High-Throughput DNA melt measurements enable improved models of DNA folding thermodynamics

Yuxi Ke[1], Eesha Sharma[2,3], Hannah K. Wayment-Steele[4], Winston R. Becker[5], Anthony Ho[2], Emil Marklund[2,6], and William J. Greenleaf[2,7*]

[1]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA

[2]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

[3]Lamar Health, Palo Alto, CA 94305, USA

[4]Department of Chemistry, Stanford University, Stanford, CA 94305, USA

[5]Program in Biophysics, Stanford University, Stanford, CA 94305, USA

[6]Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, 17165, Solna, Sweden

[7]Department of Applied Physics, Stanford University, Stanford, CA 94305, USA

*Correspondence: wjg@stanford.edu (W.J.G.)

Abstract

DNA folding thermodynamics are central to many biological processes and biotechnological applications involving base-pairing. Current methods for predicting stability from DNA sequence use nearest-neighbor models that struggle to accurately capture the diverse sequence-dependency of elements other than Watson-Crick base pairs, likely due to insufficient experimental data. We introduce a massively parallel method, Array Melt, that uses fluorescence-based quenching signals to measure equilibrium stability of millions of DNA hairpins simultaneously on a repurposed Illumina sequencing flow cell. By leveraging this dataset of 27,732 sequences with two-state melting behavior, we derived a refined NUPACK-compatible nearest-neighbor model, a richer parameterization nearest-neighbor model that exhibits higher accuracy, and a graph neural network (GNN) model that identifies relevant interactions within DNA beyond nearest neighbors. All models provide improved accuracy in predicting DNA folding thermodynamics, providing improvements relevant for *in silico* design of qPCR primers, oligo hybridization probes, and DNA origami.

Introduction

The thermodynamics of DNA secondary structure formation is fundamental to understanding diverse biological processes, such as DNA replication[1] and repair[2], and is widely used in biotechnological applications, including the design of PCR primers[3], the engineering of in-situ hybridization (ISH) probes[4], the design of guide sequences for genome editing tools[5], and for DNA nanotechnology[6,7]. Numerous algorithms have been developed to predict DNA secondary structure thermodynamics, many of which rely on nearest-neighbor models[8–10].

Briefly, nearest-neighbor models assume that the total folding energy of DNA can be calculated by summing up the energies of each pair of neighboring base pairs. With this assumption, the "folded state" can be defined as the minimum

free energy (MFE) base pairing configuration, identified using dynamic programming[11]. Additionally, by using partition functions, it is possible to calculate probabilities of the DNA in alternative secondary structures based on these computed free energies. However, nearest-neighbor models often struggle to accurately capture the sheer diversity and complexity of DNA secondary structure motifs, including hairpin loops, mismatches, and bulges. One reason for this limitation is the lack of experimental data spanning those loop features upon which these models are built, leading to inaccuracies in prediction. For example, the most widely used parameter set from *SantaLucia et al. 2004* used data from 108 sequences to derive 12 parameters for Watson-Crick (WC) base pairs, and 174 sequences to derive 44 parameters for internal single mismatches[8].

This data bottleneck has been due to the laborious nature of optical melting[12] and differential scanning calorimetry[13] experiments, traditionally considered the gold standards of DNA secondary structure dynamics. Several groups have attempted to overcome this bottleneck, usually with fluorescent melting in bulk solutions in well plates[14–17], reaching a throughput on the scale of thousands. Still, due to the extremely large combinatorial DNA sequence space, these works are limited to a specific type of DNA structural element in a fixed sequence scaffold.

To address this data generation bottleneck, we developed a method for systematic, accurate, high-throughput measurements of nucleic acid secondary-structure motif thermal stability, enabling large-scale, quantitative measurements of the thermodynamics of nucleic acid secondary structure. We demonstrate that nearest-neighbor thermodynamic parameters inferred from these data are nearly identical to those observed from bulk thermal melting experiments for WC base pairs but are improved for loop sequence motifs. The improved nearest-neighbor model parameter sets are capable of accurately predicting DNA duplex folding energy on orthogonal unseen data sets. Furthermore, by deploying advanced computational methods including deep learning to model DNA thermodynamics, we develop a state-of-the-art thermodynamic model for DNA hairpin folding with accuracy comparable to measurement uncertainties.

Results

Array melt: A high-throughput technique for quantifying the melting behavior of nucleic acids.

We developed the Array Melt technique, a method that allows for the rapid and comprehensive quantification of DNA folding thermodynamics in high throughput. This technique is based on an Illumina MiSeq chip repurposed for high-throughput measurements[18,19]. To assay the melting behavior of a nucleic acid construct, we engineered a common region for annealing a 3'-fluorophore-labeled oligonucleotide to the 5'-end of a hairpin to be investigated. Similarly, we designed another region for annealing a 5'-quencher-labeled oligonucleotide to the 3'-end of the hairpin (**Fig. 1a**). After determining the position and sequence of the variant on the MiSeq chip through sequencing, we annealed Cy3-labeled and Black Hole Quencher (BHQ)-labeled oligonucleotides to these respective binding sites. These annealed oligos have predicted melting temperatures of 74°C, much higher than the highest temperature in our experiments (**Table S1**). At room temperature, this fluorophore-quencher pair results in low levels of fluorescence in each cluster, as the bulk of our hairpin library is folded at room temperature. As the sequence-variable hairpin regions are exposed

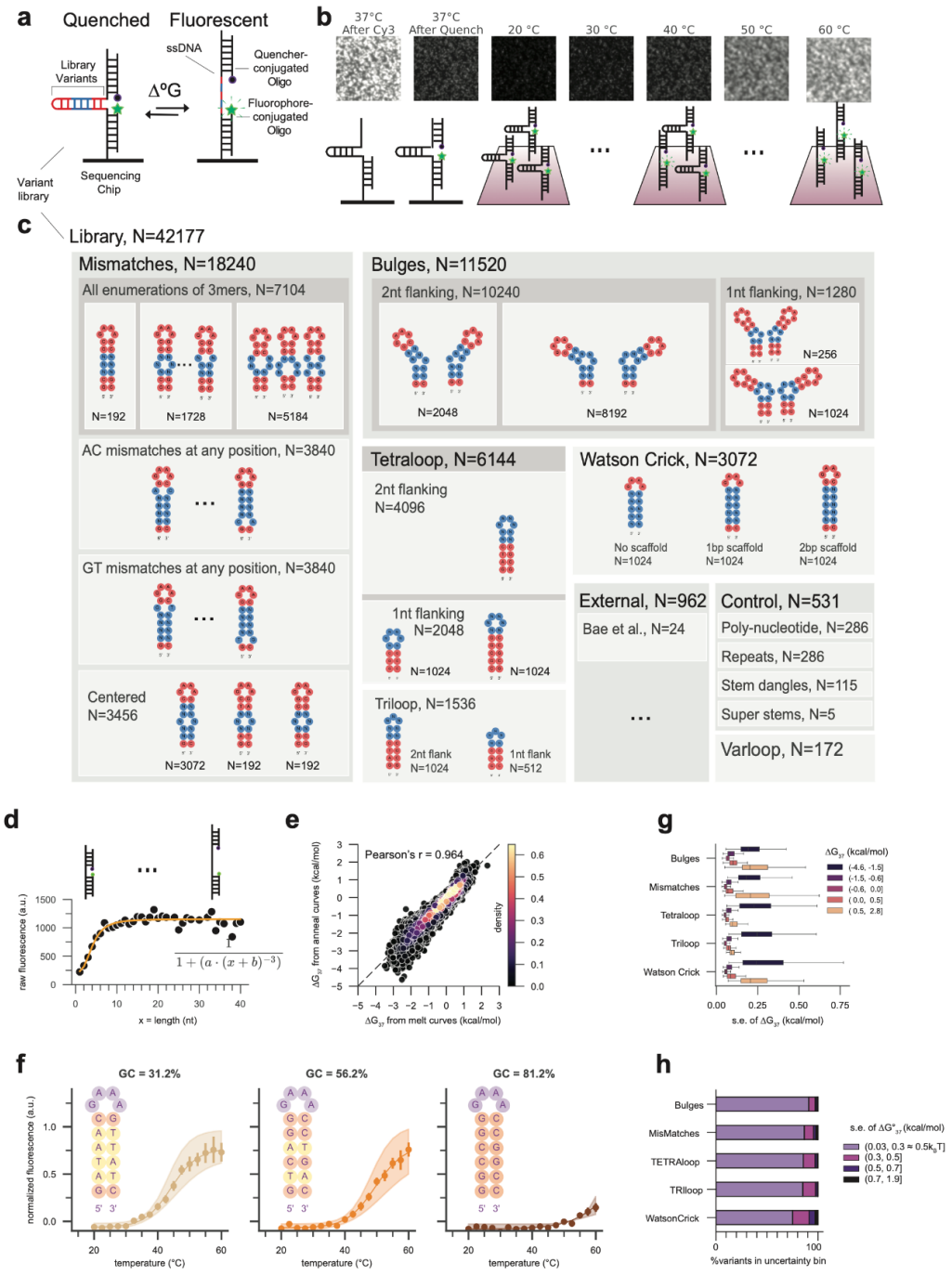
to increasing temperatures between 20 °C to 60 °C, the distance between Cy3 and BHQ increases, leading to brighter fluorescence signals (**Fig. 1a-b**). Using this measurement scheme, we synthesized a comprehensive library of 42,177 unique hairpin sequences for measurement (**Fig. 1c**). These sequences integrated diverse structural elements into a hairpin scaffold, including Watson-Crick pairing (WC), mismatches, bulges, hairpin loops of various lengths, and other control sequences, including super-stable stem, polynucleotide repeats, and single-strand fluorescence controls (see **Fig. 1c** and Methods).

We performed thorough quality control on the dataset by first requiring variants to accurately fit a two-state model with melting behavior observable within our temperature measurement range (see Methods for our filtering strategy). Total variants that pass these stringent thresholds include 6,393,050 individual melt curves (including technical replicates) spanning 27,732 sequence variants for subsequent analysis with a standard two-state melt model.

Sensitivity and precision of the Array Melt method

To calibrate the amount of fluorescence observed as a function of the distance between Cy3 and BHQ, we evaluated the fluorescence signals of a series of repeat sequence variants. These variants were designed to systematically increase the distance between the fluorophore and quencher by incorporating increasing numbers of mono-, di- or tri-nucleotide repeats. The fluorescent signals, aggregated over repeats of the same length, showed a nearly linear response to distance variations up to approximately 8 nucleotides (**Fig. 1d**). Our observed length-fluorescence curve closely aligns with a theoretical static quenching curve (Methods). We normalized the fluorescent signals using unfolded and folded controls designed within the library (**Fig. S1b**). Three example melt curves in **Fig. 1f** illustrated increasing observed melting temperatures and increasing GC content of the variable region. Technical replicate measurements from Array Melt data showed high correlation, with $R > 0.94$ (**Fig. S1c**). Melt curves were also highly correlated with anneal curves ($R = 0.964$), confirming that the DNA molecules were measured at equilibrium (**Fig. 1e**). Upon analyzing our measurement precision (**Methods**), we found that nearly 90% of the tested variants displayed ΔG_{37} uncertainties within 0.3 kcal/mol, or 0.5 k_BT (**Fig. 1h)**. Overall, these data demonstrate that the equilibrium, aggregate fluorescence signals obtained from the Array Melt method provide highly accurate melting information and can serve as a "molecular ruler" capable of resolving single nucleotide distance variations.

FIGURE 1


a
Quenched
Fluorescent
Library Variants
ssDNA
Quencher-conjugated Oligo
ΔºG
Fluorophore-conjugated Oligo
Variant library
Sequencing Chip
b
37°C After Cy3
37°C After Quench
20 °C
30 °C
40 °C
50 °C
60 °C
c
Library, N=42177
Mismatches, N=18240
All enumerations of 3mers, N=7104
N=192
N=1728
N=5184
AC mismatches at any position, N=3840
GT mismatches at any position, N=3840
Centered N=3456
N=3072
N=192
N=192
Bulges, N=11520
2nt flanking, N=10240
N=2048
N=8192
1nt flanking, N=1280
N=256
N=1024
Tetraloop, N=6144
2nt flanking N=4096
1nt flanking N=2048
N=1024
N=1024
Triloop, N=1536
2nt flank N=1024
1nt flank N=512
Watson Crick, N=3072
No scaffold N=1024
1bp scaffold N=1024
2bp scaffold N=1024
External, N=962
Bae et al., N=24
Control, N=531
Poly-nucleotide, N=286
Repeats, N=286
Stem dangles, N=115
Super stems, N=5
Varloop, N=172
d
raw fluorescence (a.u.)
x = length (nt)
e
Pearson's r = 0.964
density
ΔG from anneal curves (kcal/mol)
ΔG from melt curves (kcal/mol)
g
Bulges
Mismatches
Tetraloop
Triloop
Watson Crick
ΔG (kcal/mol)
s.e. of ΔG (kcal/mol)
f
GC = 31.2%
GC = 56.2%
GC = 81.2%
normalized fluorescence (a.u.)
temperature (°C)
h
Bulges
MisMatches
TETRAloop
TRIloop
WatsonCrick
s.e. of ΔG° (kcal/mol)
%variants in uncertainty bin

93

Figure 1. High-Throughput thermodynamic measurement of DNA secondary structures using a fluorophore–quencher system.

a. Schematic representation of DNA molecules in a folded (quenched) and an unfolded (fluorescent) state.

b. Images of fluorescent DNA clusters immobilized on the sequencing chip surface. Top: Image with only fluorophore-conjugated oligo (Cy3), and with both fluorophore- and quencher-conjugated oligo at increasing temperatures. Bottom: Schematics of DNA molecules in each image. All images are normalized to super-stable stem and repeat control variants for temperature-dependent effects on fluorescence and quenching during the experiment, as shown in **Fig. S1d**.

c. Design of library variants between the constant-sequence binding sites for fluorophore- and quencher-conjugated oligonucleotides. Red represents scaffold nucleotides constant within each type, and blue represents the variable nucleotides systematically permuted ('N'). Numbers under each class indicate the number of unique sequences in each class.

d. Fluorescence measurement of control constructs where the single-strand distance between the fluorophore and the quencher increases in single nucleotide steps. The orange line shows the theoretical fit.

e. Correlation of ΔG_{37} derived from library variants under increasing temperatures (melt curve, x-axis) and decreasing temperatures (annealing curve, y-axis)

f. Representative examples of melt curves for constructs that vary in GC content.

g. Pearson correlation coefficient (R) of ΔG_{37} for replicates of three melt and one annealing curve experiment.

h. Standard error of ΔG_{37} for various construct classes as a function of ΔG_{37}.

UV melting validation

To further validate our method, we compared the results from the Array Melt method to those obtained using conventional UV melting methods on a cloud lab platform. We selected random representative variants from each category within the Array Melt library and sent the synthesized oligonucleotides to Emerald Cloud Lab (ECL) for analysis. Using the variant information and UV melting experiment parameters as input, scripts written in Symbolic Lab Language (SLL) automatically generated experiments as ECL protocol objects. These protocols were then queued for remote execution in a physical laboratory. Raw melting curve data from ECL were automatically analyzed by custom Python code, and samples that failed to pass quality control were resubmitted to ECL for re-measurement (**Fig. 2a**, **Fig. S2a**, **Methods**). A comparison of the melting profiles of random library variants showed good agreement between the two methods (**Fig. 2c-e,** R=0.85 for ΔG_{37}, or R^2= 0.62 after the linear adjustment described later). The melting temperature was found to be independent of DNA oligo concentration (**Fig. S2b**), suggesting intramolecular folding. Notably, we observed a systematic offset between the UV measurements and our Array Melt results, which could be due to the presence of the hairpins on a flow-cell surface, or the additional sequence elements included to facilitate our quenching-based readout. To account for these potential systematic differences, we applied a single offset to the melting points observed in our dataset and used the corrected results for subsequent analysis (**Methods**; on average, a correction of 9.35°C in Tm or approximately 0.7 kcal/mol in ΔG_{37} was applied). Compared to standard nearest-neighbor model predictions from NUPACK software, the measured values showed moderate improvement for ΔG and Tm, and considerable improvement for ΔH (**Fig. 2e, Fig. S2c**). This UV melting validation demonstrated the ability of the Array Melt method to accurately measure the folding energies of DNA and generalize to in-solution systems.

