Native-Contact Guided Replica Exchange Simulations of Structured RNA

A. Voronin, A. Schug

published in

NIC Symposium 2022
M. Müller, Ch. Peter, A. Trautmann (Editors)

Forschungszentrum Jülich GmbH,
John von Neumann Institute for Computing (NIC),
Schriften des Forschungszentrums Jülich, NIC Series, Vol. 51,
ISBN 978-3-95806-646-5, pp. 65.
http://hdl.handle.net/2128/31840

Native-Contact Guided Replica Exchange Simulations of Structured RNA

Arthur Voronin[1] **and Alexander Schug**[2,3]

[1] Steinbuch Centre for Computing, Karlsruhe Institute of Technology,
76344 Eggenstein-Leopoldshafen,Germany
E-mail: arthur.voronin@kit.edu

[2] John von Neumann Institute for Computing, Jülich Supercomputing Centre,
Forschungszentrum Jülich, 52428 Jülich, Germany
E-mail: al.schug@fz-juelich.de

[3] Faculty of Biology, University of Duisburg-Essen, 45117 Essen, Germany

Proteins and other biomolecules such as RNA and their interactions determine life at the molecular level. Similar to proteins, RNA plays central roles in living organisms including genetic regulation or information transfer. RNA structure and function are closely coupled yet structurally resolving RNAs is experimentally challenging. *In silico* structure prediction techniques can complement experimental efforts to gain structural information. Here, we explore how replica-exchange molecular dynamics (REX) enhanced with contact data can be used to predict RNA tertiary structure. Such contact data can be obtained by statistical inference of spatial adjacency in the exponentially growing sequence databases via methods such as direct coupling analysis. Our computational approach can thus complement experimental efforts and predict biomolecular structure for a class of biomolecules, which is structurally less explored than proteins. We test this approach on three different structured RNA. We find that contact-guided REX is capable of improving models of a RNA structure and can deliver physically reasonable 3D structures.

1 Introduction

The incredible advances in sequencing techniques have lead to an exponential growth of RNA sequence data[1]. The development of improved statistical methods and alignment software allows building multiple sequence alignment data (MSA) based on inferring structural or even functional relations within and between phylogenetic trees[2–8]. Freely accessible databases such as UniProt, the protein family (Pfam) and RNA family databases (Rfam) offer a treasure trove of sequence information on both proteins and ribonucleic acids (RNA)[9–11]. In the last decade, direct coupling analysis (DCA) and similar methods have distilled this information to predict spatial adjacency based on co-evolution by statically identifying linked mutational patterns[12–18]. Similar to using experimental information as restraints[19–21], contact information can be used as structural restraints to build biomolecular models of, e.g., protein complexes[12] or conformational transitions[22] even on a large scale[23] or to redesign protein signalling[24]. More recently, data-driven machine-learning (ML) methods have shown to astonishingly enhance protein 3D structure prediction[25–27]. These ML approaches, however, rely on deep neural networks with an accordingly massive number of free parameters. Naturally, training these deep networks require equally very large datasets of structurally known 3D structures and accompanying MSAs. As only a limited number of available experimentally resolved structures exist for RNA, these methods cannot be directly transferred to them[28,29]. It is, however, possible to improve RNA

contact prediction[16,18] by shallow learning approaches[30] with limited amounts of parameters.

In the current study, we tested to what extend molecular dynamics can structurally predict RNA or stabilise its native state. We will compare approaches based on regular molecular dynamics (MD), replica exchange (REX) and contact enhanced molecular dynamics (REX bias). REX is a parallel implementation of MD by running multiple non-interacting copies of the system in parallel at different temperatures[31–33]. Switching between different temperatures allows overcoming entrapment in local minima. As REX maintains thermodynamically correct ensembles during the simulation, such simulations offer not only a structure prediction method but also the possibility to derive, e.g., partitions sums. As partition sums allow to derive basically arbitrary thermodynamic properties of a system, such simulations can also provide insight into the physical properties that drive structure adoption. The conducted study is thus similar to our prior work on proteins, where we compared MD vs. REX and carefully investigated optimal weight and bias potentials[34a]. We additionally compare our results to another prior study[30], where we used SimRNA[35] for RNA structure prediction. SimRNA is based on a Monte-Carlo approach for the sampling of the RNA conformational space and a statistical potential to evaluate the energy of the configuration.

