Predicting the Influence of System Parameters on DNA Origami Self-Assembly Using a Monte Carlo Lattice Model

Anastasia Ershova & Andrew Sullivan

May 13, 2019

1 Introduction

In a biological context, deoxyribonucleic acid (DNA) serves as the genetic blueprint giving rise to the entire complement of cellular proteins, which mediate a myriad of functions ranging from intracellular communication and migration to enzymatic reactions involved in apoptosis or energy production and storage. This information is encoded within individual nucleotides which are assembled into long polymeric chains in the double helical structure of DNA. Each nucleotide is composed of a backbone of alternating deoxyribose sugars and phosphate groups along with one of four nitrogenous bases: adenine, thymine, guanine, and cytosine (A, T, G, and C). Each base forms specific hydrogen bonds with its complement (A and T or C and G) to assemble the antiparallel double-stranded molecule. In living cells, complementary base pairing allows either strand of DNA to be accurately replicated, repaired, and translated into specific sequences of amino acids.

In recent years, synthetic biologists and bioengineers have recognized the potential for DNA as a structural and functionally active molecule in addition to its canonical role in storing cellular information. The regular repeating structure of the double helix and the precise and simple nature of complementary Watson-Crick base pairing have been shown to be well suited for the self-assembly of complex macromolecular constructs [1]. The field of DNA nanotechnology was revolutionized with the introduction of the DNA origami paradigm, in which specific DNA sequences are designed in the form of a scaffold strand and multiple staple strands [2]. The staple strands bind to specific sites on the scaffold or other staples, inducing the formation of three-dimensional structures upon mixing of the components [3] (Figure 1A).

Since its introduction in 2006, DNA origami has found applications in a wide variety of fields, including molecular robotics, functional studies of proteins, assembly of nanoparticle structures, drug delivery, and biomarker detection [4, 5, 6]. The presence of a scaffold strand sets apart the origami method from previous schemes of designing DNA nanostructures, which instead relied only on short oligonucleotides and thus made the system highly sensitive to stochiometry. Still, a number of system parameters mediate the strength of these interactions, the speed at which the origami structure assembles, and the accuracy of the final product, including temperature, concentration of cations such as sodium or magnesium, and staple concentration. While the influence of such system parameters may be uncovered experimentally, molecular modelling provides a more efficient path to identifying proper system parameters for optimal origami assembly.

Like most polymeric molecules, DNA is too large to be modeled at an atomistic level. An individual nucleoside – the fundamental unit of the DNA backbone lacking any of the four bases – contains 25 atoms by itself. Adding on either of the smaller bases (cytosine or thymine) increases this number to almost 40. Clearly, even a short sequence of nucleotides would be computationally expensive to model using density functional theory (DFT) or molecular dynamics (MD). As a result, coarse graining techniques have been the primary means of computational investigation of DNA nanostructures. Within the context of these coarse-grained methods, there still exists a large range of complexity. The oxDNA method [7, 8, 9], which considers the interaction between chains of rigid bodies, in which each body is a single nucleotide, allows for unprecedented interrogation of the self-assembly characteristics of DNA nanostructures. Even so, for small DNA sequences the computational time is still prohibitively long. In [9], each simulation run for a 384 base pair-long structure required several months on a GPU. While this may be suitable for certain applications requiring high precision in small structures, many DNA origami designs involve sequences at least an order of magnitude larger (for example, see the initial structures created in [2], which are 7,000 nucleotides long).

Conversely, low-resolution models are commonly designed by extending the nearest-neighbor model developed by SantaLucia Jr. and Hicks in 1998 [10]. In the SantaLucia model, the driving factor for geometric configurations is the free energy of hybridization between complementary sequences, which can be parameterized into nearest-neighbor pairs such that each sequence has an associated hybridization free energy that is the sum of the hybridization of adjacent free-energy pairs in the nucleotide sequence of the strand. While this method has been incorporated into Markov chain and chemical equilibrium schemes to model self-assembly, consideration of only free energy differences lacks geometrical considerations essential to proper assembly and does not consider interactions between non-complementary binding sequences, despite a favorable computational cost.

In this work, we employ a novel strategy for modeling DNA self-assembly based on a Monte Carlo lattice model introduced by Cumberworth et al [11]. This method has been shown to give reasonably accurate results for DNA origami systems comparable to those obtained by oxDNA but with a computational cost more closely aligned to the SantaLucia model. While initial investigations have validated the results against the same 384-base pair model tested with oxDNA, its applicability to larger structures, such as those initially designed by Rothemund in 2006 (Figure 1B and 1C), has yet to be confirmed. Therefore, we utilize this model to simulate the self-assembly of larger DNA origami structures and investigate the influence of temperature, cation concentration, and staple concentration on the order parameters of the system: the number of bound domain pairs, misbound domain pairs, bound or misbound staples, and stacked pairs.

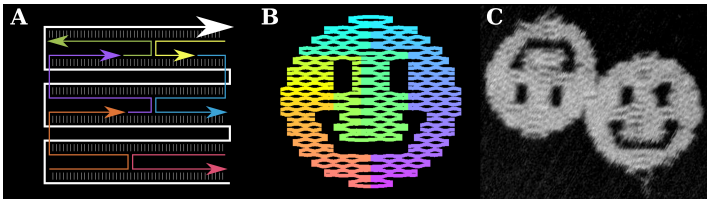

Figure 1: (A) Schematic of a simple DNA origami structure with a single long scaffold strand (white) and several shorter staple strands (coloured), the binding of which folds the scaffold strand to impose a defined structure. (B and C) Smiley face DNA origami (reproduced from Rothemund (2006) [2]). (B) Schematic illustrating the bending of helices at and away from crossovers, with the colour corresponding to the base-pair index along the folding path (red = 1, purple = 7,000). (C) Atomic force microscopy image of the folded structure.

2 Methods

Model overview

Rather than considering each nucleotide individually, the Cumberworth model considers binding domains as the fundamental unit of assembly. In the context of DNA origami, a binding domain is a sequence of fixed length which is fully bound to a complementary sequence in the assembled state. These domains are represented as particles on a lattice, a representation justified by the constraints imposed by the double helix and used to substantially reduce the configuration space. Each lattice site is defined by its occupancy – zero for unoccupied sites, one for an occupied but unbound site, and two for a site which contains a bound or misbound domain pair. Misbound states are those in which the occupying binding domains are not perfectly complementary. Hybridization energies are introduced to account for binding energy and hybridization entropy when two domains occupy a single lattice site and are computed using the SantaLucia model [12]. In the case of misbound pairs, the energy is computed for the longest complementary sequence within the binding domain. The total free energy associated with hybridization is the sum of four individual free energies, each composed of an enthalpic and entropic contribution. Three of these terms account for initiation, symmetry, and the identity of the final pair. The fourth is sequence-dependent and is the sum of each nearest-neighbor pair free energy, such that an N base pair-long sequence will have N nearest-neighbor terms in its free energy of the form ΔG_{1-2}, ΔG_{2-3}, ..., $\Delta G_{(N-2)-(N-1)}$, $\Delta G_{(N-1)-N}$, where each entry $1, 2, ..., N$ is one of the four bases, A, T, C, or G.

Each binding domain is associated with two unit vectors that define the state of that domain: orientation vectors point orthogonally from the helical axis to the strand position at the end of the binding domain to account for the constraints imposed on possible locations for strand crossovers by the double helical geometry, and next-binding-domain vectors point from the current binding domain to the next one along the chain. Stacking interactions are also introduced for contiguous domains in the same helix that are in bound states resulting from interactions between aromatic rings in consecutive base pairs.

This formulation imposes a number of required constraints on the assembly of a DNA nanostructure, discussed further in [11]. These constraints relate to the definition of a helical axis and the identification of kinks in the structure when adjacent domain pairs do not satisfy certain criteria. It is at these sites where a crossover, a site where one strand of a double helix switches its binding partner to a different complementary strand, may occur. Further rules are introduced to restrict certain kink configurations based on steric hindrance. A final constraint is introduced which forces two helices with at least two crossovers between them to be parallel.

Monte Carlo methods

With the above model, it is possible to fully describe the configuration of a DNA origami assembly based on the binding states of its staple and scaffold strands. In order to sample different configurations, a Monte Carlo scheme is employed by describing a set of moves and their acceptance probabilities. It is assumed that staple concentrations are large enough relative to the scaffold to be constant and that staple-staple binding is not important, except near the scaffold due to a local elevation in staple concentration. The grand canonical (μVT) ensemble is chosen to sample states only with bound staples. Replica-exchange Monte Carlo (REMC) is used to improve sampling efficiency by varying simulation parameters and allowing for exchange across different replicas of the system at different conditions.