FIGURE 2

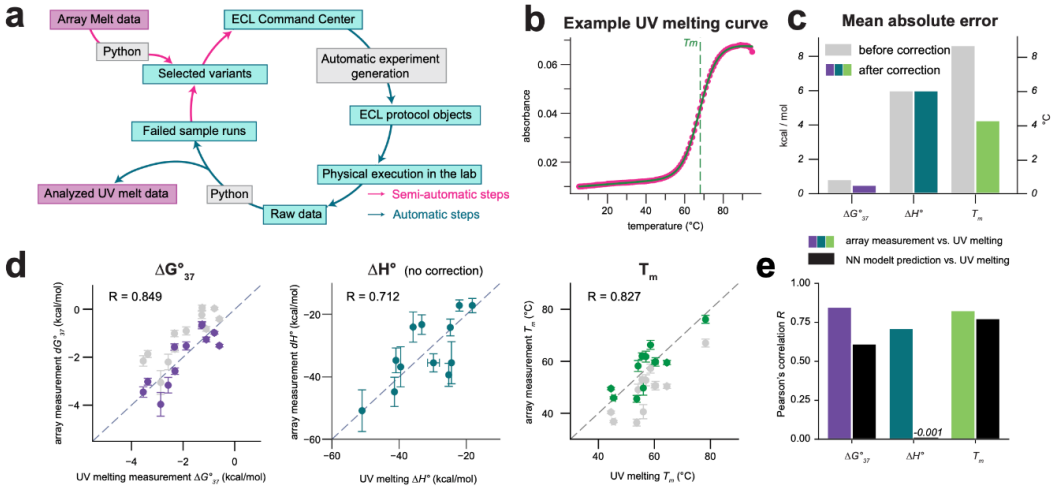


Figure 2. Cloud lab UV melting validation of the method

a. Schematic representation of UV melting data collection and analysis using the Emerald Cloud Lab (ECL) platform. Magenta arrows indicate semi-automatic steps, and teal arrows indicate automatic steps.
b. Example UV melting curve of the variant WC2318. Pink dots represent measured data points of the cooling curve; the green solid line represents the fitted line; the green dotted line represents the fitted melting temperature.
c. Mean absolute error (MAE) of ΔG, ΔH and Tm between Array Melt data and UV melting data before and after systematic error correction. Gray bars represent errors prior to correction; colored bars represent errors after applying correction with a single offset parameter.
d. Direct comparison of measured ΔG, ΔH and Tm between Array Melt and UV melting data. Each datapoint represents one hairpin variant. The gray dots represent measurements prior to correction; the colored dots represent measurements after applying the correction.
e. Pearson's correlation coefficient (R) between UV melting data and either Array Melt data or nearest-neighbor model predictions. The colored bars represent the correlation between Array Melt and UV melting data; the black bars represent that between nearest-neighbor model predictions and UV melting data.

Quantifying nearest-neighbor model performance on non-WC hairpins

After validation of our experimental system, we aimed to examine the performance of the standard nearest-neighbor model on hairpins whose stems are not perfect Watson-Crick pairs. When predicting folding free energies associated with different single mismatch types (**Fig. 3a)**, the nearest-neighbor model achieved an R^2 of 0.61 on the average free energy for each of the types (implying it can explain 61% of the observed experimental variance), or $R^2 = 0.31$ on all individual variants. However, this model was unable to consistently predict significant variances in folding energies that arise from the same mismatch type with differing flanking sequences ($R^2 = -0.61$ for T>C mismatches, as shown on the right side panel, $R^2 = 0.09$ for A>G, $R^2 = -0.48$ for C>T, and $R^2 = -1.66$ for G>A) (**Fig. 3b, Fig. S3a**). Indeed, for three out of the four mismatches we analyzed, the nearest-neighbor model introduced extra variance (i.e. negative R^2 value) compared to the class-average prediction, implying that the variance in the nearest-neighbor model's prediction residual was greater than the actual variance present in the data (**Fig. 3c, Fig. S3b**). While the nearest-neighbor model succeeded in capturing variance for some variant groups with the same mismatch type and nearest-neighbor base pairs, it is unable to capture others (**Fig. S3c**). This limitation of the nearest-neighbor model led us to conclude that there are significant variances in folding energy that are not accounted for by the nearest-neighbor model.

Influence of sequences on energy across different secondary structures

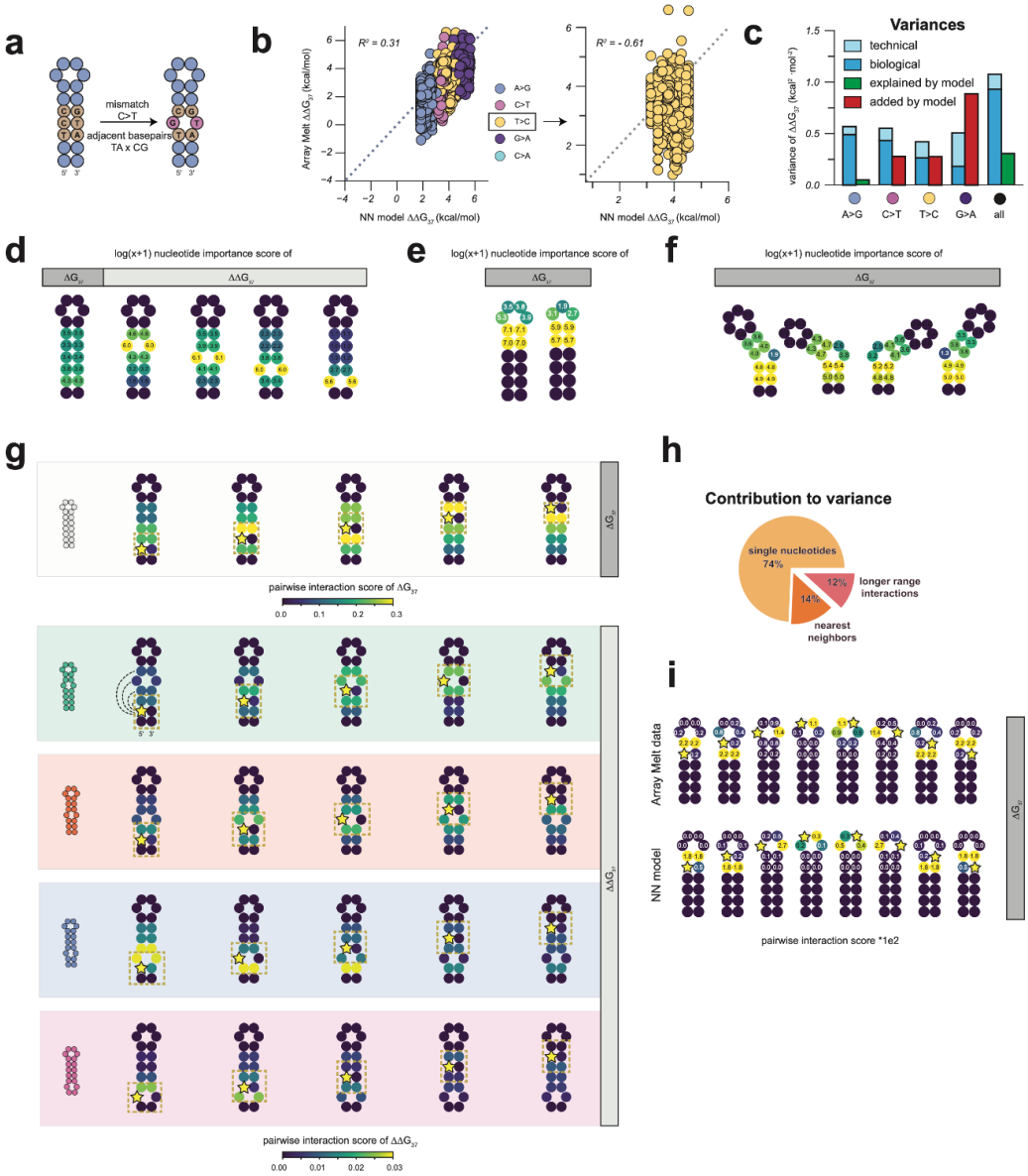
We hypothesized that some of the observed variance in melting behavior not explained by the nearest-neighbor model may be due to the influence of sequences adjacent to mismatches. To assess this possibility, we calculated the impact of each nucleotide position on folding energy. We first determined a nucleotide importance score at each nucleotide position in different sequence groups with the same target secondary structures by calculating the variance in either ΔG_{37} or $\Delta\Delta G_{37}$ values for the four different identities of the nucleotide at the target position (see **Methods**). While it is possible that variants might adopt alternative secondary structures yet maintain two-state behavior, our analysis still provides valuable insights into how sequences affect folding energy. For single mismatches, we used $\Delta\Delta G_{37}$ values

173 relative to Watson-Crick pairing "parent" variants to regress out the energy expected from Watson-Crick stacks; for
174 other secondary structures, we directly used ΔG_{37} values. As anticipated, nucleotides along the hairpin stem displayed
175 consistent nucleotide importance scores for Watson-Crick pairing variants, and the two mismatching nucleotides have
176 the highest contribution to this folding energy for single mismatch variants, both in the Array Melt data and the nearest-
177 neighbor model (**Fig. 3d, Fig. S3d**). Extending our analysis to hairpin loops and bulges, we observed that specific
178 sequences within these structures critically influence their thermodynamic energy (**Fig. 3e-f**). In triloops and
179 tetraloops, loop center nucleotides had a more pronounced influence on folding energy in the Array Melt data than
180 nearest-neighbor model predictions. We also noted a slight asymmetry between the 5' and 3' ends of the loop,
181 regardless of loop size (**Fig. 3e, Fig. S3e**). In bulges, flanking Watson-Crick pairs had greater influence than the bulge
182 sequence itself, with double bulges showing more sequence dependence than singles (**Fig. 3f, Fig. S3f**).

183 These observations prompted us to examine the energetic interactions between bases by calculating the additional
184 proportion of variance explained in a linear model that considered base interactions (i.e. allowing free parameters for
185 pairwise combinations of bases in the two probed positions) compared to one that treats each position independently
186 (i.e. only allowing free parameters for each individual position). Notably, nucleotides opposite the mismatch showed
187 interactions unaccounted for in the nearest-neighbor model (**Fig. 3g, Fig. S3g**). This interaction score decays with
188 distance along the stem (**Fig. S3h**). Our quantification showed that individual nucleotides, nearest neighbors, and
189 longer-range interactions contribute 74%, 14%, and 12% respectively in this specific case of single mismatches (**Fig.
190 3h**). In hairpin loops, the interactions between the loop ends and the loop center nucleotides were stronger than
191 expected in tetraloops, but not in triloops (**Fig. 3i, Fig. S3i-j**).

192 Subsequently, we fit linear regression models to Watson-Crick and single-mismatch variant Array Melt data, varying
193 the length of stack features from one (base pairs only), to two (nearest neighbors) and three (next nearest neighbors).
194 Although prediction errors were similar for nearest- and next-nearest-neighbor models in Watson-Crick variants,next-
195 nearest-neighbor models outperformed the nearest-neighbor ones in single-mismatch variants (**Fig.4a**).

FIGURE 3


a
mismatch
C>T
adjacent basepairs
TA x CG
b
R² = 0.31
Array Melt ΔΔG37 (kcal/mol)
NN model ΔΔG37 (kcal/mol)
A>G
C>T
T>C
G>A
C>A
R² = - 0.61
c
Variances
technical
biological
explained by model
added by model
variance of ΔΔG37 (kcal² mol⁻²)
A>G
C>T
T>C
G>A
all
d
log(x+1) nucleotide importance score of
ΔG37
ΔΔG37
e
log(x+1) nucleotide importance score of
ΔG37
f
log(x+1) nucleotide importance score of
ΔG37
g
ΔG37
pairwise interaction score of ΔG37
ΔΔG37
pairwise interaction score of ΔΔG37
h
Contribution to variance
single nucleotides
74%
12%
longer range
interactions
14%
nearest
neighbors
i
Array Melt data
NN model
ΔG37
pairwise interaction score *1e2

196

Figure 3. Decomposition of sources of variances in measured energies

a. Schematic representation of mismatch types and ΔΔG calculation for single mismatches.
b. Comparison of single mismatch ΔΔG between nearest-neighbor model prediction and Array Melt measurements. Each variant is colored by the identity of the mismatch (left), and the T>C mismatches are zoomed in on the right.
c. Unexplained expected biological variance for each group of mismatches and combined. Variance explained or introduced by the nearest-neighbor model are shown in green and red, respectively.
d. Nucleotide importance scores for each individual nucleotide in mismatch constructs, grouped by secondary structure. The importance score is calculated as the sum of squares introduced by varying the given nucleotide position.
e. Nucleotide importance scores for tetraloops and triloops.
f. Nucleotide importance scores for bulges.
g. Pairwise interaction scores for each pair of nucleotide locations for Watson-Crick and single mismatch variants. A yellow star indicates the anchor point in each subplot, and the color of each position indicates the strength of the interaction with the anchor. The interaction score is calculated as the R-squared of a linear model given all combinations of the pair as features, subtracted by that of the two individual nucleotides.
h. Percentage contribution of single nucleotides, nearest neighbors, and longer-range interactions to the total variance in $\Delta\Delta G_{37}$ for single mismatches.
i. Pairwise interaction scores for tetraloops and triloops.