2 Results

Method	Backbone RMSD (Å)		
	3d2g	4tzx	4yaz
MD ref	7.6	3.8	20.4
REX ref	7.4	3.6	20.3
REX biased†	5.0	4.0	15.8
SimRNA	15.7	7.5	16.1
SimRNA + mfDCA†	8.5	3.0	20.1
SimRNA + CoCoNet†	16.2	7.8	14.0

Table 1. **Lowest observed RMSD values of performed RNA MD simulations.** Listed are the method, PDB ids and backbone RMSDs with respect to the native fold. Values are obtained via own simulations (upper half) or other methods (lower half, obtained from SI of Ref. 30). Best cases are highlighted for each RNA target. †: Cases with L bias contacts (L: sequence length).

The tested RNA targets were the TPP riboswitch (PDB id: 3d2g[36]), Adenine riboswitch (PDB id: 4tzx[37]), and the 3',3'-cGAMP riboswitch (PDB id: 4yaz[38]). To mimic DCA-based contacts, we added L native contacts (L is the sequence length) into the biased simulations. These contacts were randomly chosen from the native contact map, as seen in Fig. 4. While "real" DCA data would also include erroneous contact data, using only native contacts serves as a first baseline to test the possible improvement. In Tab. 1 we

[a]It should be noted that biasing potentials can be corrected after simulation to regain unbiased thermodynamic equilibrium/properties.

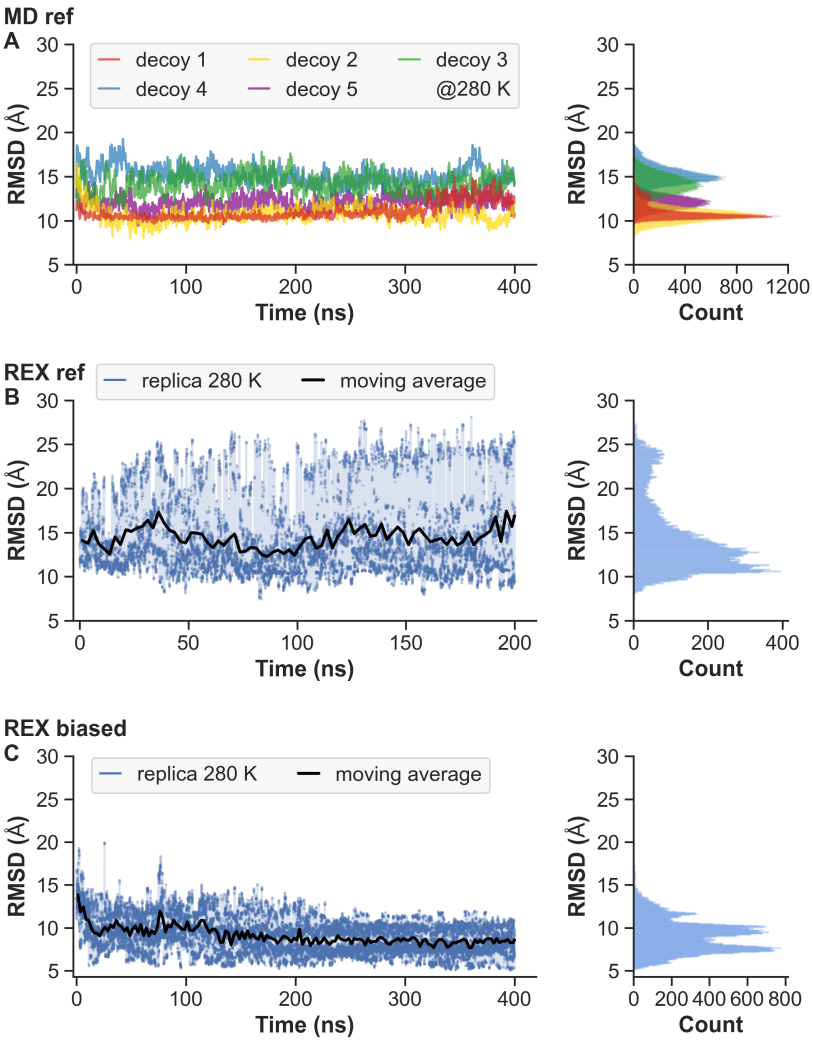

Figure 1. **RMSD comparison of simulations with TPP riboswitch (PDB id: 3d2g[36]).** Backbone RMSD of MD trajectories at 280 K. MD simulations were performed with five unique starting conformations (*decoys*). **(A)** MD without bias. **(B)** REX without bias. **(C)** REX with bias.

summarised the lowest observed RMSD values with respect to the native fold we found in our simulations. Figs. 1 to 3 show a comparison of MD and REX simulations.