Many of the allowed MC move types involve regrowth of the polymer chain, in which the binding domains in a selected set are unassigned from their current lattice positions and then reassigned in accordance with the move type acceptance criteria. Three growth processes are considered in the model: symmetric, configurational bias, and recoil growth. The trial configuration and acceptance probabilities are detailed in the Cumberworth paper [11]. Briefly, symmetric move variants randomly choose position difference vectors and orientation vectors from a uniform distribution of all possible unit vectors, and thus the acceptance probability is identical to that for the classical Metropolis algorithm. In configurational bias methods, the trial configuration is no longer chosen uniformly, but is instead biased by the resultant energy change. The corresponding acceptance criterion is given based on the Rosenbluth weights of each configuration. Recoil growth allows for previously set configurations in the current growth sequence to be unassigned via "recoil." For each binding domain, the probability of a new configuration being "open" is defined using the Metropolis criterion. If acceptance fails, new configurations can be tested using the same criterion up to a total of k_{max} times. If all tests should fail, the growth recoils to the previous binding domain and the process repeats up to a maximum of l_{max} recoil events.

The above variants are used in the algorithm to define four classes of Monte Carlo moves that may be performed. In most cases, the defined acceptance criteria are modified slightly to account for changes in number of bound domains. Like the previous definitions, we refer to [11] for more detail. The simplest is an orientation vector move, in which the symmetric variant is used for a selected binding domain in the system. The move is always accepted if the domain is unbound. Otherwise, the relevant energy change includes that of the selected binding domain and its partner. Staple regrowth begins by selecting a staple with uniform probability. One of the bound

domains in the staple is selected and the remainder of the staple is grown out using the configurational bias method. Staple exchange moves include both addition and removal of staples from the scaffold, and the staple type is selected from a uniform distribution. The symmetric scheme is used for acceptance. For insertions, a lattice site is selected from all sites with non-zero occupancies (i.e. within the "system volume"). From the chosen staple, one of the sites is selected. If the occupancy of the selected site is 1, the only acceptable configuration is chosen, i.e. the one in which the sum of the orientation vectors is zero. The staple is grown from this inserted domain. For deletion, the case is similar. The move is rejected if the chosen staple is not a connector, i.e. a staple which will cause the complete removal of other staples if it is removed. Otherwise, an acceptance probability is defined as for other moves.

The final move type, scaffold regrowth, is the most complex, as it involves the regrowth of entire segments of the scaffold strand. Variants of configurational bias and recoil growth methods are introduced which preserve the number of bound domains in the system, so changes in binding state are restricted only to the previous two move types. Scaffold regrowth moves involve the selection of endpoints between which regrowth is performed. Within each move, single or multiple segments may be selected. In either case, a length of the segment is chosen from a uniform distribution, and a random binding domain is selected. For the latter case, maximal lengths of the individual domains are chosen as well. Domains are grown out in a randomly chosen direction from the selected domain, continuing until the maximum length is reached for the entire move or for the segment itself. In the non-contiguous case, segments are then added by jumping across staple strands within the chosen segment and continuing from the domain on another binding site of the staple. Regrowth is performed in the same order that the domains were selected. Acceptance probabilities are scaled by indicator functions to account for formation of new pairings and changes in the number of ideal random walks between the two endpoints as a result of the change in configuration.

Structure generation and simulation

Simulations were performed using the latticeDNAOrigami software package presented in [11], using the most recent version of the code obtained from Github. Considering the relatively recent release date of the software package, bug fixes needed to be performed on the associated Makefile to allow for successful installation of the software, and pull requests have been submitted to share these fixes.

The structure used for the simulations was the smiley face DNA origami from Rothemund (2006) (Figure 1B and 1C) [2]. Output staple strands were generated to be of length 15, 16, 31 or 32 nucleotides to ensure compatibility with the model by the exclusion, truncation, or division of original staple strand sequences. Every binding domain was assigned a unique number to create linear scaffold positions and orientations, which were subsequently cyclized to create the unbound structure for the smiley face.

Custom Python wrappers were written for generating configuration files, running simulations, and analysing results. Default parameter values for the simulations are listed in Appendix A1. Simulations performed were either serial tempering or annealing simulations, and were executed in part on a local machine and in part on the Google cloud computing platform (GCP), scanning across temperature, cation concentration and staple concentration. Model acceptance probabilities and convergence with regards to stacking energy were explored.

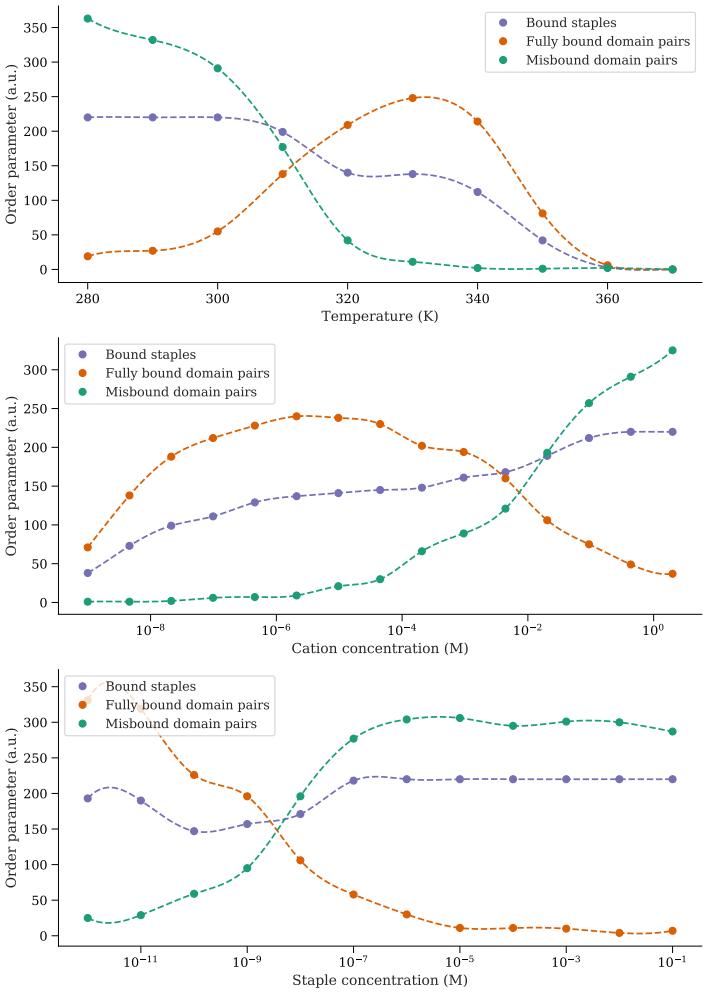

Figure 2: Final order parameters as a function of system temperature (top), cation concentration (middle) and staple concentration (bottom) for the smiley face DNA origami. Based on serial tempering simulations that were run for 2 hours at each condition on a local machine.

3 Results

Optimisation of temperature, cation concentration and staple concentration

The three parameters most commonly varied in order to optimise folding of an origami structure are temperature, cation concentration, and staple concentration. Thus, serial tempering simulations were performed to model various order parameters of self-assembly – the number of staple strands bound at a given condition, the number of scaffold domains with a staple domain properly bound, and the number that have an incorrect staple bound (among others, as seen in Appendix A2). Simulations were run for 2 hours at each condition on a local machine, using a temperature of 300 K, 10^{-7} M staple concentration and 0.5 M cation concentration as the default values. The full simulation results are presented in Appendix

A2 (Figures 7, 8 and 9); at certain conditions, convergence was not attained for many of the parameters explored. Thus, in the summary results in Figure 2 the points represent the endpoint of the simulation. Based on these curves, a temperature of approximately 330 K, a cation concentration of 10^{-5} M, and a staple concentration of 10^{-9} M might be optimal for the folding of this structure, as these are the conditions at which the maximum number of fully bound domain pairs can be seen (peak of the orange curves in Figure 2).

This parameter scan was used to subsequently determine the default parameter values for a similar optimisation performed for 3 hours at each condition on GCP. From here, we get a shift in the optimal parameter values, as the optimal temperature now appears to be 320 K, cation concentration closer to 0.5 M, and staple concentration 10^{-6} M. Thus, it is challenging to conclude based on these results what the optimal parameter values for folding the smiley face would actually be. However, even in this case the simulations were not fully converged (see Appendix A2 Figures 10, 11 and 12 for full simulation results).

