HARVARD UNIVERSITY
Graduate School of Arts and Sciences



DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Chemistry & Chemical Biology

have examined a dissertation entitled:

Nonhydrolyzable Ribonucleotide Phospohoroimidazolide Analogues for Mechanistic Studies of Nonenzymatic RNA Replication

presented by : Chun Pong Tam

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____
Typed name: Professor Jack W. Szostak

Signature _____
Typed name: Professor Gary B. Ruvkun

Signature _____
Typed name: Professor Luke H. Chao

Date: 12 April 2018

Nonhydrolyzable Ribonucleotide Phosphoroimidazolide Analogues for Mechanistic Studies of Nonenzymatic RNA Replication

A dissertation presented

by

Chun Pong Tam

to

The Department of Chemistry and Chemical Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Chemistry

Harvard University

Cambridge, Massachusetts

April, 2018



© 2018 – Chun Pong Tam

This work is licensed under a Creative Commons

Attribution-NonCommercial 4.0 International License.

Dissertation Advisor: Prof. Dr. Jack W. Szostak *Chun Pong Tam*

Nonhydrolyzable Ribonucleotide Phosphoroimidazolide Analogues for Mechanistic Studies of Nonenzymatic RNA Replication

Abstract

Efficient and faithful replication of genomic information is perhaps one of the most important hallmarks of cellular life. At some early point, such processes must have occurred in the absence of elaborate nucleic acid or protein-based enzymatic catalysts. Template-directed nonenzymatic RNA primer extension represents an important experimental model which aims to illustrate how functional ribonucleotide sequences could have transferred their genetic contents to their offspring. Prototypical nonenzymatic RNA primer extension experiments involve imidazole-activated monoribonucleotides; these activated monomers bind reversibly to an RNA template and polymerize to form the complementary strand, and such processes are catalyzed by a high concentration of divalent metal cations. In Ch. 1, I review recent chemical strategies that enable nonenzymatic RNA copying reactions to proceed efficiently.

Extensive efforts have been invested in determining the thermodynamic parameters and the optimal reaction conditions for regioselectivity and fidelity of nonenzymatic RNA polymerizations, yet some mechanistic aspects of this reaction remain elusive. One of the most perplexing features of nonenzymatic RNA replication is that parent strand copying typically does not proceed to completion. Specifically, the copying of the last template nucleotide is dramatically more sluggish than prior nucleotide incorporations. Orgel and coworkers proposed that the leaving group of the activated monomer downstream of the extending monomer plays important catalytic role(s), but the

mechanistic origin(s) of this effect remained elusive until recently. In Chapter 2, I detail my synthetic chemistry efforts to develop nonhydrolyzable analogues of 2-methylimidazole-activated monoribonucleotides, and the ensuing RNA crystallographic efforts to ascertain the mechanistic nature of the leaving group-leaving group interactions between consecutively-bound activated monomers. Notably, in these RNA-analogue complexes, noncovalent interactions are not observed, challenging the long-held assumption that noncovalent leaving group-leaving group interactions – presumably π-π ,cation-π, or steric in nature – are likely to be responsible for the catalytic effect.

An alternate mechanistic proposal, wherein primer extension is driven by highly-activated 5′-5′-linked imidazolium-bridged dinucleotides formed by two imidazole-activated monomers, began to emerge. In Chapter 3, I describe the RNA crystallography efforts, using P^1,P^3-diguanosine-5′-triphosphate (GpppG) as a stable analogue of the activated imidazolium-bridged dinucleotides, to reveal the possible binding modes of the activated imidazolium-bridged dinucleotides with cognate RNA templates. In chapter 4, I detail the development of guanosine 5'-(4-methylimidazolyl)-phosphonate (ICG) as the closest nonhydrolyzable isosteric analogue of 2-methylimidazole-activated guanosine mononucleotide (2-MeImpG). Using NMR and mass spectrometric techniques, I demonstrate that ICG and activated monomers react to form 5'-5'-linked imidazolium-bridged dinucleotides, with the imidazolium bridge flanked by both labile nitrogen-phosphorus and stable carbon-phosphorus linkages. Purified mixed-linkage dinucleotides react with cognate RNA primer-template complexes to afford extended primers, but only when the labile nitrogen-phosphorus linkage, and not the nonhydrolyzable carbon-phosphorus linkage, is in close proximity with the upstream primer 3′-hydroxyl group.

These analogue-based model systems strongly bolster the hypothesis that the 5'-5'-linked imidazolium-bridged dinucleotides are important intermediates in primer extensions driven by imidazole-activated monoribonucleotides.

The presence of activated “helper oligonucleotides” that bind complementarily downstream of the extending monomer encourages facile RNA copying with high fidelity, but the thermodynamic and/or kinetic origins of this phenomenon have not been well elucidated. In Chapter 5, I quantitatively demonstrate that complementary helper oligonucleotides that bind downstream of the extending dramatically enhance monomer binding: if the monomer is GMP and the downstream helper has a 5′-G terminus, GMP binding is enhanced 100-fold. Nonetheless, *un*activated helper oligomers only modestly accelerate upstream primer extensions, by ~10 fold, thus suggesting that the primer-extension rate enhancement conferred by *activated* helper oligonucleotides must be a combined effect of thermodynamic (enhanced monomer binding affinity) and kinetic (leaving group-leaving group interaction) effects. In Ch.6, using the approach of X-ray crystallography, I investigate the potential structural and conformational effect(s) that activated (and unactivated) helper oligonucleotides could exert on the incoming activated monomer or the overall duplex structure. The helper oligonucleotides utilized in Ch. 6 have either a 5′-ICG cap, a 5′-phosphate or a 5′-hydroxyl group, and the ICG monomer (Ch. 4) is sandwiched between the primer and the helper oligonucleotides. Surprisingly, all three downstream oligonucleotides rigidify the conformation of the bound ICG monomer, forces that monomer to adopt the 3′-*endo* conformation and to bind to the template exclusively with Watson-Crick geometry, all while aligning the phosphonate group of the sandwiched ICG closer to the primer 3′-hydroxyl group for the subsequent

nucleophilic attack. These results suggest that a downstream helper oligonucleotide can catalyze upstream primer extension by placing additional steric and conformational restraints on the incoming monomer relative to the primer.

These results afford new insights into the mechanistic underpinnings of nonenzymatic RNA primer extensions, which will in turn aid in the optimization of this reaction to afford longer and more accurately copied RNA sequences in a reasonable timescale.

Table of Contents

Abstract	iii
Table of Contents	vii
Acknowledgements	x
List of Publications	xviii
Chapter 1: On the Plausibility of the RNA-First Scenario	1
1.1 – Introduction	2
1.2 – RNA as dual information carrier and catalyst	4
1.3 – How do we address the RNA-first question experimentally?	7
1.4 – The problem of monoribonucleotides synthesis	8
1.5 – Brief comment on the pre-RNA world	12
1.6 – Recent progress on monomer synthesis	15
1.7 – Nonenzymatic RNA replication	21
1.8 – Nonenzymatic template-directed RNA copying	23
1.9 – New understandings on the mechanism	31
1.10 – Unmet challenges for nonenzymatic RNA copying	38
1.11 – References	43
Chapter 2: Unusual Base-Pairing Interactions in Monomer-Template Complexes	51
2.1 – Introduction	53
2.2 – Results	56
2.3 – Discussion	72
2.4 – References	78

Chapter 3: Insight into the Mechanism of Nonenzymatic RNA Primer Extension from the Structure of an RNA-GpppG Complex 82
3.1 – Introduction 84
3.2 – Results 89
3.3 – Discussion 97
3.4 – References 101
Chapter 4: Synthesis of a Nonhydrolyzable Nucleotide Phosphoroimidazolide Analogue that Catalyzes Nonenzymatic RNA Primer Extension 103
4.1 – Introduction 105
4.2 – Results 109
4.3 – Discussion 129
4.4 – References 133
Chapter 5: Downstream Oligonucleotides Strongly Enhance the Affinity of GMP to RNA Primer-Template Complexes 137
5.1 – Introduction 138
5.2 – Results 139
5.3 – Discussion 146
5.4 – References 149
Chapter 6: Structural Rationale for the Enhanced Catalysis of Nonenzymatic RNA Primer Extension by a Downstream Oligonucleotide 151
6.1 – Introduction 153
6.2 – Results 156
6.3 – Discussion 176

6.4 – References 181

Appendix A: General Methods 183

Appendix B: Supplementary Materials for Chapter 2 188

Appendix C: Supplementary Materials for Chapter 3 204

Appendix D: Supplementary Materials for Chapter 4 226

Appendix E: Supplementary Materials for Chapter 5 312

Appendix F: Supplementary Materials for Chapter 6 352

Acknowledgements

Quite frankly, as I sit in front of my computer, typing up this dissertation, I am in a state of *disbelief.* My first two years in the chemistry graduate program have been a bit rockier than I have anticipated, and thus – until this very moment, two weeks before my final oral examination – I have never expected that I would be so close to obtaining a doctorate, let alone from a prestigious institution packed with some 380 years of illustrious history behind it. All this is thanks to the saving grace of my dissertation advisor, Prof. Dr. Jack W. Szostak. I joined his group towards the end of my second year, anxious about my scientific abilities. Jack gave me the much-needed opportunity to complete my doctoral training. I am particularly indebted to Jack not only for this chance – none of the work presented herein would have been possible without it – but also for the extraordinary scientific insights, guidance, discussions and resources that come with his splendid mentorship. The opportunity to advance our understanding towards the chemical origins of life, which I considered a problem of paramount *philosophical* importance, is also an incredibly rewarding experience. I genuinely would not have expected an established Nobel Prize winner *always* to have an open office door for his trainees to come in and talk about science and technical chores. I have been constantly anxious when my projects were not progressing as fast as I wanted; yet he has been unbelievably kind and patient with me, giving me too much room to “fool around in the lab,” to fail, to troubleshoot, to learn, and to discover. All in all, I cannot thank Jack enough for being such an incredible and inspirational mentor. I am sure my graduate research experience with Jack will continue to shape who I am as a scientist in the future, and I hope someday I can act as a good mentor for my trainees. I would also like to thank the members of my examination

committee, Prof. Dr. Gary Ruvkun and Prof. Dr. Luke Chao, as well as the members of my graduate advising committee, Prof. Dr. Daniel Kahne and Prof. Dr. David Liu, for their most incisive critiques on my work.

What is truly inspiring about the Szostak group is how the *entire* laboratory strives to be supportive and helpful with one another. The collaborative and cordial atmosphere, as well as the openness of scientific discussion of the group, truly exceeds my every expectation; all of the work presented herein is a result of collaboration in one way or another. I am indeed honored and humbled to have been a part of this scientific family for five years, and I am very fortunate to be able to contribute something to the group.

That being said, there are individuals who I considered to be truly instrumental in my scientific training. I would like to take this opportunity to thank the following extraordinary and awe-inspiring scientific role models (listed alphabetically) who make my Ph.D. possible. First, Prof. Dr. Albert C. Fahrenbach collaborated with me to study the thermodynamics of RNA-monomer binding (Chapter 5). The collaboration started when I was in my fourth-year, when none of my original two projects seemed to be going anywhere, and I desperately needed a new direction. I had very little experience in calculating binding constants and deriving binding isotherms back then, yet he agreed to co-author the paper with me. This amazing act of generosity will stick with me for the rest of my life, and will always serve as my *singular* reminder that I should pay that generosity forward. Dr. Victor Lelyveld – a true guru in nucleic acids chemistry, protein biochemistry and analytical instrumentation – is my go-to person when I needed the most incisive scientific advice and feedback when I needed it the most. The numerous midnight scientific discussions and Viva Burrito dinners would also be cherished in many years to

come, as I am not sure if I will ever find a colleague to eat steak burrito plate with me at *those* late hours. Prof. Dr. Shenglong Zhang was my earliest mentor in the Szostak lab, and he has been crucial in getting me up to speed with the research projects of the lab when I first joined. We have had several opportunities to write NIH grant proposals and research statements on RNA sequencing technologies together; those exercises are integral to my doctoral training. Dr. Wen Zhang is a beyond-phenomenal RNA structural biologist; many of the structures shown in this thesis – and the mechanistic insights hence derived from them – would not be possible without his perspicacious acumen in RNA crystal growing and structure elucidation. Finally, Dr. Lijun Zhou is, to me, a formidable nucleic acids chemist and a sisterly figure who collaborated with me on the nonhydrolyzable analogue project. The fact that I can report my primer extension triplicates in time in the forthcoming pages is due, in no small part, to her work; her involvements in my projects shortened the duration of my Ph.D. by at least half a year. Her resilience and commitment to science is second to none, and I know she would be a great professor who publishes scientific *tour de force*. It is my sincere wish that I can continue to learn from these talented individuals, and develop a solid friendship that go well, *well* beyond our tenure in the Szostak lab.

As for other Szostak lab members, I would first like to acknowledge Dr. Fanny Shui-ying Ng, Ms. Pamela Svec and Mr. Chris Kelleher for being amazing laboratory managers; without their hard work, the lab would cease being functional in no time. I am particularly appreciative of Fanny speaking Cantonese with me in lab; it is incredibly hard to find Cantonese speakers in the Boston area (at least within my social circle), so our Cantonese conversations serve as my reminders that I *do* know my native dialect. Next,

I want to mention my dear friend, colleague, and ski instructor, Prof. Dr. Seung Soo Oh. I hope you are doing well at Pohang University, and can find some time skiing…just don't wear your cursed racing skis and boots; your knees and legs will certainly appreciate it. Scientific discussions (and non-scientific chit-chatting) with hardcore chemists like Dr. Li Li, Dr. Derek K. O'Flaherty, Dr. Tom Wright, Dr. Ayan Pal, Dr. Kyle Strom, Mr. Constantin Giurgiu and Mr. Seohyun Chris Kim, have led to fruitful synthetic routes and novel insights on multiple occasions, and shall always be cherished in my heart. Special kudos to Mr. Travis Walton; his novel mechanistic model has been instrumental in the interpretation of some data presented in this thesis; as well as Ms. Lydia Pazienza, for her involvements in the thiophosphoroimidazolide project. On the other hand, while none of my projects involved vesicles, I have had great friendships with the team members of the vesicle subgroup, including but not limited to Prof. Dr. Neha Kamat, Dr. Anna Wang, Dr. Lin Jin, Dr. Tivoli J. Olsen, Dr. Knut Anders Bjorkbom and Dr. Christian Hentrich. I thank them for discussing with me the in-and-outs of vesicle-related science, and for broadening my scientific horizon. I would also like to thank Dr. Tony Z. Jia, Dr. Noam Prywes, Ms. Zoe R. Todd (thanks a million to your brilliant weekend cookies), Dr. Claudia Bonfio, Mr. Weicheng Zhang, Dr. J. Craig Blain, Dr. Aaron T. Larsen, Dr. Daniel Duzdevich, Dr. Aaron Engelhart, and Prof. Dr. Enver C. Izgu, who have all been helpful with scientific discussions, instrumentations, and manuscript critiquing. I especially enjoyed my interactions with Tony; thanks for introducing me to your circle of nice pals (Dr. Christopher Miller, Dr. George Xu, Dr. Aileen Li, George Somi, Kevin Tian, etc.). I will try to be consistent with my commas, and to take a couple of chill-pills when I need to.

As I reflected on my extended history of schooling, I realized that my college years – spent mostly inside the Shain Chemistry Building of the University of Wisconsin-Madison – have been formative in my decision to pursue graduate studies in chemistry, so I would take this opportunity to thank a couple of key personnel as well. For one, I thoroughly thank Prof. Dr. Shannon S. Stahl and Dr. Richard I. McDonald for being my phenomenal research mentors. I joined the laboratory of Prof. Stahl as an ignorant sophomore with no research experience, but I matured in the lab quickly because of his expert supervision and (almost) one-on-one guidance. Without that rigorous training and wholehearted assistance, I am certain I would not be where I am today. Prof. Dr. Steven D. Burke exposed me to the world of natural product total synthesis in his graduate seminar course, and Dr. Allen D Clauss offered me the initial, wonderful taste of organic chemistry. I thoroughly thank them for exposing me to the ingenuity, beauty and rigor of synthetic organic chemistry. Thanks for holding me to the highest standard when it comes to my undergraduate chemistry training. All the hard work back in those days most certainly helped in my grad school years.

Next, I wish to mention that Mr. Siu-hoi Lui and Mr. Chiu Marco Lam (soon to be Dr. Lam) has cardinal influences in the shaping of my career trajectory. When I was in grade 10, I joined a group of grade 12 students – Marco being one of them – and study the preparation and energy content of biodiesels derived from spent cooking oil, and Mr. Lui was the chemistry lab technician. I do not remember what kinds of ugly data we got back in those days, but I know for sure I have been enthralled by organic chemistry ever since. My decision to pursue chemistry studies overseas was largely shaped by this

experience. Thanks for exposing me to science, and thanks for making my debut in research such an exhilarating one.

Graduate school can be incredibly stressful at times. Special thanks go to the following individuals, who have worked hard to keep my sanity intact throughout these years: Tianming Fu (and Yiyi Liu), for being my loyal roommate, ski-mate, dinner-mate, and *ad hoc* academic advisor; George Junjie Hao, for trying to convince me that it is not *that* difficult to develop a romantic relationship; Fanny Ng, Anna Wang, Chris Kim and Lin Jin, for *really* paying attention to my psychological well-being towards my final months of graduate school. I am eternally grateful to have met you all.

I would genuinely like to thank all my friends from Harvard, University of Wisconsin-Madison and King's College (Hong Kong) who have crossed path with me. Thank you for your friendship, loyalty, and continued support throughout these years. I look forward to the upcoming exciting work that you will be doing next: Prof. Dr. Peng Ding, Dr. Xiabing Lou (& Yue Qian), Dr. Leon Xueliang Liu, Dr. Lei Zhu (& Ruizhi Dai), Dr. Ruihua Ding, Prof. Dr. Yang Chen (& Dr. Tong Qin), Dr. Liuchuan Tong, Xizhu Zhao, Dr. Jun Feng, Xian Gong, Peng Qu, Xiaoyu Zhang, Dr. Yuan Qiao, Mengfei Zhang, Dr. Zhenhua Lai, Dr. Yang Li, Dr. David Lee, Dr. Nathan Ip, Joseph Wing-hei Li, Ronald Yue-him Cheng, Cheuk Lam "CCL" Chan, Sam Yip-keung Siu, Hon-chung "JC" Chan, William Yi-ming Lam, Wai-man Chan, Raymond Man-kit Wu, Dr. Yufan Wu, Lu Lu, Barbara Wang, Dr. Pan San Chan, Dr. Sheryl Man, Ms. Nicole Sicard and many, *many* others.

Finally – *and most importantly* – I must acknowledge my father and mother, Mr. Yip-man Tam and Ms. Yiu-kwan Chi, respectively. I could have easily written another thesis to convey my heartfelt appreciation towards their unwavering support,

unconditional love and boundless patience…I would just say that I could not have persevered through graduate school if not for their timely interventions, solaces and encouragements. It suffices to say that when I am down and when my soul is weary, my parents are my bedrock of faith and confidence; when troubles come and my heart is burdened, the fact that you two are always there, always ready to sit awhile with me and listen, is incredibly comforting and heartening. I have been truly blessed to have both of you be my secular guardian angels, the mountains on which I can stand and rely on. *You raise me up to more than I can be.* Thank you. I could not have done this without you two. *I love you both.*

Friends and colleagues: my pilgrimage towards a doctorate in chemistry – a seven-year odyssey – has been quite a roller-coaster ride indeed, and I will for sure be fond of this experience for the rest of my life. Yet, I am fully aware that my scientific journey has just begun, and I am so ready to learn and discover whatever interesting chemistry and science out there. So, for now, I will use the following line of lyric to conclude my acknowledgments section:

“Life has its pains…it’s good to have you at my side.”

(“生命有痛…有你真好。”)

Dedicated to my parents.

You are my shield

and

my inspiration.

List of Publications (* denotes co-first authors)

1. Zhang, W.; Tam, C. P.; Zhou, L.; Oh, S. S.; Wang, J.; Szostak, J. W. Structural Rationale for the Enhanced Catalysis of Nonenzymatic RNA Primer Extension by a Downstream Oligonucleotide. *J. Am. Chem. Soc.* **2018**, *140*, 2829–2840.

2. Tam, C. P.*; Zhou, L.*; Fahrenbach, A. C.; Zhang, W.; Walton, T.; Szostak, J. W. Synthesis of a Nonhydrolyzable Nucleotide Phosphoroimidazolide Analogue That Catalyzes Nonenzymatic RNA Primer Extension. *J. Am. Chem. Soc.* **2018**, *140*, 783–792.

3. Giurgiu, C.; Li, L.; O'Flaherty, D. K.; Tam, C. P.; Szostak, J. W. A Mechanistic Explanation for the Regioselectivity of Nonenzymatic RNA Primer Extension. *J. Am. Chem. Soc.* **2017**, *139*, 16741–16747.

4. Zhang, W.; Tam, C. P.; Fahrenbach, A. C.; Walton, T.; Birrane, G.; Szostak, J. W. Insight into the Mechanism of Nonenzymatic RNA Primer Extension from the Structure of an RNA-GpppG Complex. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 7659–7664.

5. Fahrenbach, A. C.*; Giurgiu, C.*; Tam, C. P.; Li, L.; Hongo, Y.; Aono, M.; Szostak, J. W. Common and Potentially Prebiotic Origin for Precursors of Nucleotide Synthesis and Activation. *J. Am. Chem. Soc.* **2017**, *139*, 8780–8783.

6. Tan, L.; Gurbani, D.; Weisberg, E. L.; Hunter, J. C.; Li, L.; Jones, D. S.; Ficarro, S. B. Mowafy, S.; Tam, C. P.; Rao, S.; Du, G.; Griffin, J. D.; Sorger, P. K.; Marto, J. A.; Westover, K. D.; Gray, N. S. Structure-Guided Development of Covalent TAK1 Inhibitors. *Bioorg. Med. Chem.* **2017**, *25*, 838–846.

7. Li, L.; Prywes, N.; Tam, C. P.; O'Flaherty, D. K.; Lelyveld, V. S.; Izgu, E. C.; Pal, A.; Szostak, J. W. Enhanced Nonenzymatic RNA Copying with 2-Aminoimidazole-Activated Nucleotides. *J. Am. Chem. Soc.* **2017**, *139*, 1810–1813.

8. Tam, C. P.*; Fahrenbach, A. C.*; Björkbom, A.; Prywes, N.; Izgu, E. C.; Szostak, J. W. Downstream Oligonucleotides Strongly Enhance the Affinity of GMP to RNA Primer–Template Complexes. *J. Am. Chem. Soc.* **2017**, *139*, 571–574.

9. Zhang, W.*; Tam, C. P.*; Wang, J.; Szostak, J. W. Unusual Base-Pairing Interactions in Monomer–Template Complexes. *ACS Cent. Sci.* **2016**, *2*, 916–926.

10. Björkbom, A.*; Lelyveld, V. S.*; Zhang, S.*; Zhang, W.; Tam, C. P.; Blain, J. C.; Szostak, J. W. Bidirectional Direct Sequencing of Noncanonical RNA by Two-Dimensional Analysis of Mass Chromatograms. *J. Am. Chem. Soc.* **2015***, 137,* 14430–14438.

11. Wang, Y.; Jimenez, M.; Sheehan, P.; Zhong, C.; Hung, A. W.; Tam, C. P.; Young, D. W. Selective Access to Trisubstituted Macrocyclic *E*- and *Z*-Alkenes from the Ring-Closing Metathesis of Vinylsiloxanes. *Org. Lett.* **2013**, *15*, 1218–1221.

12. McDonald, R. I.; White, P. B., Weinstein, A. B., Tam, C. P.; Stahl, S. S. Enantioselective Pd(II)-Catalyzed Aerobic Oxidative Amination of Alkenes and Insights into the Role of Electronic Asymmetry in Pyridine-Oxazoline Ligands. *Org. Lett.* **2011**, *13*, 2830–2833.

Chapter 1

On the Plausibility of the RNA-First Scenario

[*C. P. Tam wrote the text. An abridged version of this manuscript will be submitted to Acc. Chem. Res. for peer review and open-access publication.*]

Abstract. There is much evidence that there was once a period in which RNA molecules acted as the central molecule of life, not only being able to transmit genetic information but also to catalyse important chemical reactions. However, the existence of such an "RNA world" does not answer the question as to whether life began with RNA (the "RNA-first" scenario) or whether the RNA world is the evolutionary product of earlier, simpler genetic molecules (the "RNA later" scenario or the "pre-RNA world" hypothesis). Indeed, the scientific plausibility of the RNA-first model has been challenged, given that numerous unsolved problems stand in the way of a prebiotically-plausible and robust synthesis of monoribonucleotides and oligoribonucleotides, as well as subsequent nonenzymatic RNA copying. Herein, we review the remaining challenges associated with the nonenzymatic chemistry of RNA and highlight recent advances in RNA monomer synthesis, activation and nonenzymatic copying of RNA polymers. A plausible mechanistic hypothesis leading to the nonenzymatic origin of RNA catalysis, and hence the "RNA World", is presented.

1.1 Introduction

The problem of the origin of life is of paramount scientific and philosophical importance. The Earth was formed approximately 4.54 billion years ago,[1] replete with simple chemical feedstocks yet aseptic. Simple cellular microorganisms could have existed on Earth as early as 3.9-3.8 billion years ago or even earlier, as suggested by carbon isotope evidence.[2] Recently recovered fossilized microorganism samples embedded within ferruginous sedimentary rocks from Nuvvuaggittuq, Quebec, Canada,[3] as well as stromatolite samples recovered from the Australian Warrawoona formation suggest that relatively simple microbial life could have existed on earth no later than 3.5 Gyr ago.[4] Thus, the crux of the origins of life problem is to propose a scientifically-plausible mechanism that addresses *how* the inanimate chemical matter available on the prebiotic Earth[5] could have reacted to form more complicated biological building blocks.[6] Additionally, how these building blocks self-assembled into the simplest "protocell" which can: 1) synthesize and transfer genetic information to progeny, 2) capture energy and nutrients, and 3) maintain cellular integrity in order to distinguish itself from the environment[7] should also be answered. The challenges associated with the elucidation of this incredible journey, featuring an inordinate number of steps and a timespan across many millions of years, are not to be understated.

Numerous efforts have been devoted to determining the chemical composition of this initial cellular life. These efforts can be broadly characterized as either "top-down"[8] or "bottom up".[7,9] Top-down approaches focus on minimizing the genetic complexity of extant simple bacteria using various genetic analysis, knock-out and mutation techniques to deduce the features of the "minimal" genome[10] of the last universal common ancestor

(LUCA). The most notable effort in recent years is that achieved by Venter, Hutchison and coworkers, in which the 1079-kbp genome analog of *Mycoplasma mycoides* was truncated by half to achieve a minimally-viable bacterial cell with a 531 kilo base pairs (kbp) genome.[11] While efforts are being undertaken to further simplify this genome, using the nonenzymatic RNA polymerization chemistries reported to date, only short (much less than 100 bp) RNA chains are afforded, even under favorable chemical environments (*vide infra*). Thus, given the genomic complexity of LUCA, other researchers – us included – have approached the abiogenesis problem from the "bottom-up": using simple and prebiotically plausible organic molecules to assemble *de novo* a minimal protocell that is capable of replication and evolution.[7,12]

A successful protocell likely had a primitive yet robust mechanisms for genome synthesis, replication and compartmentalization.[7,12] In light of these requirements, RNA is a plausible genomic molecule in the proposed minimal cell, due to the fact that it can undergo faithful genome replication *via* Watson Crick base-pairing rules, as well as perform catalysis in extant molecular biology and *in vitro*-selected aptamers.[13] On the other hand, many so-called "Xeno Nucleic Acids" (XNAs) with alternate backbone linkages have been evaluated for their potential promise as the genomic molecule of the first protocell.[14–16] Finally, fatty-acid based compartments are likely integral components in the earliest cells,[17] due to the fact that these molecules could have been available on the surface of the prebioctic Earth either from Fischer-Tropsch fatty acid synthesis,[18] oxidative cross coupling conditions,[19] or be delivered from extraterrestrial sources.[20] Thus, the key goals at the heart of the bottom-up approach are to elucidate the chemical pathways through which: 1) simple molecules react to form the proposed building blocks

of life, and 2) how these simple building blocks self-assemble to form higher order structures or aggregates that are capable of replication, division, and evolution, all while ensuring that these proposed chemical pathways are not mutually exclusive.[9]

Herein I review key experimental advances derived from “bottom-up” or “chemistry up” approaches towards the facile nonenzymatic synthesis, replication and evolution of an RNA-based genetic material *de novo*. We review recent experimental advances demonstrating prebiotically-plausible synthesis of monoribonucleotides from simple chemical precursors, monoribonucleotide activation and oligonucleotide synthesis, and nonenzymatic RNA copying. These advances, all being central to the eventual advent of a RNA-based genetic material in the earliest cells, render the RNA-first scenario more plausible.

1.2 RNA as dual information carrier and catalyst

The idea of the “RNA world”,[21] as proposed by Crick,[22] Orgel[23] and Woese,[24] posits that RNA-catalyzed RNA replication ensured the genetic continuity of the very earliest protocellular species, a feat made possible by the crucial Watson-Crick base-pairing properties of RNA. Extensive protein-based catalysis is not believed to be critical to the survival of this initial RNA-based protocell.[25] The important discovery by Cech[26] and Altman[27] regarding *T. thermophila* group I autocatalytic intron in 1981,[26] and the RNA catalytic core of RNAse P being the endonuclease of *E. coli* tRNA precursor in 1983,[27] irrefutably demonstrates the catalytic prowess of RNA in extant biology and strongly bolstered the argument for the RNA world. One can now envision a scenario in which an RNA enzyme (ribozyme henceforth) can catalyze its own replication, thus alleviating the philosophical “chicken-and-egg” dilemma of whether catalysis or genome came first.[13]

Since then, many more ribozymes of biological importance have been discovered, and the catalytic repertoire of RNA continues to expand. Among the biologically relevant ribozymes, the most important one is the ribosome; RNA accounts for the bulk of its molecular mass, and it is now clear that the peptidyltransferase center is composed of rRNA. Amino acids are completely absent in an area as wide as 18 Å encircling the catalytic center, suggesting that rRNA is ultimately responsible for the enzymatic activity of the ribosome, while the surrounding protein residues serve as a scaffold and confer structural rigidity to the catalytic RNA core.[28] On the other hand, while RNA accounts for as much as only 5% of the molecular weight of the eukaryotic spliceosome,[29] RNA is again the actual catalytic species: Piccirilli and coworkers demonstrated that the splicing activity of phosphorothioate-substituted U6 snRNA was restored in the presence of thiophilic cadmium(II) catalysts instead of the oxophilic magnesium(II).[30] Other ribozymes of prokaryotic or viral origins that can cut RNAs[31] include the hammerhead,[32] Varkud satellite,[33] Hepatitis D virus,[34] and *glmS* ribozymes.[35] Finally, riboswitches are structured noncoding RNA domains that could regulate gene expression[36] as they bind to small molecules such as thiamine pyrophosphate,[37] adenosylcobalamin[38] or S-adensoylmethionine,[39] to name a few.

In vitro methodologies, the most important one being Systematic Evolution of Ligands by Exponential Enrichment (SELEX)[40] have been used to isolate RNA molecules that bind with high specificity to small organic molecules or peptides,[41] or to select for molecules that can perform certain catalytic functions.[42,43] The RNA-catalyzed synthesis of 4-thiouridine-5-phosphate from 4-thiouracil and 5-phosphoribosyl-1-pyrophosphate,[44] as well as oligonucleotide triphosphorylation using prebiotically plausible

trimetaphosphate,[45] is perhaps indicative of the relevance of ribozymes in the RNA-first scenario. Remarkably, RNA-based RNA polymerases have also been evolved *in vitro,* demonstrating that RNA molecules may indeed be able to catalyze their own replication in the broad sense.[43] Following the discovery of class I ligase by Bartel and Szostak,[46] Bartel and his coworkers were able to evolve the 189 nt long R18, the world's first *in vitro*-evolved RNA polymerase based upon the class I ligase scaffold.[47] This rather complicated molecule, when presented with an arbitrary, single-stranded RNA template, an RNA primer, and a mixture of the four nucleoside triphosphates, will synthesize the complement of a template of up to 14 nucleotides long; the efficiency of template-dependent polymerization drops dramatically beyond this length. Further *in vitro* evolution has led to the isolation of much more powerful ribozymes such as tC19Z,[48] tC9Y[49] and, very recently, 24-3.[50] Notably, tC9Y can, for the first time, copy sequences longer than the polymerase itself (up to 206 nts) under optimal conditions, such as favorable template sequences and sub-zero, ice-water eutectic conditions, albeit at very low yield (0.01% yield).[49] On the other hand, 24-3 is a *trans*-acting ribozyme that can not only copy a 41-nt-long, purine-rich GTP-binding aptamer in good yield, but is also able to amplify the concentration of a 24-nt-long template, in the presence of its complementary 12-nt primers, by as much as 100-fold.[50] While these structurally complex ribozymes are typically well over 150 nt long, well over the range of length currently assessible by nonenzymatic RNA copying, Lincoln and Joyce have demonstrated that long oligonucleotide strands could potentially be accessible by ribozyme-catalyzed ligation. This was demonstrated using an evolved R3C ligase ribozyme that can cross-catalytically ligate two oligonucleotide fragments, one 52 nt and the other one 13 nt long, together to

form longer, more complex structures.[51] The isolation of these ribozymes constitutes a major advance in attempts to understand the origin of life because we can now begin to envision how RNA-catalyzed RNA ligation and/or polymerization could have proceeded. These “smoking guns” have led to a general acceptance of the RNA World hypothesis.

1.3 How do we address the RNA-first question experimentally?

Supposing that the RNA world was the first organized biochemical world on the primitive earth, Orgel argued that the RNA-first problem could be solved by demonstrating robust prebiotic chemical pathways for the following four processes:[13]

1. Nonenzymatic synthesis of nucleotides
2. Polymerization of nucleotides to give random sequence oligoribonucleotides
3. Nonenzymatic Copying and Replication of RNA
4. Emergence of a set of functional RNA catalysts that could sustain exponential growth in the prebiotic environment *via* natural selection.

In order for the RNA world hypothesis to be considered convincing, each of these steps must have a satisfying chemical solution. However, upon delving deeper into the chemistry catalogued for each of these stages, one inevitably runs into substantial issues: classical prebiotic syntheses of the RNA components – nucleobases, ribose, and subsequently phosphorylated monoribonucleotides – often have disappointing yields and come with significant side reactions, thus producing intractably complex mixtures of molecules that are structurally related to but are not exclusively ribonucleotides.[13] Such a “prebiotic chemist’s nightmare”[25] (briefly reviewed below) has led many to believe that the *de novo* synthesis of RNA monomers is too complex to arise from prebiotically plausible reactions on the early earth; without a stable supply of ribonucleotide monomers,

subsequent evolutionary steps are stalled. Consequently, researchers have contemplated the possibility that life could have begun with alternate genetic systems with chemical structures resembling, but simpler than, RNA, that can transfer genetic information *via* specific and complementary recognition, yet with simpler chemical structures.[14,15] We will address these questions in a systematic manner, and see that the field has progressed significantly over the years regarding the nonenzymatic synthesis of RNA, lending plausibility to the RNA-first model.

1.4 The problem of monoribonucleotide synthesis

How could the first nucleotide arise from the prebiotic soup? There are several criteria guiding our search towards a robust and prebiotically plausible synthetic pathway. First of all, the starting materials must be chemically simple and postulated to be abundant during the early geology of the Earth; the yield of the synthetic steps should be "reasonable" and the reactions should be able to be conducted in water,[13] although reactions occurring in other prebiotically-plausible solvents have been probed recently.[52,53]

The key difficulty experienced by prebiotic chemists is essentially that most reported strategies historically afford complicated mixtures of materials, with the target compound in low yield, and many side products that are chemically, structurally or stereochemically related to the target compound. Additionally, the preparation of certain biological building blocks (*e.g.* nucleotides) may require multiple steps, yet the experimental conditions for these steps may be mutually incompatible. Numerous reviews have extensively surveyed the chemistry available for the synthesis of various prebiotic chemical feedstocks,[5,6,13] so only a brief summary will be offered herein. Take the

legendary Miller-Urey spark discharge experiment as an example.[54] Several gas mixtures of different compositions have been heated and subjected to high voltage, and the resultant organic material has been analyzed by paper chromatographic[54] and, more recently, liquid chromatography-mass spectrometry techniques.[55] A complex mixture of structurally related butyric acids, alkyl amines and simple amino acids was detected. In a related spark discharge experiment wherein hydrogen sulfide was present, sulfur- and sulfoxide- containing amino acids were detected.[56] However, many of the detected compounds detected in the mixture were only present at parts per million level, far from reaching a synthetically useful concentration.

The classical approaches to purine and pyrimidine synthesis have been based on simple but reactive prebiotically plausible feedstocks such as cyanide, cyanoacetylene, cyanates, as well as their hydrolyzed derivatives.[5] These reactive small molecules could have been generated by meteoritic impacts,[9] delivered by extraterrestrial bodies[57,58] or from geochemical reactions of the early earth.[59] It has been first demonstrated by Oro, and Miller that cyanides could afford adenine[60] and guanine[61] in low yields, presumably *via* polymeric HCN conjugates[62] such as 4-amino-5-cyanoimidazole (AICN).[63] On the other hand, condensation of cyanoacetaldehyde and urea (hydrolysis products of cyanoacetylene and cyanamide) could afford pyrimidines in moderate yield.[64] Ice-water eutectic[65] conditions favor the formation of purine, while pyrimidine formations require concentrated reagent conditions.[66] Alternatively, formamide can also be used as a starting material for the formation of many of these nucleobases, when strongly heated in the presence of mineral catalysts (such as montmorillonite).[67] However, due to the innate reactivity of many of the starting materials and ensuing intermediates under aqueous

conditions (hydrolysis), the four canonical nucleobases are far from being the only products in these reactions. Numerous analogues of the four canonical bases were detected, many of them having structural scaffolds that were expected to display significant base pairing capabilities with the four canonical bases.[6] These numerous non-biogenic nucleobases have raised questions whether the first biopolymer to evolve is RNA[68] (the notion of pre-RNA world; *vide infra*). However, recent experimental evidence suggests that noncanonical ribonucleosides with modified nucleobases[69] or aberrant nucleoside anomeric stereochemistry (α instead β-anomer)[70] have decreased photo-stability under strong UV irradiation compared to their canonical counterparts, thus providing a strong chemical selection towards the canonical RNA structure.

For decades, the prebiotic synthesis of ribose was assumed to proceed *via* the formose reaction (Butlerow synthesis).[71] Formaldehyde as a starting material reacts with itself to afford glycoaldehyde[72] in the presence of UV,[73] or heterogeneous or homogeneous metal cation catalysts under aqueous alkaline conditions.[74] The glycoaldehyde product can then autocatalyze its own formation, and can homologate to form a complex mixture composing of various stereochemically related trioses, erythroses, pentoses and hexoses in both aldose, ketose, and their branched isomeric forms.[6] Ribose, the integral chemical component of ribonucleotides, is formed with a miniscule yield of less than 1%.[75] There exist strategies which steer this reaction network towards the synthesis of pentoses (of which ribose is a member). For one, phosphorylated glycoaldehyde and glyceraldehyde could be used;[76] alternatively, borate minerals[77] are known to stabilize pentoses, while silicate minerals[78] stabilize carbohydrates with four to six carbons. However, to date there are no formose-reaction based strategies that give

ribose as a major product. Even if ribose sugars are formed in high yield, they decompose very rapidly on the geological timescale, especially under high temperature conditions, wherein the half-life of ribose at 100°C and pH 7 is on the scale of hours and 44 years at 0°C and pH 7.[79] Stabilization of ribose has been demonstrated on borate and silicate minerals; alternatively, cyanamide could selectively react with ribose in a complex monosaccharide mixture of stereochemically-related pentoses and hexoses, under moderate temperature and basicity, to give 5,5-fused bicyclic ribo-aminooxazoline which selectively crystalizes out of the aqueous sugar mixture; this bicyclic ribose-cyanamide product also decomposes more slowly under the formation conditions than ribose alone.[80] However, the robustness of these stabilization strategies remains to be probed upon stronger heating or more basic conditions.

Finally, even in the most optimistic scenario wherein all these problems are satisfactorily solved, ribosylation reactions involving purines and pyrimidines to form complete nucleosides have been problematic.[13] Among the nucleobases, only purines such as adenine and guanine could be condensed with ribose and, even then, at poor yields (<10%).[81] For pyrimidines, direct ribosylation did not yield any useful nucleoside products. Perhaps even more importantly, direct condensation of nucleobases with ribose will afford products with both α- and β-anomers, as well as sugar rings in both furanose and pyranose configurations.[82]

Taking all these synthetic challenges into consideration, when these reactions are required to occur sequentially with minimal sequestration or “purification” mechanisms to afford the prebiotic molecule of interest, these reactions quickly expand into an intertwined and irresolvable network of related reactions, giving inextricably complex

products with very little means of purification.[83] These difficulties have inspired ongoing efforts to find more effective and prebiotically plausible synthetic pathways for the preparations of ribonucleotides.

1.5 Brief Comment on the pre-RNA world

In order to circumvent the various aforementioned problems posed by the "RNA-first" model, some believe that a simpler chemical polymer arose prior to RNA; thus, the central tenet of the "RNA-later" model is that RNA is not the original but the evolutionary outcome of another genetic polymer.[14–16] Numerous informational macromolecules have been studied under this context; it was hypothesized that the pre-RNA polymer would have the following characteristics:[84]

1) phenotype and genotype both reside in the same polymer, such that catalytic functions can evolve during repeated rounds of replication;
2) the chemical synthesis of the alternative polymer should be compatible with the prebiotic chemistry of the early Earth, ideally simpler than RNA;
3) be capable of promoting template copying by sequence-dependent, nonenzymatic polymerization
4) be able to undergo specific and high affinity base-pairing recognition with RNA, so to facilitate genetic handover at a later stage of RNA evolution.

Among these posited alternate genetic polymers include ones that have backbone structures that are different, sometimes to a dramatic extent, from the canonical ribose-phosphate backbone.[16] For example, the furanose ring of RNA has been substituted by other carbohydrate rings, such as threose,[85] pyranosyl-(deoxy)ribose, β-allopyranosyl-ribose, locked ribose or hexitol. These non-ribose carbohydrate backbones may either

have hydroxyl or amino groups for enhanced reactivity in nonenzymatic copying.[86,87] Alternatively, the ribose ring may be deconstructed altogether and be substituted with flexible alkyl chains to afford flexible nucleic acid (FNA) or glycol nucleic acid (GNA).[16] The negatively charged phosphate linkages could be replaced as well, by either a negatively-charged glyoxylate or aspartate backbone, or neutral peptide-, sulfone-, acetal-, or imine-linkages.[16] Finally, the nucleobases (the recognition unit) may deviate from the canonical purine and pyrimidine nucleobases. Reported examples of recognition units include purines or pyrimidines with alternate atomic connectivity or substitution patterns, or nucleobases could even be entirely replaced with other prebiotically plausible heteroarenes, such as triaminopyrimidines, barbiturates, or cyanuric acids.[68]

From an evolutionary standpoint, if a pre-RNA polymer did indeed exist at some point, the most important properties of that polymer will be the ability to catalyze their own replication, as well as the ability to transfer their genetic content to RNA at a later stage (genetic handover). In this regard, threofuranosyl nucleic acid (TNA)[85] has been investigated as a potential RNA progenitor, and it has been demonstrated by Szostak and coworkers a decade ago that TNA and DNA can be copied into each other by protein enzymes, albeit inefficiently.[88,89] More efficient DNA-dependent TNA polymerase and TNA-dependent reverse transcriptase enzymes have since been developed.[90,91] On the other hand, Holliger and colleagues have expanded the scope of DNA-templated XNA synthesis and XNA reverse transcription to HNA, CeNA, ANA, FANA and LNA by means of engineered protein polymerases. In the same report, Holliger and coworkers demonstrated that HNA aptamers could be *in vitro* selected and evolved to display ligand binding capabilities.[92] A recent paper from Holliger and colleagues demonstrated the

selection of XNAzymes that can cleave and ligate RNA oligomers; thus, transfer of genetic information from XNA to RNA is possible without the involvement of extensively-evolved protein catalyst.[93] The remaining question, therefore, is whether XNAzymes can replicate themselves in a fashion similar to ribozymes such as R18 or tC19Z. Towards this end, in the very same work, Holliger and coworkers were able to evolve an XNA-based *trans*-acting ligase which could join two XNA-fragments together, thereby demonstrating a promising proof-of-concept example of how XNA can possess both phenotypic and genotypic functions as a self-contained system.[93]

While these XNAs display interesting chemical and catalytic properties, it is important to bear in mind that they are not present in extant biology. Additionally, with the possible exception of TNA, none of these XNAs have synthetic pathways more facile and prebiotically plausible than those of RNA. Perhaps these XNAs never preceded the RNA world – implying that RNA was the first biogenic polymer – or perhaps an XNA world did exist, yet RNA has completely displaced XNA at a later stage of evolution because of its functional superiority, and XNA was eliminated from modern biology without a single trace in any catalogued organisms. The question of *how* biology could lead to the complete extinction of a genetic material remains a key question of the pre-RNA world scenario. On the other hand, should an XNA-based protocell be possible – implying that XNA could both encode and catalyze intergenerational transfer of genetic information – what are the possible selection pressures which drove the protocell to ultimately adopt RNA as its genome? What are the possible functional superiorities of RNA over XNA, such that survival or catalytic advantages are conferred to the protocell during each of the evolutionary steps towards RNA-based genome?[9] It is important to realize that the

evolution of an XNA world merely delays the necessary evolution of chemistries for a prebiotically plausible and high-yielding synthesis of RNA monomers and polymers. Until these questions are satisfactorily answered, the “RNA first” scenario will remain the most straight forward solution towards the chemical origins of life.

1.6 Recent Progress on monomer synthesis

The discouragingly long list of chemical challenges posed by prebiotic nucleotide synthesis led many to conclude that the *de novo* synthesis of RNA during the inception of cellular life was perhaps impossible. Orgel once famously proclaimed that "the abiotic synthesis of RNA is so difficult that that it is unclear that the RNA world could have evolved *de novo* on the primitive earth".[13] However, as of the last decade, there have been multiple seminal advances which elegantly address some of the aforementioned concerns. The most notable of all are the innovative chemistries leading to pyrimidine and purine nucleosides; yields and regio- and stereospecificities are high when compared to previously reported chemistries. Equally important is the progress made in the activation of nucleotides, especially in the context of nonenzymatic RNA synthesis. Taking these advances into consideration, there is now renewed optimism that the nonenzymatic synthesis of monoribonucleotides, and subsequently oligoribonucleotides, may not be as far-fetched a goal as it once seemed. Due to the fact that excellent reviews have been published on this subject, we will only offer a brief summary of the recent progress achieved by Sutherland, Powner and Carrell and their coworkers revolving around the challenges of nucleotide synthesis.

As explained in the previous sections, the traditional retrosynthetic strategies developed for nucleotide synthesis are problematic in terms of product complexity and

low yield. In 2009, Sutherland and coworkers reported an elegant synthesis of cytidine and uridine nucleotides which did not involve formation of free nucleobase and ribose sugar, and their subsequent condensation.[70] Their alternate synthetic strategy utilizes glycoaldehyde, glyceraldehyde, cyanamide, cyanoacetylene, and phosphate; all these compounds are available either from samples of Murchison meteriorite (albeit in low concentration),[58] yet more recent work addresses the possible synthesis of these starting materials on Earth's surface.[9] Cyanamide first condenses with glycoaldehyde to give 2-aminooxazole, which then condenses with glyceraldehyde in the presence of concentrated phosphate buffer at pH 7 to yield 5,5-fused pentose aminooxazoline, with either arabino- or ribo- stereochemistry. The arabino-pentose aminooxazoline was reacted further with cyanoacetylene in the presence of concentrated pH 7 phosphate buffer to afford a 5,5,6-fused tricyclic anhydroarabinocytidine. Selective phosphorylation of the 3′-hydroxyl group in the presence of concentrated urea (*not* the more reactive 5′-hydroxyl, due to a subtle stereoelectronic effect[94]) and nucleophilic cleavage of the C-O bond connecting the pentose and the nucleoside then afford β-cytidine-2′-3′-cyclic phosphate in unprecedented overall yield and regioselectivity. β-cytidine-2′-3′-cyclic phosphate could then be hydrolyzed to β-uridine-2′-3′-cyclic phosphate photochemically in the presence of 254 nm UV light. Note that the same photochemical irradiation of products derived from the stereochemically incorrect anhydroribocytidine gives products that are suscepitible to photodegradation, thus guaranteeing the stereoselectivity of the overall synthetic pathway.[70]

Recently, Powner and coworkers have expanded this strategy towards the synthesis of noncanonical purine nucleosides, 8-oxoadenosine and 8-oxoinosine.[95] The

synthetic pathway is mostly analogous to the one reported for cytidine and uridine, but instead of the initial construction of 2-aminooxazole, glycoaldehyde was first condensed with thiocyanate – a chemical feedstock reported from geothermal springs[96] – to afford 2-thiooxazole. The exocyclic sulfur of thiooxazole was then regioselectively activated by cyanoacetylene, followed by the nucleophilic displacement of the derivatized sulfur by HCN trimer or its hydrolyzed primary amide product to afford the 5,5,5-fused tricylic aminoimidazole. The tricyclic aminoimidazole was formylated by a formamidine/formamide solution at 100°C to furnish the 5,6-fused purine rings of 8-oxoinosine and 8-oxoadenosine. Subsequent phosphorylation and nucleophilic C2′-purine 8-oxygen bond cleavage then afford the corresponding nucleoside-2′,3′-cyclic phosphates. Unfortunately, a prebiotically-plausible pathway which converts these 8-oxo-derivatized purine ribonucleotides to adenosine and guanosine has not yet been reported, thus further work remains warranted.

The success of both of these chemistries depend critically upon the purity of the starting material and the specific pH conditions under which the reaction is conducted. For example, it has been noted that a high yielding synthesis of bicyclic aminooxazoline is critically dependent on the sequential reaction and spatial resolution of cyanamide, glycoaldehyde, the resultant 2-aminooxazole and glyceraldehyde. Should these molecules be mixed together in the same pool without appropriate temporal and spatial resolution, the reaction mixture quickly descends into a clutter of isomers and homologues of the desired 2-aminooxazole and 5,5-fused aminooxazoline products. Thus, some have expressed concerns over the prebiotic relevance of this chemistry, while other dismissed this chemistry as bearing an excessive degree of "*deux ex machina*."[97]

This criticism has recently been addressed by Powner and coworkers, who recently demonstrated synthesis of the critical bicyclic aminooxazoline intermediate in a highly heterogeneous pool of sugars.[98] It was found that glycoaldehyde and glyceraldehyde react with 2-aminothiazole to form their corresponding heterocyclic aminal products, which crystallize in aqueous solution with dramatically different rates. Taking advantage of such time-resolved crystallization of glycoaldehyde- and glyceraldehyde-aminal, Powner demonstrated that purification and physical separation of the two substrates is possible. The now spatially-resolved glycoaldehyde and glyceraldehyde could react with cyanamide and 2-aminooxazole separately to afford the key bicyclic aminooxazoline intermediate in high yield.

How could the necessary glycoaldehyde and glyceraldehyde starting materials selectively be afforded without forming the clutter resultant from alkaline formose chemistry (*vide supra*)? Sutherland and coworkers reported that both of these sugars could be prepared from a mixture of hydrogen cyanide (HCN) and copper cyanide in the presence of 254 nm ultraviolet light.[99] The authors demonstrate that a mixture of cyanocuprates and excess cyanide ions could initiate a catalytic Cu(I)-Cu(II) photoredox cycle, in which the solvated electrons thus produced could then reduce hydrogen cyanide to methanimine, which hydrolyzed in water to form ammonia and formaldehyde. Glycoaldehyde, and subsequently glyceraldehyde, was formed by one-carbon homologation of formaldehyde *via* a Kiliani–Fischer type mechanism, using cyanide as the carbon donor, followed by nitrile reduction. Remarkably, in the presence of hydrogen sulfide as the terminal electron sink, the same experimental conditions also created precursors for some amino acids and lipids, such as acetaldehyde, acetone and

glycerol.[19] Todd and coworkers have recently shown that such a synthetic process remains plausible in a broad range of UV wavelengths, thus supporting the role of this chemistry in a wide range of atmospheric scenarios.[100]

There are two potential concerns with this elegant chemistry that are not to be overlooked. First, the availability of copper (I) ions in the prebiotic scenario remains a matter of debate. Secondly, the reliance of this synthetic pathway on highly concentrated, and sometimes freshly prepared, reagent solutions (dissolved hydrogen sulfide, cyanamide, cyanoacetylene, etc.) remains a concern. To illustrate this concern, it has been noted that cyanoacetylene quickly degrades into an intractable tar following its synthesis under aqueous environment,[97] and that efficient copper-catalyzed reductive synthesis of glycoaldehyde from cyanide requires a concentrated (>100 mM) stock of cyanide substrates.[99] Very recently, Sutherland and colleagues reported an alternate reductive cyanide homologation strategy which, instead of using copper(I) salts as catalysts and hydrogen sulfide as stoichiometric reductants, utilizes aqueous ferrocyanide and sulfite anions as the catalyst and stoichiometric reductant, respectively. The appeal of this newfound protocol is two-fold. First, the authors report a higher yield of 68% total glycoaldehyde synthesized (as cyanohydrin and sulfite adducts of glycoaldehyde) when the ferrocyanide-sulfite catalytic system is used in lieu of the copper(I)-hydrogen sulfide system (42% yield). (Xu et al., *Chem. Comm.*, in press). Second, both ferrocyanide and sulfite are more readily available than copper(I) cation and hydrogen sulfide: Ferrocyanide, a water-soluble anion readily available from aqueous ferrous ions, is known to sequester atmospheric HCN with ease,[101] and cyanide could potentially be recovered from ferrocyanide by geothermal processing of sodium or

potassium ferrocyanide salts.[102] On the other hand, soluble sulfite salts are the product of volcanic sulfur dioxide which disproportionate in aqueous media.[103] That being said, the requirement of concentrated phosphate for the phosphorylation of nucleotides in Sutherland's proposed nucleotide synthesis pathway remains a major obstacle.

In parallel to the efforts of Sutherland, Powner and coworkers, Carrell and coworkers have explored a different strategy for the synthesis of purine nucleosides[104] and, more recently, of their noncanonical analogues under wet-dry conditions.[105] Their first retrosynthetic step follows the traditional nucleobase – sugar glycosidic bond disconnection, yet Carell and coworkers used tri- or tetraamino substituted pyrimidines in lieu of the regular 5,6-fused purines for subsequent ribosylations. Using 4,5,6-triaminopyrimidine as an example, the exocyclic amino group at C5 was first regioselectively formylated in neat formic acid under reflux temperature (100 °C). With the exocyclic pyrimidine C5 amino group protected, the exocyclic amino groups at pyrimidine C4 and C6 positions – the masked purine N9 nitrogen – are now identical because they are correlated by C2 symmetry. Under mildly basic conditions (in the presence of borax, or alkyl amines) imine formation involving the pyrimidine exocyclic amines with excess ribose ensued, followed by intramolecular ribose cyclization and purine imidazole ring closure and aromatization to afford purine furanosides and pyranosides as both the α- and β-anomers. This chemistry, while ingeniously taking advantage of the symmetry and enhanced nucleophilicity of aminopyrimidines to side step the regioselectivity problem of direct purine ribosylation, possesses a few drawbacks. First, the C5 formylation of aminopyrimidine requires neat formic acid or formamide, but the prebiotic plausibility of having pools of neat formic acid and formamide seems low,

especially in light of the fact that these reagents are very hygroscopic. Small amounts of concentrated formamide solution (>85%) could be prepared under laboratory settings, but only under highly circumstantial conditions.[106] Second, the regioselectivity advantage brought by regioselective N9 ribosylation is quickly diminished in the subsequent ribose cyclization steps, wherein the formation of non-biogenic pyranoside and α-anomeric side products is severe.

1.7 Nonenzymatic RNA Polymerization

In order for the RNA world to be kickstarted, there must be a mechanism by which a library of RNA strands with different sequences formed spontaneously on the primitive Earth. These random sequences might then be the starting pool for the subsequent evolution of useful ribozymes, including catalysts able to support self-replication of RNA, eventually leading to an RNA polymerase.

Nonenzymatic polymerization of unactivated nucleotides under aqueous conditions does not occur spontaneously because the process is thermodynamically unfavorable.[107] The most conceptually simple reaction for nucleotide polymerization is the intermolecular dehydration/condensation reaction between the ribose (2′-) 3′-OH group and the 5′-phosphate group, with concomitant release of a water molecule. By subjecting acidic solutions of 2′(3′)-nucleotides under evaporation conditions to elevated temperature, the formation of complex mixtures of short oligonucleotides was detected, abut with regioselectivity of 2′-5′- vs. 3′-5′-phosphodiester linkages.[108] Deamer and coworkers have studied whether lipid encapsulation could enhance the formation of longer chain products from monoribonucleotides. Indeed, by encapsulating nucleotide monophosphates inside phosphatidylcholine-based lipid vesicles and subjecting such to

repeated wet-dry cycles at elevated temperatures, a broad band of RNA-like polymers, with lengths ranging from 20-mer to 50-mer, could be obtained.[109] It was hypothesized that nucleotides were trapped in between the lipidic lamelle as the vesicles were dehydrated,[110] thus forming a highly concentrated layer of mononucleotides that promoted RNA synthesis.[111] However, it was later determined that the RNA-like polymer thus formed contained a large number of depurinated and depyrimidinated products.[112] Thus, until now, oligonucleotide polymerization from direct monophosphate condensation remains a challenge.

Consequently, attempts to polymerize nucleotides from aqueous solution have generally used chemically activated nucleotides. Nucleotide 5′-triphosphates have been the chemically activated nucleotide of choice in extant enzyme-driven oligonucleotide polymerization. However, these molecules are exceptionally unreactive in solution because of the diffuse negative charge surrounding the thermodynamically-labile pyrophosphate linkages, thus precluding nucleophiles from nucleophilic attack with concomitant pyrophosphate release. Nonenzymatic polymerization utilizing polyphosphate or triphosphate-activated mononucleotides has proven to be unproductive. Only short oligonucleotide fragments are afforded, and the yield of polynucleotide thus formed tends to be poor.[113] Even when amino-substituted activated ribonucleotides are polymerized, only dimers and trimers are formed.[114]

The most promising strategies reported thus far tend to use ribonucleotides chemically pre-activated ribonucleotides with heterocycles such as imidazoles,[115] 1-methyladenines,[116] and electron rich pyridines[117] as the leaving group. Metal ion catalysis has proven to be indispensable in enhancing the reactivity of the activated monomers

towards polymerization. In particular, nucleotide phosphoroimidazolides require divalent metal ions such as lead (Pb^{2+})[118] or uranyl (UO_2^{2+}) ions[119,120] in order for efficient oligonucleotide polymerization to occur in solution phase. However, it is noteworthy that the resultant oligomers thus formed tend to be short, and are predominantly 2′-5′-linked.

Monomer sequestration and concentration in confined spaces appeared to be another promising strategy for the polymerization of long chain oligonucleotides from activated monomers. In one example, Deamer and coworkers demonstrated that long homo- and heteronucleotide polymers could be obtained if polymerization is carried out in the presence of Mg^{2+} and Pb^{2+} in ice-water eutectic solution.[121,122] Alternatively, Montmorillonite, a volcanic ash derived aluminosilicate mineral, could be used as a catalyst. Polymers thus derived could be up to 50 nucleotides long, depending on the activating group and experimental conditions used.[123,124] Montmorillonite likely catalyzes mononucleotide polymerization by allowing activated monomers to enter the interstitial spaces in between the aluminosilicate layers.[125] Montmorillonite is known to steer the regiochemistry of phosphodiester bind linkage towards the biogenic 3′-5′-linkages, but the effect is highly leaving group and sequence dependent.[116,126]

1.8 Nonenzymatic Template-directed RNA Copying

Since the 1960s, single-stranded RNA has been known to facilitate the assembly of short complementary oligonucleotide strands from activated mononucleotides, at room temperature, in the absence of any enzymes. Such template-directed nonenzymatic synthesis, in many cases involving a low-melting double- or triple-helical complex,[127] brings complementary activated monomers or short oligomers together and facilitates their ligation or polymerization. Naylor and Gilham reported perhaps the earliest example

of such experiments when they demonstrated the ligation of two hexathymidine strands across a dodecaadenosine template stand, using 1-ethyl-3-(dimethylamino)carbodiimide (EDC) as the condensing reagent.[128] Later, Sulston et al. demonstrated the polymerization of monomers across complementary homonucleotide templates using EDC, affording complex mixtures of oligonucleotides with heterogeneous phosphodiester backbones.[129–131] Shabarova and coworkers extended these early results by using different deoxy-templates, deoxy-substrates, and activating agents.[132] Cyanogen bromide,[133,134] as well as *N*-cyanoimidazole,[135] have also been used as activating reagents for the ligation of oligonucleotides under a wide variety of reaction conditions.