Revised parameter set for NUPACK integration

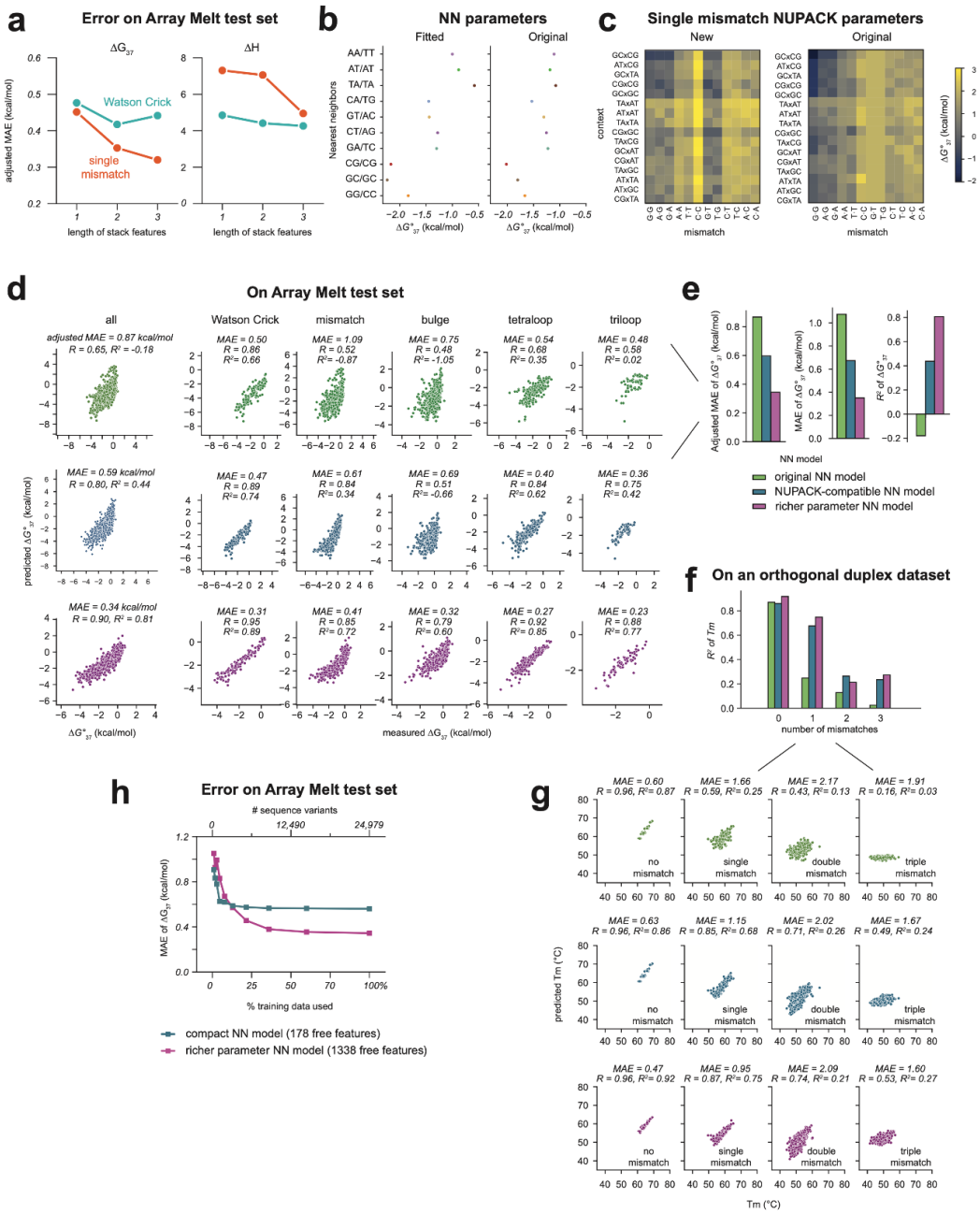
We aimed to develop a refined set of parameters to integrate our findings with the NUPACK software[7,20], thereby enabling more accurate DNA thermodynamic predictions that are easily accessible. We divided Array Melt data into training, validation, and test sets, respectively with 24979, 1315, and 1438 variants, stratified by major variant types as shown in **Fig. 1c**. Initially, we fitted compact linear regression models to the Array Melt data, focusing on structural elements prevalent in our variant library. By allowing the 10 Watson-Crick pairing stack parameters to vary freely (totaling 188 free parameters), we reproduced the established nearest-neighbor parameters set[8] (**Fig. 4b**). Subsequently, we fixed the values of the 10 Watson-Crick stack parameters and fitted a compact 178-parameter nearest-neighbor model. We compared the single mismatch parameters of our new parameter set with the original *dna04* set and observed similar overall trends. Our new parameters also captured context-dependent variations of GT mismatches and highlighted the instability of CC mismatches, which are shown to be the most thermodynamically unstable DNA single mismatch[21] (**Fig. 4c, Fig. S4c**). Next, we obtained parameters for tetraloops and triloops by aggregating thermodynamic measurements of relevant Array Melt variants and converting the aggregated ΔΔH and $\Delta\Delta G_{37}$ values to ΔH and ΔG_{37} parameters (**Methods**). We combined these corrected hairpin loop parameters with the compact 178-parameter nearest-neighbor model. This combined model, thereafter referred to as the NUPACK-compatible model, is easily accessible by pointing NUPACK to our parameter set file (**Fig. S4a**; parameter file available in supplementary material). Note that when using NUPACK to predict free energy, we assumed a two-state model in alignment with the Array Melt dataset, which was filtered based on the two-state criterion. Additionally, the ensemble model yielded inferior results compared to the two-state model, both with the original *dna04* and with our new parameter sets (**Fig. S4b**).

236 We evaluated the models on held-out Array Melt data and two orthogonal DNA duplex melting datasets from the
237 literature. To assess model performance, we used both the raw mean absolute error (MAE) and the adjusted MAE,
238 which is more relevant for assessing ΔΔG errors (**Supplementary Notes**). Testing our NUPACK-compatible model
239 on the held-out Array Melt test data resulted in an adjusted MAE of 0.59 kcal/mol for ΔG_{37} (**Fig. 4d-e**). This new
240 parameter set also improved predictions for DNA hairpins designed with random sequences measured with UV
241 melting (**Fig. S4d**). We further asked whether the new parameter set, trained on DNA hairpin data, extends to unseen
242 DNA duplexes. Using NUPACK to predict unseen duplex melting data, we observed a similar performance for
243 Watson-Crick strands with both the original *dna04* and the new parameter set. However, the new parameter set showed
244 improved performance on more complex variants. For a UV melting data set consisting of 384 Watson-Crick pairing
245 DNA duplexes[22], split evenly into validation and test sets, our parameter set performed comparably to *dna04* (**Fig.
246 S4e**). On a dataset from *Oliveira et al.* containing DNA duplex melting temperatures[15] with varying numbers of
247 mismatches (also split evenly between validation and test sets), both parameter sets had a similar adjusted MAE of
248 around 0.6 °C for Watson-Crick duplexes on the test set; while for single mismatches, the new parameter set achieved
249 a much lower adjusted MAE of 1.15 °C ($R^2 = 0.68$) compared to 1.66 °C ($R^2 = 0.25$) with *dna04*. The improvement
250 was even greater for double and triple mismatches (**Fig. 4f-g**). In conclusion, this updated parameter set, while simple,
251 enhances the prediction accuracy of NUPACK for DNA secondary structures, particularly those with mismatches and
252 hairpin loops. This improvement extends to orthogonal DNA duplex data that the model was not trained on.

253 ***Enhanced linear model with rich parameterization***

254 Building on this NUPACK-compatible model, we next developed an enhanced linear model that incorporates a richer
255 parameter set ("richer parameter model"), allowing still more accurate and comprehensive prediction of DNA
256 thermodynamics on unseen validation data. Because this model has more parameters than are implemented in the
257 NUPACK, it is not immediately compatible with the software. We fit this model with 1338 free parameters,
258 incorporating still more context information (including interactions between the two stacks flanking a mismatch, bulge
259 sequence identity, and full hairpin loop sequence for triloops and tetraloops, see **Methods**), to the Array Melt training
260 data. This rich parameter nearest-neighbor model yields an even better performance both on the Array Melt test data
261 (adjusted $MAE = 0.34\ kcal/mol$, $R^2 = 0.81$) and the external duplex datasets (**Fig. 4d-g, Fig. S4e**). Notably, when
262 comparing the prediction error on Array Melt validation dataset as we titrate the percentage of training data used to
263 fit the model, the 178 parameter linear regression model used to build the NUPACK-compatible model plateaus with
264 just a small fraction of training data used, while the 1338 parameter model requires more measurements to plateau,
265 but still does so after being fit on less than half of the training data set (**Fig. 4h**). Given the relatively overdetermined
266 nature of even this rich parameter nearest-neighbor model, this behavior is expected, and clearly demonstrates that
267 increasing data size allows for more variance to be captured by a more expressive model.

FIGURE 4


a
Error on Array Melt test set
ΔG37
ΔH
Watson Crick
single mismatch
adjusted MAE (kcal/mol)
length of stack features
b
NN parameters
Fitted
Original
Nearest neighbors
c
Single mismatch NUPACK parameters
New
Original
context
mismatch
d
On Array Melt test set
all
Watson Crick
mismatch
bulge
tetraloop
triloop
adjusted MAE = 0.87 kcal/mol
R = 0.65, R² = -0.18
MAE = 0.50
R = 0.86
R² = 0.66
MAE = 1.09
R = 0.52
R² = -0.87
MAE = 0.75
R = 0.48
R² = -1.05
MAE = 0.54
R = 0.68
R² = 0.35
MAE = 0.48
R = 0.58
R² = 0.02
MAE = 0.59 kcal/mol
R = 0.80, R² = 0.44
MAE = 0.47
R = 0.89
R² = 0.74
MAE = 0.61
R = 0.84
R² = 0.34
MAE = 0.69
R = 0.51
R² = -0.66
MAE = 0.40
R = 0.84
R² = 0.62
MAE = 0.36
R = 0.75
R² = 0.42
MAE = 0.34 kcal/mol
R = 0.90, R² = 0.81
MAE = 0.31
R = 0.95
R² = 0.89
MAE = 0.41
R = 0.85
R² = 0.72
MAE = 0.32
R = 0.79
R² = 0.60
MAE = 0.27
R = 0.92
R² = 0.85
MAE = 0.23
R = 0.88
R² = 0.77
measured ΔG37 (kcal/mol)
e
NN model
original NN model
NUPACK-compatible NN model
richer parameter NN model
f
On an orthogonal duplex dataset
number of mismatches
g
MAE = 0.60
R = 0.96, R² = 0.87
MAE = 1.66
R = 0.59, R² = 0.25
MAE = 2.17
R = 0.43, R² = 0.13
MAE = 1.91
R = 0.16, R² = 0.03
no mismatch
single mismatch
double mismatch
triple mismatch
MAE = 0.63
R = 0.96, R² = 0.86
MAE = 1.15
R = 0.85, R² = 0.68
MAE = 2.02
R = 0.71, R² = 0.26
MAE = 1.67
R = 0.49, R² = 0.24
MAE = 0.47
R = 0.96, R² = 0.92
MAE = 0.95
R = 0.87, R² = 0.75
MAE = 2.09
R = 0.74, R² = 0.21
MAE = 1.60
R = 0.53, R² = 0.27
predicted Tm (°C)
Tm (°C)
h
Error on Array Melt test set
sequence variants
% training data used
compact NN model (178 free features)
richer parameter NN model (1338 free features)

268

Figure 4. Enhanced linear regression models for DNA folding energetics

a. Adjusted mean absolute error (adjusted MAE) of linear regression models with features based on single base pair, nearest neighbor, or next nearest neighbor stacks, applied to Array Melt test set data.

b. Comparison of nearest-neighbor parameters for Watson-Crick pairs fitted from Array Melt data and those reported in the literature[8].

c. Visualization of new and the original *dna04* free energy parameters for single mismatches. "Context" and "mismatch" definitions are consistent with those in **Figure 3a**.

d-e. Performance comparison of the original (*dna04*), NUPACK-compatible, and richer parameter nearest-neighbor models on held-out Array Melt data. All calculations were performed at 1M Na^+ concentration. NN: nearest-neighbor.

f-g. Performance comparison of the original (*dna04*), NUPACK-compatible, and richer parameter nearest-neighbor models on an orthogonal dataset of DNA duplexes with varying numbers of mismatches (*Oliveira et al.*[15]).

h. Adjusted MAE on Array Melt data plotted as a function of the percentage of training data used to fit the linear regression model.

Graph neural network (GNN) model for DNA energy prediction

Because our rich linear model likely had insufficient power to capture the biological variance in our data, we developed a graph neural network model[23,24] specifically designed for predicting DNA energy. In this model, a DNA molecule is represented as a graph where nodes represent nucleotides and edges represent chemical bonds. Nucleotides (A, T, C, or G) and chemical bonds (either 5' to 3' backbone, 3' to 5' backbone, or hydrogen bond) are one-hot encoded as node and edge features, respectively. Taking these graph representations as inputs, the GNN iteratively refines node embeddings across graph convolutional layers. Essentially, nodes exchange information through different edge types, allowing, over multiple layers, for the embedding of one node to encapsulate information from the entire graph. Investigation of various types of graph convolutional layers revealed that the graph transformer layers[25] outperformed other architectures, such as graph attention layers[26] (**Fig. S5a**). After graph convolution, each node possesses a learned embedded feature, reflecting not only the original node features, but also the edges and overarching graph structures. To embed this graph into a fixed-length vector representation, we employed a global pooling layer employing the Set2Set algorithm[27], which used a Long Short-Term Memory (LSTM) processing block. This block systematically ingests node embeddings while updating its internal state, resulting in a vector representation of the entire graph that is order-invariant to the input node features. This processed summary vector is then fed into a small, fully connected neural network, which regresses for ΔH and Tm values (**Fig. 5a**).

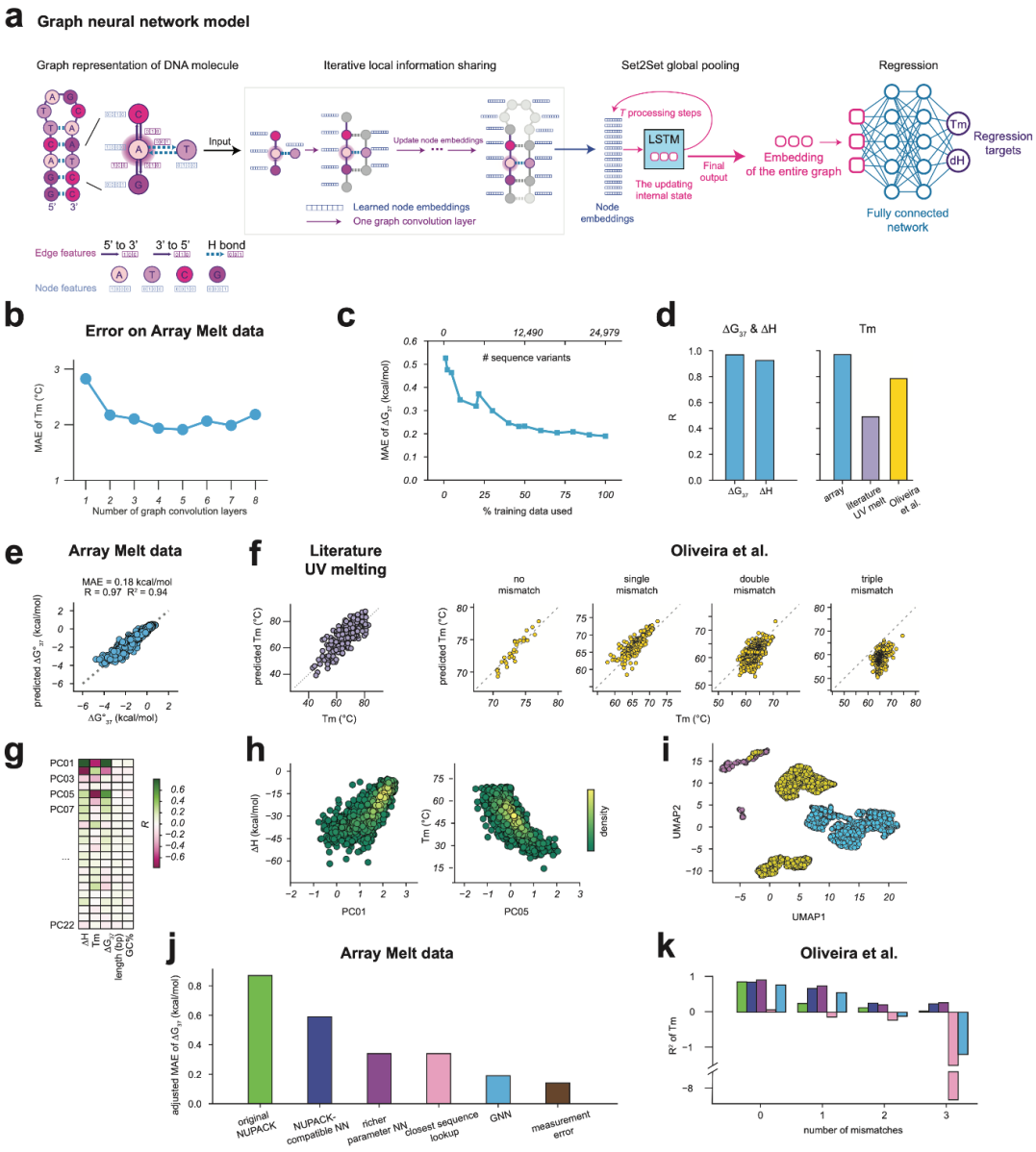
We trained this end-to-end model on Array Melt data. A portion of the Array Melt data was held out as validation and test set. We first explored the effect of the number of graph convolutional layers on prediction accuracy (**Fig. 5b, Fig. S5b**). Performance converged at around 4 graph convolutional layers, and did not improve beyond 4 layers, consistent with our observations in **Fig. S3h** that pairwise nucleotide interactions span around a distance of 4 nt. When evaluating model performance over an increasing percentage of training data used as in **Fig. 4h**, we observed a slower plateauing of the curve, indicating that the model can make use of this large-scale data set (**Fig. 5c**). The trained GNN achieved prediction error of 0.18 kcal/mol in dG_{37} ($R = 0.97$, $R^2 = 0.94$), compared to 0.34 kcal/mol of a baseline model that simply looks up the k=8 closest variants in the training set and gives a weighted average of the lookup values as prediction (**Fig. 5d-e, Fig. S5c-d**), and the GNN predictions are consistently high quality for different types of

sequence variants (**Fig. S5e**). Interestingly, although the GNN was trained entirely on graph representations of DNA hairpins without knowing the nature of backbones and base pairing, it is able to extend to the annealing and melting properties of DNA duplexes (**Fig. 5f**). While the GNN performance is slightly worse than our rich parameter nearest-neighbor model (**Fig. 4f-g, Fig. S4e**) on DNA duplexes, this level of performance suggests that the neural network has grasped the "gist" of the physical system of DNA folding thermodynamics and can extrapolate to non-hairping annealing contexts. Compared to the slow decrease of prediction error on the Array Melt data, prediction error rapidly decreased within around the first 20 epochs and then plateaued for the two DNA duplex datasets (**Fig. S5f**), likely because of the relatively low complexity of those datasets.