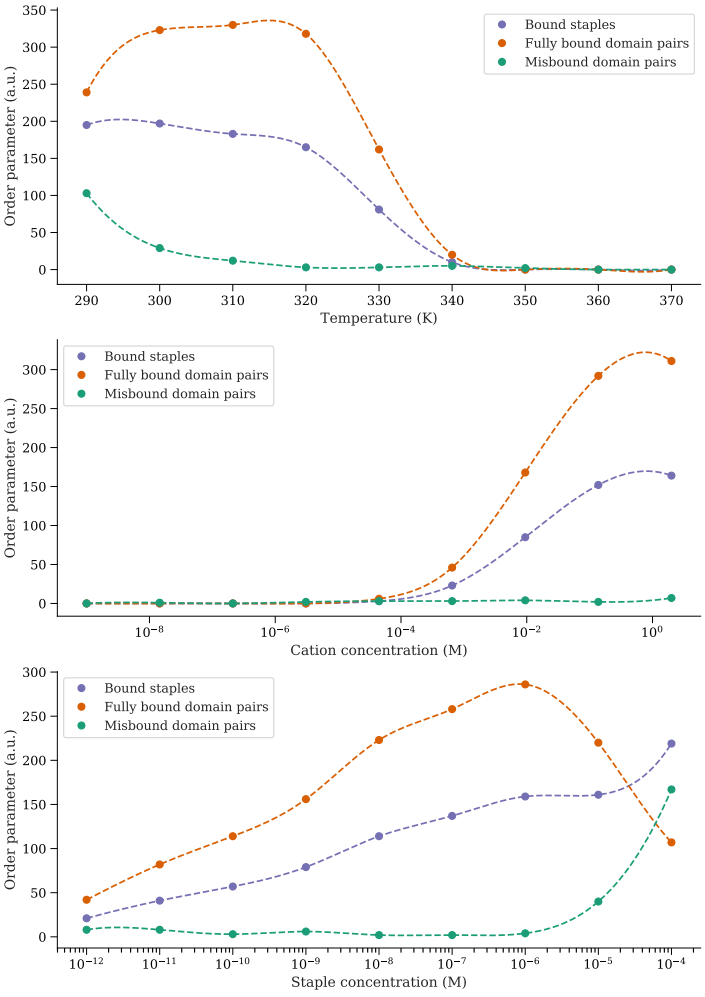

Figure 3: Final order parameters as a function of system temperature (top), cation concentration (middle) and staple concentration (bottom) for the smiley face DNA origami. Based on serial tempering simulations that were run for 3 hours at each condition on GCP, using optimised parameters from Figure 2 as defaults.

In order to address this issue, serial tempering simulations for a single condition using the optimal values from Figure 3 were run for 24 hours on GCP. A representative run from three total runs is presented in Figure 4 (see Appendix A2 Figures 13 and 14 for the other two runs). Here, it is clear that convergence was not reached even after this long runtime. Given that longer runtimes were not feasible in this project and would generally prevent the practical use of the model in a research setting, this direction was not pursued any further.

Of note is that in a properly folded structure, the number of unique bound staples should equal the number of bound staples. As can be seen in Figure 4, in these simulations the number of unique bound staples asymptotes at a much lower value than the number of bound staples. This suggests that this Monte Carlo lattice model in its current implementation is not capable of correctly capturing the folding of a larger origami structure, though this result may be due to the manipulations performed on the staples to make them compatible with the model, and not intrinsic to the model itself.

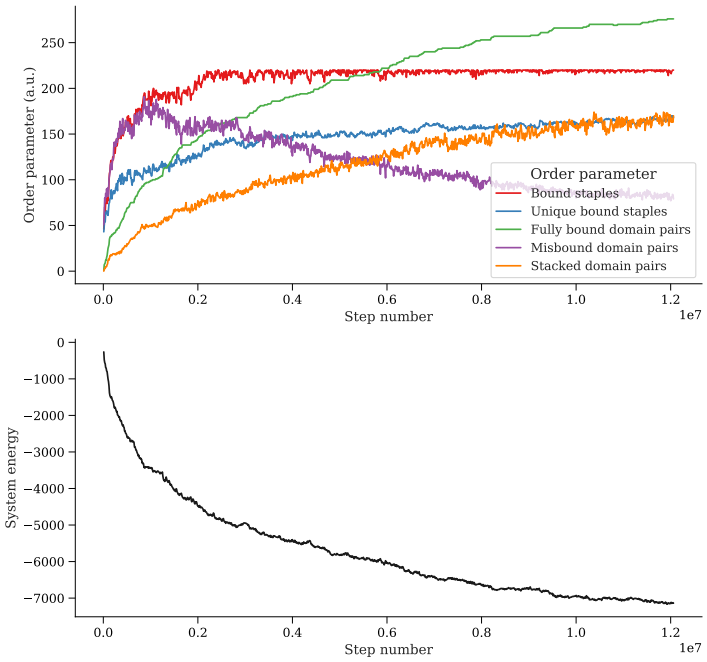

Figure 4: Overall results for one of three serial tempering simulations performed at the optimised parameter values from Figure 3 for 24 hours each on GCP.

Exploration of other model parameters

Scans across different values for stacking energy were performed, as this was the only remaining system input parameter not tested above. Stacking energy represents the total energy of the base stacking interactions of adjacent pairs of nucleotides. However, instead of calculating this *de novo* for each configuration, the current model implementation takes a single value for stacking energy as an input to use throughout the simulation. Given that this value has no physical meaning, it was reassuring to find that it did not seem to significantly alter the order parameter outputs, although it did change total system energy in simulations performed at 300 K, 10^{-7} M staple concentration and 0.5 M cation concentration for 2 hours at each condition (Figure 5). Testing even higher values of stacking energy caused the simulation to crash, suggesting that stacking energy does not affect results across the full range of possible inputs. As above, the simulation end-points

were considered in the analysis due to lack of convergence (see Appendix A2 Figure 15 for full simulation results).

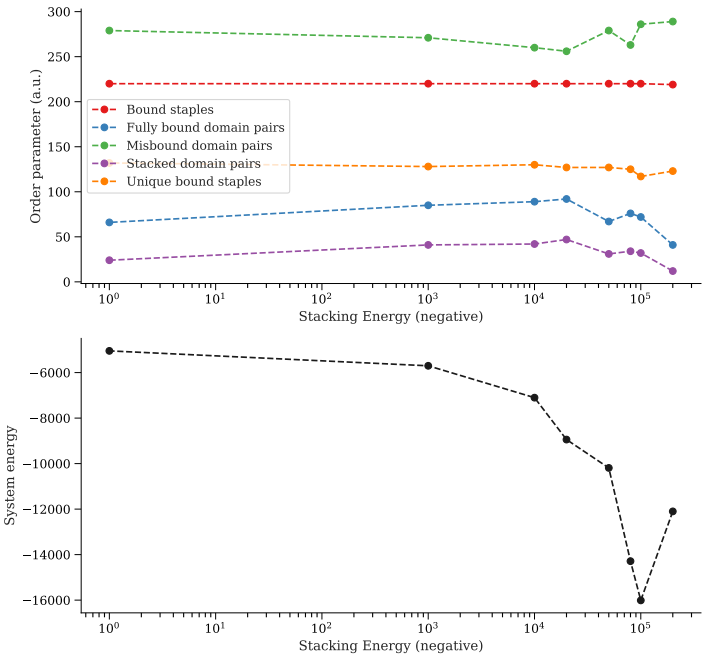

Figure 5: Order parameters and system energy as a function of stacking energy. Simulations were performed for 2 hours each on GCP, at the optimised parameter values from Figure 3.

Movetype frequencies and other model parameters were not altered as the overall acceptance probabilities seemed to fall within the desired range – for example, for the three simulations in Figure 4 the total acceptance probability was 0.54 ± 0.01.

Performance of annealing simulations

Most commonly, origami growth is performed with a temperature ramp from high temperatures (typically up to 80 °C) down to room temperature [13]. This is because the ideal folding temperature is not known, and thus by holding at various temperatures for a shorter period of time, the hope is that sufficient folding will occur during the permissive temperature regime. Running simulations as above to identify the ideal folding temperature would help mitigate this need, and thereby improve the yield and accuracy of folded structures. To explore this possibility, we simulated folding with a temperature ramp from 350 K to 290 K using a 24 hour annealing simulation and the optimised parameters from Figure 3.

It can be seen from the results in Figure 6 (and Appendix A2 Figure 16) that the simulation levels off at <250 fully bound domain pairs, as at the lower temperatures towards the end of the simulation no further folding can occur. This is in contrast to Figure 4, where the number of fully bound domain pairs continues to increase throughout the simulation, and despite a lower total number of simulation steps, attains higher (>250) numbers of fully bound domain pairs. These results validate the benefit of using constant temperature conditions as opposed to temperature ramps for the folding of DNA origami.

4 Conclusion

In this project, we attempted to determine whether the Cumberworth model for DNA origami self-assembly is capable of scaling to larger and thus more experimentally relevant origami structures, compared to the simplified structures tested in the original paper. While we were unable to attain full convergence within feasible runtimes, it is possible that the qualitivative predictions of the simulations would still be valid, although the dependence on input parameters may be a limitation (Figures 2 and 3).

In Figure 4 we found that the model does not appear capable of generating the fully folded structure. While it is possible that even longer runtimes would solve this problem, the convergence of bound staple and unique bound staple values suggests otherwise. A likely explanation, however, is that the pre-processing carried out on the original staple strands in order to make them compatible with the model implementation could have affected the foldability of the structure itself, meaning that even in infinitely long simulations we would still not be able to recapitulate proper folding. This is purely a limitation of the current model implementation, and could in principle be solved by allowing staples of different lengths and with more than two domains.