In the case of TPP riboswitch (cf. Fig. 1) RMSD values reached approx. 5 Å during our simulations. Our REX bias strongly outperformed all other methods, whereas normal REX reached values of 7.4 Å and biased SimRNA with added mean field DCA[14] only 8.5 Å. In the case of the second RNA target, i.e. Adenine riboswitch (cf. Fig. 2), normal REX achieved 3.6 Å and performed slightly better than the biased variant with 4.0 Å.

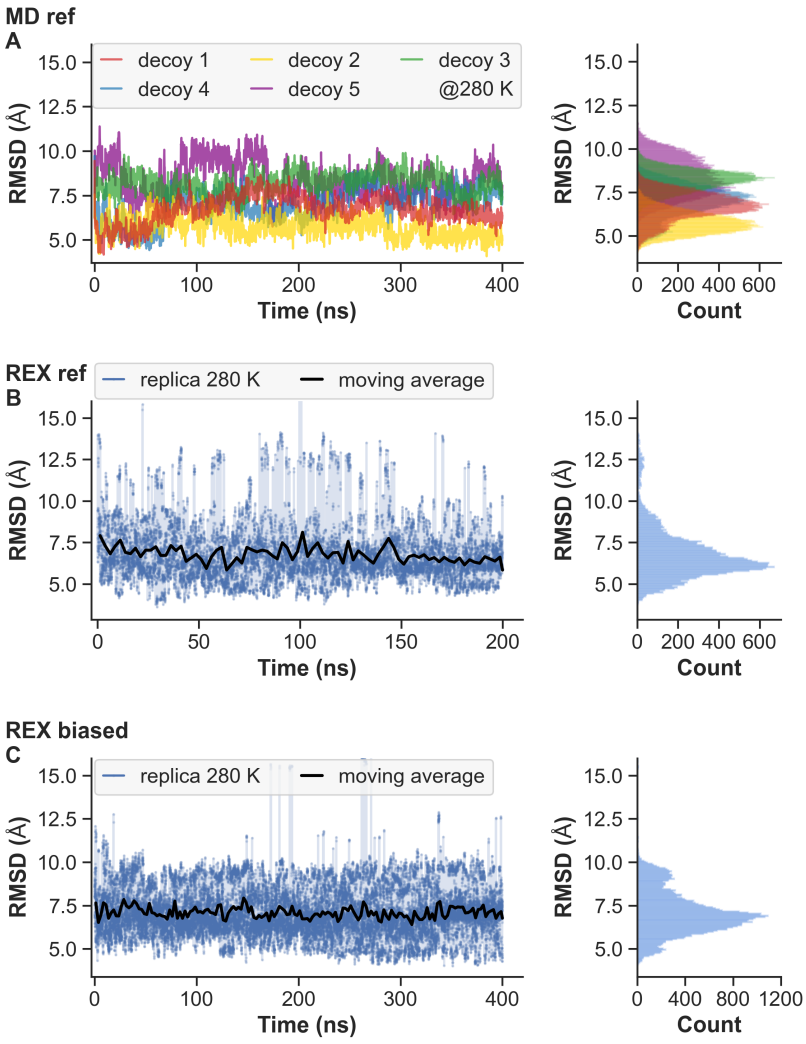

Figure 2. **RMSD comparison of simulations with Adenine riboswitch (PDB id: 4tzx[37]).** Backbone RMSD of MD trajectories at 280 K. MD simulations were performed with five unique starting conformations (*decoys*). **(A)** MD without bias. **(B)** REX without bias. **(C)** REX with bias.