To better understand how the GNN makes predictions, we analyzed the intermediate activation from the output of the Set2Set aggregation layer. The GNN projects each DNA molecular graph into a 250-dimensional embedding vector. Many of the top principal components of the embeddings correlate strongly with predicted thermodynamics parameters ($R = 0.8$ between PC1 and ΔH, $R = -0.77$ between PC5 and Tm, and $R = 0.79$ between PC1 and ΔG_{37}), but none with other qualities of the sequence variants including DNA length or GC content (**Fig. 5g-h**), underscoring the GNN's focus on thermodynamics. When examining embeddings from different datasets (Array Melt, literature UV, and Oliveira et al.), each dataset's distinct design structure mapped to unique areas in a two-dimensional UMAP visualization (**Fig. 5i**). This observation indicates that the embeddings capture more than just predicted thermodynamics like melting temperatures (**Fig. S5g**), and that the learned embedding space encompasses sequence and structure details beyond mere parameter prediction.

Finally, we compared prediction errors across various models. The graph neural network is the top performer on Array Melt test data, while the 1778-parameter richer parameterization nearest-neighbor model performs best when applied to duplex data. Precision of the GNN is noteworthy, nearing the uncertainty range of bootstrapped experimental measurements (mean of uncertainty is 0.14 kcal/mol in ΔG_{37}), highlighting its effectiveness in capturing complex relationships in DNA sequences and structures (**Fig. 5j, Table 1**).

FIGURE 5



331

332

Figure 5. Graph neural network model for DNA folding thermodynamics

a. Schematic representation of the graph neural network (GNN) architecture.
b. Mean absolute error (MAE) of melting temperature on Array Melt data as a function of the number of graph convolution layers in the GNN.
c. MAE of ΔG_{37} on Array Melt data as a function of the percentage of training data used to fit the GNN.
d. Pearson's *R* correlation coefficient between GNN predictions and held-out Array Melt data, literature UV melting data, or Oliveira et al. duplex melting data.
e. Scatter plot comparing GNN predictions with held-out Array Melt data. Each dot represents a single sequence variant.
f. Scatter plot comparing GNN predictions with literature UV melting data or Oliveira et al. duplex melting data.
g. Pearson's *R* between principal components of the learned variant embeddings and measured thermodynamic parameters or variant properties.
h. Comparison between the most correlated principal components and ΔH or Tm.
i. UMAP visualization of learned variant embeddings after aggregation, colored by dataset (Array Melt, literature UV melting, or Oliveira et al. duplex melting data).
j. Benchmark of all models on held-out Array Melt data, compared to the measurement error of Array Melt.
k. Benchmark of all models on the orthogonal Oliveira et al. dataset, plotted by the number of mismatches. Colors correspond to models as in **Fig. 5j**.

		Model	Original NUPACK	NUPACK-compatible	Rich parameterization	Closest lookup	Graph neural network	Measurement
Array	ΔH	Adjusted MAE	8.59	6.27	4.97	5.27	3.13	3.09
		MAE	14.33	10.79	5.03	4.03	3.13	3.09
		R^2	0.12	0.48	0.64	0.77	0.86	NA
	Tm	Adjusted MAE	7.94	6.52	5.09	3.72	1.80	1.28
		MAE	8.38	6.53	5.65	3.72	1.80	1.28
		R^2	-0.06	0.26	0.20	0.78	0.94	NA
	ΔG_{37}	Adjusted MAE	0.87	0.59	0.34	0.34	0.19	0.14
		MAE	1.07	0.67	0.35	0.34	0.18	0.14
		R^2	-0.15	0.44	0.81	0.81	0.94	NA
Oliveira et.al	Tm (all)	Adjusted MAE	2.06	1.77	2.76	5.12	5.30	NA
		MAE	5.43	4.70	10.20	9.01	5.56	NA
		R	0.81	0.86	0.66	0.35	0.78	NA
		R^2	0.65	0.73	0.42	-1.06	0.49	NA
	R^2 of Tm	No mismatch	0.87	0.86	0.92	0.07	0.78	NA
		Single	0.25	0.68	0.75	-0.14	0.56	NA
		Double	0.13	0.26	0.21	-0.23	-0.12	NA
		Triple	0.03	0.24	0.27	-8.33	-1.21	NA

Table 1. Model benchmarks.

MAE: mean absolute error. Units of adjusted MAE and MAE are kcal/mol for ΔH and ΔG_{37} and °C for Tm.

Numbers in the "Measurement" column are means of bootstrapped uncertainties from measurement.

Discussion

In this work, we developed an accurate, scalable fluorescence-based method for measuring the melting of DNA hairpins (Array Melt) and used it to generate the largest-to-date DNA folding thermodynamics dataset. The Array Melt method, with its high throughput and precise measurement capabilities, allows for a more detailed exploration of the thermodynamics of diverse DNA structures, including those with mismatches, bulges, and various loop motifs. The scale of this data set revealed the significant role played by sequences adjacent to mismatches and within loops to the thermodynamic stability of hairpins. This observation implies that traditional models that parameterize only the immediate nearest neighbors must be underpowered to capture the variance observed in these molecular variants. To this end, we developed three quantitative folding energy models of increasing complexity, including revised parameterization of the linear regression model compatible with NUPACK, a richer nearest-neighbor model especially accurate for duplex annealing of mismatched molecules, and graph neural network (GNN) model that accells at predicting the melting behaviors of hairpins. To our knowledge, this GNN model is the first deep learning model applied to DNA folding. These results support the notion that while traditional nearest-neighbor models are accurate in predicting Watson-Crick stack behaviors, they fall short in accurately modeling the thermodynamics of more complex DNA structures, such as hairpins and interior loops. We anticipate that future library designs will be more oriented towards machine learning, enabling a more comprehensive and less biased sampling of the DNA sequencing space.

Overall, by providing a more accurate model for DNA folding thermodynamics, we pave the way for improved design of biotechnological tools and applications. For example, our models will provide a more accurate starting point for *in silico* design in areas such as qPCR primer design, DNA oligo hybridization probe design, and DNA origami. However, the study's focus on two-state folding presents a notable limitation, although our comparison between the two-state and ensemble methods (**Fig. S4b**) revealed greater accuracy of the former in predicting Array Melt measured thermodynamic parameters. The inferior performance of the ensemble method may be attributed to the inference methods of parameters. Additionally, the experimental design of the datasets used to derive nearest-neighbor models may also play a role. Therefore, interpreting nearest-neighbor model predictions and dynamic programming algorithm outputs as "true" energies associated with monolithic secondary structures warrants careful consideration. In this study, the two-state folding energies are better interpreted as an aggregated average across multiple folding states, rather than that of a single folded microstate within this ensemble. We anticipate that future analysis incorporating experimental design, data preprocessing, and parameter inference[28] techniques that consider ensemble folding could result in further improvements in prediction accuracy. This is particularly relevant when predicting probabilities of microstates of dynamic molecules is important.

This work also indicates that cloud labs can provide researchers with practical access to specialized instruments and support, addressing traditional challenges in equipment access and maintenance. These platforms also facilitate streamlined data collection and metadata tracking, improving open data and experiment reproducibility. This study

390 represents an early instance of using commercial cloud labs in research setting[29], suggesting potential broader
391 academic use.

392 Finally, while our measurements were carried out at a single monovalent salt concentration, large-scale measurements
393 exploring the stability of DNA structure to the concentration and type of monovalent or divalent cation would be
394 relatively straightforward. Furthermore, the potential application of our method to RNA structures presents an exciting
395 avenue for future research. Given the diverse and critical biological functions of RNA, which are heavily dependent
396 on its secondary structure, applying our high-throughput and precise thermodynamic measurement technique could
397 unveil new dimensions in RNA biology and facilitate the development of RNA-based therapeutics.

398 Acknowledgement

399 We thank Dr. Rhiju Das (Stanford University) for comments on the manuscript; Dr. Mark Schnitzer (Stanford
400 University) for supporting the project; Dr. Niles Pierce and Dr. Mark Fornace (Caltech) for answering NUPACK-
401 related questions. We thank Emerald Cloud Lab (ECL) for the UV melting experiments, with special attention to
402 Malav Desai and Ben Kline for tutorials and technical support of Symbolic Lab Language (SLL) programming and
403 Brian Frezza for funding.

404 This work was supported in part by NIH grants R01GM111990, P50HG007735, R01HG009909, P01GM066275,
405 UM1HG009436 and R01GM121487 to W.J.G. W.J.G. acknowledges support as a Chan Zuckerberg Investigator.
406 E.M. acknowledges support from the Swedish Research Council (grant 2020-06459), the Foundation Blanceflor, and
407 the Science for Life Laboratory (SciLifeLab).

408 Author contributions

409 Conceptualization, W.J.G., Y.K., E.S., and H.K.W.; Methodology, Y.K, E.S.; Software, Y.K. and H.K.W.; Validation,
410 Y.K; Formal Analysis, Y.K.; Investigation, Y.K, E.S., H.K.W., W.R.B., and A.H.; Data Curation, Y.K.; Writing -
411 Original Draft, Y.K. and W.J.G.; Writing - Review & Editing, Y.K., E.S., H.K.W., E.M., and W.J.G.; Visualization,
412 Y.K.; Supervision, E.M. and W.J.G.; Funding Acquisition, W.J.G.

413 Declaration of Interest

414 Stanford filed a patent application on aspects of this work with H.K.W., E.S., R.D., W.J.G., A.H., W.B., and Y.K. as
415 inventors (WO2023028618). W.J.G. is a consultant and equity holder for 10x Genomics, Guardant Health, Quantapore
416 and Ultima Genomics, and cofounder of Protillion Biosciences. The other authors declare no competing interests.

417 Methods

418 **Library assembly and sequencing**

419 Designed library variants were synthesized into DNA by Twist Biosciences (South San Francisco, CA). Sequence
420 map of the library is provided in supplementary information as *array_melt_library.gb*, where the "filler sequence" is

421 a 5' partial segment of AACAACAACAACATACTAACAACAACATAACAAATCAAAA, its length and the variable
422 region adding up to 40 bp, such that all synthesized fragments have the same length despite different variable region
423 length. Reverse complement of the sequence map was ordered as a synthesized oligo pool, amplified using internal
424 primers to enrich for full-length library variants. The PCR reaction consisted of 1 in 8 dilution of the synthesized oligo
425 pool (final concentration 10 nM, but concentration of full-length library variants could be much lower), 200 nM of
426 each primer (T7A1library, D-TruSeqR2 Table S1), 1x Phire Hot Start II PCR Master Mix (ThermoFisher Scientific
427 F125L). The reaction proceeded for 9 cycles of 98˚C for 10 seconds, 56˚C for 30 seconds, and 72˚C for 30 seconds.
428 Reaction mixtures were purified using QIAquick PCR Purification Kit (Qiagen 28104) to remove primers and
429 proteins, then eluted into 20 uL elution buffer.

430 After initial amplification, the library was amplified with primers to bring in sequences compatible with Illumina
431 sequencing. This five-piece assembly PCR included two outside primers and two adapter sequences. The PCR reaction
432 consisted of 1 µl of the previous reaction, 137 nM of outside primers (*short_C* and *short_D*; Table S1), 3.84 nM of
433 the adapter sequences (*C-i7pr-bc-T7A1* and *D_TruSeqR2*; Table S1), 1x Phire Hot Start II PCR Master Mix
434 (ThermoFisher Scientific F125L). The reaction proceeded for 14 cycles of 98˚C for 10 seconds, 56˚C for 30 seconds,
435 and 72˚C for 30 seconds. Reactions were purified using the QIAquick PCR Purification Kit and quantified with a
436 Qubit Fluorometer (ThermoFisher Scientific).

437 Sequencing was run with MiSeq Reagent Kit v3, 150 Cycles (MS-102-3001). The amplified library was diluted and
438 pooled to 1.16 nM final concentration in the pool (27% sample), together with a final concentration of 2.8 nM PhiX
439 control v3 (70% sample, Illumina), 0.04 nM fiducial marker (1% sample), 0.04 nM Cy3 no RNA control (1% sample),
440 and 0.04 nM single-strand fluorescence control sequences (1% sample, not included in analysis), adding up to a total
441 sample concentration of 4 nM. Paired-end sequencing was done with a custom read 1 primer (*stall_R1_primer*) and a
442 standard read 2 primer to cover the variable region in both read 1 and read 2 directions.

443 **Imaging station setup**

444 An imaging station was used to image the Miseq chip at increasing temperatures. This station was built from a
445 combination of custom-designed parts from a disassembled Illumina genome analyzer IIx[18,30]. Two channels were
446 employed: the "red" channel used the 660 nm laser and 664 nm long pass filter (Semrock BLP01-664R-25), and the
447 "green" channel used the 532 nm laser and 590 nm band pass filter (Semrock FF01-590/104-25). All images were
448 taken with 600 ms exposure times at 150 mW fiber input laser power. Focusing at each temperature was achieved by
449 sequentially adjusting the z-position and re-imaging the four corners of the flow cell; the adjusted $z_{\text{-positions}}$ were then
450 fit to a plane. We chose Cy3 as the fluorophore because the "green" channel is relatively more stable in our instrument
451 setup.

