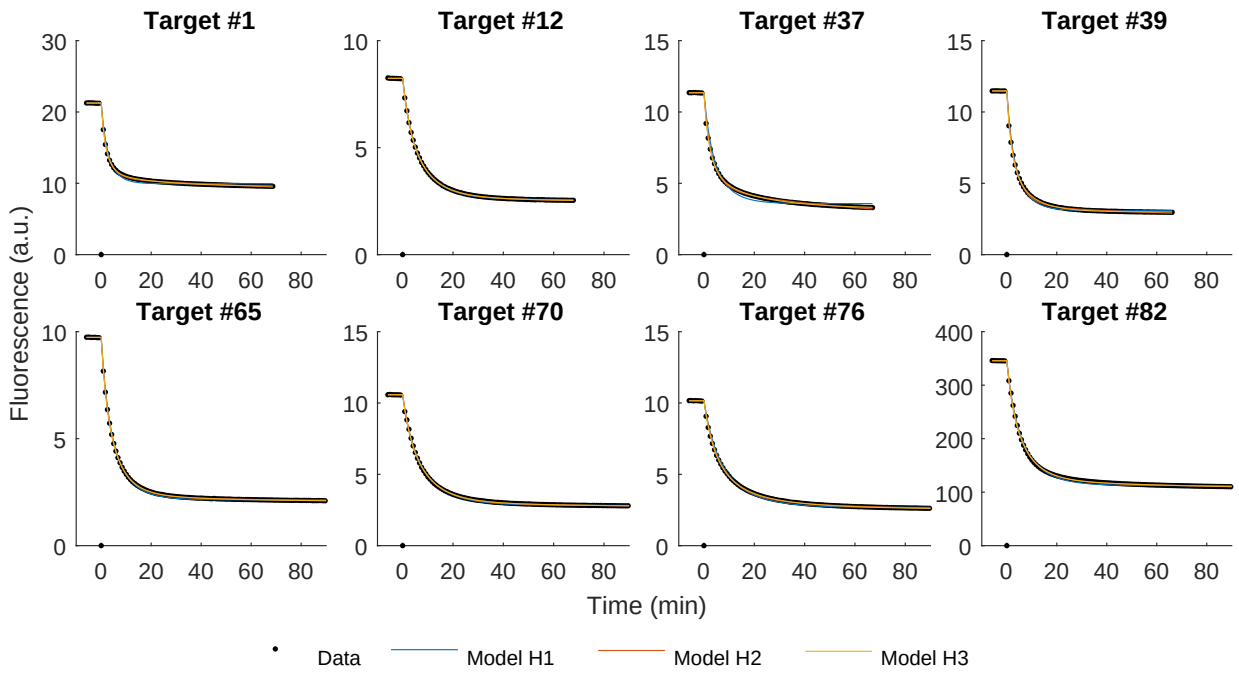

FIG. 2-7: Fluorescence data and best-fit traces for hybridization experiments performed at 46 °C.

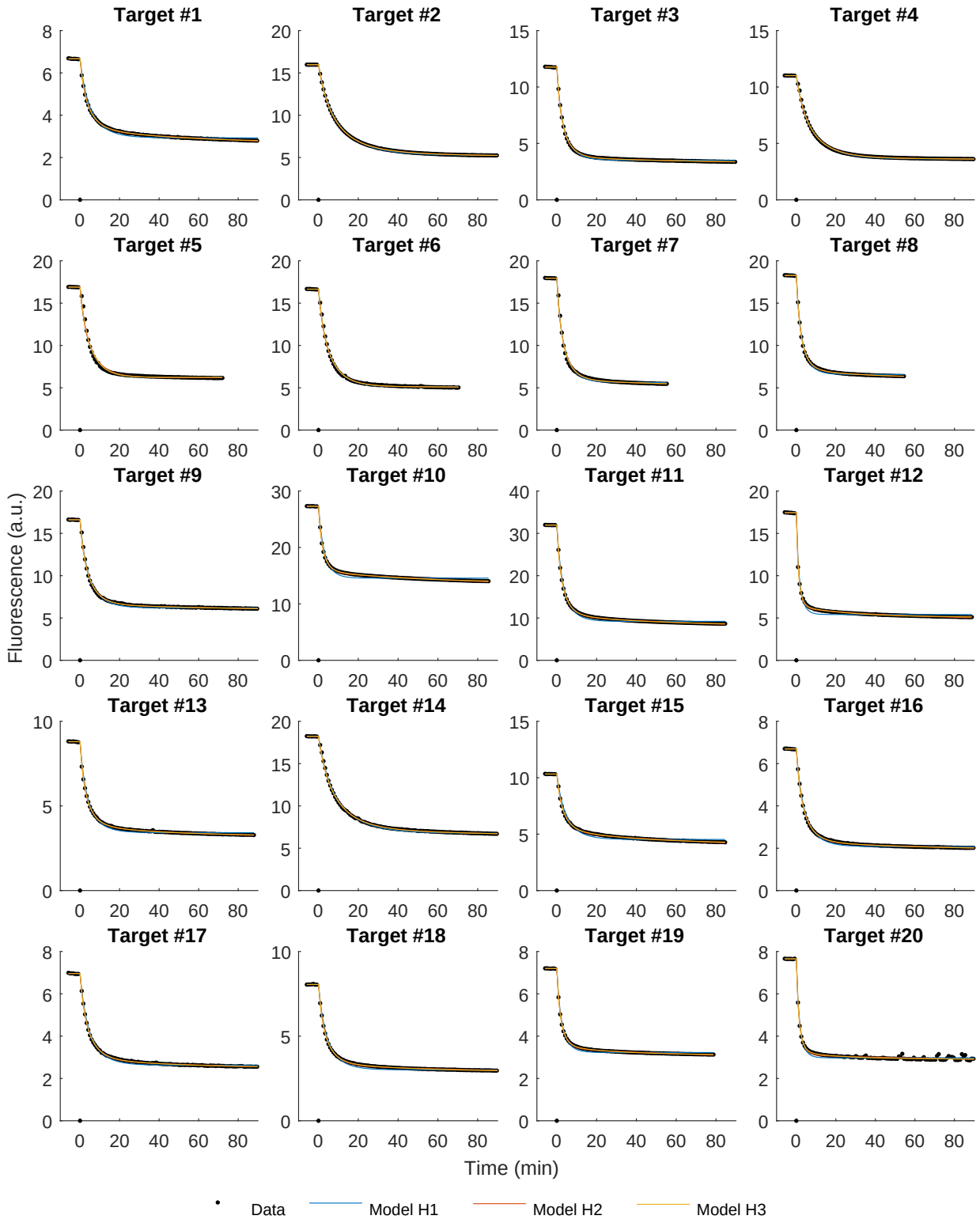

FIG. 2-8: Fluorescence data and best-fit traces for hybridization experiments performed at 55 °C, target sequences 1-20.

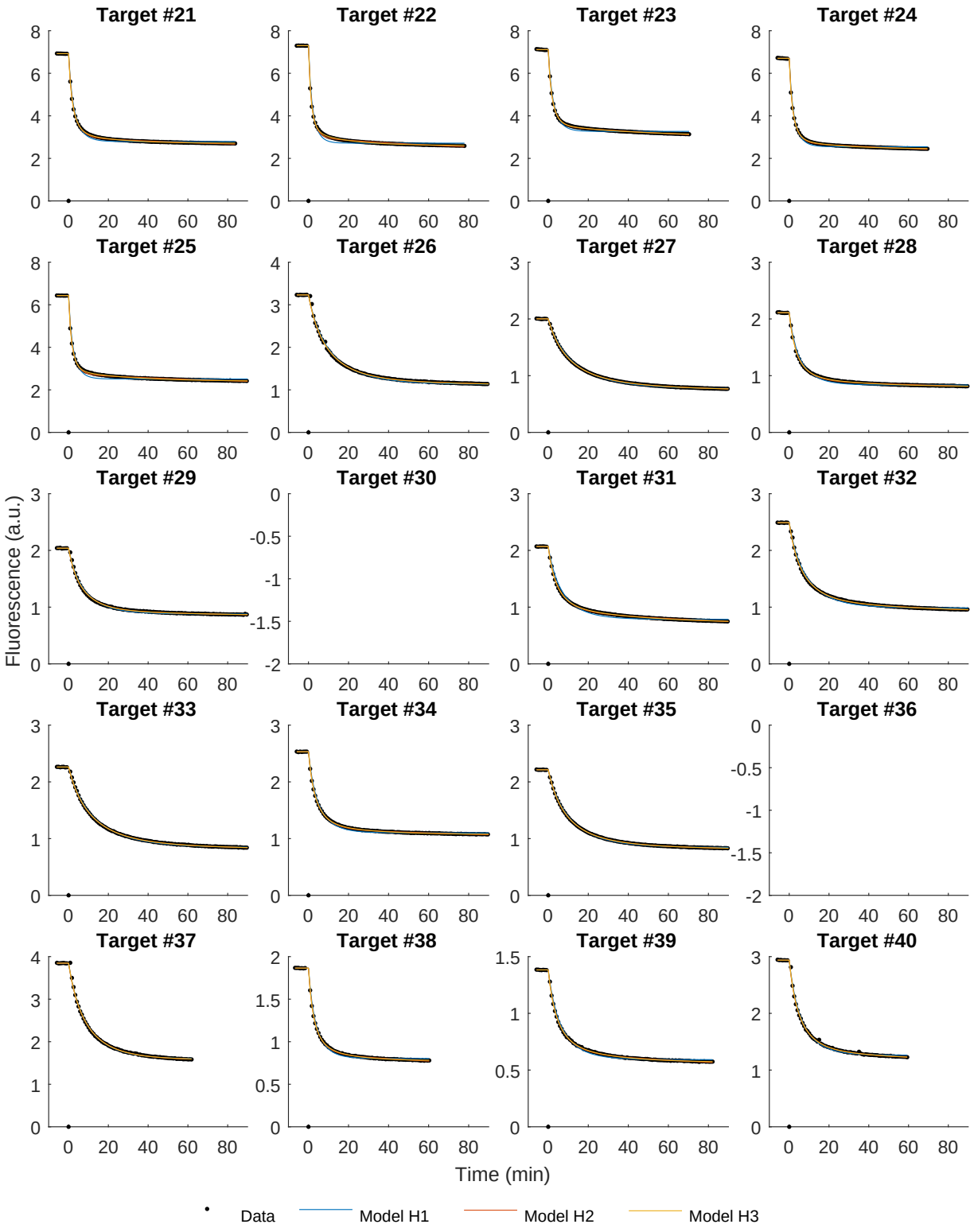

FIG. 2-9: Fluorescence data and best-fit traces for hybridization experiments performed at 55 °C, target sequences 21-40.

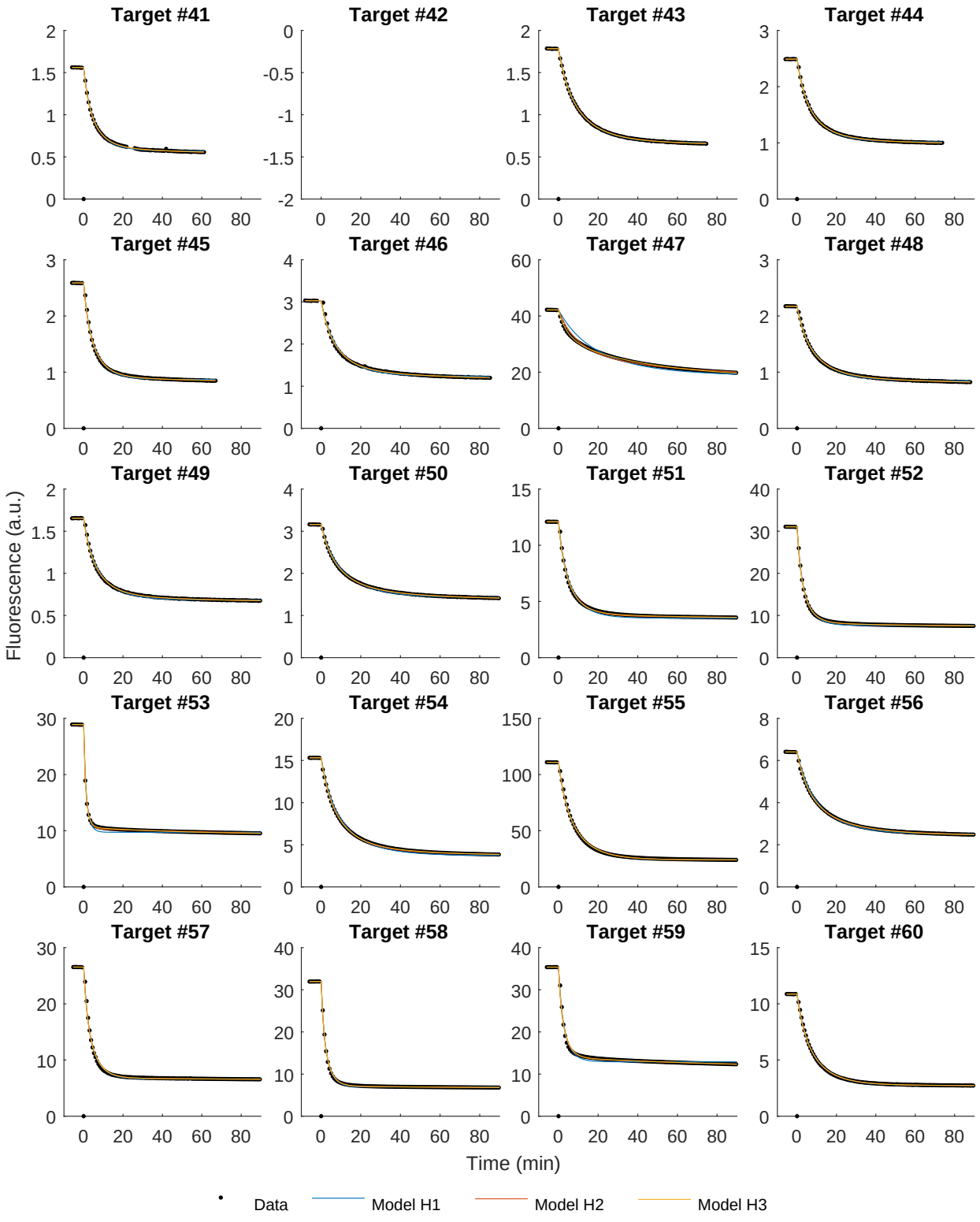

FIG. 2-10: Fluorescence data and best-fit traces for hybridization experiments performed at 55 °C, target sequences 41-60.