SimRNA + mean field DCA yielded the best results with an RMSD of 3.0 Å. We want to emphasise that our applied force field seems for these cases to be very accurate and reliable, which is reflected in the good RMSD statistics of the performed MD simulations for both TPP riboswitch and Adenine riboswitch. Furthermore, the Adenine riboswitch seems to be very flexible by nature. The attractive force resulting from the applied bias potential was found to be too strong and resulted in a slight local bending of the RNA structure, which is reflected in the higher RMSD values as compared to the unbiased case. Lastly,

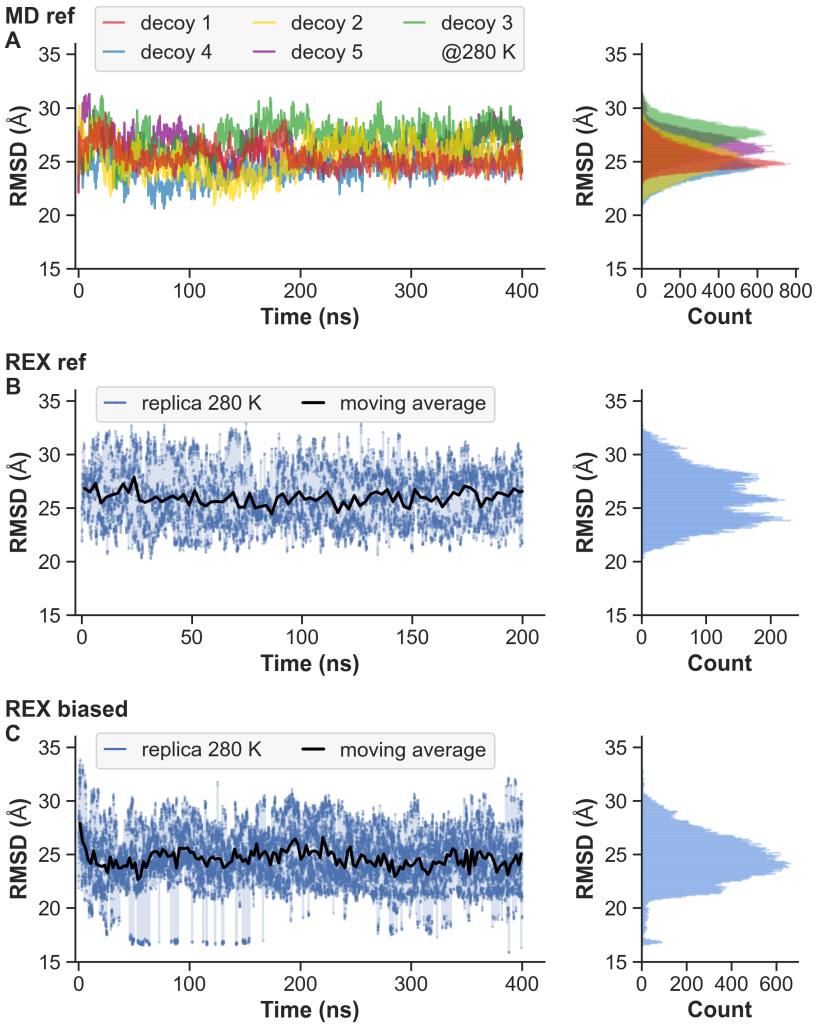

Figure 3. **RMSD comparison of simulations with 3',3'-cGAMP riboswitch (PDB id: 4yaz[38]).** Backbone RMSD of MD trajectories at 280 K. MD simulations were performed with five unique starting conformations (*decoys*). **(A)** MD without bias. **(B)** REX without bias. **(C)** REX with bias.

the comparison of the 3',3'-cGAMP riboswitch simulations are shown in Fig. 3. Our performed simulations yield the best results for the biased REX case, yet again. We want to stress, however, that the overall poor performance might have two courses. Either a) the force fields struggle with this particular example or b) the structure of the 3',3'-cGAMP riboswitch was experimentally measured as a dimer which is not reflected in the simulations, where we used only one molecule chain. Due to the missing counterpart and their physical interactions, the simulations would more likely to adopt non-native other struc-