452 Post-sequencing, the chip was washed with Cleavage buffer (100 mM Tris-HCl, 125 mM NaCl, 0.05% Tween20, 100
453 mM TCEP, pH 7.4) to remove residual fluorescence from the reversible terminators used in the sequencing reaction
454 at 60˚C for 5 minutes. Any strands of DNA not covalently attached to the surface of the chip was removed by washing
455 in 100% formamide at 55˚C. The resulting single-stranded DNA fragments were incubated with 500 nM of the oligo

456 *Biotin_D_Read2* and *red_oligo* (Table S1) in Hybridization buffer (5x SSC buffer (ThermoFisher 15557036), 5 mM
457 EDTA, 0.05% Tween20) for 15 minutes at 60 °C, subsequently the temperature was lowered to 40 °C for another 10
458 minutes. The *red_oligo* is conjugated to Alexa647 and binds a sparse subset of "fiducial" sequences on the chip, whose
459 fluorescence signal is used for initial image registration to the sequenced tiles.

460 **Measuring melt curves on the chip**

461 The Cy3 labeled *fluor_oligo* was annealed at 40˚C for 5 minutes, then the Black Hole Quencher (BHQ)-labeled oligo
462 *quench_oligo* (**Table S1**) using the same protocol. The chip was imaged after each step to ensure that hybridization
463 had occurred through increase or quenching of signal in the corresponding channel. The chip was then rinsed with
464 melt buffer (50 mM Na-HEPES pH 8.0, 25 mM NaCl, total Na^+ concentration about 88 mM at pH 8.0). We chose
465 HEPES buffer because it has one of the lowest sensitivities to temperature (-0.014 ΔpK_a/°C) among the most
466 commonly used buffers in biological experiments[31].

467 To quantify the melt curves, image station temperature was then lowered to 20˚C. For each temperature point, the
468 system was allowed to equilibrate to the new temperature (approximately 5 minutes) before refocusing and imaging.
469 The temperature was raised in 2.5˚C increments to a maximum temperature of 60˚C. Detailed protocol can be found
470 in XML imaging scripts in the supplementary information.

471 **Processing sequencing data**

472 Sequencing data from Illumina Miseq was processed to extract tile and coordinates of each sequenced cluster. Forward
473 and reverse paired-end reads were aligned using FLASH[32] with default settings. Consensus sequences from FLASH
474 were aligned to the reverse complement sequences of *fluor_oligo* and *quench_oligo* using a Needleman-Wunsch
475 alignment (nwalign3). For consensus sequences that successfully aligned to both with p-value <1e-3, evaluated as
476 described in[33], the variable region was extracted as the region between the two flanking regions and aligned to
477 sequences in the reference library. A reference sequence was assigned based on the best-scoring alignment with a p-
478 value < 1e-6. The resulting clusters were used for fluorescence quantification in the downstream data analysis.

479 **Fluorescence data processing and image fitting**

480 Fluorescent melt curve images were mapped to sequencing data from the Illumina Miseq. First, sequencing data was
481 processed to extract the tile and coordinates of each sequenced cluster. The coordinates were used to generate synthetic
482 images of clusters, which were registered to the fluorescent images iteratively to map the coordinates to the images at
483 sub-pixel resolution[34]. Once the identities of the clusters were determined, each cluster was fitted to a 2D normal
484 distribution to quantify its fluorescence[30].

485 **Distance sensitivity curve**

486 The repeat controls used to plot the distance sensitivity curve in Fig. 1d are repeats of varied lengths of the following
487 sequences: poly-A, T, AT, AAC, AAG, AAT, AC, ACC, AG, AGG, TC, TG, TTA, TTC, and TTG.

488 The theoretical curve fitted to the distance-to-fluorescence curve in Fig. 1d is $f(x) = \frac{1}{1+(a\cdot(x+b)^{-3})}$, where x is the
489 length in nt, $f(x)$ is the dimensionless fluorescence at length x, normalized to maximum fluorescence, b is a fitted
490 parameter in nt, which accounts for systemic distance offset, and a is another fitted parameter:

491 $$a = \frac{2\pi^{2/3}}{3}(K_{D,BHQ} \cdot N_A \cdot k^3)^{-1} \qquad (1)$$

492 where $K_{D,BHQ}$ is the dissociation constant for the fluor-quench pair, N_A is Avogadro's number, and $k =$
493 $0.64 \times 10^{-9}\, m \cdot nt^{-1}$ is a distance conversion factor from nt to physical distance for ssDNA. The fitted parameter
494 values are $a = 71.19$, $b = 0.08$, and $K_{D,BHQ} = 145.6\, mol \cdot L^{-1}$. Using $\Delta G = RT \cdot ln\, K_{D,BHQ}$, we have $dG_{37} =$
495 $3\, kcal/mol$.

496 **Fluorescence normalization**

497 ***Size normalization***

498 To reduce inter-cluster variation in fluorescence measurements, we normalized the Cy3 fluorescence during the melt
499 curve measurements by the total amount of ssDNA in that cluster, measured as Cy3 fluorescence right after
500 hybridization of *fluor_oligo* but before hybridization of *quench_oligo*. The normalizing factor Cy3 signal was clipped
501 to the 1st and 99th percentile of the total distribution.

502 ***Temperature dependence and photobleaching normalization***

503 Across the timespan of the experiment, fluorescent signals are dependent on temperature and undergo photobleaching.
504 To correct for these factors, we further normalized the size-normalized signals of clusters to median signals of control
505 variants designed to have constant distances between the Cy3 fluorophore and the quench, including 1) "super stable
506 stem" variants with long GC stems that are not supposed to melt within the experimental temperature range and show
507 minimum fluorescence signals, and 2) "long repeat" controls that are at least 39 nt long, not supposed to form a hairpin,
508 and show maximum fluorescence signals. For both control variant groups, only the variants whose median signal
509 across clusters fall within the 5th and 95th percentile of all median signals at each temperature point within that control
510 variant group are used. Each variant is allowed to fall outside the percentile range at 1 temperature point at most. This
511 step filters out the variants not exhibiting the expected behavior. Afterwards, the median values at each temperature
512 point of both groups are taken and used to normalize the entire dataset. **(Fig. S1d)**

513 **Direct melt curve fitting**

514 The melt curve fitting pipeline was adapted from[34]. Briefly, the direct fitting process was carried out in the following
515 steps: 1) Directly fit normalized signals of single clusters; 2) Estimate the distribution of f_{max} and f_{min} from single
516 cluster fit results; 3) Refine the fit at the variant level with bootstrapping. The details are described below. The fitting
517 process is available as a snakemake[35] pipeline at https://github.com/keyuxi/array_analysis.

518 *1) Single cluster fitting*

519 **Equation (2)** was directly fitted to normalized signals of single clusters with minimal constraints, where T and T_m are
520 temperatures in Kelvin, and $k_B = 0.0019872$ is the Boltzmann constant. For initial values, Tm was set to the
521 temperature point where the normalized signal is closest to 0.5, ΔH to -40, f_{max} to 1, and f_{min} to 0. The python
522 package *lmfit* (v1.0.3)[36] was used for least-squares fitting.

523 $$f[T] = f_{min} + \frac{f_{max} - f_{min}}{1 + exp(\frac{\Delta H}{k_B} \cdot \left(\frac{1}{T_m} - \frac{1}{T}\right)} \qquad (2)$$

524 *2) Estimating distributions of f_{max} and f_{min}*

525 We used the following empirically gated criteria to decide if a single cluster fit was successful: R^2 of the fit at least
526 0.5, fitted f_{min} between -1 and 2, the estimated standard error of the best-fitted parameter value $\sigma_{f_{min}}$ less than $f_{min} +$
527 1, fitted f_{max} between 0 and 3, $\sigma_{f_{max}}$ less than f_{max}, σ_{Tm} less than 10 K, and $\sigma_{\Delta H}$ less than 100 kcal/mol. A p value is
528 calculated for each variant with a null hypothesis of 25% successful single cluster fits as in[34]. The single clusters are
529 then aggregated to the variant level. Median values of the parameters of each variant were calculated. To select for
530 variants used for estimation of f_{max}, the previously mentioned single cluster criteria were combined with the following
531 extra criteria: 1) be a good fitter, defined as having a success p value less than 0.01; 2) be unfolded at high
532 temperatures, defined as having a fitted f_{max} value of at least 0.5, and $p_{unfolded}^{60°C}$, the portion of unfolded molecules in
533 the cluster calculated from the fitted parameters, of at least 0.975. Similarly, for f_{min} we select for variants with a p
534 value less than 0.01, f_{min} less than 0.05, and $p_{unfolded}^{20°C}$ at most 0.025. We then fitted cluster-per-variant-dependent
535 Gamma distributions to f_{max} and f_{min} as in[34].

536 *3) Fit refinement*

537 For each variant, we separately decided whether to enforce f_{max} and f_{min} distributions to correct for variants that
538 never become fully unfolded or folded during the experiment. Specifically, we calculated the lower bound for f_{max} as
539 $\mu_{f_{max}} - 10 \cdot \sigma_{f_{max}}$, where $\mu_{f_{max}}$ and $\sigma_{f_{max}}$ are the mean and standard deviation calculated from the Gamma
540 distribution with the according number of clusters per variant from the previous step, and the upper bound for f_{min} as
541 $\mu_{f_{min}} + \sigma_{f_{min}}$. Here, we set the margin for f_{max} larger to enforce global f_{max} less, as f_{max} is more variable between
542 variants than f_{min} because of sequence-dependent effects. If the median signal at the highest temperature point is less
543 than the lower bound for f_{max}, we enforce f_{max}; likewise, if the median signal at the lowest temperature point is
544 greater than the upper bound for f_{min}, we enforce f_{min}.

545 We fitted each variant by bootstrapping the clusters within that variant. At each resampling step, we draw a f_{max} or
546 f_{min} value from the global distribution if enforced or keep f_{max} or f_{min} floating otherwise, and fit **equation (2)** to the
547 median signal of the drawn clusters, using the single cluster fit results as the initial parameters. Then, we calculate
548 ΔG_{37} and ΔS for each resampling step:

549 $$\Delta G[T] = \Delta H - T \cdot \Delta S = \Delta H \cdot \left(1 - \frac{T}{T_m}\right) \qquad (3)$$

550 $$\Delta S = \frac{\Delta H}{Tm} \qquad (4)$$

551 where T is temperature in Kelvin. We repeated the process 100 times and obtained a 95% confidence interval and
552 standard deviation for each parameter, ΔH, Tm, ΔG_{37} and ΔS.

553 **ΔG Line fitting**

554 We also fitted the data with an alternative line fitting to obtain heuristics for two-state behavior. We first re-normalize
555 the signal $f[T]$, previously normalized to initial fluorescence after hybridization as aforementioned, with fitted f_{max}
556 and f_{min}. We clip $f_{norm}[T] = (f[T] - f_{min})/(f_{max} - f_{min})$ between 0 and 1 with an $\epsilon = 0.01$ to avoid problems in
557 the next step. We transformed $f_{norm}[T]$ to $\Delta G[T]$ with

558 $$\Delta G[T] = -k_B T ln\left(\frac{1}{f_{norm}[T]} - 1\right) \qquad (5)$$

559 Error propagation was calculated with the uncertainties python library. Finally, we fitted with a line with free
560 parameters ΔH and Tm:

561 $$\Delta G[T] = \Delta H \cdot \left(1 - \frac{T}{T_m}\right) \qquad (6)$$

562 If the variant exhibits two-state behavior, $\Delta G[T]$ should fit well to a linear line. We first fit with RANSAC[37] to identify
563 outliers, then re-fit the inliers with ordinary least squares linear regression for the final parameters.

564 **Filtering and combining data replicates**

565 Each variant measured and fitted to curves in a replicate experiment was first filtered by the criteria: at least measured
566 in 5 clusters, bootstrapped error in ΔG_{37} is less than 2 kcal/mol, error in Tm less than 25°C, error in ΔH less than 25
567 kcal/mol, and curve fitting RMSE less than 0.5. They were then compared to the line fitting results. To be labeled as
568 passing two-state criteria, the fitted ΔH value of a variant between the 2 methods need to agree within 20%, Tm within
569 2°C, the reduced χ^2 of the line fit below 1.5, and the number of inlier data points in line fitting greater than 10. The
570 replicates were then combined as described in[34]. Briefly, for each parameter (ΔH, Tm, ΔG_{37}, or ΔS), its measurements
571 X_i and bootstrapped error σ_i in each replicate, the combined error is $\sigma = \sqrt{\left(\sum_i \ \sigma_i^{-2}\right)^{-1}}$ and the combined parameter
572 $X = \sigma^2 \cdot \sum_i \ X_i/\sigma_i^2$, where the sum is over the replicates whose data are not missing for the given variant. The missing
573 values were ignored. We then apply a second filter for two-state behavior. A given variant needs to be labeled as
574 passing two-state criteria in at least 2 replicates to be kept. We combined 4 such replicate experiments, including 3
575 melting and 1 cooling curve measurements from 2 different chips (see table for detail). We apply a final filter on the

576 combined variants to only keep those whose Tm are between 25 and 55°C and fall within the dynamic range of the
577 system.

578 **UV melting**

579 ***Designing sequences for UV melting and quality control***

580 We randomly selected hairpin sequences from each category of the Array Melt library for validations in **Fig. 2** and
581 designed random 16-mers that would fold into a hairpin with the secondary structure ((((((....)))))) for **Fig. S4d**,
582 filtering those with misfold secondary structures with NUPACK. DNA oligo samples were synthesized by Integrated
583 DNA Technologies (IDT) and were run through Ion Exchange High Pressure Liquid Chromatography (IE-HPLC)
584 remotely on Emerald Cloud Lab (ECL) with DNAPac PA200, 9x250 mm Semi-Prep column (Thermo Fisher
585 Scientific 063421), and the fractions were measured by a UV-Vis detector at 260 nm. Only the oligos with a single
586 peak were used for data analysis.

587 ***Running cloud lab UV melting experiments***

588 UV melting experiments were run remotely on ECL. The Manifold data objects could be found at the data object links
589 in supplementary table 2. DNA oligos ordered from IDT were shipped to ECL and resuspended in high concentration
590 stock solutions. For hairpins, the DNA stock solutions are first snap freezed right before each experiment, being placed
591 at 90°C for 5 min and moved to 4°C for 30 min before diluted into working concentrations; for duplexes, the DNA
592 stock solutions are first diluted into working concentrations, then placed on 90°C for 5min and left annealing for 30
593 min to 1 hour. Melt curves are measured at varied concentrations, where the range is 3 to 12 uM for hairpins and 72
594 to 96 uM for duplexes. Hairpin data used for method validation in **Fig. 2** and model comparison in **Fig. S4d** were all
595 measured at 6 uM or 9 uM. Samples were measured on a Cary 3500 UV/Vis spectrophotometer, using micro scale
596 black walled UV quartz cuvettes with a pathlength of 10 mm. Samples were measured for 2 to 3 melt-cooling cycles
597 with a temperature ramp rate of 0.8°C and 15 min equilibrium time between each melt and cool sweep. Temperatures
598 were measured with an immersion probe. Data points were acquired every 1 °C at 260 nm from 5°C to 85 or 95 °C.