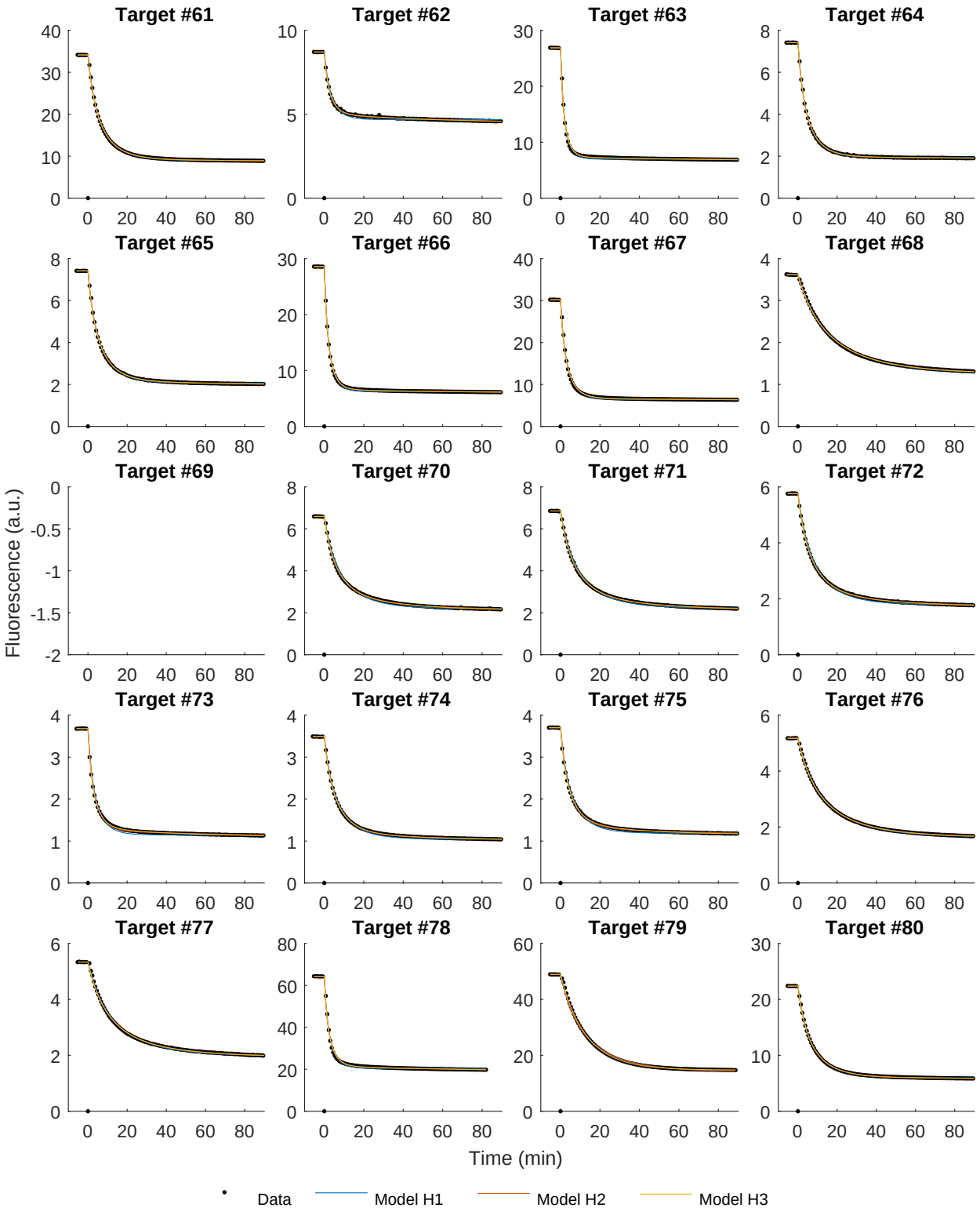

FIG. 2-11: Fluorescence data and best-fit traces for hybridization experiments performed at 55 °C, target sequences 61-80.

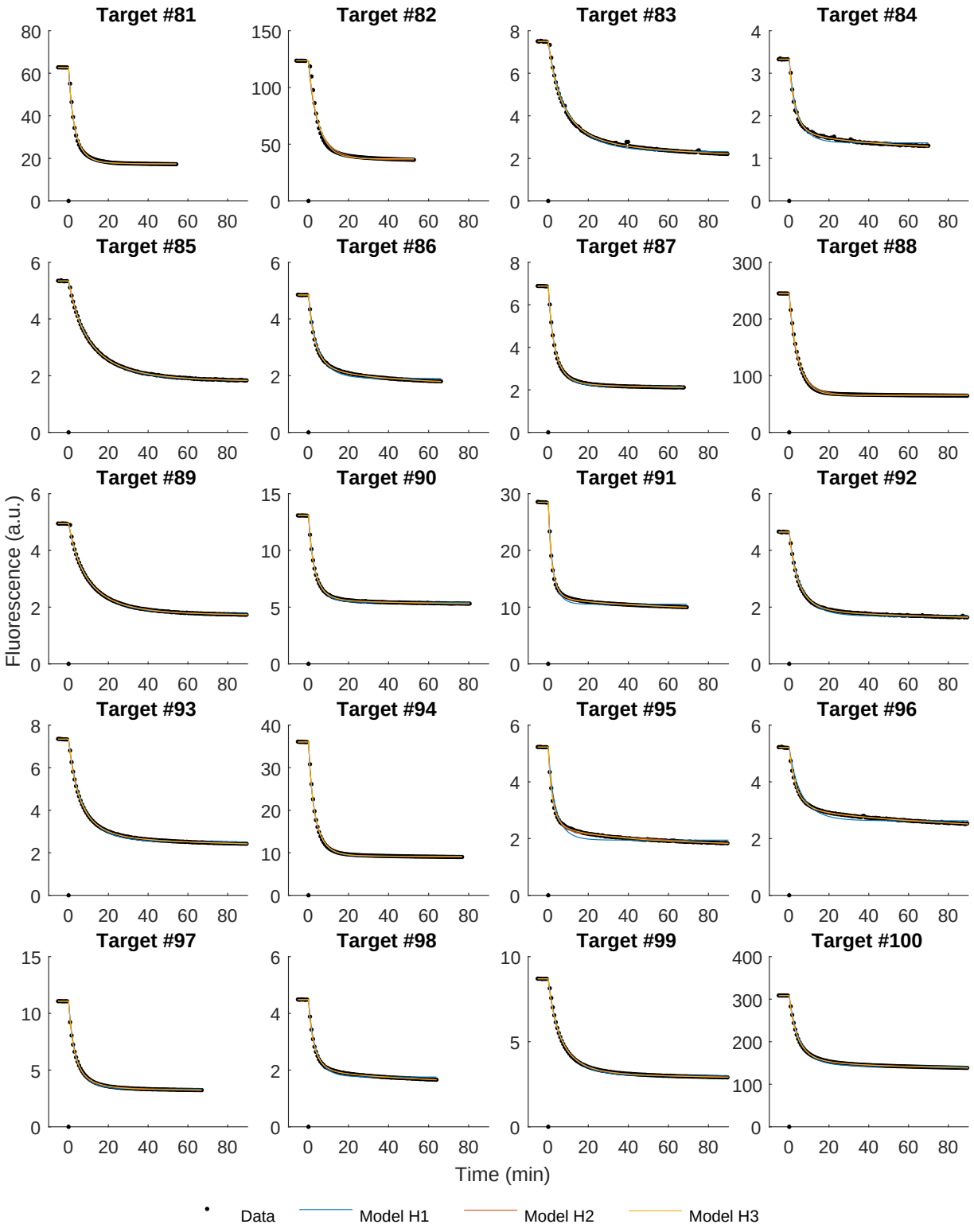

FIG. 2-12: Fluorescence data and best-fit traces for hybridization experiments performed at 55 °C, target sequences 81-100.

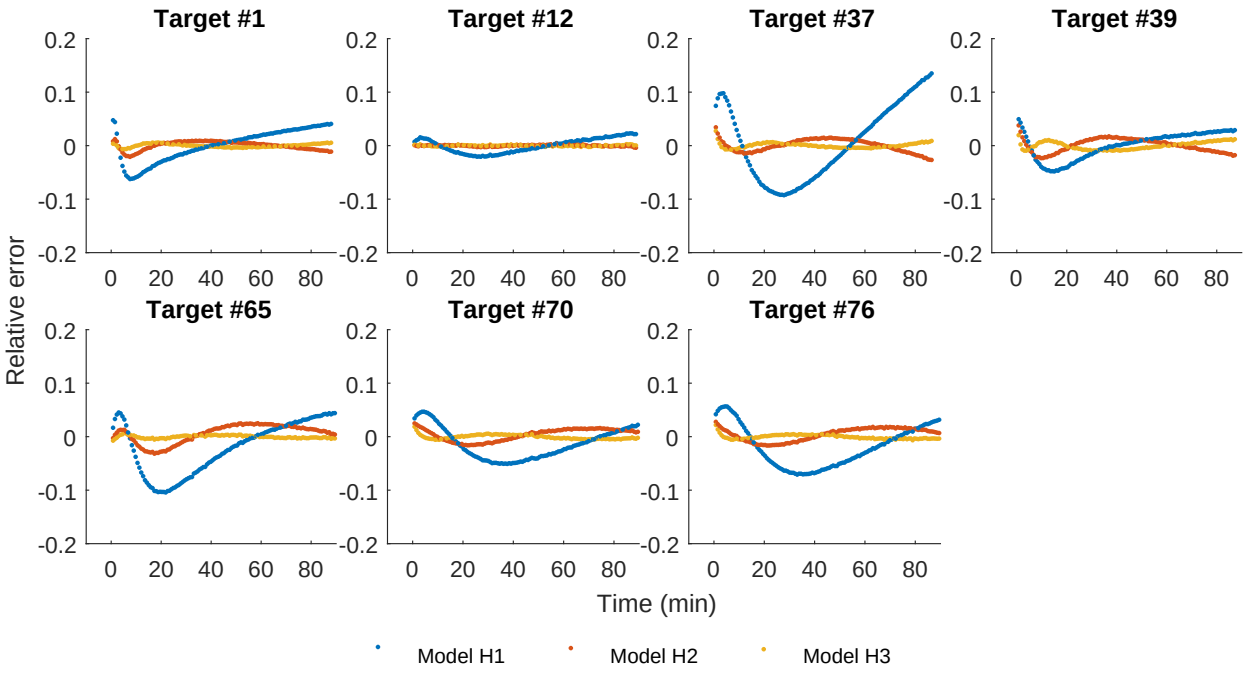

FIG. 2-13: Relative errors of best-fit simulations for hybridization experiments performed at 28 °C.

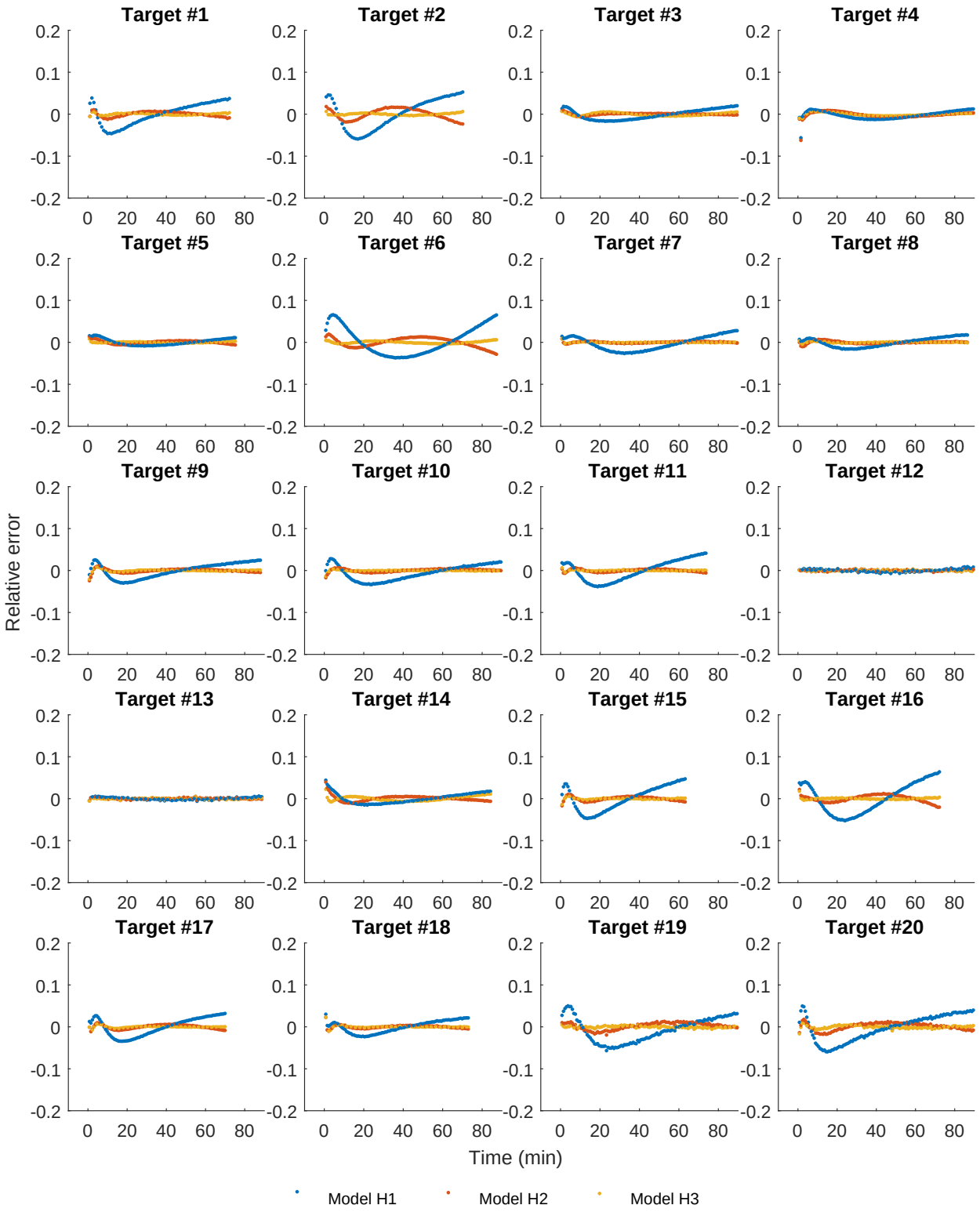

FIG. 2-14: Relative errors of best-fit simulations for hybridization experiments performed at 37 °C, target sequences 1-20.

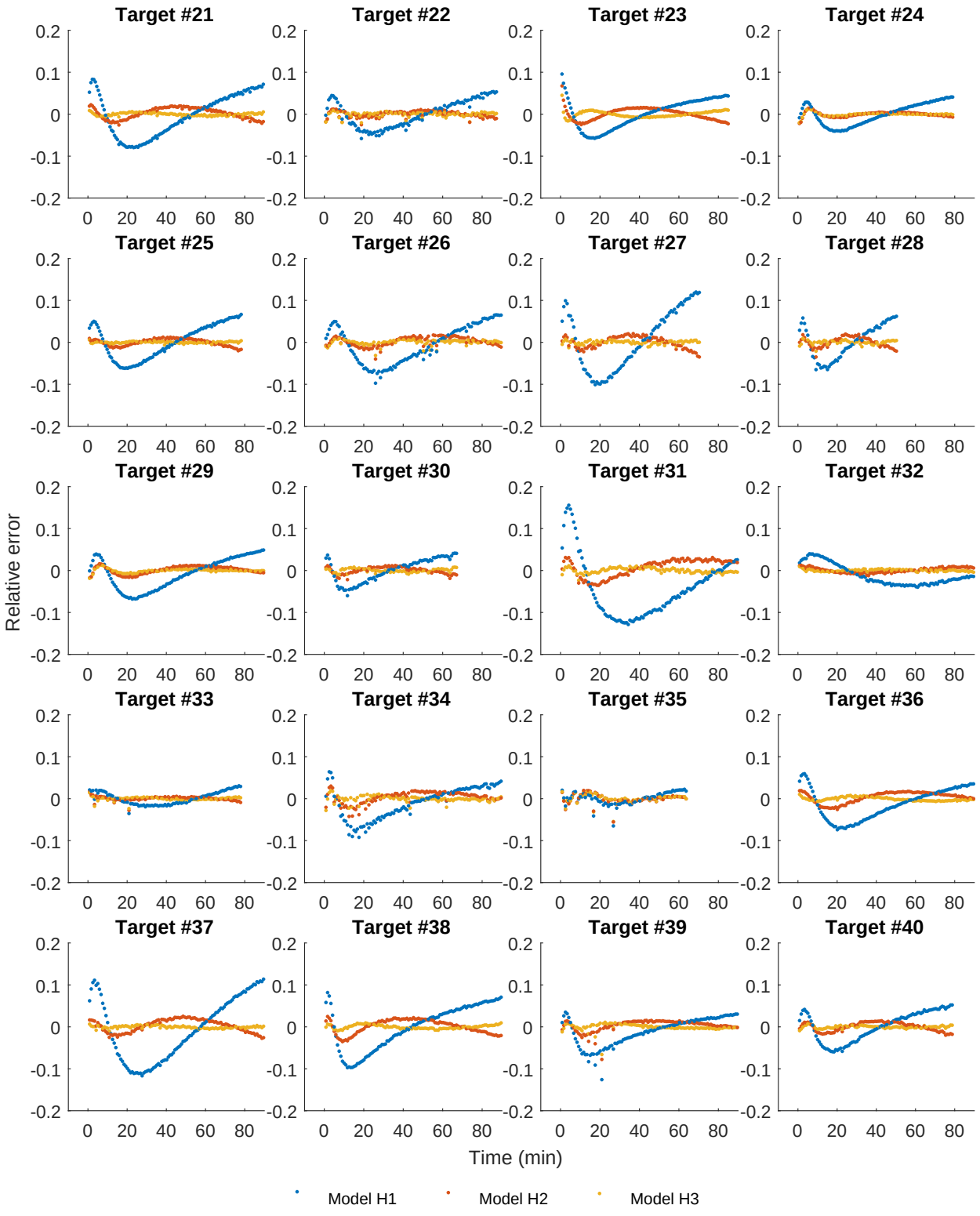

FIG. 2-15: Relative errors of best-fit simulations for hybridization experiments performed at 37 °C, target sequences 21-40.