On the other hand, the favored reactive nucleotides of extant biology - nucleoside-5′-triphosphates – proved to be suitable activated substrates only under enzyme-catalyzed conditions, wherein divalent metal ions are strategically placed to neutralize the negative charge of the leaving group such that subsequent nucleophilic attack can ensue.[136] Additionally, the primer 3′-OH nucleophile is often conformationally restrained by the enzyme active site such that the nucleophile is in close proximity to the antibonding orbital of the leaving group. For example, Yang and coworkers have used time-resolved X-ray crystallographic structures to demonstrate that human DNA polymerase eta requires at least two[137] and possibly three[138] Mg^{2+} ions for efficient polymerization of deoxynucleotide triphosphates. Negatively-charged aspartate and glutamate residues within the catalytic center chelate the divalent magnesium ions, such that they can neutralize the negative charge of the activated triphosphate and facilitate subsequent primer extension.[137] Likewise, the active site of the Class I ligase ribozyme involves at least two catalytic Mg^{2+} ions for ligation to occur efficiently.[139] In both cases, the primer

3′-OH is almost antiperiplanar (180° away) to the antibonding orbital of the labile pyrophosphate P-O linkage between the α- and β-phosphates of the NTP. In the absence of enzymatic catalysis, triphosphate activated mono- or oligonucleotides react too slowly. For instance, template-directed ligation of 5′-triphosphate-capped oligonucleotides displayed extremely slow kinetics, even when the primer and the downstream oligonucleotide substrate are structurally preorganized by the complementary template.[140] Thus, in order for triphosphate activated mono- or oligonucleotides to be useful substrates for nonenzymatic RNA copying, the key chemical challenge is design strategies such that either a chelated magnesium is “delivered” to and held close to the reaction site (increasing the effective concentration of Mg^{2+} to a level on par of that seen in biological catalysis), or that a strongly cationic charged peptide sequence could bind to the pyrophosphate and facilitate its displacement.

Another approach to nonenzymatic RNA copying is to use nucleotides that are chemically activated by imidazoles. Imidazoles have been proposed as prebiotically plausible because imidazoles can be synthesized in high yield from very simple organic feedstocks such as ammonia, formaldehyde and glyoxal[141] under aqueous conditions, and that the activated phosphoroimidazolides could be prepared from either nucleotide 5′-polyphosphates[142] or 5′-monophosphoramidate under wet-dry cycles at slightly elevated temperatures in the presence of urea. Nonenzymatic RNA copying utilizing nucleotide phosphoroimidazolide substrates was first reported by Orgel and coworkers in 1968, following the demonstration that nucleotide phosphoroimidazolides can polymerize in a template-directed manner into short oligonucleotide strands.[115] It was found that facile phosphodiester bond formation required divalent metal cations; in their absence,

polymerization is at least 100-fold slower, phosphodiester bond regioselectivity is poor, and the resultant oligonucleotide products are very short. For long-chain polyguanosine to be formed on a polycytidine template, a cocatalyst combination of Mg^{2+} and either Pb^{2+} (favoring 2′-5′ regioselectivity),[143] or Zn^{2+} is required (favoring 3′-5′ regioselectivity).[144]

Substitution patterns on the imidazole leaving group dramatically alter the rate of nonenzymatic RNA copying. It was found that methyl group substitution on the 2-position of imidazole improves the rate of nonenzymatic RNA copying and the 3′-5′-regioselectivity of phosphodiester backbone, and that the copying reaction requires only Mg^{2+} as the cocatalyst.[145] Tiny changes to this “privileged scaffold” – such as one carbon homologation of the methyl group, or changing the methyl group substitution pattern,[145] or altering the sterics of the leaving group (*vide infra*), strongly impacts the catalytic effect. A prebiotically plausible imidazole, 4-amino-5-cyanoimidazole (AICN), has been used to activate nucleotides, but the rate of nonenzymatic RNA copying is slow. (Prywes, unpublished). It was also noticed that the identity of the leaving group of the downstream stacking activated monomer/oligomer had a dramatic effect on polymerization efficiency: 2-methylimidazole activated monomer catalyzes primer extension much more efficiently than the imidazole activated variant. Finally, the rate of 2-MeImpN-driven nonenzymatic template copying depends on the structure of the template. Orgel and coworkers demonstrated that template strands that favor an overall A-form duplex – wherein the mononucleotides adopt the 3′-*endo* conformation and the 3′-hydroxyl is modelled to point directly to the downstream 5′-activated phosphates – facilitate nonenzymatic copying. For instance, template copying using HNA and RNA as templates – both of which favor an A-form duplex – display faster rates than a DNA template.[146] In line with observations

reported by Orgel, we used Transferred Nuclear Overhauser Effect Spectroscopy (TrNOESY) methods to demonstrate that during nonenzymatic RNA copying, the ribose pucker of activated monoribonucleotides undergoes structural rearrangement from the native (unbound) C2′-*endo* to the C3′-*endo* state, allowing the 3′-OH hydroxyl group to point directly towards the activated phosphate of the downstream monomer.[147]

For nonenzymatic RNA copying, the incorporation of guanosine and cytidine residues into the extending primer proceeds with satisfactory rates,[148] yet incorporation of uridine across templating adenosine is extremely challenging:[149] the rate of uridine incorporation is on par with the rate of RNA degradation. Incorporation of adenosine across uridine is of intermediary efficiency, yet guanosine-uridine[149] or cytidine-(protonated) adenosine[150] wobble pairing becomes a significant issue in a pool of mixed activated mononucleotides. Thus, adenosine and uridine-rich templates are difficult sequences to copy effectively. In this regard, 2-thiouridine or 2-thioribothymidine can substitute for uridine either as a more reactive activated monomer, or as templating residues in primer extensions: these noncanonical, sulfur substituted nucleotides display enhanced kinetics of primer extension. The fidelity and the mean lengths of nonenzymatic primer extension products on mixed-sequence RNA templates are also improved.[151]

Replacing uridine with 2-thiouridine or 2-thioribothymidine has resulted in similar improvements in other experimental systems, e.g. for NP-DNA, where the ribose-phosphodiester backbone has been replaced with 3′-amino-2′,3′-dideoxyribose-phosphoramidate backbone,[87] or when RNA primer extension is driven by a template-bound b1-233t ribozyme.[152] It is worth noting that the b1-233t ribozyme retained partial catalytic function even when all the uridine nucleotides were replaced with 2-thiouridine

(although a slightly lower k_{cat} is observed); also, 2-thiouridine could be converted to the canonical uridine under oxidative conditions. Therefore, these thiolated ribonucleotides could potentially pave the way towards future evolution of complex catalytic ribozymes, followed by further evolution towards the advent of uridine-based ribozymes.

Another approach to efficient nonenzymatic RNA copying is to use helper oligonucleotides to assist with monomer binding. The Richert group has pioneered the usage of complementary "helper" oligonucleotides that bind downstream of the extending monomer to improve the yield of primer extension.[153] In the initial reports by Richert and coworkers, these ten-nucleotide-long "helper" RNAs allowed facile addition of thymidine monomers to the extending primer strand. Richert and coworkers have further developed this idea to achieve high-yielding mixed-sequence template copying by utilizing "microhelper" RNAs – tri- or tetranucleotides that could well be the products of randomized monomer polymerization – that bind downstream of the incoming monomer.[154] By immobilizing the primer-template duplex and periodically replacing the activated monomer and microhelper, Richert and coworkers achieved four consecutive template copying steps at near quantitative yield on a template containing all four nucleotides, despite the fact that complete reaction took almost a month.[154] By utilizing activated trinucleotide microhelpers in lieu of the unactivated ones, the rate of primer extension is dramatically enhanced, while achieving a comparable conversion.[155] For a four-nucleotide, mixed sequence primer extension, it took us about 48 hours to achieve conversion comparable to that reported by Richert and coworkers, instead of a month; we were also able to demonstrate the partial synthesis of the hammerhead ribozyme strand in 6% yield, *via* 12 consecutive extension steps, in ~6 days.

How do activated microhelper oligonucleotides confer their catalytic effect? We have recently looked into this question from a thermodynamic standpoint and have concluded that the activated microhelpers most likely catalyze upstream primer extensions by a combination of thermodynamic and kinetic effects.[156] We found that, in a primer-template-helper complex wherein the GMP binds in between the 3′-terminus of the primer and the 5′-terminus of the helper, the presence of a stably-bound downstream oligonucleotide could dramatically increase the affinity of GMP binding by >100 fold, reaching a K_a in the mid 10^3 M^{-1} range instead of the usual the mid-teens M^{-1} range in the absence of a downstream helper oligonucleotide, most likely by affording additional base stacking for the incoming monomer. Nonetheless, the presence of downstream *un*activated oligomers only modestly accelerates upstream primer extensions by ~10 fold. We conclude, therefore, that the large degree of primer-extension rate enhancement conferred by *activated* short oligonucleotides, which typically exceeds 100 fold,[155] must be a combined effect of thermodynamic (enhanced monomer binding affinity) and kinetic effects, originating from the leaving group-leaving group interactions.[148,149] Note that these activated oligonucleotides, while capable of being excellent catalysts, are poor substrates themselves: despite that they form very stable A-form duplex with the template, they typically exhibit primer ligation rates that are >100 fold slower than that driven by consecutively-bound activated monomers.[157]

2-methylimidazole has been the leaving group of choice for monoribonucleotide activation since Orgel's seminal report in 1981[145], yet the pursuit of an activating group imparting better reactivity than 2-methylimidazole has continued. To that end, we have studied the structure-reactivity relationship of the imidazole leaving group in

nonenzymatic RNA copying. The chemical groups and substitution patterns were varied to generate nucleotide phosphoroimidazolides with a range of leaving group pK_a and steric bulkiness. The 2-substituted imidazoles are genuinely the "privileged scaffold" in nonenzymatic RNA replications: installation of groups bulkier than methyl groups on the imidazole ring, or changing the imidazole substitution pattern, all led to activated monomers that are less reactive than the phosphoro-2-methylimidazolide parent. Notably, we observed that the template-copying rate increased as the pK_a of the leaving group increased. In light of these chemical and steric requirements, we established 2-aminoimidazole-activated monoribonucleotides as a more reactive monoribonucleotide substrate for nonenzymatic RNA copying, owing to its adherence to the "privileged scaffold" and the electron-donation properties of the exocyclic amino group.[158] Using a combination of 2-aminoimidazole activated monomers and microhelper oligonucleotides, the overall yield of primer + UU extension (with AA template sequence) is now dramatically enhanced, achieving an overall yield of 92 % instead of the previously documented 10% when 2-methylimidazole monomers and microhelpers are used. In the same report, we successfully demonstrated primer extension by seven consecutive nucleotides, on a mixed sequence template containing consecutive A and U residues. The yield of each nucleotide incorporation now approaches >97 %.[158] Of note, 3′-amino-2′,3′-dideoxyribonucleotides that were activated with 2-aminoimidazole not only displayed enhanced kinetics over their 2-methylimidazole activated counterparts,[87] but also were more stable against monomer decomposition by intramolecular cyclization between the 3′-amino group and the activated phosphate.[159] Using the 2-aminoimidazole activated 3′-amino-2′,3′-dideoxyribo-nucleotides, one-pot mixed template copying of a 26-nt overhang

containing all four nucleotides is now achievable, without the use of activated microhelpers or reagent replacement. The yield of individual nucleotide extension can be as high as 96-97%. (O'Flaherty and Szostak, to be published) Thus, together with template immobilization or other orthogonal catalytic strategies, high-yielding mixed template copying of extended, ribozyme-length RNA or analogous oligonucleotide products is now within reach.

Given that 2-aminoimidazole-activated monoribonucleotides enhance the rate of nonenzymatic RNA copying, we wondered whether the synthesis of 2-aminoimidazole is facile under plausible prebiotic conditions. Towards this end, Fahrenbach et al. have recently found that coincubating 1M glycoaldehyde, cyanamide and ammonia (or ammonium chloride) for three hours at 60 °C affords both 2-aminoimidazole and 2-aminoxazole, which has recently been demonstrated as an integral heterocyclic compound *en route* to cytidine and uridine. It was also found that the ratio of 2-aminoimidazole to 2-aminooxazole could be varied by altering the concentration of ammonium chloride as well as the reaction pH. This common synthetic origin reveals a possible chemical link between nucleotide synthesis and activation.[160]

1.9 New understandings of the mechanism

In old models of nonenzymatic RNA copying chemistry, activated nucleotides or short oligonucleotides bind reversibly to cognate positions of a single-stranded template. The 3′-hydroxyl group of the priming oligomer then attacks the activated phosphate of the downstream activated monomer/oligomer *via* an in line S_N2 nucleophilic substitution to form a phosphodiester bond, concomitantly releasing a leaving group (typically 2-methylimidazole) in this process.[12] Multiple studies have used phosphoroimidazolide

hydrolysis as a mechanistic surrogate of the phosphodiester bond formation reaction. Earlier work by Kanavarioti et al. suggested that protonation of imidazole leaving group – forming a phosphoroimidazolide zwitterion as a result – is a necessary but insufficient condition for hydrolysis within pH 4-10.[161] To characterize the transition state of phosphoryl transfer reaction involving phosphoroimidazolides, we determined the leaving group kinetic isotope effect of phosphate-N15-labeled- and O18-labeled- guanosine phosphoroimidazolide; various guanosine phosphoroimidazolide derivatives involving methylated imidazoles have been prepared, the hydrolysis rates of these derivatives were plotted to the leaving group pK_a and fitted linearly to obtain the Bronsted plot. The observed KIE for the hydrolysis of guanosine phosphoroimidazolides relative to its phosphate N15 and O18 isotopologues, as well as the βLG value of the Bronsted plot, are strongly indicative of an S_N2-like concerted loose A_ND_N mechanism, without the involvement of pentavalent phosphorane intermediate. The transition state features a deprotonated phosphate and significant P-N bond fission with a weakly buckled (10° deviation from fully trigonal planar metaphosphate) phosphoryl group.[162]

As explained earlier, divalent metal cations enhance the rate of phosphodiester bond formation; but the mechanistic origin of divalent metal catalysis remains unclear. Using proton NMR spectroscopy, we have shown that Mg^{2+} does not interact significantly with the aromatic π electron cloud of 2-methylimidazole, thus it cannot mediate catalysis by altering the electron density of the leaving group. On the other hand, we have determined the K_M of Mg^{2+} and $Co(NH_3)_6^{3+}$ in primer extension reactions by plotting the rate of primer extension against ion concentration. Both fits are highly suggestive of single-site binding, while $Co(NH_3)_6^{3+}$ binding saturated at a lower concentration and

affords a lesser degree of catalysis than Mg^{2+}. Finally, the catalytic effect afforded by Mg^{2+} and $Co(NH_3)_6^{3+}$ is synergistic, which is indicative of distinct catalytic roles. Based on these data, we suggest that Mg^{2+} can catalyze primer extension in two possible ways: one is to neutralize the repulsive charge interactions between RNA phosphate backbone and the phosphoroimidazolides through outer shell ionic interactions, thus stabilizing monomer binding; the ability of Mg^{2+} to engage in direct inner shell contacts with ligands also suggests that it could coordinate the primer 3′-hydroxyl group and the activated phosphate simultaneously, thus forming the active site complex.[157] Whether Mg^{2+} catalyzes primer extension *via* one, or both proposed mechanistic pathways, is a matter for further investigation.

However, understanding the mechanistic details of the phosphodiester bond formation step only reveals part of the full mechanistic story. It was known that the presence of an activated downstream monomer or oligomer downstream of the extending monomer dramatically catalyzes the reaction.[148,149] This effect was initially believed to result from a noncovalent interaction between the leaving groups of adjacent nucleotides, such that the reactive site becomes preorganized for in-line displacement,[148] yet the precise chemical nature of such noncovalent interactions, be it π-π slip stacking between leaving groups, hydrogen bonding or cation-π interactions between a protonated and unprotonated leaving group, or simply steric and/or columbic interactions between adjacent leaving groups, remained unclear. Our laboratory sought to ascertain the precise mechanism for such noncovalent leaving group-leaving group contacts *via* X-ray crystallography. We recently developed a series of nonhydrolyzable chemical analogues of 2-MeImpG, in which the labile N–P bond is replaced with an unreactive C–P bond.

Pyrazolyl-[163] and imidazolylphosphonate[164] groups have been used as the pseudo-activating group of these analogues, and the two analogues were cocrystallized with rigidified RNA primer-template duplexes to afford high-resolution structural information.[163,164] We hoped that the high-resolution crystal structure would reveal the ground state conformation of the bound monomers relative to one another and with the RNA duplex, thus affording insights into the possible pathways of leaving group catalysis. In all the solved structures, if the RNA primer-template contains a single nucleotide overhang, the phosphate group of the extending monomer is typically 6 Å away from the primer 3′-OH group, which is too far for any bond formation to occur.[163] This is indicative of significant conformational rearrangement on the part of monomer preceding phosphodiester bond formation. Surprisingly, in all of the solved crystal structures, the leaving group mimicking moieties of the nonhydrolyzable analogues were severely disordered, and structural information pertaining to the leaving group could not be obtained. The observed disorder pertaining to the leaving group of the extending monomer was not improved by using a stably bound, 5′-imidazolylphosphonate capped complementary oligonucleotide downstream of the extension site.[164] Finally, it appears general that the presence of a sterically bulky group – be it heteroaryl phosphonates (for analogues) or even unactivated phosphate (for mononucleotides) – downstream of the extending monomer induces a closer approach of the “reacting” phosphate to the primer 3′-OH by as much as 2 Å. The same observation is obtained when the stacking monomer is replaced by a stably bound complementary oligonucleotide. Since both phosphates and the nonhydrolyzable phosphonates reorient the primer-adjacent phosphate group in very similar ways, any noncovalent interactions between consecutive leaving groups are

likely steric or Coulombic in nature, and the π system associated with leaving group is likely not involved in catalysis. It also suggests that the closer distance between primer 3′-OH and activated phosphate of the extending monomer in the presence of downstream mono- or oligonucleotide is not due to leaving group-leaving group interactions, but is merely a mechanistic byproduct of steric interactions.

Our laboratory has recently proposed a radically different mechanistic model to account for the leaving group catalytic effect. We observed that the rate of 2-MeImpG-driven extension on a -CCCC template was dependent on a few factors, namely the pH, concentration, and preincubation time of activated monomer stock solution, before the initiation of primer extension. It was found that the initial rate of primer extension is the fastest if the activated monomer stock solution was preincubated at pH 7 for roughly 20 min before the induction of stock solution to the primer extension mixture. The fact that a lengthy activated monomer preincubation was necessary to enhance the primer-extension rate suggested the involvement of a covalent intermediate. Since the pK_a of the 2-methylimdiazole leaving group is close to 7, it follows that half of the 2-MeImpG had a protonated leaving group (thus forming a zwitterionic 2-MeImpG), while half of the 2-MeImpG had an unprotonated 2-MeIm leaving group. Thus, it was hypothesized that the neutral leaving group of anionic 2-MeImpG could nucleophilically displace the protonated leaving group of the zwitterionic 2-MeImpG, thus forming a highly activated, imidazolium-bridged dinucleotide, which then bound with cognate templating positions as a 5′-5′-bridged dinucleotide, and then engaged in primer extension.[165] Richert and colleagues also noted the high reactivity of such intermediates.[166] This mechanism could also explain our prior observation that using nucleotide 2-methylpyrrolylphosphonate as a stacking

monomer abolishes the catalytic effect, since that the pyrrole group of the pyrrolylphosphonate analogue cannot form the 5′-5′-bridged dinucleotide due to insufficient nucleophilicity.[158] This also indicates that neither π-stacking nor hydrogen bonding interactions between neighbouring leaving groups are sufficient to catalyze primer extension.

The covalent catalysis mechanism has been validated by multiple follow-up studies from our laboratory. Zhang et al. used diguanosine-5′-triphosphate (GpppG) as a kinetically stable triphosphate-bridged analogue of the 5′-5′-imidazolium-bridged diguanosine dinucleotide, and demonstrated that the molecule could bind with a cognate -CC RNA template *via* exclusive Watson-Crick geometry and with high affinity. Additionally, a model of the actual 5′-5′-imidazolium-bridged diguanosine dinucleotide was docked into the RNA-GpppG crystal structure, and the nucleobase and sugar of the bridged dinucleotide could fit well within the corresponding GpppG observed electron density, indicating that imidazolium bridged dinucleotide could potentially bind to cognate RNA duplexes in a conformation similar to GpppG.[167] Later that year, using the chemically similar yet nonhydrolyzable imidazolyl phosphonate analog of 2-MeImpG, we conclusively demonstrated that the analogue could react with cytidine phosphoroimidazolide with very high chemo- and regioselectivity, resulting in the formation of a 5′-5′-imidazole-bridged GC dinucleotide product which contains a bridging imidazole flanked by both labile N–P and unreactive C–P bonds. This mixed linkage GC dinucleotide could give primer-extension rates comparable to that displayed by the corresponding monomers, but at a concentration 10-fold lower. These data thus recapitulated the prior data that the imidazole leaving group of 2-MeImpN is nucleophilic

enough to engage in covalent catalysis, and that the resultant bridged dinucleotide is more reactive.[168]

The most convincing evidence that is in direct agreement with our proposed covalent catalysis mechanism is very recent. By using pre-crystallized RNA-dGMP complexes and soaking such crystals in concentrated solutions of pre-formed 2-aminoimidazolium-bridged diguanosine dinucleotide, we were able to obtain a crystal structure of the RNA-duplex-bridged dinucleotide complex, with an overall resolution determined to 1.8 Å. The imidazolium-bridged dinucleotide is well-ordered in the crystal structure. The electron density map clearly showed that the bridging 2-aminoimidazolium bridge is flanked by two phosphoryl groups; unfortunately, the electron density corresponding to the 2-aminoimidazole bridge is tubular, suggesting that the conformation of the imidazolium bridge is not perfectly ordered, and that the imidazolium bridge could likely rotate along the axis formed in between the two flanking phosphates. Despite this observation, the exocyclic amino group of 2-aminoimidazolium bridge is likely hydrogen bonded to the to the nonbridging oxygens. Additionally, Zhang et al. were able to trace the progress of imidazolium bridge formation and subsequent phosphodiester bond formation, by incubating the pre-formed RNA-dGMP crystals in a concentrated solution containing both Mg^{2+} and activated monomer, and terminating the *in crystallo* polymerization at different time points throughout a total duration of 3 h. In the crystal structures, the 2′-OH group of the primer is at least 2Å further apart from the reactive phosphoryl group than the 3′-OH groups, explaining the enhanced 3′-5′-backbone regioselectivity if such a mechanism is at play.[169] This crystallographic observation is in direct agreement with recent biochemical studies performed by Giurgiu et al., who used

RNase digestion studies to elucidate the factors favoring 3′-5′-linkage during nonenzymatic primer extensions. It was found that the formation of imidazolium-bridged dinucleotides that bind strongly to the template in a Watson-Crick fashion favors the formation of 3′-5′-phosphodiester linkage during nonenzymatic copying.[170]

On top of the aforementioned crystallographic efforts, members of our laboratory have performed detailed kinetic and studies on 1) the rates of imidazolium-bridged dinucleotide formation, from its constitutive nucleotide phosphoroimidazolides, in both the forward and backward direction; 2) and 3) hydrolysis rates of imidazolium-bridged dinucleotide and nucleotide phosphoroimidazolides, 4) the K_M of imidazolium-bridged dinucleotide binding, and 5) the inhibitory effect of nucleotide phosphoroimidazolide monomers and hydrolyzed by-products. These determined rate constants, K_M and K_i values were then integrated into a coherent kinetic model, which satisfactorily accounts for the rate of primer extension at earlier time points of the primer extension reaction. As primer extension progresses, the actual observed extension rates are lower than the modelled rates, implying that there may be as yet-uncovered competing inhibitory processes that are poisoning the rate of extension.[171]

1.10 Unmet Challenges for nonenzymatic RNA copying

In 2012, we published a perspective on the eight significant challenges one must address to demonstrate robust and multiple-cycle nonenzymatic genomic replication within reproducing protocells.[172] Experimental solutions for some of these challenges, such as phosphodiester backbone regioselectivity, fidelity and rate of template copying chemistry, have been at least partially demonstrated. However, some of the challenges, as explained below, remained obscure and will require significant experimental work.

Reactivation chemistry. While the high reactivity of nucleotide phosphoroimidazolides is ideal for nonenzymatic template copying, it also leads to side reactions such as hydrolysis (forming nucleoside monophosphates), or, in the case of longer oligonucleotides, intramolecular cyclization (forming small, circular RNAs).[6] These competitive processes lower the concentration of activated monomers and oligomers for nonenzymatic RNA copying. This is especially problematic in that most of these side products, being able to base-pair with templates but not able to serve as substrates for the extending primer, could act as competitive inhibitors of nonenzymatic RNA replication. It is, therefore, important to have some mechanism to convert these hydrolyzed products back to their activated states. Activating agents such as EDC,[173] cyanogen bromide and *N*-cyanoimidazole (NCI),[174] and *N*-carboxyanhydrides[175] have been studied for their ability to perform *in situ* phosphate activation and oligonucleotide ligation, yet all these reagents could also nonspecifically conjugate to nucleobases and hydroxyl termini of RNA primers or templates, thus leading to premature chain termination.[176] A prebiotically plausible synthetic/reactivation pathway which chemoselectively targets phosphates to afford heterocycle-activated nucleotides is therefore needed. Towards this end, Fahrenbach and colleagues have devised an *in situ* activation method based on cyanide, imidazole and hypochlorous acid.[177] Based on prior literature that diimidazole imine could activate phosphate groups of oligo- and mononucleotides, they discovered that hypochlorous acid and cyanide could react to form cyanogen chloride – an activator chemically analogous to cyanogen bromide – which in turn react with imidazole under neutral medium to form the desired diimidazole imine activator, as well as trace amounts of *N*-cyanoimidazole byproducts. This aqueous activator mix could convert nucleotide

monophosphates to their corresponding nucleotide phosphoroimidazolides with an excellent yield of ~70% after one hour at pH 6.[177] While this protocol affords an exceptional yield of nucleotide phosphoroimidazolides, it is important to bear in mind that the concentration of molecular chlorine – the precursor of hypochlorous acid – on the prebiotic earth is likely limited by the geochemical pathways *via* which it could be generated. The authors proposed that high concentration of chloride salts could be radiolyzed with ionizing radiation to generate chlorine, but the robustness of hypochlorous acid generation and sequestration remains to be seen. Additionally, nonenzymatic RNA copying using nucleotide phosphoroimidazolide as substrates favor alkaline conditions. Thus, the slightly acidic environment under which nucleotide phosphoroimidazolides are generated is not the optimal environment for subsequent nonenzymatic RNA polymerization or copying chemistry. Fortunately, it has been reported that pH values can vary within the same geological environment (such as a surface lagoon), thus it remains possible that the nucleotide phosphoroimidazolides synthesized at an acidic microenvironment could be transported to an alkaline environment by simple flow or mixing.

Metal ions. High concentrations of divalent metal cations, such as Mg^{2+}, Zn^{2+}, and Mn^{2+} are required for nonenzymatic RNA chemistry to proceed with satisfactory rate (> 50 mM for Mg^{2+}, > 10 mM for Zn^{2+} and Mn^{2+}),[172] yet such concentrations to cause aggregation of fatty acids,[172] thus destabilizing the structural integrity of the primitive, fatty acid-based membranes. In the presence of a high concentration of divalent cation, the RNA phosphodiester backbone could also endo- or exonucleolytically degrade into their corresponding 2′-3′-cyclic phosphate terminated fragment and 5′-hydroxyl terminated

fragment *via* transesterification. Therefore, a prebiotically plausible mechanism which facilitates nonenzymatic RNA replication within a fatty acid vesicle at a low concentration of divalent metal cation is needed. Towards this end, we have previously reported that the presence of excess citrate as a magnesium chelator permits nonenzymatic RNA replication to proceed within vesicles (at a slower rate compared to the non-chelated reaction) without significant compromising the structural integrity of fatty-acid membranes, all while diminishing the side reaction of RNA strand cleavage.[178] While a prebiotically plausible synthesis of citrate is yet to be reported, and that citrate has not been discovered in geological sources, this advance suggests that prebiotically plausible small molecules could serve as protective "cofactors" and form a catalytic complex with divalent metal cations, which could then catalyze primer extension without the detrimental membrane- and RNA-destabilizing effects. On the other hand, it is important that the small molecule chelator neither chelate Mg^{2+} too strongly, nor saturate the six ligand sites of Mg^{2+}; this is evident from the observation that high-affinity polycarboxylate Mg^{2+} chelators, such as ethylenediaminetetraacetic acid (EDTA) and nitrilotriacetic acid (NTA) could protect fatty acid protocells from the destabilizing effects of Mg^{2+}, yet completely abolish the catalytic effect of Mg^{2+} in nonenzymatic RNA copying chemistry.[178] These ligands could potentially saturate all ligand sites for Mg^{2+} inner-sphere contact, thus abolishing the catalytic effect of the Mg^{2+} cation. An interesting possibility will be to develop peptide-based catalysts from prebiotically available amino acids that can bind with Mg^{2+} in a fashion reminiscent of citrate anion. These peptides could either be short, linear oligopeptides containing multiple aspartate or glutamate residues, which flank the active site of certain RNA polymerases and bind Mg^{2+} with high affinity. Alternatively, these oligopeptides could

cyclize to form macrocycles and encage the Mg^{2+} cation within the cyclic peptide[179] in a fashion similar to the potassium(I)-valinomycin complex.[180] Very recently, Bonfio et al. have demonstrated that the cysteine residues of four glutathione molecules could bind with ferrous ions to form a tetra-coordinated iron complex, which could undergo further redox chemistry to form iron-sulfur-clusters upon UV irradiation.[181] Apparently, peptide-metal interactions are prebiotically plausible and may be catalytically relevant, and hence warrant a more comprehensive study.

If a protocell were to originate in environments wherein the overall concentration of catalytic divalent cations was low, there must be a plausible mechanism wherein catalytic ions were sequestered and localized close to the site wherein phosphodiester bond formation occurs. In other words, the rate of nonenzymatic template copying chemistry could be enhanced by increasing the effective molarity of catalytic cations in the proximity of the bond formation site beyond a certain catalytic threshold. A plausible strategy for solving this problem is to develop a small molecule cofactor which could not only sequester aqueous divalent metal cations, but also place the cation near the RNA primer-template complex being copied. While metal ions could potentially be sequestered by carboxylate-rich ligands, how could they be localized close to RNA duplexes? Hud and coworkers have reported that nucleic acid intercalators, such as proflavine, could stack in between adenine-thymidine Watson Crick pairs nonspecifically, and the resultant polyA•polyT DNA duplexes are more stable.[182] Members of our laboratory are actively exploring the plausibility of using intercalator-chelator conjugates to enhance the rate of nonenzymatic RNA copying.

1.11 References

(1) Manhes, G.; Allègre, C. J.; Dupré, B.; Hamelin, B. *Earth Planet. Sci. Lett.* **1980**, *47*, 370–382.

(2) Mojzsis, S. J.; Arrhenius, G.; McKeegan, K. D.; Harrison, T. M.; Nutman, A. P.; Friend, C. R. L. *Nature* **1996**, *384*, 55–59.

(3) Dodd, M. S.; Papineau, D.; Grenne, T.; Slack, J. F.; Rittner, M.; Pirajno, F.; O'Neil, J.; Little, C. T. S. *Nature* **2017**, *543*, 60–64.

(4) Schopf, J.; Packer, B. *Science* **1987**, *237*, 70–73.

(5) Kitadai, N.; Maruyama, S. *Geosci. Front*. **2017**. ASAP.

(6) Ruiz-Mirazo, K.; Briones, C.; De La Escosura, A. *Chem. Rev*. **2014**, *114*, 285–366.

(7) Szostak, J. W.; Bartel, D. P.; Luisi, P. L. *Nature* **2001**, *409*, 387–390.

(8) Woese, C. R. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 8392–8396.

(9) Sutherland, J. D. *Angew. Chem. Int. Ed*. **2016**, *55*, 104–121.

(10) Reuß, D. R.; Commichau, F. M.; Gundlach, J.; Zhu, B.; Stülke, J. *Microbiol. Mol. Biol. Rev.* **2016**, *80*, 955–987.

(11) Hutchison, C. A.; Chuang, R. Y.; Noskov, V. N.; Assad-Garcia, N.; Deerinck, T. J.; Ellisman, M. H.; Gill, J.; Kannan, K.; Karas, B. J.; Ma, L.; et al. *Science* **2016**, *351*, aad6253

(12) Blain, J. C.; Szostak, J. W. *Annu. Rev. Biochem.* **2014**, *83*, 615–640.

(13) Orgel, L. E. *Crit. Rev. Biochem. Mol. Biol.* **2004**, *39*, 99–123.

(14) Eschenmoser, A. *Science* **1999**, *284*, 2118.

(15) Krishnamurthy, R. *Isr. J. Chem.* **2015**, *55*, 837–850.

(16) Engelhart, A. E.; Hud, N. V. *Cold Spring Harb. Perspect. Biol*. **2010**, *2*, a002196.

(17) Chen, I. A.; Walde, P. *Cold Spring Harb. Perspect. Biol*. **2010**. *2*, a002170.

(18) Nooner, D. W.; Gibert, J. M.; Gelpi, E.; Oró, J. *Geochim. Cosmochim. Acta* **1976**, *40*, 915–924.

(19) Patel, B. H.; Percivalle, C.; Ritson, D. J.; Duffy, C. D.; Sutherland, J. D. *Nat. Chem.* **2015**, *7*, 301–307.

(20) Deamer, D. W. *Nature* **1985**, *317*, 792–794.

(21) Gilbert, W. *Nature* **1986**, *319*, 618.

(22) Crick, F. H. *J. Mol. Biol.* **1968**, *38*, 367–379.

(23) Orgel, L. E. *J. Mol. Biol.* **1968**, *38*, 381–393.

(24) Woese, C. R. *The Genetic Code: The Molecular Basis for Genetic Expression*; Harper & Row: New York, 1967.

(25) Robertson, M. P.; Joyce, G. F. *Cold Spring Harb. Perspect. Biol.* **2012**, *4*, a003608.

(26) Cech, T. R.; Zaug, A. J.; Grabowski, P. J. *Cell* **1981**, *27*, 487–496.

(27) Guerrier-Takada, C.; Gardiner, K.; Marsh, T.; Pace, N.; Altman, S. *Cell* **1983**, *35*, 849–857.

(28) Steitz, T. A.; Moore, P. B. *Trends in Biochem. Sci.* **2003**, 411–418.

(29) Shi, Y. *Nat. Rev. Mol. Cell Biol*. **2017**, 655–670.

(30) Fica, S. M.; Tuttle, N.; Novak, T.; Li, N. S.; Lu, J.; Koodathingal, P.; Dai, Q.; Staley, J. P.; Piccirilli, J. A. *Nature* **2013**, *503*, 229–234.

(31) Jimenez, R. M.; Polanco, J. A.; Lupták, A. *Trends Biochem. Sci*. **2015,** *40*, 648–661.

(32) Uhlenbeck, O. C. *Nature* **1987**, *328*, 596–600.

(33) Saville, B. J.; Collins, R. A. *Cell* **1990**, *61*, 685–696.

(34) Sharmeen, L.; Kuo, M. Y.; Dinter-Gottlieb, G.; Taylor, J. *J. Virol.* **1988**, *62*, 2674–2679.

(35) Winkler, W. C.; Nahvi, A.; Roth, A.; Collins, J. A.; Breaker, R. R. *Nature* **2004**, *428*, 281–286.

(36) Breaker, R. R. *Cold Spring Harb. Perspect. Biol.* **2012**, *4*, a003566.

(37) Winkler, W.; Nahvi, A.; Breaker, R. R. *Nature* **2002**, *419*, 952–956.

(38) Nahvi, A.; Sudarsan, N.; Ebert, M. S.; Zou, X.; Brown, K. L.; Breaker, R. R. *Chem. Biol.* **2002**, *9*, 1043–1049.

(39) Epshtein, V.; Mironov, A. S.; Nudler, E. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 5052–5056.

(40) Ellington, A. D.; Szostak, J. W. *Nature* **1990**, *346*, 818–822.

(41) Puglisi, J. D.; Williamson, J. R. In *The RNA World*; Gesteland, R. F., Cech, T. R., Atkins, J. F., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 1999; 403–425.

(42) Breaker, R. R.; Joyce, G. F. *Chem. Biol*. **2014**, *21*, 1059–1065.

(43) Wachowius, F.; Attwater, J.; Holliger, P. *Q. Rev. Biophys.* **2017**, *50*, e4.

(44) Unrau, P. J.; Bartel, D. P. *Nature* **1998**, *395*, 260–263.

(45) Lorsch, J. R.; Szostak, J. W. *Nature* **1994**, *371*, 31–36.

(46) Bartel, D.; Szostak, J. *Science* **1993**, *261*, 1411–1418.

(47) Johnston, W. K.; Unrau, P. J.; Lawrence, M. S.; Glasner, M. E.; Bartel, D. P. *Science* **2001**, *292*, 1319–1325.

(48) Wochner, A.; Attwater, J.; Coulson, A.; Holliger, P. *Science* **2011**, *332*, 209–212.

(49) Mutschler, H.; Wochner, A.; Holliger, P. *Nat. Chem.* **2015**, *7*, 502–508.

(50) Horning, D. P.; Joyce, G. F. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 9786–9791.

(51) Lincoln, T. A.; Joyce, G. F. *Science* **2009**, *323*, 1229–1232.

(52) He, C.; Gállego, I.; Laughlin, B.; Grover, M. A.; Hud, N. V. *Nat. Chem.* **2017**, *9*, 318–324.

(53) Burcar, B.; Pasek, M.; Gull, M.; Cafferty, B. J.; Velasco, F.; Hud, N. V.; Menor-Salván, C. *Angew. Chem. Int. Ed.* **2016**, *55*, 13249–13253.

(54) Miller, S. L. *Science* **1953**, *117*, 528–529.

(55) Johnson, A. P.; Cleaves, H. J.; Dworkin, J. P.; Glavin, D. P.; Lazcano, A.; Bada, J. L. *Science* **2008**, *322*, 404.

(56) Parker, E. T.; Cleaves, H. J.; Dworkin, J. P.; Glavin, D. P.; Callahan, M.; Aubrey, A.; Lazcano, A.; Bada, J. L. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 5526–5531.

(57) Pasek, M.; Lauretta, D. *Orig. Life Evol. Biosph.* **2008**, *38*, 5–21.

(58) Pizzarello, S.; Shock, E. *Cold Spring Harb. Perspect. Biol*. **2010**, *2*, a002105.

(59) Cleaves, H. *Life* **2013**, *3*, 331–345.

(60) Oró, J. *Biochem. Biophys. Res. Commun.* **1960**, *2*, 407–412.

(61) Levy, M.; Miller, S. L.; Oró, J. *J. Mol. Evol.* **1999**, *49*, 165–168.

(62) Oró, J.; Kimball, A. P. *Arch. Biochem. Biophys.* **1961**, *94*, 293–313.

(63) Sanchez, R. A.; Ferris, J. P.; Orgel, L. E. *J. Mol. Biol.* **1968**, *38*, 121–128.

(64) Shapiro, R. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 4396–4401.

(65) Miyakawa, S.; Cleaves, H. J.; Miller, S. L. *Orig. Life Evol. Biosph.* **2002**, *32*, 209–218.

(66) Robertson, M. P.; Miller, S. L. *Nature* **1995**, *375*, 772–774.

(67) Saladino, R.; Crestini, C.; Ciambecchini, U.; Ciciriello, F.; Costanzo, G.; Di Mauro, E. *ChemBioChem* **2004**, *5*, 1558–1566.

(68) Cafferty, B. J.; Hud, N. V. *Curr. Opin. Chem. Biol*. **2014**, *22*, 146–157.

(69) Beckstead, A. A.; Zhang, Y.; de Vries, M. S.; Kohler, B. *Phys. Chem. Chem. Phys.* **2016**, *18*, 24228–24238.

(70) Powner, M. W.; Gerland, B.; Sutherland, J. D. *Nature* **2009**, *459*, 239–242.

(71) Butlerow, A. *Justus Liebigs Ann. Chem.* **1861**, *120*, 295–298.

(72) Breslow, R. *Tetrahedron Lett.* **1959**, *1*, 22–26.

(73) Shigemasa, Y.; Matsuda, Y.; Sakazawa, C.; Matsuura, T. *Bull. Chem. Soc. Jpn.* **1977**, *50*, 222.

(74) Delidovich, I. V.; Simonov, A. N.; Taran, O. P.; Parmon, V. N. *ChemSusChem*. **2014**, *7,* 1833–1846.

(75) Shapiro, R. *Orig. Life Evol. Biosph.* **1988**, *18*, 71–85.

(76) Pitsch, S.; Eschenmoser, A.; Gedulin, B.; Hui, S.; Arrhenius, G. *Orig. Life Evol. Biosph.* **1995**, *25*, 297–334.

(77) Ricardo, A.; Carrigan, M. A.; Olcott, A. N.; Benner, S. A. *Science* **2004**, *303*, 196–196.

(78) Lambert, J. B.; Gurusamy-Thangavelu, S. A.; Ma, K. *Science* **2010**, *327*, 984–986.

(79) Lazcano, A.; Miller, S. L. *Cell* **1996**, *85*, 793–798.

(80) Springsteen, G.; Joyce, G. F. *J. Am. Chem. Soc.* **2004**, *126*, 9578–9583.

(81) Fuller, W. D.; Sanchez, R. A.; Orgel, L. E. *J. Mol. Evol.* **1972**, *1*, 249–257.

(82) Sutherland, J. D. *Cold Spring Harb. Perspect. Biol*. **2010**, *2*, a005439.

(83) Islam, S.; Powner, M. W. *Chem*. **2017**, 470–501.

(84) Carothers, J. M. Dissertations, Harvard University. **2006**.

(85) Schoning, K. U.; Scholz, P.; Guntha, S.; Wu, X.; Krishnamurthy, R.; Eschenmoser, A. *Science* **2000**, *290*, 1347–1351.

(86) Schrum, J. P.; Ricardo, A.; Krishnamurthy, M.; Blain, J. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2009**, *131*, 14560–14570.

(87) Zhang, S.; Blain, J. C.; Zielinska, D.; Gryaznov, S. M.; Szostak, J. W. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 17732–17737.

(88) Chaput, J. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2003**, *125*, 9274–9275.

(89) Chaput, J. C.; Ichida, J. K.; Szostak, J. W. *J. Am. Chem. Soc.* **2003**, *125*, 856–857.

(90) Horhota, A.; Zou, K.; Ichida, J. K.; Yu, B.; McLaughlin, L. W.; Szostak, J. W.; Chaput, J. C. *J. Am. Chem. Soc.* **2005**, *127*, 7427–7434.

(91) Ichida, J. K.; Horhota, A.; Zou, K.; McLaughlin, L. W.; Szostak, J. W. *Nucleic Acids Res.* **2005**, *33*, 5219–5225.

(92) Pinheiro, V. B.; Taylor, A. I.; Cozens, C.; Abramov, M.; Renders, M.; Zhang, S.; Chaput, J. C.; Wengel, J.; Peak-Chew, S. Y.; McLaughlin, S. H.; et al. *Science*

2012, *336*, 341–344.

(93) Taylor, A. I.; Pinheiro, V. B.; Smola, M. J.; Morgunov, A. S.; Peak-Chew, S.; Cozens, C.; Weeks, K. M.; Herdewijn, P.; Holliger, P. *Nature* **2015**, *518*, 427–430.

(94) Choudhary, A.; Kamer, K. J.; Powner, M. W.; Sutherland, J. D.; Raines, R. T. *ACS Chem. Biol.* **2010**, *5*, 655–657.

(95) Stairs, S.; Nikmal, A.; Bučar, D.-K.; Zheng, S.-L.; Szostak, J. W.; Powner, M. W. *Nat. Commun.* **2017**, *8*, 15270.

(96) Mukhin, L. *Nature* **1974**, *251*, 50–51.

(97) Benner, S. A.; Kim, H. J.; Yang, Z. *Cold Spring Harb. Perspect. Biol.* **2012**, *4*, a003541.

(98) Islam, S.; Bučar, D.-K. K.; Powner, M. W. *Nat. Chem.* **2017**, *9*, 584–589.

(99) Ritson, D.; Sutherland, J. D. *Nat. Chem.* **2012**, *4*, 895–899.

(100) Todd, Z. R.; Fahrenbach, A. C.; Magnani, C. J.; Ranjan, S.; Björkbom, A.; Szostak, J. W.; Sasselov, D. D. *Chem. Commun.* **2018**, *54*, 1121–1124.

(101) Keefe, a D.; Miller, S. L. *Orig. Life Evol. Biosph.* **1996**, *26*, 111–129.

(102) Seifer, G. B. *Russ. J. Inorg. Chem.* **1962**, *7*, 640–643.

(103) Ranjan, S.; Todd, Z. R.; Sutherland, J. D.; Sasselov, D. D. *arXiv* **2018**, 1801.07725.

(104) Becker, S.; Thoma, I.; Deutsch, A.; Gehrke, T.; Mayer, P.; Zipse, H.; Carell, T. *Science* **2016**, *352*, 833–836.

(105) Becker, S.; Schneider, C.; Okamura, H.; Crisp, A.; Amatov, T.; Dejmek, M.; Carell, T. *Nat. Commun.* **2018**, *9*, 163.

(106) Bada, J. L.; Chalmers, J. H.; Cleaves, H. J. *Phys. Chem. Chem. Phys.* **2016**, *18*, 20085–20090.

(107) Dickson, K. S.; Burns, C. M.; Richardson, J. P. *J. Biol. Chem.* **2000**, *275*, 15828–15831.

(108) Morávek, J. *Tetrahedron Lett.* **1967**, *8*, 1707–1710.

(109) Rajamani, S.; Vlassov, A.; Benner, S.; Coombs, A.; Olasagasti, F.; Deamer, D. *Orig. Life Evol. Biosph.* **2008**, *38*, 57–74.

(110) Deamer, D. W.; Barchfeld, G. L. *J. Mol. Evol.* **1982**, *18*, 203–206.

(111) Walde, P. *Orig. Life Evol. Biosph*. **2006,** *36*, 109–150.

(112) Mungi, C.; Rajamani, S. *Life* **2015**, *5*, 65–84.

(113) Orgel, L. E.; Lohrmann, R. *Acc. Chem. Res.* **1974**, *7*, 368–377.

(114) Lohrmann, R.; Orgel, L. E. *Nature* **1976**, *261*, 342–344.

(115) Weimann, B. J.; Lohrmann, R.; Orgel, L. E.; Schneider-Bernloehr, H.; Sulston, J. E. *Science* **1968**, *161*, 387–387.

(116) Prabahar, K. J.; Ferris, J. P. *J. Am. Chem. Soc.* **1997**, *119*, 4330–4337.

(117) Prabahar, K. J.; Cole, T. D.; Ferris, J. P. *J. Am. Chem. Soc.* **1994**, *116*, 10914–10920.

(118) Sleeper, H. L.; Orgel, L. E. *J. Mol. Evol.* **1979**, *12*, 357–364.

(119) Sawai, H.; Karoda, K.; Hojo, T. *Bull. Chem. Soc. Jpn.* **1989**, *62*, 2018–2023.

(120) Sawai, H.; Higa, K.; Kuroda, K. *J .Chem. Soc. Perkin Trans. I* **1992**, 505–508.

(121) Kanavarioti, A.; Monnard, P.-A.; Deamer, D. W. *Astrobiology* **2001**, *1*, 271–281.

(122) Monnard, P. A.; Kanavarioti, A.; Deamer, D. W. *J. Am. Chem. Soc.* **2003**, *125*, 13734–13740.

(123) Ferris, J. P.; Ertem, G. *J. Am. Chem. Soc.* **1993**, *115*, 12270–12275.

(124) Huang, W.; Ferris, J. P. *J. Am. Chem. Soc.* **2006**, *128*, 8914–8919.

(125) Ertem, G.; Ferris, J. P. *Orig. Life Evol. Biosph.* **1998**, *28*, 485–499.

(126) Ertem, G.; Ferris, J. P. *J. Am. Chem. Soc.* **1997**, *119*, 7197–7201.

(127) Howard, F. B.; Frazier, J.; Singer, M. F.; Miles, H. T. *J. Mol. Biol.* **1966**, *16*, 415–439.

(128) Naylor, R.; Gilham, P. T. *Biochemistry* **1966**, *5*, 2722–2728.

(129) Sulston, J.; Lohrmann, R.; Orgel, L. E.; Miles, H. T. *Proc. Natl. Acad. Sci. U. S. A.* **1968**, *59*, 726–733.

(130) Sulston, J.; Lohrmann, R.; Orgel, L. E.; Miles, H. T. *Proc. Natl. Acad. Sci. U. S. A.* **1968**, *60*, 409–415.

(131) Sulston, J.; Lohrmann, R.; Orgel, L. E.; Schneider-Bernloehr, H.; Weimann, B. J.; Miles, H. T. *J. Mol. Biol.* **1969**, *40*, 227–234.

(132) Shabarova, Z. A. *Biochimie*. **1988,** *70*, 1323–1334.

(133) Sokolova, N. I.; Ashirbekova, D. T.; Dolinnaya, N. G.; Shabarova, Z. A. *FEBS Lett.* **1988**, *232*, 153–155.

(134) Vogel, H.; Gerlach, C.; Richert, C. *Nucleos. Nucleot. Nucl.* **2013**, *32*, 17–27.

(135) Bowler, F. R.; Chan, C. K. W. W.; Duffy, C. D.; Gerland, B.; Islam, S.; Powner, M. W.; Sutherland, J. D.; Xu, J. *Nat. Chem.* **2013**, *5*, 383–389.

(136) Westheimer, F. *Science* **1987**, *235*, 1173–1178.

(137) Biertümpfel, C.; Zhao, Y.; Kondo, Y.; Ramón-Maiques, S.; Gregory, M.; Lee, J. Y.; Masutani, C.; Lehmann, A. R.; Hanaoka, F.; Yang, W. *Nature* **2010**, *465*, 1044–1048.

(138) Gao, Y.; Yang, W. *Science* **2016**, *352*, 1334–1337.

(139) Shechner, D. M.; Bartel, D. P. *Nat. Struct. Mol. Biol.* **2011**, *18*, 1036–1042.

(140) Rohatgi, R.; Bartel, D. P.; Szostak, J. W. *J. Am. Chem. Soc.* **1996**, *118*, 3340–3344.

(141) Oró, J.; Basile, B.; Cortes, S.; Shen, C.; Yamrom, T. *Orig. Life* **1984**, *14*, 237–242.

(142) Lohrmann, R. *J. Mol. Evol.* **1977**, *10*, 137.

(143) Lohrmann, R.; Bridson, P. K.; Orgel, L. E. *Science* **1980**, *208*, 1464–1465.

(144) Bridson, P. K.; Orgel, L. E. *J. Mol. Biol.* **1980**, *144*, 567–577.

(145) Inoue, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1981**, *103*, 7666–7667.

(146) Kozlov, I. A.; Orgel, L. E. *Mol. Biol.* **2000**, *34*, 781–789.

(147) Zhang, N.; Zhang, S.; Szostak, J. W. *J. Am. Chem. Soc.* **2012**, *134*, 3691–3694.

(148) Wu, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1992**, *114*, 5496–5501.

(149) Wu, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1992**, *114*, 7963–7969.

(150) Rajamani, S.; Ichida, J. K.; Antal, T.; Treco, D. A.; Leu, K.; Nowak, M. A.; Szostak, J. W.; Chen, I. A. *J. Am. Chem. Soc.* **2010**, *132*, 5880–5885.

(151) Heuberger, B. D.; Pal, A.; Del Frate, F.; Topkar, V. V.; Szostak, J. W. *J. Am. Chem. Soc.* **2015**, *137*, 2769–2775.

(152) Prywes, N.; Michaels, Y. S.; Pal, A.; Oh, S. S.; Szostak, J. W. *Chem. Commun.* **2016**, *52*, 6529–6532.

(153) Vogel, S. R.; Deck, C.; Richert, C. *Chem. Commun.* **2005**, *39*, 4922–4924.

(154) Deck, C.; Jauker, M.; Richert, C. *Nat. Chem.* **2011**, *3*, 603–608.

(155) Prywes, N.; Blain, J. C.; Del Frate, F.; Szostak, J. W. *Elife* **2016**, *5*, e17756.

(156) Tam, C. P.; Fahrenbach, A. C.; Björkbom, A.; Prywes, N.; Izgu, E. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 571–574.

(157) Blain, J. C. Dissertation, Harvard University. **2013**.

(158) Li, L.; Prywes, N.; Tam, C. P.; O'Flaherty, D. K.; Lelyveld, V. S.; Izgu, E. C.; Pal, A.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, 139, 1810–1813.

(159) Zhang, S.; Zhang, N.; Blain, J. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2013**, *135*, 924–932.

(160) Fahrenbach, A. C.; Giurgiu, C.; Tam, C. P.; Li, L.; Hongo, Y.; Aono, M.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 8780–8783.

(161) Kanavarioti, A.; Bernasconi, C. F.; Doodokyan, D. L.; Alberas, D. J. *J. Am. Chem. Soc.* **1989**, *111*, 7247–7257.

(162) Li, L.; Lelyveld, V. S.; Prywes, N.; Szostak, J. W. *J. Am. Chem. Soc.* **2016**, *138*, 3986–3989.

(163) Zhang, W.; Tam, C. P.; Wang, J.; Szostak, J. W. *ACS Cent. Sci.* **2016**, *2*, 916–926.

(164) Zhang, W.; Tam, C. P.; Zhou, L.; Oh, S. S.; Wang, J.; Szostak, J. W. *J. Am. Chem. Soc.* **2018**, *140*, 2829–2840.

(165) Walton, T.; Szostak, J. W. *J. Am. Chem. Soc.* **2016**, *138*, 11996–12002.

(166) Kervio, E.; Sosson, M.; Richert, C. *Nucleic Acids Res.* **2016**, *44*, 5504–5514.

(167) Zhang, W.; Tam, C. P.; Walton, T.; Fahrenbach, A. C.; Birrane, G.; Szostak, J. W. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 7659–7664.

(168) Tam, C. P.; Zhou, L.; Fahrenbach, A. C.; Zhang, W.; Walton, T.; Szostak, J. W. *J. Am. Chem. Soc.* **2018**, *140*, 783–792.

(169) Zhang, W.; Walton, T.; Li, L.; Szostak, J. W. *Manuscript Submitted.*

(170) Giurgiu, C.; Li, L.; O'Flaherty, D. K.; Tam, C. P.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 16741–16747.

(171) Walton, T.; Szostak, J. W. *Biochemistry* **2017**, *56*, 5379.

(172) Szostak, J. W. *J. Syst. Chem.* **2012**, *3*, 2.

(173) Jauker, M.; Griesser, H.; Richert, C. *Angew. Chem. Int. Ed.* **2015**, *54*, 14559–14563.

(174) Kanaya, E.; Yanagawa, H. *Biochemistry* **1986**, *25*, 7423–7430.

(175) Pascal, R.; Boiteau, L.; Commeyras, A. In *Topics in Current Chemistry: Prebiotic Chemistry*; 2005; 69–122.

(176) Röthlingshöfer, M.; Richert, C. *J. Org. Chem.* **2010**, *75*, 3945–3952.

(177) Yi, R.; Hongo, Y.; Fahrenbach, A. C. *Chem. Commun.* **2018**, 54, 511–514.

(178) Adamala, K.; Szostak, J. W. *Science* **2013**, *342*, 1098–1100.

(179) Kartha, G.; Varughese, K. I.; Aimoto, S. *Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 4519–4522.

(180) Neupert-Laves, K.; Dobler, M. *Helv. Chim. Acta* **1975**, *58*, 432–442.

(181) Bonfio, C.; Valer, L.; Scintilla, S.; Shah, S.; Evans, D. J.; Jin, L.; Szostak, J. W.; Sasselov, D. D.; Sutherland, J. D.; Mansy, S. S. *Nat. Chem.* **2017**, *9*, 1229–1234.

(182) Jain, S. S.; Anet, F. A. L.; Stahle, C. J.; Hud, N. V. *Angew. Chem. Int. Ed.* **2004**, *43*, 2004–2008.

Chapter 2

Unusual Base-Pairing Interactions in Monomer−Template Complexes

[*W. Zhang, C.P. Tam and J. Szostak conceived of experiments, analyzed the data and wrote the text. W. Zhang performed crystallography experiments and C. P. Tam developed synthesis for PZG. J. Wang provided technical assistance. This chapter has previously been published with minor changes as the following publication:* Zhang, W.*; Tam, C. P.*; Wang, J.; Szostak, J. W. Unusual Base-Pairing Interactions in Monomer–Template Complexes. *ACS Cent. Sci.* **2016***, 2* (12)*,* 916–926. ** denotes co-first author.*]

Abstract. Many high-resolution crystal structures have contributed to our understanding of the reaction pathway for catalysis by DNA and RNA polymerases, but the structural basis of nonenzymatic template-directed RNA replication has not been studied in comparable detail. Here we present crystallographic studies of the binding of ribonucleotide monomers to RNA primer−template complexes, with the goal of improving our understanding of the mechanism of nonenzymatic RNA copying, and of catalysis by polymerases. To explore how activated ribonucleotides recognize and bind to RNA templates, we synthesized an unreactive phosphonate-linked pyrazole analogue of guanosine 5′-phosphoro-2-methylimidazolide (2-MeImpG), a highly activated nucleotide that has been used extensively to study nonenzymatic primer extension. We cocrystallized this analogue with structurally rigidified RNA primer−template complexes

carrying single or multiple monomer binding sites, and obtained high-resolution X-ray structures of these complexes. In addition to Watson−Crick base pairing, we repeatedly observed noncanonical guanine:cytidine base pairs in our crystal structures. In most structures, the phosphate and leaving group moieties of the monomers were highly disordered, while in others the distance from O3′ of the primer to the phosphorus of the incoming monomer was too great to allow for reaction. We suggest that these effects significantly influence the rate and fidelity of nonenzymatic RNA replication, and that even primitive ribozyme polymerases could enhance RNA replication by enforcing Watson−Crick base pairing between monomers and primer−template complexes, and by bringing the reactive functional groups into closer proximity.

2.1 Introduction

In the RNA world hypothesis as first proposed by Crick,[1] Orgel,[2] and Woese,[3] the genetic continuity of the earliest earthly life forms was ensured by RNA replication. Template directed nonenzymatic RNA replication[4−6] has unique significance in the RNA world hypothesis in that it facilitates the transition from the prebiotic chemistry of simple organic building blocks to the emergence of the earliest ribozyme catalysts.[4] During nonenzymatic RNA replication, chemically activated mono- or oligoribonucleotides bind to their complementary binding sites on an RNA primer−template complex, and the polymerization or ligation of these monomers and oligomers then follows to furnish the complementary RNA strand.[6] An example of such nonenzymatic RNA template copying is the purely chemically driven but template-directed synthesis of short, 3′−5′-linked guanosine- and cytosine-rich RNA segments from 2-methylimidazole-activated ribonucleotide monomers in the presence of divalent metal ions.[7] Subsequently, the thermodynamic principles that govern the reversible noncovalent association between RNA templates and activated ribonucleotides,[8,9] the kinetic parameters of the subsequent oligomerization,[10,11] and the reaction conditions that ensure the regioselectivity of the 3′- to 5′-RNA polymerization,[12] have been explored.

The hydrogen-bonding interactions between the nucleobases of the template and the incoming activated mononucleotides are instrumental in ensuring complementarity between template and monomers.[6] Intense experimental efforts have been devoted to the observation of the structural nature of such interactions. Despite the weakness of noncovalent template−monomer interactions in aqueous media,[13,14] studies of complexes composed of a polypyrimidine template and monomeric purine nucleotides have provided

evidence, including infrared absorption spectra, that the template−monomer complexes resemble fully complementary RNA-double helices.[15,16] This observation, together with the high fidelity of G:C template copying, implies that monomers associate with the template with Watson−Crick geometry. Surprisingly, Fujita and colleagues have recently demonstrated that a self-assembling, tripyridyl-triazine capped hydrophobic molecular cage encourages formation of nonconventional anti- Hoogsteen base pairing interactions between adenosine and uridine monophosphate monomers (AMP and UMP), as well as between two self-complementary thymidylyl-adenosine (TA) dinucleotides within the hydrophobic cavity of the cage.[17]

These observations raise the question of whether Watson−Crick interactions indeed dominate the noncovalent association between complementary nucleobases, or whether other types of interaction geometries could be of comparable importance in base pairing. We therefore decided to probe the structural nature of template−monomer recognition under aqueous conditions that are as close as possible to standard reaction conditions, in order to elucidate the structural basis of nonenzymatic RNA replication.

Recent studies have elucidated some of the thermodynamic aspects of the binding interactions between mononucleotides and oligonucleotides. For example, studies from our lab and the Richert lab have used proton NMR spectroscopy to probe the affinity and context dependence of the reversible binding between all four deoxy- and ribonucleotides with complementary oligonucleotide templates, as well as the G:T(U) and C:T mismatches.[8,9] We have also applied transferred nuclear Overhauser effect spectroscopy (TrNOESY) to demonstrate that the sugar pucker of activated guanosine ribonucleotides switches from predominantly C2′-endo in the free state to predominantly C3′-endo when

bound to a complementary poly-C RNA template.[18] Interestingly, this phenomenon is observed only in the case of an RNA template and activated ribonucleotide monomers, and not with a DNA template or deoxynucleotide monomers. Despite the fact that these studies provide insight into certain structural aspects of monomer−template binding, the noncovalent assembly of ribonucleotides onto an RNA template has not yet been probed by X-ray crystallography.