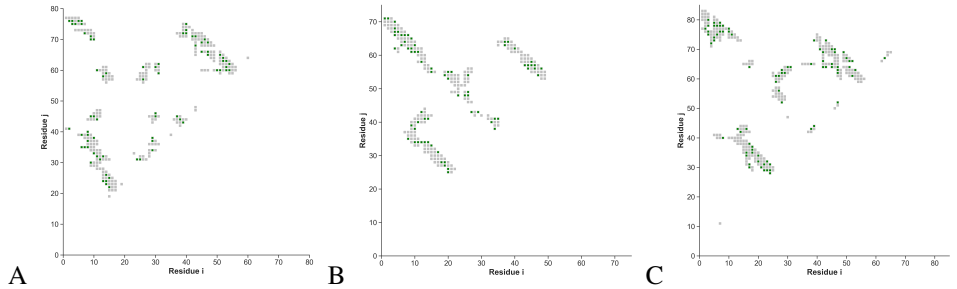

Figure 4. **Contact map of (A) TPP riboswitch (PDB id: 3d2g[36]), (B) Adenine riboswitch (PDB id: 4tzx[37]) and (C) 3',3'-cGAMP riboswitch (PDB id: 4yaz[38]).** Displayed are the native contacts (gray) and 75/70/80 randomly selected true-positive bias contacts (green) used in the respective REX-Bias simulations.

tures and therefore all methods led to relatively high RMSD values. While the biased REX simulations yielded values of 15.8 Å, SimRNA + CoCoNet was capable to achieve 14.0 Å. As all compared methods struggle, we tend to assume that b), i.e. the dimeric nature, is the source of the poor quality predictions.

3 Discussion and Summary

Overall, the application of contact-guided biased REX was able to provide low RMSD structures in two cases with one case only leading to poor quality structure. The poor performance of the third case can likely be attributed to its dimeric nature. There are several future avenues. First, DCA or related methods cannot reliably predict high numbers of error-free contacts. Future simulations should probe the methods robustness towards being fed erroneous contact information. Second the applied bias potential was optimised for proteins and not for RNAs. Hence, this method might or even should yield even better results once the bias potential has been optimised/ fine tuned specifically for RNA targets. Nevertheless, contact-guided REX provides good results which are on par with the best performances of the other mentioned methods, as shown in Tab. 1.

4 Methods

Each REX simulation had 60 replicas and a total of 20 unique starting conformations/ decoys (cf. Tab. 2). The temperatures were distributed covering a range of 280K to 370K. Time steps were set to 2 fs and exchange attempts each 500 MD steps. Exchange rates over the duration of the simulation were in the order of $10 - 15\%$. The all-atom simulations of this study utilised the OL15 nucleic force field[39,40] and TIP3P[41] explicit water model. All MD simulations are performed using GROMACS 2020[42] with cube-shaped systems boxes with periodic boundary conditions (PBC). For all simulations, we use a leap-frog integrator, V-rescale thermostat and Parrinello-Rahman barostat. For evaluation we used our toolkit pyrexMD[43].

Replica number	Decoy RMSD (Å) 3d2g	4tzx	4yaz	Replica number	Decoy RMSD (Å) 3d2g	4tzx	4yaz
1/21/41	10.96	9.37	21.45	11/31/51	12.64	7.67	23.35
2/22/42	13.55	7.49	23.31	12/32/52	14.13	6.57	22.89
3/23/43	12.41	6.23	23.28	13/33/53	12.60	7.88	22.58
4/24/44	13.95	9.68	21.56	14/34/54	13.20	8.65	21.54
5/25/45	12.45	7.47	22.48	15/35/55	14.32	9.37	23.57
6/26/46	13.03	11.24	24.68	16/36/56	11.15	7.78	22.53
7/27/47	13.08	8.49	20.86	17/37/57	10.97	7.78	24.90
8/28/48	10.94	8.90	24.58	18/38/58	12.58	10.57	24.35
9/29/49	13.21	9.41	22.35	19/39/59	13.58	9.19	21.29
10/30/50	11.85	10.30	21.80	20/40/60	12.79	9.19	21.45

Table 2. **Starting decoy accuracy of performed REX simulations.** Table shows the corresponding replica numbers, PDB ids of RNA targets and backbone root-mean-square-deviation (RMSD) before the simulation started.

Acknowledgements

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC).