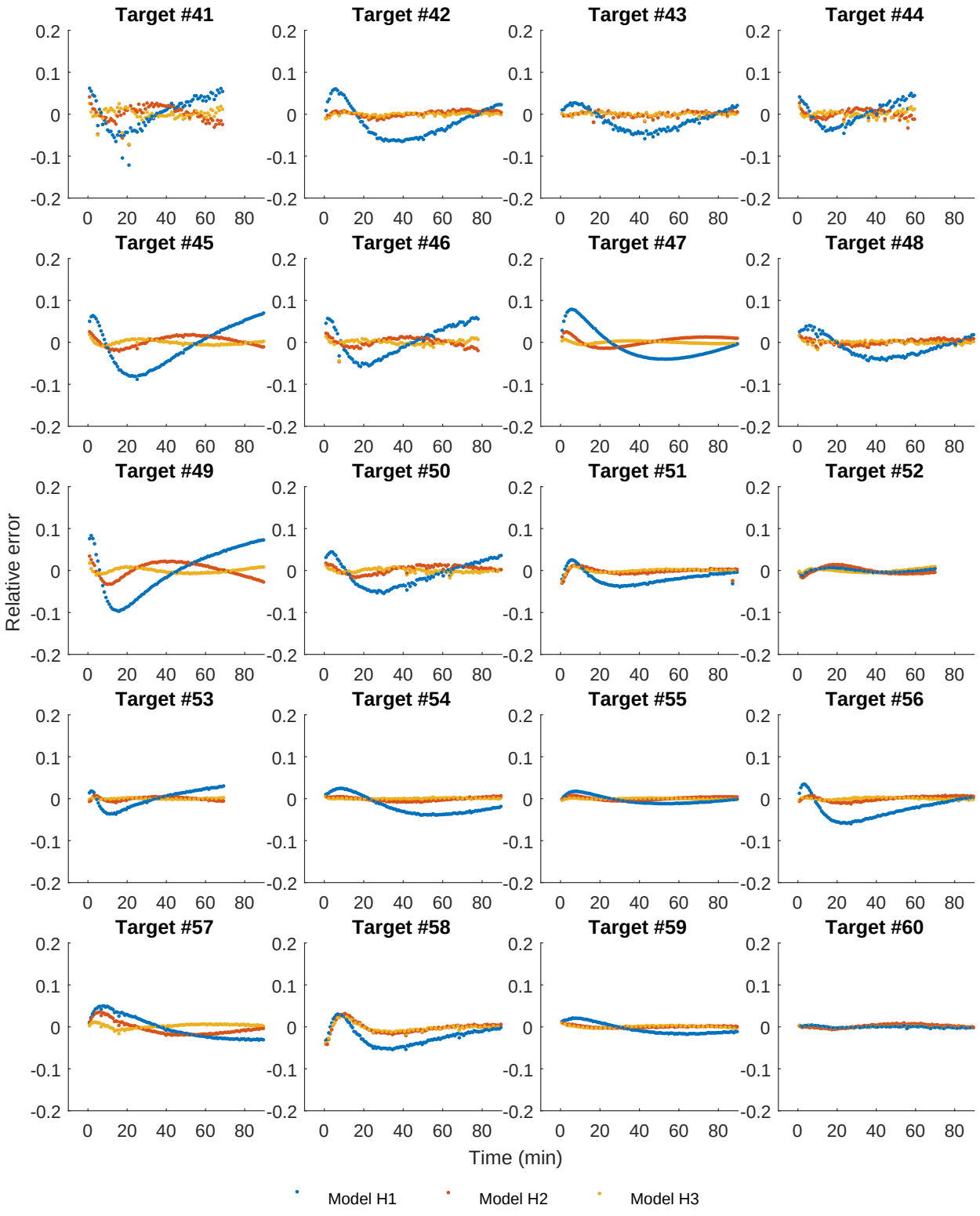

FIG. 2-16: Relative errors of best-fit simulations for hybridization experiments performed at 37 °C, target sequences 41-60.

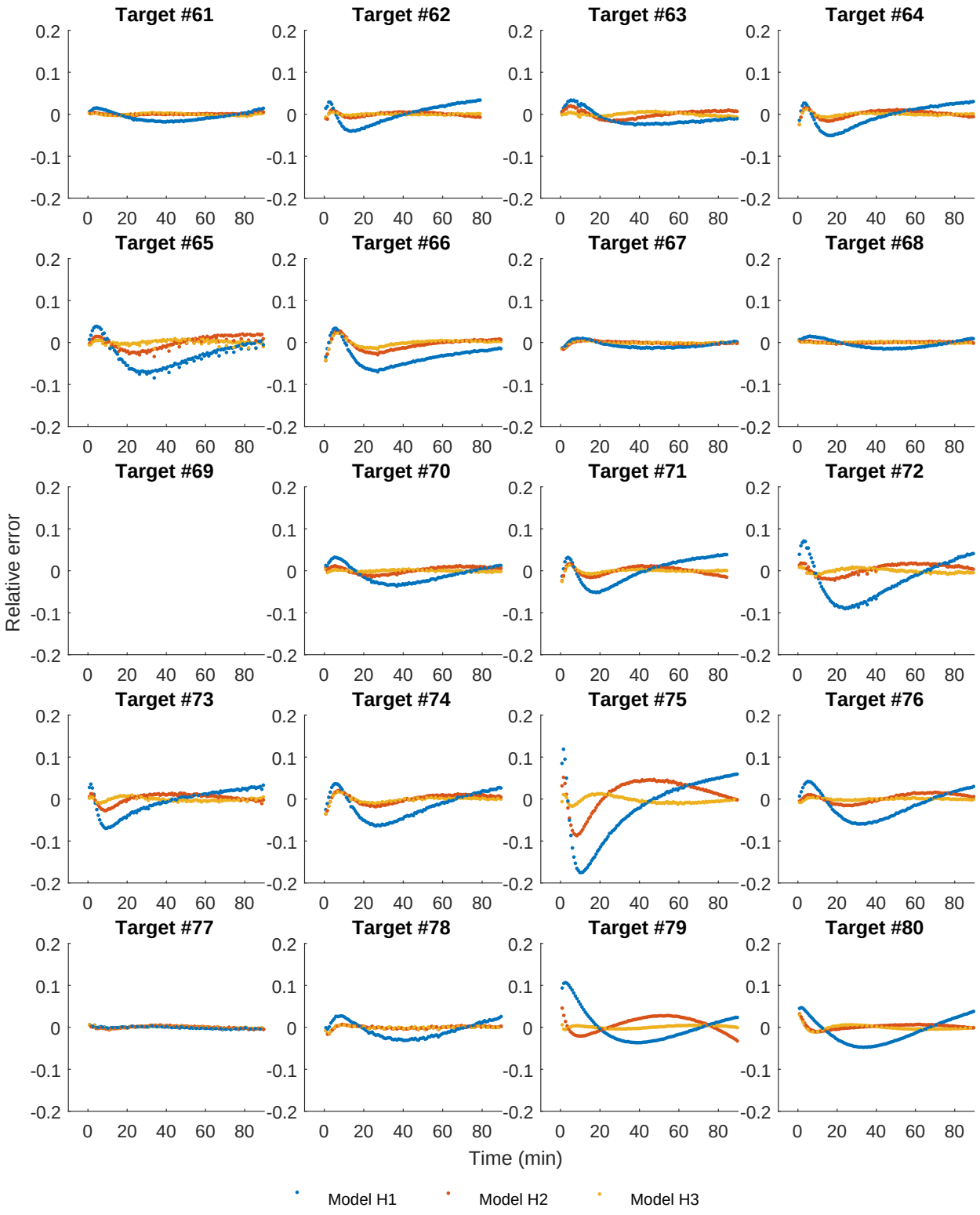

FIG. 2-17: Relative errors of best-fit simulations for hybridization experiments performed at 37 °C, target sequences 61-80.

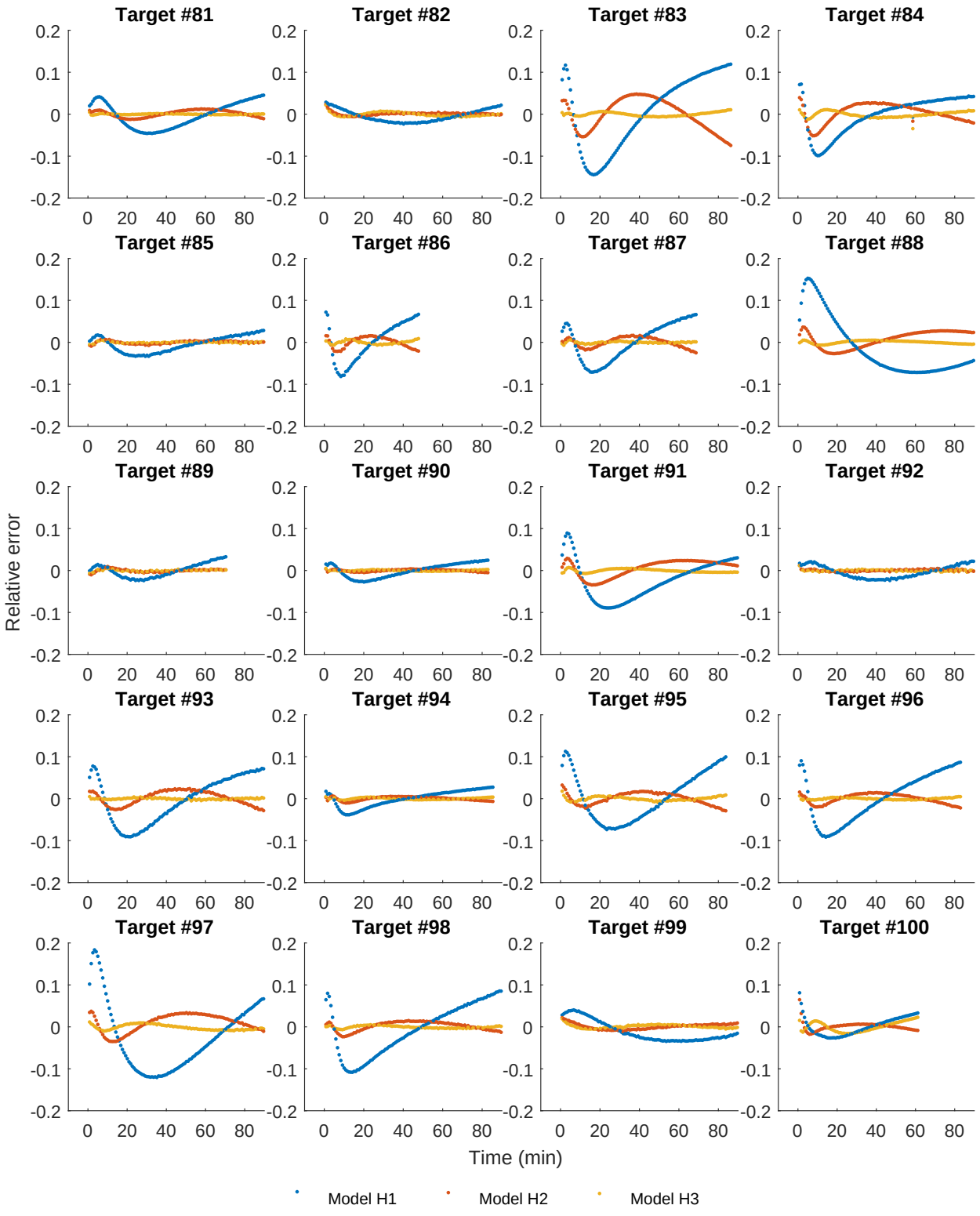

FIG. 2-18: Relative errors of best-fit simulations for hybridization experiments performed at 37 °C, target sequences 81-100.

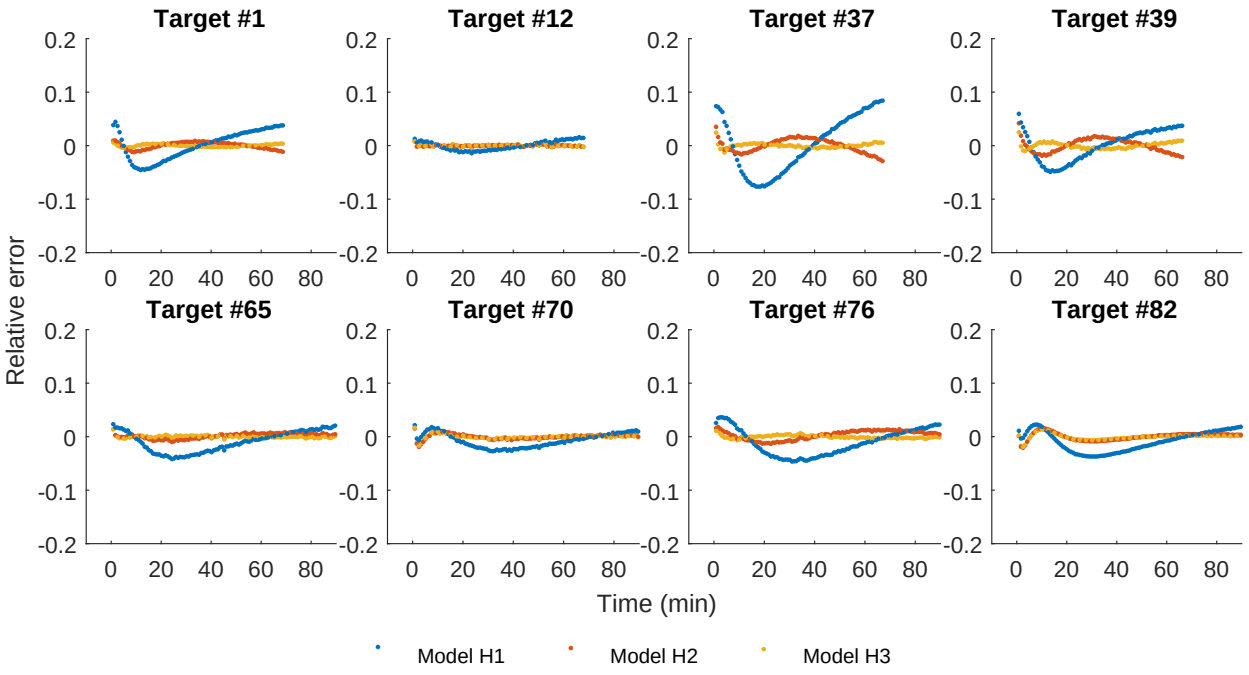

FIG. 2-19: Relative errors of best-fit simulations for hybridization experiments performed at 46 °C.