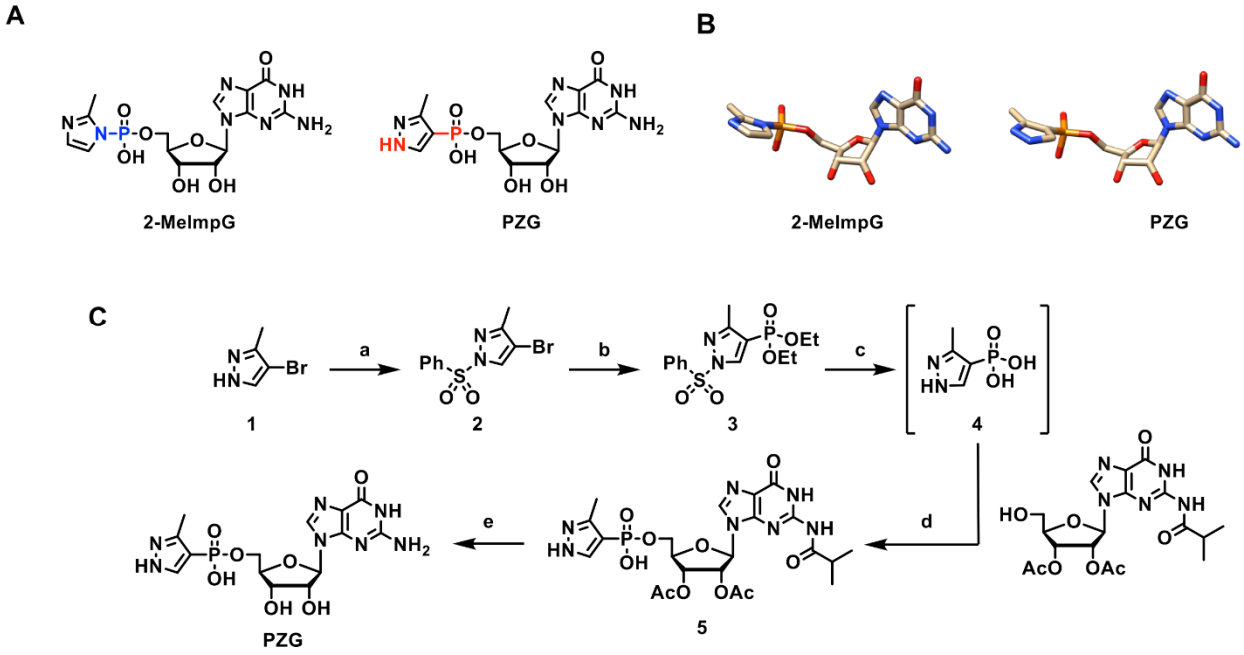
X-ray crystallography remains the dominant technique for revealing the three-dimensional structures of nucleic acids and nucleic acid−ligand complexes at atomic resolution.[19] We have recently used crystallography to study modified RNA duplexes, with the goal of obtaining structural information that could lead to more efficient and accurate nonenzymatic RNA replication. This approach has been successful in demonstrating how the structural perturbations resulting from interspersed 2′−5′-phosphodiester linkages are accommodated by local conformational adjustments in the region flanking the aberrant linkage.[20] We also demonstrated by crystallography that the stabilization of the U:A Watson−Crick base pair by replacing U with 2-thio-U causes minimal perturbation of the overall RNA duplex structure.[21] In order to accurately reflect the structural characteristics of monomer−template binding under conditions resembling standard nonenzymatic RNA replication protocols, we designed and synthesized a stable nonhydrolyzable isosteric 2-MeImpG analogue, and cocrystallized it with LNA-rigidified RNA primer−template complexes. We obtained atomic resolution X-ray crystal structures of this ligand complexed with three variants of self-complementary LNA-modified RNA duplexes that contain either one, two, or three 5-methylcytidine (5-mC) LNA overhangs at their 5′-termini, as binding sites for the guanosine analogue. In addition to canonical Watson−Crick base

pairing, we frequently observed base pairing of the guanosine analogue with the templating 5-mC via noncanonical interactions. The activated phosphate moiety of a template bound monomer was frequently highly disordered, or, in more ordered structures, was too far from the 3′-hydroxyl group of the upstream nucleotide to react without significant conformational change. Our findings suggest that noncanonical base pairing may occur alongside standard Watson−Crick base pairing interactions in solution, potentially leading to slower and less accurate primer extension. Our observations further suggest that primitive ribozyme (and protein) RNA polymerases may enhance the rate and fidelity of template copying by restricting monomer−template interactions to those with Watson−Crick geometry and by imposing structural order and specific conformational changes on the reacting primer terminus and incoming monomer.

2.2 Results

2.2.1 Synthesis of the 2-MeImpG Pyrazole Analogue PZG. In 2-MeImpG, the 2-methylimidazole leaving group is linked to guanosine monophosphate via a labile nitrogen−phosphorus bond (Scheme 1A, highlighted in blue). Under nonenzymatic RNA copying conditions, divalent metal cations, such as Mg(II) and Mn(II), can behave as Lewis acids and may coordinate with an oxygen atom on the phosphate of 2-MeImpG,[22,23] thus chemically activating the phosphorus for nucleophilic attack by either the 3′- or 2′-OH group of the RNA primer (leading to primer extension), or by ambient water molecules (leading to hydrolysis), thereby breaking the N−P bond and liberating the 2-methylimidazole leaving group. We designed a novel pseudo-activated monomer by replacing the labile N−P bond with a stable C−P bond. We achieved this by changing the 2-methylimidazoyl group of 2-MeImpG into a 3-methylpyrazoyl group and by connecting

the leaving group mimic to the phosphorus via a stable and unreactive phosphonate linkage (Scheme 1A, highlighted in red). The new molecule, guanosine 5′-(3-methyl-1*H*-pyrazol-4-yl)phosphonate (PZG), is isosteric to the original 2-MeImpG, as well as having a similar overall structure and aromaticity (Scheme 1B).



Scheme 2.1. (A) Structures of 2-MeImpG and PZG analog compared side-by-side. One of the nitrogen atoms of the 3-methylpyrazolyl group (labeled red) is displaced from its corresponding position in 2-MeImpG (labeled blue). (B) Stick model representations of 2-MeImpG and PZG. (C) Synthesis of PZG. Reaction conditions: (a) $PhSO_2Cl$, Et_3N, DCM; (b) $(EtO)_2P(O)H$, $Pd(PPh_3)_4$, PPh_3, Et_3N, DMSO, 115 °C, 2 h; (c) TMSBr, Et_3N, DCM, 4 h; then MeOH; (d) DIAD, PPh_3, DCM, 3 h; (e) conc. ammonium hydroxide, MeOH, 65 °C, 4 h.

The synthetic scheme used to prepare PZG is shown in Scheme 1C; the longest linear sequence from commercially available starting materials was five steps. To synthesize PZG, two bond-forming events were critical: the C−P linkage of the pyrazolyl-phosphonate, and the O−P bond that bridges the guanosine ribonucleoside and the pyrazolyl-phosphonate. We first protected commercially available bromopyrazole **1** with

a benzenesulfonyl protecting group to generate the *N*-phenylsulfonyl-protected **2**; the protecting group served to temporarily mask the acidic pyrazole proton, and also to activate the carbon−bromine bond for subsequent C−P bond formation. Next, the protected pyrazolyl-phosphonate **3** was produced by a palladium(0)-catalyzed C−P coupling between **2** and diethyl phosphite, furnishing the critical carbon−phosphorus bond.[24] To connect the pyrazolyl-phosphonate to guanosine, the two ethyl groups on **3** must be removed to unveil the aryl phosphonate in its nucleophilic anionic form. This was achieved by treating **3** with bromotrimethylsilane, followed by decomposing the resultant bis(trimethylsilyl)phosphonate ester with methanol to generate the dianionic aryl phosphonate **4**.[25,26] The guanosine−phosphonate ester linkage was prepared via a Mitsunobu reaction to afford protected guanosine pyrazolyl-phosphonate **5**, followed by global protecting group aminolysis to afford PZG.

2.2.2 Structure Determination of RNA Duplexes with Bound PZG. We cocrystallized PZG with self-complementary RNA oligonucleotide duplexes which share a common ten-nucleotide duplex core, flanked on both ends by 5′ single stranded overhangs that provided one, two, or three PZG binding sites.[27] The adapted oligonucleotide sequence was originally reported by Beal and co-workers for probing the effects of nucleobase alkynyl and triazolyl modifications on the overall duplex structure.[27] The 5′-overhangs were synthesized with 5-mC LNA nucleotides (denoted as mC in oligomer sequences, *vide infra*). The LNA modification (denoted as bolded nucleotides, *vide infra*) locks the sugar into the 3′-endo conformation and helps to preorganize the single-stranded overhang in the A-form conformation. The pK_a of N3 of 5-mC is close to that of native cytidine (4.45 for C and 4.6 for 5-mC),[28] and 5-mC can form a Watson−Crick base pair

with a guanine nucleobase much like canonical C. Together, these modifications facilitated crystallization and allowed us to obtain high-resolution X-ray structures of the resulting oligonucleotide−monomer complexes.

Table 2.1. Parameters for X-ray data collection and structural refinement.

	PZG-RNA 15mer	**PZG**-RNA 7mer	**PZG**-RNA 14mer	**PZG**-RNA 13mer	GMP-RNA 15mer	GMP-RNA 7mer	GMP-RNA 14mer
PDB code	5KRG	5HBW	5HBX	5HBY	5L00	5DHB	5DHC
Space group	P3	$P2_12_12_1$	P3	C_12_1	P3	P3	P3
RNA duplexes per asymmetric unit	2	1	2	0.5	2	2	2
Resolution range, Å	50-1.60	50-1.90	50-1.70	50-1.18	50-1.25	50-1.80	50-1.55
Unique reflections	41492	7254	38632	23597	47622	20608	25146
Completeness, %	95.5 (71.3)	99.6 (99)	99.8 (100)	93.6 (54.2)	92.7 (64.5)	98.1 (94.3)	99.9 (100)
R_{work}, %	20.6	21.7	21.3	15.3	20.8	18.7	17.0
R_{free}, %	23.4	25.4	24.9	16.6	23.8	19.2	19.5
Average B-factors, $Å^2$	27.54	33.77	29.32	13.29	16.85	31.38	20.59

We began our exploration of monomer−template binding by examining the crystal structure of the complex formed by PZG and the self-dimerizing RNA sequence 5′-mCmCmCGACUUAAGUCGG-3′, with the single overhanging 5-mC locked ribonucleotide at the 5′-terminus serving as a binding site for the PZG monomer.[27] The key crystallographic parameters are listed in **Table 2.1** (PDB: 5KRG). There were two RNA duplex−PZG molecular complexes per asymmetric unit, thus providing four independent views of PZG:template binding. Each PZG bound to the 5-mC LNA overhang was sandwiched between the upstream guanine at the 3′-end of the oligonucleotide, as well as the "downstream" 5′-overhanging 5-mC of a neighboring slip-stacked RNA duplex (**Figure 2.1a**). The internucleobase spacing between the PZG guanine base and both the upstream guanine and the "downstream" 5-mC nucleobase was approximately 3.3 Å based on analysis using the CONTACT program from CCP4.[29] At one end of each duplex, the PZG monomer was bound to the templating 5-mC via Watson−Crick base pairing. Remarkably, a noncanonical G:C base pair was observed at the other end of each duplex, mediated by two hydrogen bonds: the guanine N3 of the PZG was 3.0 Å from the exocyclic amine of the 5-mC, while the exocyclic amine of the guanine was 3.1 Å from the N3 of the 5-mC[30] (**Figure 2.1b**). Unfortunately, the sugar and aryl phosphonate moieties of the PZG monomers at both ends of both RNA duplexes were too disordered (nucleobase B factor, 30.6; ribose B factor, 49; phosphonate B factor, 109) to obtain useful information regarding the conformation of the sugar and leaving group mimic moieties of the PZG monomer.

In an effort to obtain a structure in which a single bound PZG would be more highly ordered, we crystallized a shorter RNA duplex (5′-mCTGUACA-3′).[31] The structure was

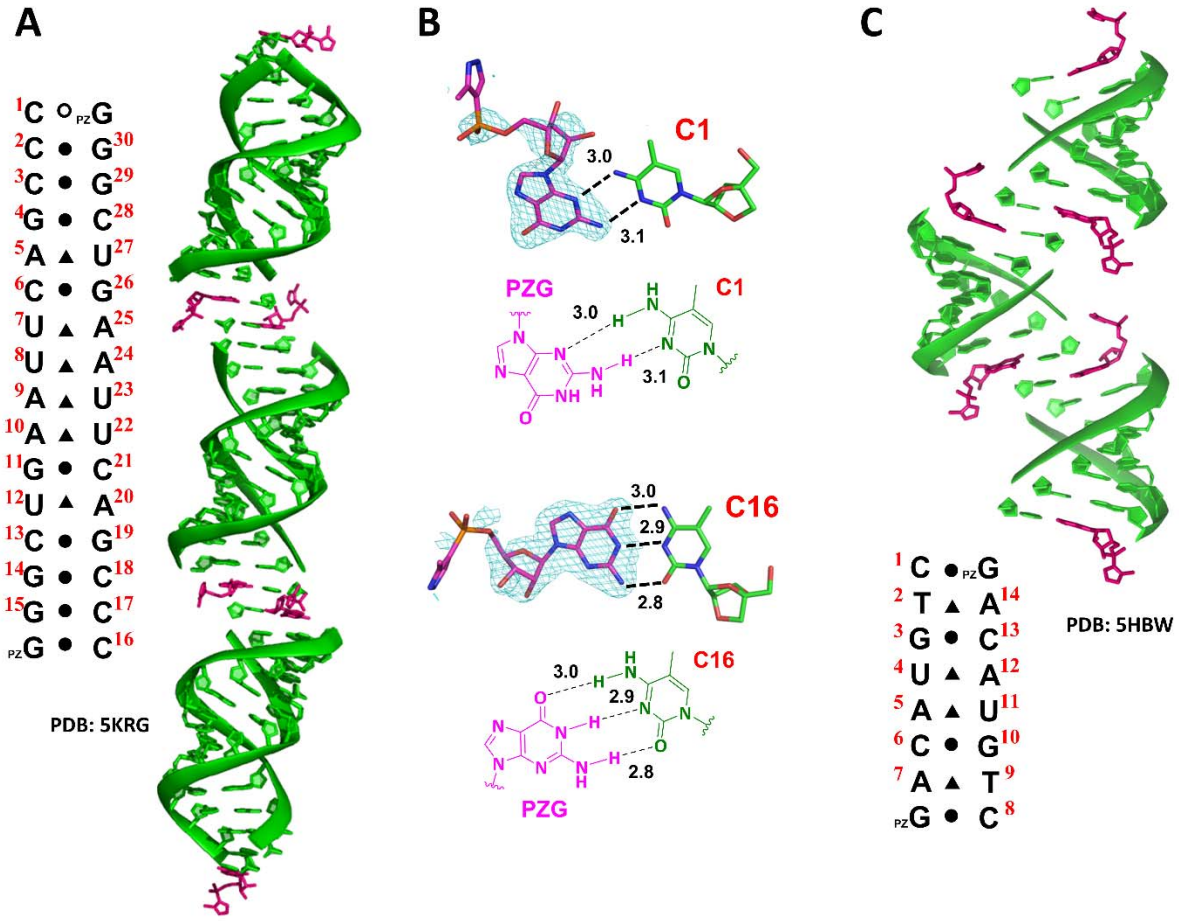


Figure 2.1. Crystal structures of single binding site RNA-PZG complexes. Each duplex has a one nucleotide 5′-overhang at each end that binds one PZG residue. Green: RNA duplex; Magenta: PZG monomer. Black filled circle: GC Watson-Crick pairs; Black triangle: AU Watson-Crick pairs; Black open circle: noncanonical base pairs with two hydrogen bonds. Hydrogen bonds are labeled as black dashes and their interatomic distances are indicated. (A) Diagram and structure of duplex formed from a self-complementary RNA 15mer with one 5-mC overhang at each end. The bound PZG monomers are stacked between the ends of coaxial RNA duplexes. The sequence of the oligonucleotide is shown and numbered. (B) The PZG monomers form two different base-pairs with the templating 5-mC residues at opposite ends of the duplex. The corresponding 2Fo-Fc omit maps contoured at 1.0 σ (cyan mesh) indicate the ordered nucleobases and the disorder of the sugar, phosphate and pyrazolyl moieties. (C) A self-complementary RNA 7mer with one 5-mC binding site at each end was co-crystallized with PZG monomers. The monomers are stacked between coaxial RNA duplexes and Watson-Crick paired to the template.

determined at 1.9 Å resolution ($P2_12_12_1$ space group, PDB: 5HBW). An end-to-end helical stacking interaction was also observed as in the previous one-binding-site structure. However, in this case, the PZG monomer was bound to the templating 5-mC via Watson−Crick base pairing at both ends of the duplex (H-bond distances: 2.8 to 3.0 Å). Furthermore, the improved electron density fitting of the PZG monomers in this structure allowed observation of different ribose conformations at the two termini of the duplex (**Figure 2.1c**). At one end, the PZG sugar was in the canonical 3′-endo conformation, as were the sugars in the rest of the A-form RNA duplex. At the other end, however, the ribose was in the 2′-endo conformation. This observation highlights the conformational flexibility of the PZG monomer, even when it is bound to the template, and is consistent with our prior TrNOESY spectroscopic studies which showed that 2-methylimidazolide activated guanosine mononucleotide was in the 2′-endo conformation a fraction of the time even when template bound.[18] The distances between the 3′-hydroxyl groups at the two duplex 3′-termini and the phosphorus atoms of PZG aryl phosphonates were over 6 Å at both ends (6.3 Å for 3′-endo PZG and 6.5 Å for 2′-endo PZG); at this distance, the phosphodiester-bond-forming reaction between the RNA primer and the activated monomer would require a major conformational change.

We next cocrystallized the PZG monomer with a partially self-complementary RNA 14mer oligonucleotide, 5′- mCmCAGACUUAAGUCU-3′,[27] which forms a 12 base-pair duplex with two-nucleotide 5-mCmC 5′-overhangs at both ends. The sequence of the duplex core was similar to that of the 15mer duplex reported above. The structure was determined to 1.7 Å resolution by molecular replacement using the central core of the 12mer RNA duplex as a search model, with an overall B-factor of 29.32 (PDB: 5HBX). As

in the previous structures, the individual double helices were slip-stacked with one another end-to-end (**Figure 2.2a**). Overall, two PZG monomers were clearly bound at each end of each RNA duplex, fully occupying all available binding sites. The guanine nucleobases of both monomers were bound to the two consecutive 5-mC template sites with clearly defined electron densities, while the sugars and pyrazolyl phosphonate groups of the monomers were highly disordered (B factors >40).

At one end of each RNA duplex, the guanine nucleobases of the two consecutively bound PZGs were well-ordered, coplanar, and Watson−Crick base paired with the consecutive 5-mC LNA nucleotides in the 5′-overhang (hydrogen-bond distances, 2.7 to 2.9 Å; the interplanar distances between the monomer nucleobases and their flanking nucleotides, ∼3.4 Å). The sugars and aryl phosphonates of both monomers were disordered. At the other end of the RNA duplex, however, an alternative base-recognition motif was observed. The nucleobase of the first bound monomer (adjacent to the 3′-end of the oligonucleotide) was base paired with the template in Watson−Crick fashion (hydrogen-bond lengths, 3.1, 2.8, 2.6 Å; nucleobase B factor, 26.8), and the ribose sugar was in the 3′-endo A-form conformation (B factor: 38.9). The second bound monomer was paired with 5-mC via the previously described noncanonical G:C N3-amino, amino-N3 interactions (**Figure 2.2c**).[30] Unfortunately, the sugar and aryl phosphonate moieties of the second bound PZG monomer were disordered and not clearly visible. B factor calculation indicated that the first monomer was more structurally ordered than the second one at both ends of the RNA duplex. This phenomenon is consistent with our NMR spectroscopic data[32] suggesting that, when nucleotide monomers bind sequentially

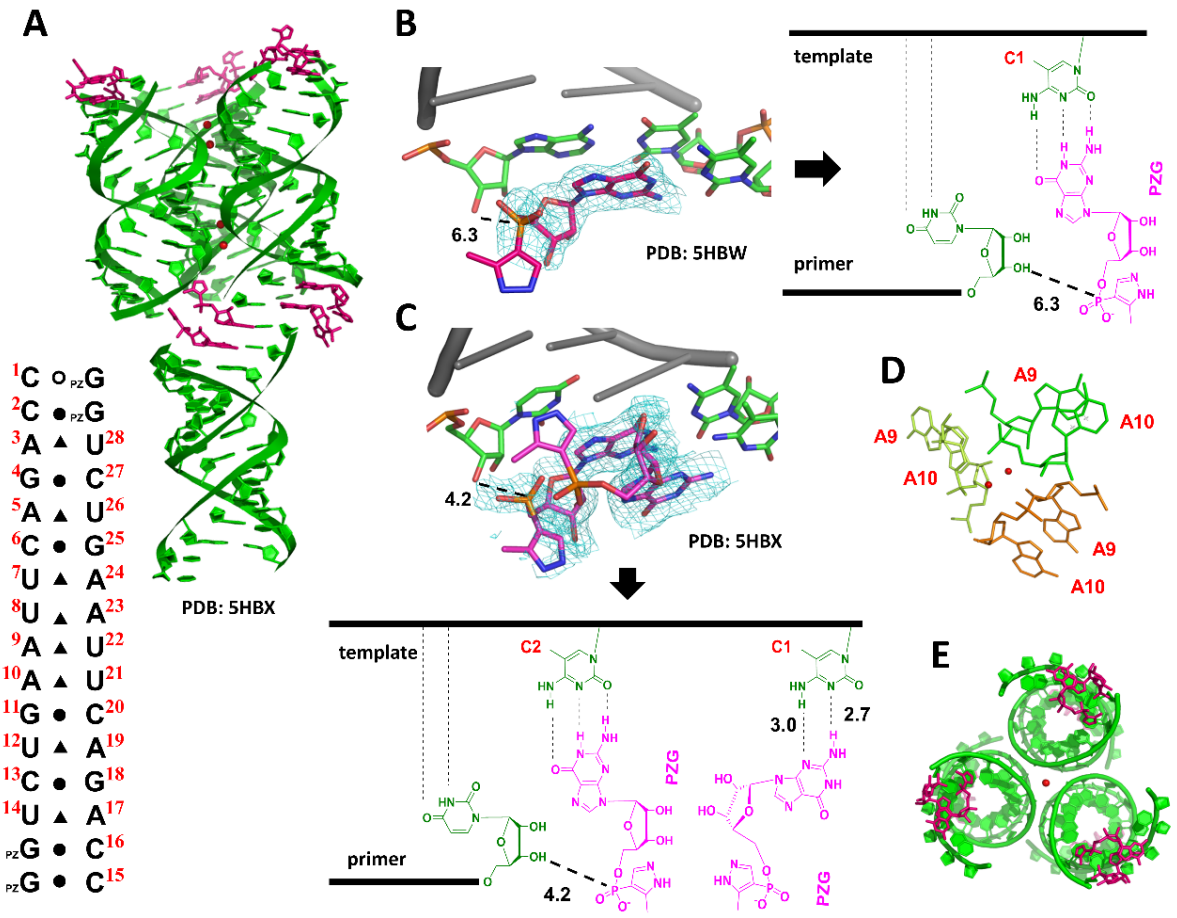


Figure 2.2. Crystal structures of RNA-**PZG** complexes with 5′-overhangs with two **PZG** binding sites. Green: RNA duplex; Magenta: **PZG** monomer; Red sphere: water molecules. Black filled circle: GC Watson-Crick pairs; Black triangle: AU Watson-Crick pairs; Black open circle: noncanonical base- pairs with two hydrogen bonds. (A) A self-complementary RNA 14mer with two 5-mC binding sites was co-crystallized with **PZG** monomer. The monomers are stacked with one another and with the ends of two RNA duplexes. Highly ordered water molecules were observed to bridge and bundle three RNA duplexes into a larger prism-like complex. (B) In the RNA 7mer-**PZG** complex, the distance between the primer 3′-hydroxyl group and the **PZG** phosphorus atom was 6.3 Å (black dashes). The corresponding Chemdraw schematic is shown on the right. (C) In the RNA 14mer-**PZG** complex, the distance between primer 3′-hydroxyl group and the first bound **PZG** phosphorus atom was 4.2 Å (black dashes). The corresponding $2F_o$-F_c omit maps contoured at 1.0 σ (cyan mesh) indicate the ordered nucleobases and partial disorder of the sugar, phosphate and pyrazolyl moieties. The corresponding Chemdraw schematic is shown below. (D) Water molecules formed hydrogen bonds with three different neighboring RNA duplexes to facilitate molecular assembly. (E) Bird's-eye view of the prism-like complex, with three symmetric RNA duplexes forming a central hollow channel. Water molecules are centrally located within the channel.

to two consecutive binding sites, the affinity of the second bound monomer is lower than that of the first bound monomer. It is noteworthy that the P−O interatomic distance between the aryl-phosphonate of the first bound monomer and the 3′-hydroxyl group of the primer 3′-terminus was only 4.2 Å, compared to the 6.3 Å distance as observed in the one-binding-site 7mer duplex structure (**Figures 2.2b & c**). While there are significant differences in sequence composition and in the crystallization conditions for the two duplexes, we suggest that the stacking interactions afforded by the second bound monomer may have favored a structural reorganization such that the activated phosphate moiety of the first monomer moved closer to the upstream primer terminus.

In order to explore the possible influences of the pyrazolyl phosphonate groups on the diversity of monomer−template base-pairing motifs, we decided to revisit the crystal structures of the above RNA duplexes, using guanosine monophosphate (GMP) in lieu of PZG. The structures of the 15-, 7-, and 14mer RNA:GMP cocrystals were determined to 1.25 Å, 1.8 Å, and 1.55 Å resolution, respectively (PDB: 5L00, 5DHB, and 5DHC respectively). These structures were all very similar to their respective RNA−PZG complex counterparts, including the GMP-assisted inter-duplex helical slip-stacking, as well as the presence of both Watson−Crick and noncanonical GC base-pairing motifs (**Figure B.1**). For the 15mer:GMP complex, the ribose and phosphate moieties of GMPs at the terminal binding sites remained disordered, as was the case of the corresponding RNA−PZG complex. For the 7mer:GMP complex, however, the conformation of the ribose and phosphate moieties of both template-bound GMPs were much more clearly defined. Unlike the bound PZGs, the two monomers were both in the 3′-endo conformation, as were the ribose sugars in the RNA duplex region. Finally, in the 14mer RNA−GMP

structure, with two GMP molecules binding at each end of the duplex, the ribose and phosphate of the first bound monomer was better ordered than the second bound GMP on both ends of the duplex. As in the corresponding RNA−PZG complexes, the phosphate group of the first bound GMP was found to be much closer to the primer 3′-terminus (roughly 4.2 Å) than in the one-binding-site structure (6.3 Å) (**Figure B.2**). Our results indicate that our observations of both (i) diverse inter-nucleobase binding motifs between the template and bound monomer and (ii) the decreased primer−monomer distance afforded by a downstream bound monomer can probably occur irrespective of the chemical constituents present on the phosphate moiety of the bound monomers.

2.2.3 Three Distinct PZG Binding Modes on a Three-Nucleotide Template. We cocrystallized the PZG monomer with a partially self-complementary RNA 13mer oligonucleotide with a triple 5-mC 5′-overhang (5′-mCmCmCGACUUAAGUC-3′, **Figure 2.3a**). To our surprise, the crystal diffracted to 1.2 Å resolution despite having a relatively flexible three nucleotide 5′-overhang. The high-resolution diffraction data allowed us to solve the structure by *ab initio* methods instead of by molecular replacement as performed previously.[33] An initial model, comprising a single chimeric RNA strand and two GMP molecules, was derived solely from our high-resolution diffraction data, and was subsequently refined. The overall RMSD value between the initial model and the final structure was 1.6 Å. In contrast to the one- and two-binding-site structures elucidated above, the RNA 13mer−PZG complex crystallized with space group C_12_1; one asymmetric unit contained one single-stranded RNA and two PZG monomer molecules (PDB: 5HBY). Each self-complementary RNA duplex bound to a total of six PZG

molecules, three at each end, with all six monomers fitting very well to their corresponding electron densities (B factor <20 for all six monomers) and with all ribose moieties in the

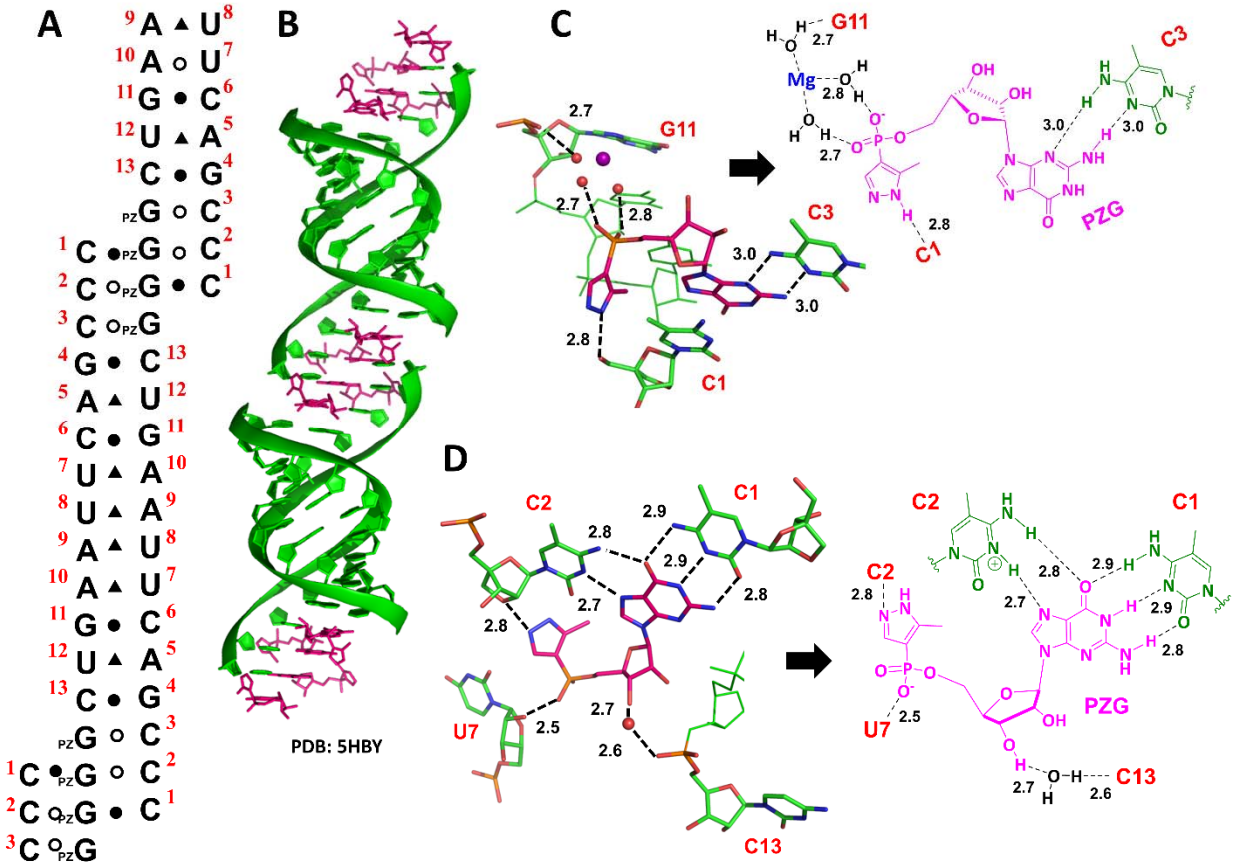


Figure 2.3. Crystal structures of RNA-**PZG** complexes with 5′-overhangs with three **PZG** binding sites. (A) Sequence of the RNA duplex with three 5-mC binding sites at each end, to which **PZG** monomers could bind. Black filled circle: GC Watson-Crick pairs; Black triangle: AU Watson-Crick pairs; Black open circle: noncanonical base-pairs with two hydrogen bonds. (B) The self-complementary RNA duplex was co-crystallized with **PZG** monomers. The monomers bound to the second and third 5-mC binding site bridge two duplexes, constituting a local triplex region. (C) At the first binding site, the **PZG** monomer pairs with a templating 5-mC via a noncanonical base pair, and was connected to the 5′-terminus of another downstream duplex via a hydrogen bond contact with the pyrazolyl group of **PZG**. A magnesium (II) cation and its chelating water molecules were observed in the major groove, and coordinated the **PZG** monomer with the RNA backbone. The base pairing motif, as well as the additional stabilizing hydrogen-bonding interactions are highlighted in the corresponding Chemdraw schematic. (D) At both the second and the

third 5-mC binding sites, a base-triple was observed, with the **PZG** monomer forming both Watson-Crick and Hoogsteen base-pairs with 5-mC on two different template strands. The conformation of the **PZG** is stabilized by hydrogen bond contacts with neighboring RNAs. The chemical structure of the base-triple is illustrated.
3′-endo conformation. Two of the three PZG molecules bound to each end of the duplex were shared with an adjacent duplex, thus explaining the ratio of one oligonucleotide to two PZG monomers.

We observed multiple noncanonical base-pairing interactions between the bound PZG monomers and the template 5-mC residues. At the first binding site, adjacent to the 3′-end of the oligonucleotide (mC3 in **Figure 2.3a**), the guanine nucleobase of PZG paired with the 5-mC template via the same N3-amino, amino-N3 interaction[34] as described above. However, in this higher resolution structure, we were able to observe additional interactions that may also contribute to the stabilization of this unusual structure. The N2 of the pyrazolyl group made a contact with the 5′-OH of the downstream 5-mC nucleobase from another strand (distance 2.8 Å). In addition, a highly ordered hexaaqua-magnesium(II) cation was observed in the major groove of the duplex. Two of the magnesium-coordinated water molecules hydrogen bonded with the two oxygen atoms of the phosphonate moiety of PZG, while another coordinating water molecule hydrogen bonded with the *pro*-R_P oxygen atom of the A10−G11 phosphodiester linkage (**Figure 2.3c**). These observations demonstrate that divalent magnesium ions, together with inner shell coordinated water molecules, could serve to stabilize the binding of the monomer via interactions with the backbone of the duplex RNA to which the monomer is bound.

At the second 5-mC binding site (mC2 in **Figure 2.3a.**), the guanine nucleobase of the PZG was base paired with the template 5-mC in a Hoogsteen fashion (the two hydrogen bonds are 2.7 and 2.8 Å, **Figure 2.3**). In addition, N1 of the PZG pyrazolyl

phosphonate was within hydrogen-bonding distance of O2′ of the 4′-C-methylene-linked bridge of the mC-LNA ribose. Additional hydrogen-bond contacts from nucleotides of adjacent RNA strands also appeared to stabilize this unusual mode of monomer:template binding: the *pro*-R_P nonbonded oxygen from the PZG aryl phosphonate interacted with the 2′-hydroxyl oxygen of U7 from an adjacent duplex, and a water molecule bridged the 3′-OH of PZG and the *pro*-S_P nonbonded oxygen of the U12−C13 phosphodiester linkage from another duplex. Only at the third and terminal binding site (mC1 in **Figure 2.3a**) was the nucleobase of the PZG monomer observed to base pair with the template mC in Watson−Crick fashion (with distances of ∼2.8 Å, **Figure 2.3d**). This PZG monomer was shared by two RNA strands in the same manner as the PZG at the second binding site. It was striking that, at the primer +2 and +3 positions, the PZG monomers bridged two RNA strands by utilizing all available hydrogen-bond donors and acceptors on the Watson−Crick and Hoogsteen faces of the guanine nucleobase (N1, N7, the exocyclic amine, and the carbonyl group), so as to constitute two stacked planar $C{\cdot}G{\cdot}C^+$ triplex motifs.[13,35,36] These two base triples are related by symmetry, and therefore possess identical H-bond interactions and geometry. This local triple helix motif, consisting of both Hoogsteen and Watson−Crick base pairing, a local hydrogen-bonding network, and expanded internucleotide stacking interactions, constitute a highly ordered platform for monomer binding at the +1 position, and for nucleation of the crystallization process (**Figure 2.2d**).

2.2.4 Overall Topology of the One-, Two-, and Three- Binding-Site Complexes. Following elucidation of various monomer−template base-pairing motifs, we have compared the overall topology and crystallization pattern of the duplexes containing one,

two, and three 5-mC binding sites with the corresponding native full-length RNA duplex. The RNA 15mer and 14mer cocrystallized with either PZG or GMP in an almost identical manner, with the resulting RNA−monomer complexes sharing the same space group and asymmetric unit presumably because of the overall sequence and length identity. For example, in the RNA 14mer−PZG complex structure, groups of three adjacent duplexes coordinated four highly ordered water molecules (**Figure 2.2a**, red dots) within the central channel formed by the triangular prism (**Figure 2.2e**). These structured water molecules bridged adjacent duplexes by hydrogen bonding, thus ensuring facile assembly of these duplexes into larger structures. Two of the four water molecules formed three 2.6 Å hydrogen-bond contacts with the three surrounding duplexes via the 2′-hydroxyl of A9 of the RNA sequence, while the remaining two water molecules each formed three 2.8 Å hydrogen-bond contacts with the *pro*-S_P nonbonded oxygen atoms of the A10−G11 phosphodiester linkages of the three prism-forming duplexes. The local structure and coordination environment for these water molecules is depicted in **Figure 2.2d**. The RNA 15mer−PZG complex crystallized in a similar manner, sharing comparable molecular contacts and overall structural assembly.

In contrast to the 15mer and 14mer structures, the 13mer RNA duplex, with three binding sites on its two ends, crystallized with different symmetry and a distinctive packing pattern. Instead of forming water-bridged RNA prisms, four highly ordered water molecules were observed to bridge and stabilize PZG monomers, as well as connecting them to both a neighboring RNA and the RNA duplex they were bound to (**Figure 2.3c & d**).

A detailed comparison of the overall topology of these RNA−monomer structures against the original full length RNA[27] was then generated by superimposing their RNA duplex backbones. **Figures 2.4a & b** show the superimposed structures of the

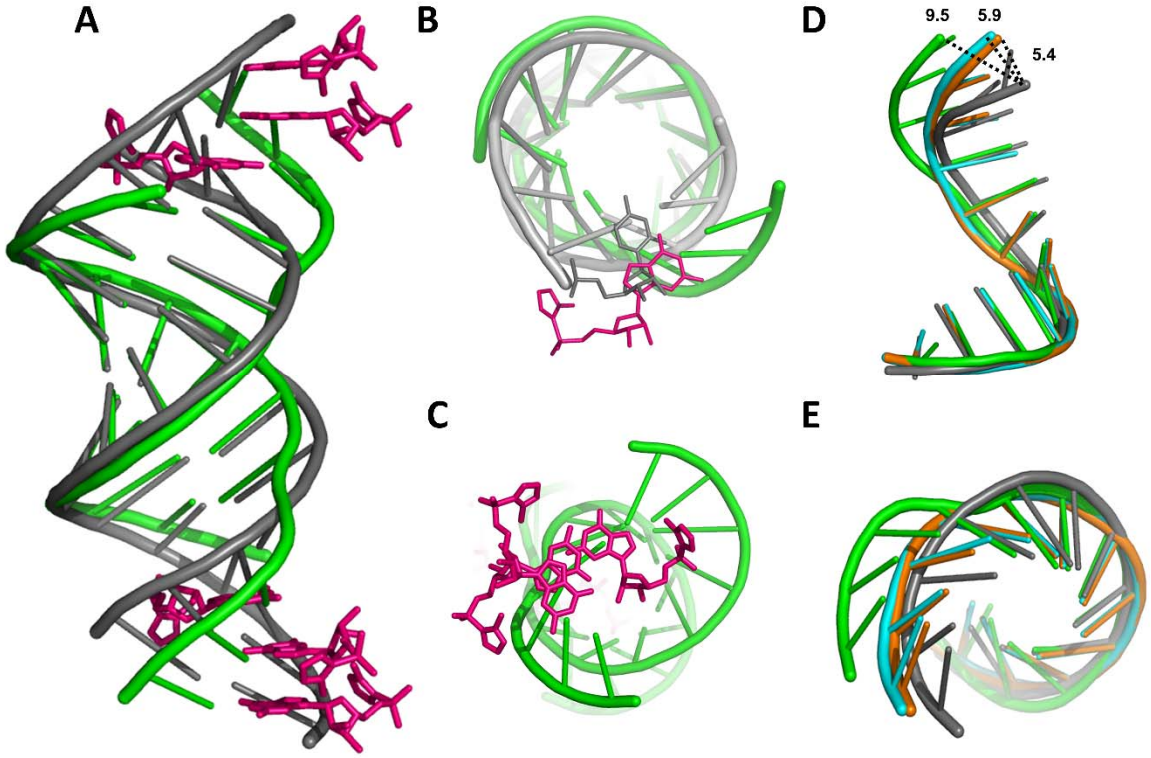


Figure 2.4. Superimposed structures of the native full-length RNA and the RNA-**PZG** complex. Magenta: **PZG** monomer; Gray: native full-length RNA; Green: RNA duplex with three 5-mC binding sites; Cyan: RNA duplex with two 5-mC binding sites; Orange: RNA duplex with one 5-mC binding site. (A) Superimposed structures of the native full-length RNA and the RNA-**PZG** complex with 3 binding sites. (B) Bird's-eye view of the superimposed structures showing location of the second bound **PZG** monomer within the minor groove. (C) The three bound **PZG**s orient their leaving group mimics in different trajectories. (D) and (E): (D) Side view and (E) bird's-eye view of the superimposed structures comprising of full-length RNA, complex with one binding site, complex with two binding sites, and complex with three binding sites. The 5′-hydroxyl termini of the **PZG**-RNA complexes are unwound from the central helical axis, and are shifted by 5.4, 5.9 and 9.5 Å from the corresponding position in the native full-length structure (one, two and three **PZG**s bound, respectively). This generated a wider helix diameter with more open space for monomers.

previously determined full-length RNA 16mer (PDB ID: 4NFO) and our three-binding-site RNA:PZG complex. **Figures 2.4a** (side view), **2.4b**, and **2.4c** (bird's-eye view) show that the first incoming PZG monomer bound to the template by having the guanine base slide directly underneath the primer in a nucleobase-first, "head-on" fashion to engage in non-Watson−Crick base−base recognition. The second and third monomers clearly reside on the wide and shallow minor groove of the RNA duplex, with the nucleobases of PZG twisted and tilted so to open up the minor groove and unwind the duplex (**Figures 2.4a & b**). The ribose and pyrazolyl phosphonate groups of the first bound monomer remained buried inside the major groove of the duplex, while those of the second and third monomers were solvent-exposed and pointed toward the periphery of the duplex with different trajectories (**Figure 2.4c**). In this conformation, the PZG monomers can interact with neighboring RNAs or water molecules. Compared with the native and fully complementary RNA duplex, the 5′-terminus of the one-binding-site duplex overhang shifted by 5.4 Å, and that of the two-binding-site complex shifted by 5.9 Å, while the 5′-terminus of the three-binding-site complex shifted by 9.5 Å to create additional space in the minor groove for monomer binding. The superimposed structures, together with the measured distances between terminal 5′-hydroxyl groups, are shown in **Figures 2.4d & e**.

2.3 Discussion

Given the dominance of the Watson−Crick base-pairing motif in nucleic acid replication chemistry, it has long been thought that the binding of activated monomers to RNA templates would occur predominantly via Watson−Crick base pairing.[6] Indeed, in the context of protein polymerase-assisted RNA or DNA copying, a multitude of

strategically located hydrophobic, electrostatic, and hydrogen-bonding interactions constrain the noncovalent interaction of the RNA template and incoming mononucleotide substrates,[37] and only Watson−Crick base pairing geometry is seen, because this recognition motif is the basis of information transfer during nucleic acid replication. However, under the prebiotic conditions of nonenzymatic RNA replication, where these enzyme-assisted monomer−template interactions are absent, the many hydrogen-bond donor and acceptor groups on the nucleobases allow pairing to occur in a multitude of ways in addition to the usual Watson−Crick constraints. This is especially true when multiple mononucleotides are binding to a single-stranded template, and the lack of spatial and/or conformational constraints allows nucleotide binding in unusual structural modes. In fact, hydrogen bonding between complementary base pairs can occur via nonconventional motifs even when spatial constraints and/or assistive hydrophobic interactions are present: for pairs of complementary monoribonucleotides and deoxyribodinucleotides that were stabilized within the hydrophobic cavity formed by the triazine caps and bipyridyl pillars of a molecular cage, Fujita and co-workers exclusively observed non-Watson−Crick base pairing between the nucleobases of the interacting molecules.[17] Arguably, in the context of monomer:template interactions, the perceived energetic and/or conformational preference of Watson−Crick base pairing over other noncanonical mechanisms may not be as pronounced as previously thought. Our crystallographic studies suggest that the binding of activated monomers to RNA templates might involve, at least transiently, non-Watson−Crick base pairing during nonenzymatic RNA copying in solution. If they do occur in solution, some of the various binding modes of incoming monomers could, in the worst cases, lead to errors in

nonenzymatic RNA polymerization and subsequent stalling of primer extension.[38] Alternatively, some non-Watson−Crick binding modes could also potentially lead to unproductive conformations which displace the activated phosphate of the incoming monomer away from the primer terminus, thus retarding the rate of nonenzymatic RNA polymerization. Strategies to reduce the high error rates seen in nonenzymatic RNA replication are necessary to enable the expansion of the genome, so that larger numbers of ribozymes could be encoded. Ribozyme polymerases, by enforcing Watson−Crick pairing of incoming monomers to the template via geometrical constraints, could significantly increase the fidelity of early RNA replication, thus increasing the coding capacity of primitive RNA genomes.

The experiments described herein represent our initial attempt to derive a structural basis for understanding some of the fundamental mechanisms involved in nonenzymatic RNA copying. In the partial duplexes with one or two nucleotide overhangs, the bound PZG monomers are in some (but not all) cases bound in Watson−Crick fashion to the template C residue. In these cases, the distance from the O3′ of the last nucleotide of the RNA oligonucleotide to the phosphorus of the incoming monomer ranges from 4.5 to 6.5 Å. This distance is far too long to allow for attack of the 3′-hydroxyl on the activated phosphate, as required for primer extension. This suggests that our structures represent a snapshot of the ground state conformations of the bound monomer, and that a significant conformational rearrangement that brings O3′ closer to the reactive P atom must precede initiation of the primer extension reaction. This conformational change could be driven, for example, by binding of the catalytic divalent metal ion, which could potentially bridge the 3′-hydroxyl of the primer and the phosphate of the incoming

monomer. Another contributing factor could be a difference in conformation when one vs two monomers are bound to the template. It is well known that the rate of the primer extension reaction is much faster when two or more monomers are bound to the template than when only a single monomer is present.[39−41] In our structure of the 14mer duplex with single nucleotide overhangs at either end, the monomer pyrazolyl-phosphate is too disordered to measure the O3′−P distance, however in the shorter 6mer duplex with single nucleotide overhangs this distance is 6.5 Å. In our structure of the 12mer duplex flanked by two nucleotide overhangs, the end in which both PZGs are Watson−Crick paired with the template is again too disordered to determine the O3′−P distance. However, at the other end, the first PZG is Watson−Crick paired, with an O3′−P distance of 4.5 Å, still far too long for any reaction to occur, but significantly shorter than in the case with a single bound PZG. The presence of the second, non-Watson−Crick-paired PZG appears to have pushed the first PZG closer to the 3′-OH of the “primer”, possibly representing an initial step along the reaction coordinate. Another possibility is that the reactive conformation is only achieved upon binding, or synthesis on the template, of the imidazolium-bridged dinucleotide reaction intermediate that we have recently described.[42] Additional structures may clarify the role of downstream monomers or oligonucleotides and intermediates, in preorganizing the primer−template−monomer complex for reaction.

We have recently shown that primer extension with an activated monomer can be catalyzed by a downstream activated monomer or oligonucleotide.[41] These observations, together with related earlier observations by Orgel,[39,40] strongly suggested the possibility of a physical interaction between the leaving groups of adjacent template-bound activated monomers. In the primer−template−PZG complexes described above, the pyrazolyl-

phosphonate moieties of most template-bound PZG monomers are disordered, making it impossible to observe any potential noncovalent interactions between the pyrazolyl moieties of consecutively bound monomers. Such interactions, if they occur, may be too transient to see in a crystal structure. Alternatively, the PZG analogue, in spite of mimicking the overall structure of 2-MeImpG (**Figures 1a & b**), may still be too different from 2-MeImpG to exhibit the catalytically relevant interactions. For example, the pK_a values for pyrazoles and imidazoles are quite different.[43] In addition, pyrazole has a charge distribution skewed toward the NH−N single bond, while the charge distribution in the imidazole ring is symmetrically shared between the two nitrogen atoms. The electronic and chemical differences between PZG and 2-MeImpG could prevent adjacent PZGs from forming stable leaving group−leaving group interactions *in crystallo*. We are currently examining other inert 2-MeImpG structural analogue that better reflect the chemical and electronic characteristics of 2-MeImpG, and may serve to illustrate the precise role of leaving group−leaving group interactions in primer extension reactions. Finally, as noted above, we have recently obtained evidence suggesting that primer extension proceeds through a covalent intermediate generated by the reaction of two monomers to form an imidazolium-bridged dinucleotide, which is subsequently attacked by the primer 3′-hydroxyl, resulting in primer extension by one nucleotide with concomitant release of an activated nucleotide as the leaving group.[42] We are currently examining the structure of stable analogues of the proposed intermediate when bound to an RNA template.

In this report, we used locked nucleic acid and 5-methylcytidine-modified RNA as the monomer binding template to enhance crystallization efficiency and complex stability. Although these modifications provide useful models for the study of monomer−template

binding during nonenzymatic RNA replication, it is also possible that these structural modifications may introduce artifacts that do not necessarily reflect solution-phase activated monomer-driven RNA replication. For example, the higher conformational rigidity of the LNA overhang could facilitate the formation of weak non-Watson−Crick base pairs between PZG and the template. Finally, the high effective molarity of RNA duplexes and guanosine monomers in the RNA crystals implies that some of the *in crystallo* interactions we observed—for example, the stabilization of PZG monomers in the 15mer RNA complex by a neighboring slip-stacked RNA duplex, or the simultaneous noncovalent interactions observed between the PZG monomer and two RNA strands in the three-binding-site complex—result from the close intermolecular contacts during crystallization, and may not be relevant to monomer−template recognition in dilute aqueous solution.

One of the noncanonical base pairs that we observed several times is the C·G N3-amino, amino-N3 interaction.[30,34] This base-pairing motif belongs to the family of trans Watson−Crick/sugar-edge interactions,[34] and has been observed in larger RNA complexes such as the group I self-splicing intron of *Tetrahymena thermophila*,[44,45] and the peptidyl transferase center[46] of the 23S rRNA embedded within the 50S large ribosomal subunit from *Haloarcula marismortui*.[47] On the other hand, Hoogsteen base pairing,[48] as observed in the three binding-site RNA complex, is arguably more prevalent in nature, as evidenced by its integral roles in the formation of various triplex and quadruplex motifs. Recently, NMR studies have demonstrated that, at any given time, ∼0.1−1% of Watson−Crick base pairs in a duplex DNA can transiently open up and switch to a Hoogsteen-like conformation, with the Hoogsteen-like base pairs having lifetimes on

the order of milliseconds.[48] Considered together, these experimental observations demonstrate the possible omnipresence of noncanonical base-pairing motifs in the prebiotic world and contemporary biology.

In summary, we have designed and synthesized the pyrazolyl phosphonate nucleotide PZG, which is an inert structural mimic of 2-MeImpG. High-resolution X-ray crystal structures of this ligand bound to RNA duplexes with various overhang lengths revealed an unexpected variety of monomer−template binding motifs. Our work reveals that, in addition to canonical Watson−Crick base pairing, the guanine base of PZG can engage with RNA templates via noncanonical binding motifs by taking advantage of all available hydrogen-bond donor and acceptor groups of the nucleobase. Our work demonstrates the utility of rational analogue design and X-ray crystallography in teasing out mechanistic aspects of nonenzymatic RNA replication. We are currently expanding our use of crystallography to explore the mechanistic roles of divalent metal ions and monomer−monomer interactions during phosphodiester bond formation.

2.4 References

(1) Crick, F. H. *J. Mol. Biol.* **1968**, *38*, 367.

(2) Orgel, L. E. *J. Mol. Biol.* **1968**, *38*, 381.

(3) Woese, C. R. *The Genetic Code: The Molecular Basis for Genetic Expression*; Harper & Row: New York, 1967; pp 179−195.

(4) Szostak, J. W. *J. Syst. Chem.* **2012**, *3*, 2.

(5) Robertson, M. P.; Joyce, G. F. *Cold Spring Harbor Perspect. Biol.* **2012**, *4*, a003608.

(6) Orgel, L. E. *Crit. Rev. Biochem. Mol. Biol.* **2004**, *39*, 99.

(7) Inoue, T.; Orgel, L. E. *J. Mol. Biol.* **1982**, *162*, 201.

(8) Izgu, E. C.; Fahrenbach, A. C.; Zhang, N.; Li, L.; Zhang, W.; Larsen, A. T.; Blain, J. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2015**, *137*, 6373.

(9) Kervio, E.; Claasen, B.; Steiner, U. E.; Richert, C. *Nucleic Acids Res.* **2014**, *42*, 7409.

(10) Kanavarioti, A.; Bernasconi, C. F.; Alberas, D. J.; Baird, E. E. *J. Am. Chem. Soc.* **1993**, *115*, 8537.

(11) Kanavarioti, A.; Bernasconi, C. F.; Baird, E. E. *J. Am. Chem. Soc.* **1998**, *120*, 8575.

(12) Inoue, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1981**, *103*, 7666.

(13) Saenger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, 1984; pp 116−158.

(14) Philp, D.; Stoddart, J. F. *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 1154.

(15) Howard, F. B.; Frazier, J.; Lipsett, M. N.; Miles, H. Y. *Biochem. Biophys. Res. Commun.* **1964**, *17*, 93.

(16) Howard, F. B.; Frazier, J.; Singer, M. F.; Miles, H. T. *J. Mol. Biol.* **1966**, *16*, 415.

(17) Sawada, T.; Yoshizawa, M.; Sato, S.; Fujita, M. *Nat. Chem.* **2009**, *1*, 53.

(18) Zhang, N.; Zhang, S.; Szostak, J. W. *J. Am. Chem. Soc.* **2012**, *134*, 3691.

(19) Egli, M. *Curr. Opin. Chem. Biol.* **2004**, *8*, 580.

(20) Sheng, J.; Li, L.; Engelhart, A. E.; Gan, J.; Wang, J.; Szostak, J. W. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 3050.

(21) Sheng, J.; Larsen, A.; Heuberger, B. D.; Blain, J. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2014**, *136*, 13916.

(22) Lohrmann, R.; Bridson, P. K.; Orgel, L. E. *Science* **1980**, *208*, 1464.

(23) Dunn M.F. (1975) Mechanisms of zinc ion catalysis in small molecules and enzymes. In: Biochemistry. Structure and Bonding, vol 23. Springer, Berlin, Heidelberg.

(24) Kalek, M.; Ziadi, A.; Stawinski, J. *Org. Lett.* **2008**, *10*, 4637.

(25) McKenna, C. E.; Higa, M. T.; Cheung, N. H.; McKenna, M.-C. *Tetrahedron Lett.* **1977**, *18*, 155.

(26) Błażewska, K. M. *J. Org. Chem.* **2014**, *79*, 408.

(27) Phelps, K. J.; Ibarra-Soza, J. M.; Tran, K.; Fisher, A. J.; Beal, P. A. *ACS Chem. Biol.* **2014**, *9*, 1780.

(28) Fox, J. J.; Van Praag, D.; Wempen, I.; Doerr, I. L.; Cheong, L.; Knoll, J. E.; Eidinoff, M. L.; Bendich, A.; Brown, G. B. *J. Am. Chem. Soc.* **1959**, *81*, 178.

(29) Krissinel, E.; Winn, M.; Ballard, C.; Ashton, A.; Patel, P.; Potterton, E.; McNicholas, S.; Cowtan, K.; Emsley, P. *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60*, 2250.

(30) Lemieux, S.; Major, F. *Nucleic Acids Res.* **2002**, *30*, 4250.

(31) Thota, N.; Li, X. H.; Bingman, C.; Sundaralingam, M. Acta Crystallogr. D Biol. Crystallogr. **1993**, *49*, 282.

(32) Tam, C. P.; Fahrenbach, A. C.; Björkbom, A.; Prywes, N.; Izgu, E. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 571.

(33) Foadi, J.; Woolfson, M. M.; Dodson, E. J.; Wilson, K. S.; Jia-xing, Y.; Chao-de, Z. *Acta Crystallogr. D Biol. Crystallogr.* **2000**, *56*, 1137.

(34) Leontis, N. B.; Westhof, E. *RNA* **2001**, *7*, 499.

(35) Bloomfield, V. A.; Crothers, D. M.; Tinoco, I. *Nucleic Acids: Structures, Properties, and Functions*; University Science Books: Sausalito, CA, 2000; pp 259−334.

(36) Neidle, S. *Principles of Nucleic Acid Structure*, 1st ed.; Academic Press: London, 2008; pp 81−131.

(37) Westover, K. D.; Bushnell, D. A.; Kornberg, R. D. *Cell* **2004**, *119*, 481.

(38) Rajamani, S.; Ichida, J. K.; Antal, T.; Treco, D. A.; Leu, K.; Nowak, M. A.; Szostak, J. W.; Chen, I. A. *J. Am. Chem. Soc.* **2010**, *132*, 5880.

(39) Wu, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1992**, *114*, 5496.

(40) Wu, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1992**, *114*, 7963.

(41) Prywes, N.; Blain, J. C.; Del Frate, F.; Szostak, J. W. *eLife* **2016**, e17756.

(42) Walton, T.; Szostak, J. W. *J. Am. Chem. Soc.* **2016**, *138*, 11996.

(43) Joule, J. A.; Mills, K. *Heterocyclic Chemistry*, 5th ed.; Wiley-Blackwell: Hoboken, NJ, 2013; pp 56−57.

(44) Juneau, K.; Podell, E.; Harrington, D. J.; Cech, T. R. *Structure* **2001**, *9*, 221.

(45) Battle, D. J.; Doudna, J. A. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 11676.

(46) Puglisi, E. V.; Green, R.; Noller, H. F.; Puglisi, J. D. *Nat. Struct. Biol.* **1997**, *4*, 775.

(47) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. *Science* **2000**, *289*, 905.

(48) Nikolova, E. N.; Zhou, H.; Gottardo, F. L.; Alvey, H. S.; Kimsey, I. J.; Al-Hashimi, H. M. *Biopolymers* **2013**, *99*, 955.

Chapter 3

Insight into the Mechanism of Nonenzymatic RNA Primer Extension from the Structure of an RNA-GpppG Complex

[*W. Zhang and J. Szostak conceived of experiments; W. Zhang performed experiments. W. Zhang, C. P. Tam and J. Szostak analyzed the data and wrote the texts. C. P. Tam performed NMR assays for GpppG-RNA binding, and T Walton performed saturation kinetics analysis for Gp-Im-pG-RNA binding. A. C. Fahrenbach and G. Birrane provided technical assistance. This chapter has previously been published with minor changes as the following publication:* Zhang, W.; Tam, C. P.; Walton, T.; Fahrenbach, A. C.; Birrane, G.; Szostak, J. W. Insight into the Mechanism of Nonenzymatic RNA Primer Extension from the Structure of an RNA-GpppG Complex. *Proc. Natl. Acad. Sci. U. S. A.* **2017***, 114* (29), 7659–7664.]

Abstract. The nonenzymatic copying of RNA templates with imidazole-activated nucleotides is a well-studied model for the emergence of RNA self-replication during the origin of life. We have recently discovered that this reaction can proceed through the formation of an imidazolium-bridged dinucleotide intermediate that reacts rapidly with the primer. To gain insight into the relationship between the structure of this intermediate and its reactivity, we cocrystallized an RNA primer–template complex with a close analog of the intermediate, the triphosphate-bridged guanosine dinucleotide GpppG, and solved a high-resolution X-ray structure of the complex. The structure shows that GpppG binds the

RNA template through two Watson–Crick base pairs, with the primer 3′-hydroxyl oriented to attack the 5′-phosphate of the adjacent G residue. Thus, the GpppG structure suggests that the bound imidazolium-bridged dinucleotide intermediate would be preorganized to react with the primer by in-line S_N2 substitution. The structures of bound GppG and GppppG suggest that the length and flexibility of the 5′-5′ linkage are important for optimal preorganization of the complex, whereas the position of the 5′-phosphate of bound pGpG explains the slow rate of oligonucleotide ligation reactions. Our studies provide a structural interpretation for the observed reactivity of the imidazolium-bridged dinucleotide intermediate in nonenzymatic RNA primer extension.

3.1 Introduction

In the RNA world hypothesis, the emergence of RNA-catalyzed RNA replication is thought to have been preceded by a stage in which RNA replication was driven purely through chemical processes[1,2]. The nonenzymatic RNA-templated polymerization of activated nucleotides or oligonucleotides has been extensively studied, with the intent of optimizing the rate, extent, and fidelity of nonenzymatic RNA/DNA polymerization[3,4]. Numerous phosphate-activating groups have been studied in the context of nonenzymatic RNA replication. For example, imidazoles such as 2-methylimidazole[5] and, more recently, 2-aminoimidazole[6], have been found to be useful phosphate activators. The potentially prebiotic synthesis of imidazoles under primitive Earth conditions has been investigated.[7] On the other hand, Richert and coworkers reported the use of benzotriazole-activated monomers to improve the rate of primer extension.[8] In an alternative approach, the *in situ* activation of monoribonucleotides, and subsequent template-guided polymerization, has been achieved by Richert and coworkers,[9] using a carbodiimide reagent together with *N*-alkyl-imidazole catalysts.