References

1. B. Hwang, J-H. Lee, and D. Bang, *Single-cell RNA sequencing technologies and bioinformatics pipelines*, Experimental & Molecular Medicine, **50**, 1–14, 2018.
2. S. R. Eddy, *Hidden Markov models*, Curr. Opin. Struct. Biol., **6**, 361–365, 1996.
3. E. P. Nawrocki and S. R. Eddy, *Infernal 1.1: 100-fold faster RNA homology searches*, Bioinformatics, **29**, 2933–2935, 2013.
4. S. R. Eddy and R. Durbin, *RNA sequence analysis using covariance models*, Nucleic acids research, **22**, no. 11, 2079–2088, 1994.
5. J. D. Thompson, T. J. Gibson, and D. G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX*, Curr. Protoc. Bioinf., **00**, 2.3.1–2.3.22, 2002.
6. H. McWilliam, W. Li, M. Uludag, S. Squizzato, Y. M. Park, N. Buso, A. P. Cowley, and R. Lopez, *Analysis Tool Web Services from the EMBL-EBI*, Nucl. Acids Res., **41**, W597–W600, 2013.
7. D. J. Lipman and W. R. Pearson, *Rapid and sensitive protein similarity searches*, Science, **227**, 1435–1441, 1985.
8. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *Basic local alignment search tool*, J. Mol. Biol., **215**, 403–410, 1990.
9. The UniProt Consortium, *UniProt: a worldwide hub of protein knowledge*, Nucl. Acids Res., **47**, D506–D515, 2019.
10. S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, *Rfam: an RNA family database*, Nucl. Acids Res., **31**, 439–441, 2003.

11. R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Everhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta, *Pfam: the protein families database*, Nucl. Acids Res., **42**, D222–D230, 2014.
12. A. Schug, M. Weigt, J. N. Onuchic, T. Hwa, and H. Szurmant, *High-resolution protein complexes from integrating genomic information with molecular simulation*, Proceedings of the National Academy of Sciences, **106**, no. 52, 22124–22129, 2009.
13. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, *Identification of direct residue contacts in protein-protein interaction by message passing*, Proc. Natl. Acad. Sci. U.S.A., **106**, 67–72, 2009.
14. F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, *Direct-coupling analysis of residue coevolution captures native contacts across many protein families*, Proc. Natl. Acad. Sci. U.S.A., **108**, E1293–E1301, 2011.
15. D. S. Marks, T. A. Hopf, and C. Sander, *Protein structure prediction from sequence variation*, Nature Biotechnol., **30**, 1072–1080, 2012.
16. E. De Leonardis, B. Lutz, S. Ratz, S. Cocco, R. Monasson, A. Schug, and M. Weigt, *Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction*, Nucleic acids research, **43**, no. 21, 10444–10455, 2015.
17. S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, *Inverse statistical physics of protein sequences: a key issues review*, Rep. Prog. Phys., **81**, 032601, 2018.
18. M. B. Zerihun, F. Pucci, E. K. Peter, and A. Schug, *pydca v1. 0: a comprehensive software for Direct Coupling Analysis of RNA and Protein Sequences*, Bioinformatics, **36**, no. 7, 2264–2265, 2020.
19. M. Weiel, I. Reinartz, and A. Schug, *Rapid interpretation of small-angle X-ray scattering data*, PLoS Computational Biology, **15**, no. 3, e1006900, 2019.
20. I. Reinartz, C. Sinner, D. Nettels, B. Stucki-Buchli, F. Stockmar, P. T. Panek, C. R. Jacob, G. U. Nienhaus, B. Schuler, and A. Schug, *Simulation of FRET dyes allows quantitative comparison against experimental data*, The Journal of Chemical Physics, **148**, no. 12, 123321, 2018.
21. M. Weiel, M. Götz, A. Klein, D. Coquelin, R. Floca, and A. Schug, *Dynamic particle swarm optimization of biomolecular simulation parameters with flexible objective functions*, Nature Machine Intelligence, **3**, no. 8, 727–734, 2021.
22. A. E. Dago, A. Schug, A. Procaccini, J. A. Hoch, M. Weigt, and H. Szurmant, *Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis*, Proceedings of the National Academy of Sciences, **109**, no. 26, E1733–E1742, 2012.
23. G. Uguzzoni, S. J. Lovis, F. Oteri, A. Schug, H. Szurmant, and M. Weigt, *Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis*, Proceedings of the National Academy of Sciences, **114**, no. 13, E2662–E2671, 2017.
24. R. R. Cheng, F. Morcos, H. Levine, and J. N. Onuchic, *Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information*, Proceedings of the National Academy of Sciences USA, **111**, no. 5, E563–E571, 2014.