599 ***Analyzing UV melting curves***

600 The UV melting curves were fitted to a modified version of **equation (2)** with an extra term to account for baseline
601 slope:

602 $$f[T] = f_{min} + a \cdot T + \frac{f_{max} - f_{min}}{1 + exp(\frac{\Delta H}{k_B} \cdot \left(\frac{1}{T_m} - \frac{1}{T}\right)} \qquad (7)$$

603 where f[T] is the absorbance, a is a free parameter for the slope, and the rest the same as in **equation (2)**. For each
604 individual melt or cooling curve, we first found the temperature where the derivative $\frac{df[T]}{T}$ is maximum, that is, $T_m^{diff} =$
605 $argmax_T \frac{df[T]}{T}$, as the initial value of T_m. We fixed the value of T_m when fitting **equation (7)** in the first round and

let the parameters a, f_{max}, f_{min}, and ΔH float. Afterwards, we use the best-fitted values from the previous round as the initial values of the parameters and allow T_m to float in the second round of fitting. This two-step fitting process empirically improved the fitting RMSE. We selected the curves with an RMSE less than 0.015, $\sigma_{\Delta H} < 10$ And $\sigma_{T_m} < 5$ and aggregated the fitted ΔH and Tm values of the melt and cooling curves by each sample tube with $X = \sigma^2 \cdot \sum_i X_i/\sigma_i^2$ and $\sigma = \sqrt{\left(\sum_i \sigma_i^{-2}\right)^{-1}}$ as for the array experiments. We then calculated ΔG_{37} and ΔS from the aggregated ΔH and T_m using equations (3) and (4), and used the uncertainty propagation equations $\sigma_{\Delta S} = -\Delta S \cdot \sqrt{\left(\frac{\sigma_{\Delta H}}{\Delta H}\right)^2 + \left(\frac{\sigma_{T_m}}{T_m}\right)^2}$ and $\sigma_{\Delta G_T} = \sqrt{\sigma_{\Delta H}^2 + (T \cdot \sigma_{\Delta S})^2}$ to calculate the measurement uncertainties.

Salt correction

All salt corrections used the equation from[38]:

$$\frac{1}{T_m(2)} = \left(\frac{1}{T_m(1)} + (4.29 \cdot f(GC) - 3.95) \times 10^{-5} \cdot ln\frac{[Na^+]_2}{[Na^+]_1} + 9.4 \times 10^{-6}(ln^2[Na^+]_2 - ln^2[Na^+]_1)\right. \quad (8)$$

where $T_m(1)$ and $T_m(2)$ are the melting temperatures before and after salt correction in Kelvin, $[Na^+]_1$ and $[Na^+]_2$ the corresponding sodium concentrations, and $f(GC)$ the fraction of GC in the sequence.

Variance analysis

Mismatch $\Delta\Delta G_{37}$ calculation

For each mismatch, we find its parent sequence variant, which is a hairpin with Watson-Crick pairing stem and can derive into the mismatch by mutating a single nucleotide. The mismatches are categorized by the identities of those nucleotide substitutions. We then calculated $\Delta\Delta G_{37}$ of mismatches by subtracting ΔG_{37} of the parent variant from that of the mismatch variant. C>G single mismatches were excluded from the variance calculations because of insufficient number of data points (only 2 variants).

Expected biological variance and model explained or introduced variance

For a dataset of n sequence variants with n corresponding $d\Delta$ values $y_1, \dots, y_n$, measurement uncertainty values $\sigma_1, \dots, \sigma_n$, and nearest-neighbor model prediction values $\hat{y}_1, \dots, \hat{y}_n$, we have total variance ${\sigma^2}_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2/n$ and technical variance ${\sigma^2}_{tech} = \sum_{i=1}^{n} {\sigma_i}^2/n$. The expected biological variance ${\sigma^2}_{bio} = {\sigma^2}_{tot} - {\sigma^2}_{tech}$. Model variance is the variance of prediction values ${\sigma^2}_{model} = \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2/n$, and residual is ${\sigma^2}_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n$. From these we calculate variance explained ${\sigma^2}_{explained} = {\sigma^2}_{tot} - {\sigma^2}_{res}$. In the case where ${\sigma^2}_{res} > {\sigma^2}_{tot}$, we have variance added by model $\sigma_{\text{added}}^2 = -\sigma_{explained}^2 = \sigma_{res}^2 - \sigma_{\text{tot}}^2$.

Nucleotide importance score

We calculated nucleotide importance score (NIS) using $\Delta\Delta G_{37}$ for mismatches, and ΔG_{37} for triloop, tetraloop, and bulges. For a dataset of n sequence variants with n corresponding energy values $y_1, \dots, y_n$, either ΔG_{37} or $\Delta\Delta G_{37}$, we

group the n data points into four groups based on the identity of the nucleotide of interest. For each of the four groups, we calculated the mean, $\overline{y_A}$, $\overline{y_T}$, $\overline{y_C}$, and $\overline{y_G}$. The grand mean $\overline{y_{tot}}$ is the mean of $y_1, \ldots, y_n$. The NIS is simply the between group sum of squares (SSB):

$$NIS = \sum_{i=\{A,T,C,G\}} (\overline{y_i} - \overline{y_{tot}})^2 \qquad (9)$$

Pairwise interaction score

For a pair of nucleotides of interest, M and N, we compare two linear models that predict the energy value Y, model A with independence assumption and model B without. Model A is in the form of $M + N \sim Y$, and model B $MN \sim Y$. The pairwise interaction score (PIS) is the difference of R^2 between the two linear models

$$PIS = R^2{}_{MN\sim Y} - R^2{}_{M+N\sim Y} \qquad (10)$$

Calculations with NUPACK

Melting temperature

Melting temperatures of DNA duplexes were calculated from the fraction of molecules unpaired. We first calculated structure free energies of the paired state at temperature T, $\Delta G_{paired,T}$, then got the equilibrium constant

$$K = \frac{[dsDNA]}{[ssDNA]^2} = exp\left(-\beta \cdot \Delta G_{paired,T}\right) \qquad (11)$$

where $\beta = 1/RT$ is the thermodynamic beta, $[dsDNA]$ is the concentration of double-strand DNA, and [ssDNA] that of the single-strand DNA. At melting temperature, $[dsDNA] = [ssDNA] = [DNA]/2$ for distinguishable strands and $[dsDNA] = [DNA]/4$ for indistinguishable strands, where $[DNA]$ is the total concentration of DNA strand species. This gives $K = \frac{2}{[DNA]}$ or $K = \frac{1}{[DNA]}$ for distinguishable and indistinguishable strands respectively. Solving this equation with bisection gives us the melting temperature.

For DNA hairpins, melting temperature is independent of DNA strand concentration. The Boltzmann factor becomes:

$$e^{-\beta E} = exp\left(-\beta \cdot \Delta G_{folded}\right) \qquad (12)$$

T_m is then simply the temperature when $dG_{folded} = 0$. Therefore, we get structure free energies ΔG_1 and ΔG_2at a lower temperature T_1 and a higher one T_2, then calculate $\Delta S = -(\Delta G_2 - \Delta G_1)/(T_2 - T_1)$, $\Delta H = \Delta G_1 + T_1 \cdot \Delta S$, and $T_m = \Delta H/\Delta S$.

$$T_m = \Delta H / \left(\frac{\Delta H - \Delta G_{37}}{T_{37}} - R \cdot ln\, K\right) \qquad (13)$$

Data splitting for machine learning models

After filtering with the two-state criteria as described in "ΔG line fitting", we splitted the clean Array Melt into train, validation, and test sets, stratified by variant type. We first splitted 5% of the data as the test set (1438 variants), then splitted 5% of the remaining data as the validation set (1315 variants), and 95% of the rest as the training set (24979 variants). We splitted the external 384-duplex dataset and Oliveira et al. dataset to a 1:1 split of validation and test data. The cloud lab UV melting data were used for correction between Array Melt and bulk solution and were not used for model testing to prevent circularity.

Linear models

Weighted linear regression model

We used a singular value decomposition (SVD) solver to solve the linear regression solutions, taking measurement error of each sequence into account. For a training dataset of n sequence variants with n corresponding ΔH or ΔG_{37} values $y_1, \dots, y_n$, and measurement uncertainty values $\vec{\sigma} = [\sigma_1, \dots, \sigma_n]$, we have a n by $n_{feature}$ feature count matrix X derived from the sequences. We then did SVD on matrix $A = X/\vec{\sigma}$ and get U, S, V. Any diagonal value in S that is smaller than a relative numeric threshold 10^{-15} was set to zero. Then the estimated parameter values for the features are $V^T \cdot S_{inv} \cdot U^T \cdot \vec{b}$, and the estimated errors for the features are $\sqrt{\sum \; V^T \cdot S^2{}_{inv}}$, where $\vec{b} = \vec{y}/\vec{\sigma}$, S_{inv} is S but taking $\frac{1}{s}$ for all diagonal elements, and $S^2{}_{inv}$ is S but taking $\frac{1}{s^2}$ for all elements.

The NUPACK-compatible nearest-neighbor model

We trained the improved nearest-neighbor model on the Array Melt training data set, containing 24,971 of the 27,732 sequences that passed quality control and two-state criteria, and tuned the model on a validation set of 1,315 sequences. We first adjusted salt concentration to 1M Na^+ to be compatible with NUPACK and other softwares, and then corrected systematic errors by linearly shifting Tm and ΔG, such that the medians of each construct type match those of NUPACK predictions. The underlying assumption is that NUPACK predictions have minimal bias but could be improved by explaining more variance. Improved parameters for NUPACK were fitted with weighted linear regression. We allowed an extra intercept parameter to adjust mean absolute error. From sequence-structure pairs in the training set, we extracted features of stacks and loops. Stacks include nearest-neighbor Watson-Crick pairs and terminal stack penalty, as in NUPACK; loops include the following:

1) Hairpin loops, defined as a subsequence $\phi_{i:j}$ where $i \cdot j$ is the closing base pair. We extracted "hairpin_mismatch" features as in NUPACK, $\Delta G^{hairpinmm}_{\phi_{j-1},\phi_j,\phi_i,\phi_{i+1}}$, and introduced an intermediate "hairpin_loop_mid" feature, $\Delta G^{hairpinmid}_{\phi_{[i+2:j-2]}}$ to account for sequence variations in the middle of hairpin loops for triloops and tetraloops. The final energy of hairpin is $\Delta G^{hairpin}_{\phi_{[i:j]}} = \Delta G^{hairpinsize}_{n} + \Delta G^{hairpinseq}_{\phi_{[i:j]}}$ as in[20], where

$$\Delta G^{hairpinseq}_{\phi_{[i:j]}} = \Delta G^{hairpinmm}_{\phi_{j-1},\phi_j,\phi_i,\phi_{i+1}} + \Delta G^{hairpinmid}_{\phi_{[i+2:j-2]}} \left(+\Delta G^{terminalbp}_{\phi_j,\phi_i} if triloop\right) \qquad (14)$$

692 2) Interior loops, defined from 5' to 3' as two subsequences $\phi_{[i:d]}$ and $\phi_{[e:j]}$. Any other interior loop types were
693 defined as in[20], apart from the $\Delta G^{interiormm}_{\phi_{j-1},\phi_j,\phi_i,\phi_{i+1}}$, which were fitted separately instead of shared with terminal
694 mismatches as in NUPACK, and that we introduced an intermediate parameter for the interaction between
695 the two base pairs surrounding 1x1 mismatches $\Delta G^{interiormmstacks}_{\phi_i,\phi_j,\phi_{i+2},\phi_{j-2}}$.

696 This results in 178 features for ΔH and ΔG each. Parameters for hairpin and interior loop sizes were taken from
697 NUPACK and clamped during fitting. Lookup tables of $\Delta G^{hairpinseq}_{\phi_{[i:j]}}$ and $\Delta G^{interiorseq}_{\phi_{[i:d]},\phi_{[e:j]}}$ for "interior_1_1",
698 "interior_1_2", and "interior_2_2" were calculated from the fitted parameters. All updated parameters were then
699 written to a NUPACK-compatible json file.

700 Afterwards, we obtained parameters for $\Delta G^{hairpinseq}_{\phi_{[i:j]}}$ for tetraloops and triloops directly from the '2nd flanking'
701 construct types of either tetraloop or triloop parts of the library (**Fig. 1c**). These variants covered all possible sequences
702 for hairpin loops of length 3 or 4, as well as the nearest-neighbor and next-nearest-neighbor closing pair sequences.
703 We first regressed out the next-nearest-neighbor closing pair of the loops from measured ΔH and ΔG values, then
704 aggregated the mean values for variants with identical hairpin loop and nearest-neighbor closing pair sequences. These
705 aggregated values have the correct ΔΔH and ΔΔG relationships between each other but need a constant to be added to
706 get the ΔH and ΔG values needed for the nearest-neighbor model. To solve this problem, we fitted two correction
707 values $\Delta\Delta H$ and $\Delta\Delta G$ in kcal/mol. We searched for values that minimize the equation

708 $$f(\Delta\Delta H, \Delta\Delta G) = w_{Tm} \cdot RMSE(Tm) + w_{\Delta H} \cdot RMSE(\Delta H) + w_{\Delta G} \cdot RMSE(\Delta G) + \alpha \cdot (w_{\Delta H} \cdot \Delta\Delta H^2 + w_{\Delta G} \cdot \Delta\Delta G^2)$$
709 (15)

710 where $w_{Tm} = 4, w_{\Delta H} = 0.3, w_{\Delta G} = 3.4$ and $\alpha = 0.0001$ are weights determined by validation on 100 sampled
711 variants in the training set. Calculating this equation is computationally expensive, as it involves directly calling
712 NUPACK in each iteration. We minimized it on 25 randomly sampled variants from the training set, repeating the
713 process 4 times on different subsets using the BFGS algorithm implemented in the 'scipy.optimize.minimize' function,
714 with parameters tol=1e-1 and maxiter=20. The resulting $\Delta\Delta H$ and $\Delta\Delta G$ values were mostly consistent between the
715 runs (**Fig. S4g**). We took the median values from the 4 runs ($\Delta\Delta H = -28\ kcal/mol$, $\Delta\Delta G = -4\ kcal/mol$) and
716 added them to the aggregated $\Delta\Delta H$ and $\Delta\Delta G$ values of hairpin loops.

717 Empirically, we found that model performance on the validation set was improved when we augmented the compact
718 178-parameter nearest-neighbor model with these curated hairpin loop parameters. This improvement might be due
719 to the dominantly overrepresentation of 'GAAA' hairpin loops, which led to highly skewed data distribution and
720 caused directly regressed hairpin loop parameters to be less accurate.

721 ***The richer parameter nearest-neighbor model***

722 We trained the richer parameter model on the same Array Melt training data set as for the NUPACK-compatible
723 nearest-neighbor model. For feature extraction, we extracted nearest-neighbor Watson-Crick pairs and terminal stack
724 penalty as in NUPACK, and the following parameters for loops:

1) Hairpin loops. For triloops and tetraloops, $\Delta G^{hairpinseq}_{\phi_{[i:j]}} = \Delta G^{hairpinmm}_{\phi_{j-1},\phi_j,\phi_i,\phi_{i+1}} + \Delta G^{hairpinloop}_{\phi_{[i+1:j-1]}}$, where $\Delta G^{hairpinloop}_{\phi_{[i+1:j-1]}}$ is the full sequence of the hairpin loop, in contrast to only the middle of loops in the NUPACK-compatible model. For hairpin loops longer than tetraloop, we applied a single loop sequence parameter $\Delta G^{hairpinloop}_{long}$. The final energy of hairpin is also $\Delta G^{hairpin}_{\phi_{[i:j]}} = \Delta G^{hairpinsize}_{n} + \Delta G^{hairpinseq}_{\phi_{[i:j]}}$.

2) Interior loops. For single and double bulges or single mismatches, the sequence-dependent parameter is fitted for each sequence, $\Delta G^{interiorseq}_{\phi_{[i:d]},\phi_{[e:j]}}$. For double and triple mismatches, $\Delta G^{interiorseq}_{\phi_{[i:d]},\phi_{[e:j]}} = \Delta G^{interiormmstacks}_{\phi_i,\phi_j,\phi_{i+2},\phi_{j-2}} + \Delta G^{interiormm}_{\phi_{i+1},\phi_{d-1},\phi_{e+1},\phi_{j-1}}$.