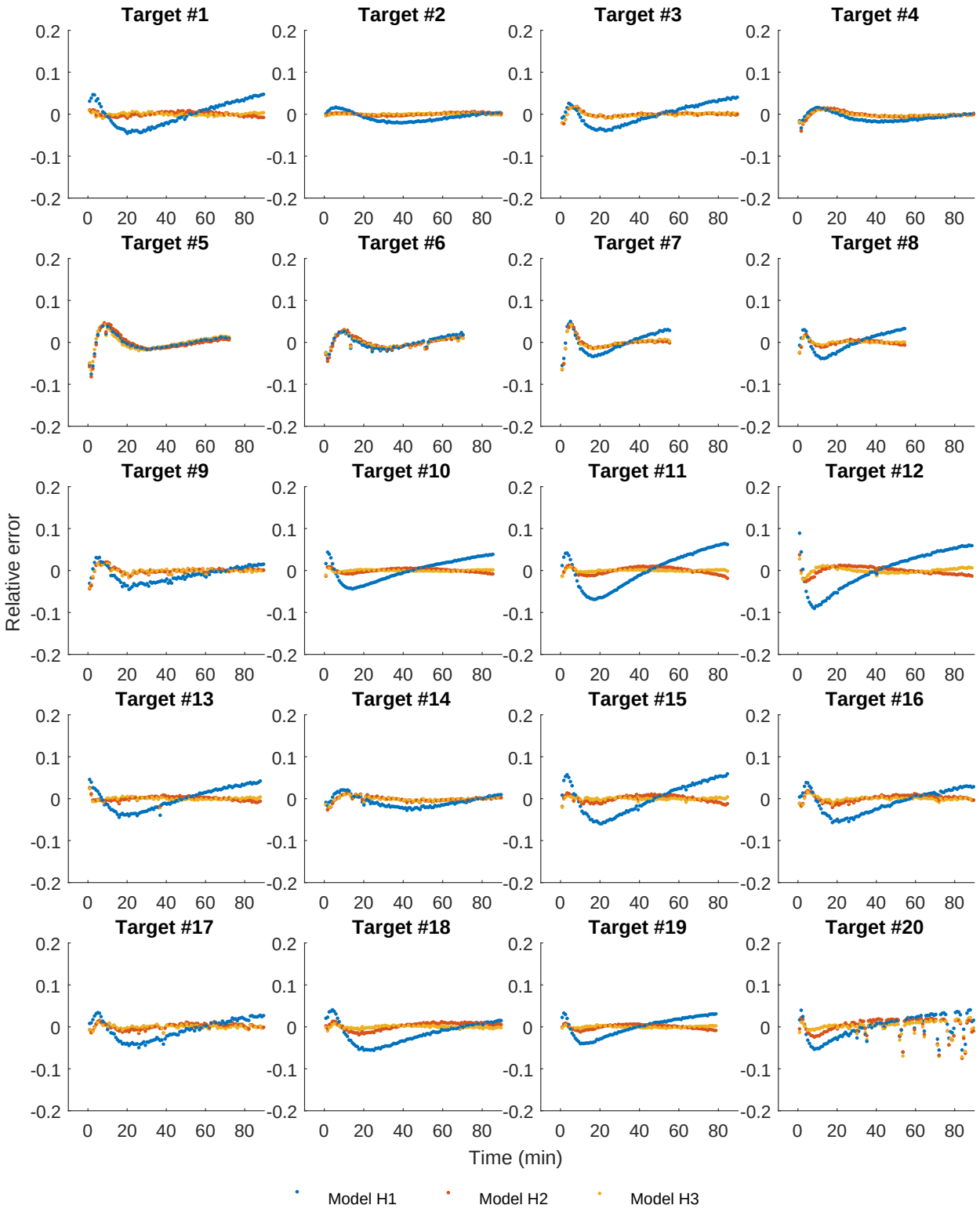

FIG. 2-20: Relative errors of best-fit simulations for hybridization experiments performed at 55 °C, target sequences 1-20.

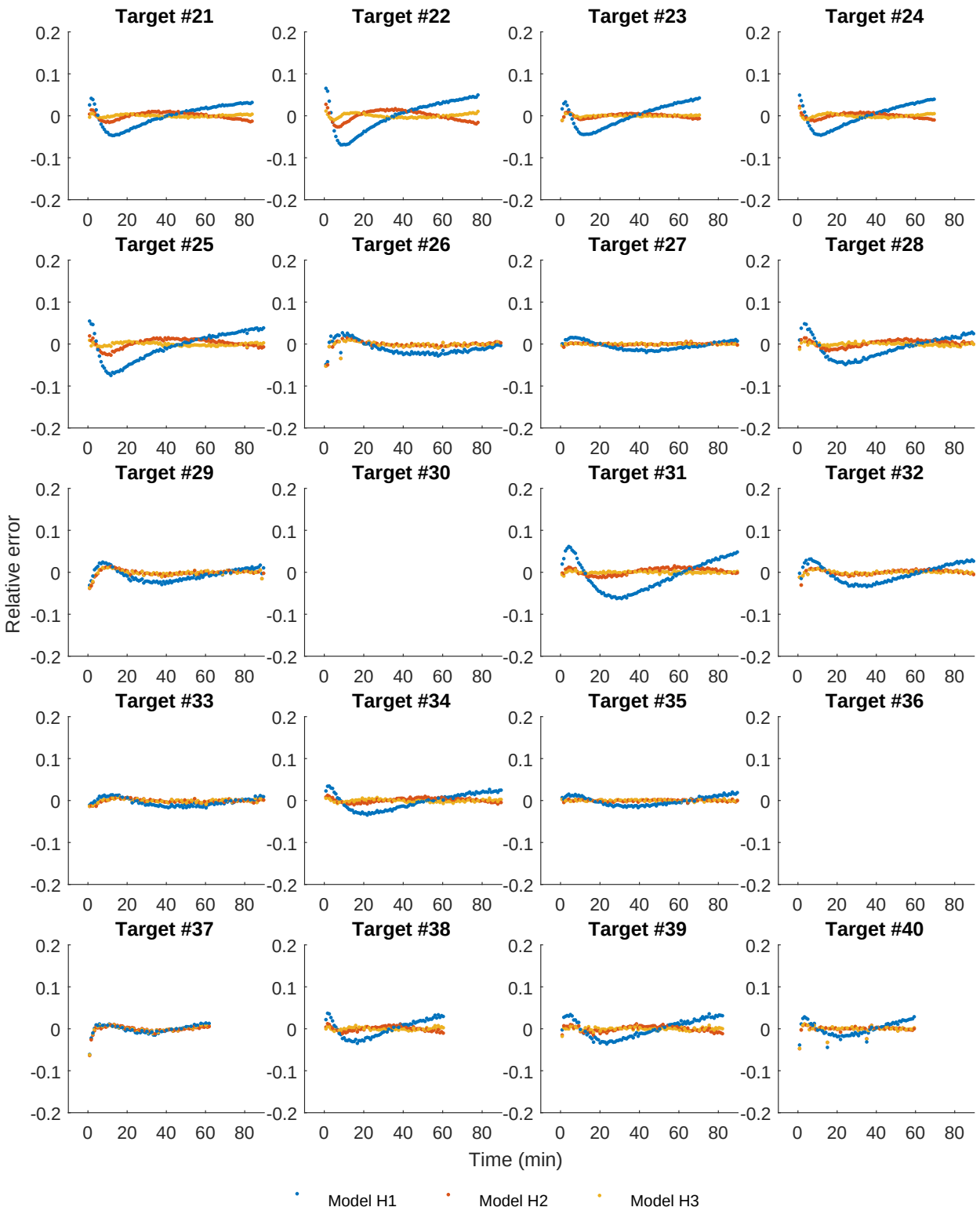

FIG. 2-21: Relative errors of best-fit simulations for hybridization experiments performed at 55 °C, target sequences 21-40.

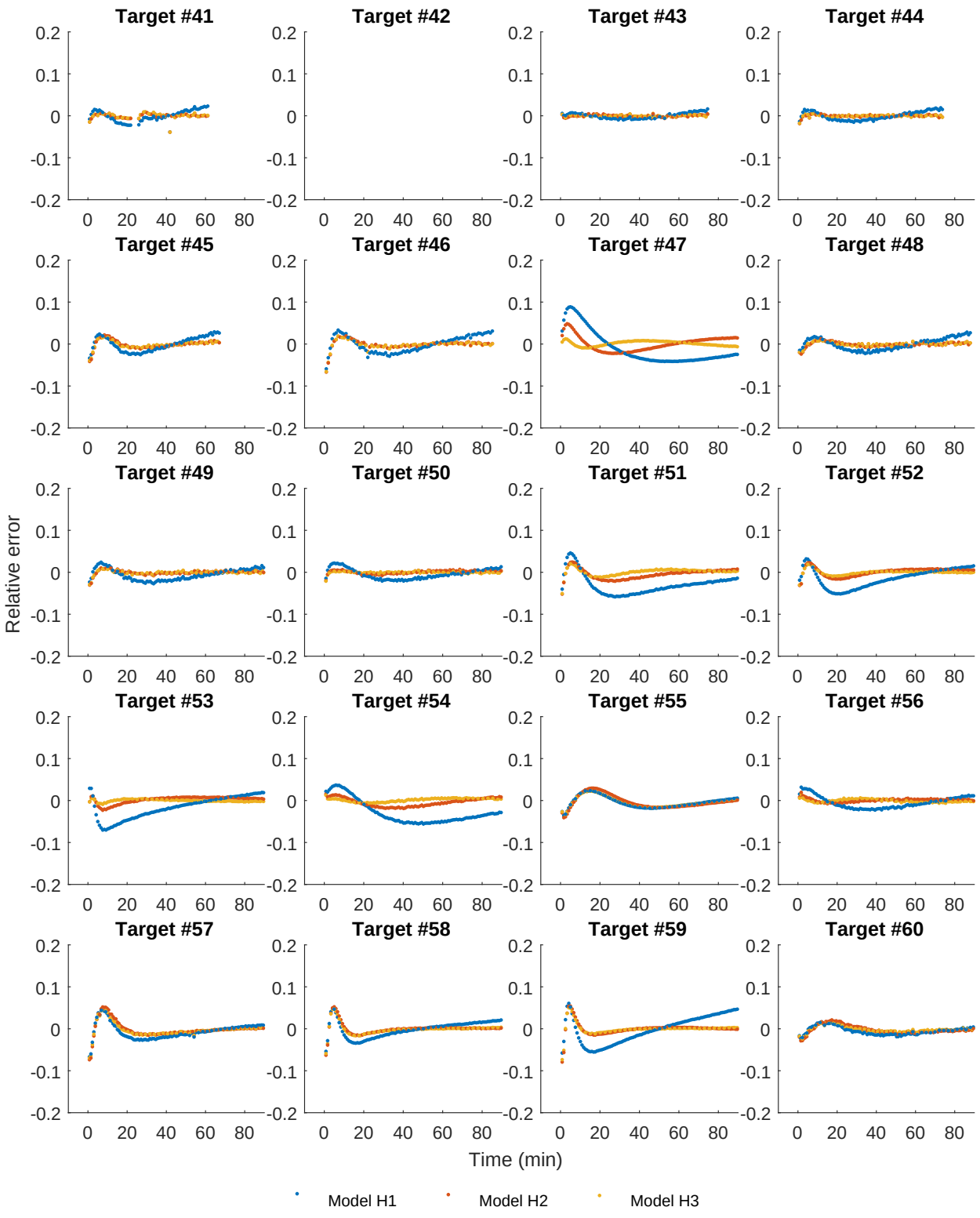

FIG. 2-22: Relative errors of best-fit simulations for hybridization experiments performed at 55 °C, target sequences 41-60.

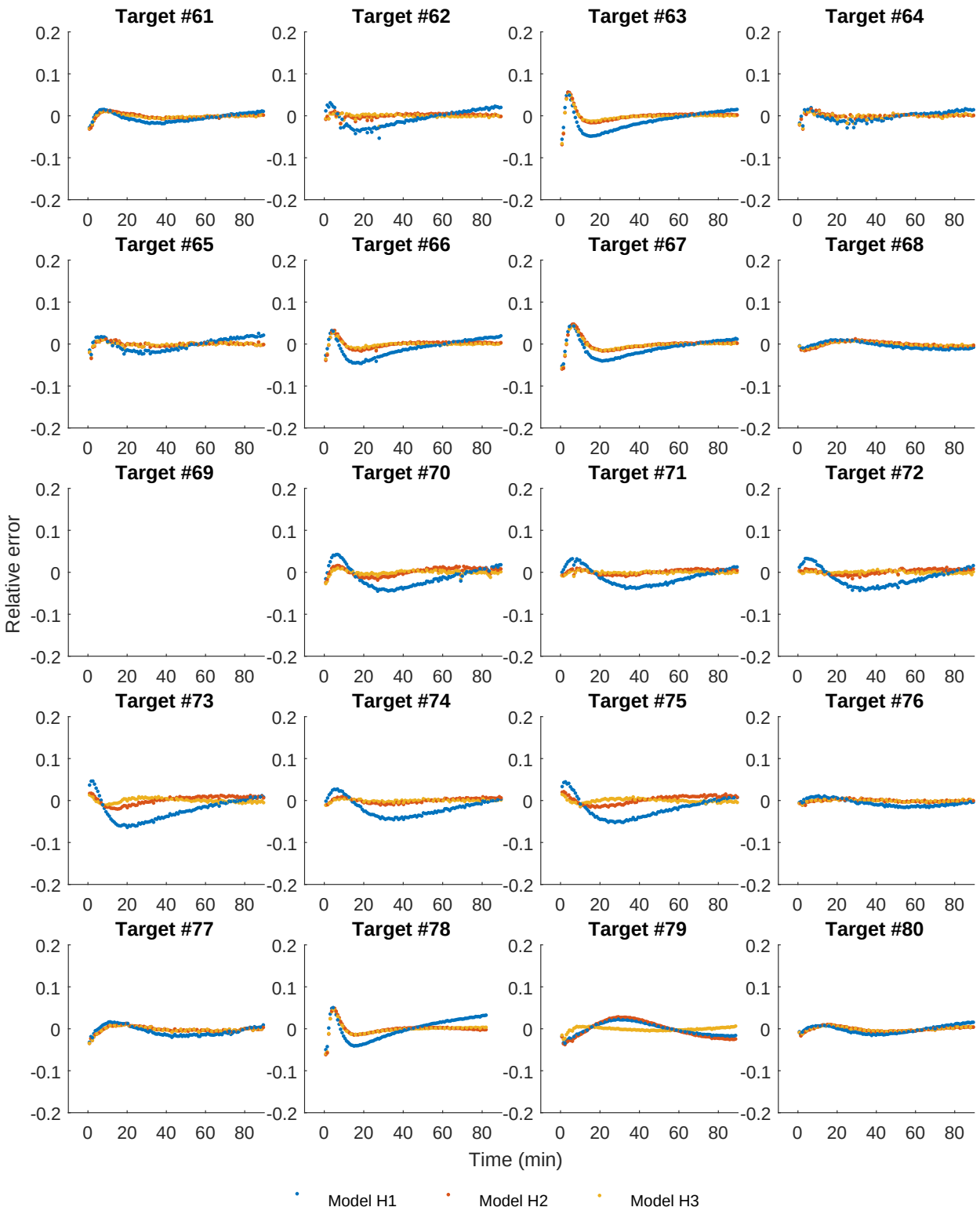

FIG. 2-23: Relative errors of best-fit simulations for hybridization experiments performed at 55 °C, target sequences 61-80.