Finally, we showed that folding at a constant temperature is likely to be more effective than the temperature ramps usually used experimentally (Figure 6). While this model is not able to predict the real time it would take to fold a structure, it has the potential to generate the optimal conditions for folding, which would not be feasible to screen for experimentally. Ultimately, these results would need to be validated experimentally before any conclusions as to the suitability of the model can be made. If the trends do indeed hold, this model can become a powerful tool available to experimentalists to streamline the process of DNA origami design and folding, while also improving its accuracy and helping understand self-assembly from a more theoretical perspective.

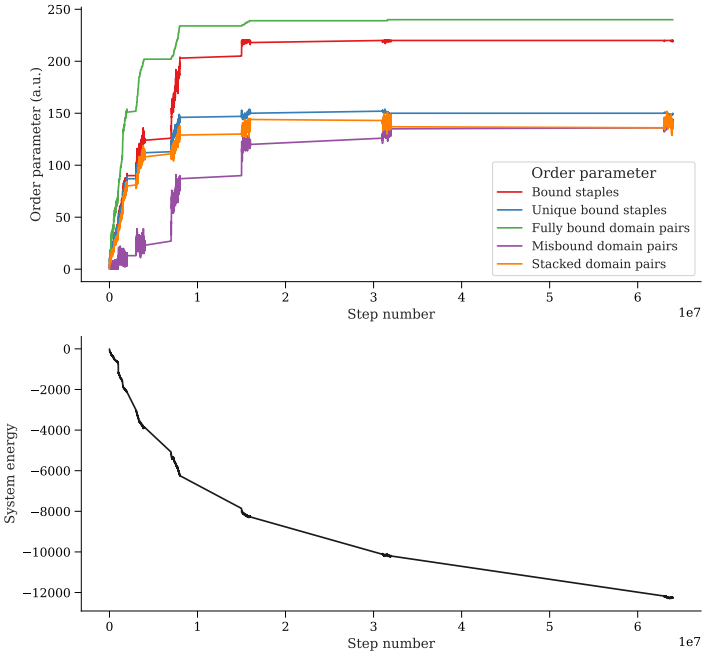

Figure 6: Annealing simulations for a ramp from 350 K to 290 K, at a staple concentration of 10^{-6} M and cation concentration of 0.5 M. Representative result from two runs of the simulation.

References

[1] N. C. Seeman, "Nucleic acid junctions and lattices," *Journal of Theoretical Biology*, vol. 99, pp. 237–247, Nov. 1982.

[2] P. W. K. Rothemund, "Folding DNA to create nanoscale shapes and patterns," *Nature*, vol. 440, pp. 297–302, Mar. 2006.

[3] S. M. Douglas, H. Dietz, T. Liedl, B. Hgberg, F. Graf, and W. M. Shih, "Self-assembly of DNA into nanoscale three-dimensional shapes," *Nature*, vol. 459, pp. 414–418, May 2009.

[4] F. Hong, F. Zhang, Y. Liu, and H. Yan, "DNA origami: Scaffolds for creating higher order structures," *Chemical Reviews*, vol. 117, no. 20, pp. 12584–12640, 2017.

[5] P. Wang, T. A. Meyer, V. Pan, P. K. Dutta, and Y. Ke, "The beauty and utility of DNA origami," *Chem*, vol. 2, no. 3, pp. 359 – 382, 2017.

[6] Q. Zhang, Q. Jiang, N. Li, L. Dai, Q. Liu, L. Song, J. Wang, Y. Li, J. Tian, B. Ding, and Y. Du, "DNA origami as an in vivo drug delivery vehicle for cancer therapy," *ACS Nano*, vol. 8, no. 7, pp. 6633–6643, 2014.

[7] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, "Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model," *The Journal of Chemical Physics*, vol. 134, no. 8, p. 085101, 2011.

[8] B. E. K. Snodin, F. Randisi, M. Mosayebi, P. ulc, J. S. Schreck, F. Romano, T. E. Ouldridge, R. Tsukanov, E. Nir, A. A. Louis, and J. P. K. Doye, "Introducing improved structural properties and salt dependence into a coarse-grained model of DNA," *The Journal of Chemical Physics*, vol. 142, p. 234901, June 2015.

[9] B. E. K. Snodin, F. Romano, L. Rovigatti, T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, "Direct simulation of the self-assembly of a small DNA origami," *ACS Nano*, vol. 10, no. 2, pp. 1724–1737, 2016.

[10] J. SantaLucia, "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," *Proceedings of the National Academy of Sciences*, vol. 95, pp. 1460–1465, Feb. 1998.

[11] A. Cumberworth, A. Reinhardt, and D. Frenkel, "Lattice models and Monte Carlo methods for simulating DNA origami self-assembly," *The Journal of Chemical Physics*, vol. 149, p. 234905, Dec. 2018.

[12] J. SantaLucia and D. Hicks, "The thermodynamics of DNA structural motifs," *Annual Review of Biophysics and Biomolecular Structure*, vol. 33, no. 1, pp. 415–440, 2004.

[13] K. F. Wagenbauer, F. A. S. Engelhardt, E. Stahl, V. K. Hechtl, P. Stmmer, F. Seebacher, L. Meregalli, P. Ketterer, T. Gerling, and H. Dietz, "How We Make DNA Origami," *ChemBioChem*, vol. 18, no. 19, pp. 1873–1885, 2017.

5 Appendix

A1 Default simulation parameter values

temp=300
hybridization_pot=NearestNeighbour
binding_pot=ConKinkLinearFlexible
misbinding_pot=Opposing
stacking_pot=Constant
staple_M=1e-7
cation_M=0.5
temp_for_staple_u=300
staple_u_mult=1
binding_h=0
binding_s=0
misbinding_h=0
misbinding_s=0
stacking_ene=-1000
max_total_staples=220
max_type_staples=220
domain_update_biases_present=false
simulation_type=constant_temp
max_duration=7200

Movetype frequencies:
Orientation rotation=2/6
Met staple rotation=1/6
CB staple regrowth=1/6
Contiguous CTRG scaffold regrowth = 1/6
Non-contiguous CTRG scaffold regrowth = 1/6

A2 Full simulation results

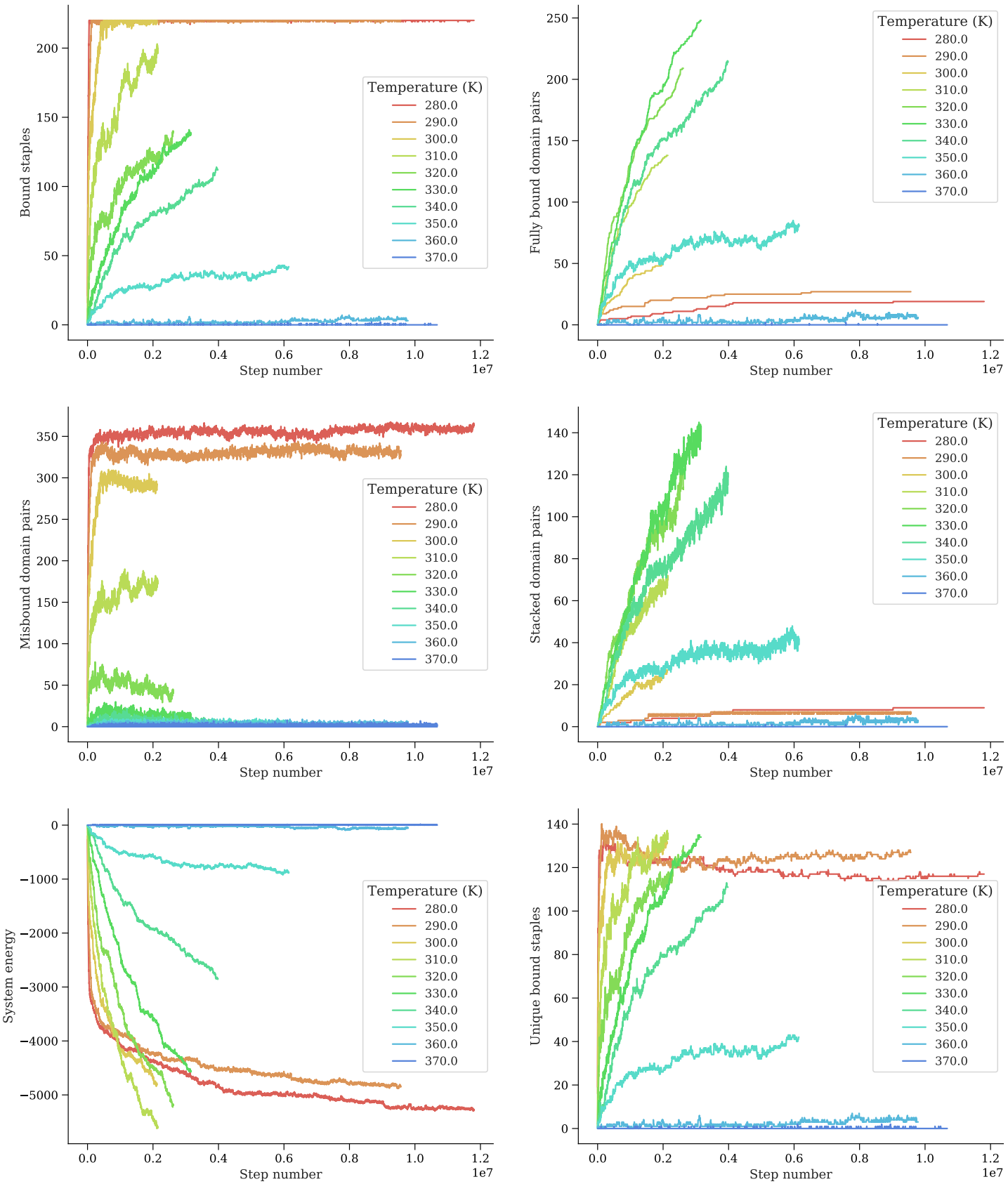

Figure 7: Full simulation results for temperature optimisation in Figure 2.