Deep learning models

Closest sequence lookup baseline model

We defined the distance between any two sequence-structure pairs by summing up the edit distance between the sequences and the dot-bracket secondary structure representation. For every query sequence-structure pair, we find the k closest sequence-structure pairs in the training set, whose distance-weighted sum becomes the predicted value for the query using python package *scipy*[39]. We found the best value for k using the held-out Array Melt validation data.

Graph neural network model

We implemented the graph neural network using python packages PyTorch and PyTorch Geometric[40]. The final model has 4 graph transformer layers with a 125-vec node attribute at each layer, stacked with dropout of 0.01 at each layer, a Set2Set module with 10 processing steps, and a 1 fully connected hidden layer with 128 channels and a dropout ratio of 0.49. We normalized ΔH and Tm of the training data to fall within 0 and 1 and used the sum of the root mean square error (RMSE) of both ΔH and Tm as the loss function and trained with Adam optimizer for 200 epoches. The model has 287,136 parameters. The validation model registry was monitored and stored on weight and biases (wandb.ai).

Using the NUPACK-compatible model

The parameter file for the NUPACK-compatible nearest-neighbor model is available in the supplementary material as 'NUPACK_compatible_NN_model.json'. To use this file, download and point to it when using the python package NUPACK 4[41]:

```
my_model = nupack.Model(material='/path/to/parameter/file/NUPACK_compatible_NN_model.json', …)
```

Code availability

Code for array data preprocessing from raw sequencing and imaging data plus curve fitting could be found at https://github.com/GreenleafLab/array_analysis.

All subsequent analysis could be found at https://github.com/GreenleafLab/nnn_paper.

References

1. Kunkel, T. A. DNA Replication Fidelity. *J. Biol. Chem.* **279**, 16895–16898 (2004).

2. Rajski, S. R., Jackson, B. A. & Barton, J. K. DNA repair: models for damage and mismatch recognition. *Mutat.*

758 *Res. Mol. Mech. Mutagen.* **447**, 49–72 (2000).
759 3. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
760 4. Kibbe, W. A. OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Res.* **35**, W43–W46
761 (2007).
762 5. Labun, K. *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids*
763 *Res.* **47**, W171–W174 (2019).
764 6. Majikes, J. M. *et al.* Revealing thermodynamics of DNA origami folding via affine transformations. *Nucleic*
765 *Acids Res.* **48**, 5268–5280 (2020).
766 7. Zadeh, J. N. *et al.* NUPACK: Analysis and design of nucleic acid systems. *J. Comput. Chem.* **32**, 170–173
767 (2011).
768 8. SantaLucia, J. & Hicks, D. The Thermodynamics of DNA Structural Motifs. *Annu. Rev. Biophys. Biomol. Struct.*
769 **33**, 415–440 (2004).
770 9. SantaLucia, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor
771 thermodynamics. *Proc. Natl. Acad. Sci.* **95**, 1460–1465 (1998).
772 10. SantaLucia, John, Allawi, H. T. & Seneviratne, P. A. Improved Nearest-Neighbor Parameters for Predicting
773 DNA Duplex Stability. *Biochemistry* **35**, 3555–3562 (1996).
774 11. Zuker, M. On Finding All Suboptimal Foldings of an RNA Molecule. *Science* **244**, 48–52 (1989).
775 12. Schroeder, S. J. & Turner, D. H. Optical Melting Measurements of Nucleic Acid Thermodynamics. *Methods*
776 *Enzymol.* **468**, 371–387 (2009).
777 13. Spink, C. H. Differential Scanning Calorimetry. in *Methods in Cell Biology* vol. 84 115–141 (Academic Press,
778 2008).
779 14. Hadiwikarta, W. W., Walter, J.-C., Hooyberghs, J. & Carlon, E. Probing hybridization parameters from
780 microarray experiments: nearest-neighbor model and beyond. *Nucleic Acids Res.* **40**, e138 (2012).
781 15. Oliveira, L. M., Long, A. S., Brown, T., Fox, K. R. & Weber, G. Melting temperature measurement and
782 mesoscopic evaluation of single, double and triple DNA mismatches. *Chem. Sci.* **11**, 8273–8287 (2020).
783 16. Bae, J. H., Fang, J. Z. & Zhang, D. Y. High-throughput methods for measuring DNA thermodynamics. *Nucleic*
784 *Acids Res.* **48**, e89–e89 (2020).
785 17. You, Y., Tataurov, A. V. & Owczarzy, R. Measuring thermodynamic details of DNA hybridization using

786 fluorescence. *Biopolymers* **95**, 472–486 (2011).

787 18. Buenrostro, J. D. *et al.* Quantitative analysis of RNA-protein interactions on a massively parallel array reveals
788 biophysical and evolutionary landscapes. *Nat. Biotechnol.* **32**, 562–568 (2014).

789 19. Marklund, E., Ke, Y. & Greenleaf, W. J. High-throughput biochemistry in RNA sequence space: predicting
790 structure and function. *Nat. Rev. Genet.* **24**, 401–414 (2023).

791 20. Fornace, M. E., Porubsky, N. J. & Pierce, N. A. A Unified Dynamic Programming Framework for the Analysis
792 of Interacting Nucleic Acid Strands: Enhanced Models, Scalability, and Speed. *ACS Synth. Biol.* **9**, 2665–2678
793 (2020).

794 21. Tikhomirova, A., Beletskaya, I. V. & Chalikian, T. V. Stability of DNA Duplexes Containing GG, CC, AA, and
795 TT Mismatches. *Biochemistry* **45**, 10563–10571 (2006).

796 22. Khandelwal, G. & Bhyravabhotla, J. A Phenomenological Model for Predicting Melting Temperatures of DNA
797 Sequences. *PLOS ONE* **5**, e12433 (2010).

798 23. Yan, Z., Hamilton, W. L. & Blanchette, M. Graph neural representational learning of RNA secondary structures
799 for predicting RNA-protein interactions. *Bioinformatics* **36**, i276–i284 (2020).

800 24. Zhao, Q. *et al.* Review of machine learning methods for RNA secondary structure prediction. *PLOS Comput.*
801 *Biol.* **17**, e1009291 (2021).

802 25. Shi, Y. *et al.* Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification.
803 Preprint at https://doi.org/10.48550/arXiv.2009.03509 (2021).

804 26. Veličković, P. *et al.* Graph Attention Networks. Preprint at https://doi.org/10.48550/arXiv.1710.10903 (2018).

805 27. Vinyals, O., Bengio, S. & Kudlur, M. Order Matters: Sequence to sequence for sets. Preprint at
806 http://arxiv.org/abs/1511.06391 (2016).

807 28. Spasic, A. *et al.* Improving RNA nearest neighbor parameters for helices by going beyond the two-state model.
808 *Nucleic Acids Res.* **46**, 4883–4892 (2018).

809 29. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language
810 models. *Nature* **624**, 570–578 (2023).

811 30. She, R. *et al.* Comprehensive and quantitative mapping of RNA–protein interactions across a transcribed
812 eukaryotic genome. *Proc. Natl. Acad. Sci.* **114**, 3619–3624 (2017).

813 31. Good, N. E. *et al.* Hydrogen Ion Buffers for Biological Research*. *Biochemistry* **5**, 467–477 (1966).

32. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

33. Altschul, S. F. & Gish, W. [27] Local alignment statistics. in *Methods in Enzymology* vol. 266 460–480 (Academic Press, 1996).

34. Denny, S. K. *et al.* High-Throughput Investigation of Diverse Junction Elements in RNA Tertiary Folding. *Cell* **174**, 377-390.e20 (2018).

35. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

36. Newville, M., Stensitzki, T., Allen, D. B. & Ingargiola, A. LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python. Zenodo https://doi.org/10.5281/zenodo.11813 (2014).

37. Fischler, M. A. & Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**, 381–395 (1981).

38. Owczarzy, R. *et al.* Effects of Sodium Ions on DNA Duplex Oligomers: Improved Predictions of Melting Temperatures. *Biochemistry* **43**, 3537–3554 (2004).

39. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

40. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. Preprint at https://doi.org/10.48550/arXiv.1903.02428 (2019).

41. Fornace, M. E. *et al. NUPACK: Analysis and Design of Nucleic Acid Structures, Devices, and Systems*. https://chemrxiv.org/engage/chemrxiv/article-details/636c7089b588507d0045f283 (2022) doi:10.26434/chemrxiv-2022-xv98l.

Supplementary Figures

SUPP FIGURE 1

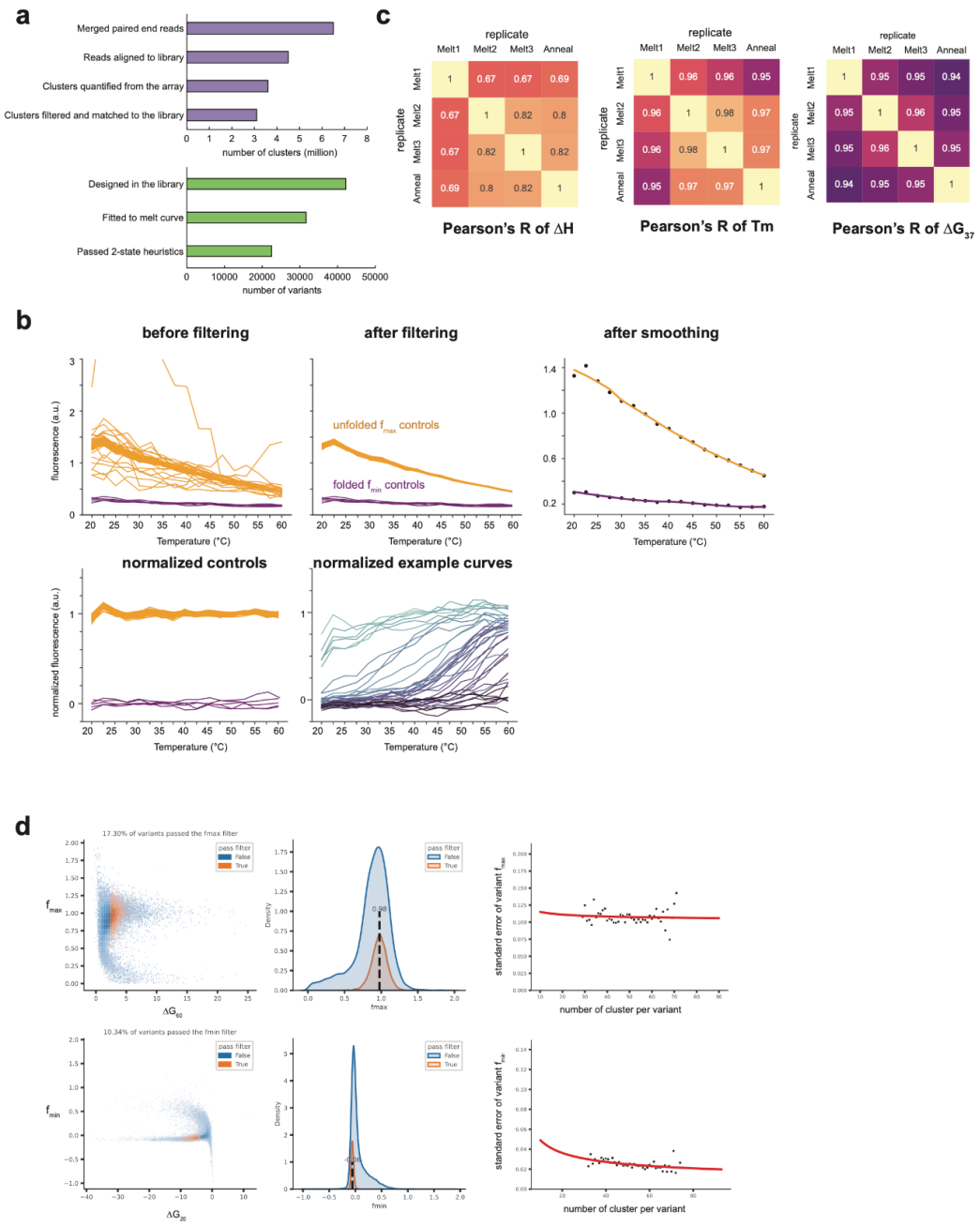


Figure S1

a. Number of measurements in a single Array Melt experiment.
b. Pearson's R of ΔH, Tm, and ΔG_{37} between replicates.
c. Controls used to normalize fluorescence across different temperatures, including always unfolded constructs for f_{max} (maximum fluorescence) and folded constructs for f_{min} (minimum fluorescence). Some Watson-Crick curves are randomly selected to visualize after the normalization, colored by their expected ΔG_{37} values predicted by the nearest-neighbor model.
d. Estimation of f_{max} and f_{min} after the initial round of curve fitting. Of all the fitted variants (blue), only a fraction that passed quality control filters (orange) were used to estimate f_{max} and f_{min} parameters used for the next round of refinement fitting. Uncertainties of f_{max} and f_{min} were corrected as a function of number of clusters per variant by curve fitting (right).