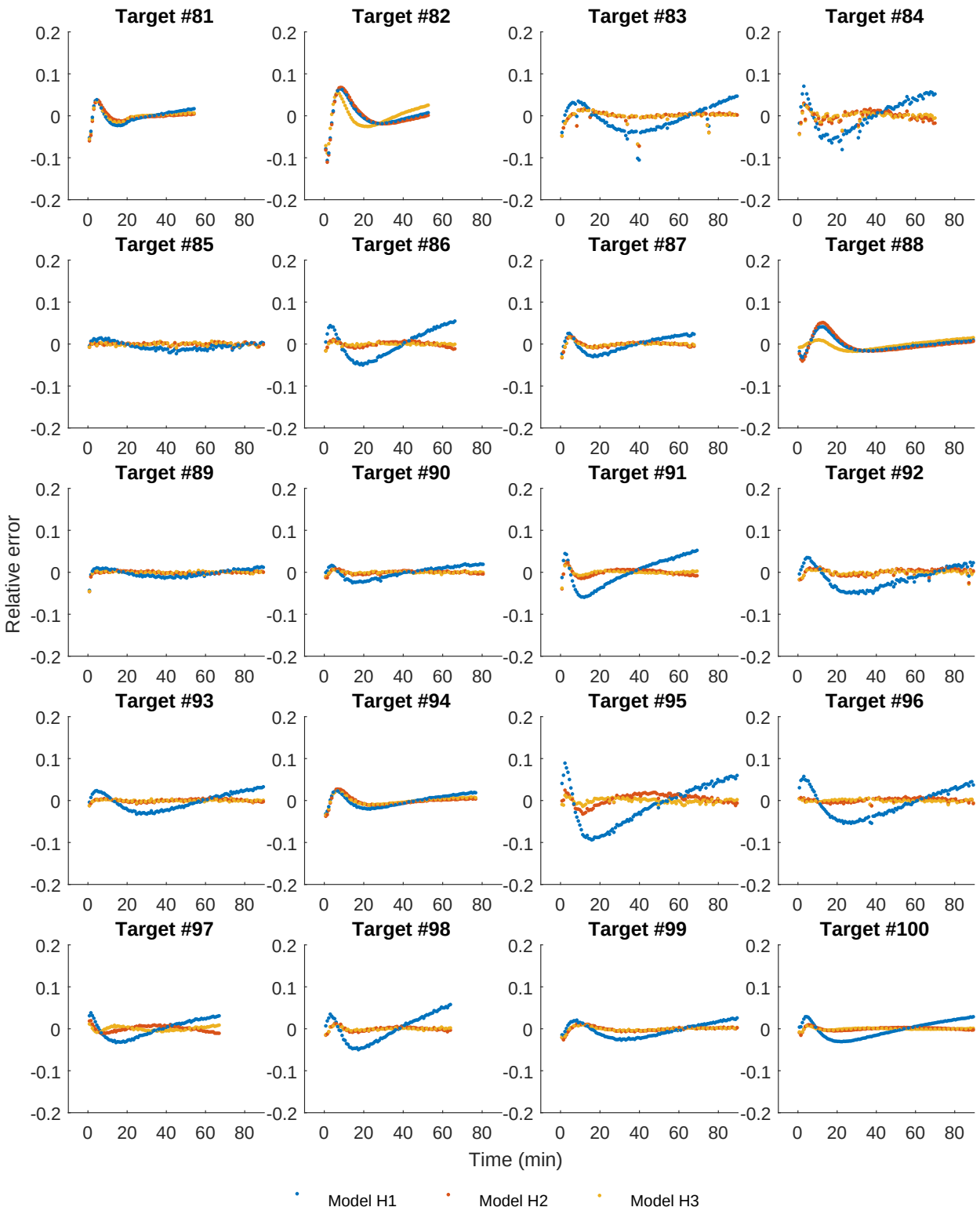

FIG. 2-24: Relative errors of best-fit simulations for hybridization experiments performed at 55 °C, target sequences 81-100.

3. WNV Model Details

WNV Model Overview. To quantitate the similarity or dissimilarity between two hybridization reactions, we abstract each reaction into a number of features. The value of each feature for a particular hybridization reaction is computable based on the sequences of the target and probe, and the reaction temperature and buffer conditions (although buffer is the same for all experiments in this work). Each hybridization reaction can thus be represented by a point in high-dimensional feature space. With an optimally designed and weighted set of features, the two points close in feature space should exhibit similar k_{Hyb} values. The converse is not necessarily true: two hybridization reactions with coincidentally similar k_{Hyb} values may possess very different feature values.

Mapping the hybridization reactions into feature space is important because targets that are similar in sequence space may not have similar hybridization kinetics, and vice versa, due to the sensitivity of secondary structure to small changes in DNA sequence in certain regions, but not in others. As a simple example, oligonucleotide (2) with sequence "ACACACACTTAAAATTGTGTGTGTCCC" has higher Hamming distance [1] to oligo (1) with sequence "ACACACACTTTTTTTTTGTGTGTGTCCC" than oligo (3) with sequence "ACTCAGACTTTTTTTTTGTGTGTGTCCC". However, sequences (1) and (2) differ in the loop region of a hairpin, while sequences (1) and (3) differ in the stem, so we may expect k_{Hyb} to be more differ for the former pair than the latter.

In a WNV model using k different features, the feature space distance between two different hybridization reactions j and m is defined as:

$$d_{j,m} = \sqrt{\sum_{i=1}^{k} (f_i(j) - f_i(m))^2}$$

where $f_i(j)$ is the value of weighted feature i for reaction j.

Comparison of WNV to linear regression in predicting interpolation and extrapolation values. The WNV model we present outperformed simpler models such as multilinear regression (MLR) in predicting DNA hybridization rate constants k_{Hyb}. To explain to the reader why WNV outperforms linear regression, here we describe a simple interpolation/extrapolation problem, and show the predictions produced by WNV vs. by linear regression.

Fig. 3-1a shows a semi-circular relationship between the feature value and the desired parameter. MLR in the context of a single feature is simply linear regression (black line); MLR poorly captures the relationship between the feature and the desired parameter despite using 2 fitting parameters (slope and intercept). WNV, in contrast, uses only a single fitting parameter (feature weight), and produces predictions that captures the relationship between the feature and the desired parameter.

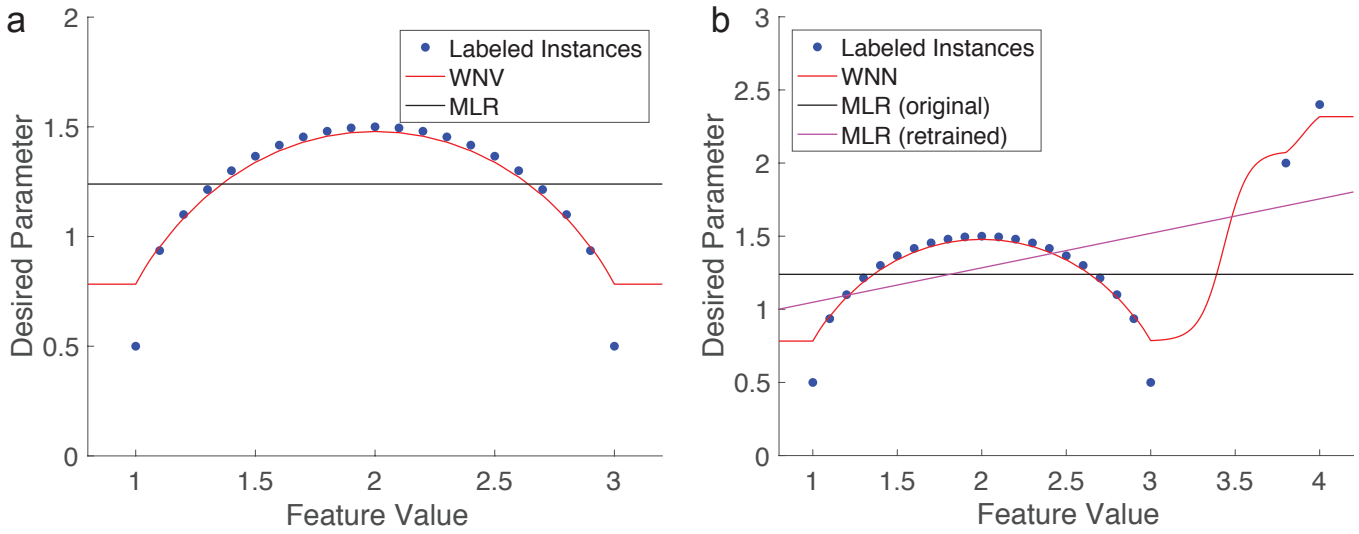

FIG. 3-1: Dependence of k_{Hyb} on initial features 1 through 6.

WNV is also more scalable to addition of new data. Fig. 3-1b shows the introduction of two additional data points to the right. The original MLR model does not dynamically adjust based on new information; model

retraining is needed to produce a new MLR model (slope and intercept values) that fits the new data better. In contrast, WNV adapts to new data and changes predictions for larger feature values based on the new information, without requiring any adjustment to the feature weight fitting parameter.

Feature Weighting. The distributions of values for many features were distinctively non-Gaussian for our 210 reactions; furthermore, different features had value spreads that varied dramatically. Feature weighting was performed based on the interquartile range: the 75th percentile feature value is mapped to a score of $+\frac{w(i)}{2}$, and the 25th percentile value is mapped to $-\frac{w(i)}{2}$ (Fig. 4d), where $w(i)$ is the weight of feature i. Because a feature i with larger weight $w(i)$ allows a larger range of scores, it can contribute more to the distance between two hybridization reactions.

Hybridization Kinetics Feature Selection and Weighting. We started by rationally designing over 50 potential features, each based on some aspect of DNA biophysics that we believed may influence kinetics. Supplementary Section 4 shows the 35 features that showed significant correlation with k_{Hyb} by itself; other features were removed from consideration. There may not be good physical interpretations of all effective features—in these cases, the feature in question is likely correlated with a yet-undiscovered complex feature with a firm physical basis. From this second list of 35 features, we then implemented a greedy algorithm:

First, we start constructing 1-feature WNV models for each of the 35 features. For single-feature WNV model, we performed numerical optimization to find the feature weight that yields the best leave-one-out (LOO) prediction Badness for k_{Hyb}. The different single-feature WNV models, using each's optimized feature weight, were then compared against each other, in order to select the best single-feature WNV model. The nGp feature (Nupack ensemble standard free energy of the probe) emerged as the single feature that best-predicts k_{Hyb} in a 1-feature WNV setting.

Next, we considered the 34 possible 2-feature WNV models including nGp and each of the remaining features. Once again, we performed numerical optimization to determine feature weights that product the best LOO prediction Badness, and Pap was selected as the second feature. This process continued to 10-feature WNV models. We observed that prediction Badness did not significantly improve after the 6-feature WNV model, so the 6-feature WNV model was selected as the final prediction model. Finally, the 6-feature WNV model was subjected to intensive numerical optimization in order to determine final feature weights.

---

[1] Hamming distance is defined as the number of single nucleotide replacements needed to convert the first sequence into the second.

4. Feature List

Details of the 6 Final Features.

The 6 features used were nGp, Pap, Temperature, wPat, GavgMSR1, and Gb, with respective feature weights of 7.96, 15.12, 10.55, 4.44, 10.90, and 18.69. The mathematical definitions for each feature shown displayed below.

1. nGp. This feature denotes the partition function (ensemble) standard free energy of folding of the probe oligonucleotide P. For a given sequence, this can be directly computed using Nupack via the "pfunc" command.

2. Pap. This feature denotes the probability-averaged standard free energy of hybridization of the strongest continuous unpaired stretch of nucleotides in the probe oligonucleotide P, in possible states of P with above 0.5% probability at equilibrium. Mathematically,

$$\text{Pap} \;=\; \sum_{s} \Pr(P_s) \cdot \text{H}(\Pr(P_s) > \mathcal{S}) \cdot \min\{\Delta G^{\circ}(P_{s,i:j})\}$$

$\Delta G^{\circ}(P_{s,i:j})$ is the standard free energy of hybridization of the subsequence of P from positions i through j to its perfect complement, given that all nucleotides from i through j are unpaired in state P_s at equilibrium. $\text{H}(\cdot)$ is the Heaviside step function that returns 1 when the input is positive and 0 when the input is negative. For example, for a particular state of P with the dot-paren notation of

```
.....(((...)))...
```

positions 1 through 5, 9 through 11, and 15 through 17 are unpaired, so $\min\{\Delta G^{\circ}(P_{s,i:j})\}$ would consider only $\Delta G^{\circ}(P_{s,1:5})$, $\Delta G^{\circ}(P_{s,9:11})$, and $\Delta G^{\circ}(P_{s,15:17})$.

3. Temperature. This is simply the temperature of the hybridization reaction, in Celsius.

4. wPat. This feature denotes the probability-averaged standard free energy of hybridization of the strongest continuous unpaired exposed stretch of nucleotides in the target oligonucleotide T, in possible states of T with above 0.5% probability at equilibrium. Mathematically,

$$\text{Pap} \;=\; \sum_{s} \Pr(T_s) \cdot \text{H}(\Pr(T_s) > \mathcal{S}) \cdot \min_{i,j}\{\Delta G^{\circ}(T_{s,i:j})\}$$

$\Delta G^{\circ}(T_{s,i:j})$ is the standard free energy of hybridization of the subsequence of T from positions i through j to its perfect complement, given that all nucleotides from i through j are unpaired in state T_s at equilibrium. For example, for a particular state of T with the dot-paren notation of

```
.....(((...)))...
```

positions 1 through 5, 9 through 11, and 15 through 17 are unpaired. However, the nucleotides between positions 9 through 11 are not considered exposed, since they are an internal part of a hairpin secondary structure. Therefore, $\min\{\Delta G^{\circ}(T_{s,i:j})\}$ would consider only $\Delta G^{\circ}(T_{s,1:5})$ and $\Delta G^{\circ}(T_{s,15:17})$.

5. dGavgMSR1. This feature denotes the standard free energy of hybridization of all subsequences of nucleotides i through j with $1 \le i < j \le 36$, weighted by the probability of all nucleotides being in unbound states at equilibrium. Mathematically,

$$\begin{aligned}
\text{dGavg} \;&=\; \sum_{i,j} \Pr_{\text{free}}(T_{i:j}) \cdot \Delta G^{\circ}(T_{i:j}) \\
\Delta G^{\circ}(T_{i:j}) \;&=\; \sum_{a=i}^{j-1} \Delta G^{\circ}_{\text{stack}}(T_{a:(a+1)}) + \Delta G_{\text{term}} \\
\Pr_{\text{free}}(T_{i:j}) \;&=\; \Pr_{\text{bound}}(T_{i-1}) \cdot \Pr_{\text{bound}}(T_{j+1}) \cdot \prod_{a=i}^{j} \Pr_{\text{free}}(T_a)
\end{aligned}$$

$\Delta G^{\circ}_{\text{stack}}(T_{a:(a+1)})$ corresponds to the standard free energy of binding of a base stack and ΔG_{term} is an terminal AT penalty. These values are based on parameters found in ref. [1] and temperature and salt correction is performed as

described therein. $\Pr_{\text{free}}(T_i)$ corresponds to the probability of the ith nucleotide in target T being in unhybridized state at equilibrium (in the absence of probe P), and values can be calculated using the Nupack "pairs" function [2]. Similarly, $\Pr_{\text{bound}}(T_i)$ corresponds to the probability of the ith nucleotide in target T being in hybridized state at equilibrium, which is equal to $1 - \Pr_{\text{free}}(T_i)$ if nucleotide i is within the bounds of the target T and 1 otherwise. Including the terms $\Pr_{\text{bound}}(T_{i-1})$ and $\Pr_{\text{bound}}(T_{j+1})$ ensures that the exposed regions on the target that are considered are maximal, i.e. they are terminated on each side by either a hybridized nucleotide or a terminus of the target T.

6. Gb. This feature corresponds to the probability-averaged standard free energy of formation of the TP complex for states with above 0.5% probability at equilibrium. Mathematically,

$$\text{Gb} \;=\; \sum_{s} \Pr(TP_s) \cdot \text{H}(\Pr(TP_s) > \mathcal{S}) \cdot \Delta G^{\circ}(TP_s)$$

$\Pr(TP_s)$ denotes the probability of state TP_s at equilibrium, and can be calculated using the Nupack "subopt" function. $\mathcal{S}$ denotes the substructure threshold (currently set to 0.005). $\Delta G^{\circ}(TP_s)$ is the standard free energy of formation of state TP_s, and can be calculated using the Nupack "energy" function.

List of Feature Correlations. Figs. 4-1 through 4-6 show the plot of k_{Hyb} vs. the feature values for 35 of the features with significant correlation and reasonable individual prediction performance in the WNV model; the 6 final features are boxed in green. Except for temperature, all feature units are in kcal/mol.

[1] SantaLucia, J. & Hicks, D. The Thermodynamics of DNA Structural Motifs. *Ann. Rev. Biochem.* **33,** 415-440 (2004).
[2] Zadeh, J. N. *et al.* (2011). NUPACK: analysis and design of nucleic acid systems. Journal of computational chemistry, 32(1), 170-173.