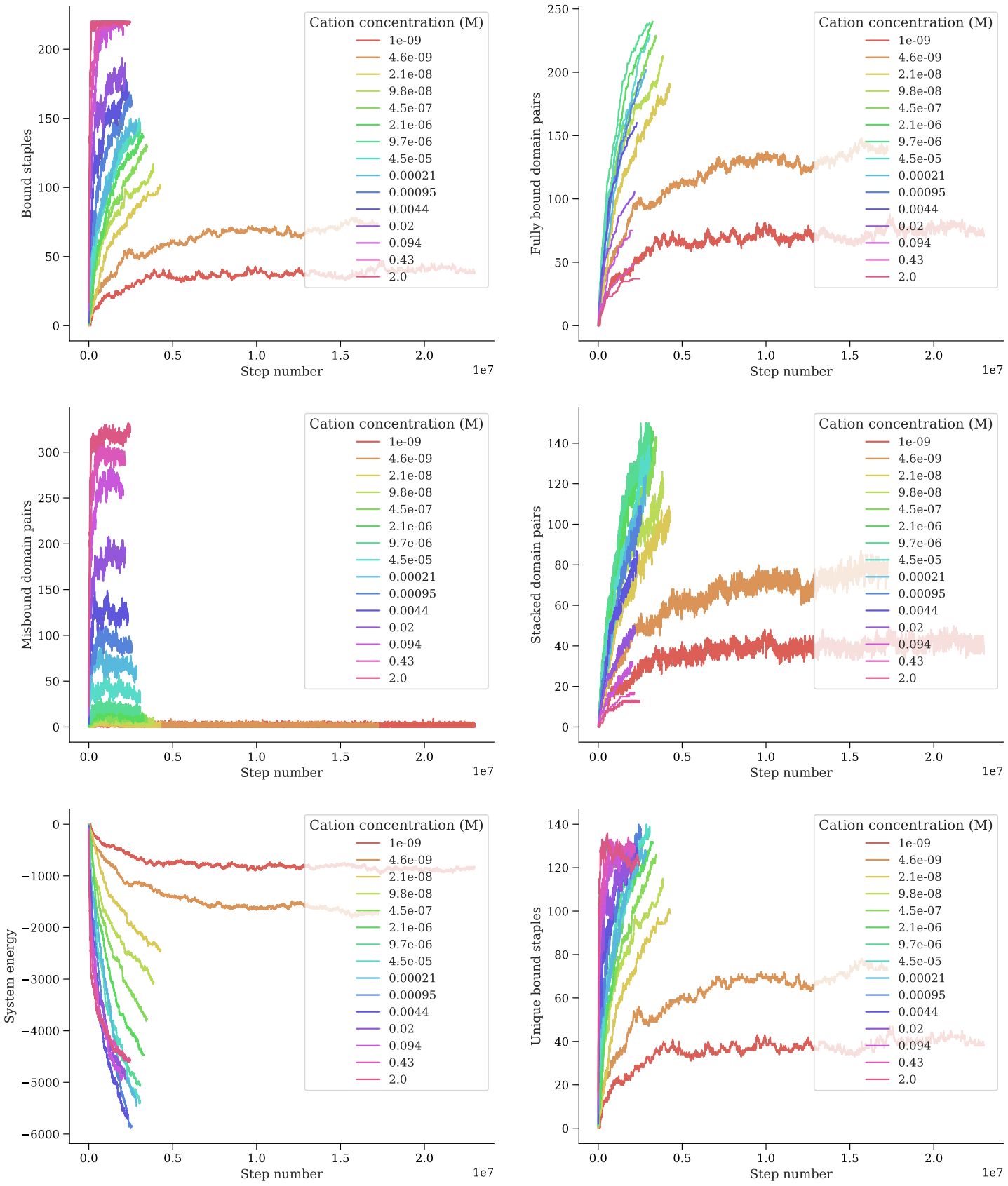

Figure 8: Full simulation results for cation optimisation in Figure 2.

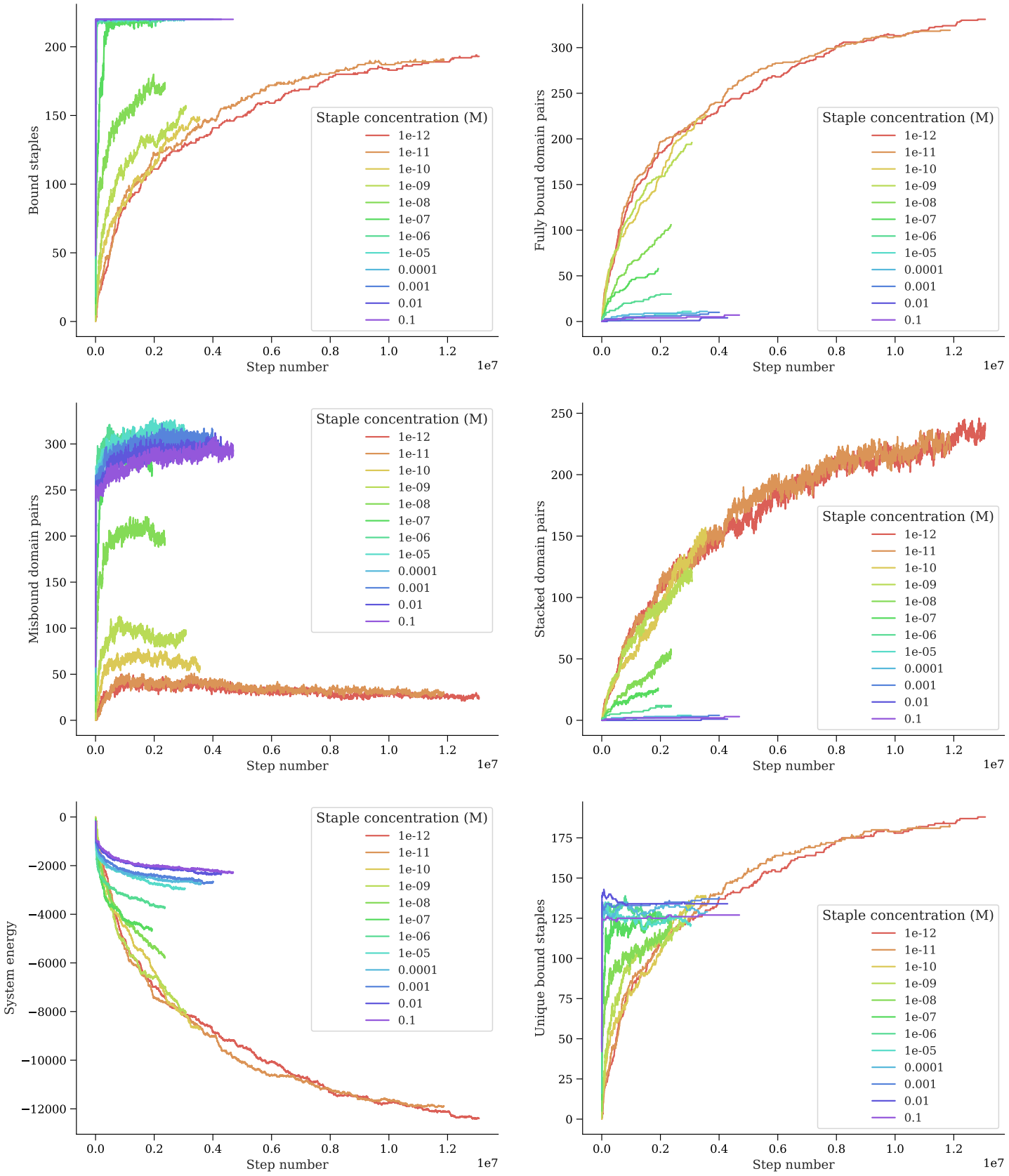

Figure 9: Full simulation results for staple optimisation in Figure 2.