Many of the thermodynamic and kinetic parameters associated with nonenzymatic RNA replication have been quantitatively determined.[10–14] Until recently, the general assumption has been that nonenzymatic primer extension with activated mononucleotides involves classical S_N2 nucleophilic substitution, in which the nucleophilic 3′-hydroxyl group of the primer attacks the activated phosphorus center of the adjacent template-bound monomer via an in-line mechanism.[15] The reaction is known to be catalyzed by the presence of an activated downstream monomer or oligomer, an effect initially believed to result from a noncovalent interaction between the leaving groups

of adjacent nucleotides, such that the reactive site becomes preorganized for in-line displacement.[16,17] However, recent work from our laboratory strongly indicates that an imidazolium-bridged dinucleotide is a covalent intermediate in the reaction[18]. In this model of primer extension, two activated monomers in solution react with each other, forming the imidazolium-bridged dinucleotide (**Figure 3.1a**), which then binds the RNA template. As a dinucleotide, the intermediate could potentially form two Watson–Crick base pairs and thus bind more tightly than a monomer. An even more important question is whether the structure of the complex could favor nucleophilic attack of the primer on the intermediate (**Figure 3.1b**). As yet, there is no structural evidence as to how the dinucleotide intermediate binds the RNA and whether the conformation would be favorable for nucleophilic attack.

Despite these recent studies, relatively few mechanistic investigations have been based on crystallographic approaches. We have previously used crystallographic approaches to investigate various aspects of nonenzymatic RNA polymerization, including the effect of 2′-5′ phosphodiester linkages and 2-thio-U substitution on RNA structure.[19–21] To study either enzymatic or nonenzymatic replication reactions, chemically stable analogs of substrates are typically necessary. For example, the nonhydrolyzable α−β imido ATP analogs have been intensively applied to structural studies of the mechanism of catalysis by DNA polymerase.[22] In our laboratory, the nonhydrolyzable guanosine 5′-(3-methyl-1*H*-pyrazol-4-yl)phosphonate (PZG) was designed and synthesized to mimic the activated monomer 2-MeImpG.[23] Our crystal structures of RNA-PZG complexes revealed both Watson–Crick and, surprisingly,

noncanonical base pairs, suggesting that mismatched template–monomer base pairing may be more common than expected. More significantly, the phosphate and leaving

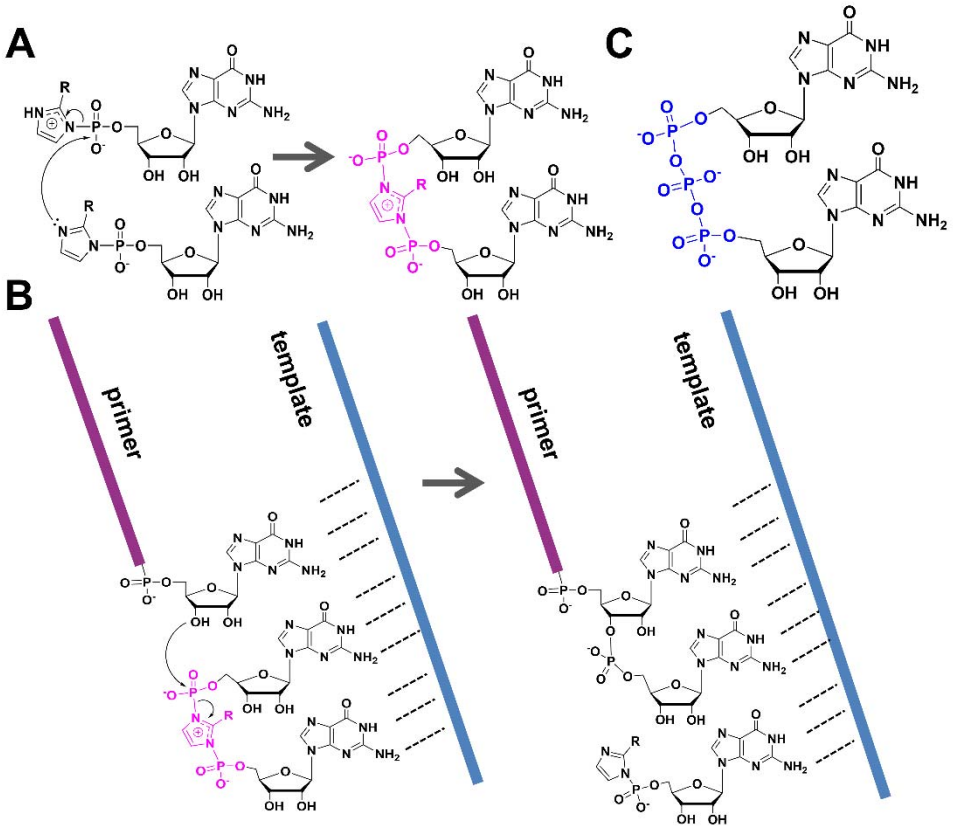


Figure 3.1. Mechanism of nonenzymatic primer extension by guanosine phosphoroimidazolide monomers. A: Two monomers react to form the imidazolium-bridged diguanosine intermediate. B: The intermediate binds to the RNA template through two Watson-Crick base pairs. The 3'-hydroxyl of the primer attacks the adjacent phosphorus centre, leading to the S_N2 reaction and primer extension. C: Structure of GpppG, a close analogue of the imidazolium-bridged intermediate. R= $-CH_3$ or $-NH_2$. The bridging moieties of the intermediate and the analogue are colored pink and blue, respectively.

group portions of these structures were always disordered. The absence of any detectable interaction between the leaving group analogs of adjacent monomers suggested that the catalytic effect of a downstream activated nucleotide might not be due to a noncovalent interaction and raised the question of whether the imidazolium-bridged

dinucleotide intermediate might exhibit a more ordered structure that could help to explain its high reactivity.

To address the mechanism of nonenzymatic RNA polymerization, we used X-ray crystallography to provide high-resolution structures of RNAs complexed with stable analogs of the reactive imidazolium-bridged intermediate. The readily available compound, P^1,P^3-diguanosine-5′-triphosphate (GpppG), is a reasonable analog of the imidazolium-bridged dinucleotide, GpImpG. GpppG has a central linker of similar length and flexibility, but the labile N-P bonds are replaced by more stable O-P bonds (**Figure 3.1c**). In GpppG, the flanking phosphates are separated by a central phosphate, which is similar in size to the central imidazolium in GpImpG. Furthermore, the net charge of the imidazolium-bridged intermediate is the same as that of GpppG coordinated to Mg^{2+} because the imidazolium moiety carries one positive charge, whereas the combination of a bound Mg^{2+} ion with the central phosphate of GpppG again results in a net single positive charge. Finally, we have recently shown that 2-aminoimidazole–activated monomers are superior substrates for primer extension. The corresponding 2-aminoimidazolium–bridged intermediate may adopt a conformation in which the exocyclic amino group forms hydrogen bonds with the adjacent nonbridging oxygens of the flanking phosphates, generating a compact p-NH_2Im-p structure that is similar to that of the triphosphate in the Mg^{2+}-GpppG complex. GpppG is also an analog of the m7G(5′)ppp(5′)G cap structure, which is present at the 5′-terminus of most eukaryotic and viral mRNAs and promotes translation initiation both in vitro and in vivo.[24,25] However, the ability of 5′-5′-linked dinucleotides to bind RNA has not been investigated through a structural approach. Here we report insights into the mechanism of nonenzymatic RNA

primer extension derived from high-resolution crystal structures of GpppG and three other G-G dinucleotides complexed with the same RNA primer–template complex.

3.2 Results

3.2.1 Overall Structural Features of the RNA-GpppG Complex. We cocrystallized a series of guanosine dinucleotides with the RNA 5′-**mCmCmC**GACUUAAGUCG-3′[23] (Appendix C provides experimental details). The first three nucleotides in bold are 5-methylcytidine LNA residues, designed to favor and rigidify the A-form strand conformation[26] and thereby facilitate RNA crystallization. At each end of the RNA duplex, the **mCmC** overhang serves as the binding site for G-G dinucleotides because the **mC** residues form Watson–Crick base pairs with G.[23,27] The RNA cocrystallized with the GpppG ligand with hexagonal symmetry, as in our previous RNA-GMP and RNA-PZG complexes.[23] We determined the structure to 1.9-Å resolution by molecular replacement, with an overall B-factor of 32.42. The space group is $P3_121$, and there is one RNA duplex with two bound GpppG molecules per asymmetric unit. At each end of the RNA duplex, the overhanging **mCmC** binding sites are fully occupied through Watson–Crick base pairs with GpppG (**Figure 3.2a & b**). As in our previously determined RNA–monomer complex structures, the RNA double helices are A-form, all sugars are in the C3′-*endo* conformation, and the duplexes slip-stack on each other to form extended columns. Groups of three RNA duplex–ligand complexes form triangular prisms, and the central channel accommodates at least three water molecules that bridge the neighboring duplexes (**Figure 3.2c**). Two symmetry-related water molecules form three 2.5-Å H-bond contacts with the three surrounding duplexes via the 2′-hydroxyls of their G4 residues (**Appendix C, Figure C.2a**). The other water (and a possible fourth water molecule at a

symmetry-related position, but with very weak density) appears to engage in three 3.1-Å hydrogen bonds with the *pro*-S_P nonbonded oxygen atoms of three G4–A5 phosphodiester linkages. However, this may be a time-averaged view of a water molecule that at any given moment is acting as an H-bond donor to two nonbridging phosphate oxygens. In addition, Mg^{2+} ions link the three adjacent complexes by coordinating with both the 2′- and 3′-hydroxyl groups of the guanosine at the primer +1 position of three adjacent GpppG ligands, forming a total of six ~2.4-Å electrostatic interactions. (**Figure 3.2d**).

3.2.2 Structure of GpppG Bound to RNA. A GpppG ligand is Watson–Crick base paired to the 5′-**mCmC** overhang at each end of the RNA duplex; the six hydrogen bond contact distances range from 2.8 Å to 3.0 Å (**Figure 2e & f**). The entire GpppG molecule is well-ordered. The two guanine nucleobases are coplanar and stacked with the upstream primer and the neighboring duplex with interplanar distances of ~3.3 Å. Both ribose sugars of the GpppG are in the C3′-*endo* A-form conformation, consistent with our previous observation of a C3′-*endo* sugar pucker when activated guanosine ribonucleotides bind to a template in solution.[28] Moreover, the GpppG triphosphate linkage is well-ordered due to a Mg^{2+} ion that is coordinated with three nonbridging oxygen atoms, one from each phosphate of GpppG. The three electrostatic interaction distances are 2.3 Å, 2.5 Å, and 3.0 Å. This coordination with Mg^{2+} results in the triphosphate bridge of GpppG being buckled in a well-defined manner (**Figure 2e & f**).

Because GpppG was chosen as a close analog of the imidazolium-bridged dinucleotide GpImpG, we were interested in whether GpppG is bound in a conformation

consistent with the observed high reactivity of the GpImpG intermediate with the primer.

The distance between the primer 3′-hydroxyl and the phosphorus atom of the closest

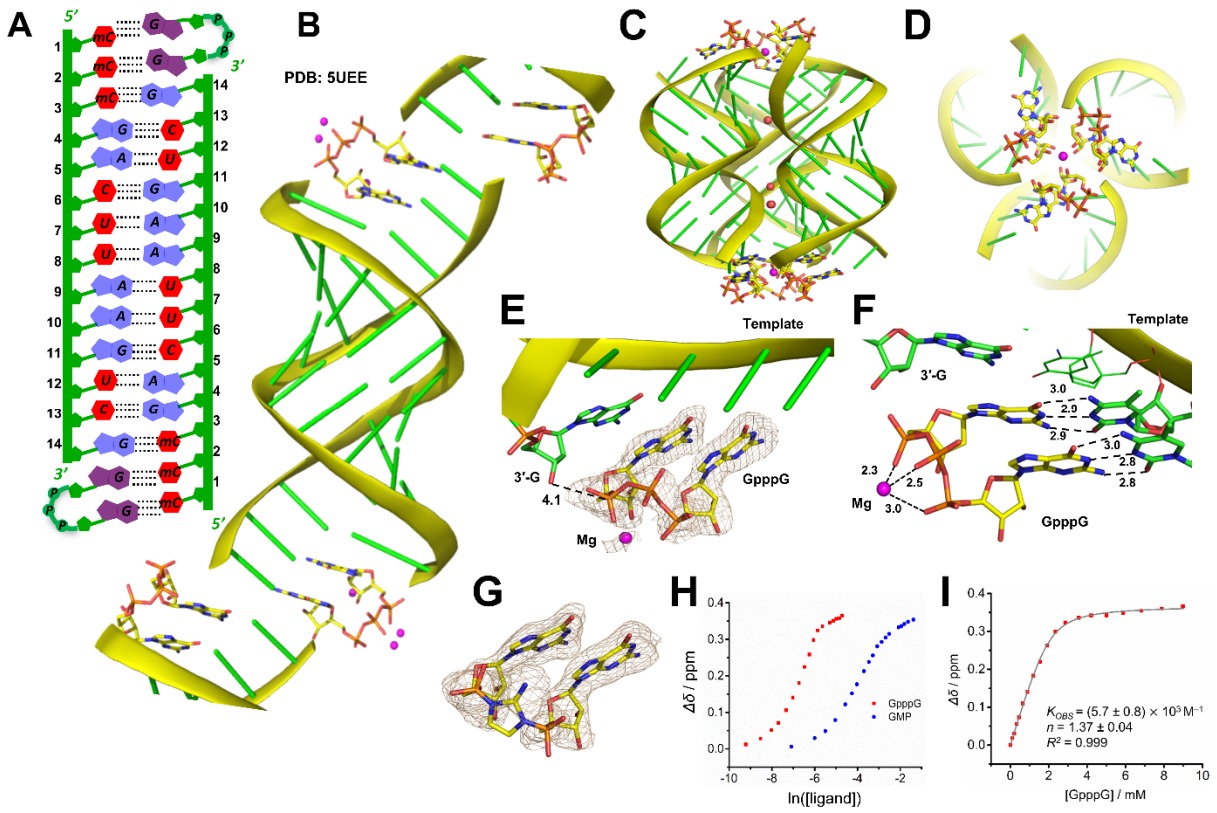


Figure 3.2. RNA-GpppG crystal structure and affinity measurement. Yellow stick: GpppG ligand. Green stick: RNA molecule. Red dot: water molecule. Magenta dot: Mg atom. All the wheat meshes indicate the $F_o - F_c$ omit map contoured at 1.5σ. (A) Schematic of the RNA-GpppG complex. (B) Overall structure of the RNA-GpppG complex. (C) and (D) Ordered water molecules and Mg atoms bridge sets of three adjacent RNA-GpppG complexes. (E) The local structure of the GpppG ligand bound to RNA template. The corresponding omit map indicates the ordered GpppG ligand. (F) GpppG forms two Watson-Crick base pairs with the template through six hydrogen bonds. (G) Gp-NH_2Im-pG ligand is constructed and applied in refinement to fit the density of GpppG. (H) The chemical shift of the G imino proton changes significantly as ligand binding takes place. The significant leftward shift of the sigmoidal plot for GpppG shows that GpppG binding is significantly stronger than GMP. (I) The change in chemical shift of the same G imino proton plotted against the concentration of GpppG, fitted to a modified single-site binding isotherm.

phosphate of GpppG is 4.1 Å, and the angle between the 3′-OH and the bridging P-O bond of GpppG is 126° (**Table 3.1**). We then asked whether the actual imidazolium-bridged intermediate could potentially adopt a similar conformation to the observed conformation of GpppG when bound to the RNA primer–template. We constructed a 2-aminoimidazolium–bridged diguanosine model ($GpNH_2ImpG$) and applied it in the restrained refinement in place of GpppG.[29] We discovered that the GpNH2ImpG ligand fit reasonably well to the electron density of GpppG (**Figure 3.2g**). The observed electron density fits both the nucleobases and sugars of Gp-NH2Im-pG very well (B-factors: nucleobase, 27.5; sugar, 36.4), and the Gp-NH2Im-pG molecule formed two Watson–Crick base pairs with the template in the same manner as GpppG. The B-factors of the two phosphorus atoms in $GpNH_2ImpG$ (37.6 and 45.6) were comparable with the corresponding atoms in GpppG (31.6 and 43.4), and the distance between the P1 and P3 atoms in GpNH2ImpG is only marginally longer than that in GpppG (4.8 Å vs. 4.5 Å). The fact that $GpNH_2ImpG$ can be modeled to fit the density corresponding to GpppG demonstrates the potential structural similarity between the imidazolium-bridged diguanosine intermediate and the stable GpppG analog, when bound to an RNA primer–template complex.

Table 3.1. Crystallographic structure features.

	5'-5' linkage	binding motifs	3'-O--P distance, Å	3'-O--P-O angle
RNA-GpppG	triphosphate	Watson-Crick	4.1	126°
RNA-GppG	pyrophosphate	Watson-Crick/ noncanonical	4.8	170°/ND
RNA-GppppG	tetraphosphate	Watson-Crick	~4.4	ND
RNA-pGpG	Monophosphate (3'-5')	Watson-Crick	4.9	ND

In order for the proposed dinucleotide intermediate in nonenzymatic RNA polymerization to bind and react with the primer, it must compete with the more abundant free monomer for template occupancy. We therefore set out to measure the strength of GpppG binding to the RNA template relative to GMP. NMR methods have been used to measure the affinity of GMP for RNA primer–templates.[10,12–14] Here we used the same methods and RNA primer–template system[10] to determine the affinity of GpppG for an RNA template overhang consisting of two consecutive cytidines. In the experiment, the concentrated GpppG solution was titrated into the RNA primer–template complex solution, and the binding constant was then calculated from the change in the chemical shift of the imino proton of the primer 3′-G vs. the concentration of GpppG (Appendix C provides detailed procedures). Strikingly, the observed K_d of GpppG is ~0.2 mM, which is ~100-fold greater than our previously determined measurement of the affinity of GMP, which has a K_d of ~20 mM (**Figure 3.2h & i**). The stronger binding of GpppG most likely reflects its two Watson–Crick base pairs with the template. We then asked whether the aminoimidazolium-bridged diguanosine intermediate $GpNH_2ImpG$ would have a similar affinity to the same RNA substrate. To address this question, we isolated $GpNH_2ImpG$ in ~80% purity, with 2-NH_2ImpG as impurity, and used it as the substrate in a primer extension reaction using the same primer–template complex as for the K_d measurement of GpppG. Interestingly, the K_M is ~0.6 mM, which is comparable to the K_d of the GpppG analog, given the different conditions required for each assay (Appendix C provides details). This affinity measurement suggests that the GpNH2ImpG intermediate and the structural analog GpppG may compete effectively with monomers for binding to the template during nonenzymatic RNA polymerization reactions.

3.2.3 G(5′)pp(5′)G Binds to RNA Template Through Two Different Motifs. To understand the properties of the 5′-5′-linkage that make the imidazolium-bridged dinucleotide an appropriate intermediate, we cocrystallized the same RNA sequence with other guanosine dinucleotides that have different lengths or types of bridging linkages. Crystals of all the RNA–dinucleotide complexes grew with hexagonal symmetry (**Appendix C, Tables S2 and S3**). The overall structures were similar to that of the RNA-GpppG complex, with the same molecular packing patterns, including the slip-stacked RNA double helices, the triangular-prism structure formed by groups of three RNA–ligand complexes, and the binding of the dinucleotide ligand with the **mCmC** binding sites at the RNA terminus. We first examined diguanosine-5′-5′-pyrophosphate (GppG), which is known to form in trace amounts in activated monomer solutions as a result of the reaction of GMP, formed by hydrolysis, with an activated monomer. Complementary templates can catalyze the synthesis of pyrophosphate-linked dinucleotides from activated monomers.[30] These pyrophosphate-bridged dinucleotides likely inhibit nonenzymatic RNA polymerization, presumably through competitive binding of the RNA template.

The mode of binding and the rigidity of the pyrophosphate linkage of GppG are subtly different from that of GpppG. At one end, GppG forms two Watson–Crick base pairs with the templating **mC**s, but the pyrophosphate linkage and sugars are disordered (**Figure 3.3b**), making it difficult to define the geometry of the pyrophosphate and the conformation of the GppG sugars. Remarkably, GppG binds to the template in a distinctly different manner at the other end. The guanosine adjacent to the primer is Watson–Crick base paired with the template **mC** through three hydrogen bonds as expected. However, the second guanosine forms a noncanonical G:C base pair with two hydrogen bonds: a

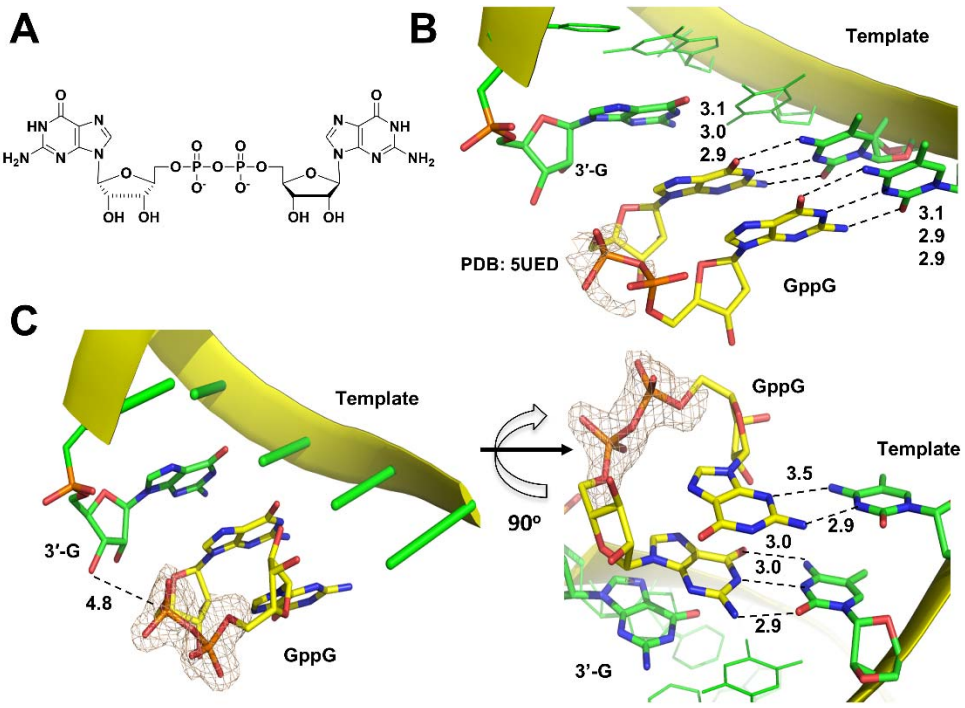


Figure 3.3. Structure of the RNA-GppG complex. RNA and ligand structures and $F_o - F_c$ omit map are represented as metioned above. (A) Chemical structure of GppG ligand. (B) At one end of RNA duplex, GppG forms two Watson-Crick base pairs with template, but the diphosphate linkage and sugar moieties are disordered. (C) At the other end of the duplex, GppG forms two different base pairs with the template. The diphosphate linkage is highly ordered.

weak 3.5-Å H-bond between the guanine N3 and the exocyclic amine of **mC** and a second 2.9-Å H-bond between the exocyclic amine of the guanine and the N3 of the **mC**. We previously observed the same type of noncanonical G:C base pair in our RNA-GMP and RNA-PZG structures, suggesting that this structure could play an important role in modulating the efficiency and fidelity of nonenzymatic RNA replication.[23] This GppG is well-ordered overall, including nucleobases, sugars, and the pyrophosphate linkage. The first ribose is in the C3'-endo conformation, whereas the second one is in the C2′-endo conformation. The distance between the 3′-hydroxyl group of the primer and the phosphorus center of pyrophosphate linkage is 4.8 Å (**Table 3.1** and **Figure 3.3c**),

significantly longer than for the GpppG complex. Although the poor solubility of GppG prevented us from measuring the affinity of GppG for the RNA primer–template, the overall structure of the RNA-GppG complex suggests that GppG likely binds the RNA template with a similar affinity as GpppG.

3.2.4 Tetraphosphate of G(5′)pppp(5′)G Bound to an RNA Template is Disordered. We then cocrystallized the same RNA duplex with diguanosine-5′,5′-tetraphosphate (GppppG), in which the linkage between the two nucleosides is longer and possibly more flexible than that of GpppG. In the complex structure, GppppG binds to the RNA through two well-ordered G:C Watson–Crick base pairs at both ends, but the tetraphosphate linkage is disordered. The electron density associated with the tetraphosphate linkage (**Figure 3.4b**) is significantly larger and more globular than that of GpppG and can be modeled as two GppppG ligands with different conformations (each with an occupancy of 0.5). The distance between the 3′-hydroxyl group of the primer and the adjacent phosphorus center is about 4.4 Å. The disordered structure of the tetraphosphate bridge suggests that any bridge with similar length and flexibility is unlikely to be optimal for the primer extension reaction due to unfavorable reaction angles for in-line attack by the primer.

3.2.5 pGpG Binds the RNA Template Through Watson–Crick Base Pairs. We cocrystallized the RNA with the 5'-phosphorylated GG dimer pGpG to compare this 3′-5′ linked oligonucleotide with the previously studied 5′-5′ bridged dinucleotides. At both ends of the RNA duplex, the pGpG dimer binds to the **mCmC** binding sites through two Watson–Crick base pairs. However, the two sugars are partially disordered, making it impossible to define the ribose conformation. The 5′-phosphate of the dimer is displaced

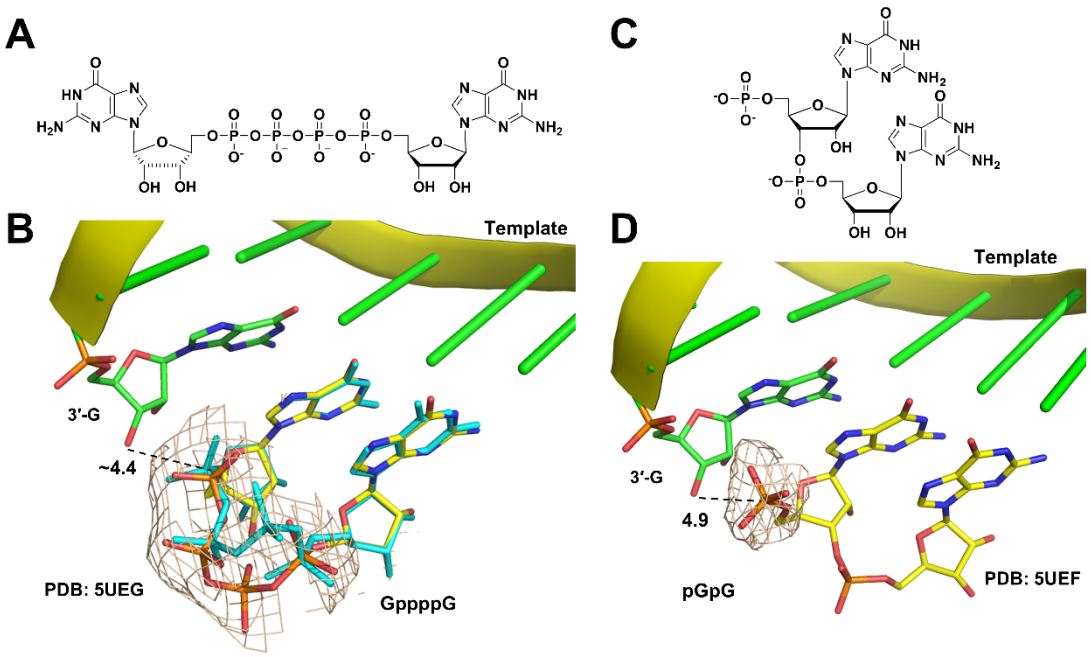


Figure 3.4. Chemical and crystal structures of RNA-GppppG and RNA-pGpG complexes. (A) Chemical structure of GppppG. (B) Local crystal structure of GppppG. The tetraphosphate is disordered. Two potential conformations with 0.5 occupancy each are presented by the yellow and cyan sticks. (C) Chemical structure of pGpG. (D) Local crystal structure of pGpG bound to the RNA template.

toward the major groove, and the distance between the 3′-hydroxyl of the primer and the phosphorus atom of the phosphate is 4.9 Å (**Figure 3.4d**). This distance is the longest of all the complex structures and is consistent with the previous observation that primer extension by ligation is much slower than by polymerization with activated monomers in the presence of an activated downstream helper monomer or oligonucleotide.[16]

3.3 Discussion

Understanding the mechanism of template-directed nonenzymatic primer extension may lead to improved and/or more prebiotically realistic ways of driving RNA replication without enzymes. We have recently shown that nonenzymatic primer extension with 2-methylimidazole–activated nucleotide[5] can proceed via a two-step process, in which two

activated monomers first react with each other to form an imidazolium-bridged dinucleotide intermediate.[18] Once formed and bound to the template, this intermediate reacts rapidly with the primer. This surprising reaction mechanism and the unusual structure of the intermediate raise several important questions. Here we have focused on the interaction of the intermediate with the template: Can the intermediate bind to the template through Watson–Crick base pairing of both of its nucleobases? Once bound, is there some aspect of its conformation that would favor reaction with the primer? To address these questions, we have taken advantage of the structural similarity between the triphosphate-bridged dinucleotide, GpppG, and the imidazolium-bridged diguanosine intermediate, GpImpG. The 2-methylimidazole/2-aminoimidazole–bridged dinucleotides are not stable enough to directly cocrystallize with RNA, but the triphosphate analog is quite stable. As a result, we were able to obtain a high-resolution crystal structure of a close analog to the true reaction intermediate bound to an RNA primer–template complex.

The crystal structure of the RNA-GpppG complex shows that both G nucleotides interact with the template via Watson–Crick base pairing. We also showed that GpNH2ImpG could be modeled to fit the electron density of GpppG. Based on the structural similarities of GpppG and the imidazolium-bridged intermediate, it seems highly likely that the true intermediate also binds to the template through two Watson–Crick base pairs. The formation of two Watson–Crick base pairs leads to a much higher affinity of GpppG than GMP for a CC template (K_d of ~0.2 mM vs. ~20 mM, respectively). Additionally, the K_M of purified Gp-NH2Im-pG in primer extension (~0.6 mM) suggests comparable affinity of the imidazolium-bridged intermediate. The high affinity of the imidazolium-bridged intermediate helps to explain its effectiveness in primer extension

reactions because even if only a small fraction of monomer is converted to intermediate, the intermediate would still be able to bind to the primer–template complex in the presence of a large excess of activated monomer. Interestingly the GppG dinucleotide, which has a shorter linker between the two G residues, binds to the template via two Watson–Crick base pairs at one end of the RNA duplex, but at the other end by one Watson–Crick and one noncanonical base pair, suggesting that there is some strain involved in folding GppG into the conformation necessary to allow both Gs to Watson–Crick pair with the template.

Once the dinucleotide, GpppG, is bound to RNA, its overall structure, including the sugars and the triphosphate linker, becomes highly ordered. In the RNA-GpppG complex, the triphosphate linkage is structured due to a Mg^{2+} ion that coordinates with one nonbridging oxygen from each phosphate. As a result, the local structure is preorganized for S_N2 reaction with the primer 3′-hydroxyl, which in the case of the imidazolium-bridged intermediate would result in primer extension by one nucleotide. In the RNA-GpppG structure, the distance between the primer 3′-OH and the adjacent phosphate is 4.1 Å, and the 3′-O–P-O angle is 126°. The distance and angle seen here are comparable to the 4.1 Å and 107° for the S_N2 reaction catalyzed by the eukaryotic RNA polymerase II during transcription initiation, as seen in the corresponding crystal structure.[31] Assuming that the imidazolium-bridged intermediate adopts a similar conformation, it is clear that further adjustments in the distance and angle of attack would have to occur before any reaction could take place, suggesting a possible role for divalent metal ion catalysis. Nevertheless, comparison with the structures of dinucleotides with different linkages (either a shorter pyrophosphate linkage or a longer tetraphosphate linkage) revealed greater 3′-O to P

distances and/or greater disorder of the phosphate. Therefore, if the imidazolium-bridged intermediate adopts a conformation similar to that of GpppG under primer extension conditions, it will be partially preorganized for S_N2 attack by the primer hydroxyl. One caveat is that the intermediate generated from 2-methylimidazole–activated monomers could not bind Mg^{2+} in the same manner at the GpppG triphosphate, and the p-2MeIm-p bridge might be less ordered and therefore less than optimally reactive. However, 2-aminoimidazole–activated monomers, which lead to 10–100 times faster primer extension than 2-methylimidazole–activated monomer,[6] would generate a 2-aminoimidazolium–bridged intermediate. We propose that the 2-amino group of the imidazole could hydrogen bond to the nonbridging oxygens of both flanking phosphates, potentially resulting in a highly-ordered structure that is preorganized for nucleophilic attack by the primer hydroxyl. It is also of note that when the dinucleotide pGpG is bound to the same primer–template complex, the primer 3′-O to P distance is 4.9 Å, which is significantly greater than the distance for any of the 5′-5′–linked dinucleotides. This long distance may contribute to the slow rate of oligonucleotide ligation reactions, compared with primer extension with activated monomers.

Finally, we note that the 5′-5′–linked dinucleotides with pyrophosphate, triphosphate, and tetraphosphate linkages could all be synthesized prebiotically, given the availability of imidazole-activated nucleotides. Pyrophosphate-linked dinucleotides are readily formed by attack of the 5′-phosphate of an unactivated monomer on the activated phosphate of a second monomer.[30,32] Similarly, attack of the β-phosphate of GDP on the phosphate of an activated GMP would generate GpppG. The tight binding of GpppG to a CC template suggests that GpppG might be an ideal primer for the initiation

of template copying by primer extension. The resulting RNAs would begin with a 5′-caplike GpppG moiety, suggesting a potential evolutionary origin for the eukaryotic mRNA 5′-cap structure.

In summary, our structural studies of template-bound GpppG support the model that the structurally similar imidazolium-bridged intermediate binds the template tightly through two Watson–Crick base pairs and that the conformational constraint imposed by the covalent internucleotide bridge helps to preorganize the bound complex for in-line nucleophilic attack by the primer 3′-hydroxyl.

3.4 References

(1) Crick, F. H. *J. Mol. Biol.* **1968**, *38*, 367.

(2) Orgel, L. E. *J. Mol. Biol.* **1968**, *38*, 381.

(3) Szostak, J. W. J. Syst. Chem. **2012**, *3*, 2.

(4) Robertson, M. P.; Joyce, G. F. *Cold Spring Harb. Perspect. Biol.* **2012**, *4*, a003608.

(5) Inoue, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1981**, *103*, 7666.

(6) Li L, et al. *J. Am. Chem. Soc.* **2017**, *139*, 1810.

(7) Oró, J.; Basile, B.; Cortes, S.; Shen, C.; Yamrom, T. *Orig. Life* **1984**, *14*, 237.

(8) Hagenbuch, P.; Kervio, E.; Hochgesand, A.; Plutowski, U.; Richert, C. *Angew. Chem. Int. Ed. Engl.* **2005**, *44*, 6588.

(9) Jauker, M.; Griesser, H.; Richert, C. *Angew. Chem. Int. Ed. Engl.* **2015**, *54*,14559.

(10) Tam, C. P., et al. *J. Am. Chem. Soc.* **2017**, *139*, 571.

(11) Röthlingshöfer, M., et al. *Angew. Chem. Int. Ed. Engl.* **2008**, *47*, 6065.

(12) Kervio, E.; Claasen, B.; Steiner, U. E.; Richert, C. *Nucleic Acids Res.* **2014**, *42*, 7409.

(13) Kervio, E.; Sosson, M.; Richert, C. *Nucleic Acids Res.* **2016**, *44*, 5504.

(14) Izgu, E. C., et al. *J. Am. Chem. Soc.* **2015**, *137*, 6373.

(15) Blain, J. C., Szostak, J. W. *Annu. Rev. Biochem.* **2014**, *83*, 615.

(16) Prywes, N.; Blain, J. C.; Del Frate, F.; Szostak, J. W. *eLife* **2016**, *5*, e17756.

(17) Wu, T., Orgel, L. E. *J. Am. Chem. Soc.* **1992**, *114*, 5496.

(18) Walton, T.; Szostak, J. W. *J. Am. Chem. Soc.* 2016, 138, 11996.

(19) Sheng, J.; Larsen, A.; Heuberger, B. D.; Blain, J. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2014**, *136*, 13916.

(20) Sheng, J., et al. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 3050.

(21) Shen, F., et al. *Nucleic Acids Res.* **2017**, *45*, 3537.

(22) Biertümpfel, C., et al. *Nature* **2010**, *465*, 1044.

(23) Zhang, W.; Tam, C. P.; Wang, J.; Szostak, J. W. *ACS Cent. Sci.* **2016**, *2*, 916.

(24) Banerjee, A. K. *Microbiol. Rev.* **1980**, *44*, 175.

(25) Furuichi, Y.; Shatkin, A. J. *Adv. Virus Res.* **2000**, *55*, 135.

(26) Kaur, H.; Arora, A.; Wengel, J.; Maiti, S. *Biochemistry* **2006**, *45*, 7347.

(27) Fox, J. J., et al. *J. Am. Chem. Soc.* **1959**, *81*, 178.

(28.) Zhang, N.; Zhang, S.; Szostak, J. W. *J. Am. Chem. Soc.* **2012**, *134*, 3691.

(29) Murshudov, G. N., et al. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 355.

(30) Majerfeld, I.; Puthenvedu, D.; Yarus, M. *RNA* **2016**, *22*, 397.

(31) Cheung, A. C.; Sainsbury, S.; Cramer, P. *EMBO J.* **2011**, *30*, 4755.

(32) Kanavarioti, A.; Rosenbach, M. T.; Hurley, T. B. *Orig. Life Evol. Biosph.* **1992**, *21*, 199.

Chapter 4

Synthesis of a Nonhydrolyzable Nucleotide Phosphoroimidazolide Analogue that Catalyzes Nonenzymatic RNA Primer Extension

[*C. P. Tam and J. W. Szostak conceived of experiments, analyzed the data and wrote the text. L. Zhou performed primer extension experiments; C. P. Tam performed all other experiments. A. C. Fahrenbach, W. Zhang and T. Walton provided technical assistance. This chapter has previously been published with minor changes as the following publication:* Tam, C. P.; Zhou, L.; Fahrenbach, A. C.; Zhang, W.; Walton, T.; Szostak, J. W. Synthesis of a Nonhydrolyzable Nucleotide Phosphoroimidazolide Analogue that Catalyzes Nonenzymatic RNA Primer Extension. *J. Am. Chem. Soc.* **2018**, *140* (2), 783–792.]

Abstract. We report the synthesis of guanosine 5′-(4-methylimidazolyl)phosphonate (ICG), the third member of a series of nonhydrolyzable nucleoside 5′-phosphoro-2-methylimidazolide (2-MeImpN) analogues designed for mechanistic studies of nonenzymatic RNA primer extension. The addition of a 2-MeImpN monomer to a primer is catalyzed by the presence of a downstream activated monomer, yet the three nonhydrolyzable analogues do not show catalytic effects under standard mildly basic primer extension conditions. Surprisingly, ICG, which has a pK_a similar to that of 2-MeImpG, is a modest catalyst of nonenzymatic primer extension at acidic pH. Here we show that ICG reacts with 2-MeImpC to form a stable 5′–5′-imidazole-bridged guanosine-

cytosine dinucleotide, with both a labile nitrogen–phosphorus and a stable carbon–phosphorus linkage flanking the central imidazole bridge. Cognate RNA primer–template complexes react with this GC-dinucleotide by attack of the primer 3′-hydroxyl on the activated N–P side of the 5′-5′-imidazole bridge. These observations support the hypothesis that 5′–5′-imidazole-bridged dinucleotides can bind to cognate RNA primer–template duplexes and adopt appropriate conformations for subsequent phosphodiester bond formation, consistent with our recent mechanistic proposal that the formation of activated 5′–5′-imidazolium-bridged dinucleotides is responsible for 2-MeImpN-driven primer extension.

4.1 Introduction

For the earliest cellular organisms,[1,2] replication must have occurred through primitive mechanisms without elaborate enzymatic machinery.[3,4] Template-directed nonenzymatic primer extension[5] is an important experimental model system for the study of chemically driven genome replication. In this model, chemically activated mononucleotides or short oligonucleotides[6] bind to a template,[7] followed by monomer polymerization[8] or oligomer ligation[9] to assemble the template complement. Imidazoles, among other heteroarenes,[10,11] have been broadly utilized in the chemical activation of mono-[12,13] and oligonucleotides.[6,14] Much effort has been invested in enabling mixed sequences, especially those containing adenosine and uridine,[15] to be copied efficiently and with high fidelity. Among these efforts, a small, focused screen of substituted imidazoles as monomer leaving groups revealed 2-aminoimidazole-activated monoribonucleotides as superior primer extension substrates compared to imidazole- and 2-methylimidazole-activated monomers.[14] On the other hand, activated trinucleotides that bind downstream of the extending nucleotide were discovered to be much better catalysts than activated monomers, not only providing rate enhancements of at least 2 orders of magnitude, but also enabling copying of templates containing all four canonical nucleotides in one pot.[6] Additional work has been performed to determine and enhance the fidelity,[16,17] regioselectivity,[18-20] and thermodynamics[21-23] of these reactions. Plausible prebiotic conditions conducive to imidazole synthesis have also been probed.[24, 25]

For decades, the classical S_N2-type mechanism was thought to constitute the sole mechanistic pathway for primer extension by nucleotide phosphoroimidazolides:[3] the nucleophilic 3′-hydroxyl group of the primer would attack the imidazole-activated

phosphate of the incoming monomer via an in-line mechanism, and concomitantly displace the imidazole leaving group. Activated monomers[26] and oligomers[6, 14] binding downstream of the polymerization site play important catalytic roles, as their presence strongly accelerates the rate of primer extension. This rate acceleration was initially hypothesized to result from noncovalent leaving group–leaving group interactions between consecutively bound activated substrates, which preorganize the upstream monomer and lower the activation barrier for subsequent phosphodiester bond formation. However, emerging evidence suggests that two activated monomers can react with each other to form a highly activated imidazolium-bridged dinucleotide intermediate (Np-MeIm-pN, **Figure 4.1**), and the observed rate acceleration afforded by a downstream activated monomer (or oligomer) may be due to covalent nucleophilic catalysis.[27] Hence, we have sought to ascertain which of these two potential pathways is primarily responsible for catalysis by downstream activated nucleotides.

Nonhydrolyzable nucleotide analogues have been indispensable tools in crystallographic and mechanistic studies of polymerases and ligases.[28] For example, nonhydrolyzable nucleotide triphosphates, such as α,β-methylene[29] and α,β-difluoromethylene[30] nucleoside triphosphates have been cocrystallized with T7 RNA polymerase[29] and DNA polymerase β,[30] respectively. Also, Raines and co-workers used α,β-methylene GTP to show that the scission of the GTP α–β-phosphoanhydride bond is critical for RNA ligation by the archaeal RNA ligase RtcB.[31] Drawing inspiration from these strategies, we have developed several nonhydrolyzable 2-MeImpN analogues as mechanistic probes of nonenzymatic RNA template copying. We previously reported the synthesis of guanosine 5′-(3-methylpyrazolyl) phosphonate (PZG, **Figure 4.1**)[32] and

guanosine 5′-(2-methylpyrrolyl)phosphonate (PYG, **Figure 4.1**)[14] as nonhydrolyzable analogues of 2-MeImpG. Our crystal structures of RNA-PZG complexes revealed[32] that the noncovalent interactions between monomers and RNA templates do not always obey canonical Watson–Crick geometry under crystallization conditions. More importantly, we were not able to observe any noncovalent interactions between the leaving groups of consecutively bound PZG nucleotides. We suspected that the low pK_a of the pyrazolyl group of PZG might have led to chemical differences from 2-MeImpG that prevented the formation of catalytically relevant interactions in the crystal structure. In a parallel approach, we used P^1,P^3-diguanosine-5′-triphosphate (GpppG) as a stable analogue of the corresponding activated Gp-MeIm-pG dinucleotide for crystallographic studies, and showed that it could have the appropriate geometry to bind to cognate RNA primer–template complexes via Watson–Crick base pairing, with its 5′–5′-bridge preorganized for subsequent S_N2 attack by the primer 3′-hydroxyl group.[33]

Here we report the synthesis of guanosine 5′-(4-methylimidazolyl)phosphonate (ICG, **Figure 4.1**), the third and closest nonhydrolyzable isosteric analogue of 2-MeImpG. For all three analogues (ICG, PYG, and PZG), we report the pK_a of the leaving group mimics in the context of a full nucleotide, along with the binding constants of 2-MeImpG and all three pseudo-activated nucleotides for RNA primer–template complexes. Finally we report measurements of the pH-dependent catalytic activity of all four nucleotides in nonenzymatic primer extension. None of the three analogues shows catalytic effects under mild alkaline primer extension conditions, yet ICG, which has a pK_a and RNA affinity similar to those of 2-MeImpG, is a modest catalyst under acidic pH. Using NMR and mass spectrometry, we demonstrate that ICG reacts with an activated C monomer to form the

5′–5′-imidazole-bridged guanosine-cytosine dinucleotide (CpICG) with very high regioselectivity. The purified GC-dinucleotide, with a labile nitrogen–phosphorus and a stable carbon–phosphorus linkage flanking the imidazole bridge, reacts with cognate RNA primer–template complexes to afford extended primers only when the labile nitrogen–phosphorus linkage, but not the stable carbon–phosphorus linkage, is proximal to the primer 3′-hydroxyl group. These results are consistent with our recent mechanistic proposal that 2-MeImpN-driven primer extension proceeds via the formation of 5′–5′-imidazolium-bridged dinucleotide intermediates.

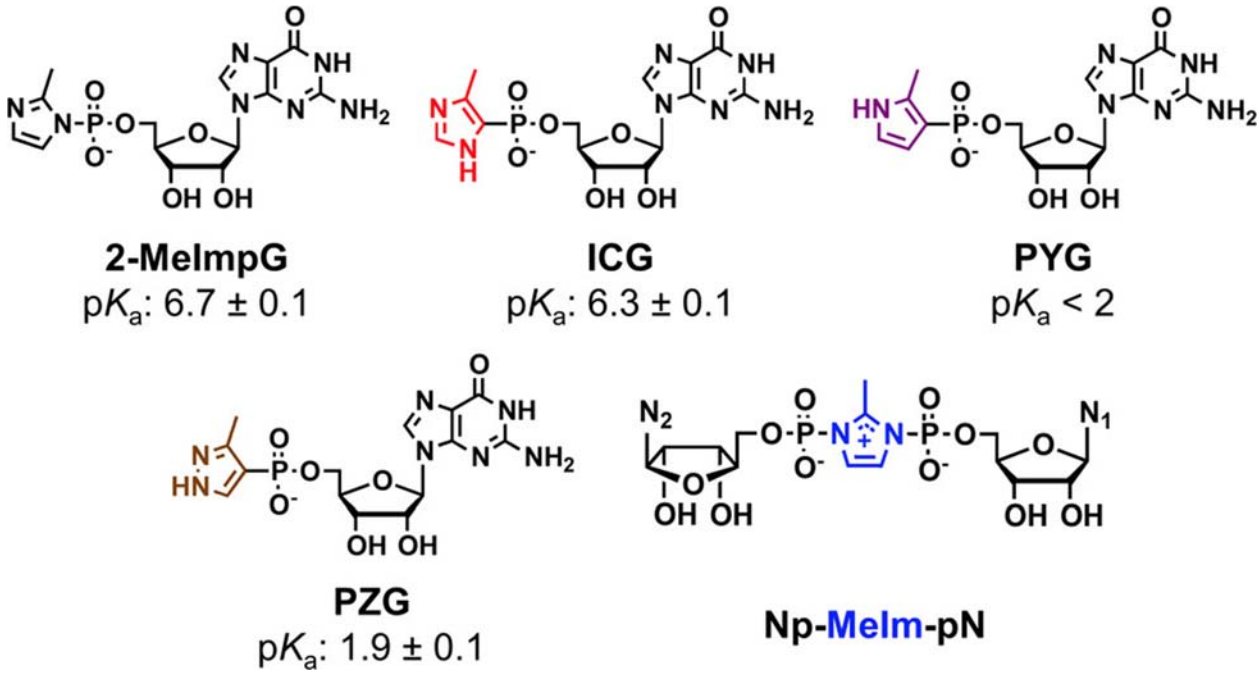
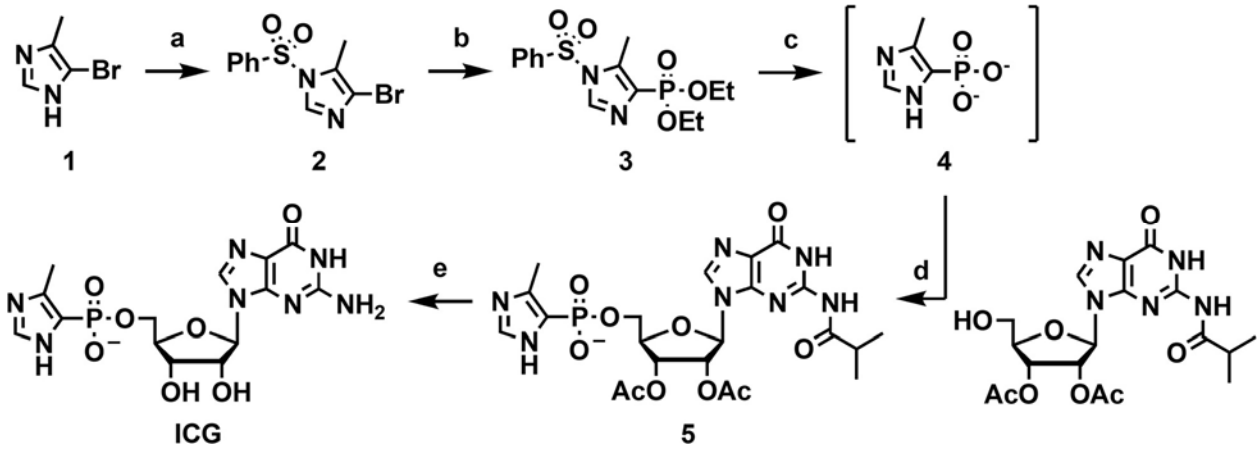


Figure 4.1. Chemical structures of 2-MeImpG, ICG, PYG, PZG, and Np-MeIm-pN, together with the pK_a values of the heteroaryl leaving groups in water, as estimated using NMR spectroscopy (mean and range of duplicate measurements). The pK_a of the pyrrolyl group of PYG (purple) could not be confidently determined due to extensive proton–deuterium exchange at strongly acidic pHs (**Figure D.3**, **Appendix D**).

4.2 Results

4.2.1 Synthesis and pK_a Determination of Guanosine 5′-(4-Methylimidazolyl)phosphonate (ICG). Our synthetic strategy for the construction of ICG was parallel to our previously described synthesis of PZG.[32] Briefly, the commercially available 4-bromo-5-methylimidazole **1** was protected on N3 with a benzenesulfonyl group to give **2**,[34] followed by installation of the phosphonate moiety via a palladium-catalyzed carbon–phosphorus coupling[35] reaction to afford **3**. The protected C–P-linked imidazolylphosphonate was then coupled to the 5′-hydroxyl of guanosine to afford **5**, by first removing the phosphonate alkyl groups (**4**),[36] followed by Mitsunobu phosphonylation.[37] Subsequent deprotection of **5** then furnished the desired ICG product (**Scheme 4.1**).



Scheme 4.1. Synthesis of ICG. Reaction conditions: (a) $PhSO_2Cl$, Et_3N, DCM, 48%; (b) $(EtO)_2P(O)H$, $Pd(PPh_3)_4$, PPh_3, Et_3N, DMSO, 115 °C, 2 h, 35%; (c) TMSBr, Et_3N, DCM, 4 h; then MeOH; (d) N^2-isobutyryl-2′,3′-diacetylguanosine, DIAD, PPh_3, DCM, 3 h; (e) conc. NH_3 in H_2O, MeOH, 65 °C, 4 h (22% over c–e). See **Figures D.34–D.36** (**Appendix D**) for the ^1H, ^{13}C, and ^{31}P NMR spectra of ICG.

The pK_a of the leaving group of monomers in the downstream (i.e., primer+2) position has been shown to strongly influence the rate of reaction of the primer with an activated monomer in the adjacent (i.e., primer+1) position,[14] suggesting that the protonation state of the downstream monomer plays an important role in catalysis. Hence, we sought to determine the protonation pK_a of the leaving group of 2-MeImpG and the heteroaryl-phosphonate moieties of nonhydrolyzable 2-MeImpG analogues in the context of the full nucleotide (**Figure 4.2.A**), and to correlate the experimentally determined pK_a values with catalysis under primer extension conditions. We used proton NMR spectroscopy for pK_a determination[38,39] by monitoring the chemical shifts of the aromatic protons of 2-MeIm and its mimics as a function of pD (**Figure 4.2.B**). The resulting sigmoidal curves (**Figure 4.2.C**) were fit to an equation derived from the Henderson–Hasselbach formula (**Section D.3**, **Appendix D**), followed by correction as reported by Krezel and Bal[40] (**Section D.3.1**, **Appendix D**) to estimate the pK_a values of the four monomers (**Figure 4.1**). Of the nonhydrolyzable analogues, the protonation pK_a of the imidazolyl moiety of ICG (pK_a = 6.3) is most similar to that of the 2-MeIm leaving group of 2-MeImpG (pK_a = 6.7, literature value[41] = 7.09; **Figure D.1**, **Appendix D**). The pK_a values of the heteroaryl groups of PZG and PYG are at least 4 units lower than that of ICG (**Figures D.2** and **D.3**, **Appendix D**). It is interesting to note that the pK_a values of ICG and 2-MeImpG are at least 1 unit lower than the reported pK_a (7.86) of free 2-methylimidazole[42] despite the expected stabilization of the protonated state by the adjacent negative charge of the phosphate or phosphonate. Apparently, the inductive electron withdrawing effect of phosphate or phosphonate groups has a greater influence and serves to lower the pK_a of the leaving groups relative to free 2-methylimidazole.

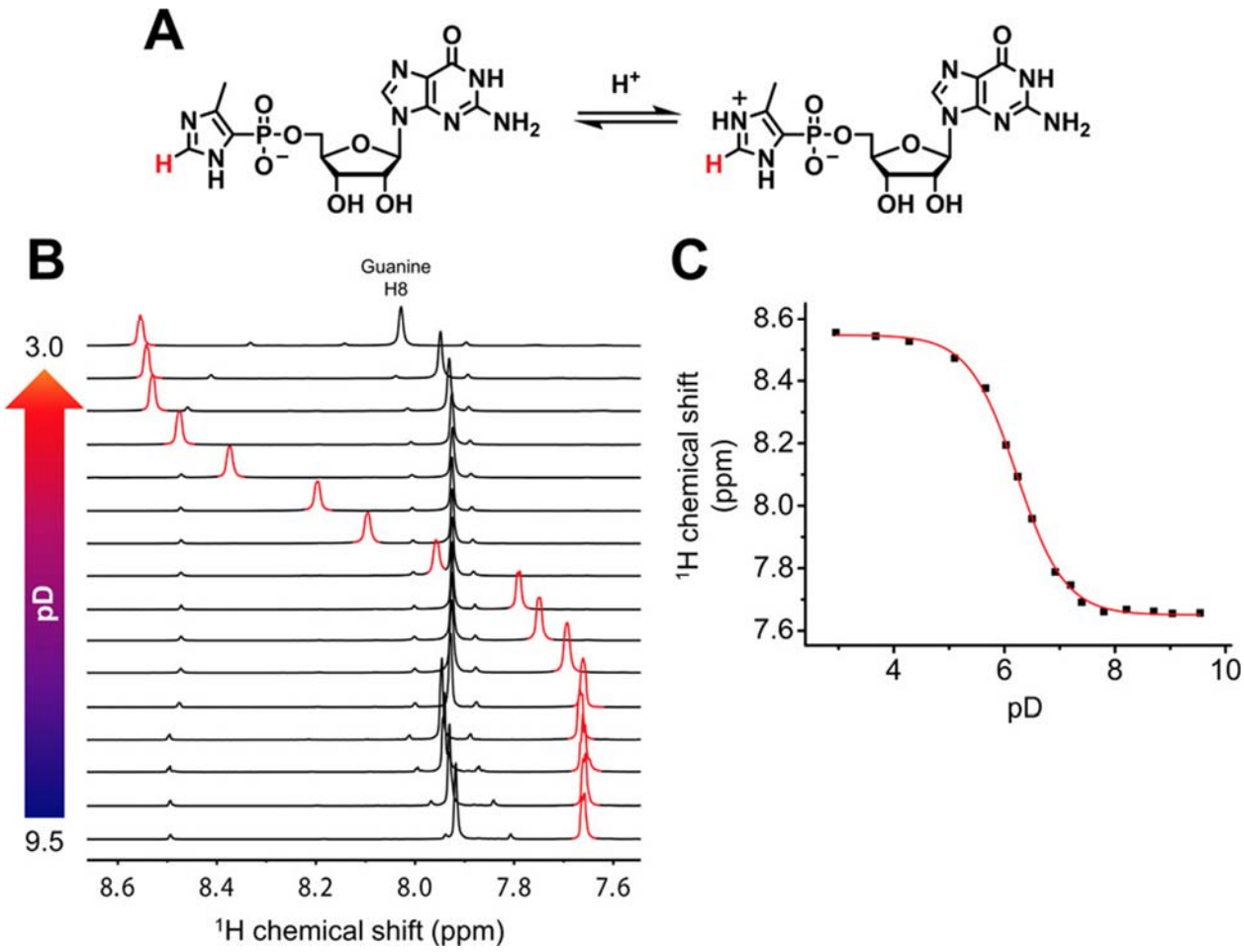
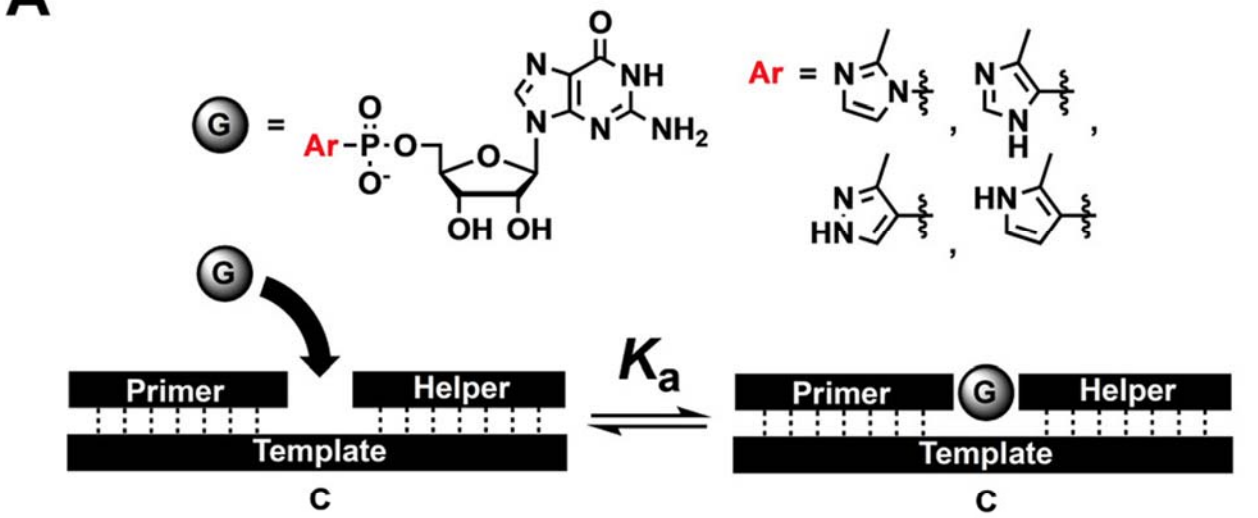


Figure 4.2. (A) Protonation of the imidazolylphosphonate of ICG. The chemical shift of the ICG imidazole methine proton (colored red) is monitored throughout the pD titration experiment. (B) The stacked variable pD-proton NMR spectra of ICG. The aromatic region of the proton spectra (7.5–8.7 ppm) is shown. The peaks in the stacked proton spectra corresponding to the ICG imidazole methine proton are colored red. All spectra are referenced to the residual HOD peak at 4.79 ppm (not shown). As the monomer solution was acidified, the methine proton shifted downfield dramatically, crossing over the proton peak of the guanine C8 proton. Experiments were performed in duplicate. (C) The corresponding chemical shift vs monomer pD plot. The inflection point of the sigmoidal fit curve reflects the pK_a value of the leaving group moieties of monomer or analogue.

4.2.2 Affinity of 2-MeImpG and Nonhydrolyzable Analogues to RNA Duplex. We assessed the affinity of 2-MeImpG and its nonhydrolyzable analogues for cognate binding sites on an RNA primer–template complex by isothermal titration calorimetry.[43,44] Specifically, 4.4 equiv of 2-MeImpG or analogue was titrated into 1.5 mM of a previously

reported interrupted RNA duplex in which the binding site for guanosine mononucleotides is sandwiched between the primer 3′-terminus and a downstream helper oligonucleotide, in the presence of pH 7-buffered NaCl solution (500 mM Na^+) at 20 °C (**Figure 4.3.A**). The helper oligonucleotide primarily serves to enhance the affinity of monomer or analogue to the RNA primer–template complex, thus enabling affinity be measured by ITC.[21] We now show that the affinities of the nonhydrolyzable 2-MeImpG analogues for the sandwich duplex are comparable, with estimated K_a values between 2.8×10^3 and 4.5×10^3 M^{-1}; this is slightly higher than the K_a of 2-MeImpG ($(2.1 \pm 0.4) \times 10^3$ M^{-1}, **Figure 4.3.B**). The estimated stoichiometric constant *n* of RNA:analogue binding for all three monomer analogues was ~0.8, consistent with 1:1 analogue-to-RNA binding. The free energy of RNA binding for the analogues ranged between −4.6 and −4.9 kcal mol^{-1}, suggesting that the chemical characteristics of the heteroaryl groups of the analogues have only minimal effects on the free energy of binding with RNA. Notably, the presence of 50 mM Mg^{2+} did not alter the affinity of ICG for RNA. We suspect that the high ionic strength of both solutions negates any effect Mg^{2+} may have on the binding of the analogue to RNA duplexes.