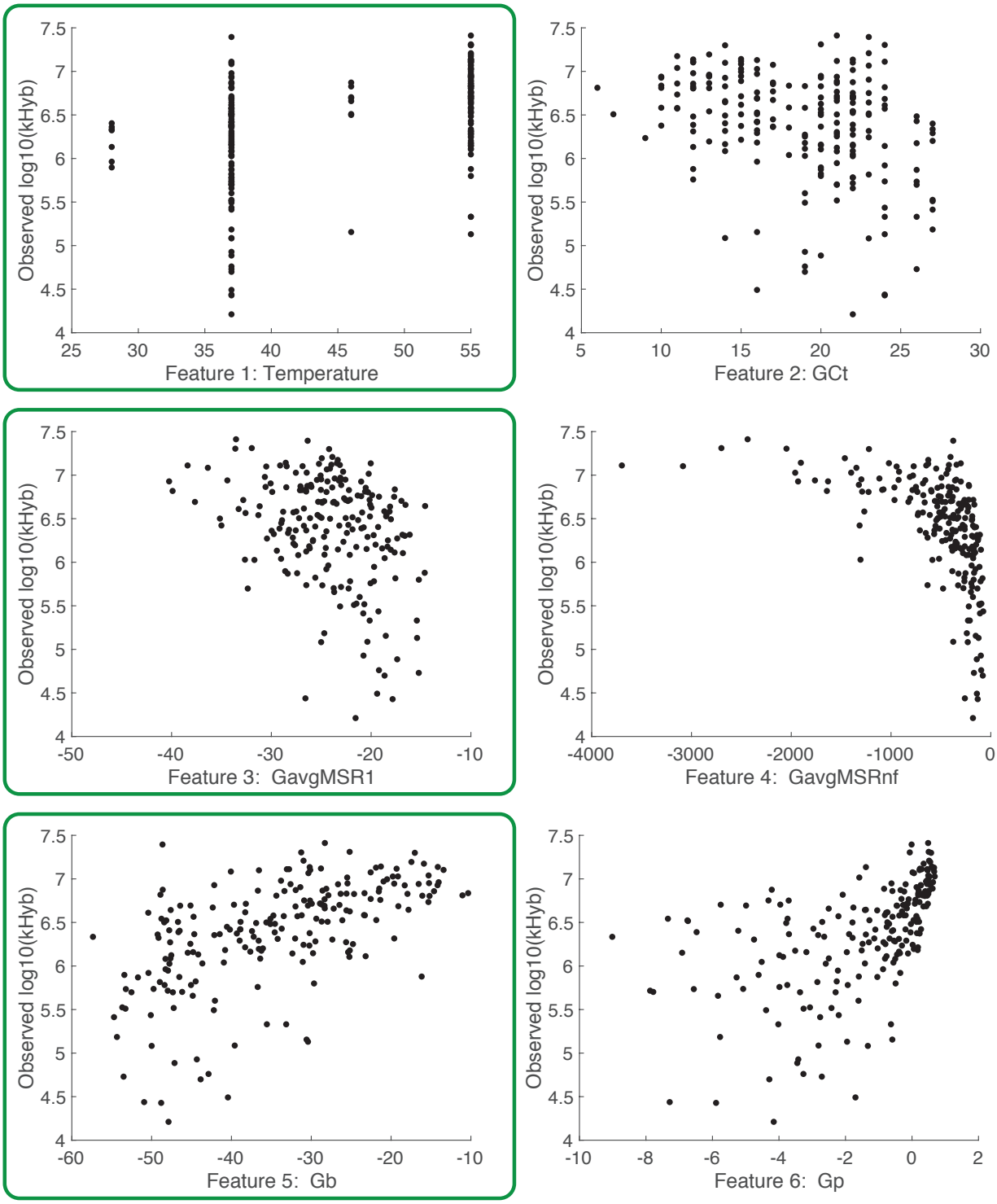

FIG. 4-1: Dependence of k_{Hyb} on initial features 1 through 6.

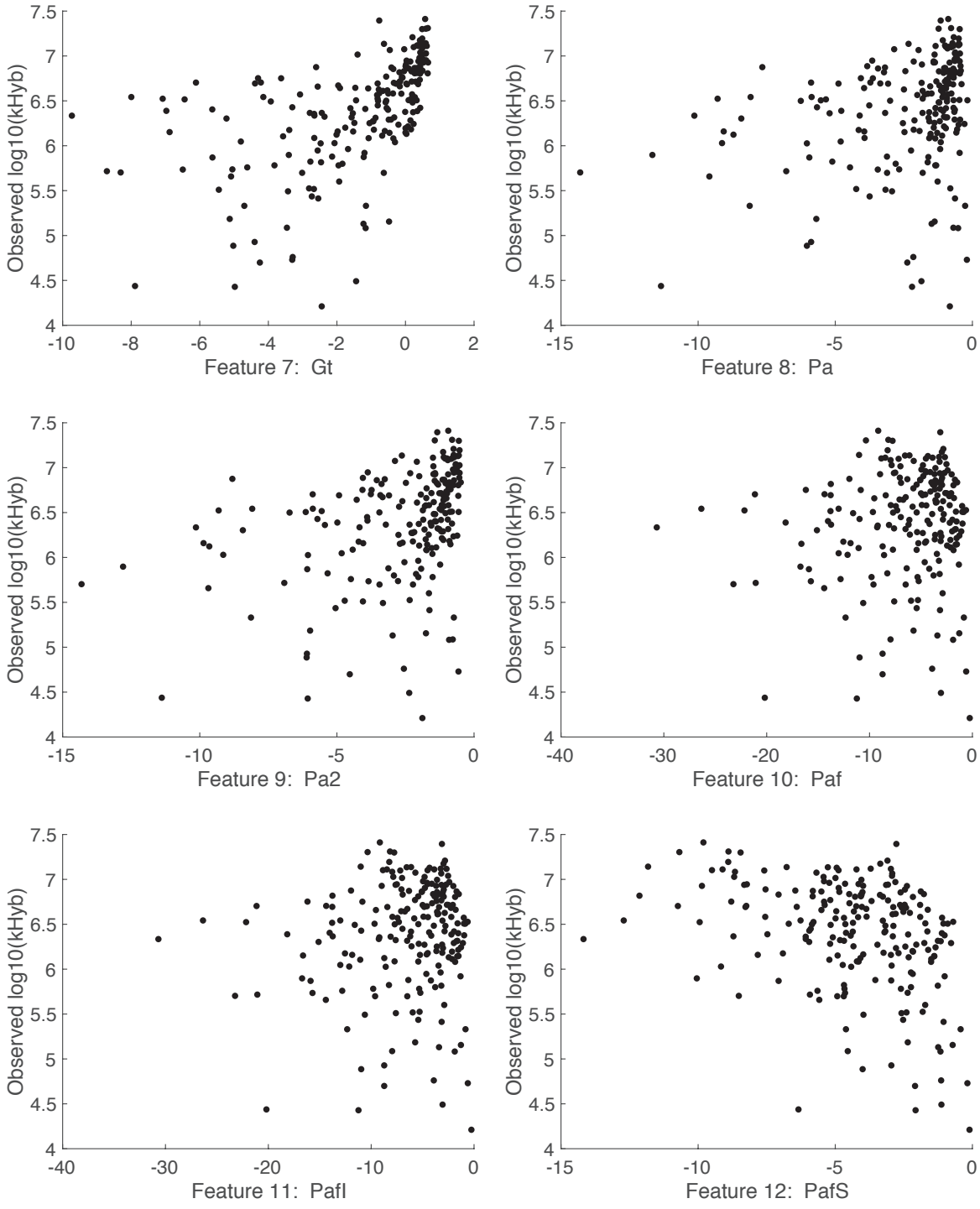

FIG. 4-2: Dependence of k_{Hyb} on initial features 7 through 12.

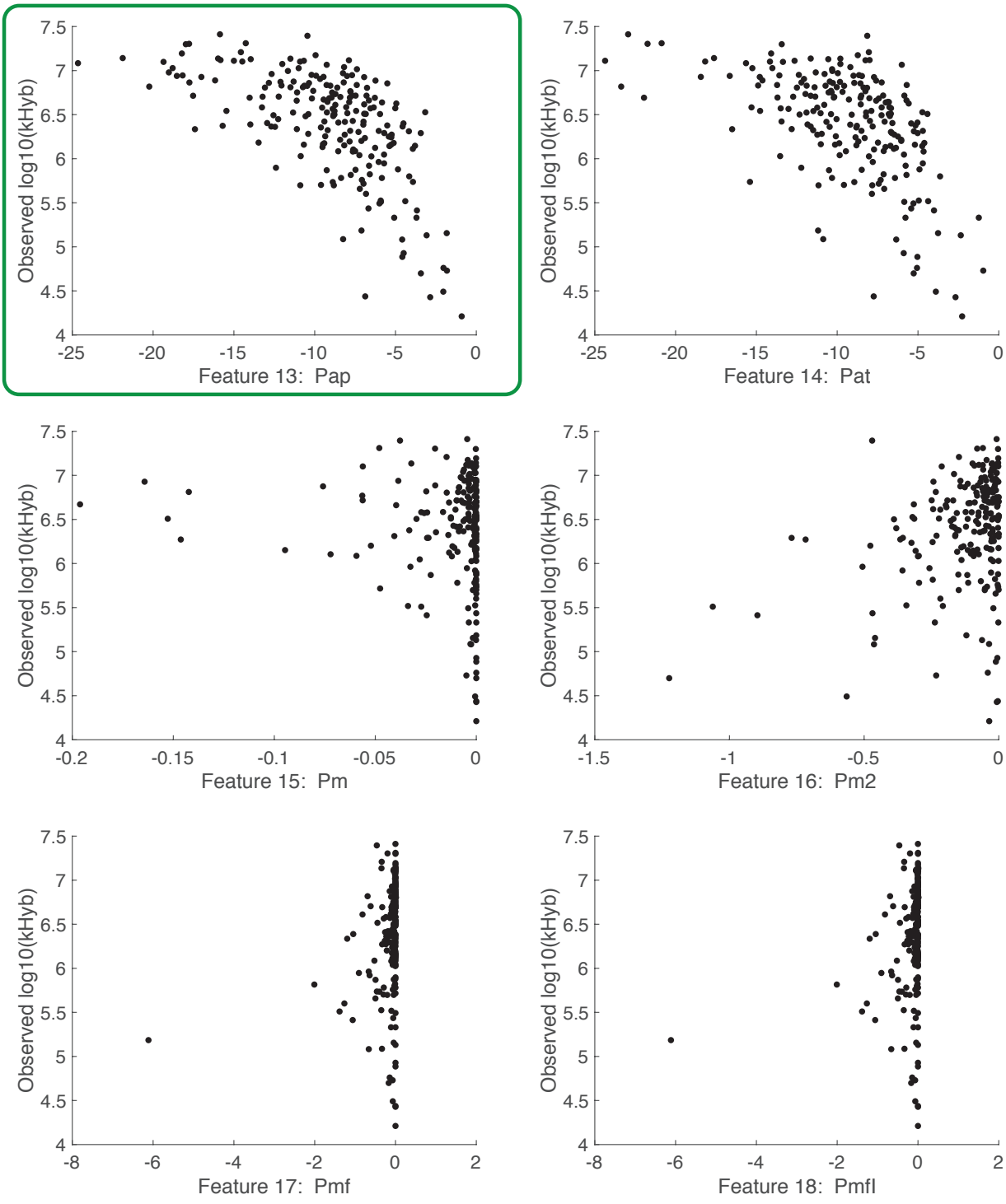

FIG. 4-3: Dependence of k_{Hyb} on initial features 13 through 18.

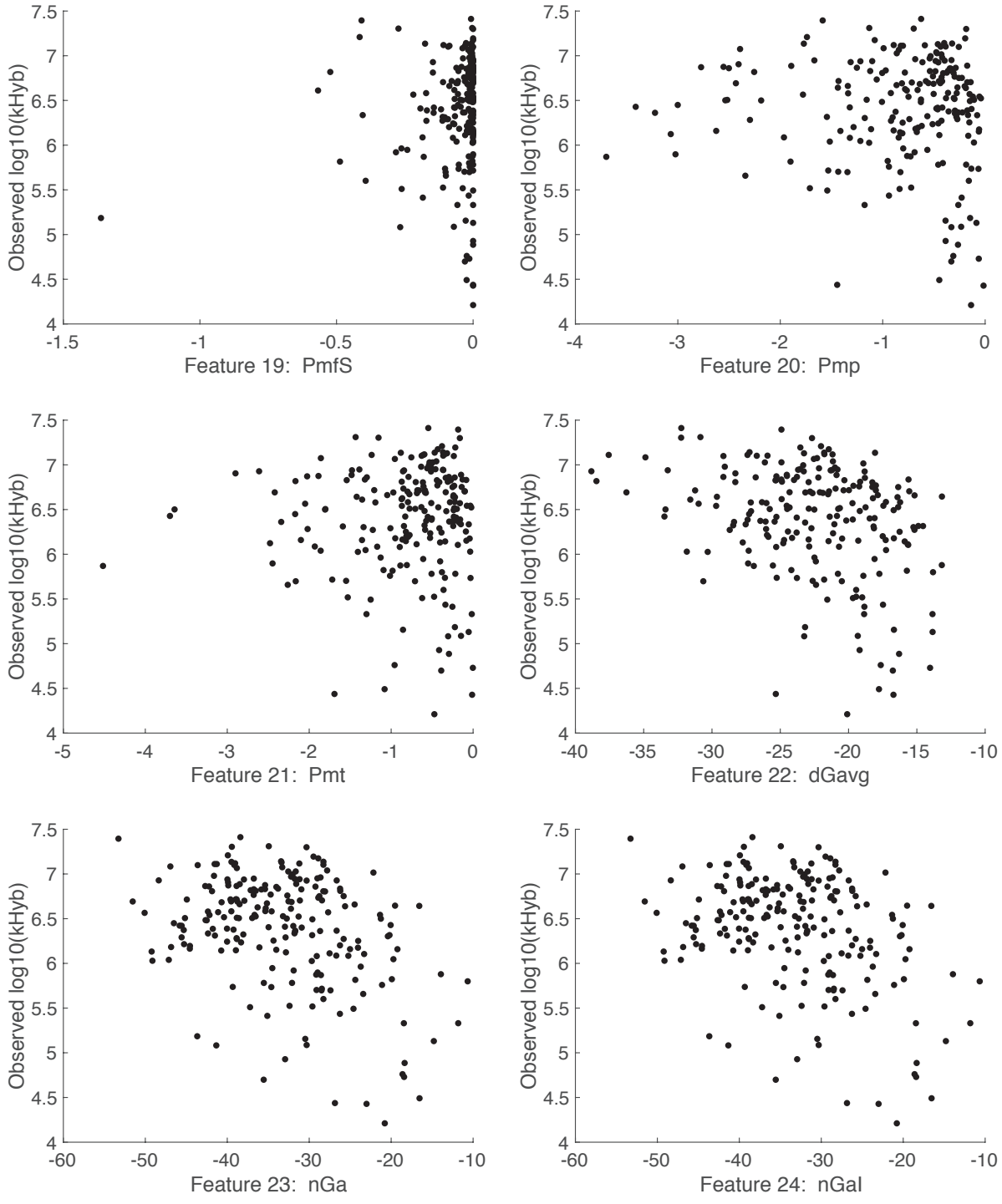

FIG. 4-4: Dependence of k_{Hyb} on initial features 19 through 24.