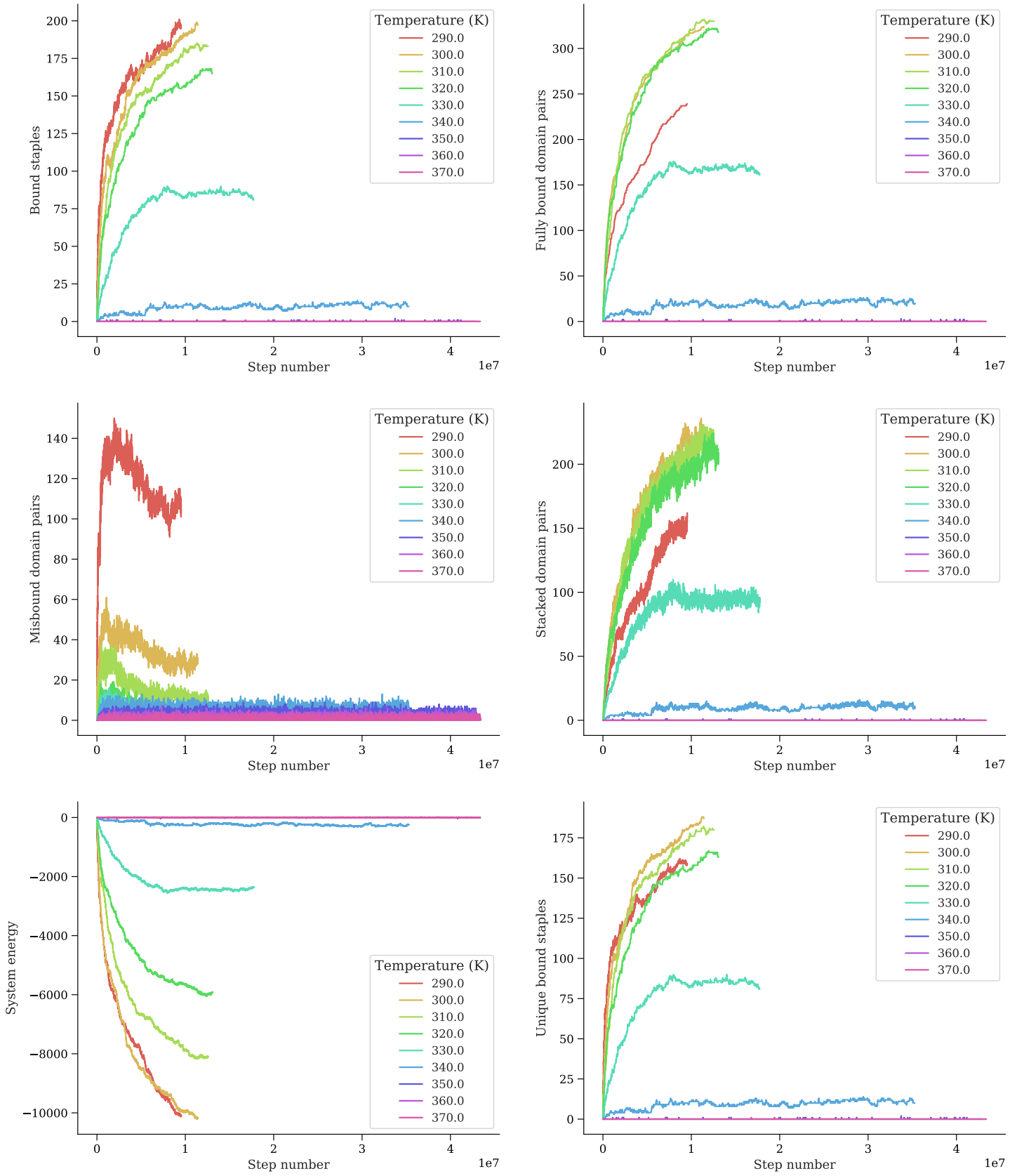

Figure 10: Full simulation results for temperature optimisation in Figure 3.

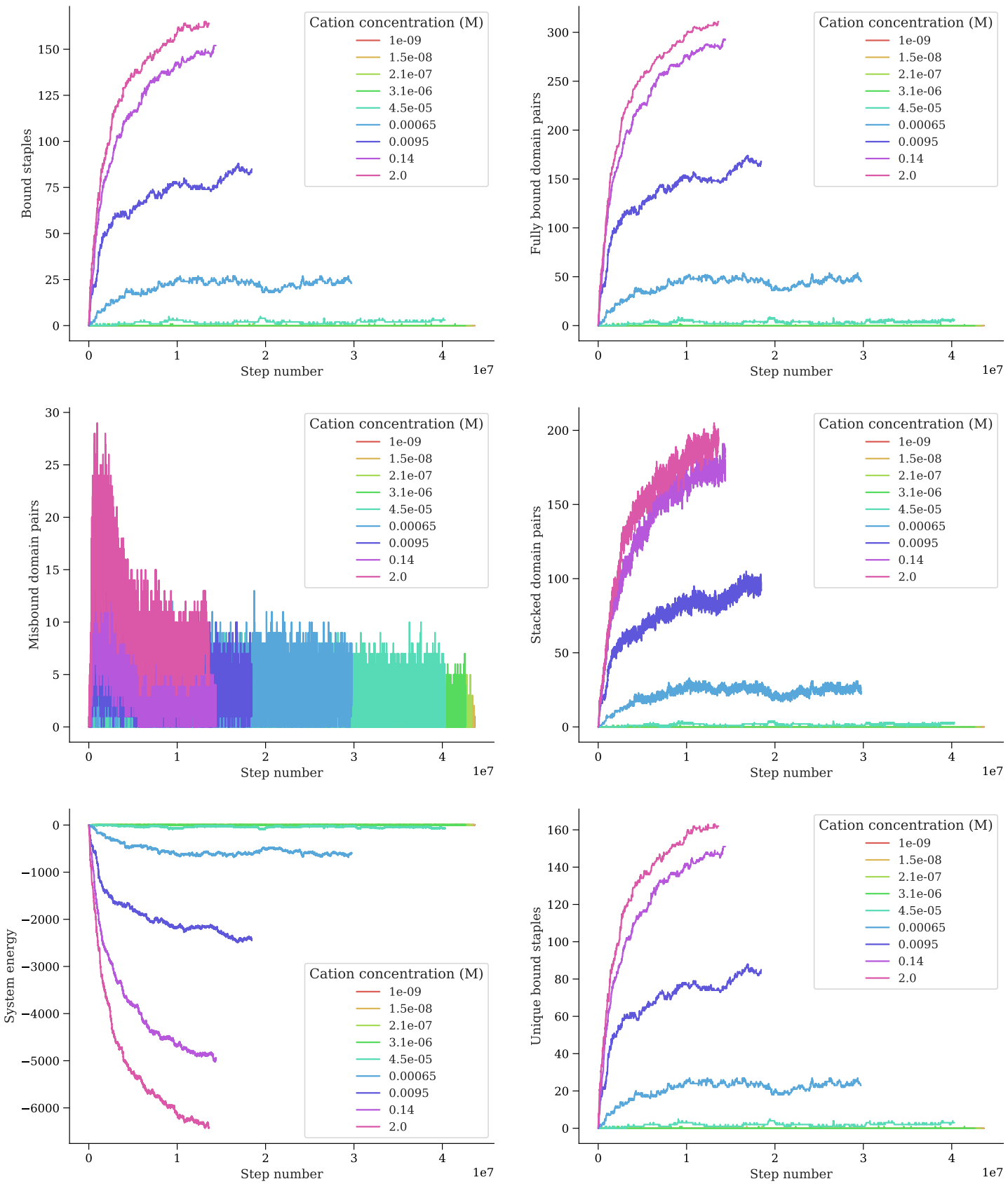

Figure 11: Full simulation results for cation optimisation in Figure 3.