A



B

	K_a (M^{-1})	ΔG (kcal mol^{-1})	ΔH (kcal mol^{-1})	ΔS (cal mol^{-1} K^{-1})
2-MeImpG	$(2.1\pm0.4)\times10^3$	–4.45±0.12	–17.1±3.0	–43.1±10.7
ICG	$(4.1\pm0.5)\times10^3$	–4.84±0.07	–14.1±0.1	–31.5±0.5
ICG with Mg^{2+}	$(4.5\pm1.0)\times10^3$	–4.89±0.13	–15.2±0.5	–35.1±1.9
PZG	$(2.8\pm0.3)\times10^3$	–4.63±0.06	–14.2±0.6	–32.7±1.8
PYG	$(3.6\pm0.2)\times10^3$	–4.78±0.03	–14.8±0.3	–34.2±0.9

Figure 4.3. Affinity of 2-MeImpG and nonhydrolyzable analogues with the primer–template–helper (P/TS/H) sandwich duplex. (A) The binding scheme between 2-MeImpG and its analogues with the P/TS/H duplex. The analogue is sandwiched between two flanking G residues at the primer 3′-terminus and the helper 5′-terminus, respectively. (B) Associated thermodynamic parameters, ± standard error, *n* = 3. See **Figure D.4** for the LC-HRMS analysis of the sandwich oligonucleotides; see **Figure D.5** and **Table D.1** (**Appendix D**) for data derived from individual ITC titrations. All ITC experiments were performed at 20 °C.

4.2.3 Effects of 2-MeImpG and Nonhydrolyzable Analogues on RNA Primer Extension. With these insights into the chemical and thermodynamic properties of the 2-MeImpG analogues, we turned to the question of catalysis of the reaction between the primer and an adjacent monomer, by analogues bound to the template at the primer +2 position. We measured the pseudo-first-order rate of conversion of primer to extended products on a template with a GCCCAA-5′ overhang (**Figure 4.4.A**; **Figure D.6**, **Appendix D**), using a reaction mixture that contained a large excess of the 2-MeImpC (30 mM). 2-MeImpG and its nonhydrolyzable analogues were used at 30 mM, because higher concentrations led to observable precipitation, especially for ICG and GMP; the catalytic metal ion Mg^{2+} was present at 100 mM. We varied the pH of the primer extension reactions from mildly acidic (pH 5.5) to slightly alkaline (pH 8) to probe the effect of pH on catalysis by the analogues. Interestingly, the rate of primer extension with 2-MeImpC in the absence of any downstream guanosine nucleotides (blue bars, **Figure 4.4.B**) was slightly higher than in the presence of unactivated GMP (black bars, **Figure 4.4.B**). We attribute this small effect to the potential for weak binding of 2-MeImpC at the primer +2 position in the absence of a competing G nucleotide. To avoid this confounding effect, we compare the catalytic effects of 2-MeImpG and the three analogues to the baseline rate observed in the presence of GMP. As expected 2-MeImpG in the primer +2 position exhibits a strong catalytic effect at all pH values from 5.5 to 8.0 (red bars, **Figure 4.4.B**). In general, all primer extension reactions exhibit higher rates at alkaline pH than in neutral or acidic pHs, presumably due to the need to deprotonate the primer 3′-hydroxyl. We then asked whether binding of ICG, PYG or PZG at the primer +2 position could enhance the rate of primer extension with 2-MeImpC (green, purple, and orange bars, respectively;

Figure 4.4.B). Regardless of pH, the rate of primer extension was not improved in the presence of PYG or PZG, compared with GMP. Therefore, neither PZG nor PYG affords any catalysis to primer extension.

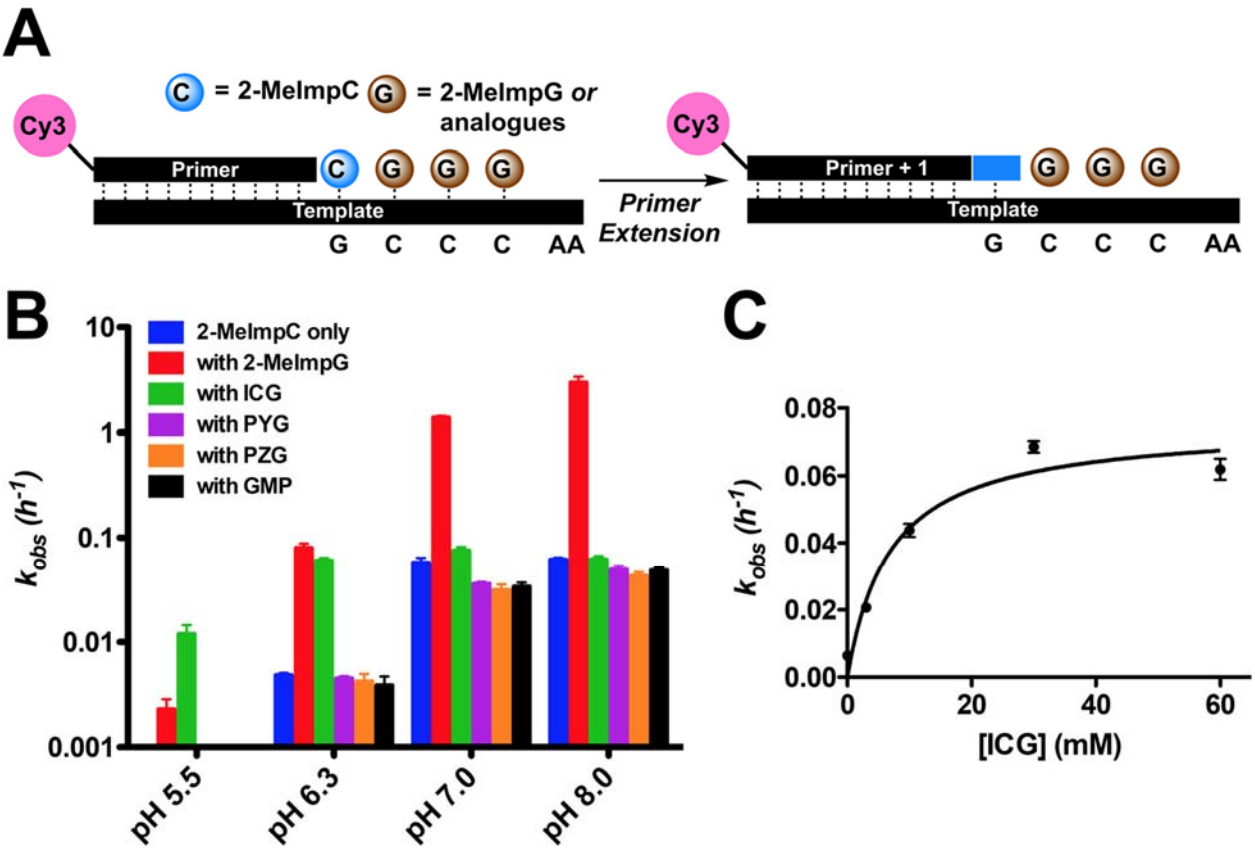


Figure 4.4. (A) Schematic of the primer extension assays used herein. 2-MeImpC binds to the cognate primer +1 binding site, followed by the consecutive binding of multiple 2-MeImpG or its nonhydrolyzable analogues. (B) k_{obs} of the primer extension shown in (A) under different reaction pH and different stacking guanosine monomers. (C) Saturation curve of the ICG-assisted primer extensions at pH 6.3. Monomer precipitation was not observed, even at the highest concentration of ICG (60 mM). In (B) and (C), all primer extensions were performed in triplicate. See **Figures D.7–D.14** (**Appendix D**) for representative denaturing PAGE gel images and first-order rate plots for the primer extensions depicted in (B), as well as the saturation curves of 2-MeImpG- and ICG-catalyzed primer extensions at pH 6.3 and 8.0.

On the other hand, ICG, which has an imidazolyl pK_a close to the pK_a of 2-MeImpG, did exhibit a pH-dependent catalytic effect on primer extension (**Figure 4.4.B**). At a primer

extension pH of 8, the presence of ICG led to only a very small increase in the rate of primer extension of 20% (compared to the 60-fold effect of 2-MeImpG). However, when the extension reactions were carried out at neutral pH, the catalytic effect of ICG increased to ~2.5-fold (vs 40-fold for 2-MeImpG), while at pH 6.3, ICG exhibited a 15-fold rate enhancement relative to GMP, rivaling the ~20-fold effect of 2-MeImpG. The k_{obs} of ICG-assisted primer extension is similar from pH 8.0 to 6.3, while the k_{obs} of 2-MeImpG-assisted primer extension decreased ~37 fold from pH 8.0 to pH 6.3. At pH 5.5, primer extension in the presence of GMP could barely be detected even after 20 h. Even 2-MeImpG in the primer +2 position yielded less than 5% extended primer after 20 h. However, the catalytic effect of ICG surpassed that of 2-MeImpG and resulted in readily observable primer extension, leading to 20% of the primer +1 product after 20 h (**Figure D.7**, **Appendix D**). Thus, the rate enhancement afforded by ICG is much more pronounced at acidic pH, while the rate enhancement afforded by 2-MeImpG is most pronounced at alkaline pH.

To see if the increasing catalytic effect of ICG at lower pH (and conversely, the declining catalytic effect of 2-MeImpG at lower pH) was due to changes in template binding, we measured the catalytic effect as a function of concentration at pH 6.3 and 8.0. Primer extensions at pH 6.3 were repeated at four different concentrations of ICG (0, 3, 10, 30, 60 mM). The estimated k_{max} was 0.075 ± 0.008 h^{-1}, and the K_M was 7 ± 3 mM. From the rate vs concentration plot (**Figure 4.4.C**), it is apparent that the rate-enhancement afforded by ICG approaches a maximum when the concentration of ICG is at, or above 20 mM. Since both ICG and 2-MeImpG achieve their maximal catalytic effect at 30 mM at pH 6.3 (**Figure D.13**, **Appendix D**), their different catalytic abilities cannot

be attributed to differential binding. In contrast, at pH 8.0, the rates of primer extension are unaffected by ICG concentration (0 mM ICG; rate ≈ 0.07 h^{-1}; **Figure D.12**, **Appendix D**). Since at pH 8 ICG and 2-MeImpG bind with similar affinities to primer/template complexes, the pH-dependence of catalysis by ICG does not reflect pH-dependent changes in binding, but likely reflects a direct pH-dependence of the catalytic mechanism.

How can we explain the observation that ICG becomes a better catalyst of primer extension at progressively lower pH, while 2-MeImpG gradually loses catalytic ability at low pH? Given the observation of a 5′–5′-imidazolium-bridged dinucleotide in primer extension reactions, we wondered if ICG might catalyze primer extension in a similar fashion, with the imidazolyl group of ICG (highlighted in blue, **Figure 4.5**) reacting with 2-MeImpG to form a mixed N–P- and C–P-linked imidazole-bridged intermediate (**Figure 4.5**, middle panel).[27] Due to the lower pK_a of ICG than 2-MeImpG, the ratio of protonated 2-MeImpG to unprotonated ICG is increased at lower pH, while a higher proportion of ICG remains in the neutral, nucleophilic form. Therefore, we sought to examine whether the proposed 5′–5′-imidazole-bridged dinucleotide was formed during the ICG-catalyzed primer extension.

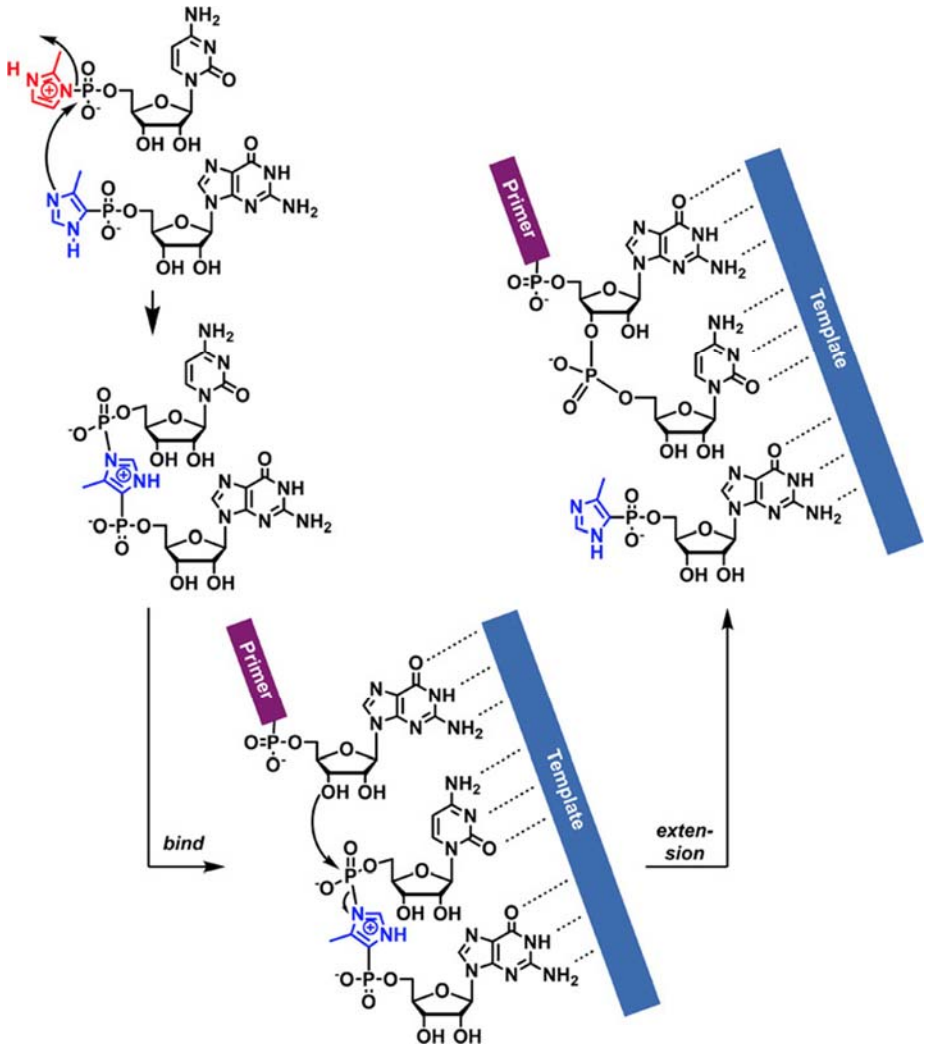


Figure 4.5. Proposed mechanism of the ICG-assisted primer extension via the formation of the 5′–5′-imidazole-bridged dinucleotide (CpICG).

4.2.4 Detection of 5′–5′-Imidazole-Bridged Mixed-Linkage Dinucleotide via LCMS. Initial attempts to identify the 5′–5′-imidazole-bridged mixed-linkage dinucleotide were performed with analytical reversed-phase liquid chromatography-coupled mass spectrometry (LCMS), using ICG and dideoxycytidine 5′-phosphoro-2-methylimidazolide (2MeImpddC) as the substrate. 2MeImpddC was chosen in lieu of 2-MeImpC because its dideoxyribose moiety affords enhanced hydrophobicity, allowing 2MeImpddC, as well as its ICG-conjugated products to be chromatographically resolved from ICG. Additionally, removal of the cytidine ribose 2′- and 3′-hydroxyl groups eliminates the possibility of 2′-5′-

and 3′-5′-linked CC dinucleotide formation, thus facilitating chromatographic separation of possible products. 2MeImpddC was conveniently prepared from Zalcitabine with a three-step protection–activation–deprotection sequence (**Figure 4.6.A**). First, 30 mM ICG and 30 mM 2MeImpddC were incubated at pH 6.3 in the absence of Mg^{2+}. After 2 h, the incubated sample was chromatographically analyzed under reverse phase conditions (300 pmol each for ICG and 2MeImpddC). ICG and 2MeImpddC were well resolved, and only one new peak appeared, with a retention time of 19.11 min (**Figure 4.6.B**) and an *m/z* of 699.33 (**Figure 4.6.C**), which corresponds to the expected *m/z* of monoanionic ddCMP-ICG dinucleotide. To confirm that the ddCMP-ICG dinucleotide carries the desired 5′–5′-imidazole bridge, we performed tandem MS fragmentation on the precursor ion with the *m/z* = 699.33 (with a *m/z* window of ±1). The fragmentation voltage was chosen such that the precursor ion was only partially fragmented so that the major fragment *m/z* peak could indicate which bond within the precursor ion is the most labile (**Figure 4.6.D**). Notably, the major fragment peak, with an *m/z* of 416.17 (peak 1, **Figure 4.6.D**), is suggestive of fragmentation of the 5′–5′-imidazole-bridged ddCMP-ICG dinucleotide, with heterolytic cleavage occurring at the O–P bond between the guanosine and the imidazole bridge moieties (**Figure 4.6.E**). Other possible regioisomers of ddCMP-ICG dinucleotide, such as the 3′-5′-, 2′–5′-, and nucleobase-linked dinucleotide, could not afford fragments with *m/z* of 416.17. This observation thus provides the first clue that ICG could react with activated monomers to afford 5′–5′-imidazole-bridged dinucleotides.

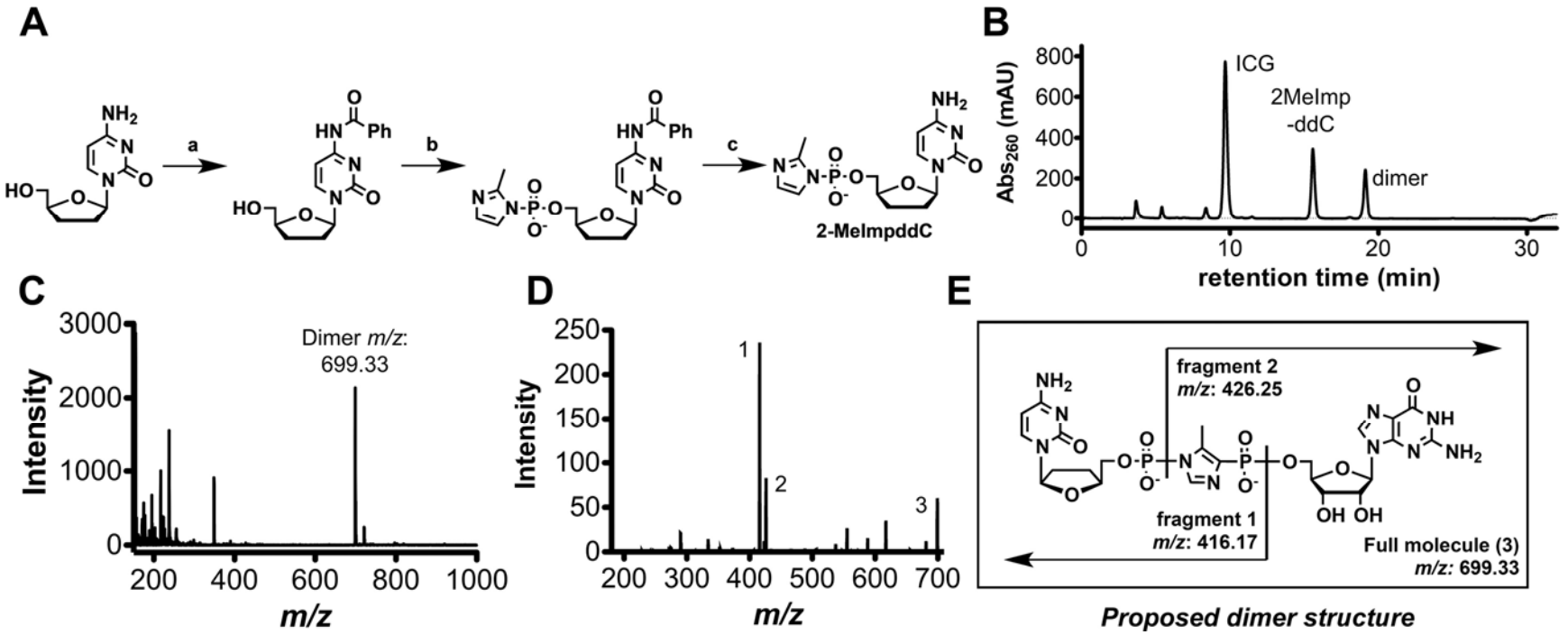


Figure 4.6. (A) Synthesis of 2-MeImpddC. Conditions: (a) TMSCl, pyr; then BzCl; then conc. NH_4OH; (b) $POCl_3$, DIPEA, $PO(OMe)_3$; then 2-MeIm; then 1 M TEAB; (c) conc NH_4OH, 65 °C. (B) Liquid chromatography analysis of the products of 2MeImpddC-ICG coincubation at pH 6.3 after 2 h. UV absorbance of the products at 260 nm was monitored. (C) ESI-MS analysis of the LC peak at ~19 min shown in (B). The major *m/z* peak corresponds to the monoisotopic monoanionic mass of the proposed ddCMP-ICG dinucleotide; the structure of the proposed dinucleotide is shown in (E). (D) The MS^2 spectrum obtained by fragmentation of the *m/z* = 699.33 precursor ion. Peak 3 is the partially fragmented parent ion; peak 2, with a *m/z* of 426.25, corresponds to the ICG fragment of CpICG, resultant from P–N bond cleavage during collision-induced dissociation. See **Figure D.17** of **Appendix D** for a more detailed assignment of the MS^2 spectrum. (E) The proposed structure of the 5′–5′-imidazole-linked ddCMP-ICG dinucleotide, the observed *m/z* signals, and the corresponding CpICG fragments. 2-MeImpddC had been coincubated with ICG, PYG, and PZG, and the formation of 5′–5′-heteroarene-bridged dinucleotide had been monitored by LCMS and NMR. See **Figures D.15–D.20** (**Appendix D**) for detailed analyses.

4.2.5 Structure Elucidation of CpICG by NMR. Encouraged by the LC-MS results, we then asked whether 2-MeImpC and ICG could react to form the 5′–5′-imidazole-bridged, mixed linkage CG dinucleotide (CpICG, **Figure 4.5**, middle panel). 50 mM ICG was incubated at pH 6.3 with 50 mM 2-MeImpC in the absence of Mg^{2+}, and the progress of dinucleotide formation was monitored by ^{31}P NMR spectroscopy overnight (**Figure 4.7.A**). As seen in the *t* = 0 h spectrum, acquired immediately after mixing, two major ^{31}P singlet peaks were seen at −9.89 and 2.30 ppm, corresponding to 2-MeImpC and ICG, respectively. Three additional minor singlet peaks were observed at −8.95, 2.25, and 7.54 ppm, and no other significant ^{31}P peaks are observed. These peaks increased in intensity as the incubation proceeded, with the peaks at −8.95 and 7.54 ppm maintaining similar intensity throughout the entire NMR time course, suggesting that they might arise from the same molecule. The peak at 2.25 ppm is assigned to CMP by internal referencing, which results from the hydrolysis of 2-MeImpC (**Figure D.21**, **Appendix D**). The remaining two ^{31}P peaks at −8.95 and 7.54 ppm were tentatively assigned to the two phosphorus atoms of the newly formed GC dinucleotide. To establish the regiochemistry of the newly formed GC dinucleotide, and rule out the possibility of 2′-5′-, 3′-5′-, or guanine-phosphate linkages, we examined the reaction of PYG with 2-MeImpC. The pyrrolylphosphonate group of PYG is known to be non-nucleophilic under near-neutral conditions. Throughout an overnight ^{31}P NMR time course, the only new detectable product was CMP (**Figure 4.7.B**), suggesting that the nucleophilic ribose hydroxyls and nucleobase exocyclic amine of PYG do not react with 2-MeImpC over the time course of primer extension assays. Since the only reactive site on ICG is the imidazolylphosphonate group, the CpICG dinucleotide should indeed contain a 5′–5′-imidazole bridge.

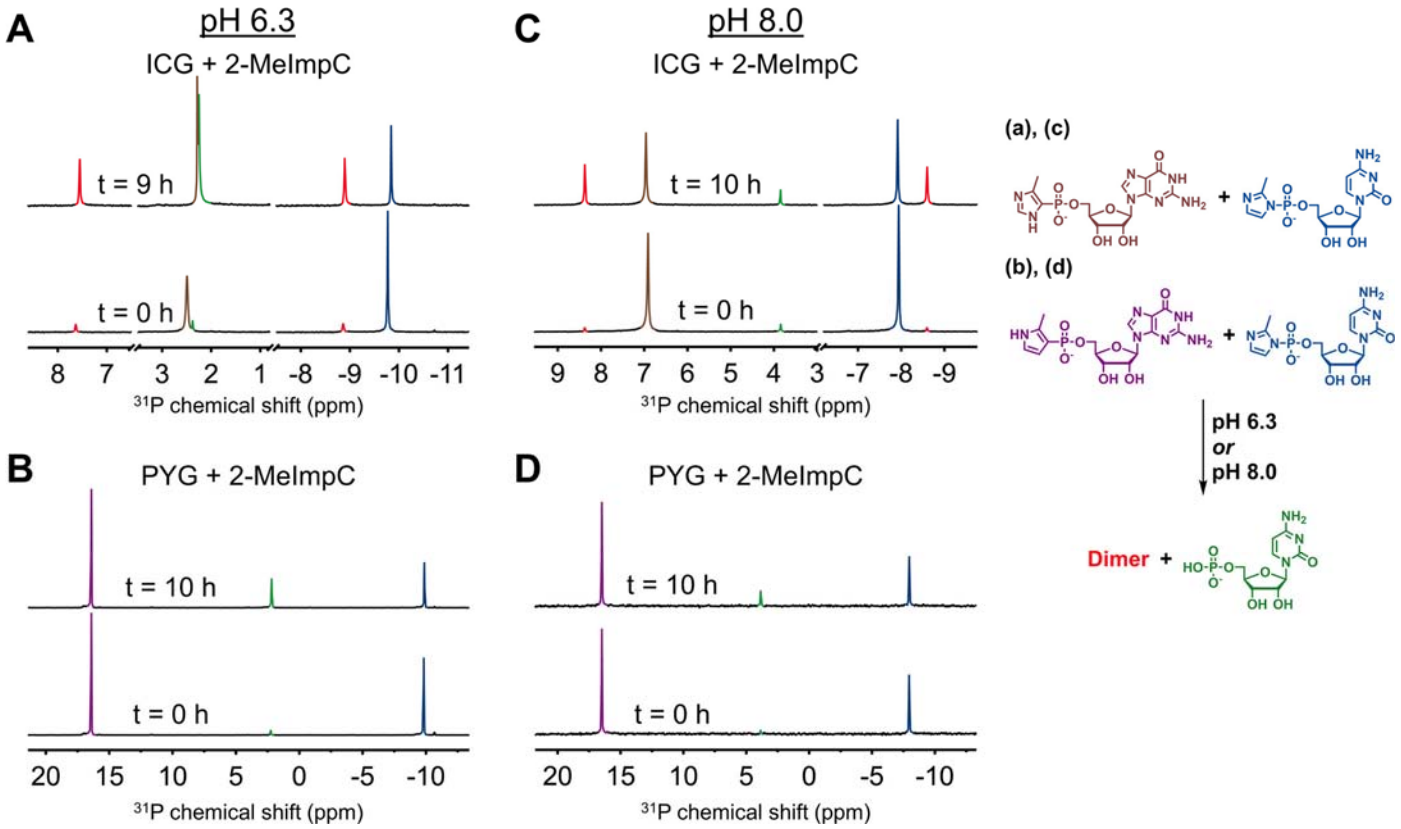


Figure 4.7. (A) ICG coincubated with 2-MeImpC at pH 6.3. (B) PYG coincubated with 2-MeImpC at pH 6.3. (C) ICG coincubated with 2-MeImpC at pH 8.0. (D) PYG coincubated with 2-MeImpC at pH 8.0. The phosphorus resonances of ICG, 2-MeImpC, CMP, PYG, and GC-dinucleotide were colored as maroon, naval blue, green, purple, and red, respectively. The lack of reactivity between PYG and 2-MeImpC strongly suggests that the ribose hydroxyls and guanine exocyclic amine of ICG do not react significantly with 2-MeImpC. See **Figures D.21–D.24** (**Appendix D**) for the complete ^{31}P stacked spectrum of the NMR time course analyses depicted in panels (A)–(D), respectively. See **Figure D.25** (**Appendix D**) for the complete ^{31}P NMR time course analyses of 2-MeImpC-PZG coincubation. See **Figures D.37–D.40** (**Appendix D**) for detailed LCMS time course analyses of the formation and hydrolysis of CpICG at pH 6.3 and 8.0. See **Figure D.41** (**Appendix D**) for the fitted rate of CpICG formation and degradation derived from the LCMS time course analyses.

Because ICG catalyzed primer extension at pH 6.3 but not at pH 8.0, we asked whether the 5′–5′-imidazole-bridged CpICG dinucleotide could form under basic conditions. Using ^{31}P NMR spectroscopy, we monitored the overnight reaction of coincubated ICG and 2-MeImpC at pH 8.0 in the absence of Mg^{2+}. As with the pH 6.3 NMR time course, three new singlet peaks increased in intensity over time. The three new peaks were shifted downfield to varying degrees, presumably due to the change in pH (**Figure 4.7.C**). The peak at 3.85 ppm was assigned to CMP (**Figure D.22**, **Appendix D**) suggesting that the remaining two peaks belong to CpICG. This was confirmed with the same PYG control experiment as used above (**Figure 4.7.D**). To confirm that the GC dinucleotide formed at pH 6.3 and 8.0 is indeed the same chemical entity, the dinucleotide prepared from overnight 2-MeImpC-ICG coincubation at pH 8.0 was chromatographically purified under reverse-phase conditions, followed by lyophilization and ^{31}P NMR acquisition at pH 6.3. The two singlet ^{31}P peaks thus obtained have chemical shift values close to those obtained in the pH 6.3 ^{31}P NMR time course (within ~±0.2 ppm; **Figure D.26**, **Appendix D**), thus confirming that the change in dinucleotide phosphorus chemical shifts is due to pH change.

The proposed structure of CpICG, as well as the presence of the 5′–5′-imidazole-bridged connectivity, was validated by a combination of 1D homonuclear and 2D heteronuclear coupled NMR spectroscopy (**Figures 4.8.A, 4.8.B**). We looked specifically for the additional ^{31}P-induced splitting of the carbon resonances associated with the bridging imidazole of CpICG. To identify these carbon resonances, we used both ^{1}H–^{13}C coupled gc2HSQCse and gHMBCAD spectroscopy. The doublet ^{1}H peak at 7.87 ppm was assigned to the imidazolyl methine proton, while the ^{1}H peak at 2.44 ppm was

assigned to the methyl protons exocyclic to the imidazole bridge (**Figure D.27**, **Appendix D**). As shown in the partial $^{1}H–^{13}C$ coupled gc2HSQCse spectrum, the imidazolyl methine proton is correlated with the doublet of doublet resonance at 141.8–142.0 ppm (**Figures 4.8.C, 4.8.F**). This is consistent with the expected splitting pattern induced by both the N–P and C–P phosphorus atoms. On the other hand, the partial $^{1}H–^{13}C$ coupled gHMBCAD spectrum clearly demonstrates that the exocyclic methyl protons of the bridging imidazole are coupled to only two carbon resonances, which must be the imidazole bridge carbons that are two and three bonds away (**Figures 4.8.D, 4.8.F**). These resonances at 137.0–137.4 and 131.5–133.4 ppm were also doublets of doublets, consistent with the 5′–5′-phosphate-imidazole-phosphonate linkage regiochemistry. The corresponding carbon signals of ICG only have doublet splitting because the molecule only contains one phosphorus atom (**Figure D.35**, **Appendix D**). Finally, $^{1}H–^{31}P$ coupled gHMBCAD spectroscopy of CpICG (estimated $^{n>1}J_{P,H}$ = 8 Hz), clearly showed that the three protons of the imidazole exocyclic methyl group are coupled to both phosphorus atoms of the imidazole-flanking phosphate and phosphonate (**Figure 4.8.E**). These data, together with the ^{31}P NMR time courses shown in **Figure 4.7**, strongly suggest that ICG reacts with 2-MeImpC to form the 5′–5′-imidazole-bridged CpICG.

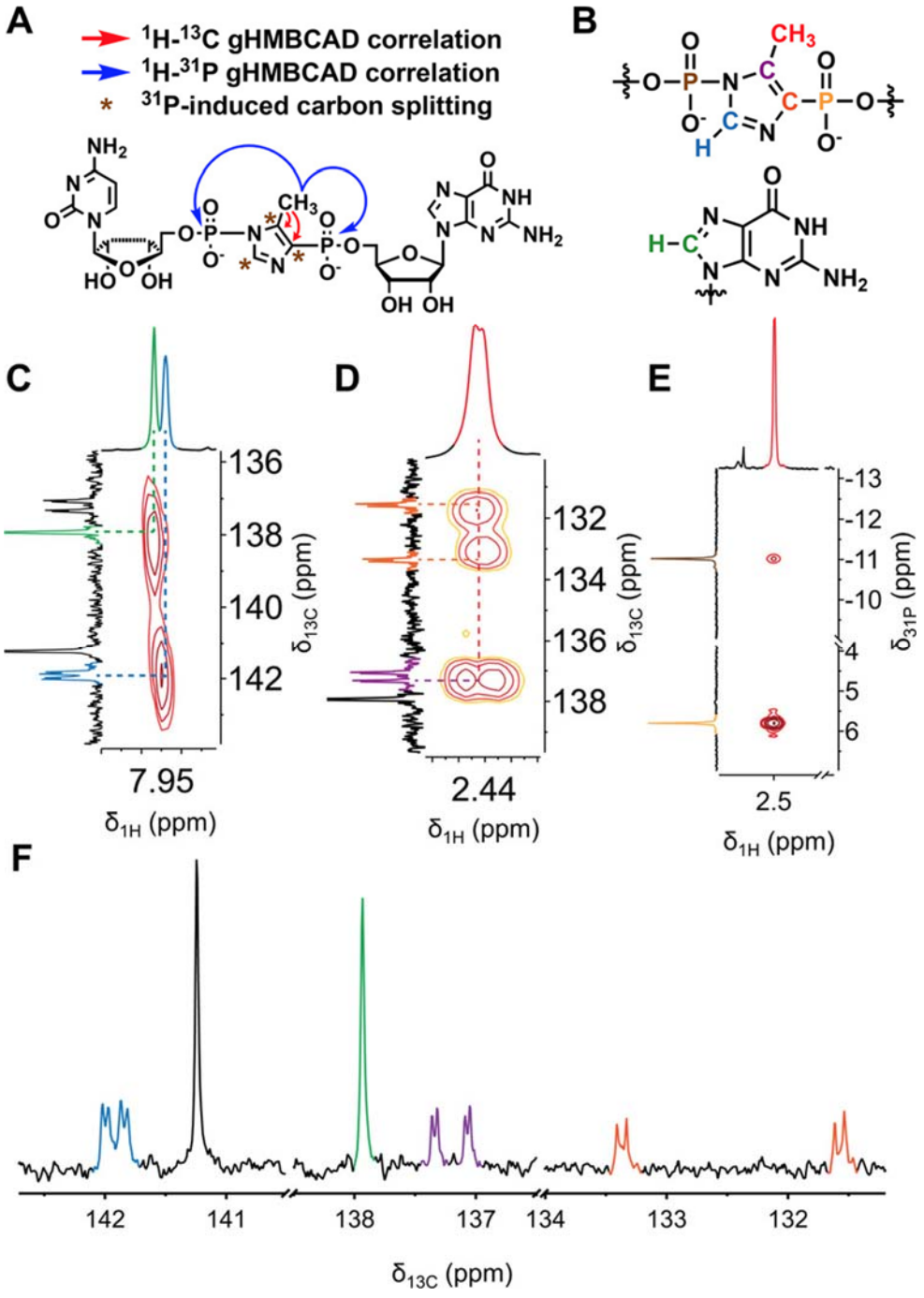


Figure 4.8. (A) Proposed structure of CpICG, and observed heteronuclear NMR correlations which support the proposed structure. (B) The proton, carbon, and phosphorus atoms associated with the CpICG 5′–5′-imidazole linkage have been assigned to their respective NMR resonances using a combination of 1D and homo- and heteronuclear 2D correlation spectroscopies. The color coding of the atoms is used in panels (C)–(F). (C,D) Expanded view of the ^{1}H–^{13}C coupled (C) gc2HSQCse and (D) gHMBCAD spectrum, covering the spectral region showing the proton–carbon

(**Figure 4.8. continued**) correlations relevant to the 5′–5′-imidazole bridge. Flanked by two phosphorus atoms, the carbon resonances of the 5′–5′-bridge are split as doublet-of-doublets, thus affirming the 5′–5′-linkage regiochemistry of CpICG. (E) Expanded view of the ^{1}H–^{31}P-coupled gHMBCAD spectrum. The exocyclic methyl protons of the bridging 2-methylimidazole are correlated with both imidazole-flanking phosphorus atoms of CpICG via long-range coupling. (F) Expanded spectral view of the CpICG ^{13}C{^{1}H} spectrum that shows the carbon signals of the bridging imidazole. All resonances are clearly split as doublets of doublets, which is consistent with the 5′–5′-imidazole-bridged dinucleotide linkage. See **Figures D.27–D.29** (**Appendix D**) for the full ^{1}H, ^{13}C, and ^{31}P spectra of CpICG, and **Figures D.30–D.33** (**Appendix D**) for the full 2D spectrum of ^{1}H–^{1}H gCOSY, ^{1}H–^{13}C gc2HSQCse and gHMBCAD, and ^{1}H–^{31}P gHMBCAD spectra of CpICG.

4.2.6 Primer Extension Reactions with CpICG. Having established that 5′–5′-linked CpICG can form by the reaction of ICG with 2-MeImpC, we asked whether this dinucleotide can bind to its cognate RNA template and react with an upstream primer. We carried out the primer extension experiment as in **Figure 4.4**, using the same RNA primer–template complex and reaction conditions as used for the pseudo-first-order kinetic analysis, but with 20 mM purified CpICG in lieu of the 2-MeImpC + ICG combination (**Figure 4.9.A**; **Figure D.42**, **Appendix D**). At both pH 6.3 and pH 8.0, the primer +1 product can be clearly observed by 15 min, with its abundance increasing at longer reaction times. This result strongly suggests that CpICG can bind to a cognate RNA template and react with the adjacent primer 3′-hydroxyl group. As a control experiment, we reversed the sequence of the templating nucleotides at the primer +1 and +2 positions (underlined nucleotides, **Figure 4.9.A**), from 3′-GCCCAA-5′ to 3′-CGAA-5′. If CpICG binds to this “reversed” template, then the unreactive C–P linkage of the imidazole bridge would be pointing toward the primer 3′-OH nucleophile, instead of the labile N–P linkage, and no primer extension should occur. Indeed, reactivity is completely

abolished with this template, even at the highest tested CpICG concentration (20 mM, **Figure 4.9.A**).

To assess the affinity of CpICG for its cognate RNA primer–template duplex, we determined the K_M of this dinucleotide under primer extension conditions (**Figure 4.9.B**). We measured the pseudo-first-order rate of primer extension as a function of concentration of CpICG at pH 6.3 and 8.0. At pH 6.3, the estimated k_{max} and K_M were 0.28 ± 0.01 h^{-1} and 10.6 ± 0.8 mM, respectively (R^2 = 0.996). At pH 8.0, the estimated k_{max} and K_M were 0.35 ± 0.02 h^{-1} and 12 ± 1 mM, respectively (R^2 = 0.994). The pH change does not significantly alter the affinity of CpICG for its cognate template, and the estimated k_{max} of CpICG-driven primer extension was comparable under both acidic and alkaline conditions, consistent with the similar rates of ICG-catalyzed primer extension with 2-MeImpC at pH 6.3 and 8.0 (k_{obs} = 0.069 ± 0.002 h^{-1} and k_{obs} = 0.066 ± 0.003 h^{-1}, respectively; **Figure 4.4.B** and **Figures D.11** and **D.12**, **Appendix D**).

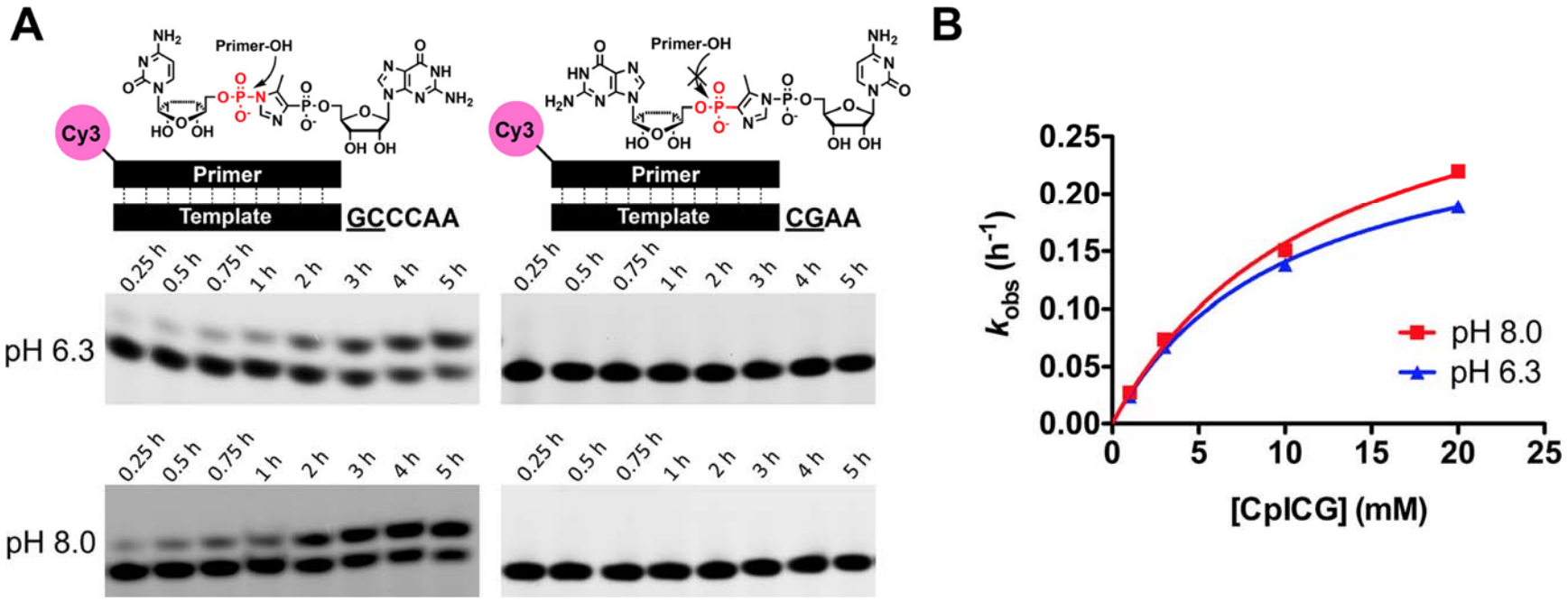


Figure 4.9. (A) Primer extension using 20 mM CpICG as activated substrate. Primer extension could be observed when the cytidylyl group of CpICG, attached to the bridging imidazole via labile N–P linkage, is adjacent to the primer; no extension could be observed when the guanidylyl group, and the stable C–P linkage, is adjacent to the primer. (B) Saturation curve of the CpICG-driven primer extensions at pH 6.3 and 8.0. See **Figures D.43** and **D.44** (**Appendix D**) for representative denaturing PAGE gel images, first-order rate plots, and the saturation curves of CpICG-driven primer extensions at pH 6.3 and 8.0 at different CpICG concentrations.

4.3 Discussion

The mechanism by which nonenzymatic RNA polymerization occurs has been of interest for decades. Since the early 90s, it has been known that the presence of an activated monomer downstream of the reacting nucleotide catalyzes primer extension.[26,45] Our laboratory recently confirmed and extended these early results, showing that the presence of an activated 5′-phosphate on the downstream nucleotide (or oligonucleotide) catalyzed the reaction between the primer and the primer-adjacent nucleotide. Noncovalent leaving group–leaving group interactions, or simply steric or Coulombic repulsions between the leaving groups, were thought to catalyze primer extension by bringing the activated phosphate of the reacting monomer closer to the primer 3′-hydroxyl, or helping to orient the leaving group for in-line nucleophilic substitution. Indeed, our previously reported RNA-PZG crystal structures show that the presence of a downstream PZG nucleotide decreases the distance from the primer 3′-OH to the adjacent PZG phosphate from 6.5 to 4.5 Å,[32] and we have obtained similar results in a more recent study with ICG.[46] However, no leaving group-leaving group interactions were observed in either study. If noncovalent leaving group interactions are a critical component of downstream monomer-induced catalysis, PYG, which has a neutral pyrrolyl leaving group mimic (pK_a < 2), should catalyze upstream primer extension to some degree but no catalysis is observed (**Figure 4.4.B**). Since 4.5 Å is too far for phosphodiester bond formation, other factors must come into play.

Our laboratory has proposed a novel mechanism for primer extension in which the phosphoroimidazolide moieties of two 2-MeImpN first react with each other to form a 5′–5′-imidazolium-bridged pdinucleotide intermediate, followed by template binding, primer

extension by one nucleotide and regeneration of the downstream 2-MeImpN.[27] The Richert lab has also noted the high reactivity of imidazolium-bridged dinucleotides in primer extension.[23] Primer extension reactions driven by partially purified Cp-MeIm-pC display extremely high (>30 h^{-1}) initial rates of primer extension, at a concentration much lower than the >50 mM concentrations of 2-MeImpC commonly employed,[27] suggesting that catalysis involving a 5′–5′-imidazolium-bridged dinucleotide may be the predominant mechanism of primer extension.

In the present study we synthesized ICG as a close nonhydrolyzable analogue of 2-MeImpG, which we initially hoped to use as a probe of catalytically relevant noncovalent interactions. However, no such interactions have been observed. In addition, we were puzzled by the fact that ICG does act as a modest catalyst of primer extension, but only at mildly acidic pH. We then found that ICG could react with 2-MeImpC to form the corresponding 5′–5′-imidazole-bridged GC dinucleotide (CpICG). Indeed, the rates of ICG-catalyzed primer extensions, carried out with 30 mM of the monomers, can be reproduced with only 3 mM of the purified 5′–5′-imidazole-bridged CpICG (k_{obs} = 0.068 ± 0.004 and 0.074 ± 0.001 h^{-1} for pH 6.3 and 8.0, respectively; **Figures D.43** and **D.44**, **Appendix D**), suggesting not only that CpICG is more reactive than the corresponding 2-MeImpC monomer under primer extension conditions, but also that the *in situ* formation of activated 5′–5′-imidazole-bridged dinucleotides could account for the rate of ICG-catalyzed primer extensions shown in **Figure 4.4**. These results complement our recent reports[27,33,47] supporting the hypothesis that the formation of 5′–5′-imidazolium-bridged dinucleotides in nonenzymatic RNA primer extensions is an integral step for 2-MeImpN-driven primer extension.

A few characteristics of CpICG are notable. First, the ^{31}P NMR analysis of the products of 2-MeImpC-ICG coincubation suggest that only one regioisomer out of the four theoretically possible GC dinucleotides (imidazole-, 2′–5′- and 3′–5′-phosphodiester-, and nucleobase-linked) is formed. As shown in the 2-MeImpC-PYG coincubation experiment (**Figures 4.7.B**, **4.7.D**), the ribose hydroxyls and guanine exocyclic amines of the analogues do not detectably react with the electrophilic phosphoroimidazolide moiety of 2-MeImpC. Therefore, formation of the C–P/N–P mixed-linkage dinucleotide occurs with remarkable regioselectivity, with the nonhydrolyzable imidazolylphosphonate of ICG being the exclusive nucleophile during dinucleotide formation. This high regioselectivity simplifies the synthesis and purification of other C–P/N–P mixed-linkage dinucleotide products. Since the identity of the nucleobase is expected to have minimal impact on the yield or regioselectivity of the mixed linkage dinucleotide formation, the preparation of all 16 possible 5′–5′-imidazole-linked, mixed-linkage dinucleotides should be possible. Second, because of the zwitterionic character of the imidazolium bridge, the 2-MeImpN-derived 5′–5′-bridged dinucleotides are inherently reactive, as evidenced by their short half-lives under acidic conditions. Despite careful control of starting material stoichiometry and rigorous purification procedures, preparation and purification of the Cp-MeIm-pC dinucleotide proved to be challenging, affording enriched samples with a maximum of 35% Cp-MeIm-pC.[27] In contrast, the CpICG dinucleotide described herein is quite stable, especially under basic conditions. Even in the presence of a high concentration of Mg^{2+}, CpICG degrades only minimally after 10 h under alkaline conditions (**Figures D.39** and **D.40**, **Appendix D**), and lyophilization of an aqueous CpICG solution did not lead to substantial hydrolysis. The stability of CpICG, which enabled our extensive 1D and 2D

NMR studies of the 5′–5′-imidazole bridge connectivity, will also enable future kinetic, thermodynamic, and crystallographic studies of primer extension. We are currently pursuing crystallographic studies of RNA–CpICG complexes, with the goal of revealing the overall conformation of template-bound CpICG, the noncovalent interactions modulating RNA–CpICG base pairing, the relative orientation and distance between the 5′–5′-imidazole-linkage and the primer 3′-terminus, and other mechanistically relevant structural features.

The most puzzling aspect of ICG catalysis of primer extension with 2-MeImpC was the observation of a moderate catalytic effect at acidic pH, together with the fact that the rate of primer extension was slow but almost pH independent from 6.3 to 8.0. Here we present a possible explanation for these observations. CpICG-driven primer extension requires at least two major mechanistic steps: (1) nucleophilic attack of the primer 3′-OH on the imidazole bridge from the labile N–P bond side, and (2) displacement of the imidazolylphosphonate group to liberate a neutral ICG analogue. At high pH, the 5′–5′-bridging 2-methylimidazole of CpICG is uncharged; direct nucleophilic displacement of the imidazole phosphonate bridge, without prior imidazole protonation or concomitant proton transfer, would result in an unfavorable anionic imidazolyl-phosphonate leaving group. Therefore, the imidazolyl group must first be protonated and become cationic, before nucleophilic attack by the primer 3′-OH takes place. Under acidic conditions, the 5′–5′-imidazole bridge of CpICG is more likely to be protonated (estimated pK_a ≈ 5.1, **Figure D.45**, **Appendix D**), but deprotonation of the primer 3′-OH will be correspondingly more difficult. We suggest that the need to deprotonate the primer 3′-OH while simultaneously protonating the imidazolyl-phosphonate leads to a low primer extension

rate regardless of overall pH, together with a catalytic effect of ICG at acidic pH when the intermediate can be protonated. On the other hand, preliminary data suggested that PZG does react with 2-MeImpC to give the 5′–5′-pyrazole-bridged CpPZG (**Figure D.25**, **Appendix D**), yet PZG is unable to catalyze primer extensions to any significant extent (**Figure 4.4.B**). This lack of reactivity may be due to the low protonation pK_a of the 5′–5′-pyrazole bridge (likely to be 1.94 or less; **Figure 4.1** and **Figure D.2**, **Appendix D**), which renders cationic activation of the pyrazole bridge unlikely when primer extension is carried out at pH ≥5.5.

In conclusion, we have described the synthesis and properties of ICG, the closest nonhydrolyzable 2-MeImpG analogue reported to date. ICG acts as a modest catalyst of primer extension under acidic conditions, primarily via the formation of a 5′–5′-imidazole-bridged dinucleotide which carries both N–P and C–P linkages. This relatively stable mixed-linkage dinucleotide serves as a substrate in primer extension reactions, consistent with our recent mechanistic proposal that the formation of 5′–5′-imidazolium-bridged dinucleotides is an integral step for 2-MeImpN-driven primer extensions. The affinity of these dinucleotides for cognate and noncognate RNA primer–templates, as well as the crystal structures of RNA–dinucleotide complexes, are currently being determined.

4.4 References

(1) Szostak, J. W.; Bartel, D. P.; Luisi, P. L. *Nature* **2001**, *409*, 387.

(2) Blain, J. C.; Szostak, J. W. *Annu. Rev. Biochem.* **2014**, *83*, 615.

(3) Orgel, L. E. *Crit. Rev. Biochem. Mol. Biol.* **2004**, *39*, 99.

(4) Robertson, M. P.; Joyce, G. F. *Cold Spring Harbor Perspect. Biol.* **2012**, *4*, a003608.

(5) Szostak, J. W. *J. Syst. Chem.* 2**012**, *3*, 2.

(6) Prywes, N.; Blain, J. C.; Del Frate, F.; Szostak, J. W. *eLife* **2016**, *5*, e17756.

(7) Kanavarioti, A.; Hurley, T. B.; Baird, E. E. *J. Mol. Evol.* **1995**, *41*, 161.

(8) Kanavarioti, A.; Bernasconi, C. F.; Baird, E. E. *J. Am. Chem. Soc.* **1998**, *120*, 8575.

(9) Rohatgi, R.; Bartel, D. P.; Szostak, J. W. *J. Am. Chem. Soc.* **1996**, *118*, 3340.

(10) Prabahar, K. J.; Ferris, J. P. *J. Am. Chem. Soc.* **1997**, *119*, 4330.

(11) Vogel, S. R.; Deck, C.; Richert, C. *Chem. Commun.* **2005**, *39*, 4922.

(12) Weimann, B. J.; Lohrmann, R.; Orgel, L. E.; Schneider-Bernloehr, H.; Sulston, J. E. *Science* **1968**, *161*, 387.

(13) Inoue, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1981**, *103*, 7666.

(14) Li, L.; Prywes, N.; Tam, C. P.; O'Flaherty, D. K.; Lelyveld, V. S.; Izgu, E. C.; Pal, A.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 1810.

(15) Deck, C.; Jauker, M.; Richert, C. *Nat. Chem.* **2011**, *3*, 603.

(16) Heuberger, B. D.; Pal, A.; Del Frate, F.; Topkar, V. V.; Szostak, J. W. *J. Am. Chem. Soc.* **2015**, *137*, 2769.

(17) Rajamani, S.; Ichida, J. K.; Antal, T.; Treco, D. A.; Leu, K.; Nowak, M. A.; Szostak, J. W.; Chen, I. A. *J. Am. Chem. Soc.* **2010**, *132*, 5880.

(18) Inoue, T.; Orgel, L. E. *J. Mol. Biol.* **1982**, *162*, 201.

(19) Bridson, P. K.; Orgel, L. E. *J. Mol. Biol.* **1980**, *144*, 567.

(20) Giurgiu, C.; Li, L.; O'Flaherty, D. K.; Tam, C. P.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 16741.

(21) Tam, C. P.; Fahrenbach, A. C.; Björkbom, A.; Prywes, N.; Izgu, E. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 571.

(22) Izgu, E. C.; Fahrenbach, A. C.; Zhang, N.; Li, L.; Zhang, W.; Larsen, A. T.; Blain, J. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2015**, *137*, 6373.

(23) Kervio, E.; Sosson, M.; Richert, C. *Nucleic Acids Res.* **2016**, *44*, 5504.

(24) Oró, J.; Basile, B.; Cortes, S.; Shen, C.; Yamrom, T. *Origins Life* **1984**, *14*, 237.

(25) Fahrenbach, A. C.; Giurgiu, C.; Tam, C. P.; Li, L.; Hongo, Y.; Aono, M.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 8780.

(26) Wu, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1992**, *114*, 5496.

(27) Walton, T.; Szostak, J. W. *J. Am. Chem. Soc.* **2016**, *138*, 11996.

(28) Korhonen, H. J.; Conway, L. P.; Hodgson, D. R. W. *Curr. Opin. Chem. Biol.* **2014**, *21*, 63.

(29) Temiakov, D.; Patlan, V.; Anikin, M.; McAllister, W. T.; Yokoyama, S.; Vassylyev, D. G. *Cell* **2004**, *116*, 381.

(30) Upton, T. G.; Kashemirov, B. A.; McKenna, C. E.; Goodman, M. F.; Prakash, G. K. S.; Kultyshev, R.; Batra, V. K.; Shock, D. D.; Pedersen, L. C.; Beard, W. A.; Wilson, S. H. *Org. Lett.* **2009**, *11*, 1883.

(31) Desai, K. K.; Raines, R. T. *Biochemistry* **2012**, *51*, 1333.

(32) Zhang, W.; Tam, C. P.; Wang, J.; Szostak, J. W. *ACS Cent. Sci.* **2016**, *2*, 916.

(33) Zhang, W.; Tam, C. P.; Walton, T.; Fahrenbach, A. C.; Birrane, G.; Szostak, J. W. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 7659.

(34) Van Den Berge, E.; Robiette, R. *J. Org. Chem.* **2013**, *78*, 12220.

(35) Kalek, M.; Ziadi, A.; Stawinski, J. *Org. Lett.* **2008**, *10*, 4637.

(36) McKenna, C. E.; Higa, M. T.; Cheung, N. H.; McKenna, M.-C. *Tetrahedron Lett.* **1977**, *18*, 155.

(37) Taylor, S. D.; Mirzaei, F.; Bearne, S. L. *J. Org. Chem.* **2008**, *73*, 1403.

(38) Rabenstein, D. L.; Sayer, T. L. *Anal. Chem.* **1976**, *48*, 1141.

(39) Bezençon, J.; Wittwer, M. B.; Cutting, B.; Smieško, M.; Wagner, B.; Kansy, M.; Ernst, B. *J. Pharm. Biomed. Anal.* **2014**, *93*, 147.

(40) Krezel, A.; Bal, W. *J. Inorg. Biochem.* **2004**, *98*, 161.

(41) Kanavarioti, A.; Rosenbach, M. T. *J. Org. Chem.* **1991**, *56*, 1513.

(42) Perrin, D. D. Dissociation constants of organic bases in aqueous solution; Butterworths: London, 1965.

(43) Wiseman, T.; Williston, S.; Brandts, J. F.; Lin, L. N. *Anal. Biochem.* **1989**, *179*, 131.

(44) Schmidtchen, F. P. In Analytical Methods in Supramolecular Chemistry; Schalley, C., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2007; pp 55– 78.

(45) Wu, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1992**, *114*, 7963.

(46) Zhang, W.; Tam, C. P.; Zhou, L.; Oh, S. S.; Wang, J.; Szostak, J. W. *Manuscript submitted.*

(47) Walton, T.; Szostak, J. W. *Biochemistry* **2017**, *56*, 5739.

Chapter 5

Downstream Oligonucleotides Strongly Enhance the Affinity of GMP to RNA Primer−Template Complexes

[*C. P. Tam, A. C. Fahrenbach and J.W. Szostak conceived of experiments, analyzed the data and wrote the text. C. P. Tam and A. C. Fahrenbach performed experiments. A. Björkbom, N. Prywes and E. C. Izgu provided technical assistance. This chapter has previously been published with minor changes as the following publication:* Tam, C. P.; Fahrenbach, A. C.; Björkbom, A.; Prywes, N.; Izgu, E. C.; Szostak, J. W. Downstream Oligonucleotides Strongly Enhance the Affinity of GMP to RNA Primer–Template Complexes. *J. Am. Chem. Soc.* **2017**, *139* (2), 571–574.]

Abstract. Origins of life hypotheses often invoke a transitional phase of nonenzymatic template-directed RNA replication prior to the emergence of ribozyme-catalyzed copying of genetic information. Here, using NMR and ITC, we interrogate the binding affinity of guanosine 5′-monophosphate (GMP) for primer−template complexes when either another GMP, or a helper oligonucleotide, can bind downstream. Binding of GMP to a primer−template complex cannot be significantly enhanced by the possibility of downstream monomer binding, because the affinity of the downstream monomer is weaker than that of the first monomer. Strikingly, GMP binding affinity can be enhanced by ca. 2 orders of magnitude when a helper oligonucleotide is stably bound downstream of the monomer binding site. We compare these thermodynamic parameters to those

previously reported for T7 RNA polymerase-mediated replication to help address questions of binding affinity in related nonenzymatic processes.

5.1 Introduction

Originally proposed by Crick[1] and Orgel2 in 1968, the nonenzymatic template-directed replication of RNA is an essential aspect of what has come to be known[3] as the RNA world hypothesis.[4−7] This model suggests that RNA, or perhaps a similar nucleic acid,[8] played a dual role, serving as both a genetic and catalytic polymer.[9−11] Chemically driven RNA replication, which permits the earliest transmission of genetic information from parent protocells to their progeny, arguably forms the foundation for the Darwinian evolution of catalytic RNA sequences, or ribozymes, that could aid in genome replication and protocell metabolism.

Although nonenzymatic template-directed RNA polymerization with activated monomers has been studied since the 1960s,[6,12−14] basic thermodynamic aspects remain unclear. The proposed mechanisms[15] start with a reversible noncovalent binding event in which the activated monomer associates with the template. These interactions can be aided by base stacking from either upstream[16−20] or downstream[19,21−24] adjacent mono- or oligonucleotides. A common experimental protocol employs a template that is annealed to a shorter primer, which can then be extended by making use of activated monomers such as 5′-phosphoro-2-methylimidazolide ribonucleotides.[14] Although the primer enhances the binding of the first monomer through adjacent upstream base stacking,[18−20,25] two questions remain: (i) how tightly do subsequent downstream monomers bind, and (ii) to what degree do downstream mono- or oligonucleotides

enhance the binding of upstream monomers? Richert et al. have used primer-extension inhibitor assays to assess the effect of downstream oligomers using DNA, for which binding constant enhancements of up to ~7-fold were observed;[19] however, direct thermodynamic measurements on RNA-based systems are still needed.