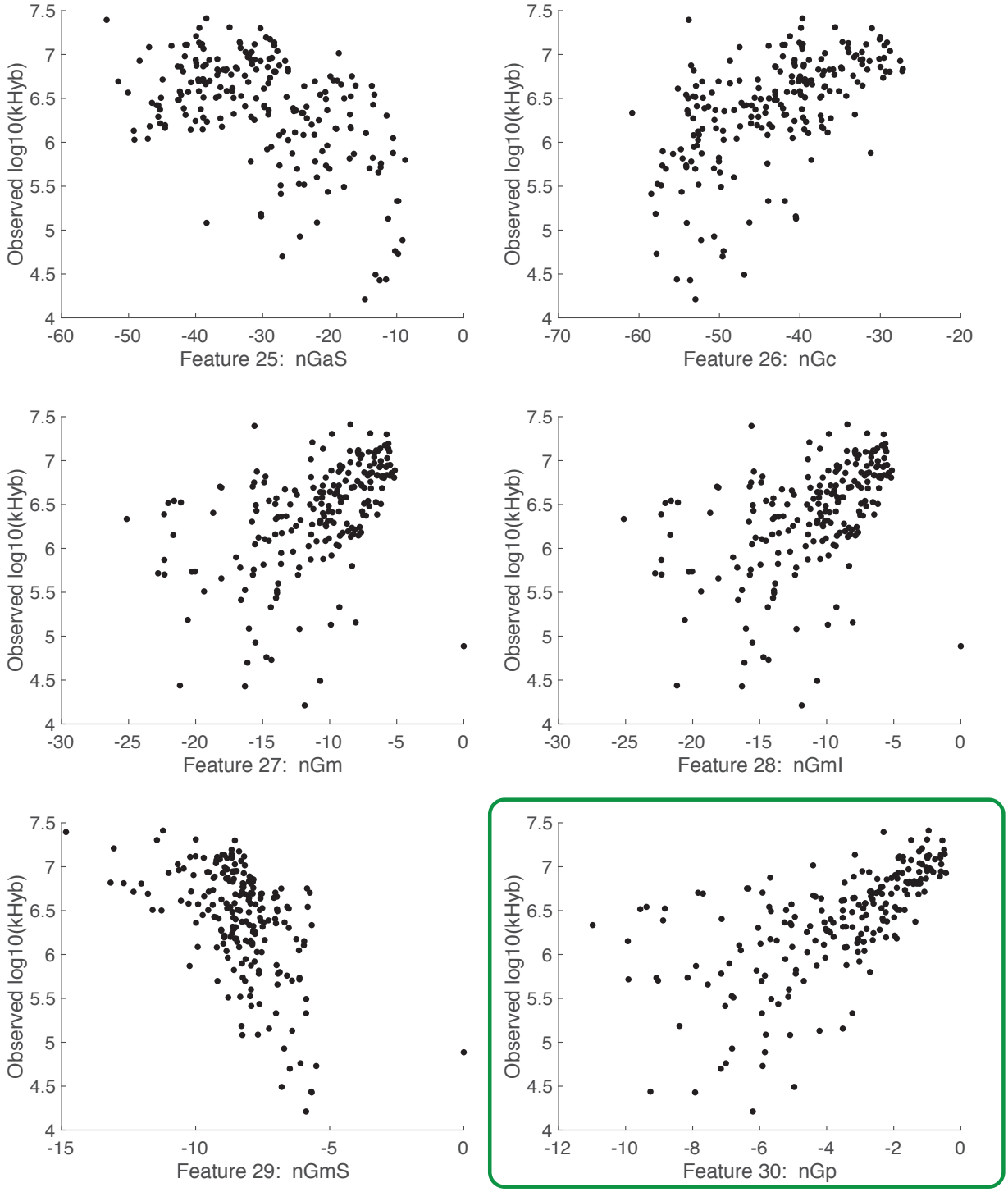

FIG. 4-5: Dependence of k_{Hyb} on initial features 25 through 30.

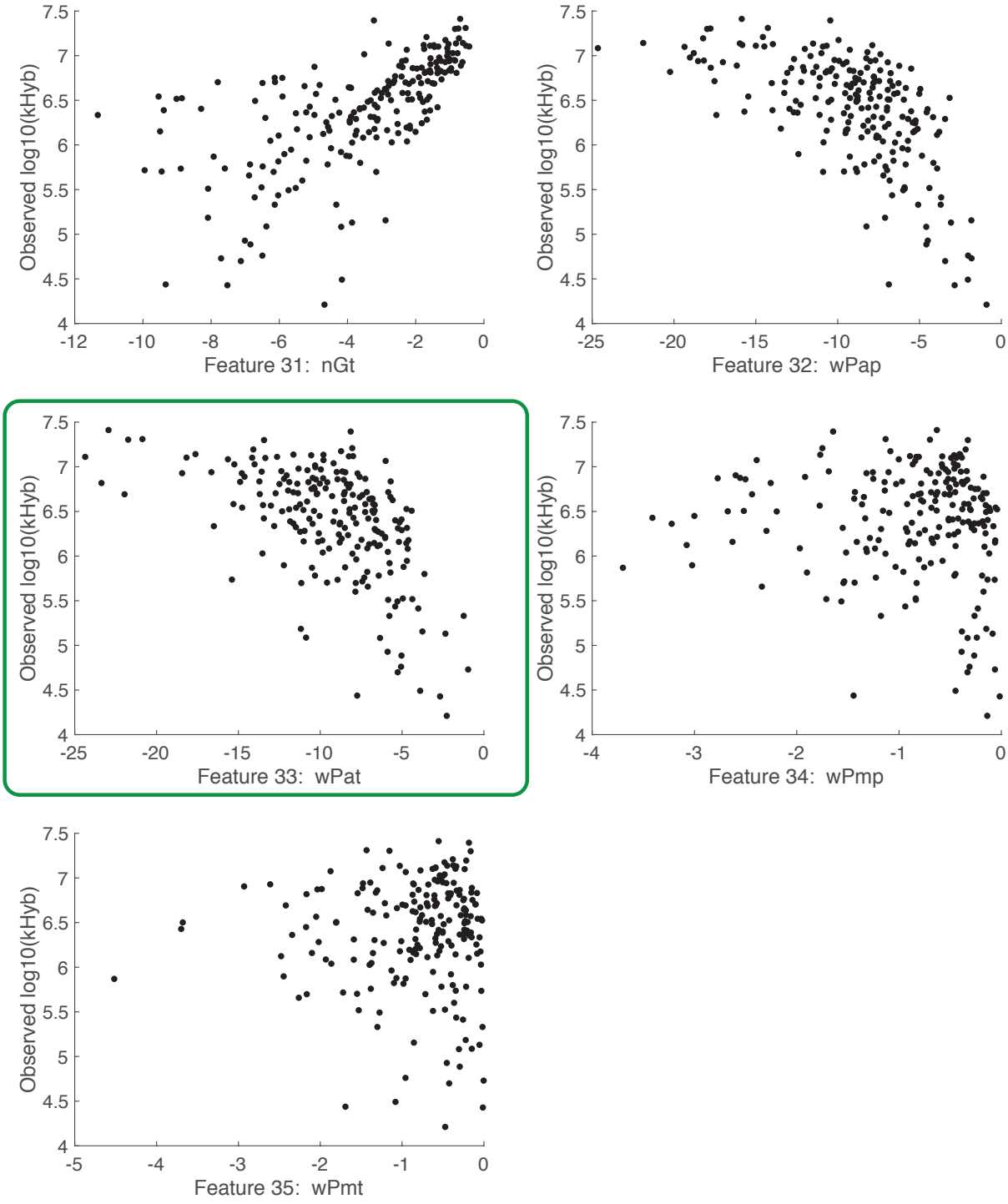

FIG. 4-6: Dependence of k_{Hyb} on initial features 31 through 35.

5. NGS Studies

We performed 2 different next-generation sequencing (NGS) studies for this work. The first study uses synthetic DNA targets bearing Illumina sequencing adaptors, in order to characterize the degree of correlation between single-plex hybridization kinetics and multiplex hybridization kinetics. The second study applies hybridization kinetics to genomic DNA (gDNA) enrichment.

Correlation Between Single-Plex and Multiplex Hybridization Kinetics. Hybridization kinetics in these multiplex enrichment assays will differ from single-plex hybridization kinetics, due to (1) nonspecific interactions between different probes and targets, (2) nonspecific interactions between targets/probes and other genomic DNA, and (3) target 5′ and 3′ dangles that can contribute to secondary structure. For example, targets and probes with minimal secondary structure typically exhibit fast binding kinetics in single-plex setting, but are most prone to nonspecific binding in genomic DNA settings. Building a predictive model for multiplex hybridization kinetics requires massive amounts of additional data due to the complexity of the problem, and is beyond the scope of this study. Instead, we performed a 52-plex hybridization reaction and assayed the kinetics using NGS to provide a high-level survey of the extent to which single-plex and multiplex hybridization kinetics are correlated.

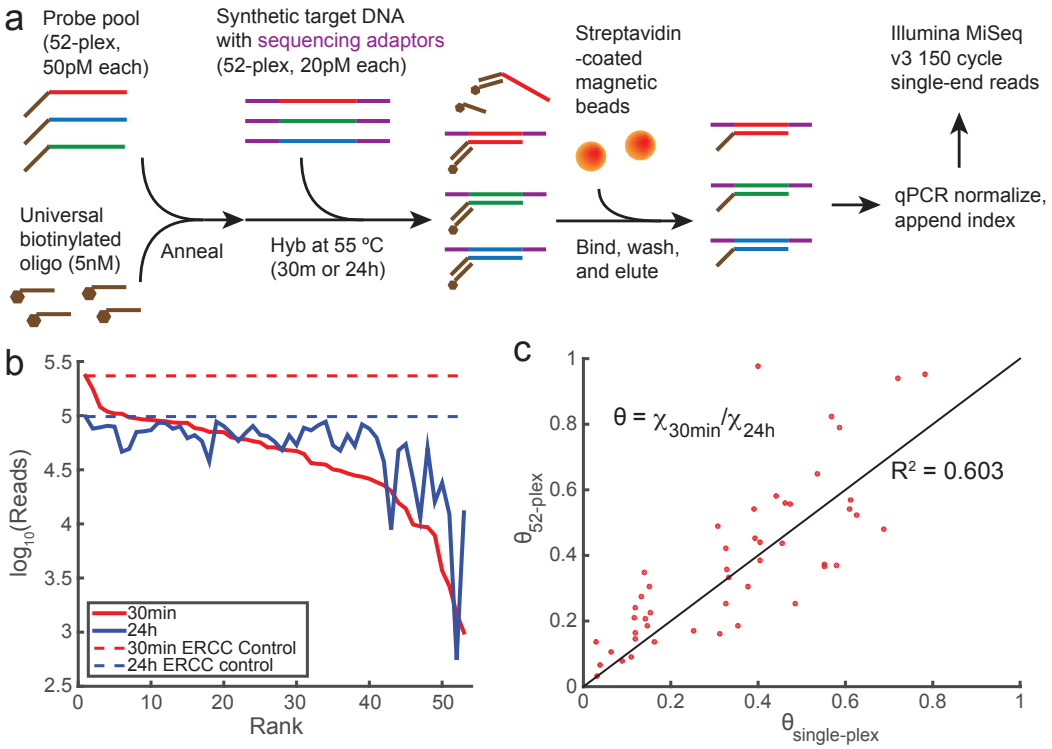

FIG. 5-1: Extensibility of single-plex hybridization rate constants k_{Hyb} to multiplexed hybrid-capture. **(a)** Multiplex hybrid capture and NGS library preparation workflow. A set of 52 non-overlapping target and probe sequences were selected from the 100 from the prior experiments. The targets were ordered as synthetic oligos with Illumina sequencing adaptors present at the 5′ and 3′ ends. Separate libraries were constructed for hybridization durations of 30 min and 24 hr. **(b)** Summary of observed NGS reads for the 52 hybridization reactions. The dotted lines show a positive control species in which the target is pre-annealed to its complement; the positive control target sequence is selected from the External RNA Controls Consortium (ERCC). **(c)** Comparison of multiplex and single-plex hybridization kinetics. θ is calculated as the ratio of hybridization yields χ at 30 minutes and 24 hours, inferred by comparison to the ERCC control. One target/probe pair with very low reads was excluded from this plot.

Fig. 5-1a shows an overview of our multiplex hybridization experiment and NGS library preparation. To minimize the unpredictable effects of stochastic fragmentation and nonspecific binding of genomic DNA, we used a set of synthetic DNA targets with Illumina sequence adaptors appended at the 5′ and 3′ ends. For solid-phase capture by streptavidin-coated magnetic beads, we used a biotin-functionalized universal oligonucleotide in place of the fluorophore-functionalized oligonucleotide for prior studies.

A pre-annealed target and probe oligo pair (with sequence selected from the External RNA Controls Consortium, ERCC) was added into each library to serve as positive control for quantitation of hybridization yield (Fig. 5-1b). Because of the low time resolution of the NGS experiments, it was not possible to infer k_{Hyb} values;

instead, we compare kinetics of single- and multiplex hybridization via the values of $\theta = \frac{\chi_{30m}}{\chi_{24h}}$ (Fig. 5-1c), where χ denotes hybridization yield. For multiplex experiments, we approximate χ as $\chi \approx \frac{\text{Reads(Target)}}{\text{Reads(ERCC)}}$ for each target. For single-plex experiments, χ is calculated based on best-fit model H3 parameters using the oligo concentrations used for multiplex experiments.

There is significant but imperfect correlation ($R^2 = 0.603$) between θ values for single-plex and multiplex hybridization. Interestingly and contrary to our initial expectations, more than half of the targets exhibited higher θ in multiplex settings. We believe there to be four main causes for the observed differences in θ: (1) The $5'$ and $3'$ adaptor sequences may form significant secondary structure with some target sequences, reducing multiplex hybridization kinetics. (2) Some target sequences may transiently bind to other target sequences, reducing the effective concentrations and multiplex hybridization kinetics of both targets. (3) Some target sequences may promiscuously bind to other (not fully complementary) probe sequences, effectively increasing probe concentration and effective multiplex kinetics. (4) Some target sequences may nonspecifically bind directly to the streptavidin-coated magnetic beads, increasing capture yield and imply higher perceived multiplex kinetics as an artifact of the solid-phase capture process. Causes (1), (2), and (3) may be captured in an expanded model built for multiplex hybridization kinetics prediction, and cause (4) may be mitigated through protocol optimization.

Two separate NGS libraries were created for multiplexed hybrid-capture experiments: one for 30 minute hybridization and one for 24 hour hybridization (see Methods). These two libraries were pooled with other libraries for other experiments in a single Miseq v3 150-cycle chip. Overall, the sequencing quality was very high, with over 99% of reads in both libraries mapped to Target or ERCC control sequence via Bowtie2 (Fig. 5-1). Furthermore, the Q scores (from FASTQ file produced by the NGS run) for each library are plotted against the read position. For the first 36 nucleotides of each read that correspond to the target sequence, mean Q scores were significantly above 30.

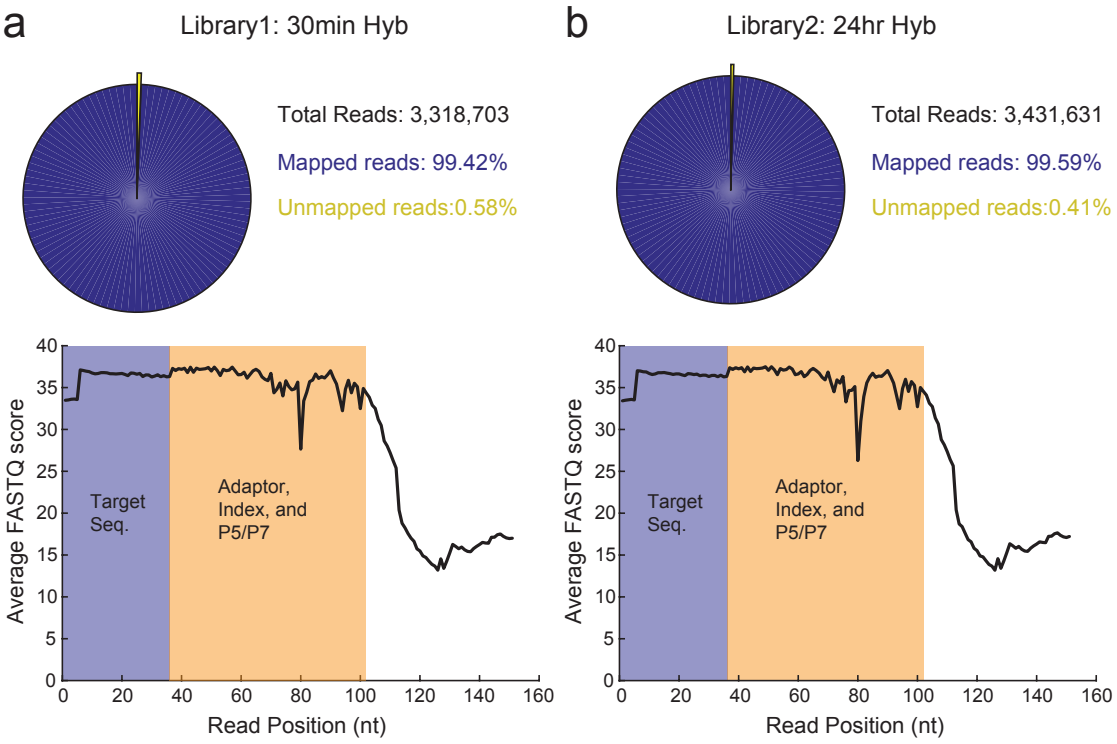

FIG. 5-2: Quality control data for NGS experiments on the **(a)** 30 minute hybridization and **(b)** 24 hour hybridization libraries. Top panels show that more than 99% of all reads were mapped to one of the 52 targets or the ERCC control via Bowtie2. Bottom panels show the mean FASTQ scores indicating quality of sequencing varying by position, averaged over all mapped reads.