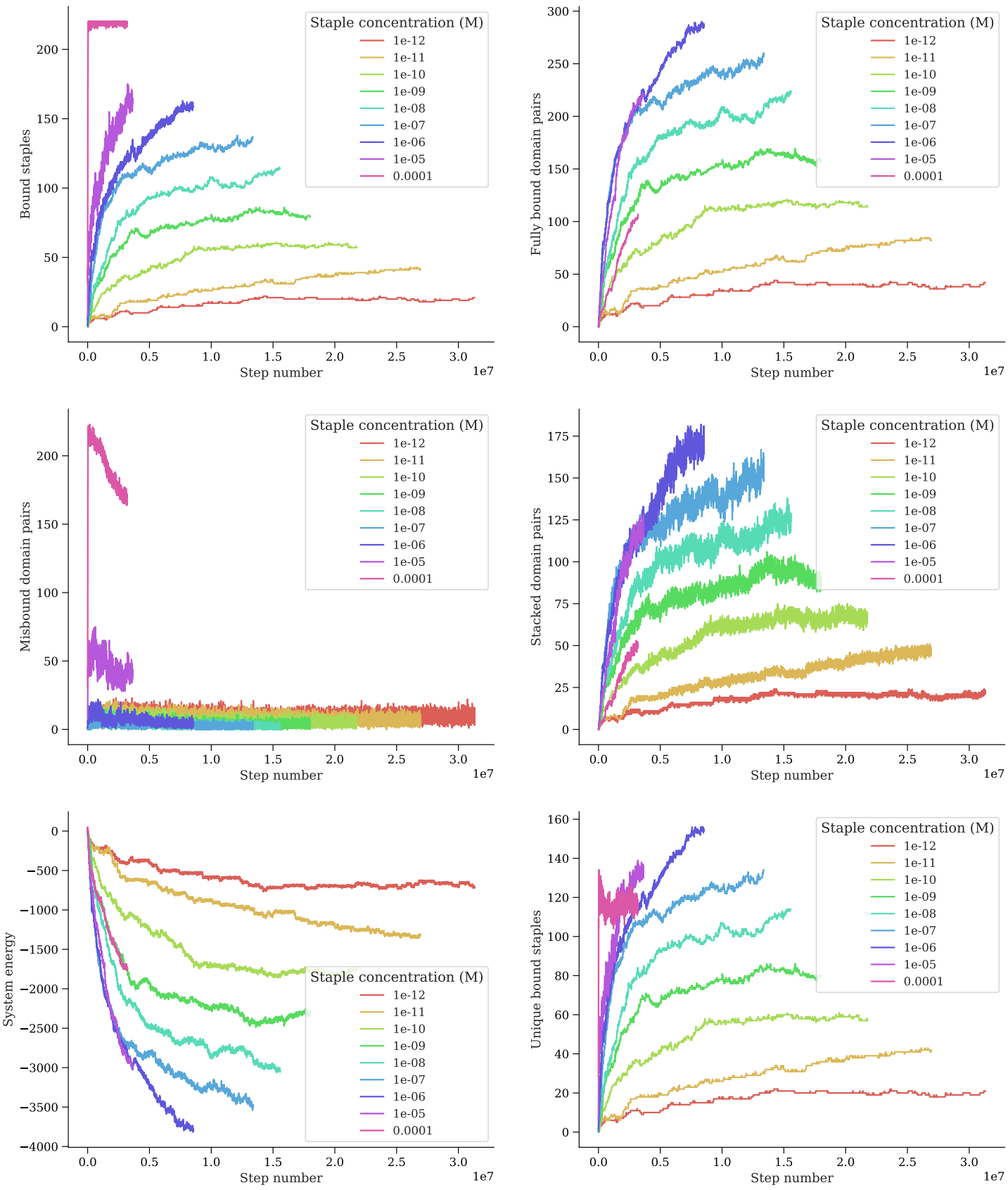

Figure 12: Full simulation results for staple optimisation in Figure 3.

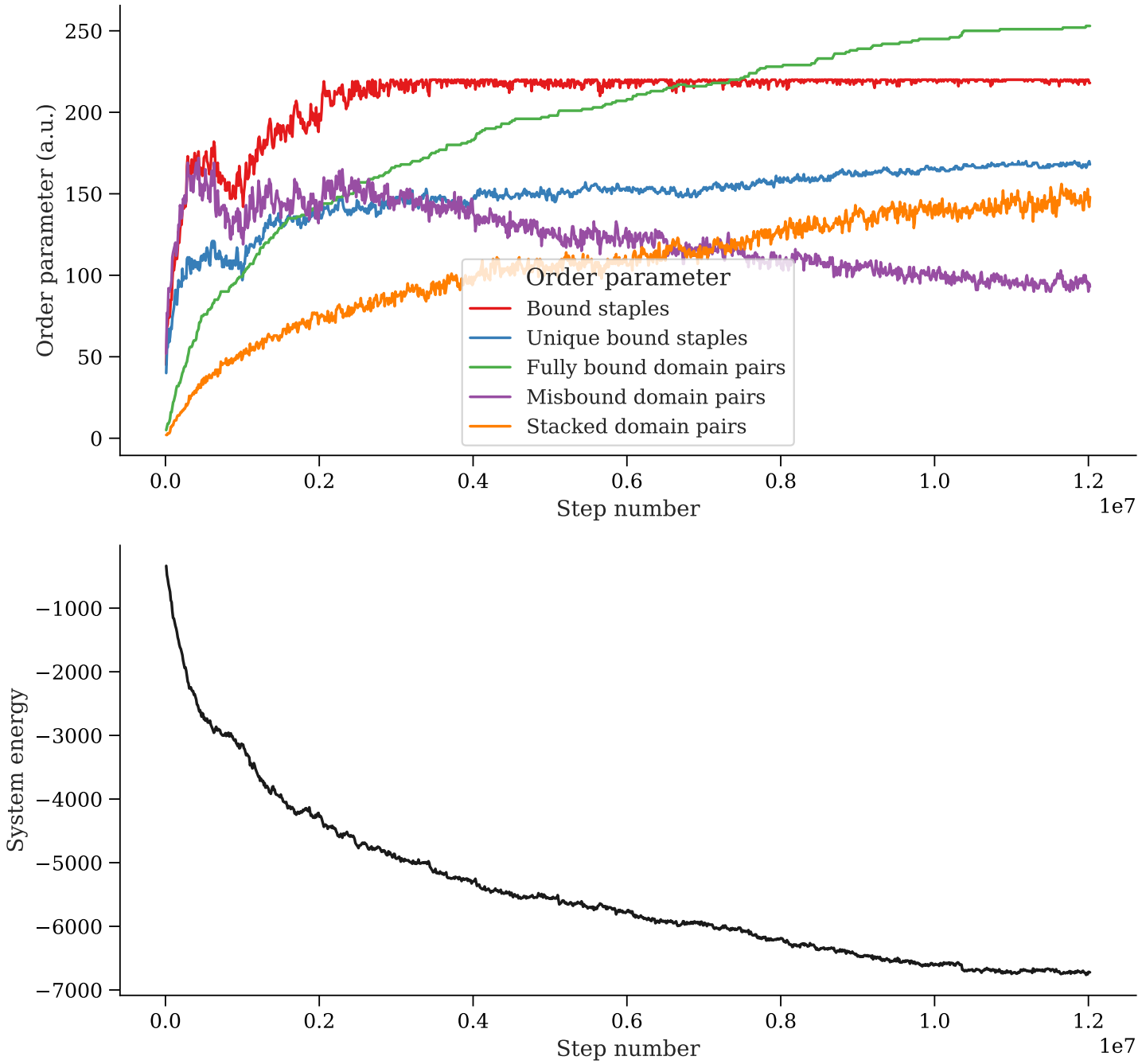

Figure 13: Second run of the same simulation as in Figure 4.

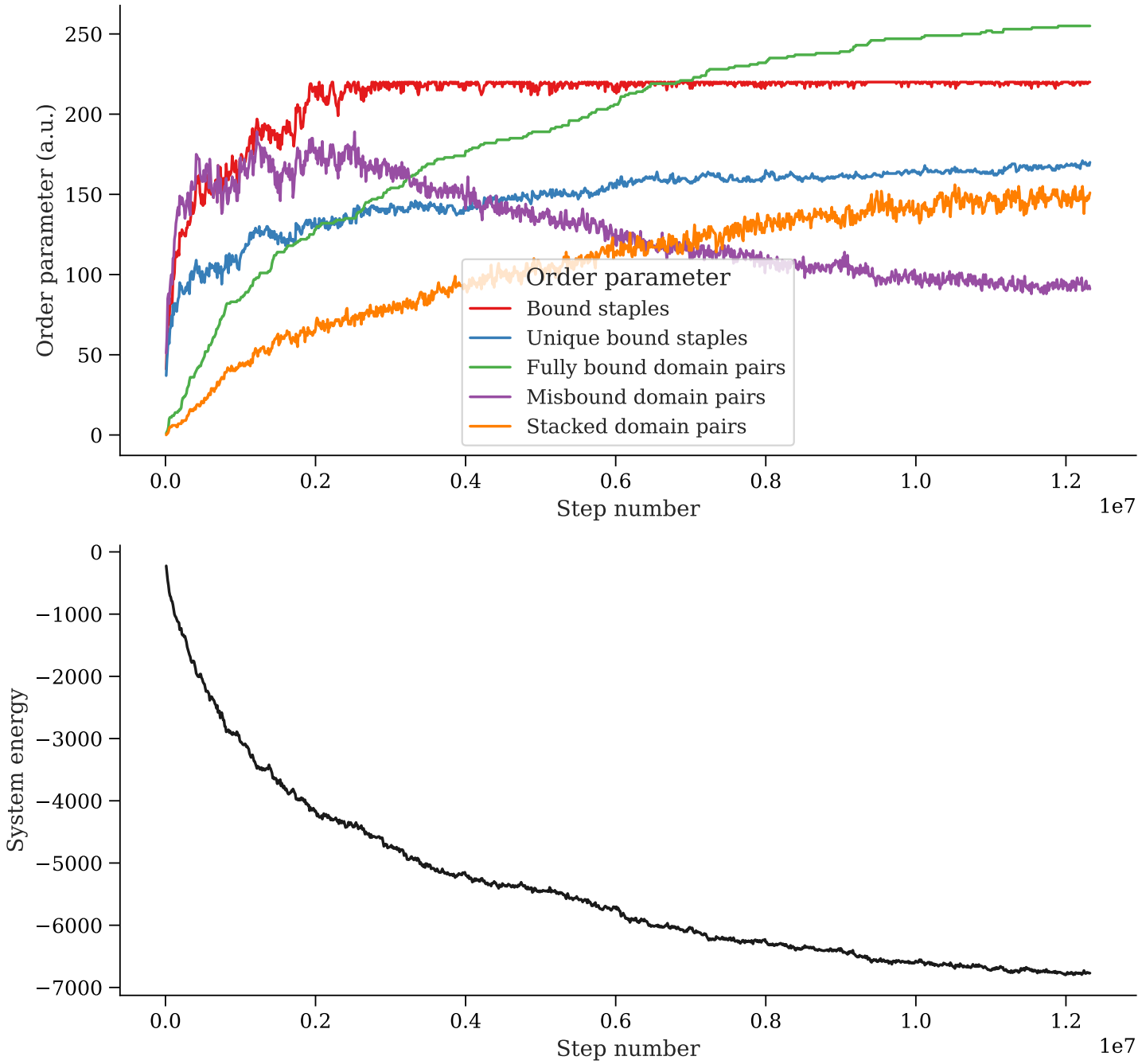

Figure 14: Third run of the same simulation as in Figure 4.