Here we describe the results of our investigations into the effects of downstream elements on the association of GMP to RNA. We measure the binding constants for the Watson−Crick association of GMP to four different primer−template complexes (**Figure 5.1**): one with a single template binding site (5′-C), another with two binding sites (5′-CC), and two more with binding sites supported by both upstream and downstream oligonucleotides. Using ^{1}H NMR spectroscopy and isothermal titration calorimetry (ITC), we characterize the increase in affinity for GMP caused by downstream stacking elements. We demonstrate that when two GMP molecules bind in a consecutive fashion, the second monomer is weaklier bound than the first, and for this reason, a downstream monomer is unable to significantly enhance the occupancy of its neighboring upstream binding site. On the other hand, replacement of the second GMP by a stably bound downstream oligomer enhances the binding affinity of the first upstream GMP by approximately 100-fold.

5.2 Results

We conducted NMR titrations[18] to quantify the binding of GMP to four different RNA primer−template complexes, each possessing distinct templating regions. The templates were designed to assess how different downstream elements affect the affinity of GMP. The sequences of these duplexes are displayed in **Figure 5.1**. The 7-nt primer shared by all duplexes is labeled **P**. The four different templating strands are labeled **T**,

with subscripts that indicate the nature of the templating regions (**T_{1C}**, **T_{2C}**, **T_D**, **T_S**). When paired with primer **P**, templates **T_{1C}** and **T_{2C}** display one and two cytidine nucleotides on their 5′-termini, respectively. Templates **T_D** and **T_S** are designed such that stably bound downstream elements can also pair with the template to create single cytidine binding site(s) for GMP that are flanked by guanidine nucleotides.

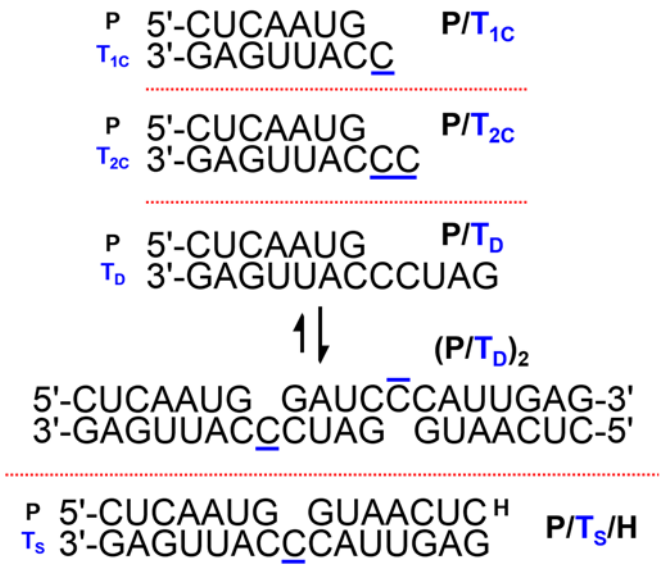


Figure 5.1. Sequences of the RNA duplexes studied by ITC and/or ^{1}H NMR. Each duplex contains a common seven-nucleotide primer **P** and a complementary template strand containing various binding sites for GMP (C, blue underlines). See **Figures E.1–5** for detailed characterizations.

Duplexes **P/T_{1C}** and **P/T_{2C}** were used to assess the thermodynamics governing the association of one and two GMP monomers, respectively. The length and sequence of these primer−template complexes were optimized such that (i) stable duplexes are formed under our NMR titration conditions and (ii) the imino proton signals are defined and well resolved[18] from one another in the 10−15 ppm region of the proton NMR spectrum. The well-resolved imino protons of these two duplexes were monitored during the titrations in order to quantify monomer binding constants. For both the titrations of **P/T_{1C}** and **P/T_{2C}** with GMP, the resonance frequency of the G7 imino proton—the primer

nucleotide closest to the GMP binding sites—exhibits the largest upfield shift, of ~0.35 ppm (**Figure 5.2**), likely because of ring-current effects[26,27] emanating from the purine base of the proximally bound GMP. The fact that G7 shifts the most is consistent with specific binding of GMP to the 5′-overhangs and agrees well with our previously reported results for the binding of GMP to a single-nucleotide template overhang both in solution[18] and the solid state.[28]

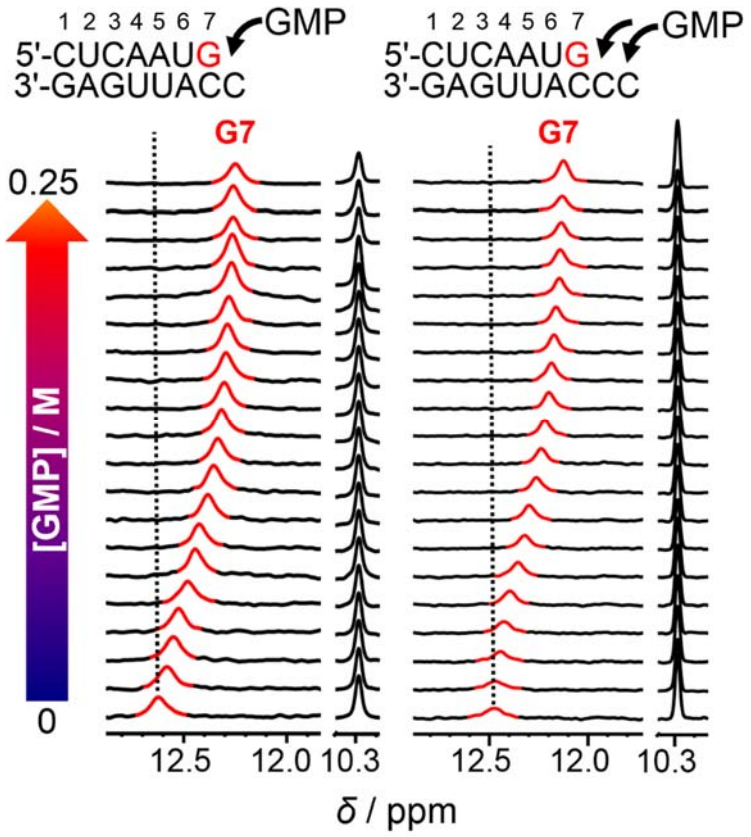
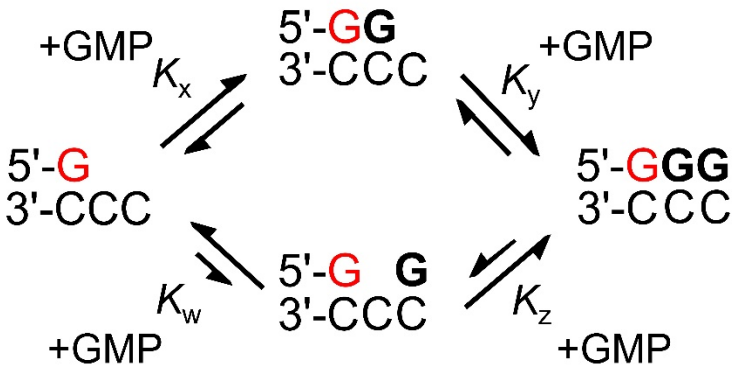


Figure 5.2. Stacked partial ^{1}H NMR spectra for the titrations of the **P/T$_{1C}$** (left) and **P/T$_{2C}$** duplexes (right) with GMP. The G7 imino proton signals are displayed. Both titrations were carried out in 10% D_2O, at 12 ^{o}C and pH 7 with 500 mM of Na^{+} from 0 to 250 mM of GMP. The assignments of the imino protons for both duplexes were carried out by a combination of ^{1}H-^{1}H 2D NOESY and variable temperature ^{1}H NMR spectroscopy (**Appendix E, Figures E.2–4**). **Figure E.6** shows the stacked spectra of both GMP– **P/T$_{1C}$** and GMP–**P/T$_{2C}$** titrations, showing the complete imino proton region (10–15 ppm) with the signals of all of the other imino protons. The change in chemical shift of all seven imino proton resonances plotted against the concentration of GMP is shown in the SI

(**Appendix E, Figure E.7** for **P/T1c**, **Figure E.8** for **P/T2c**). Circular dichroism spectroscopy was performed on these two duplexes, and both display spectral characteristics indicative of canonical A-form conformation (**Appendix E, Figure E.14**).



Scheme 5.1. The proposed GMP binding mechanism for the **P/T2c** duplex. Only the nucleotides constituting the binding sites for GMP are shown. For more details see **Appendix E**, **Scheme E.2**.

We next formulated a model to measure the binding constants for **P/T2c** with GMP employing the sequential binding processes illustrated in **Scheme 5.1**, which is described by four microscopic equilibrium constants. See **Scheme E.2** of **Appendix E** for a detailed derivation procedure. We made two assumptions: (i) sequential binding from the 3′- to the 5′-end of the overhang is the dominant pathway (i.e., $K_x >> K_w$), because G7 affords stacking interactions that increase monomer binding at the adjacent position at least 10-fold;[17] and (ii) binding of the second GMP to the 5′-terminal overhang nucleotide of **P/T2c** does not cause the G7 imino proton resonance to undergo additional upfield shifts, because the ring current arising from the nucleobase of the more distant second monomer is too weak to have any significant effect. With these assumptions, the binding constants for the first (K_x) and second (K_y) monomer associations were determined to be 45(3) and 16(5) M^{-1}, respectively. For comparison, the binding constant determined for the **P/T1c** duplex using a single-binding-site model is 72(2) M^{-1}. With respect to

assumption (ii), we tested alternative scenarios in which the influence of the second monomer on the G7 chemical shift was nonzero; all these scenarios lead to the same conclusion that K_x is larger than K_y. The values of K_x range between 40(1) and 68(2) M^{-1}, and K_y ranges between 14(2) and 42(1) M^{-1} (**Appendix E, Figure E.12**), respectively. For a detailed discussion of how these alternative scenarios of second monomer binding influence the G7 shift, see Section 5 of the Appendix E. We propose that the weaker binding of the second monomer is a consequence of the greater conformational freedom of the bound upstream monomer resulting in a less preorganized stacking surface in comparison to that of G7 in the primer.

We now turn to the question of the degree to which downstream oligonucleotides can enhance the binding of upstream monomers. To study this in depth, we designed the palindromic (**P/T$_D$**)$_2$ duplex (**Figure 5.1**), in which the **T$_D$** strand possesses a 5′-overhang that consists of a cytidine residue for GMP binding, followed by a self-complementary CUAG region that allows for dimerization of the duplex. The C_2 symmetry of the dimerized duplex allows straightforward assignment of all observed imino signals. The (**P/T$_D$**)$_2$ duplex contains two identical binding pockets that are supported by both upstream and downstream oligonucleotides, allowing us to determine the degree to which a stably bound downstream element enhances the association of GMP.

The ^{1}H NMR titration of (**P/T$_D$**)$_2$ with GMP (**Appendix E, Figures S9–10**) revealed that the association is in the tight-binding regime. To accurately quantify the thermodynamic parameters of binding, we used ITC. Since the two binding pockets of the (**P/T$_D$**)$_2$ palindromic dimer are identical, we fit the data shown in **Figure 5.3a** to a two-site model that assumes statistical binding[29] (see **Appendix E** for the derivation, **Scheme**

E.1). The observed binding constant, K_{obs}, was 7.3(3) × 10^3 M^{-1} (ΔG_{obs}(12 °C) = −5.04(2) kcal mol^{-1}), which is approximately 100-fold greater than the binding constant for **P/T$_{1C}$**, and potentially over 150-fold greater than the first binding constant (K_x) for **P/T$_{2C}$**. For detailed thermodynamic parameters for the GMP−(**P/T$_D$**)$_2$ association, see **Table 5.1**. The stoichiometric factor *n* was found to be 1.84, which is consistent with a 2:1 complex. Note that the possible incomplete dimerization or alternative secondary structures of (**P/T$_D$**)$_2$ at the beginning of the titration will lead to a lower effective concentration of (**P/T$_D$**)$_2$ under our experimental conditions. Hence, our experimentally obtained K_{obs} may represent the lower limit of the true value, and the actual binding constant *K* may be significantly larger than that observed here.[30]

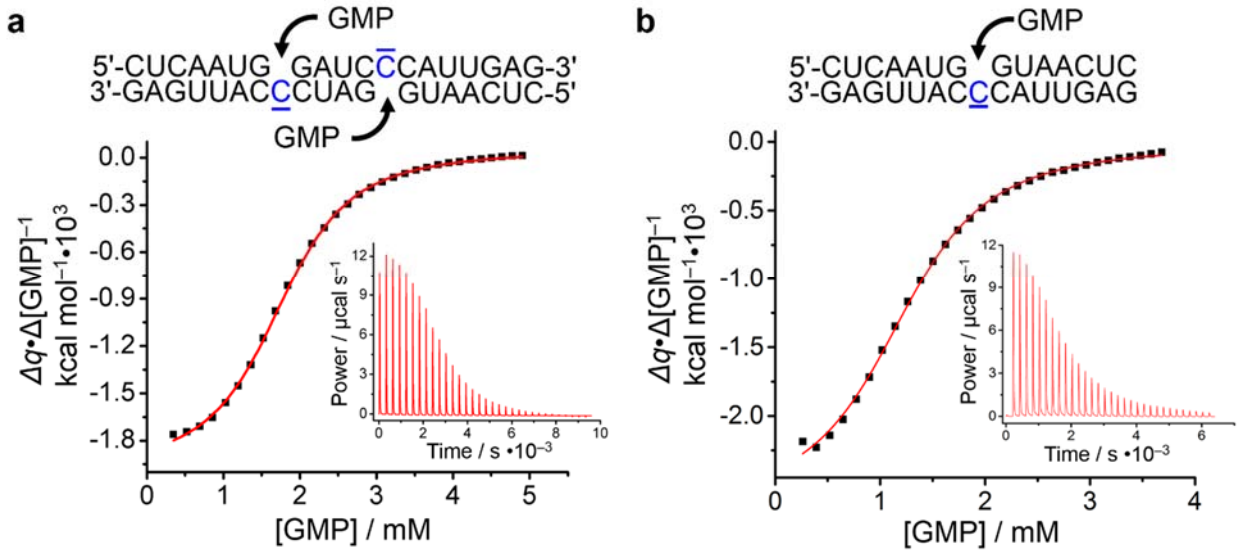


Figure 5.3. ITC for (a) (**P/T$_D$**)$_2$ and (b) **P/T$_S$/H**. Conditions: (**P/T$_D$**)$_2$: 2mM duplex in both the cell and syringe, with additional 20 mM GMP in the syringe, 12 °C; **P/T$_S$/H**: 1.5 mM duplex in both the cell and syringe, with additional 15 mM GMP in the syringe, 20 °C. **Both**: 500 mM NaCl, pH 7. The red lines are the best fits from the statistical-binding isotherm derived in Section 4b of the SI for (a) and a single-binding-site isotherm for (b). Insets: raw power versus time curves. The circular dichroism spectra of both duplexes display spectral characteristics that are indicative of canonical A-form duplexes (**Appendix E, Figure E.14**).

In order to provide insight into the possibility that incomplete dimerization of **P/T$_D$** attenuates the observed binding constant, we designed the "sandwich" duplex **P/T$_S$/H** that possesses a much more stably bound downstream helper oligonucleotide **H** (see **Figure E.11** for the NMR titration, **Appendix E**). ITC carried out on this duplex at 20 °C (**Figure 5.3b**) revealed a similar binding constant for GMP of 6.1(4) × 10^3 M^{-1} and a stochiometric factor *n* of 0.88. See **Table 5.1** for a summary of the thermodynamics for all duplexes. This rather large binding constant is consistent with a simple nearest-neighbor analysis,[31] which predicts a binding constant on the order of 10^4 M^{-1} (see **Section E.8**, **Appendix E**). A previously reported DNA-based sandwich complex,[19] which in a similar fashion flanks an incoming dGMP monomer by two dG nucleotides, leaves the binding constant nearly unchanged compared to when the downstream oligomer is absent. We propose the A-form helical structure of RNA leads to better base-stacking than B-form DNA, resulting in the ~100-fold increase in affinity.

Table 5.1. Summary of Thermodynamic Parameters.

Duplex	*K* (M^{-1})	*ΔG* (kcal mol^{-1})	*ΔH* (kcal mol^{-1})	*ΔS* (cal mol^{-1} K^{-1})
P/T$_{1C}$ [b]	72(2)	–2.42(2)	-	-
P/T$_{2C}$ [b]	K_x:45(3) K_y:16(5)	–2.16(4) –1.6(2)	-	-
(P/T$_D$)$_2$ [c]	7.3(3)•10^3	–5.04(2)	–12.1(1)	–24.8(4)
P/T$_S$/H [d]	6.1(4)•10^3	–5.08(4)	–15.7(3)	–36(1)

[a] All values determined at pH 7 and 500 mM NaCl; [b] Determined by NMR at 12 ºC; [c] Determined by ITC at 12 ºC; [d] Determined by ITC at 20 °C. Values shown in parentheses are errors for the last significant digit determined from the fits.

5.3 Discussion

The experiments described herein constitute a quantitative basis for understanding observations associated with nonenzymatic replication. For example, Richert and co-workers[19,21,22,24] observed that helper oligonucleotides that bind downstream of activated monomers moderately improve the rate of primer extension. We also observed an increase in rate when carrying out primer extension reactions comparing $(\mathbf{P/T_D})_2$ (0.11(1) h^{-1}) to $\mathbf{P/T_{1C}}$ (0.039(1) h^{-1}) (**Appendix E, Figure E.13**). At least part of this ~3-fold rate enhancement can be explained by tighter binding of guanosine 5′-phosphoro-2-methylimidazolide, a thermodynamic effect that increases the fraction of the reactive monomer-bound complex. Using the binding constants measured herein to take this effect into account, we estimate the downstream oligomer only contributes catalytically a factor of ~2-fold to this increase in observed rate.

Some other aspects of primer extension reactions, however, cannot be explained by the same thermodynamic effects. For example, Orgel and co-workers have demonstrated that the rate of nonenzymatic primer extension with respect to the first addition is increased by the presence of downstream templating sites, especially when monomers activated with 2-methylimidazole are employed; the rate of addition is on the order of 20 times faster.[32] We observe a similar increase in rate when comparing $\mathbf{P/T_{2C}}$ (0.42(1) h^{-1}) to $\mathbf{P/T_{1C}}$ (0.039(1) h^{-1}), the former of which is nearly 10 times faster under our conditions (**Appendix E, Figure E.13**). Using the binding constants measured herein, we calculate that that binding site occupancies for $\mathbf{P/T_{1C}}$ and the first position of $\mathbf{P/T_{2C}}$ are comparable (~70−80%). Therefore, the thermodynamics of sequential monomer binding cannot account for the rate enhancing effect previously hypothesized to be caused by a

neighboring downstream activated monomer.[23] We recently reported that the mechanistic origin of this kinetic effect, at least in some cases, is likely a result of the prior formation of an imidazolium-bridged dinucleotide intermediate, which binds more tightly to the template, is more reactive, or both.[33] Whether or not such an intermediate might itself be synthesized on the template, or formed only transiently as part of a concerted mechanism, is still under investigation.

Finally, how do the thermodynamic parameters described herein compare to those measured for an enzymatically assisted primer-extension reaction, e.g., one involving the T7 RNA polymerase? Such a comparison offers an important empirical way to evaluate upper bounds for binding parameters that could lead to acceptable levels of fidelity and rate in the nonenzymatic primer-extension reactions. In essence, such a side-by-side juxtaposition allows us to address the question, how tight is tight enough? The T7 RNA polymerase employs an allosteric mechanism,[34−36] switching from an “open” state, in which the complementary nucleotide triphosphate (NTP) can base-pair to the N+1 site of the DNA template, to a catalytically active “closed” state. It is informative to note that the affinities of NTP substrates to the T7 RNA polymerase ternary complex (measured as K_M constants during elongation of 76−190 μM or K_{assoc} (5.3−13) × 10^3 M^{-1})[37] are on par with the binding constants measured here for GMP association with (**P/T$_D$**)$_2$ or **P/T$_S$/H**, although alternative sandwich complexes employing pyrimidine stacking surfaces may result in much less enhancement. With this caveat in mind, downstream helper oligomers can potentially offer sufficient enhancement of monomer binding, and seeking even tighter binding of activated monomers is not necessary. On the other hand, when comparing the effective rate constant for enzymatic replication (taking into account

initiation, elongation and termination), the T7 RNA polymerase is about 106 times faster[38] than what we observe for **P/T$_{2C}$** (115 s^{-1} vs 0.4 h^{-1}). The challenge in nonenzymatic replication, at least from the perspective of this comparison, is not so much thermodynamic as it is kinetic. With respect to fidelity, the allosteric mechanism of the enzymatic reaction acts to “amplify” the selectivity of Watson−Crick pairing, as only correctly paired NTPs are able to favor the switch to the closed state,[39] thereby helping to ensure high-accuracy copying.

The thermodynamic studies that we have performed point to the fact that the two different mechanisms by which a downstream oligomer or an activated monomer increase the rate of extension are complementary to each other. We recently demonstrated in a separate report that the rates of nonenzymatic template-directed synthesis indeed can be further improved by employing activated helper oligomers.[23] Using this approach, the rate constant increases up to about 1 min^{-1} and an average fidelity of 98% can be achieved, which is sufficiently high to enable the copying of ~50 nt-long mixed-sequence templates, the length of some functional RNAs. Additional investigations into the exact mechanistic effect(s) responsible for the observed fidelity, including a means for selectivity “amplification”, are underway. Understanding the cooperative self-assembly and primer-extension kinetics of activated monomer and downstream oligonucleotides with a wide range of RNA sequences will bring the origins of life community significantly closer to achieving rapid and high-fidelity nonenzymatic copying of catalytically active RNA.

5.4 References

(1) Crick, F. H. C. *J. Mol. Biol.* **1968**, *38*, 367.

(2) Orgel, L. E. *J. Mol. Biol.* **1968**, *38*, 381.

(3) Gilbert, W. *Nature* **1986**, *319*, 618.

(4) Patel, B. H.; Percivalle, C.; Ritson, D. J.; Duffy, C. D.; Sutherland, J. D. *Nat. Chem.* **2015**, *7*, 301.

(5) Orgel, L. E. *Crit. Rev. Biochem. Mol. Biol.* **2004**, *39*, 99.

(6) Blain, J. C.; Szostak, J. W. *Annu. Rev. Biochem.* **2014**, *83*, 615.

(7) Adamala, K.; Szostak, J. W. *Science* **2013**, *342*, 1098.

(8) Taylor, A. I.; Pinheiro, V. B.; Smola, M. J.; Morgunov, A. S.; Peak-Chew, S.; Cozens, C.; Weeks, K. M.; Herdewijn, P.; Holliger, P. *Nature* **2015**, *518*, 427.

(9) Sczepanski, J. T.; Joyce, G. F. *Nature* **2014**, *515*, 440.

(10) Johnston, W. K.; Unrau, P. J.; Lawrence, M. S.; Glasner, M. E.; Bartel, D. P. *Science* **2001**, *292*, 1319.

(11) Bartel, D. P.; Doudna, J. A.; Usman, N.; Szostak, J. W. *Mol. Cell. Biol.* **1991**, *11*, 3390.

(12) Sulston, J.; Lohrmann, R.; Orgel, L. E.; Miles, H. T. *Proc. Natl. Acad. Sci. U. S. A.* **1968**, *59*, 726.

(13) Lohrmann, R.; Bridson, P. K.; Orgel, L. E. *Science* **1980**, *208*, 1464.

(14) Kozlov, I. A.; Orgel, L. E. *Mol. Biol.* **2000**, *34*, 781.

(15) Fahrenbach, A. C. *Pure Appl. Chem.* **2015**, *87*, 205.

(16) Davies, R. J. H.; Davidson, N. *Biopolymers* **1971**, *10*, 1455.

(17) Kanavarioti, A.; Hurley, T. B.; Baird, E. E. *J. Mol. Evol.* **1995**, *41*, 161.

(18) Izgu, E. C.; Fahrenbach, A. C.; Zhang, N.; Li, L.; Zhang, W.; Larsen, A. T.; Blain, J. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2015**, *137*, 6373.

(19) Kervio, E.; Claasen, B.; Steiner, U. E.; Richert, C. *Nucleic Acids Res.* **2014**, *42*, 7409.

(20) Kervio, E.; Sosson, M.; Richert, C. *Nucleic Acids Res.* **2016**, *44*, 5504.

(21) Vogel, S. R.; Deck, C.; Richert, C. *Chem. Commun.* **2005**, 4922.

(22) Röthlingshöfer, M.; Kervio, E.; Lommel, T.; Plutowski, U.; Hochgesand, A.; Richert, C. *Angew. Chem., Int. Ed. Engl.* **2008**, *47*, 6065.

(23) Prywes, N.; Blain, J. C.; Del Frate, F.; Szostak, J. W. *eLife* **2016**, *5*, e17756.

(24) Hagenbuch, P.; Kervio, E.; Hochgesand, A.; Plutowski, U.; Richert, C. *Angew. Chem., Int. Ed. Engl.* **2005**, *44*, 6588.

(25) Sawada, T.; Fujita, M. *J. Am. Chem. Soc.* **2010**, *132*, 7194.

(26) Gomes, J. A. N. F.; Mallion, R. B. *Chem. Rev.* **2001**, *101*, 1349.

(27) Wannere, C. S.; Schleyer, P. v. R. *Org. Lett.* **2003**, *5*, 605.

(28) Zhang, W.; Tam, C. P.; Wang, J.; Szostak, J. W. *ACS Cent. Sci.* **2016**, *2*, 916.

(29) Connors, K. A. *Binding Constants: the Measurement of Molecular Complex Stability*; Wiley: New York, 1987.

(30) Larsen, A. T.; Fahrenbach, A. C.; Sheng, J.; Pian, J.; Szostak, J. W. *Nucleic Acids Res.* **2015**, *43*, 7675.

(31) Turner, D. H.; Sugimoto, N.; Freier, S. M. *Annu. Rev. Biophys. Biophys. Chem.* **1988**, *17*, 167.

(32) Wu, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1992**, *114*, 5496.

(33) Walton, T.; Szostak, J. W. *J. Am. Chem. Soc.* **2016**, *138*, 11996.

(34) Temiakov, D.; Patlan, V.; Anikin, M.; McAllister, W. T.; Yokoyama, S.; Vassylyev, D. G. *Cell* **2004**, *116*, 381.

(35) Steitz, T. A. *Curr. Opin. Struct. Biol.* **2009**, *19*, 683.

(36) Steitz, T. A. In *RNA Polymerases as Molecular Motors*; Buc, H., Strick, T., Eds.; RSC Pub.: Cambridge, U.K., 2009; p 96.

(37) Vaught, J. D.; Dewey, T.; Eaton, B. E. *J. Am. Chem. Soc.* **2004**, *126*, 11231.

(38) Arnold, S.; Siemann, M.; Scharnweber, K.; Werner, M.; Baumann, S.; Reuss, M. *Biotechnol. Bioeng.* **2001**, 72, 548.

(39) Sydow, J. F.; Cramer, P. *Curr. Opin. Struct. Biol.* **2009**, 19, 732.

Chapter 6

Structural Rationale for the Enhanced Catalysis of Nonenzymatic RNA Primer Extension by a Downstream Oligonucleotide

[*W. Zhang and J. Szostak conceived of experiments; W. Zhang performed experiments. W. Zhang, C. P. Tam and J. Szostak analyzed the data and wrote the texts. L. Zhou performed the primer extension assays. S. S. Oh prepared the T7-RNA polymerase-transcribed, ICG-capped helper oligonucleotide. J. Wang provided technical assistance. This chapter has previously been published with minor changes as the following publication:* Zhang, W.; Tam, C. P.; Zhou, L..; Oh, S. S.; Wang, J.; Szostak, J. W. Structural Rationale for the Enhanced Catalysis of Nonenzymatic RNA Primer Extension by a Downstream Oligonucleotide. *J. Am. Chem. Soc.* **2018**, *140*, 2829–2840.]

Abstract. Nonenzymatic RNA primer extension by activated mononucleotides has long served as a model for the study of prebiotic RNA copying. We have recently shown that the rate of primer extension is greatly enhanced by the formation of an imidazolium-bridged dinucleotide between the incoming monomer and a second, downstream activated monomer. However, the rate of primer extension is further enhanced if the downstream monomer is replaced by an activated oligonucleotide. Even an unactivated downstream oligonucleotide provides a modest enhancement in the rate of reaction of a primer with a single activated monomer. Here we study the mechanism of these effects

through crystallographic studies of RNA complexes with the recently synthesized nonhydrolyzable substrate analog, guanosine 5′-(4-methylimidazolyl)phosphonate (ICG). ICG mimics 2-methylimidazole activated guanosine-5′-phosphate (2-MeImpG), a commonly used substrate in nonenzymatic primer extension experiments. We present crystal structures of primer-template complexes with either one or two ICG residues bound downstream of a primer. In both cases, the aryl-phosphonate moiety of the ICG adjacent to the primer is disordered. To investigate the effect of a downstream oligonucleotide, we transcribed a short RNA oligonucleotide with either a 5′-ICG residue, a 5′-phosphate or a 5′-hydroxyl. We then determined crystal structures of primer-template complexes with a bound ICG monomer sandwiched between the primer and each of the three downstream oligonucleotides. Surprisingly, all three oligonucleotides rigidify the ICG monomer conformation and position it for attack by the primer 3′-hydroxyl. Furthermore, when GpppG, an analog of the imidazolium-bridged intermediate, is sandwiched between an upstream primer and a downstream helper oligonucleotide, or covalently linked to the 5′-end of the downstream oligonucleotide, the complex is better pre-organized for primer extension than in the absence of a downstream oligonucleotide. Our results suggest that a downstream helper oligonucleotide contributes to the catalysis of primer extension by favoring a reactive conformation of the primer-template-intermediate complex.

6.1 Introduction

Template-directed nonenzymatic RNA primer extension is an important experimental model for the earliest mode of genome propagation.[1-4] In this system, mononucleotides or oligonucleotides are chemically activated and bound to a nucleic acid primer-template complex, followed by primer extension through monomer polymerization or oligomer ligation. The concept can be traced back to work by Orgel and coworkers in the late 1960s.[5] Subsequently, 2-methylimidazole-activated monoribonucleotides were discovered to be relatively efficient at template-directed nonenzymatic RNA primer extension in the presence of divalent metal cations.[6] Several significant factors have been found to further accelerate the reaction, especially for templates containing adenosine and uridine nucleotides, which are notoriously difficult to copy.[7-9] Modified bases, distinct activating groups, and downstream oligonucleotides all influence the reaction. The affinities of nucleotides for RNA primer-template complexes, together with the kinetics of the subsequent oligomerization, have been explored by our lab and others.[10-13] The reaction mechanism has been assumed for many years to be a classic in-line S_N2 nucleophilic substitution mechanism in which the 3′-hydroxyl group of the RNA primer attacks the electrophilic phosphorus atom of the incoming activated monomer to form a new covalent O-P bond, and the activating group (e.g.: 2-methylimidazole) is displaced.[14] However, it has also been known for many years that the rate of primer extension is accelerated by the binding of an activated monomer[15] or oligomer[8,9] downstream of the polymerizing nucleotide. The activating groups of the downstream monomers/oligomers play a catalytic role, which was initially thought to result from noncovalent interactions between the leaving groups of adjacent nucleotides or oligomers. To test this hypothesis

directly, we previously designed and synthesized a stable analog of an activated monomer, guanosine 5′-(3-methyl-1*H*-pyrazol-4-yl)phosphonate (PZG), and cocrystallized it with an RNA primer-template complex for structure determination.[16] Unexpectedly, the pyrazolyl-phosphonate groups of the PZGs were not only too far from the 3′-hydroxyl group of the primer to react, but also severely disordered. The absence of any detectable noncovalent interaction between the leaving groups of adjacent monomers suggested that the catalytic effect stems from some other chemical phenomenon.

Recent work from our lab strongly suggests that a 5′-5′ imidazolium-bridged dinucleotide is the actual intermediate in the RNA primer extension reaction; the Richert lab has also noted the high reactivity of this species.[17,18] Under the optimal pH environment (at the pK_a of the leaving group), two activated monomers react with each other to form an imidazolium-bridged dinucleotide, which can then bind to the template. The primer 3′-hydroxyl then attacks the phosphorus atom of the adjacent incoming nucleotide, but the displaced species is an entire activated monomer, rather than just the activating group. To examine the structural underpinning of this reaction model, we have used P^1,P^3-diguanosine-5′-triphosphate (GpppG) as a stable analog of the imidazolium-bridged dinucleotide intermediate.[19] The RNA-GpppG structure showed that both G nucleobases of the dinucleotide analog interact with the template by canonical Watson-Crick base pairing, and NMR experiments demonstrated that GpppG binds cognate RNA templates with significantly enhanced binding affinity relative to activated monomer. Meanwhile, the conformational constraints imposed by the covalent internucleotide bridge

and the two Watson-Crick base pairs preorganize the bound dinucleotide in a geometry that would favor in-line nucleophilic attack.

Chemically stable analogs of activated nucleotides are crucial for structural studies because the high reactivity of the imidazole-activated monomers and imidazolium-bridged dinucleotide precludes these molecules from being cocrystallized with RNA. We have therefore designed and synthesized several stable analogs that closely mimic the structure and basicity of the activated monomers. We recently synthesized guanosine 5′-(4-methylimidazolyl)phosphonate (ICG), which has a structure and pK_a (of its nonhydrolyzable heteroaromatic group) similar to that of 2-MeImpG (**Figure 6.1.A**).[20] Here, we report the crystallographic analysis of a series of RNA:ICG complexes in which an ICG monomer lies adjacent to the 3′-end of a primer. Because ICG residues, unlike the actual 2-MeImpG substrates, cannot react with each other to form the 5′-5′ imidazolium-bridged intermediate, we were able to specifically address the effects of downstream monomers, consecutive monomers, and oligonucleotides on the conformation of the primer relative to the adjacent ICG monomer. In addition, we describe the structures of two new RNA:GpppG complexes, in which the GpppG dinucleotide is either sandwiched between a primer and a downstream oligonucleotide, or covalently attached to the downstream oligonucleotide. In all cases, we find that in the presence of a downstream oligonucleotide, the primer 3′-hydroxyl and the phosphorus atom of the adjacent ICG or GpppG are both closer and more properly aligned for nucleophilic attack. The data presented here, together with previous work on the reactive imidazolium-bridged intermediate, suggest that the rate of nonenzymatic primer extension is increased by both covalent nucleophilic catalysis and by conformational constraint of the reactive complex.

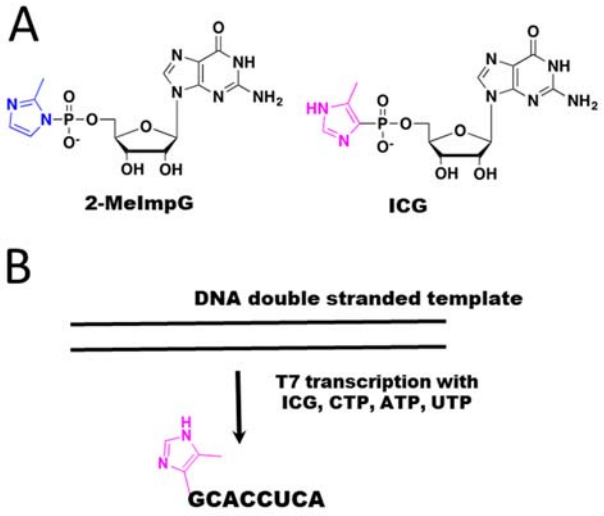


Figure 6.1. (A) Structures of 2-MeImpG and the ICG analog. (B) T7 *in vitro* transcription of RNA helper 8mer containing ICG at the 5′-end.

6.2 Results

6.2.1 RNA duplex-monomer cocrystal structures reveal multiple modes of monomer binding. We used the nucleotide analog guanosine 5′-(4-methylimidazolyl)phosphonate (ICG), which has a nonhydrolyzable heteroaryl-phosphonate moiety that mimics the structure and properties of the leaving group of the reactive monomer 2-MeImpG,[21] to investigate the binding mode of activated monomers to a primer-template complex. We began with an RNA that would generate one monomer binding site at each end of the duplex formed by annealing of the self-complementary oligonucleotide 5′-*CCCG*ACUUAAGUCGG-3′. The four italicized nucleotides at 5′-end are LNA nucleotides in which the sugar is locked in the 3′-*endo* conformation. *C* denotes locked 5-methylcytidine, which has similar basicity as native cytidine and Watson-Crick base pairs with guanosine.[22] The 1.9 Å resolution ICG-RNA structure showed one ICG molecule bound to the templating 5-methylcytidine at each end of the duplex. The RNA-monomer duplexes are slip-stacked with one another, which may stabilize monomer

binding by providing additional stacking interactions. Interestingly, ICG bound to the templating *C* residues in two different modes. At one end, canonical Watson-Crick base pairing was observed while at the other end an alternative base pair was present in which the guanine N3 of ICG was hydrogen bonded with the exocyclic amine of 5-methylcytidine, while the exocyclic amine of guanine was hydrogen bonded with N3 of 5-methylcytidine. For the Watson-Crick paired monomer, the sugar and phosphonate were moderately well-ordered, with a primer O3′-P distance of 6 Å. However, the imidazolyl moiety was entirely disordered, so that the O3′-P-C angle of attack could not be determined. For the noncanonical base pair, both the sugar and the imidazolyl-phosphonate were highly disordered. The crystallographic and structural features of determined structures are listed in **Table F.1**. Schematic diagrams and the crystal structure are shown in **Figure 6.2.A**.

To explore the effects of a second monomer bound downstream of the primer-adjacent monomer, we cocrystallized ICG with a 14mer RNA similar to that described above but with two binding sites at each terminus (5′-*CCCG*ACUUAAGUCG-3′). In this structure, each pair of available binding sites was occupied by ICG, forming *C*:G base pairs with the template and contributing to the head-to-head slip-stacked interactions between RNA-ICG complexes (**Figure 6.2.B**). Thus, consecutive available sites can be occupied by ICG monomers, and presumably 2-MeImpG monomers, prior to polymerization. As in the single binding site case above, at one end of the RNA duplex both ICG monomers bound through Watson-Crick base pairing, but at the other end the two monomers bound differently, the primer-adjacent monomer through Watson-Crick pairing and the distal ICG through noncanonical *C*:G pairing (**Figures 6.2.C** and **6.2.D**).

This again demonstrates the possibility of multiple binding modes even for otherwise identical substrate-monomer complexes. At the end where both ICGs were Watson-Crick paired, the sugars and imidazolyl-phosphonates of both residues were highly disordered, so that the O3′-P distance and the O3′-P-C angle of attack could not be determined. Interestingly, at the other end of the duplex, the presence of the distal monomer made the proximal monomer more structurally ordered, with an O3′-P distance of 4.6 Å. However, the imidazolyl moiety was disordered so the O3′-P-C angle of attack could not be determined; in addition, only the nucleobase of the distal noncanonically paired ICG was visible in the electron density. Nevertheless, this suggests that occupancy of a downstream site by even a single additional nucleotide is in some cases sufficient to introduce a degree of conformational constraint on the upstream monomer. The disorder of the aryl moieties of the monomers indicates that any noncovalent interactions between the leaving groups of adjacent monomers are insufficient to lead to detectable conformational order, and may not play a significant role in catalyzing primer extension. The above structures are similar overall to those previously observed with the pyrazolyl-phosphonate (PZG) analog of 2-MeImpG.[16] The reproducible observation of a noncanonical *C*:G base pair in both the one and two binding site RNA-monomer complexes also suggests that such pairing modes may influence the rate and fidelity of nonenzymatic primer extension.

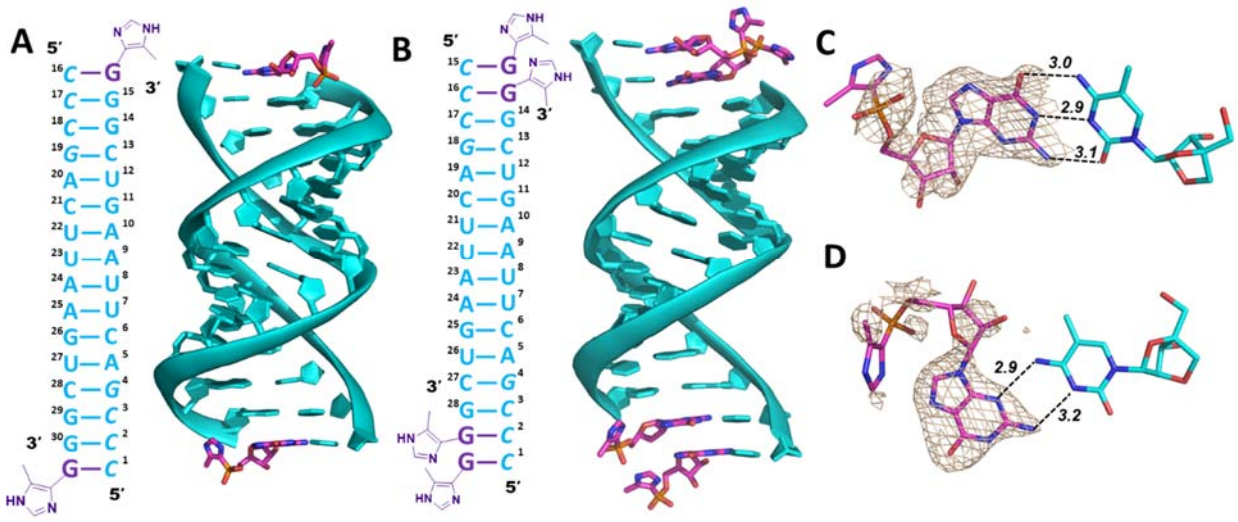


Figure 6.2. Structures of RNA duplex-ICG complexes. (A) Diagram and crystal structure of RNA duplex with one ICG monomer bound to the 5-Me-*C* overhang at each end. (B) Diagram and crystal structure of RNA duplex with two ICG monomers bound to the 5-Me-*CC* overhangs at each end. The monomers are stacked with one another and with the ends of two RNA duplexes. (C) At the top of the structure shown in panel B, the proximal ICG monomer forms a Watson-Crick base pair with the templating C. (D) At the top of the structure shown in panel B, the distal ICG monomer forms a noncanonical base pair with the templating *C*. The 4-methylimidazolyl-phosphonate and sugar moieties are disordered. Cyan: RNA duplex; Magenta: ICG monomer. Hydrogen bonds are labeled as black dashes and their interatomic distances are indicated. Wheat mesh indicates the corresponding $2F_o–F_c$ maps contoured at 1.5 σ. The sequence of the oligonucleotide is shown and numbered.

6.2.2 Downstream helper oligonucleotides facilitate RNA primer extension. We have previously reported that downstream activated oligonucleotides facilitate primer extension to a greater extent than a downstream activated monomer. To explore the structural effects of a downstream activated RNA oligonucleotide, we synthesized an RNA oligonucleotide that was capped at its 5'-end with the nonhydrolyzable ICG nucleotide to provide a stable mimic of an activated helper oligonucleotide (5′-ICG-CACCUCA-3′) (**Figure 6.1.B**, details in **Appendix F**).

We first verified that 5′-ICG-CACCUCA-3′ would pair with the RNA template and function as a downstream helper to enhance nonenzymatic primer extension. Here we used an RNA hairpin system such that the RNA strand is partially self-complementary and functions as both template and primer (**Figure 6.3.A**). We measured the rate of RNA primer extension with a single 2-MeImpC monomer, in the presence of three different downstream helpers including a 5′-hydroxyl octamer, 5′-phosphorylated octamer, and the 5′-ICG capped octamer (**Figure 6.3.A**). The rate of monomer addition without any downstream helper was measured for comparison (**Figures 6.3.B** and **6.3.C**). In the absence of any downstream helper, the nonenzymatic addition of a single activated C monomer was extremely slow, and after a 20 h incubation, the primer extension product was not detectable. A downstream 8-mer RNA with either a 5′-hydroxyl or 5'-phosphate enhanced primer extension rates to about 0.03 h^{-1}. This phenomenon is consistent with previous observations that a 5'-phosphorylated or 5′-hydroxyl RNA downstream oligo significantly enhances the rate of RNA primer extension.[7,9] These downstream helper oligonucleotides may help to constrain the primer-monomer-template geometry in the A-form conformation known to favor reactivity; in addition, base stacking and other interactions with the reactive monomer may help to preorganize the reaction center for in-line nucleophilic attack. When the downstream RNA octamer contained a 5′-(4-methylimidazolyl)phosphonate moiety, the rate of primer extension was further increased by 4-fold to ~0.12 h^{-1} compared to the reaction with the 5′-hydroxyl or 5'-phosphate helper oligonucleotide. While the specific catalytic effect of the 4-methylimidazolyl-phosphonate group is modest compared to the effect seen with a 2-methylimidazole-phosphate group,[9] we show elsewhere that this effect also occurs through the formation of an uncharged

imidazole-bridged intermediate between the activated monomer and the ICG-capped downstream oligonucleotide.[21]

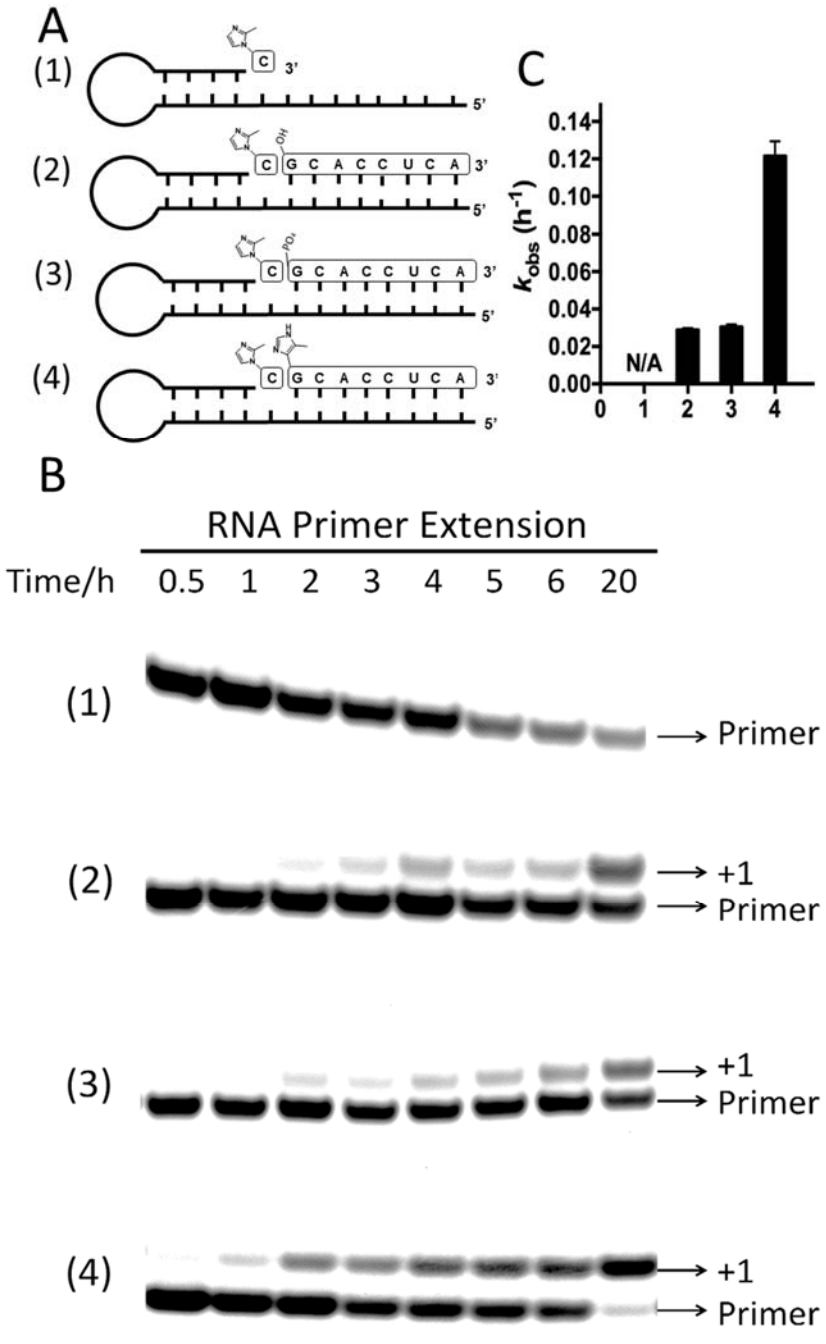


Figure 6.3. Primer extension assay. (A) Schematics of the RNA primer-template-monomer-helper oligonucleotide complexes used in primer extension experiments. Templates are complementary to the displayed monomers and oligonucleotides. (B) Primer extension assays without and with three different downstream helper oligonucleotides. (C) Pseudo-first order rates of the reactions described in A. ND: not detectable.

6.2.3 Crystal structures of RNA hairpin complexes containing different downstream helpers. To understand the structural basis for the successively increasing reaction rates seen in the presence of an unactivated and an activated downstream oligonucleotide, we cocrystallized RNA hairpin-monomer complexes together with three distinct oligonucleotides. We began by using nonhydrolyzable ICG in both the target monomer position and the 5′-terminus of the helper oligonucleotide to capture the state of the complex prior to formation of the imidazolium-bridged intermediate and primer extension. We replaced two template nucleotides that pair with the ICG monomer and the first ICG nucleotide of the 8-mer helper with locked 5-methyl *C* nucleotides (**Figure 6.4.A**, *C8* and *C9*) to rigidify the template and facilitate crystallization. The RNAs and monomer were assembled into a 34 nucleotide RNA hairpin complex (termed H-34-Im; **Figure 6.4.A**). The RNA-monomer hairpin complex crystallized in the tetragonal $P4_32_12$ space group, with one RNA-ICG complex per asymmetric unit. The previously determined structure of the GAAA tetraloop and 4bp of the adjacent stem[23] was used as the search model for molecular replacement. The unmodeled part of the 15 bp stem was visible after the first run of refinement and was gradually modeled into the density (**Figure 6.4.B**). The stem starts with a stable GC Watson-Crick base pair that closes the GAAA tetraloop, while the weak terminal AU was unpaired. The overall structure conformed to an A-form helix (except the bottom of the stem), which is expected to favor monomer binding and primer extension. Each RNA-monomer complex made significant molecular contacts with a neighboring complex. Pairs of side-by-side complexes interacted in a head-to-tail orientation, with hydrogen bonding interactions between the tetraloop (A17, A18, A19, G20) and the bottom part of the stem (G4, G5, C31, U32) (**Figure 6.4.B**). Satisfyingly,

there were no intermolecular interactions surrounding the central part of the stem (C25:G10-C28:G7), indicating that our experimental observations would predominantly reflect an unperturbed structure rather than artifacts due to molecular packing.

Based on previous computational analysis of phosphoroimidazolide hydrolysis, nucleophilic attack on the phosphorus atom of the phosphoroimidazolide likely goes through a loose A_ND_N transition state due to the weak P-N bond. If the same mechanism applies to primer extension, the alignment of the nucleophile-P-leaving group should be close to 180° in the transition state, and angles closer to 180° in the ground state will favor the reaction.[24] Therefore, in our structures, we measured both the distance between the primer 3′-OH and the electrophilic phosphorus atom, and the angle formed by the primer 3′-OH and the P-C bond of the ICG monomer or the P-O bond of GpppG. The ICG monomer (G26) was bound to *C*9 in the template, and was sandwiched between the upstream cytidine (C25) of the primer and the downstream ICG (G27) of the RNA helper oligonucleotide (**Figures 6.4.A** and **6.4.B**). The ICG monomer was Watson-Crick base paired with the template *C*9, and the ICG nucleotide at the 5'-end of the downstream helper was also Watson-Crick paired, with the template *C*8. All hydrogen bond distances ranged between 2.8 to 3.2 Å (**Figure 6.4.C**). Unlike the disordered sugars in the RNA duplex-ICG structures, the ICG electron density in the H-34-Im complex was well defined and the ribose sugar was clearly in the 3′-*endo* conformation, as was the rest of the A-form hairpin stem. Remarkably, the 4-methylimidazolyl group of the ICG monomer was also moderately well-ordered so that for the first time it was possible for us to define the orientation and geometry of the aryl phosphonate moiety. Strikingly, the distance between the 3′-hydroxyl group of the primer and the phosphorus atom of ICG was 3.8 Å. This is

the shortest primer-monomer distance we have yet observed among all RNA-monomer structures. Furthermore, the angle between the 3′-OH and the P-C bond of ICG was 100.8°. Evidently, binding of the ICG-capped downstream helper forced the sandwiched ICG monomer into a more ordered conformation, with distance and angle parameters that are favorable for S_N2 displacement. Surprisingly however, the distance between the 2′-OH and the P atom of the ICG was even shorter at 3.7 Å, and the corresponding angle of 120.6° even more favorable. Therefore, in the context of a lone RNA monomer caught between a primer and a downstream helper, S_N2 displacement by an attacking 2′-hydroxyl group appears to be at least equally likely. Significantly, we did not observe any electron density that could be fit to the 4-methylimidazolyl moiety of the helper 8-mer. The aryl group was completely disordered in the structure, and in addition no Mg^{2+} ions or water molecules were identified as making significant contributions to the reaction center geometry (**Figure 6.4.C**). The absence of detectable interactions indicates that the conventionally assumed noncovalent leaving group-leaving group interactions between incoming monomer and the downstream helper monomer/oligomer are unlikely to be responsible for helper-facilitated RNA primer extension (see Discussion).

The disordered 4-methylimidazolyl moiety of the helper 8-mer raised the question of whether this moiety contributed to the well-ordered primer-monomer structure, or whether an unactivated downstream oligonucleotide would have a similar effect. We therefore crystallized two additional RNA hairpin complexes containing bound ICG monomers (hereafter termed H-34-PO and H-34-OH) in which the downstream helpers were RNA 8-mers with only a 5′-phosphate or 5′-hydroxyl group, respectively (**Figure 6.5.A**). Both structures were similar overall to H-34-Im: the ICG monomer stacked with

upstream and downstream nucleobases and Watson-Crick paired with the templating *C*9 (**Figures 6.5.B** and **6.5.C**). However, the imidazolyl-phosphonate groups of the bound ICG monomers in both the H-34-PO and H-34-OH complexes were more disordered than in the H-34-Im complex. While the distances between the 3′-hydroxyl groups of the primer and the phosphorus atom of ICG were almost unchanged (3.8 Å), the distances between the 2′-OHs and phosphorus atoms became longer (~4.2 Å vs 3.7 Å). The corresponding angles allowing S_N2 displacement to occur also varied (~110° vs 100.8°). Again, no defined water molecules or metal ions were observed near the reaction center within either complex. The lack of downstream bulky aryl groups in both the H-34-PO and H-34-OH complexes may have led to decreased steric crowding, such that the aryl groups of bound ICG monomers are more dynamic, manifesting as disorder in our structures.

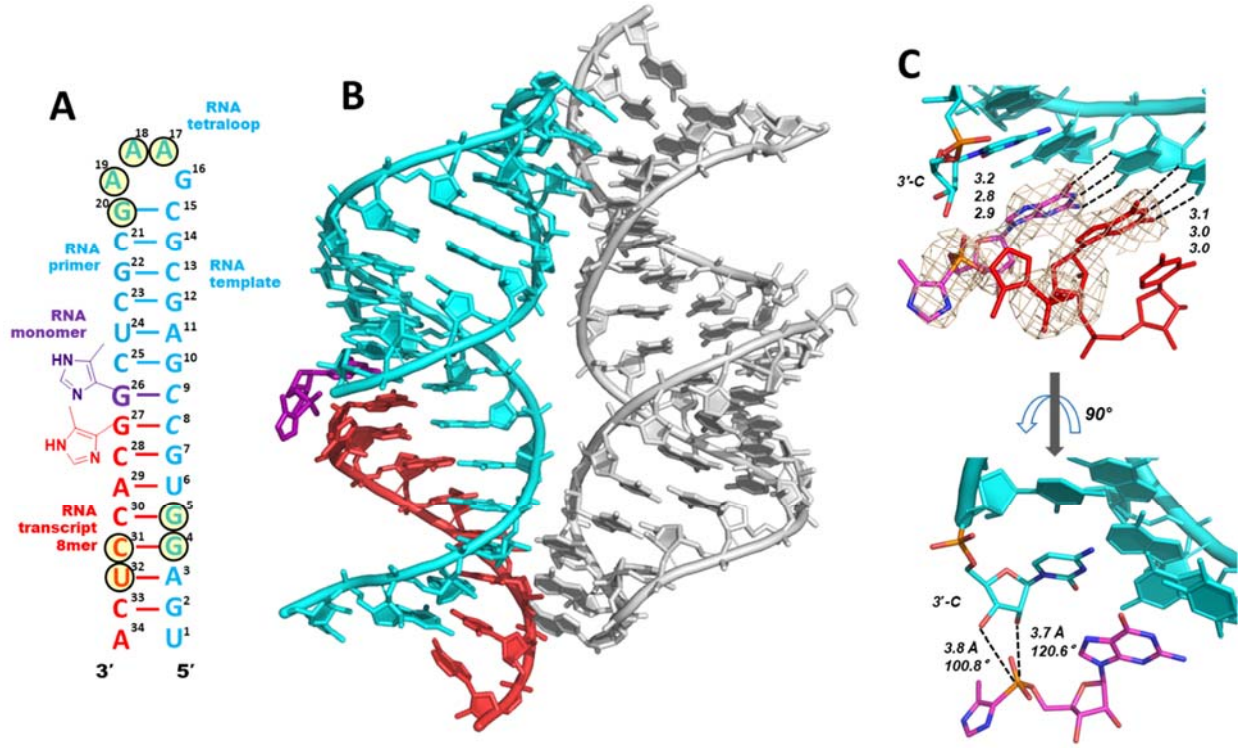


Figure 6.4. Structure of RNA hairpin-ICG H-34-Im complex. The ICG monomer is sandwiched within the hairpin structure. (A) Diagram and designed secondary structure of RNA hairpin-ICG complex. The nucleotides critical for crystallization are marked by circles. (B) Crystal structure of H-34-Im complex. Each RNA-monomer complex significantly interacts with one neighboring complex (grey color on the right): at the tetraloop and at the lower part of the helix. (C) Closer views of the bound ICG monomer. Both sandwiched ICG monomer and the first ICG nucleotide of the transcript form Watson-Crick base pairs with the template. Distinct regions of electron density demonstrate the absense of noncovalent interactions between the two 4-imidazolyl groups. Cyan: RNA template-primer; Magenta: ICG monomer; Red: transcribed ICG capped RNA. Hydrogen bonds are labeled as black dashes and their interatomic distances are indicated. Wheat mesh indicates the corresponding $2F_o–F_c$ maps contoured at 3.5 σ. The sequence of the oligonucleotide is shown and numbered.

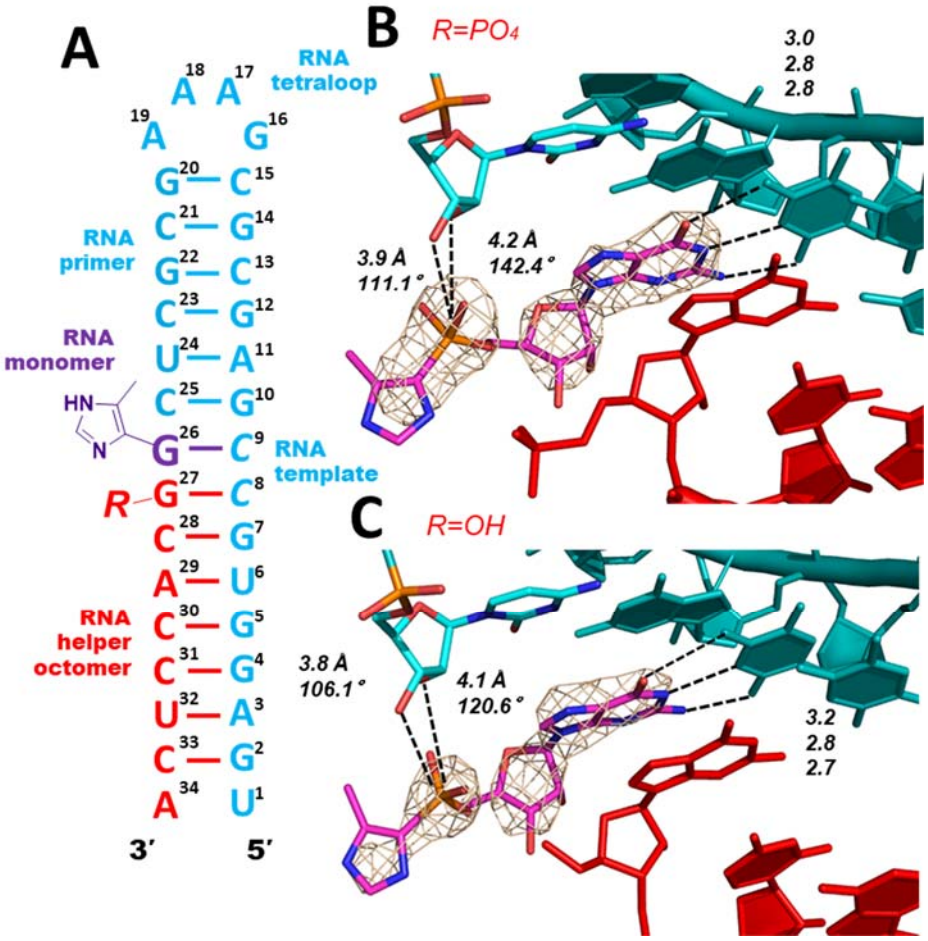


Figure 6.5. Structures of RNA hairpin-ICG H-34-PO and H-34-OH complexes. (A) Diagram and designed secondary structure of the complexes. (B) and (C) Closer views of the bound ICG monomers in both H-34-PO and H-34-OH complexes. In both cases, bound ICG monomers and the first G nucleotides of the downstream helpers form Watson-Crick base pairs with the template. Weaker electron density indicates that the ICG monomers are more disordered than in Fig. 4. Color-coding as in **Figure 6.4**. Hydrogen bonds are labeled as black dashes and their interatomic distances are indicated. Wheat mesh indicates the corresponding $2F_o–F_c$ maps contoured at 3.5 σ.