In Fig. 5-1c, $\theta = \frac{\chi_{30\text{min}}}{\chi_{24\text{hr}}}$ values for multiplex hybridization were compared against single-plex hybridization. The multiplex θ value was computed based on the reads corresponding to reads that perfectly matched the target sequence. Fig. 5-2a shows the fraction of mapped reads that were perfectly matched to a target sequence: other than 1 outlier with significant sequencing error resulting in only ≈50% of mapped reads perfectly matching the target sequence, all other targets consistently showed roughly 90% perfect match. Fig. 5-2bc shows that the decision to use perfectly-matched reads (rather than all mapped reads) does not have a significant impact on the computed value of θ.

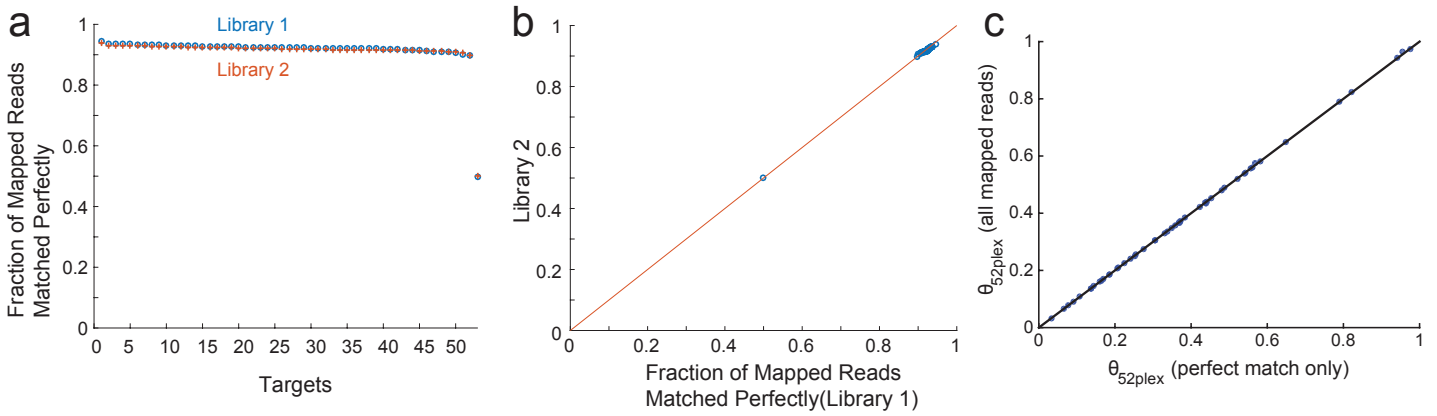

FIG. 5-3: Fraction of perfectly matched reads and influence on inferred θ values. **(a)** Distribution of the fraction of mapped reads that are perfectly matched to target sequence, ranked from high to low. One outlier exhibited high (≈50%) sequencing error; all other targets were roughly consistent at around 90% perfect sequence match fraction. **(b)** Correlation between fraction of read mapped perfectly for different targets within Library 1 and Library 2. There is a near-perfect match, indicating that the sequencing errors are independent of exact protocol, and likely intrinsic to the sequences. **(c)** Correlation between θ values for multiplexed hybridization based on total mapped reads vs. perfectly matched reads. The near-perfect match suggests that NGS errors do not significantly affect inferred θ values.

Hybrid-capture Enrichment From Human Genomic DNA. The primary results from the second NGS study are summarized in the manuscript Fig. 6. Fig. 5-4 shows the distribution of predicted hybridization rate constants for probes to different 36 nt regions of the AQP1 gene; rate constants range roughly 2 orders of magnitude. The other 20 genes show similar distributions of predicted rate constants. True rate constants for probes to different regions may exhibit a broader distribution of rate constants, due to the effects of nonspecific interactions of genomic DNA and the effects of significant overhangs on genomic DNA targets.

Genomic DNA was sheared using a Covaris sonicator to roughly 120 nt prior to end-repair and adaptor ligation. Fig. 5-5 shows the corresponding Agilent Bioanalyzer trace for the fragmented DNA. Fig. 5-7 shows the fragment lengths as determined by sequence alignment from NGS results.

The human genomic DNA hybrid-capture enrichment experiment used a total of 130 probes spanning 21 genes. Fig. 5-6 shows the reads mapped to each probe, sorted in descending order by the 24 hour library, based on the alignment workflow shown in Fig. 5-8. For 30 of the probes corresponding to 4 genes, the NGS reads were under 100 for both the 20 minute and the 24 hour libraries. This low read depth means that Poisson noise has a large affect on observed reads, so that In contrast, the median sequencing depth of the 130 probes was over 1000x.

Such sequencing bias has been well reported in literature, and could potentially be mitigated through empirical optimization (e.g. increased concentrations of probes to low-depth regions). However, as the current 21-gene panel was selected arbitrarily and does not confer important clinical information, we felt that optimizing this particular panel for sequencing uniformity is beyond the scope of the current work. As the 30 low-depth regions across the 4 genes had an equal distribution of median and fast probes, omitting these does not bias our conclusions regarding predictions of hybridization probe kinetics.

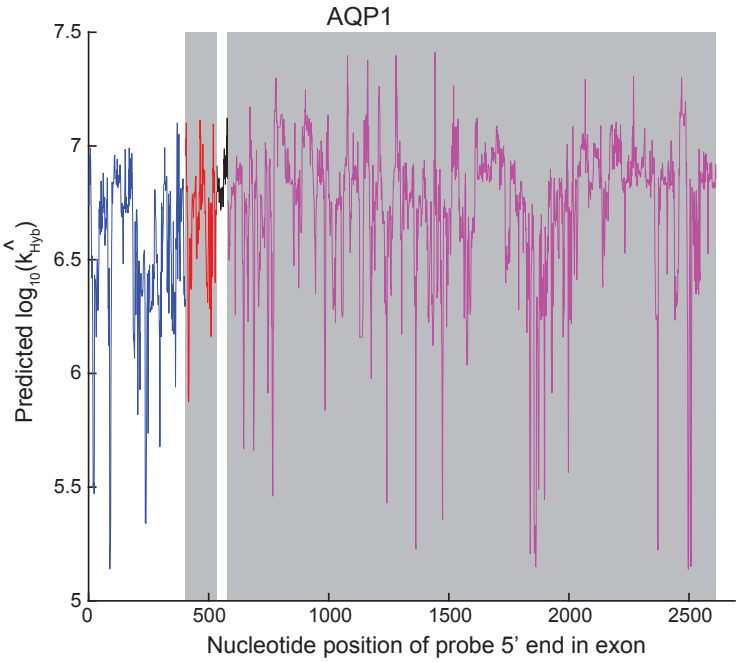

FIG. 5-4: Predicted hybridization rate constants for different 36 nt probes to exon regions of the AQP1 gene, using the final model described in manuscript Fig. 5. The different exons are separated by alternating white and gray backgrounds.

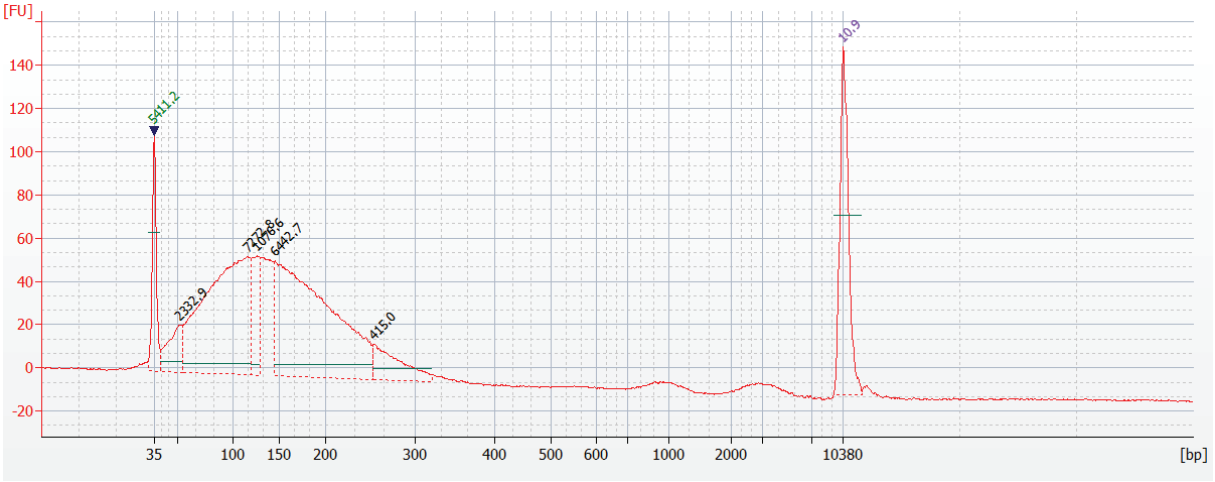

FIG. 5-5: Agilent Bioanalyzer analysis of Covaris-sheared genomic DNA fragment length. The sharp peaks at 35 and 10380 are markers, and the sheared DNA correspond to the broad peak centered at 120 base pairs.

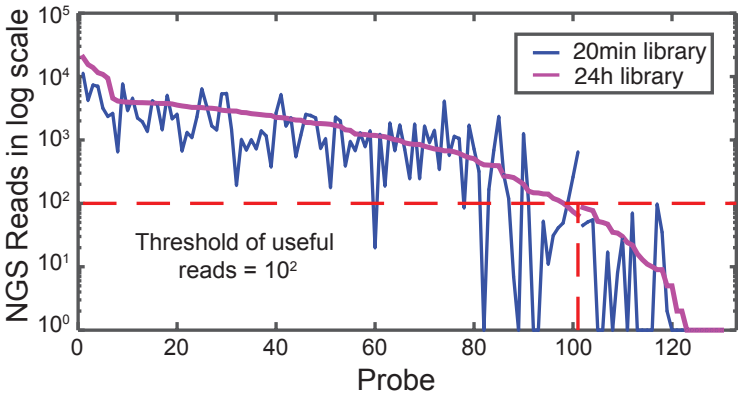

FIG. 5-6: NGS read distribution for gDNA enrichment studies. 30 out of 130 genetic loci resulted in very low NGS read depth (lower right boxed region), even for the 24 hour library; these 30 probes all target a subset of 4 out of the 21 genes we initially selected. Stochasticity means that inferring kinetic properties for these corresponding probes have unacceptable large error; data for these probes were consequently omitted from analysis. There is no observable statistical difference between median and fast probes for the 30 low-read probes.

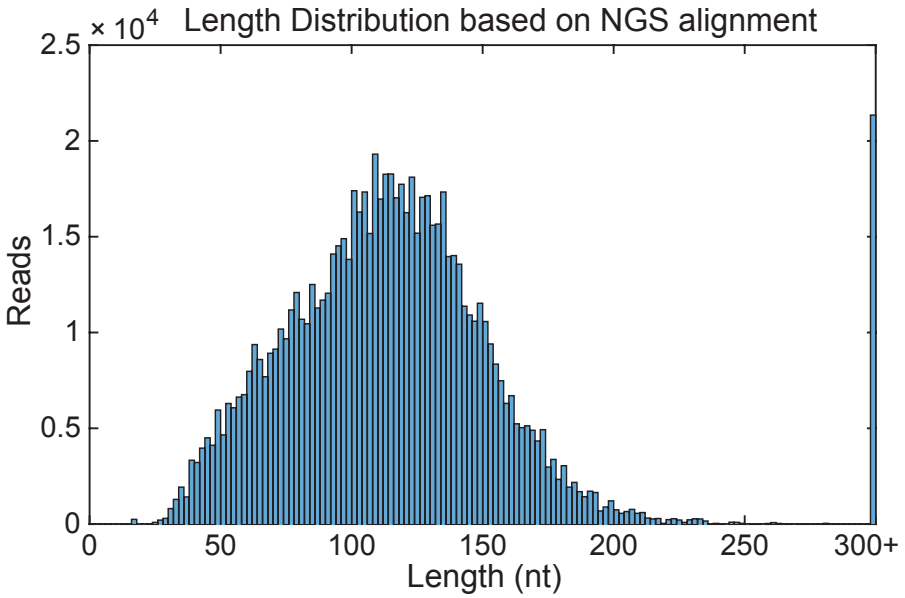

FIG. 5-7: Distribution of DNA fragment lengths based on NGS sequencing reads and alignment to probe sequences.

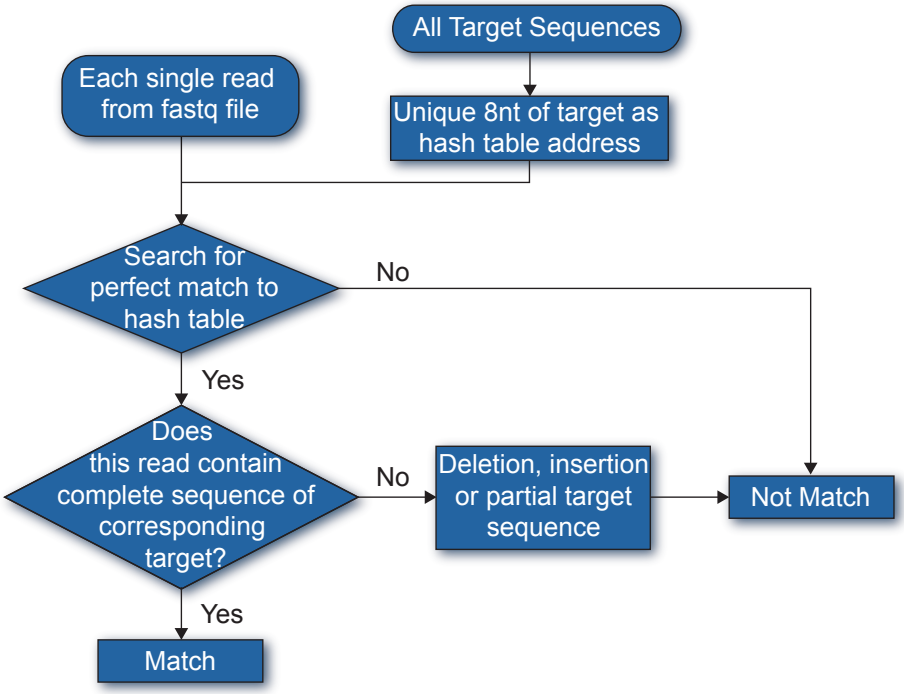

FIG. 5-8: Workflow for aligning NGS reads to probes.

6. Excel Spreadsheet Descriptions

Features.xlsx Spreadsheet.

This xlsx document contains a single sheet, titled Feature_Values. The first row on the sheet contains the number of the feature, if the column is a feature column. Each cell in the second row gives a short description of the values listed in the column. Each row after the second corresponds to a fluorescence hybridization experiment. Column A shows the target sequence ID number, and columns C and D are the observed best-fit values for k_{Hyb} and Bad Fraction. All other columns indicate the calculated feature values for the hybridization reactions.

Sequences_and_Concentrations.xlsx Spreadsheet.

The first sheet, titled "Target subsequences," shows the 36 nt target sequences and their corresponding identification (ID) numbers. These sequences and the reaction temperatures are the only inputs in our current WNV model and feature calculations.

The second sheet, titled "Seq. for Fluor. Expt.," shows the actual DNA sequences for P (Probe), T (Target), F (Fluorophore), and Q (Quencher) that were ordered commercially as synthetic oligonucleotides. The number in each sequence name corresponds to the sequence ID on the first sheet.

The third sheet, titled "Conc. for Fluor. Expt.," shows the final concentrations used for fluorescence experiments. Columns A through D show conditions for the experiments summarized in Fig. 3 of the main text; columns G through J show conditions for the experiments used in Supplementary Section 1.

The fourth sheet, titled "Seq. for gDNA NGS," shows the sequences of oligonucleotides used for NGS experiments on human genomic DNA, summarized in Fig. 6 of the main text.

The fifth sheet, titled "Seq. for Multiplex Kinetics NGS," shows the sequences of oligonucleotides used for NGS experiments summarized in Supplementary Section 5.