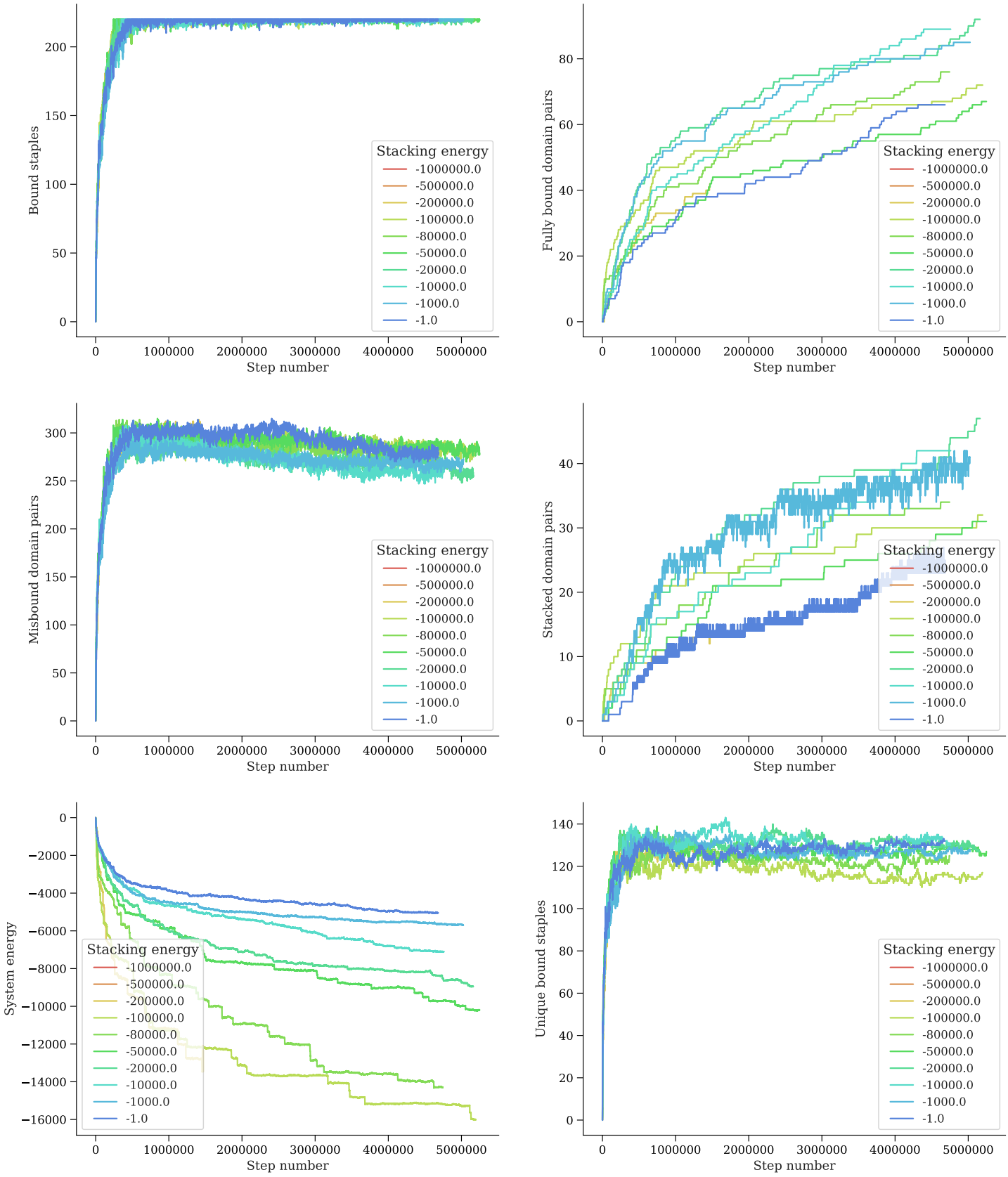

Figure 15: Full simulation results for stacking energy scanning in Figure 5.

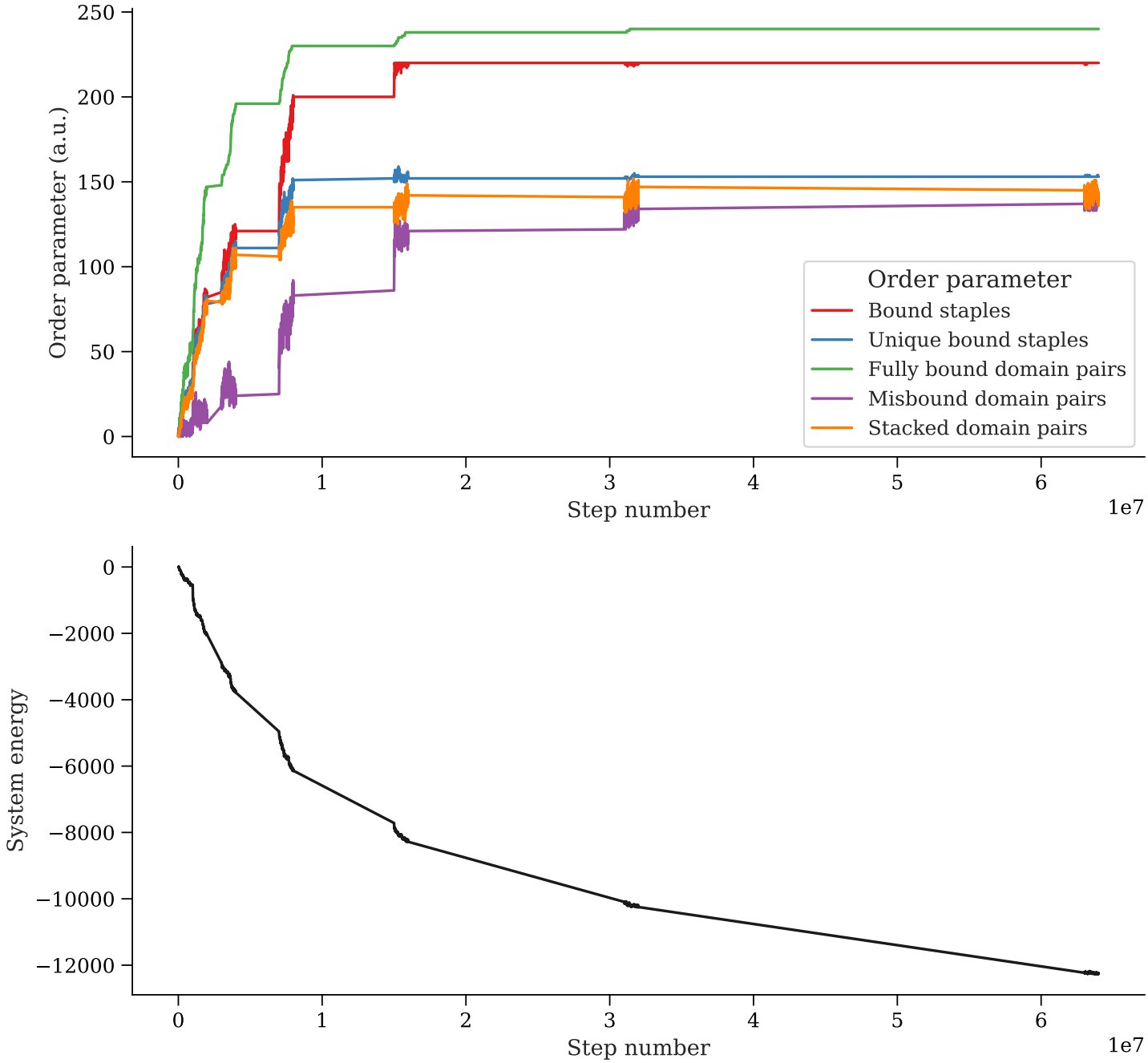

Figure 16: Second run of the same simulation as in Figure 6.