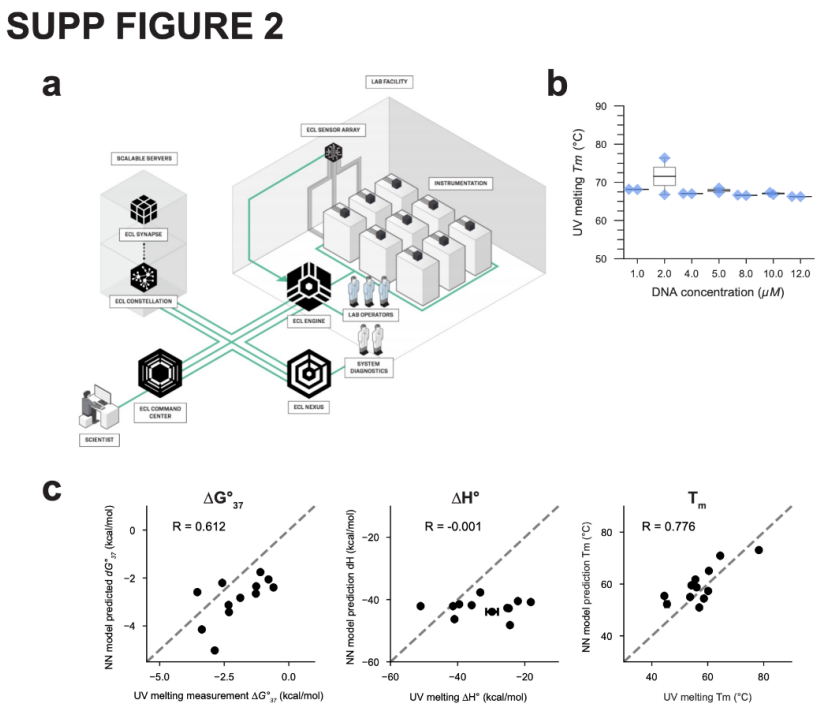
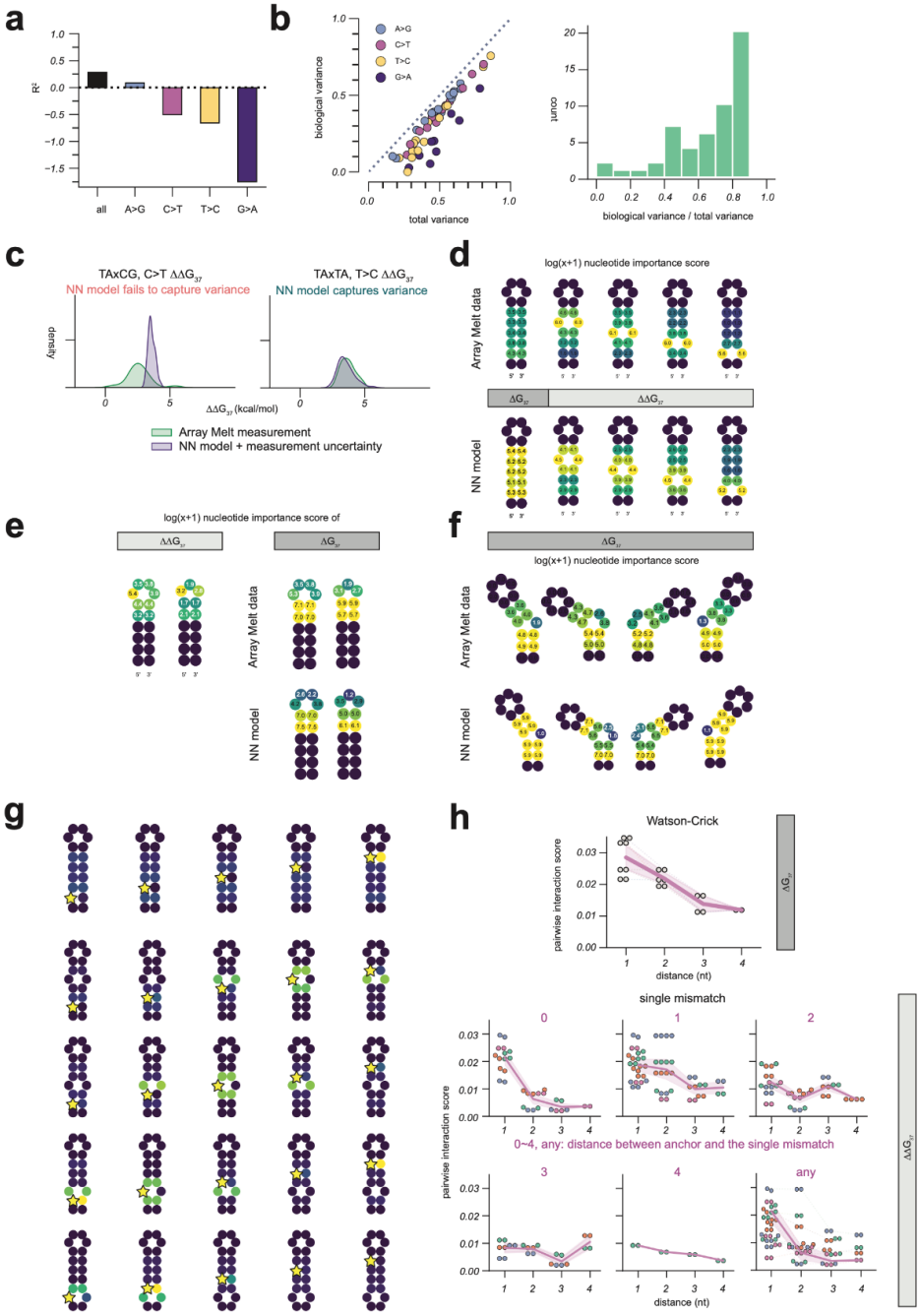


Figure S2

a. Schematic representation of Emerald Cloud Lab (ECL) workflow. Image from ECL with permission.
b. UV melting measured Tm at different DNA hairpin oligo concentrations. Oligo used is WC68.
c. Direct comparison of measured ΔG, ΔH and Tm between nearest-neighbor model prediction and UV melting data. Each datapoint is one hairpin variant.

SUPP FIGURE 3



854

SUPP FIGURE 3

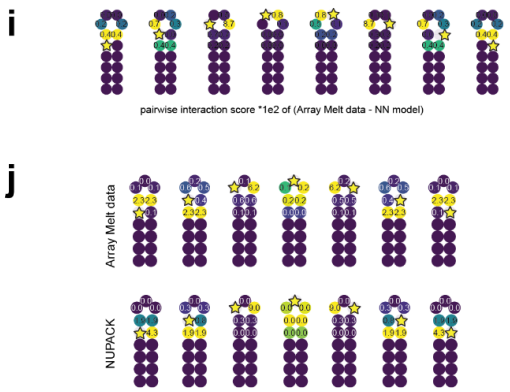


Figure S3

a. R^2 values between Array Melt data and nearest-neighbor model predictions for all single mismatches, and for each mismatch type.
b. Relationship of biological variance to total variance in Array Melt data. Each dot is a group of variants with identical mismatch type and adjacent nearest-neighbor base pairs as in **Fig. 3a**, colored by mismatched type. The biological/total variance ratio is shown on the right.
c. Examples of when the nearest-neighbor model succeeds or fails to capture total variance in data. Each subplot is calculated from a group of variants with identical mismatch type and adjacent base pairs. Green is the histogram of Array Melt measurements, and purple is what is expected from the nearest-neighbor model prediction with measurement uncertainty.
d. Comparison of nucleotide importance scores calculated from Array Melt data and nearest-neighbor model predictions for Watson-Crick and single mismatch variants.
e. Comparison of nucleotide importance scores calculated from either $\Delta\Delta G37$ or ΔG_{37}, and on Array Melt data or nearest-neighbor model predictions, for tetraloops and triloops.
f. Comparison of nucleotide importance scores calculated from Array Melt data and nearest-neighbor model predictions of bulge variants.
g. Pairwise interaction scores calculated from nearest-neighbor model predictions for Watson-Crick and single mismatch variants.
h. Pairwise interaction scores as a function of distance between the two nucleotides in the pair, for Watson-Crick variants or single mismatch variants. Those for the latter are splitted by the distance from the anchor point to the single mismatch.
i. Difference between pairwise interaction scores from Array Melt data and nearest-neighbor model prediction.
j. Comparison of nucleotide importance scores calculated from Array Melt data and nearest-neighbor model predictions of triloop variants.

SUPP FIGURE 4

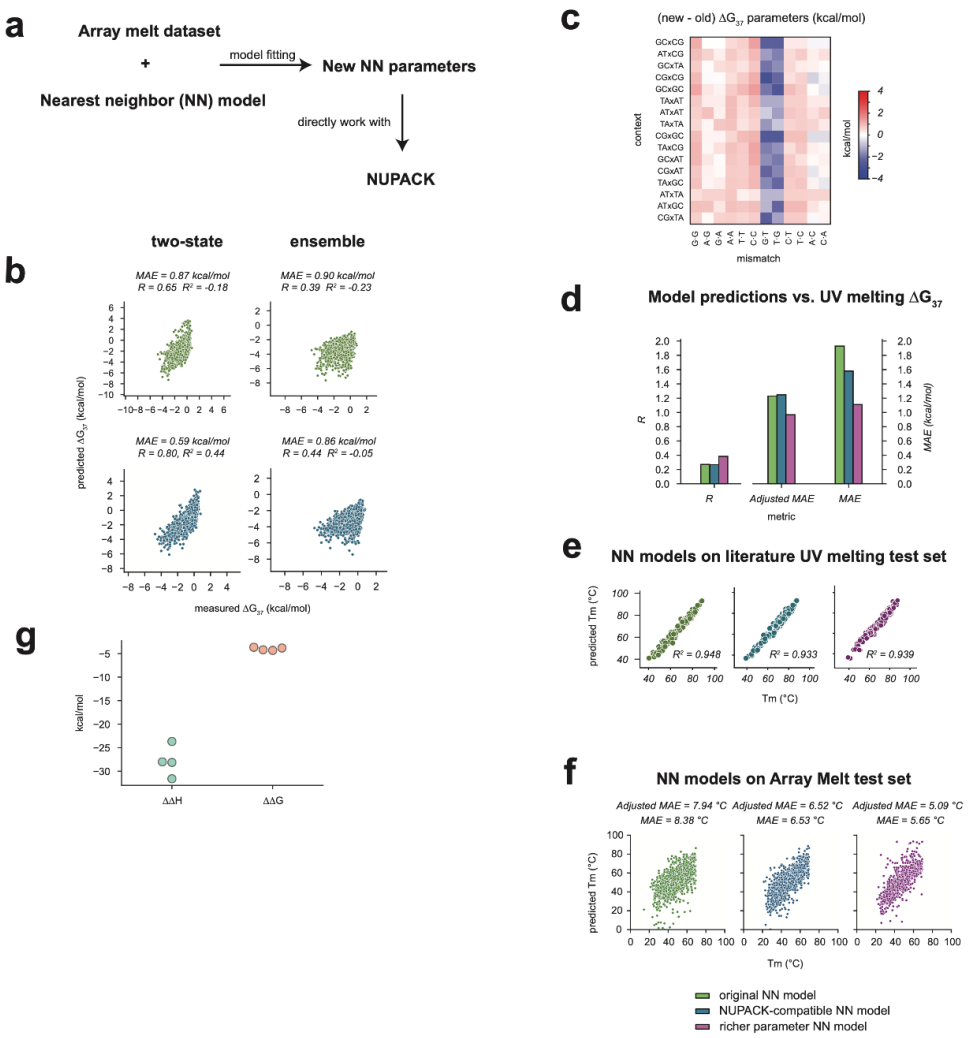
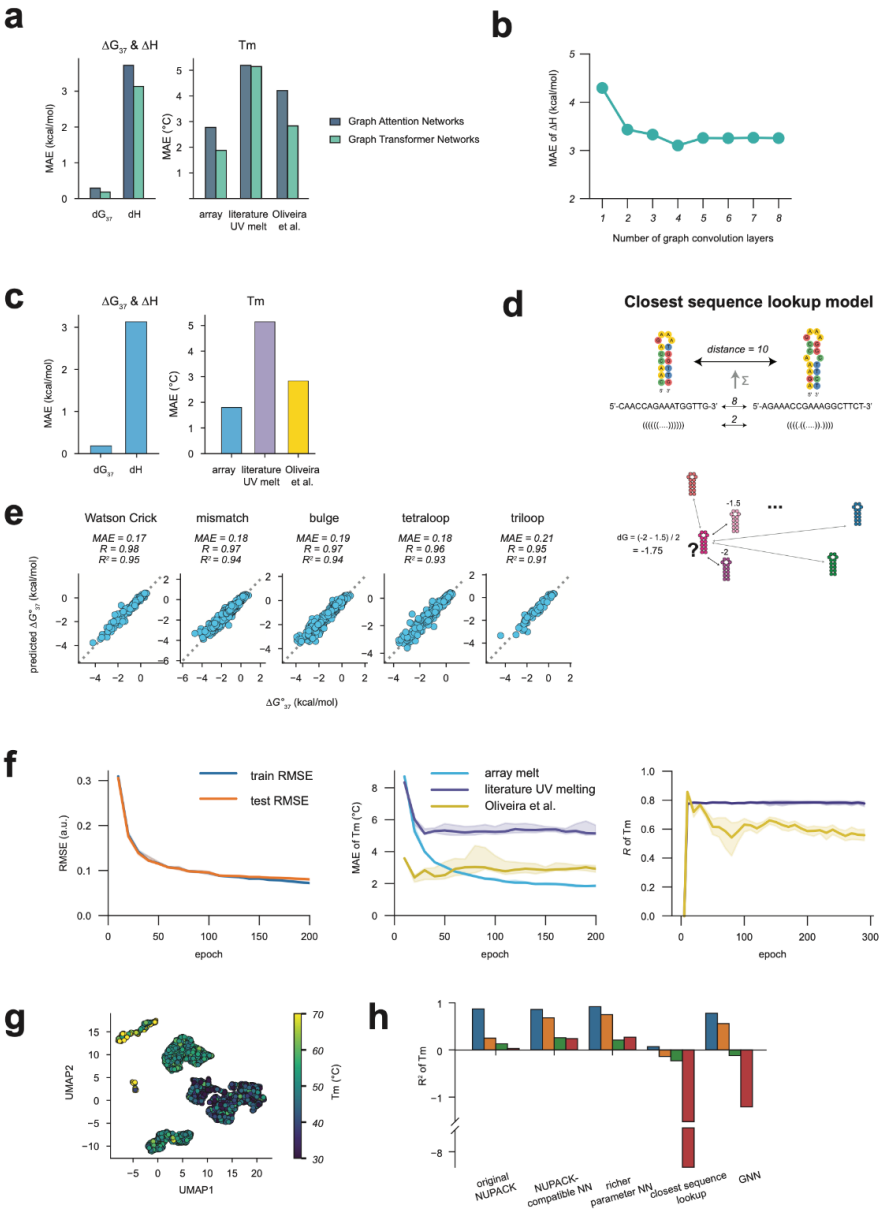


Figure S4

a. Schematics of the NUPACK-compatible nearest-neighbor model.
b. Difference between the new and original *dna04* single mismatch parameters.
c. Comparison of two-state and ensemble model predictions from NUPACK against Array Melt measurements.
d. Comparison of *dna04*, NUPACK-compatible, and richer parameterization nearest-neighbor model predictions with cloud lab UV melting measurements of randomly designed DNA hairpins.
e. Performance of the original, NUPACK-compatible, and richer parameter nearest-neighbor models on literature UV melting data used to derive the original nearest-neighbor model.
f. Performance of the original, NUPACK-compatible, and richer parameter nearest-neighbor model on Array Melt test set for Tm.
g. Fitted ΔΔH and ΔΔG values for 'hairpin_tetraloop' and 'hairpin_triloop'. Each dot is from one individual optimization run.

SUPP FIGURE 5



893

Figure S5

a. Mean absolute error of GNNs with either Graph Attention Network or Graph Transformer Network architecture at the graph convolutional layer.
b. Mean absolute error of enthalpy on Array Melt data as a function of the number of graph convolution layers in the GNN.
c. Mean absolute error of melting temperature on Array Melt data as a function of percentage of training data used to fit the GNN.
d. Schematic of the baseline closest sequence lookup model.
e. GNN predicted ΔG_{37} vs. held out Array Melt, splitted by type of variant.
f. The objective function of GNN (RMSE of normalized ΔH and Tm), mean absolute error of Tm, and Pearson's R of Tm across the GNN training process.
g. UMAP of learned embeddings of test data, colored by their predicted Tm.
h. Benchmark of all models on the orthogonal Oliveira et al. dataset as in **Fig. 5** but plotted by model. Bars for each model are for no mismatch, one mismatch, two mismatches and three mismatches, respectively.

Supplementary Notes

Adjusted mean absolute error

Nearest-neighbor models inherently have parameterization degeneracy, meaning that adding a constant to one group of parameters and subtracting from another could lead to the same prediction. This degeneracy can result in a prediction bias, but its toll on prediction accuracy could be mitigated by applying a constant offset to align the predictions with the true values. This constant offset preserves the relative $\Delta\Delta G$ values but not the absolute ΔG values. For instance, in the Array Melt test set, NUPACK with the original parameter set has a ΔG_{37} bias of -0.71 kcal/mol, and our NUPACK-compatible model parameter set has a bias of -0.42 kcal/mol. After subtracting this bias to align the predictions with the true values, the MAE of ΔG_{37} was reduced from 1.07 to 0.87 kcal/mol for the original NUPACK model, and from 0.67 to 0.59 kcal/mol for our NUPACK-compatible model (**Table 1**). This adjustment demonstrated that, with the presence of this bias, the accuracy of predicted $\Delta\Delta G$ values between different sequence-structure pairs is generally higher than that of the predicted ΔG values compared to the measured ground truth. Despite this model bias, we argue that the accuracy of nearest-neighbor models is acceptable in practice, particularly for applications that primarily rely on relative $\Delta\Delta G$ values rather than absolute ΔG values. This is because many applications, such as using dynamic programming to find the minimum free energy structure, depend on the relative $\Delta\Delta G$ values between different candidate secondary structures.

Given this context, we find it reasonable to report both the raw and the adjusted MAE model metrics. The adjusted metrics are calculated by first centering the prediction values to match the mean of the true values, thereby providing a more informative estimate of the relative errors for a given model.