6.2.4 Crystal structures of RNA hairpin complexes containing tetraloop/receptor motifs. In an effort to rule out crystal lattice artifacts as the cause of the increased order in the above primer-monomer-helper structures, we decided to try to obtain similar structures with a different packing of the RNA molecules in the crystal. The RNA hairpin structure we used as a model (PDB: 4FNJ) to design these new RNA sequences contained a tetraloop-receptor motif to facilitate crystallization. The tetraloop receptor sequence, essentially 3 G:U or U:U base pairs, generates an overall structure that is greatly different from the H-34-Im complex. We decided to use this structure to extend our view of ICG binding with different downstream oligonucleotides as helpers. An RNA 35mer complex was designed and constructed by assembling a new RNA primer-template oligonucleotide with a new downstream helper 12mer, and a sandwiched ICG monomer (**Figure 6.6**). In contrast to all of our previous RNA-monomer complexes, we used no locked nucleotides in the design of this complex. The complex contained a GAAA tetraloop, the RNA helical stem, and a tetraloop receptor motif to facilitate RNA molecular packing during crystallization (**Figure 6.6.A**). We solved two RNA-monomer complex structures with different downstream helpers: 5′-phosphate and 5′-OH RNA 12mers (unfortunately we have been unable to obtain crystals with the corresponding ICG-capped oligonucleotide). In both structures, the 35-nucleotide complexes crystallized in the space group R3, and the RNA hairpin structures formed as designed (referred to as H-35-PO and H-35-OH). The monomer-binding site was located at the upper stem (G23) to separate it from the tetraloop receptor (between C8:G28 and G12:U25) in the middle of the helix. Each RNA complex contacted four neighboring complexes: two were the designed tetraloop/receptor interactions, and the other two were stacking interactions

between the terminal base pair of the helices and the U27 nucleobase of the receptors from a symmetry related molecule (**Figures 6.6.B** and **6.6.C**).

As anticipated, the ICG monomer bound the template at the designed site on the upper stem so that it was not directly interacting with the nearby symmetry related complexes in the crystal. In both structures, the ICG monomer forms a Watson-Crick base pair with the native cytidine residue in the template (with H-bond distances ranging from 2.7 to 3.0 Å), and the guanine nucleobase of the ICG monomer is stacked with the upstream guanosine and downstream cytidine nucleotides (**Figures 6.6.D** and **6.6.E**). However, the aryl group of the ICG monomer is disordered in both H-35-PO and H-35-OH, and the phosphonate moiety is significantly more distant from the RNA primer in H-35-PO and H-35-OH than in the H-34-PO and H-34-OH complexes. In the H-35-OH structure, the distance between the 3′-OH and the phosphorus atom was 4.4 Å, while the distance between the 2′-OH and the phosphorus atom was 6.0 Å. Similar distances were observed in the H-35-PO structure (4.6 Å and 6.4 Å respectively). While the primer-monomer geometry is clearly less preorganized for nucleophilic attack than in the H-34-PO and H-34-OH complexes, possibly due to the lack of LNA nucleotides in the template, they are still better ordered than when the downstream element is a monomer instead of an oligonucleotide. These observations may help to explain why RNA primer extension is so slow without help from a downstream activated oligonucleotide.

To further explore the function of a downstream oligonucleotide in the catalysis of primer extension, we cocrystallized the above RNA structure with GpppG, an analog of the imidazolium-bridged G-G intermediate in primer extension.[19] RNA structures with different lengths of helical stems were screened, and a new 43-mer RNA complex was

constructed and crystallized. The RNA hairpin conformation, like H-35-PO and H-35-OH, contains a GAAA tetraloop, a receptor motif and a helical stem (**Figure 6.7.A**, hereafter termed H-43-GpppG). The H-43-GpppG complex crystallized in space group $P4_12_12$, with one complex per asymmetric unit. The GAAA tetraloop contacted the tetraloop receptor of a neighboring molecule, and the terminal base pair stacked with the terminal base pair of another neighboring complex. The GpppG dinucleotide stacked with the upstream and downstream cytidine nucleobases, and had no interactions with the surrounding complexes (**Figure 6.7.B**).

In the H-43-GpppG structure, the GpppG ligand pairs with the two consecutive templating cytidines C10 and C11, with hydrogen bond distances ranging between 2.8 to 3.1 Å (**Figure 6.7.C**). The two guanine nucleobases of GpppG are coplanar and stack with the upstream primer and downstream helper oligo. The triphosphate linkage of GpppG is well-ordered, but surprisingly, no bound Mg^{2+} ion was observed, in contrast to our previous structure of GpppG bound to a primer template complex, in which a well-defined Mg^{2+} ion was coordinated with three nonbridging oxygens. In the present structure, the conformation of the triphosphate bridge is different, in that the three nonbridging oxygens on the corresponding face of the triphosphate do not point to a common center, and are thus not well oriented for binding a single Mg^{2+} ion. This conformational difference likely explains the absence of a well-defined Mg^{2+} ion in the current structure. Remarkably, in the structure H-43-GpppG, the distance between the 3'-OH of the primer and the adjacent phosphorus atom of GpppG is 3.7 Å, and the angle between the 3'-OH and the bridging P-O bond of GpppG is 138.8°. These parameters are significantly improved relative to our previous observations of an RNA duplex-GpppG

complex with no downstream helper (4.1 Å and 126°).[19] This is consistent with the greater catalytic effect of a downstream helper oligonucleotide as opposed to a mononucleotide by enhancing conformational restraint. Additionally, the distance between the 2′-OH and the adjacent P atom of GpppG is 5.6 Å, and the corresponding angle is 126.6°, both of which would make the formation of a 2′-5′-linked product less favorable. Therefore, it appears that the presence of a downstream helper oligonucleotide would not only enhance the rate but also improve the regioselectivity of the primer extension reaction.

In order to further investigate the enhanced catalysis of primer extension by an activated downstream oligonucleotide, which is thought to act by forming an imidazolium-bridged intermediate with the incoming activated monomer, we crystallized an RNA complex in which the GpppG analogue is covalently attached to the downstream oligonucleotide. This RNA complex (**Figure 6.7.D**, hereafter termed H-34-GpppG-oligo) has a similar structure to H-34-Im, and is assembled from two RNA strands to form the hairpin structure. The GpppG-containing downstream RNA strand was *in vitro* transcribed using the same conditions as for the 5′-ICG capped octamer, except that ICG monomer was replaced by GpppG. The triphosphate linkage in the downstream oligonucleotide imitates the imidazolium-bridge formed between the activated monomer and activated downstream helper in an actual primer extension reaction.[19]

The H-34-GpppG-oligo complex crystallized in space group C2, with three complexes per asymmetric unit. Like the structure of H-34-Im, the hairpin has GAAA tetraloop and A-form double stranded helix, except for the last base pair in the stem. The GpppG dinucleotide at the 5′-end of the downstream helper is Watson-Crick base paired with the locked templating cytidines *C*8 and *C*9, and has no interactions with the

surrounding molecules (**Figure 6.7.E**). Like the structure of H-43-GpppG, the triphosphate linkage in H-34-GpppG-oligo is well-ordered with no coordinated metal ion (**Figure 6.7.F**). Strikingly, the distance between the 3'-OH of the primer and the adjacent phosphorus atom of GpppG is ~3.2 Å, which is the shortest distance we have observed in any of our complexes. In this structure the GpppG ligand is covalently linked to the rest of the downstream oligo, and the GpppG appears more preorganized for reaction with the primer 3'-OH. Meanwhile, the distance between the 2′-OH and the adjacent P atom of GpppG is ~4.8 Å, making 2′-5′ linkage formation unfavorable. Considering the similarity between GpppG and the imidazolium-bridged guanosine dinucleotide intermediate,[19] the present structure helps to explain our previous experimental observation of enhanced catalysis of primer extension by an activated downstream helper oligonucleotide vs. a downstream mononucleotide.[9]

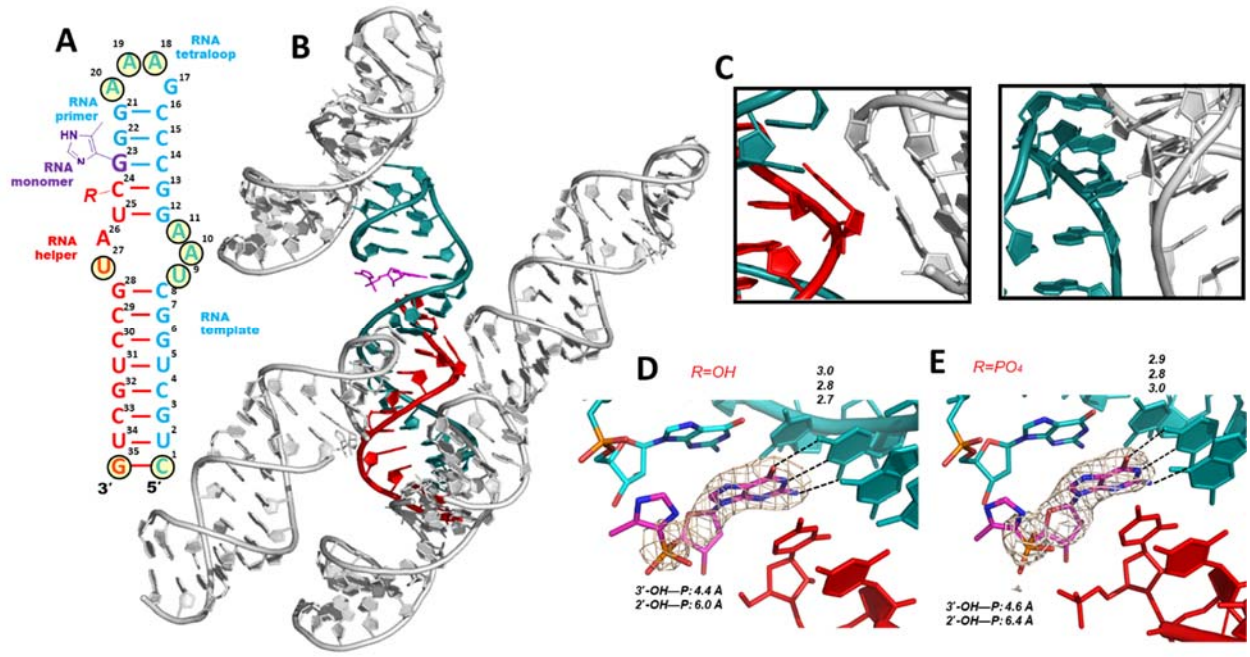


Figure 6.6. Structures of RNA hairpin-ICG H-35-OH and H-35-PO complexes. (A) Diagram and secondary structures of RNA H-35-OH and H-35-PO complexes. The nucleotides critical for crystallization are marked by circles. (B) Crystal structure of H-35-OH complex. Each RNA-monomer complex interacts with four other complexes (grey color): at the tetraloop, at the receptor, on the opposite side of the receptor, and at the bottom of the helix. (C) Two typical intermolecular interactions that are critical for molecular packing. (D) and (E) Closer views of the bound ICG monomers. The ICG monomers form Watson-Crick base pairs with the templating C in both H-35-OH and H-35-PO. Absence of electron density indicates that the 4-imidazolyl groups are disordered. Color-coding as in **Figure 6.4**. Hydrogen bonds are labeled as black dashes and their interatomic distances are indicated. Wheat mesh indicates the corresponding $2F_o–F_c$ maps contoured at 3.0 σ. The sequence of the oligonucleotide is shown and numbered.

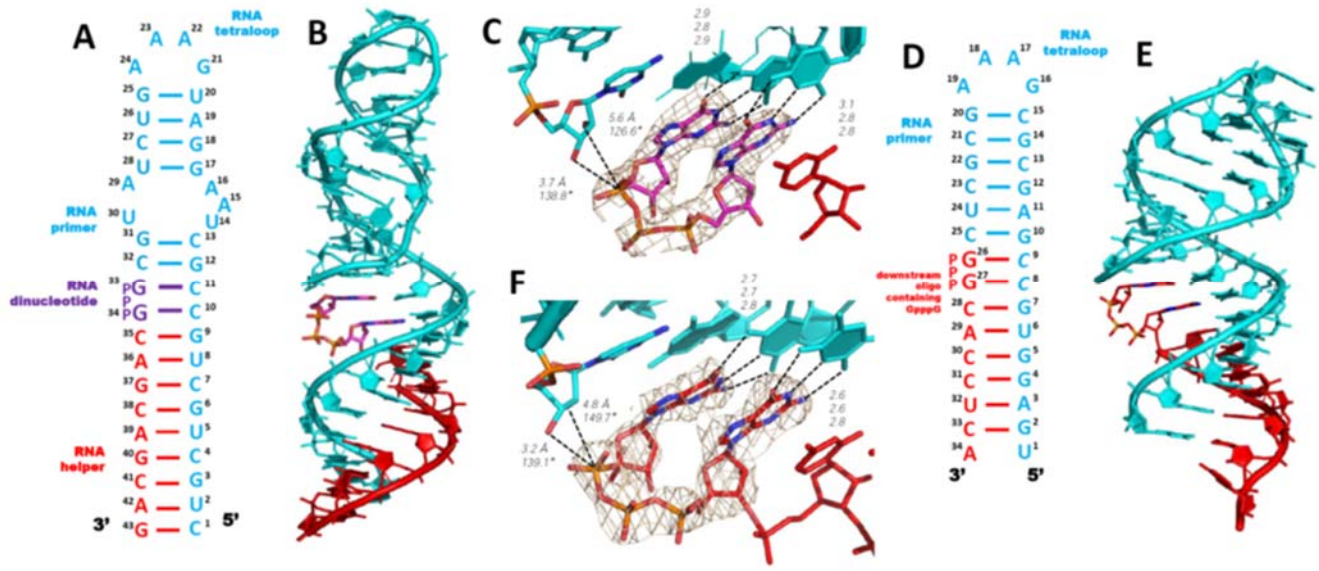


Figure 6.7. Structures of RNA H-43-GpppG and H-34-GpppG-oligo complexes. (A) Diagram and secondary structure of the H-43-GpppG complex. (B) Overall crystal structure of H-43-GpppG. (C) Closer view of the bound GpppG molecule. (D) Diagram and secondary structure of the H-34-GpppG-oligo complex. (E) Overall crystal structure of H-34-GpppG-oligo. (F) Closer view of the GpppG dinucleotide that begins downstream oligo. Color-coding as in **Figure 6.4**. Hydrogen bonds are labeled as black dashes and their interatomic distances are indicated. Wheat mesh indicates the corresponding $2F_o–F_c$ maps contoured at 2.0 σ. The sequence of the oligonucleotide is shown and numbered.

6.2.5 Ligation-like structures containing helper oligomers. In order to understand the extent to which structural phenomena associated with primer extension by monomers are relevant to nonenzymatic ligation reactions, we constructed two additional RNA hairpin structures, hereafter termed H-34-L and H-35-L. These constructs mimic intermediates in the ligation of an activated oligomer to an adjacent oligomer on the same template; there are no binding sites for ICG monomers. Similar to H-34-GpppG-oligo, in H-34-L, a 5′-imidazolyl-phosphonate-RNA is paired completely with the template through Watson-Crick base pairs and forms an A-form double stranded helix, except for the last base pair in the stem. In this structure, the distance between the 3′-hydroxyl of the primer and the

phosphorus atom of the phosphate is 4.4 Å, and the 4-methylimidazolyl group is highly disordered (**Figures 6.8.A** and **6.8.B**). In H-35-L, a 5′-phosphorylated 13-mer RNA and the primer-template hairpin assembled to form a designed hairpin structure containing the tetraloop/receptor motif. In this case, the distance between the 3′-hydroxyl of the primer and the phosphorus atom of the phosphate was 4.1 Å (**Figures 6.8.C** and **6.8.D**). Interestingly, all these distances are too long for the efficient backbone ligation. These observations, together with our previous structure of an RNA-pGpG complex,[19] provide a structural basis for our previous experimental finding that RNA nonenzymatic ligation is much slower than RNA nonenzymatic polymerization.[9] Although in the case of nonenzymatic ligation, the primer and ligating oligomer are expected to tightly bind to the template in the preferred A-form conformation,[25] there is no downstream activating group to supply either covalent or noncovalent interactions that could help to properly orient the activating group on the ligator oligonucleotide.

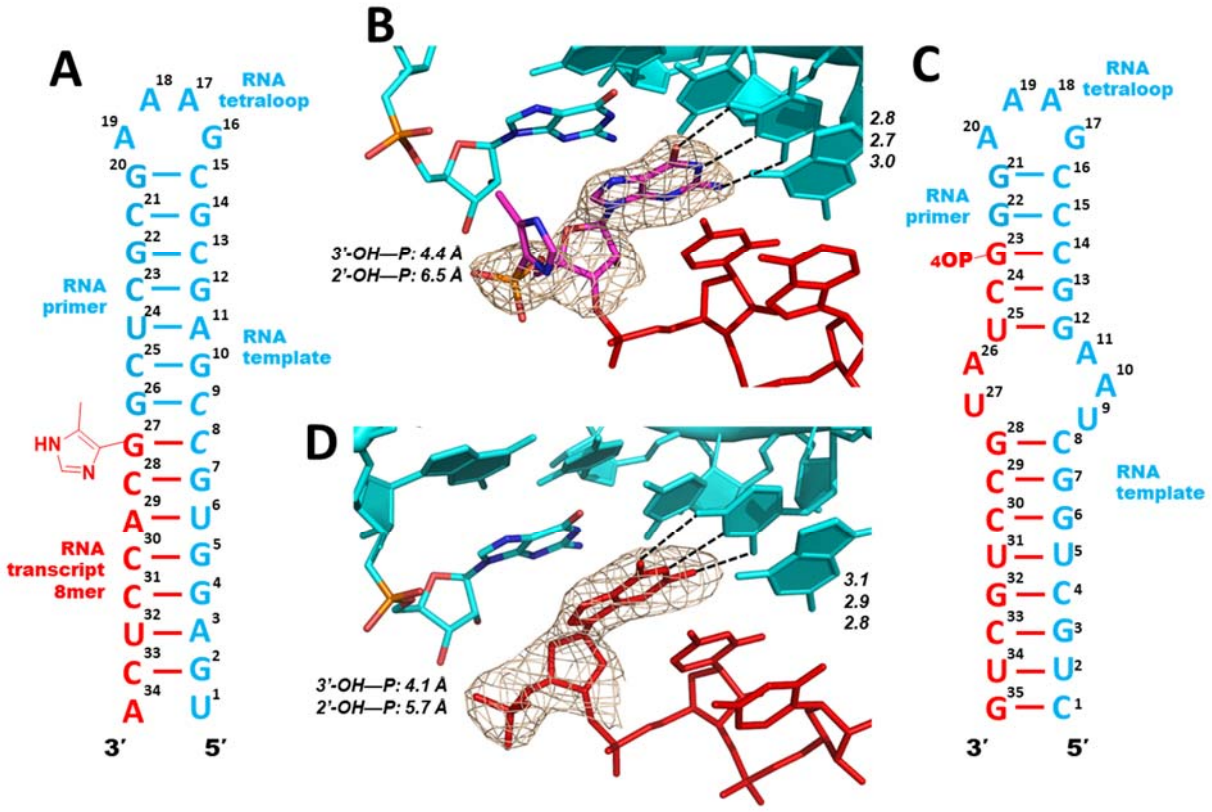


Figure 6.8. Structures of ligation-like RNA complexes H-34-L and H-35-L. (A) Diagram and secondary structure of RNA H-34-L. (B) Closer view of the 5'-ICG nucleotide of the RNA transcript, which Watson-Crick pairs with the template. Absence of electron density indicates that the 4-imidazolyl group is disordered. (C) Diagram and secondary structure of RNA H-35-L. (D) Closer view of the first G nucleotide of the downstream 13mer RNA helper. Color-coding as in **Figure 6.4**. Wheat mesh indicates the corresponding $2F_o–F_c$ maps contoured at 2.0 σ. The sequence of the oligonucleotide is shown and numbered.

6.3 Discussion

We have used crystallographic studies of RNA-ICG and RNA-GpppG complexes to investigate the structural basis for the catalysis of nonenzymatic RNA primer extension by oligonucleotides that are bound downstream of the reactive monomer. Previous studies have shown that the major catalytic effect is covalent nucleophilic catalysis, involving the formation of a covalent intermediate between the reactive monomer and an activated downstream monomer or oligonucleotide. However, other studies, from both

the Richert lab[7,26] and our lab,[9] show that a downstream oligonucleotide can provide an additional catalytic effect, which is independent of whether the downstream oligonucleotide is activated or not. Our key finding is that, in addition to the effect of enhanced affinity of bound monomer for the template, a downstream oligonucleotide (and to a lesser extent even a mononucleotide) contributes to a favorable geometry in the reaction center by inducing a closer approach of the primer 3′-hydroxyl to the phosphorous atom of the incoming monomer.

To gain further insight into the nature of primer-template-monomer complexes in the absence of a downstream oligonucleotide, we determined two new structures, in which either one or two pseudo-activated G monomers are bound to the template strand, adjacent to and downstream of the primer strand. ICG monomers closely mimic the structure of a 2-methylimidazole activated G nucleotide, but because the phosphonate linkage is stable, a single ICG monomer cannot react with the primer, and two ICG monomers cannot react with each other to form the normal imidazolium-bridged G-G dinucleotide intermediate in primer extension. Crystal structures of RNA with bound ICG residues should therefore mimic the prereaction ground state. When a single ICG monomer was bound to a primer-template complex, the ICG sugar and aryl-phosphonate were disordered, and the phosphate of the ICG monomer was quite far (6 Å) from the 3′-hydroxyl of primer. Assuming that this structure mimics that of an actual 2-MeImpG activated monomer bound to a primer-template complex, the extremely slow rate of primer extension with a single activated monomer becomes understandable. When two ICG monomers were bound to the template, the monomer adjacent to the primer was more ordered and the distance between the primer 3′-hydroxyl and the adjacent P atom

was shortened (4.6 Å), even though the imidazolyl group could not be seen in the electron density. This structure suggests that even a single downstream activated monomer would help to preorganize the complex so as to favor reaction between the primer and the adjacent monomer, in a manner that is independent of formation of the imidazolium-bridged reaction intermediate, and in a way that is also independent of noncovalent interactions between the leaving groups of adjacent monomers.

To investigate the effect of a downstream oligonucleotide, in contrast to a downstream mononucleotide, we solved a set of structures with the same primer, ICG monomer, and template, but with downstream oligonucleotides bearing either a 5′-ICG cap, a 5′-phosphate, or a 5′-hydroxyl group. Only Watson-Crick base pairing of the ICG monomer to the template was observed, in contrast to the noncanonical base pairing motifs observed at one end of both RNA duplex-ICG structures. Remarkably, in all three structures with a downstream oligonucleotide and locked cytidine as template, the bound ICG monomer became much more ordered and the distance between the phosphonate group of the ICG monomer and the 3′-hydroxyl group of the primer decreased to less than 4 Å. Even in the two structures containing native cytidine in the template part of the hairpin stem, the sandwiched ICG monomer also became more ordered in the presence of a downstream helper, although the distance between the primer 3′-OH and the phosphonate group of the ICG monomer remained unchanged. Our structures suggest that the rate enhancement afforded by the downstream oligonucleotide results from additional conformational constraints that bring the primer 3′-hydroxyl closer to the phosphate of the adjacent nucleotide. Could such constraints also explain the catalytic effect of a downstream oligonucleotide on reactions involving an imidazolium bridged

intermediate, e.g. the much faster primer extension observed with an activated downstream trimer vs. an activated downstream monomer? The reactivity of the imidazolium-bridged intermediate makes this a difficult question to address directly through crystallography. However, we have recently solved the structure of GpppG bound to an RNA primer-template complex, and showed that it appears to be a reasonable mimic of the imidazolium-bridged intermediate.[19] Here we again made use of template bound GpppG to ask whether a downstream oligonucleotide would further preorganize the complex so as to favor the geometry that would enhance reaction with the true intermediate. Compared with our previous structural observations, in which GpppG was bound to the CC overhang at the end of a duplex, without any downstream helper oligonucleotide, the presence of the downstream oligonucleotide appears to push the dinucleotide closer to the primer. The distance from the primer 3′-OH to the adjacent phosphorus atom of GpppG is shortened from 4.1 Å to 3.7 Å, and the corresponding O-P-O angle increases from 126° to 139°. Interestingly, when the GpppG-containing downstream oligo was used to mimic the intermediate formed by activated monomer and activated downstream oligomer, the distance for primer extension is further shortened to 3.2 Å. This phenomenon is consistent with our previous observation that nonenzymatic primer extension has a much faster rate with the help of an activated downstream oligonucleotide, rather than unactivated downstream oligonucleotide or activated downstream monomer. Assuming a similar effect holds in the case of the true reaction intermediate it is easy to see how the conformational constraint induced by a downstream activated or unactivated oligonucleotide would further enhance the rate of reaction between the primer and the intermediate. The observed conformations in both structures

would also disfavor formation of the incorrect 2′-5′ linkage during primer extension due to the longer O2′-P distances than O3′-P distances. Our observations are consistent with our recent finding that primer extension through the 5′-5′ imidazolium bridged intermediate strongly favors the correct 3′-5′ linkages.[27]

A consistent observation across all structures is the absence of detectable noncovalent interactions, such as π-π, cation-π or hydrogen bonding interactions, between the leaving group of the incoming monomer and that of a downstream monomer or oligomer. It is known that primer extension with an activated monomer is catalyzed by a downstream activated monomer or oligonucleotide in a manner that is distinct from the rate enhancement due to an unactivated downstream oligonucleotide. We have found that a pseudo-activated downstream helper oligonucleotide (5'-ICG) did not engage in any detectable interactions that could explain the catalytic effect of activated vs. unactivated helpers. It has been hypothesized that noncovalent leaving group-leaving group interactions could contribute to the observed catalysis, but our structures provide no evidence in support of this hypothesis.

In conclusion, using the novel nonhydrolyzable analog of activated guanosine, ICG, and an analog of the imidazolium-bridged intermediate, GpppG, we have presented structural evidence that downstream base-paired elements (monomers or oligomers) induce conformational constraints in a variety of constructs that model nonenzymatic primer extension. The observed conformational effects provide a potential explanation for the rate enhancement conferred by both unactivated and activated downstream oligonucleotides.

6.4 References

(1) Crick, F. H. C. *J. Mol. Biol.* **1968**, *38*, 367.

(2) Gilbert, W. *Nature* **1986**, *319*, 618.

(3) Szostak, J. W. *J. Syst. Chem.* **2012**, *3*, 2.

(4) Robertson, M. P.; Joyce, G. F. *Cold Spring Harb. Perspect. Biol.* **2012**, *4*, a003608.

(5) Weimann, B. J.; Lohrmann, R.; Orgel, L. E.; Schneider-Bernloehr, H.; Sulston, J. E. *Science* **1968**, *161*, 387.

(6) Inoue, T.; Joyce, G. F.; Grzeskowiak, K.; Orgel, L. E.; Brown, J. M.; Reese, C. B. *J. Mol. Biol.* **1984**, *178*, 669.

(7) Vogel, S. R.; Deck, C.; Richert, C. *Chem. Commun.* **2005**, *39*, 4922.

(8) Li, L.; Prywes, N.; Tam, C. P.; O'Flaherty, D. K.; Lelyveld, V. S.; Izgu, E. C.; Pal, A.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 1810.

(9) Prywes, N.; Blain, J. C.; Del Frate, F.; Szostak, J. W. *Elife* **2016**, *5*, e17756.

(10) Izgu, E. C.; Fahrenbach, A. C.; Zhang, N.; Li, L.; Zhang, W.; Larsen, A. T.; Blain, J. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2015**, *137*, 6373.

(11) Tam, C. P.; Fahrenbach, A. C.; Björkbom, A.; Prywes, N.; Izgu, E. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 571.

(12) Kervio, E.; Claasen, B.; Steiner, U. E.; Richert, C. *Nucleic Acids Res.* **2014**, *42*, 7409.

(13) Kanavarioti, A.; Bernasconi, C. F.; Alberas, D. J.; Baird, E. E. *J. Am. Chem. Soc.* **1993**, *115*, 8537.

(14) Orgel, L. E. *Crit. Rev. Biochem. Mol. Biol.* **2004**, *39*, 99.

(15) Wu, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1992**, *114*, 7963.

(16) Zhang, W.; Tam, C. P.; Wang, J.; Szostak, J. W. *ACS Cent. Sci.* **2016**, *2*, 916.

(17) Kervio, E.; Sosson, M.; Richert, C. *Nucleic Acids Res.* **2016**, *44*, 5504.

(18) Walton, T.; Szostak, J. W. *J. Am. Chem. Soc.* **2016**, *138*, 11996.

(19) Zhang, W.; Tam, C. P.; Walton, T.; Fahrenbach, A. C.; Birrane, G.; Szos-tak, J. W. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 7659.

(20) Kanavarioti, A.; Rosenbach, M. T. *J. Org. Chem.* **1991**, *56*, 1513.

(21) Tam, C. P.; Zhou, L.; Fahrenbach, A. C.; Zhang, W.; Walton, T.; Szostak, J. W. *J. Am. Chem. Soc.* **2018**, *140*, 783.

(22) Fox, J. J.; Van Praag, D.; Wempen, I.; Doerr, I. L.; Cheong, L.; Knoll, J. E.; Eidinoff, M. L.; Bendich, A.; Brown, G. B. *J. Am. Chem. Soc.* **1959**, *81*, 178.

(23) Coonrod, L. A.; Lohman, J. R.; Berglund, J. A. *Biochemistry* **2012**, *51*, 8330.

(24) Li, L.; Lelyveld, V. S.; Prywes, N.; Szostak, J. W. *J. Am. Chem. Soc.* **2016**, *138*, 3986.

(25) Kozlov, I. A.; Orgel, L. E. *Mol. Biol.* **2000**, *34*, 781.

(26) Hagenbuch, P.; Kervio, E.; Hochgesand, A.; Plutowski, U.; Richert, C. *Angew. Chem. Int. Ed.* **2005**, *44*, 6588.

(27) Giurgiu, C.; Li, L.; O'Flaherty, D. K.; Tam, C. P.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 16741.

Appendix A

General Methods

A.1. Small Molecule Synthesis: General Considerations

Unless otherwise noted, all reactions were conducted in oven-dried round-bottomed flasks that were fitted with appropriately-sized rubber septa, under a positive pressure of dry argon. Aldrich SureSeal™ anhydrous solvents were used for all reactions. Reaction progress was monitored either by means of thin-layer chromatography (TLC) or low-resolution electrospray ionization ion-trap mass spectrometry, on a Bruker Daltonics Esquire 6000 mass spectrometer (Billerica, MA). TLC plates were visualized by 254 nm ultraviolet light, alkaline potassium permanganate stains, or iodine/sand mixture. Normal- and reverse-phase flash chromatography was carried out on a Teledyne Isco CombiFlash Rf system (Lincoln, NE) on pre-packed silica or C18Aq (carbon-18 aqueous) columns.

A.2. Small Molecule Synthesis: NMR Structural Characterization

NMR spectra were recorded on a Varian Inova 400 MHz spectrometer equipped with a broadband PFG (z-gradient) probe (400 MHz for ^{1}H, 100 MHz for ^{13}C, 161 MHz for ^{31}P; Santa Clara, CA). Proton, carbon and phosphorus chemical shifts were reported in parts per million (ppm) values on the δ scale, and referenced to residual protium in the NMR solvents (Proton NMR: $CHCl_3$, δ = 7.26 ppm; DHO, δ = 4.79 ppm; CHD_2OD, δ = 1.94 ppm; Carbon NMR: $CDCl_3$, δ = 77.16 ppm; CD_3OD, δ = 49.0 ppm, DMSO-d_6: δ = 39.52 ppm),[1] while proton-decoupled phosphorus chemical shifts were internally referenced to trimethyl phosphate ($CDCl_3$: δ = 3.0 ppm; D_2O: δ = 3.8 ppm; CD_3OD: δ = 3.4 ppm; DMSO-d_6: δ = 3.5 ppm).[2] All NMR spectra were recorded at 25 °C. Data were

reported as follows: chemical shift, multiplicity (s = singlet, d = doublet, q = quartet, m = multiplet, br = broad), and integration. NMR spectra were processed by MestReNova 11 (Mestrelab Research, Santiago de Compostela, Spain).

A.3. Solid-Phase RNA Synthesis, Deprotection, and Purification

RNA oligonucleotides were synthesized by standard solid-phase phosphoramidite polymerization chemistry on a MerMade 6 RNA/DNA oligonucleotide synthesizer. Cleavage and elution of protected oligonucleotide products from 50 μmol universal CPG-solid support columns were performed by equilibrating and eluting the solid support material with a 1:1 mixture of ammonium hydroxide and 40% aqueous methylamine for a total of 3 times (equilibration time: 3 x 10 min; elution volume: 3 x 5 mL for 50 μmol columns). Deprotections of nucleobases and phosphates were carried out by heating the strongly alkaline eluent for 2.5 h at 65 °C; the resultant clear (or off-white) homogeneous mixtures were first evaporated *in vacuo* for 3 h at 40 °C on a Genevac EZ-2 tabletop speedvac system (Genevac, Stone Ridge, NY), then lyophilized to dryness on a VirTis Sentry 2.0 freeze-drier (SP Scientific, Warminster, PA) at <50 mTorr overnight to afford off-white solid residues. The residues were then resuspended in 2.5 mL of dimethyl sulfoxide (DMSO) and 2.5 mL of trimethylamine trihydrofluoride (TEA•3HF; *CAUTION: TOXIC*), and heated for 2.5 h at 65 °C to remove the TBDMS protecting group on the ribose 2'-hydroxyl groups. The mixtures appeared homogeneous and with pale to golden yellow color. After cooling to room temperature (~30 min), mixtures of 625 μL of 3 M sodium acetate and 15 mL of 1-butanol were used for RNA precipitation. The precipitates were spun down (4000 rpm, 5 min) and supernatants were removed by decanting. The resulting white solids were washed twice with absolute ethanol; the samples were then

dried under high vacuum overnight. Purification of the desired products was carried out by preparative-scale HPLC on a Varian Prostar 210 HPLC system equipped with Agilent ZORBAX Eclipse-XDB C18 column (Agilent Technologies, Santa Clara, CA), using 25 mM triethylammonium bicarbonate in H_2O (TEAB buffer, pH 7.5) with an increasing gradient of 0 to 15 % acetonitrile over 30 min (flow rate: 10 mL min^{-1}). Elution of RNA was monitored by UV absorption at 254 and 280 nm. The desired RNA fractions were pooled and lyophilized to afford a fluffy white powder. The resultant white residues were again purified by preparative-scale strong anion-exchange HPLC, using the ThermoFisher-Dionex DNAPac PA100 strong anion exchange column (Waltham, MA), with an increasing gradient of 0 to 100 mM aqueous sodium perchlorate solution over 30 min with a flow rate of 10 mL min^{-1}. The desired RNA fractions were collected, pooled and lyophilized to afford white solid residues. The residues were washed with acetone for 3 times, followed by overnight drying under high vacuum, to afford the desired products (in sodium cation form) as fine white grains.

A.4. Characterization of Small Molecules and Oligonucleotides with High Resolution Liquid Chromatography - Mass Spectrometry (HR-LC-MS)

The monoisotopic mass (*m/z*) and purity for all synthesized RNA oligomers and small molecules were assessed using an Agilent 1200 high-performance liquid chromatography system coupled to an Agilent 6220 time-of-flight mass spectrometer (Santa Clara, CA). The LC-MS setup was equipped with a solvent degasser, temperature controlled auto sampler, column oven, diode-array detector, and a dual electrospray ionization source. For all analyses, 100–200 pmole of samples were analyzed with

extended dynamic range using these settings: column temperature, 50 °C; scan rate, 1 spectrum s^{-1}; mass range, 239 m/z – 3200 m/z; drying gas flow, 8 L min^{-1}; drying gas temperature, 325 °C; nebulizer pressure, 30 psig; capillary voltage, –3500 V; fragmentor, 200 V; and skimmer, 65 V

A.4.1. Negative ion mode analysis

- Column: Waters 100 mm XBridge C18 column, 1 mm i.d., 3.5 µm particle size (Waters Corporation, Milford, MA).
- Solvent Conditions: Solvent A was water with 200 mM HFIP, and 1.25 mM TEAA at pH 7.0. Solvent B was methanol, and was ramped from 2.5% to 20% over 30 min at a flowrate of 0.1 mL/min.[3]

A.4.2. Positive ion mode analysis

- Column: Agilent ZORBAX 300 Rapid Resolution HD SP-C18 column, 50 mm × 2.1 mm, 1.8 µm particle size.
- Solvent Conditions: Solvent A was 0.1% formic acid in water. Solvent B was 0.1% formic acid in acetonitrile, and was ramped from 10% to 100% over 30 min at a flow rate of 0.2 mL/min.

Data analysis was performed using Agilent MassHunter Qualitative Analysis software.

A.5. Determination of Mono- and Oligonucleotide Concentrations

Concentrations of the aqueous mono- and oligonucleotide samples were determined by their UV absorption at 260 nm on a Thermo Scientific Nanodrop 2000c spectrophotometer (Waltham, MA). The theoretical molar extinction coefficients of oligonucleotides at 260 nm were calculated with the OligoAnalyzer software provided by Integrated DNA Technologies (version 3.1, Coralville, IA).[4,5] The extinction coefficients of cytidine

monomers were assumed to be 7.07 L $mmol^{-1}$ cm^{-1}, while that of guanosine monomers and their nonhydrolyzable analogues were assumed to be 12.08 L $mmol^{-1}$ cm^{-1}.

A.6. References

(1) Gottlieb, H. E.; Kotlyar, V.; Nudelman, A. *J. Org. Chem.* **1997**, *62*, 7512–7515.

(2) Streck, R.; Barnes, A. J. *Spectrochim. Acta A* **1999**, *55*, 1049–1057.

(3) Apffel, A.; Chakel, J. A.; Fischer, S.; Lichtenwalter, K.; Hancock, W. S. *Anal. Chem.* **1997**, *69*, 1320–1325.

(4) Cantor, C. R.; Warshaw, M. M.; Shapiro, H. *Biopolymers* **1970**, *9*, 1059–1077.

(5) Cavaluzzi, M. J.; Borer, P. N. *Nucleic Acids Res.* **2004**, *32*, e13.

Appendix B

Supplemental Materials for Chapter 2

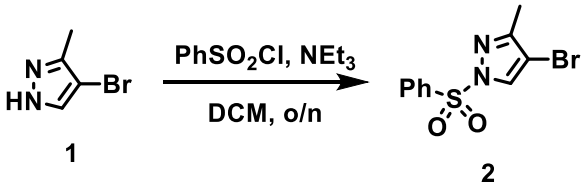
B.1. Locked nucleic acid (LNA)-modified RNA oligonucleotides.

The four LNA-modified RNA oligonucleotides (bolded letters) used for crystallographic studies (5′-**mCmCmCG**ACUUAAGUCGG-3′, 5′-**mCTG**UACA-3′, 5′-**mCmCAG**ACUUAAGUCU-3′, and 5′-**mCmCmCG**ACUUAAGUC-3′) were custom-synthesized by Exiqon Inc. (Woburn, MA), with their 5´-dimethoxytrityl (DMT) groups cleaved and samples preliminarily purified by desalting. Reverse-phase high performance liquid chromatography (HPLC)-based purification of these desalted oligonucleotides was carried out on a Varian Prostar 210 HPLC system equipped with an Agilent ZORBAX Eclipse-XDB C18 column (preparative-scale column, 21.2×250mm, 7 μm particle size; Santa Clara, CA). 25 mM triethylammonium bicarbonate in H_2O (pH 7.5; water was pre-purified by EMD Millipore's Milli-Q® water purification system, Merck KGaA, Darmstadt, Germany) was used as the primary eluent with an increasing gradient of 0% to 30% acetonitrile (HPLC grade, Fisher Scientific, Waltham, MA) over 30 mins. Elution of RNA was monitored by UV absorption at wavelengths of 254 and 280 nm. The bolded 5´-terminal regions of the oligonucleotides were modified by locked nucleic acids to rigidify the template;[1] additionally, 5-methylcytidine and thymidine replaced cytidine and uridine in the modified region of the RNA sequences. Relevant fractions were pooled and lyophilized on a VirTis Sentry 2.0 freeze-drier (SP Scientific, Warminster, PA) for ≥ 2 days at < 50 mTorr to afford the purified oligonucleotides in their triethylammonium salt form, followed by neutralization and NaCl/ethanol precipitation to furnish the corresponding sodium salts of the purified oligonucleotides. Concentrations of the aqueous RNA

samples were determined by their UV absorption at 260 nm on a Thermo Scientific Nanodrop 2000c Spectrophotometer (Waltham, MA). The theoretical molar extinction coefficients of all the RNA strands used herein at 260 nm were provided by Exiqon.

B.2. Synthesis of PZG: Experimental procedures and compound characterizations.

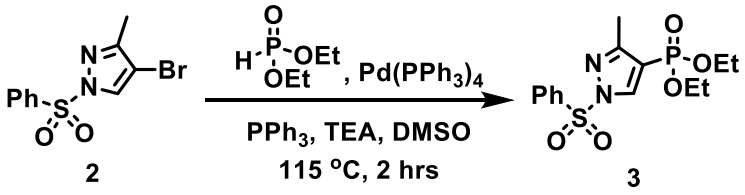
B.2.1. 4-bromo-3-methyl-1-phenylsulfonyl-1*H*-pyrazole (Compound 2)[2]



4-Bromo-3-methyl-1*H*-pyrazole (compound **1**, 5 g, 31.1 mmol) was added to an oven-dried round bottom flask and stir bar, followed by dissolution in anhydrous CH_2Cl_2 (62 mL, 0.5 M) and anhydrous triethylamine (9 mL, 62.1 mmol, 2 equiv., ρ = 0.726 g/mL) at ambient temperature. Afterwards, to that stirring solution, benzenesulfonyl chloride (5.94 mL, 46.6 mmol, 1.5 equiv., ρ = 1.384 g/mL) was added dropwise. The reaction was allowed to stir overnight (o/n) at ambient temperature. The reaction was diluted 5-fold with CH_2Cl_2 and was then washed with saturated aqueous ammonium chloride, water and brine. The organic phase was dried over sodium sulfate and evaporated *in vacuo*. The resulting crude residue was purified by normal-phase flash chromatography on a 120 g pre-packed silica column over 10 column volume (CV) of pure CH_2Cl_2 to afford the titled compound (8.70 g, 28.9 mmol, 93 %.) as a fluffy white solid.

^{1}H NMR ($CDCl_3$)	δ 7.95 (s, 1H), δ 7.89 (d, *J* = 8 Hz, 2H), δ 7.56 – 7.50 (m, 1H), δ 7.45 – 7.41 (m, 2H), δ 2.10 (s, 3H)
^{13}C NMR ($CDCl_3$)	δ 154.2, 136.6, 134.7, 131.0, 129.4, 127.9, 99.5, and 12.3
TLC (UV; $KMnO_4$)	R_f: 0.57 (20% EtOAc in Hexanes)
HRMS	Calc'd for $C_{10}H_{10}BrN_2O_2S$ $[M+H]^+$: 300.9646; Found: 300.9638

B.2.2. Diethyl (3-methyl-1-phenylsulfonyl-1*H*-pyrazol-4-yl)phosphonate (Compound 3)[3]

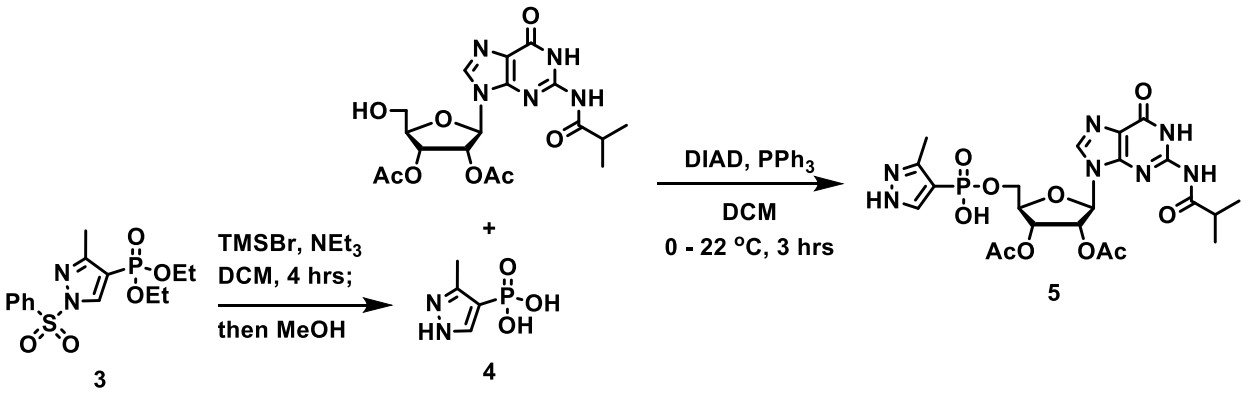


To a rigorously flame-dried round bottom flask and stir bar was added compound **2** (1.50 g, 4.98 mmol), fresh tetrakis(triphenylphosphine)palladium(0) (863 mg, 0.747 mmol, 0.15 equiv.; Strem Chemicals, Newburyport, MA), and triphenylphosphine (2.00 g, 7.47 mmol, 1.5 equiv.). The atmosphere within was evacuated to high vacuum and exchanged with dry argon at least three times. In a separate rigorously flame-dried pear-shaped flask, anhydrous dimethyl sulfoxide (49.8 mL, 0.1 M), diethyl phosphite (1.0 mL, 7.5 mmol, 1.5 equiv., ρ = 1.072 g/mL), and anhydrous triethylamine (1.4 mL, 10 mmol, 2 equiv., ρ = 0.726 g/mL) were mixed and purged by a stream of dry argon for 30 mins. The argon-purged reagent-solvent mixture was then cannula-transferred to the round bottom flask containing the starting material and the palladium catalyst, followed by 2 hrs of heating in a silicone oil bath pre-equilibrated to 115 °C. The palladium catalyst slowly dissolved in the reaction mixture upon heating, and the reaction mixture gradually turned golden orange from the initial golden yellow color. After 2 hours, the reaction was allowed to cool to ambient temperature, and diluted with 250 mL of ethyl acetate. The organic layer was

washed with water and brine, dried over sodium sulfate and evaporated *in vacuo*. The resulting crude residue was purified by normal-phase flash chromatography on an 80 g pre-packed silica column over 15 CV of solvents, with a linearly increasing gradient of 0–10% MeOH in CH_2Cl_2. The title compound (1.13 g, 3.15 mmol, 63 %.) was eluted from the column at ~6% MeOH in DCM as an off-white syrup.

^{1}H NMR ($CDCl_3$)	δ 8.30 (d, *J* = 2.9 Hz, 1H), δ 7.91 (d, *J* = 8.6 Hz, 2H), δ 7.60 – 7.56 (m, 1H), δ 7.48 – 7.44 (m, 2H), δ 4.08 – 3.92 (m, 4H), δ 2.24 (s, 3H), δ 1.20 (t, *J* = 7.1 Hz, 6H)
^{13}C NMR ($CDCl_3$)	δ 156.1 (d, *J* = 21 Hz, 1C), 138.0 (d, *J* = 39 Hz, 1C), 136.2, 134.9, 129.4, 128.2, 111.0 (d, *J* = 344 Hz, 1C), δ 62.2 (d, *J* = 9 Hz, 1C), 16.1 (d, *J* = 10 Hz, 1C), and 13.4
^{31}P NMR ($CDCl_3$)	δ 11.94 (s)
TLC (UV; $KMnO_4$)	R_f : 0.55 (15% Acetone in CH_2Cl_2)
HRMS (*m/z*)	Calc'd for $C_{14}H_{20}N_2O_5PS$ $[M+H]^+$: 359.0831; Found: 359.0829

B.2.3. N^2-isobutyryl-2′,3′-diacetylguanosine (3-methyl-1*H*-pyrazol-4-yl)phosphonate (Compound 5)



<u>*Step 1.*</u>[4] A round bottom flask containing compound **3** (513.2 mg, 1.43 mmol) and a stir bar was thoroughly dried under high vacuum overnight prior to initiating this reaction. Afterwards, anhydrous CH_2Cl_2 (19 mL, 0.075 M) and triethylamine (0.5 mL, 3.58 mmol,

2.5 equiv., ρ = 0.726 g/mL) was added to the reaction vessel to dissolve **1.2** into a pale yellow homogeneous solution. Bromotrimethylsilane (0.94 mL, 7.15 mmol, 5 equiv., ρ = 1.16 g/mL) was then added to the stirring mixture in a dropwise fashion under ambient temperature, and the reaction mixture was allowed to stir under room temperature for 4 hrs; a dense white smoke was observed when bromotrimethylsilane was added. The reaction gradually changed color, from the initial pale golden yellow to a deep yet pellucid brown color. To quench the excess bromotrimethylsilane and to decompose the resultant bis(trimethylsilyl)phosphonate esters,[5,6] an excess of methanol was added to the reaction mixture, and the overall pH of the reaction promptly dropped to < 1 (crudely determined by pH paper). Solvents were evaporated *in vacuo*, and the residual volatile impurities and hydrogen bromide were azeotropically removed by co-evaporation with toluene, repeated three times. The resultant orange-red solid residue (compound **4**) was carried forward to step 2 without further purification.

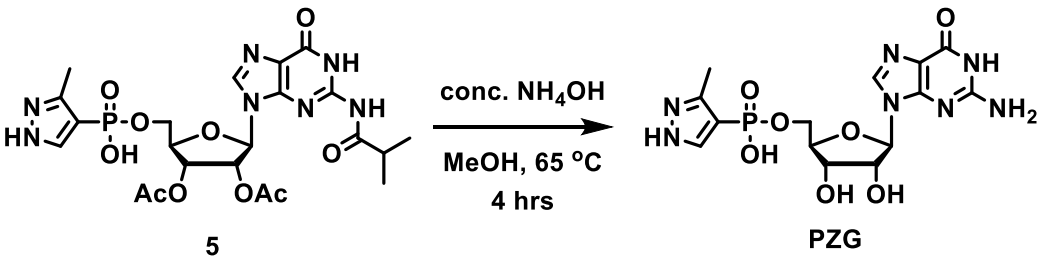
Step 2.[7] The crude material (compound **4**, 100% conversion assumed; 2.5 equiv.) obtained from step 1 was dried *in vacuo* under high vacuum overnight prior to initiating this reaction. Afterwards, the round bottom flask containing **4** was back-filled with dry argon, followed by sequential addition of N^2-isobutyryl-2′,3′-diacetylguanosine (250 mg, 0.572 mmol, 1 equiv.; obtained from ChemGenes, Wilmington, MA), triphenylphosphine (1.87 g, 7.15 mmol, 12.5 equiv.) and anhydrous CH_2Cl_2 (30 mL, ~0.05 M) under an argon atmosphere. The reaction mixture was cloudy and reddish brown. Using an ice bath, the reaction was allowed to equilibrate to 0 ^oC for 5 minutes, after which diisopropyl azodicarboxylate (DIAD; 1.45 mL, 7.15 mmol, 12.5 equiv., ρ = 1.027 g/mL) was added dropwise to the reaction mixture over a 10-minute interval. The reaction mixture was

observed to clarify into a pellucid brown mixture as DIAD was added. Afterwards, the reaction was removed from the ice bath and allowed to stir for 3 hours under ambient temperature. Subsequently, the reaction was stopped by evaporating the solvent *in vacuo* to form a light yellowish solid residue. To purify the crude product with reverse-phase flash chromatography, the crude residue was dissolved in a 3:1 mixture of dimethyl sulfoxide and water, followed by direct liquid-loading into a manufacturer-packed 50 g C18Aq column pre-equilibrated with 25 mM pH 7.5 aqueous triethylammonium bicarbonate (TEAB) buffer. The resulting crude residue was purified over 15 CV of solvents, with a linearly increasing gradient of 0–70% CH_3CN in TEAB. The title compound was eluted from the column at ~25% CH_3CN in TEAB, followed by lyophilization to afford the title compound as a white fluffy solid. Yield was not determined at this step due to the large excess of triethylammonium bicarbonate in the sample.

*Note: To afford pure compound **5**, the eluted purified product should be promptly lyophilized. The acetyl protecting groups on the 2'- and 3'-hydroxyl groups were found to be very labile under basic aqueous conditions.*

^{1}H NMR (CD_3OD)	δ 8.16 (s, 1H), δ 7.68 (s, 1H), δ 6.15 – 6.09 (m, 2H), δ 5.60 (dd, *J* = 5, 2.4 Hz, 1H), δ 4.33 (d, *J* = 1.9 Hz, 1H), δ 4.12 – 4.07 (m, 1H), δ 3.99 – 3.94 (m, 1H), δ 2.98 (q, *J* = 7.2 Hz, **TEAH$^+$**, 6H), δ 2.86 (septet, *J* = 6.8 Hz, 1H), δ 2.39 (s, 3H), δ 2.12 (s, 3H), δ 1.99 (s, 3H), δ 1.23-1.20 (t, *J* = 7.2 Hz, **TEAH$^+$**, 9H), δ 1.21 (d, 6H, overlapped with **TEAH$^+$** proton peaks)
^{13}C NMR (CD_3OD)	δ 182.3, 171.6, 171.1, 158.5, 153.1, 150.8, 150.4, 147.6, 140.6, 121.8, 111.4 (d, *J* = 327 Hz, 1C), 87.4, 83.75 (d, *J* = 15 Hz, 1C), 74.0, 73.0, 64.6 (d, *J* = 6.4 Hz, 1C), 47.1 (**TEAH$^+$**), 36.8, 20.6, 20.3, 19.6, 19.5, 11.9, 10.5 (**TEAH$^+$**)
^{31}P NMR (CD_3OD)	δ 10.55 (s)
HRMS	Calc'd for $C_{22}H_{29}N_7O_{10}P$ $[M-H]^-$: 580.1563; Found: 580.1558

B.2.4. Guanosine (3-methyl-1*H*-pyrazol-4-yl)phosphonate (PZG)



Compound **5** was added to a heavy-walled and round-bottomed borosilicate glass high-pressure reaction vessel along with a stir bar, and was then dissolved in a minimum amount of methanol. Afterwards, an excess of concentrated ammonium hydroxide (28% NH_3 in H_2O; 50 mL) was added to the vessel. The vessel was sealed, heated to 65 °C with a pre-equilibrated oil bath, and stirred for 4 hours (*CAUTION: CONTENTS UNDER PRESSURE. Prior to reaction setup, the high-pressure reaction vessel should be thoroughly inspected to ensure absence of any star-shaped glass cracks. Placement of a weighted safety shield in front of the reaction is recommended.*). Afterwards, the reaction was allowed to equilibrate to room temperature, followed by solvent evaporation *in vacuo* to afford a white solid deposit. The crude material was resuspended in water, followed by direct liquid-loading into a manufacturer-packed 50 g C18Aq column pre-equilibrated with 25 mM pH 7.5 aqueous triethylammonium bicarbonate (TEAB) buffer. The resulting crude residue was purified over 15 CV of solvents, with a linearly increasing gradient of 0–20% CH_3CN in TEAB. The title compound was eluted from the column at ~7% CH_3CN in TEAB, followed by lyophilization to afford the titled compound (in triethylammonium cation form) as a white fluffy solid. Sodium cation exchange was performed on the sample, by dissolving the sample in a minimum amount of deionized water, followed by re-precipitation in >25-fold excess (in volume) of 1:1:0.04 (v/v/v)

acetone : diethyl ether : saturated $NaClO_4$ in acetone. The precipitate was centrifuged and washed with 1:1 acetone: diethyl ether twice. Final drying of the centrifuged and washed sample on high-vacuum furnished 27.8 mg (0.062 mmol) of **PZG** (Na^+ cation form) as fine white granular solids. Final percentage yield of **PZG** was calculated by dividing the number of moles of **PZG** by that of the protected guanosine starting material, and was determined to be 10.8 % over three steps.

^{1}H NMR (D_2O)	δ 7.85 (s, 1H), δ 7.64 (s, 1H), δ 5.81 (d, *J* = 4.0 Hz, 1H), δ 4.73 – 4.70 (m, 1H), δ 4.39 – 4.37 (m, 1H), δ 4.22 – 4.20 (m, 1H), δ 3.90 – 3.93 (m, 2H), δ 3.13 (q, *J* = 7.3 Hz, **TEAH$^+$**, 6H), δ 2.18 (s, 3H), δ 1.22 (t, *J* = 7.3 Hz, **TEAH$^+$**, 9H)
^{13}C NMR (D_2O)	δ 168.4, 161.7, 152.7, 148.1 (d, *J* = 29 Hz, 1C), 140.7 (d, *J* = 29 Hz, 1C), 136.2, 118.5, 109.8 (d, *J* = 332 Hz, 1C), 87.1, 84.6 (d, *J* = 14.5 Hz, 1C), 74.5, 71.7, 64.8, 47.1 (**TEAH$^+$**), 11.8 and 9.2 (**TEAH$^+$**)
^{31}P NMR (D_2O)	δ 11.53 (s)
HRMS	Calc'd for $C_{20}H_{34}N_8O_7P$ $[M-H]^-$: 426.0933; Found: 426.0930

B.3. X-ray crystallography.

B.3.1. Crystal preparation. The Nucleic Acid Mini Screen Kit, Natrix High Throughput Kit and Index High Throughput Kit (Hampton Research, Aliso Viejo, CA) were all used for screening crystallization conditions by the sitting drop vapor diffusion method. RNA samples (1 mM) were mixed with either **PZG** (50 mM) or GMP (50 mM) at a 1:1 ratio, heated to 80^{o}C for 2 min, and then cooled down slowly to room temperature. All of the crystals grew at 18 ^{o}C, and the mother liquor containing 50% glycerol was used as a cryoprotectant during crystal mounting. All data collection was taken under a stream of nitrogen at 99 K. The data sets were collected at the SIBYLS beamline 821 and 822 at the Advanced Light Source, Lawrence Berkeley National Laboratory. The distances between the detector and the crystal were set to 200-300 mm and the collecting wavelength was set to 0.9795 Å. The crystals were exposed for 1 second per image with one degree oscillations, and 180 images were taken for each data set. The optimized crystallization conditions for the RNA-monomer complexes are listed below.

Table B.1. Optimized conditions for crystallization of RNA-monomer complexes

	Optimized crystallization conditions
RNA 15mer-PZG (PDB: 5KRG)	0.2 M potassium chloride, 0.05 M HEPES pH 7.5, 35% v/v pentaerythritol propoxylate
RNA 7mer-PZG (PDB: 5HBW)	0.040 M sodium cacodylate trihydrate pH 7.0, 0.012 M spermine tetrahydrochloride, 0.080 M sodium chloride, 0.012 M potassium chloride, 0.020 M magnesium chloride hexahydrate, 10% v/v (+/-)-2-methyl-2,4-pentanediol
RNA 14mer-2PZG (PDB: 5HBX)	0.05 M ammonium acetate, 0.01 M magnesium chloride hexahydrate, 0.05 M tris hydrochloride pH 7.5, 10% v/v (+/-)-2-methyl-2,4-pentanediol
RNA 13mer-3PZG (PDB: 5HBY)	0.01 M magnesium chloride, 0.1 M potassium thiocyanate, 30% w/v polyethylene glycol monomethyl ether 2,000
RNA 15mer-GMP (PDB: 5L00)	10% v/v (+/-)-2-methyl-2,4-pentanediol, 0.040 M sodium cacodylate trihydrate pH 7.0, 0.012 M spermine tetrahydrochloride, 0.012 M sodium chloride, 0.080 M potassium chloride
RNA 7mer-GMP (PDB: 5DHB)	0.05 M magnesium chloride, 0.1 M sodium chloride, 0.1 M HEPES pH 7.5, 30% v/v polyethylene glycol monomethyl ether 550
RNA 14mer-2GMP (PDB: 5DHC)	0.1 M potassium chloride, 0.005 M magnesium sulfate, 0.05 M sodium chloride, 0.1 M HEPES pH 7.5, 15% v/v (+/-)-2-methyl-2,4-pentanediol

B.3.2. Data collection and structure refinement. The data were processed using HKL2000.[8] All of the structures with one binding site and two binding sites were solved by molecular replacement. The RNA sequences were obtained from the PDB data bank, and the models were obtained directly from PDB or built as ideal A-form RNA.[9] All three structures were refined using Refmac.[10] The refinement protocol included simulated annealing refinement, restrained B-factor refinement, and bulk solvent correction. The RNA-PZG complex structure with three template binding sites was determined by *ab initio* methods using the program ACORN.[11] The data were artificially extended to a resolution of 1.0 Å. Anisotropy correction was performed with the program PHASER.[12] The final E-map and F-map were interpretable, and both were used to help build the initial model in COOT.[9] During refinement, the topologies and parameters for locked nucleic acids and for the ligand PZG (LCG, LCA, LCC, TLN, PZG) were constructed and applied. After several cycles of refinement, a number of highly ordered water molecules and magnesium atoms were added. The 7 crystal structures were determined to resolutions of 1.6 Å, 1.9 Å, 1.7 Å and 1.18 Å, 1.25 Å, 1.8 Å and 1.55 Å respectively (PDB ID: 5KRG, 5HBW, 5HBX, 5HBY, 5L00, 5DHB and 5DHC). Data collection, phasing, and refinement statistics of the determined structures are listed in **Tables B.2 & 3**.