25. A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult, *Critical assessment of methods of protein structure prediction (CASP)—Round XIII*, Proteins: Structure, Function, and Bioinformatics, **87**, no. 12, 1011–1020, 2019.
26. A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, *Improved protein structure prediction using potentials from deep learning*, Nature, **577**, no. 7792, 706–710, Jan 2020.
27. M. AlQuraishi, *End-to-End Differentiable Learning of Protein Structure*, Cell Systems, **8**, no. 4, 292–301.e3, 2019.
28. F. Pucci and A. Schug, *Shedding light on the dark matter of the biomolecular structural universe: Progress in RNA 3D structure prediction*, Methods, **162-163**, 68–73, 2019.
29. F. Pucci, M. B. Zerihun, E. K. Peter, and A. Schug, *Evaluating DCA-based method performances for RNA contact prediction by a well-curated dataset*, RNA, rna–073809, 2020.
30. M. B. Zerihun, F. Pucci, and A. Schug, *CoCoNet—boosting RNA contact prediction by convolutional neural networks*, Nucleic Acids Research, **49**, no. 22, 12661–12672, 2021.
31. U. H. E. Hansmann and Y. Okamoto, Curr. Opin. Struct. Biol., **9**, 177–183, 1999.
32. U. H. E. Hansmann and L. T. Wille, Phys. Rev. Lett., **88**, 068105, 2002.
33. A. Schug, W. Wenzel, and U. H. E. Hansmann, J. Chem. Phys., **122**, 194711, 2005.
34. A. Voronin, M. Weiel, and A. Schug, *Including residual contact information into replica-exchange MD simulations significantly enriches native-like conformations*, PloS one, **15**, no. 11, e0242072, 2020.
35. M. J. Boniecki, G. Lach, W. K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K. M. Rother, and J. M. Bujnicki, *SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction*, Nucleic Acids Research, **44**, no. 7, e63–e63, 12 2015.
36. S. Thore, C. Frick, and N. Ban, *Structural basis of thiamine pyrophosphate analogues binding to the eukaryotic riboswitch*, Journal of the American Chemical Society, **130**, no. 26, 8116–8117, 2008.
37. J. Zhang and A. R. Ferré-D'Amaré, *Dramatic improvement of crystals of large RNAs by cation replacement and dehydration*, Structure, **22**, no. 9, 1363–1371, 2014.
38. A. Ren, X. C. Wang, C. A. Kellenberger, K. R. Rajashankar, R. A. Jones, M. C. Hammond, and D. J. Patel, *Structural Basis for Molecular Discrimination by a 3',3'-cGAMP Sensing Riboswitch*, Cell reports, **11**, no. 1, 1–12, 2015.
39. M. Zgarbová, M. Otyepka, J. Šponer, A. Mládek, P. Banáš, T. E. Cheatham III, and P. Jurecka, *Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles*, Journal of Chemical Theory and Computation, **7**, no. 9, 2886–2902, 2011.
40. M. Zgarbová, J. Sponer, M. Otyepka, T. E. Cheatham III, R. Galindo-Murillo, and P. Jurecka, *Refinement of the sugar–phosphate backbone torsion beta for AMBER force fields improves the description of Z-and B-DNA*, Journal of chemical theory and computation, **11**, no. 12, 5723–5736, 2015.

41. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *Comparison of simple potential functions for simulating liquid water*, The Journal of chemical physics, **79**, no. 2, 926–935, 1983.
42. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, *GROMACS: fast, flexible, and free*, Journal of computational chemistry, **26**, no. 16, 1701–1718, 2005.
43. A. Voronin and A. Schug, *pyrexMD: Workflow-Orientated Python Package for Replica Exchange Molecular Dynamics*, Journal of Open Source Software, **6**, no. 68, 3325, 2021.