The structures of the RNA-GMP complexes are presented in **Figure B.1**. In **Figure B.2** we note that the phosphorus atom of the first bound monomer (out of two bound to each end of the duplex) of either the GMP-RNA structure or the PZG-RNA structure, is closer to the 3′-hydroxyl of the primer terminus (roughly 4 Å) compared to that of the one binding site structure (around 6 Å).

Table B.2. Data collection statistics.

Structure	PZG-RNA 15mer (5KRG)	PZG-RNA 7mer (5HBW)	PZG-RNA 14mer (5HBX)	PZG-RNA 13mer (5HBY)	GMP-RNA 15mer (5L00)	GMP-RNA 7mer (5DHB)	GMP-RNA 14mer (5DHC)
Space group	P3	$P2_12_12_1$	P3	C_12_1	P3	P3	P3
Unit cell parameters, Å (°)	44.05, 44.05, 84.52 (90, 90, 120)	28.08, 39.75, 42.59 (90, 90, 90)	43.54, 43.54, 84.00 (90, 90, 120)	45.87, 31.99, 33.43 (90, 122, 90)	43.85, 43.85, 86.45 (90, 90, 120)	51.25, 51.25, 38.76 (90, 90, 120)	43.11, 43.11, 83.24 (90, 90, 120)
Resolution range, Å (last shell)	50-1.60 (1.66-1.60)	50-1.90 (1.97-1.90)	50-1.70 (1.76-1.70)	50-1.18 (1.22-1.18)	50-1.25 (1.29-1.25)	50-1.80 (1.86-1.80)	50-1.55 (1.61-1.55)
Unique reflections	41492 (1867)	7254 (720)	38632 (3727)	23597 (1042)	47622 (3328)	20608 (1882)	25146 (2528)
Completeness, %	95.5 (71.3)	99.6 (99)	99.8 (100)	93.6 (54.2)	92.7 (64.5)	98.1 (94.3)	99.9 (100)
R_{merge}, %	11.6 (27.4)	5.3 (50.7)	12.8 (48.8)	4.7 (24.8)	4.4 (27.7)	7.9 (58.7)	9.6 (46.1)
$\langle I/\sigma(I)\rangle$	11.9 (2.2)	24.8 (3.3)	14.6 (2.6)	26.5 (1.8)	34.4 (2.5)	29.8 (2.1)	17.5 (2.1)
Redundancy	4.5 (1.6)	6.1 (5.9)	7.4 (5.4)	3.7 (2.0)	5.0 (2.3)	15.6 (4.6)	15.6 (4.6)

Table B.3. Structure refinement statistics.

Structure	PZG-RNA 15mer (5KRG)	PZG-RNA 7mer (5HBW)	PZG-RNA 14mer (5HBX)	PZG-RNA 13mer (5HBY)	GMP-RNA 15mer (5L00)	GMP-RNA 7mer (5DHB)	GMP-RNA 14mer (5DHC)
PDB code	5KRG	5HBW	5HBX	5HBY	5L00	5DHB	5DHC
RNA duplex per asymmetric unit	2	1	2	0.5	2	2	2
Resolution range, Å	84.52-1.60	23.44-1.90	84.00-1.70	24.70-1.18	50.00-1.25	44.38-1.80	83.25-1.55
R_{work}, %	20.6	21.7	21.3	15.3	20.8	18.7	17.0
R_{free}, %	23.4	25.4	24.9	16.6	23.8	19.2	19.5
Number of reflections	21914	3841	18463	12091	47622	20608	25146
Bond length R.M.S., Å	0.016	0.019	0.018	0.018	0.030	0.027	0.023
Bond angle R.M.S.	1.858	2.234	2.030	2.058	2.985	2.921	2.773
Average B-factors, Å^2	27.54	33.77	29.32	13.29	16.85	31.38	20.59

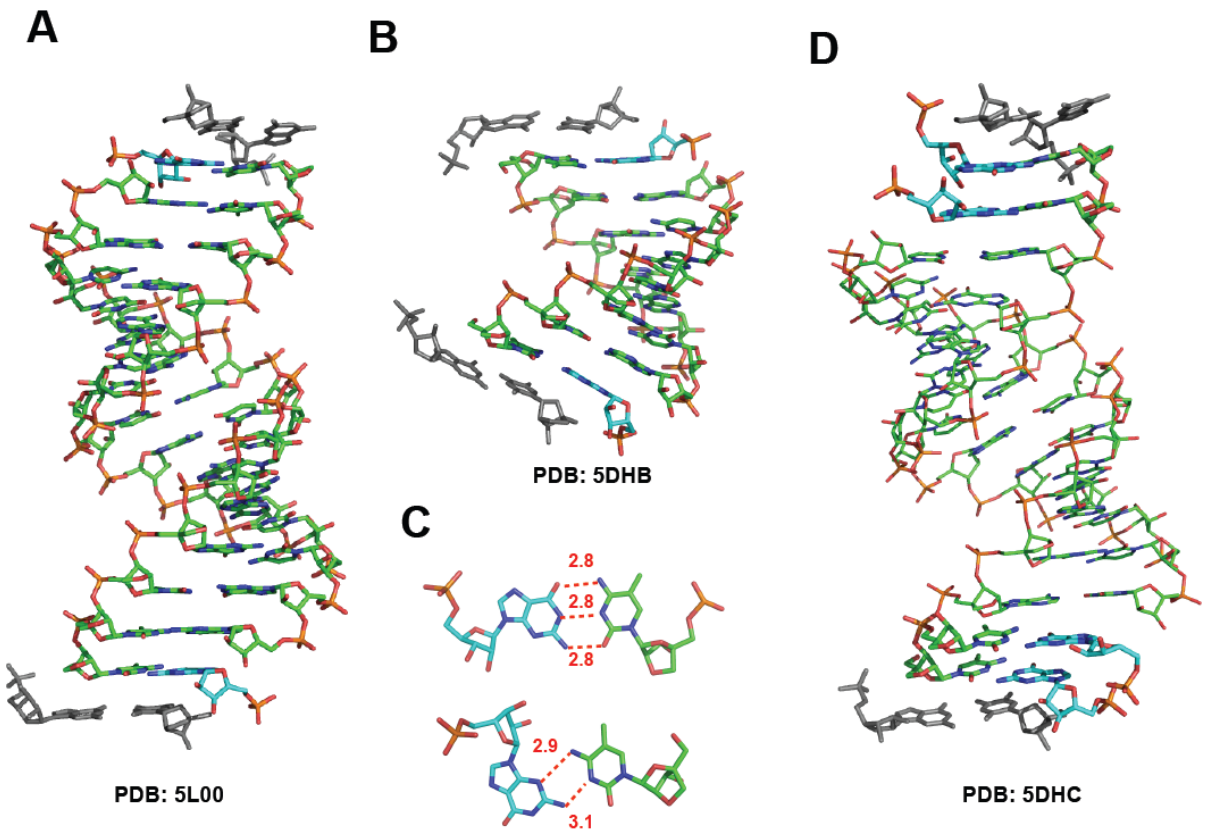


Figure B.1. X-Ray crystal structures of RNA-GMP complexes. Green: RNA duplex; cyan: GMP monomer. Nucleotides displayed in grey represent the base pairs of adjacent slip-stacked RNA duplexes. **A**: Structure of the 5′-mC overhang 15mer RNA duplex bound with one GMP on both of the termini. **B**: Structure of the 5′-mC 7mer RNA duplex bound with one GMP on both of the termini. **C**: Zoom-in of the two types of base pairs observed in GMP-RNA complex structures, including Watson-Crick fashion and an atypical geometry. **D**: Structure of the 14mer RNA duplex with the 5′-mCmC overhang bound by two GMP monomers on each end. The sugar and phosphate groups of the GMP monomers are disordered and are illustrated in arbitrary conformations.

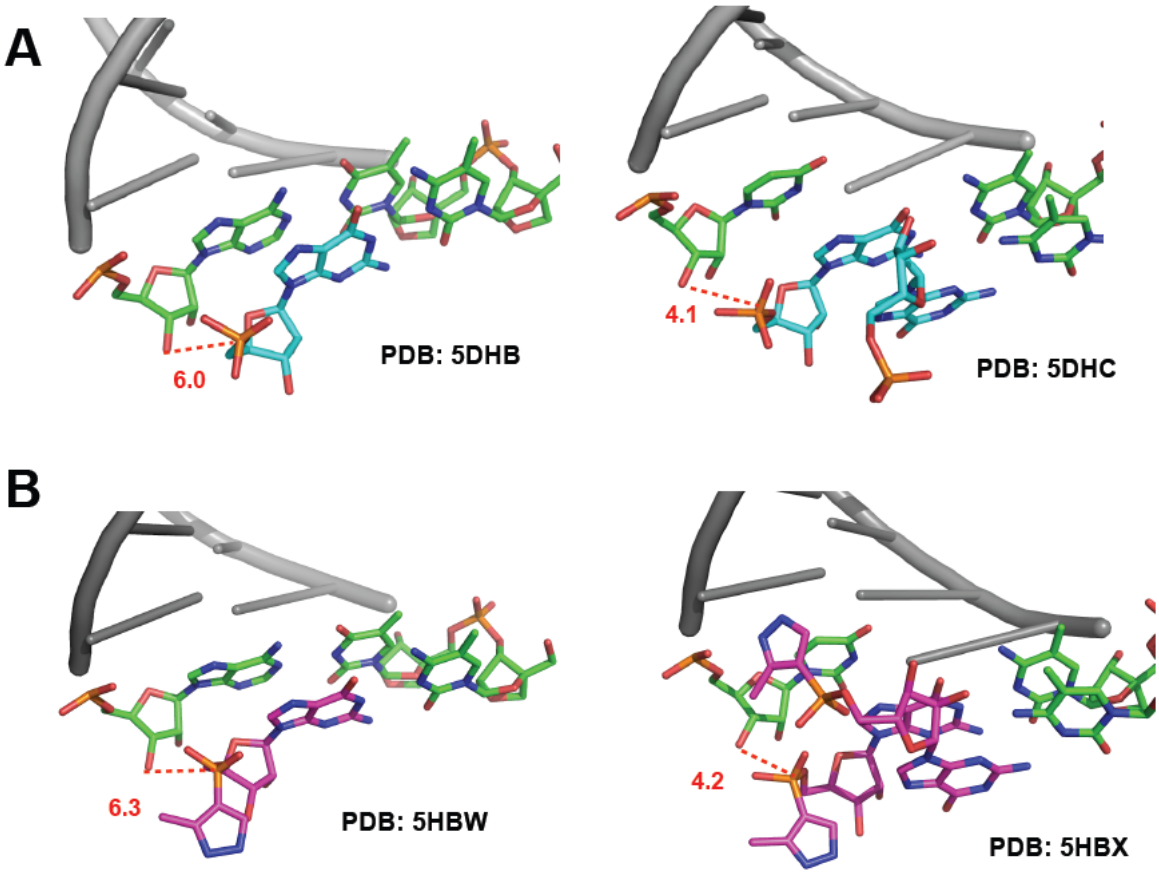


Figure B.2. Zoom-in of the monomer-binding regions around the termini of GMP-RNA and PZG-RNA complexes. Green: templating RNA and the 3′-terminus of the primer immediately upstream of the monomer; magenta color: PZG monomer; cyan: GMP monomer. **A**: For the 5′-mC:GMP duplex, the phosphorus atom of the bound GMP shown is 6.0 Å away from the 3′-hydroxyl of the primer terminus. For the first bound monomer of the 5′-mCmC:2GMP duplex, the distance shown is 4.1 Å. **B**: For the 5′-mC:PZG duplex, the phosphorus atom of the bound PZG shown is 6.3 Å away from the 3′-hydroxyl of the primer terminus. For the first bound monomer of the 5′-mCmC:2PZG duplex, the distance shown is 4.2 Å. The shorter distance could enhance the rate of primer extension. Parts of the GMP and PZG monomers are disordered and are illustrated in arbitrary conformations.

B.4. References.

(1) Kaur, H.; Arora, A.; Wengel, J.; Maiti, S. *Biochemistry* **2006**, *45*, 7347–7355.

(2) Van Den Berge, E.; Robiette, R. *J. Org. Chem.* **2013**, *78*, 12220–12223.

(3) Kalek, M.; Ziadi, A.; Stawinski, J. *Org. Lett.* **2008**, *10*, 4637–4640.

(4) Matulic-Adamic, J.; Haeberli, P.; Usman, N. *J. Org. Chem.* **1995**, *60*, 2563–2569.

(5) Błażewska, K. M. *J. Org. Chem.* **2013**, *79*, 408–412.

(6) McKenna, C. E.; Higa, M. T.; Cheung, N. H.; McKenna, M.-C. *Tetrahedron Lett.* **1977**, *18*, 155–158.

(7) Taylor, S. D.; Mirzaei, F.; Bearne, S. L. *J. Org. Chem.* **2008**, *73*, 1403–1412.

(8) Otwinowski, Z.; Minor, W. *Methods Enzym.* **1997**, *276*, 307–326.

(9) Emsley, P.; Cowtan, K. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2004**.

(10) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. *Acta Crystallogr. D Biol. Crystallogr.* **1997**, *53*, 240–255.

(11) Foadi, J.; Woolfson, M. M.; Dodson, E. J.; Wilson, K. S.; Jia-xing, Y.; Chao-de, Z. *Acta Crystallogr. D Biol. Crystallogr.* **2000**, *56*, 1137–1147.

(12) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. *J. Appl. Crystallogr.* **2007**, *40*, 658–674.

Appendix C

Supplemental Materials for Chapter 3

C.1. NMR Titration Studies.

The initial duplex solution contained 1.5 mM RNA duplex, 500 mM sodium chloride, and 10% D_2O; the ligand solution contained the appropriate concentration of GpppG (sodium salt form), as well as the same concentration of the RNA duplex (1.5 mM), D_2O (10%), and sodium cation (500 mM) as that of the starting duplex solution, in order to maintain a constant duplex concentration and ionic strength throughout the titration experiment. A monomer solution containing 30 mM of GpppG was assumed to contain 90 mM of Na^+, and 410 mM of sodium chloride was added to the monomer solution to bring the concentration of Na^+ of the monomer solution to 500 mM. The pH of both duplex and monomer solutions was adjusted to 7.0 (± 0.1) using trace amounts of either aqueous sodium hydroxide or hydrochloric acid. NMR spectra were acquired on a Varian INOVA 400 MHz NMR spectrometer equipped with a broadband PFG (z-gradient) probe. Suppression of the bulk water resonance was achieved using a Watergate pulse sequence.[1,2] Each spectrum was recorded after 128–256 scans, with an optimized delay period on pulse sequence (d1) of 1.0 s and a pulse width (pw) of 15 μs. Initial concentrations of the duplex and monomer solutions were determined by their UV absorption at 260 nm on a Thermo Scientific Nanodrop 2000c Spectrophotometer (Waltham, MA). The theoretical molar extinction coefficients at 260 nm of the RNA strands were calculated with Integrated DNA Technologies' OligoAnalyzer 3.1 (Coralville, IA);[3,4] the molar extinction coefficients of pGpG, GppG, and GpppG at 260 nm were assumed to be 24160 L mol^{-1} cm^{-1}. The chemicals shifts were referenced externally using a co-

axial NMR tube containing a solution of pentafluorobenzaldehyde (δ = 10.285 ppm) in $CDCl_3$.

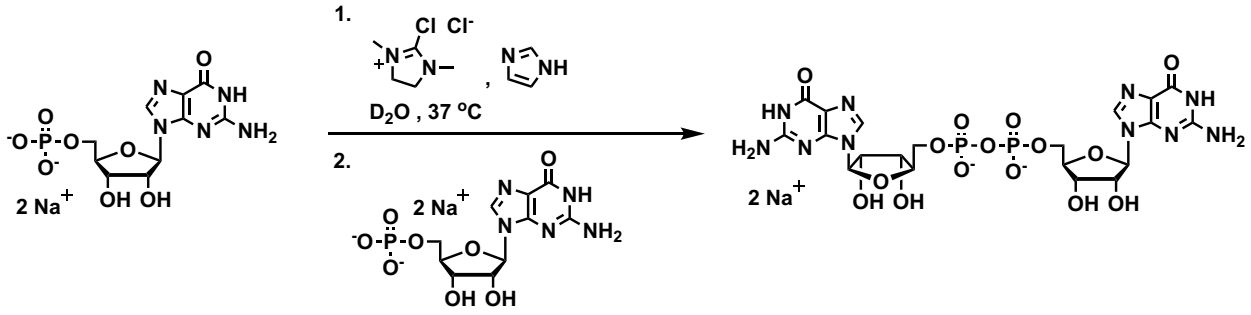
C.2. Synthesis of Diguanosine Dinucleotides: Experimental procedures and compound characterizations.

C.2.1. 5′-*O*-Phosphonoguanylyl-(3′→5′)-guanosine (pGpG)

Using standard solid-phase phosphoramidite polymerization chemistry, pGpG was prepared by four 50-μmol-scale solid-phase syntheses (total scale: 0.2 mmol) on a MerMade 6 oligonucleotide synthesizer, in accordance to procedures laid out in **Appendix A.3** of this Supporting Information.

^{1}H NMR (D_2O)	δ 8.11 (s, 1H), 7.99 (s, 1H), 5.84 (d, *J* = 6.6 Hz, 1H), 5.82 (d, *J* = 5.6 Hz, 1H), 4.7 (dd, *J* = 5.7 & 5.7 Hz), 4.46 (dd, *J* = 3.9 & 4.3 Hz, 1H), 4.41 (br s, 1H), 4.28 (br s, 1H), 4.14 (br s, 2H), 3.92 (br s, 2H). *Protons on the 2'-position of the two riboses (~ δ 4.84 – 4.80) were poorly resolved from the HOD signal at δ 4.79 ppm.*
^{13}C NMR (D_2O)	δ 162.20, 162.05, 156.89, 156.78, 152.80, 152.65, 138.02, 137.88, 117.48, 117.40, 87.71, 86.93, 84.88 (dd, *J* = 9 & 3 Hz), 84.65 (d, *J* = 9 Hz), 76.33 (d, *J* = 5 Hz), 74.51, 74.03 (d, *J* = 5 Hz), 71.38, 66.12 (d, *J* = 5 Hz), 64.5 (d, *J* = 4 Hz)
^{31}P NMR (D_2O)	δ 0.34, –4.54
HRMS (*m/z*)	Calc'd for $C_{20}H_{25}N_{10}O_{15}P_2$ $[M–H]^–$: 707.0982; Found: 707.1016

C.2.2. P^1,P^2-Diguanosine-5'-pyrophosphate (GppG)

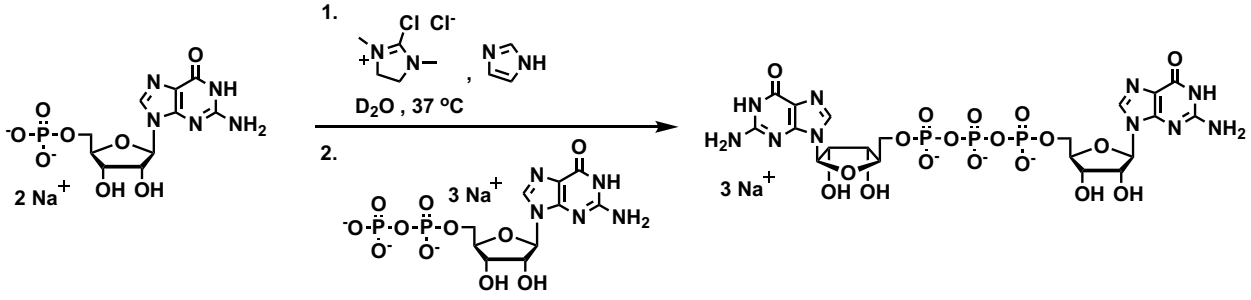


The synthetic protocol was adapted from the report of Tanaka et al.[5] 2-Chloro-1,3-dimethylimidazolinium chloride (DMC, 340 mg, 2 mmol, 2 equiv.; *very hygroscopic*), imidazole (272.3 mg, 4 mmol, 4 equiv.) and guanosine 5'-monophosphate disodium salt (GMP•2Na$^+$, 407.2 mg, 1 mmol, 1 equiv.) were charged into an oven-dried 4 mL borosilicate glass vial equipped with a stir bar and a Teflon-lined cap. D_2O (2 mL, 0.5 M) was added as a solvent. The mixture was stirred at 40 ^oC for 1 h, followed by addition of another 1 equiv. of GMP•2Na$^+$ (407.18 mg). After further stirring at 40 ^oC overnight, the reaction mixture was diluted 20-fold with water, followed by preparative-scale HPLC purification on a Varian Prostar 210 HPLC system equipped with Agilent ZORBAX Eclipse-XDB C18 column using 25 mM triethylammonium bicarbonate in H_2O (pH 7.5) with an increasing gradient of 0 % to 15 % acetonitrile over 30 m. Elution of RNA was monitored by UV absorption at 254 and 280 nm. The desired RNA fractions were collected, pooled and lyophilized to afford a fluffy white powder. The resultant white residues were again purified with preparative-scale strong anion-exchange HPLC with a ThermoFisher-Dionex DNAPac PA100 strong anion exchange column with an increasing gradient of 0 to 100 mM aqueous sodium perchlorate over 30 m. The desired fractions were collected, pooled and lyophilized at <50 mTorr to afford white solid residues. The residues were

washed with acetone for 3 times, followed by overnight drying under high vacuum, to afford the desired products (in sodium cation form) as fine white grains.

^{1}H NMR (D_2O)	δ 7.94 (s, 2H), 5.78 (d, *J* = 5.2 Hz, 2H), 4.62 (dd, *J* = 5.1 & 5.1 Hz, 2H), 4.44 (dd, *J* = 4.4 & 4.6 Hz, 2H), 4.31–4.26 (m, 4H), 4.22–4.17 (m, 2H)
^{13}CNMR (D_2O)	δ 159.5, 154.6, 152.1, 138.1, 116.9, 88.3, 84.1 (t, *J* = 4.5 Hz), 74.9, 70.9, 65.9
^{31}P NMR (D_2O)	δ –10.50
HRMS (*m/z*)	Calc'd for $C_{20}H_{25}N_{10}O_{15}P_2$ $[M–H]^-$: 707.0982; Found: 707.1013

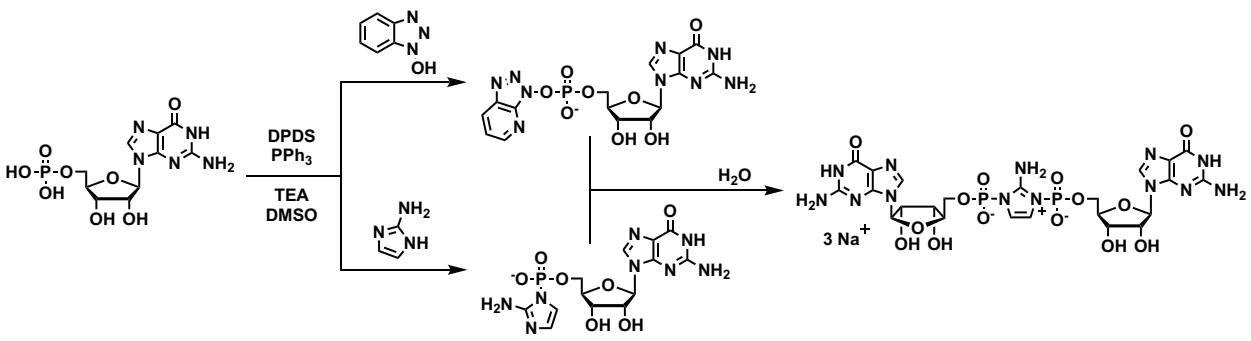
C.2.3. *P*^{1},*P*^{3}-Diguanosine-5'-triphosphate (GpppG)



The synthetic procedures leading to GpppG were highly analogous to that of GppG, except that guanosine 5'-diphosphate (sodium salt form) was added in lieu of GMP•2Na^{+} following the initial one-hour DMC-mediated GMP activation.

^{1}H NMR (D_2O)	δ 8.05 (s, 2H), 5.88 (d, *J* = 4.8 Hz, 2H), 4.67 (dd, *J* = 4.9 & 4.9 Hz, 2H), 4.5 (dd, *J* = 4.6 & 4.7 Hz, 2H), 4.36 (br, 2H), 4.33 (m, 4H)
^{13}CNMR (D_2O)	δ 159.4, 154.6, 152.1, 138.1, 116.8, 88.2, 84.0 (d, *J* = 9 Hz), 74.9, 70.7, 65.7
^{31}P NMR (D_2O)	δ –10.45 (d, *J* = 9.7 Hz), –21.76 (t, *J* = 17.7 Hz)
HRMS	Calc'd for $C_{22}H_{29}N_7O_{10}P$ $[M–H]^-$: 787.0645; Found: 787.0675

C.2.4. 1,3-di-(guanosine-5'-phosphoryl)-2-aminoimidazolium (Gp-NH_2Im-pG)



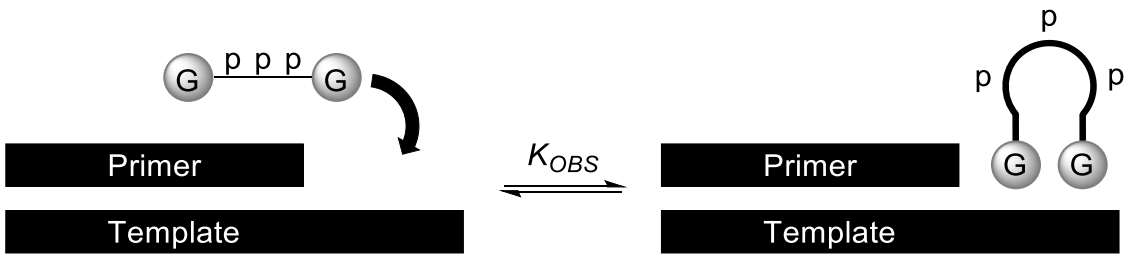
The synthesis of the di-guanosine intermediate began by first synthesizing two monomers, guanosine-5'-phosphoryl-(2-aminoimidazole) (2-AmImpG) and guanosine-5'-phosphoryl-(1-hydroxy-7-azabenzotriazole) (GMP-OAt). Aqueous mixtures of GMP•$2H^+$ (100 mg, 0.275 mmol, 1 equiv.) with either 2-aminoimidazole (dark brown oil, 91 mg, 1.1 mmol, 4 equiv.) or 1-hydroxy-7-azabenzotriazole (200 mg, 1.47 mmol, 5.3 equiv.), and triethylamine (200 µL, 1.4 mmol, 4 equiv., ρ = 0.726 g mL^{-1}) were first prepared, vortexed and sonicated until complete homogenization, then flash frozen in liquid nitrogen and lyophilized over 5 days at < 50 mTorr. Once dry, the solids were separately resuspended in DMSO (20 mL, 14 mM) and triethylamine (300 µL, 2.15 mmol, 7.8 equiv., ρ = 0.726 g mL^{-1}). 2,2'-dipyridyl disulfide (1.2 g, 5.5 mmol, 20 equiv.) and triphenylphosphine (1.2 g, 4.6 mmol, 17 equiv.) were added and the reactions were left stirring overnight (~ 12 hrs). Extra 2,2'-dipyridyldisulfide (0.62 g, 2.8 mmol, 10.2 equiv.) and triphenylphosphine (0.6 g, 2.3 mmol, 8.4 equiv.) were added to the reactions. Four hours later, both reactions were separately precipitated in precipitation solutions containing 120 mL acetone, 60 mL diethyl ether, and 4.5 g of sodium perchlorate. Pellets were washed twice with acetone, followed by house vacuum-drying overnight. Afterwards, one half of the crude monomer GMP-OAt and one half of the crude 2-AmImpG were mixed together in 5 mL H_2O. This

mixture was incubated for 70 m at room temperature and became viscous. The mixture was then purified by reverse phase flash chromatography, by direct loading onto a 30 g C18aq column on a Combiflash Rf-200 from Teledyne Isco. The products were eluted over 15 column volumes, with a gradient of 20 mM TEAB (pH = 7.5) over 0 to 20% acetonitrile at a flow rate of 20 mL/m. The fraction containing the di-guanosine intermediate was immediately flash frozen and lyophilized at -20°C on a VirTis AdVantage Plus EL-85 lyophilizer from SP Scientific. Excess triethylamine from the previous purification was removed by reverse phase chromatography again, using the same procedure, except that the aqueous solvent was water instead of TEAB. The fraction containing the intermediate was then flash frozen and lyophilized at -20°C.

^{1}H NMR (D_2O)	δ 7.68 (s, 2H), 6.56 (dd, *J* = 2.6, 1.8 Hz, 2H), 5.59 (d, *J* = 5.2 Hz, 2H), 4.51 (t, *J* = 5.2 Hz, 2H), 4.26 (dd, *J* = 5.0, 4.5 Hz, 2H), 3.97–3.95 (m, 4H), 3.88-3.83 (m, 2H)
^{13}C NMR (D_2O)	δ 159.4, 154.3, 152.2, 150.7 (weak, t, *J* = 7 Hz),138.0, 116.8, 116.4–116.6 (m), 87.9, 83.2 (d, *J* = 8.3 Hz), 73.9, 70.4, 66.5 (d, *J* = 6.4 Hz)
^{31}P NMR (D_2O)	δ –9.04
HRMS (*m/z*)	Calc'd for $C_{23}H_{28}N_{13}O_{14}P_2$ $[M–H]^-$: 772.1359; Found: 772.1371

C.3. Affinity of GpppG for the RNA duplex P/T_{2C}

C.3.1. Derivation of the NMR binding isotherm.[6]



Scheme C.1. The proposed binding scheme between **P/T_{2C}** duplex with GpppG.

We have previously reported the use of the **P/T_{2C}** duplex (Sequence: **P**: 5'–CUCAAUG–3'; **T_{2C}**: 5'–CCCAUUGAG–3') to monitor the binding affinity of two consecutive GMP molecules onto the RNA duplex.[7] Upon GMP binding, the imino proton signal of G7 (the 3'-terminal nucleotide of the primer, flanking the ligand binding sites) was observed to shift upfield, without observation of new imino proton signals. This observation is in line with our hypothesis that GMP binding with RNA duplexes is in fast exchange, with the on- and off-rates of ligand binding being faster than the NMR time scale. When these conditions are met, and when the ligand-duplex binding stoichiometry is expected to be 1:1, the following single-site binding isotherm is commonly used to numerically approximate the binding affinity (*K*):

$$\Delta\delta = \frac{\Delta\delta_{tot} K[M]}{1 + K[M]}$$

with Δδ being the measured chemical shift change induced by monomer binding, $\Delta\delta_{tot}$ being the theoretical total chemical shift change when all duplex binding sites are fully saturated, and [M] being the concentration of *free*, *unbound* monomer (*not* the concentration of monomer added to the duplex solution). Since we expected the binding of GpppG onto **P/T_{2C}** to be tight, we hypothesized that a substantial portion of the added

monomer would be bound to the duplex, with only a small fraction of the added monomer being left in the unbound state. We therefore need to derive an equation which will explicitly solve the concentration of unbound monomer at all points of the titration.

We begin the derivation by realizing that the binding of GpppG with the **P/T$_{2C}$** duplex can be represented by the following equilibrium:

$$[D] + [M] \underset{}{\overset{K}{\rightleftharpoons}} [DM]$$

where D represents **P/T$_{2C}$** duplex, M represents GpppG, and DM represents GpppG-bound RNA duplex. *K* is equal to:

$$K = \frac{[DM]}{[D] \cdot [M]}$$

At any point of the titration, the total concentration of duplex (D_{tot}) is equal to the sum of the concentration of GpppG-bound RNA duplex (DM), and free RNA duplex (D). The same applies to monomer as well. Hence:

$$[D] + [DM] = [D_{tot}]\ ;\ [M] + [DM] = [M_{tot}]$$

Substituting these equations into the equilibrium expression, we get:

$$K = \frac{[DM]}{(D_{tot} - [DM]) \cdot (M_{tot} - [DM])}$$

Rearrangement of this expression in terms of [DM] gives a quadratic equation:

$$[DM]^2 - \left(M_{tot} + D_{tot} + \frac{1}{K}\right)\ [DM]\ +\ M_{tot}D_{tot}\ = 0$$

and the roots (solution) of this quadratic expression can be expressed as the following:

$$[DM] = \frac{\left(M_{tot} + D_{tot} + \frac{1}{K}\right) \pm \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}{2}$$

As K approaches positive infinity, the two roots bifurcate into M_{tot} (larger root) and D_{tot} (smaller root); however, since $M_{tot} > D_{tot}$ at the end of the titration, [DM] must be limited by the total duplex concentration (D_{tot}). Hence [DM] cannot be equal to M_{tot}, and the larger root is rejected. Hence,

$$[DM] = \frac{\left(M_{tot} + D_{tot} + \frac{1}{K}\right) - \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}{2}$$

Subsequently, at any point of the NMR titration, the concentration of free, unbound monomer is equal to

$$[M] = [M_{tot}] - [DM]$$

$$[M] = \frac{\square_{tot} - D_{tot} - \frac{1}{K} + \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}{2}$$

This equation is substituted back into the canonical NMR binding isotherm:

$$\Delta\delta = \frac{\Delta\delta_{tot}K[M]}{1 + K[M]}$$

and we get the crude NMR binding isotherm which is then used to fit the NMR data for numerical approximation of the GpppG–RNA duplex binding affinity:

$$\Delta\delta = \frac{\frac{1}{2}\Delta\delta_{tot}K\left\{M_{tot} - D_{tot} - \frac{1}{K} + \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}\right\}}{1 + \frac{1}{2}K\left\{M_{to\square} - D_{tot} - \frac{1}{K} + \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}\right\}}$$

$$\Delta\delta = \frac{\Delta\delta_{tot}K\left\{M_{tot} - D_{tot} - \frac{1}{K} + \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}\right\}}{2 + K\left\{M_{tot} - D_{tot} - \frac{1}{K} + \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}\right\}}$$

The true total monomer concentration (M_{tot}) throughout the titration may deviate from the theoretical value, due to factors like handling error in monomer addition, or inaccuracies in the approximation of the true concentration of the ligand in the monomer solution. To account for these possible deviations, the stoichiometric factor *n* is introduced to the D_{tot} term, yielding the final form:

$$\Delta\delta = \frac{\Delta\delta_{tot}K\left\{M_{tot} - nD_{tot} - \frac{1}{K} + \left[\left(M_{tot} + nD_{tot} + \frac{1}{K}\right)^2 - 4nM_{tot}D_{tot}\right]^{\frac{1}{2}}\right\}}{2 + K\left\{M_{tot} - nD_{tot} - \frac{1}{K} + \left[\left(M_{tot} + nD_{tot} + \frac{1}{K}\right)^2 - 4nM_{tot}D_{tot}\right]^{\frac{1}{2}}\right\}}$$

C.3.2. Titration of GpppG into the P/T_{2C} duplex: NMR studies.

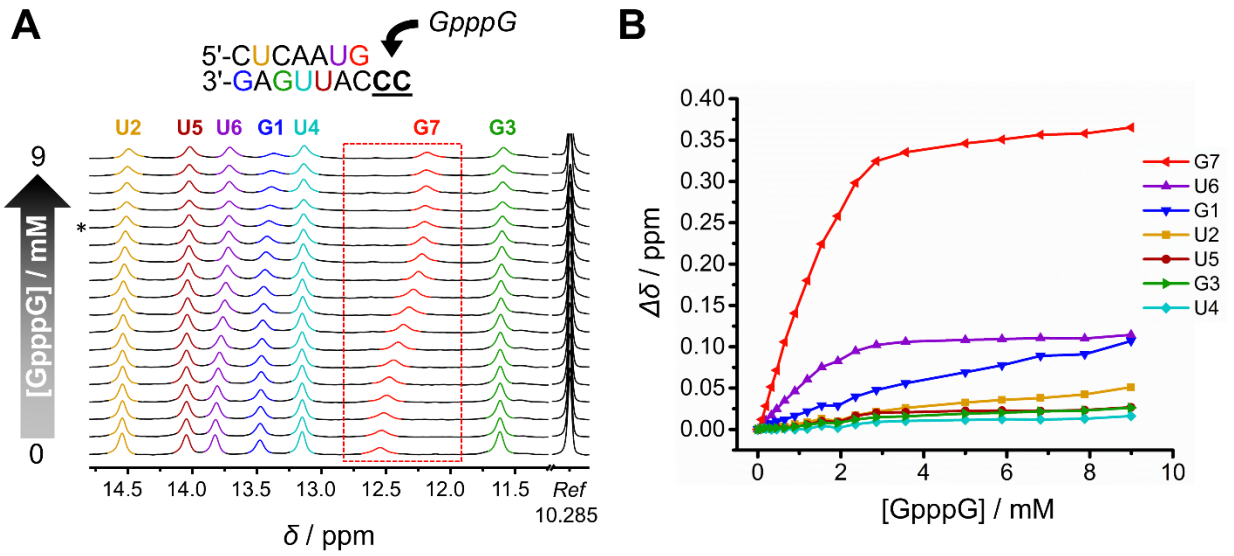


Figure C.1. Titration of the **P/T_{2C}** duplex with GpppG while monitoring the 10 to 15 ppm region of the ^{1}H NMR spectra.[8,9] **A.** The component spectra for the GpppG–**P/T_{2C}** titration were stacked for convenient visualization. The 5'-CC overhang of **P/T_{2C}** engages in binding with GpppG. All spectra were recorded at 12 °C in the presence of 500 mM Na^{+} and 10% D_2O at pH 7 with a duplex concentration of 1.5 mM. The signal at 10.285 ppm is that of the pentafluorobenzaldehyde reference. **B.** Using the spectral data shown in **Figure C.1a**, the change in chemical shifts of all **P/T_{2C}** imino protons were plotted against the concentration of GpppG. The imino proton signal of G7 (red) shifted the most, followed by the signals of U6 (purple), G1 (blue), and U2 (mustard yellow). The signals of internal imino protons U5, G3, and U4 (crimson, green, and cyan, respectively) shifted only minimally.

To assess the affinity of GpppG with cognate RNA duplexes, we used a previously-reported **P/T_{2C}** duplex[7] with a 5'-CC overhang. Under the conditions of the NMR titration experiments (500 mM Na^{+}, pH 7, 12 °C), the primer and template strands are known to be stably annealed to form an A-form duplex. Additionally, the imino protons of the **P/T_{2C}** duplex are clearly defined and baseline-resolved. The assignment of proton resonances (**Figure C.1a**) are deduced from data of previously-described variable-temperature proton and two-dimensional ^{1}H-^{1}H NOESY experiments.[7,10]

We carried out a titration from 0 to 9 mM GpppG (0 to 6 equivalents) into 1.5 mM of **P/T$_{2C}$** solution, and the change in chemical shift of the G7 imino proton was fitted to a modified single-site binding isotherm. As the concentration of GpppG increased, G7 shifted upfield remarkably (~0.35 ppm, red trace, **Figure C.1b**) and in a hyperbolic fashion, which is in line with our previous observations that purine nucleotide ligands tend to shift the G7 resonance upfield.[10] The imino proton resonance of U6 also undergoes an upfield and hyperbolic shift, although to a lesser degree (~0.10 ppm, purple trace, **Figure C.1b**). On the other hand, both G1 and U2 imino protons (blue and mustard yellow traces, **Figure C.1b**) undergo broadening and upfield shifting as GpppG is titrated into the duplex solution, although these changes are much smaller than those seen for the G7 imino proton. Since both G1 and U2 imino protons are on the blunt-end terminus of **P/T$_{2C}$** with no GpppG binding sites on the template, we hypothesize that these experimental observations likely arise from side-on, non-specific association of GpppG onto the blunt-end terminus of **P/T$_{2C}$**, and is likely unrelated to the Watson Crick based GpppG–**P/T$_{2C}$** binding of interest. Finally, the three "internal" imino protons (G3, U4, and U5; green, cyan, and crimson traces, **Figure C.1b**) only shifted upfield minimally (< 0.05 ppm), showing that the effect of GpppG binding does not propagate beyond two base pairs.

It is interesting to note that, as the concentration of GpppG in the duplex solution is increased beyond 6 mM (asterisked spectrum, **Figure C.1a**), all imino proton resonances begin to broaden into the baseline. We are reluctant to speculate on the factors or nonidealities that give rise to these observations; however, we suspect that when GpppG is in excess of available duplex binding sites, and when temperature is lowered, the nonspecific interactions between GpppG and surrounding **P/T$_{2C}$** duplexes

could be significant enough to be able to bridge multiple RNA duplexes noncovalently, to form a relatively large RNA-GpppG complex that undergoes slow tumbling on the NMR timescale.

C.4. K_M measurement for the 2-aminoimidazolium-bridged diguanosine dinucleotide (Gp-NH_2Im-pG)

C.4.1. Primer extension reaction.

Nonenzymatic template-directed RNA polymerization was monitored through primer extension assays. For accurate comparison between the RNA•GpppG and RNA•Gp-NH_2Im-pG, we used the same primer-template duplex as the one used in the NMR binding assay (Sequences: primer, 5′-FAM-CUCAAUG-3′; template, 5′-CCCAUUGAG-3′). The final concentrations of the primer extension reaction were 2 μM primer, 3 μM template, 200 mM Tris pH 8, 100 mM $MgCl_2$, and 0.125–2 mM of the Gp-NH_2Im-pG intermediate, adjusted for purity. After adding Gp-NH_2Im-pG to the reaction mixture to initiate primer extension, 1 μL reaction aliquots were removed at 1, 2, and 3 m and quenched in 7 μL of 8 M urea, 100 mM Tris-Cl, 100 mM boric acid, and 75 mM EDTA. Reaction aliquots were flash frozen on dry ice, and kept under –80 ˚C until analysis by 20% denaturing polyacrylamide gel electrophoresis (National Diagnostics, Atlanta, GA).[11] Gels were imaged on an Amersham Typhoon scanner from General Electric and quantified using the ImageQuantTL software (Little Chalfont, United Kingdom). Data were analyzed assuming pseudo-first order kinetics to determine the initial rate constant, k_{obs}.

C.4.2. Michaelis-Menten analysis of RNA:Gp-NH_2Im-pG binding

To compare the RNA binding affinity of GpppG with that of Gp-NH_2Im-pG, we sought to determine the K_M of this intermediate in primer extension reactions. We were able to obtain relatively pure Gp-NH_2Im-pG through the synthetic route laid out in **Section C.2.4**, *via* two rounds of purification, with ≥ 80% of the purified material in the form of imidazolium-bridged dinucleotide, and the remainder as the monomer (2-AmImpG). In addition, these fractions contained 1ed–2 equivalents of triethylamine. For our assays, we assumed that the RNA binding affinity of the intermediate would be greater than that of the monomer, and that the monomer would contribute insignificantly to the polymerization rate under these conditions.

For the primer extension assay, we used the same primer and template RNA sequences as that in the affinity measurement made by NMR. We observed that the buffer and pH conditions of the NMR studies (pH 7, no Mg^{2+}) were not compatible with our K_M determination because of the slow polymerization rate of primer extension. In addition, we observed that solutions containing even 100 mM NaCl resulted in precipitation of the intermediate and poor kinetic analysis of primer extension. Therefore, we used 200 mM Tris pH 8 and 100 mM $MgCl_2$ for our primer extension studies. We determined the initial rate of primer extension for 5 concentrations of the Gp-NH_2Im-pG and measured K_M = 619 ± 60 μM (**Figure C.2**). This value is 3x greater than the K_d = 175 μM of the GpppG analog by NMR, but is 27x less than the K_d = 17 mM of GMP. This suggests that the binding of the intermediate is better approximated by GpppG than GMP. The discrepancy between the K_M of the intermediate and the K_d of GpppG may be due to a variety of factors. For instance, the buffer conditions and pH of the two experimental systems differ. In

addition, the relationship between K_M and K_d may not be directly comparable. Therefore, we believe that these values are in rough agreement and support our proposal that GpppG is an analogue of the di-guanosine intermediate.

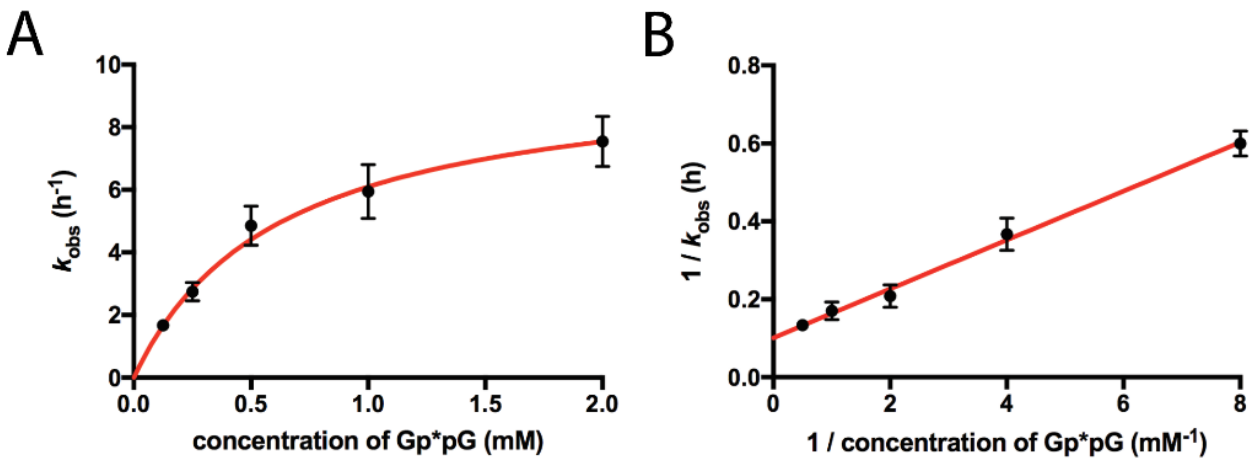


Figure C.2. Determination of the K_M of Gp-NH_2Im-pG dinucleotide by the primer extension assay. (A) Michaelis-Menten plot of the k_{obs} of primer extension versus the concentration of the di-guanosine intermediate. Black circles represent average experimental values in triplicate and error bars indicate ± 1 S.D. The red line is our fit of the experimental data using K_M = 0.619 mM and the maximum k_{obs} = 9.87 h^{-1}. (B) Lineweaver-Burke plot of the experimental data from part A. Black circles represent experimental data with error bars ± 1 S.D. The red line is the linear regression of this data.

C.5. X-ray crystallography.

C.5.1. Oligonucleotide purification and crystal preparation. The oligonucleotide used for crystallography was custom-synthesized by Exiqon Inc. Oligonucleotides for affinity measurements were prepared by solid-phase synthesis. Data was collected at the SIBYLS beamlines 8.2.1 and 8.2.2 at Lawrence Berkeley National Laboratory. The Nuc-Pro High Throughput Screen Kit (Jena Bioscience), Natrix High Throughput Kit and Index High Throughput Kit (Hampton Research, Aliso Viejo, CA) were used for screening crystallization conditions by the sitting drop vapor diffusion method. Solutions containing the RNA sample (0.5 mM) and different dinucleotide ligands (10 mM) were heated to 90 °C for 2 m, then cooled slowly to room temperature. All of the crystals grew at 18 °C, and the mother liquor containing 50% glycerol was used as a cryoprotectant during crystal mounting. All data collection was taken under a stream of nitrogen at 99 K. The data sets were collected at the SIBYLS beamline 821 and 822 at the Advanced Light Source, Lawrence Berkeley National Laboratory. The distances between the detector and the crystal were set to 200 mm and the collecting wavelength was set to 0.997 Å. The crystals were exposed for 1 second per image with one degree oscillations, and 180 images were taken for each data set. The optimized crystallization conditions for the RNA-monomer complexes are listed in **Table C.1**.

Table C.1. Optimized conditions for crystallization of RNA-dinucleotide complexes

	Optimized crystallization conditions
RNA–GpppG	0.05 M Magnesium chloride, 0.1 M Imidazole pH 6.5, 1.0 M Sodium acetate trihydrate
RNA–GppG	0.05 M Magnesium chloride, 0.2 M Ammonium citrate tribasic pH 7.0, 20% w/v Polyethylene glycol 3,350
RNA–pGpG	0.05 M Magnesium chloride, 0.1 M HEPES sodium pH 7.5, 2% v/v Polyethylene glycol 400, 2.0 M Ammonium sulfate
RNA–GppppG	0.05 M Magnesium chloride, 1.2 M Lithium sulfate, 50 mM MES pH 6.5, 2 mM Cobalt (II) chloride

C.5.3. Data collection and structure refinement.

Data sets were processed using HKL2000 and DENZO/SCALEPACK.[12] All structures were solved by molecular replacement, using structure of 5DHC as search model. All four structures were refined using Refmac.[13] The refinement protocol includes simulated annealing, positional refinement, restrained B-factor refinement, and bulk solvent correction.[13] During refinement, the topologies and parameters for locked nucleic acids (LCC) and for the ligands GpppG (GP3), GppG (GP2), GppppG (GP4) and dinucleotide intermediate (GIM) were constructed and applied. After several cycles of refinement, a number of highly ordered water molecules and magnesium ions were added. The 4 crystal structures were determined to resolutions of 1.9 Å, 1.5 Å, 2.6 Å and 2.1 Å, respectively (PDB ID: 5UEE, 5UED, 5UEG, 5UEF). Data collection, phasing, and refinement statistics of the determined structures are listed in **Tables C.2 & 3**.

Table C.2. Data collection statistics.

Structure	RNA–GpppG	RNA–GppG	RNA–GppppG	RNA–pGpG
Space group	$P3_121$	$P3_121$	P3	$P3_121$
Unit cell parameters, Å (°)	46.96, 46.96, 83.10 (90, 90, 120)	43.91, 43.91, 85.41 (90, 90, 120)	48.42, 48.42, 81.96 (90, 90, 120)	43.63, 43.63, 84.05 (90, 90, 120)
Resolution range, Å (last shell)	50-1.90 (1.97-1.90)	50-1.50 (1.55-1.50)	50-2.60 (2.69-2.60)	50-2.10 (2.18-2.10)
Unique reflections	8742	27933	12517	10461
Completeness, %	99.4 (94.4)	95.1 (100)	96.6 (74.8)	99.8 (99.6)
R_{merge}, %	6.3 (46.4)	7.9 (35.1)	10.0 (55.4)	7.4 (46.9)
$\langle I/\sigma(I)\rangle$	35.3 (1.9)	28.4 (5.5)	15.1 (1.3)	30.9 (2.8)
Redundancy	9.1 (4.5)	9.8 (7.6)	5.1 (3.6)	9.6 (7.0)

Table C.3. Structure refinement statistics.

Structure	RNA–GpppG	RNA–GppG	RNA–GppppG	RNA–pGpG
PDB code	5UEE	5UED	5UEG	5UEF
RNA duplex per asym-metric unit	1	1	2	1
Resolution range, Å	83.10-1.90	85.41-1.50	81.96-2.60	84.05-2.10
R_{work}, %	23.20	21.03	20.76	23.97
R_{free}, %	27.47	23.22	30.15	30.32
Number of reflections	8299	14252	6081	5391
Bond length R.M.S. (Å)	0.016	0.016	0.017	0.014
Bond angle R.M.S.	1.854	1.843	2.037	1.589
Average B-factors, (Å^2)	32.42	18.04	84.06	46.67

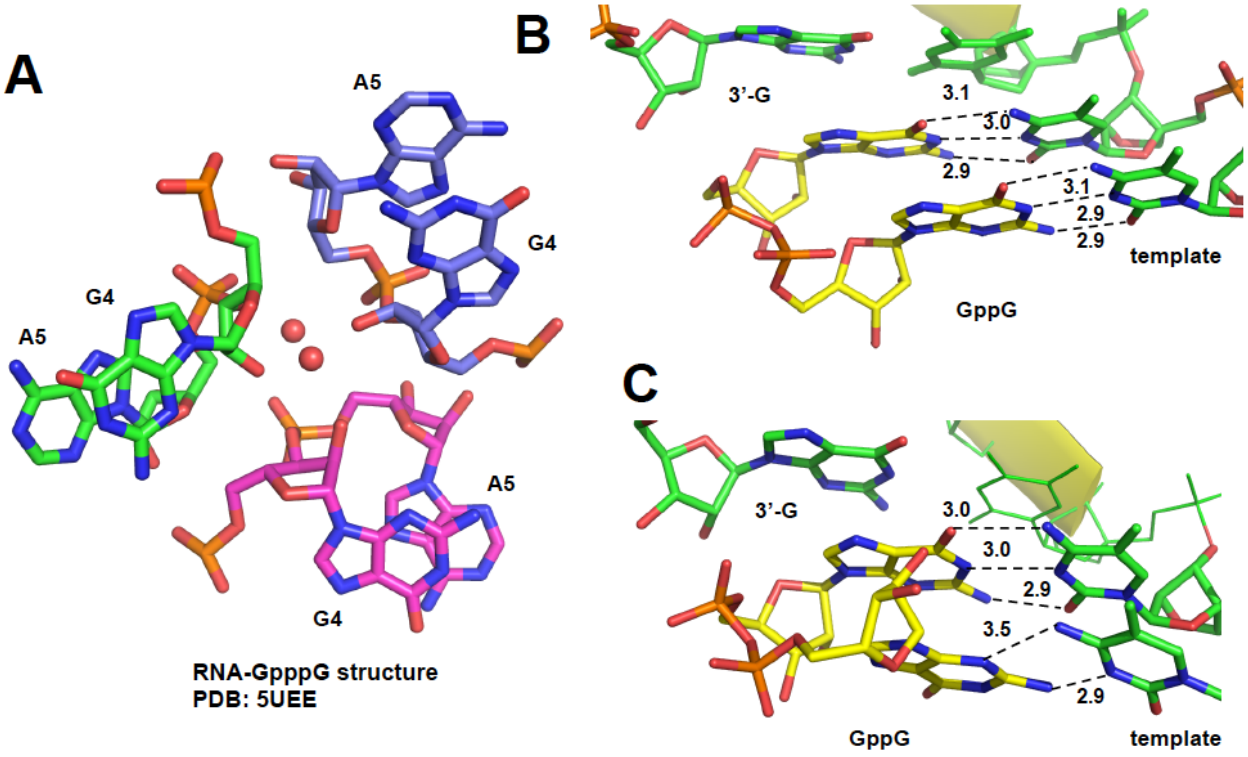


Figure C.3. (A) Local structure of RNA-GpppG complex. Two of the water molecules observed to bridge the three neighboring RNA duplexes interact by hydrogen bonding with 2'-hydroxyls of G4s and the G4-G5 phosphodiester phosphates. (B) and (C) Local structures of the RNA-GppG complex. At one end, GppG forms two Watson-Crick base pairs with the template, and at the other end, one Watson-Crick base pair and one noncanocial base pair are observed.

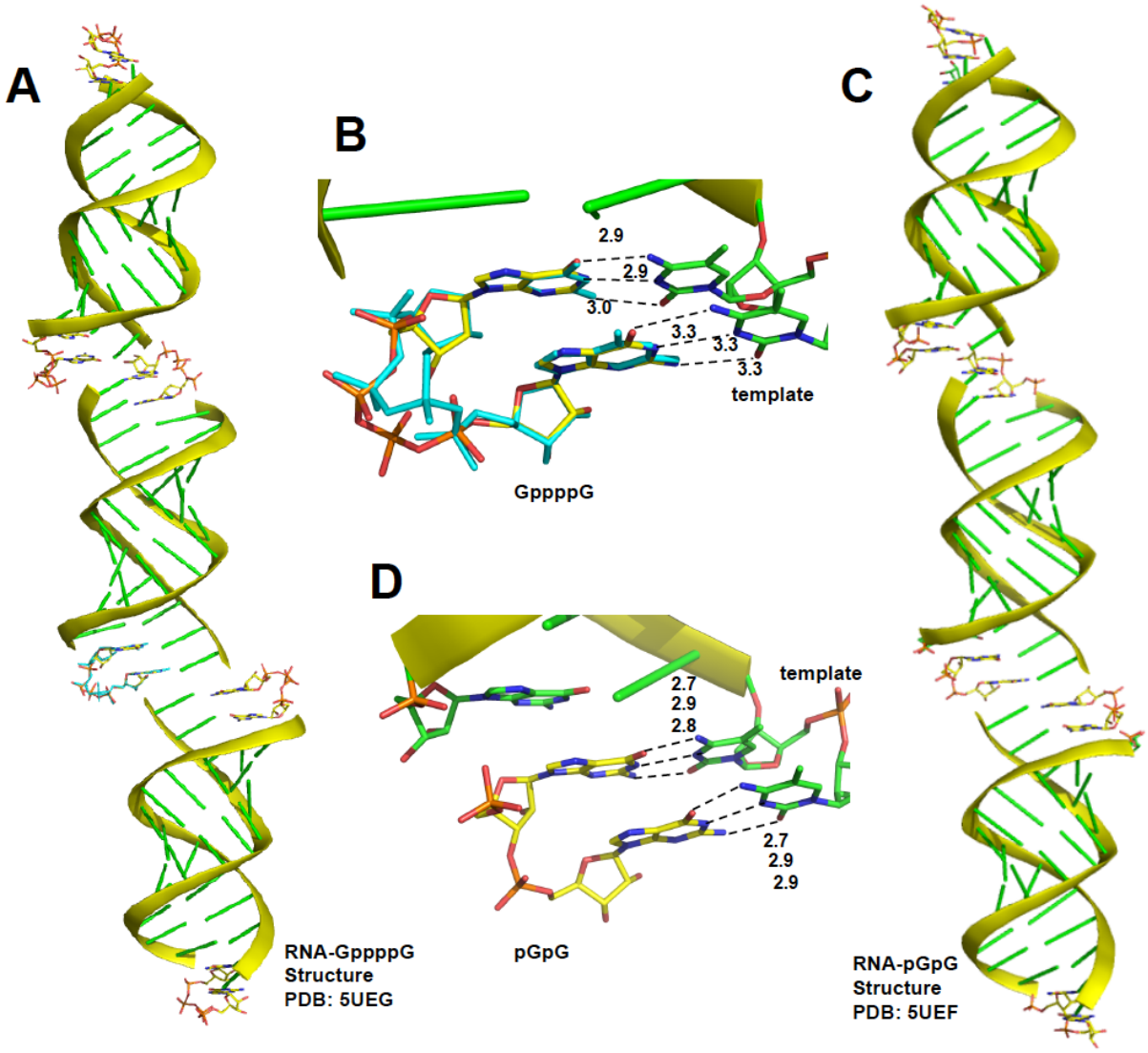


Figure C.4. (A) The RNA-GppppG complexes slip-stack to crystallize. (B) The GppppG ligand forms two Watson-Crick base pairs with the template. (C) The RNA-pGpG complexes slip-stack to crystallize. (D) The pGpG dimer forms two Watson-Crick base pairs with the template.

C.6. References.

(1) Piotto, M.; Saudek, V.; Sklenár, V. *J. Biomol. NMR* **1992**, *2*, 661–665.

(2) Lippens, G.; Dhalluin, C.; Wieruszeski, J. M. *J. Biomol. NMR* **1995**, *5*, 327–331.

(3) Cantor, C. R.; Warshaw, M. M.; Shapiro, H. *Biopolymers* **1970**, *9*, 1059–1077.

(4) Cavaluzzi, M. J.; Borer, P. N. *Nucleic Acids Res.* **2004**, *32*, e13.

(5) Tanaka, H.; Yoshimura, Y.; Jørgensen, M. R.; Cuesta-Seijo, J. A.; Hindsgaul, O. *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 11531–11534.

(6) Anslyn, E. V.; Dougherty, D. A. *Modern Physical Organic Chemistry*; University Science Books: Sausalito, CA, 2006.

(7) Tam, C. P.; Fahrenbach, A. C.; Björkbom, A.; Prywes, N.; Izgu, E. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 571–574.

(8) Conte, M. R.; Conn, G. L.; Brown, T.; Lane, A. N. *Nucleic Acids Res.* **1996**, *24*, 3693–3699.

(9) Lane, A. N.; Jenkins, T. C.; Frenkiel, T. A. *Biochim. Biophys. Acta.* **1997**, *1350*, 205–220.

(10) Izgu, E. C.; Fahrenbach, A. C.; Zhang, N.; Li, L.; Zhang, W.; Larsen, A. T.; Blain, J. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2015**, *137*, 6373–6382.

(11) Rio, D. C.; Ares Jr., M.; Hannon, G. J.; Nilsen, T. W. *Cold Spring Harb. Protoc.* **2010**, *2010*, pdb.prot5444.

(12) Otwinowski, Z.; Minor, W. *Methods Enzym.* **1997**, *276*, 307–326.

(13) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. *Acta Crystallogr. D Biol. Crystallogr.* **1997**, *53*, 240–255.

Appendix D

Supplemental Materials for Chapter 4

D.1 Synthesis of ICG, PYG and 2-MeImpddC: Experimental procedures and compound characterizations

D.1.1. 4-bromo-5-methyl-1-phenylsulfonylimidazole (Compound 2)[1]



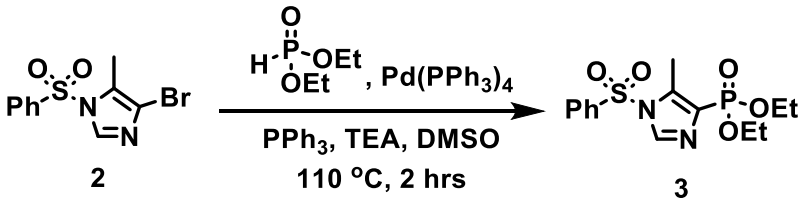
1 2

5-Bromo-4-methyl-1*H*-imidazole (compound **1**, 1 g, 6.21 mmol) was added to an oven-dried round bottom flask and stir bar, followed by addition of anhydrous dichloromethane (DCM, 25 mL, ~0.25 M). To the partially dissolved starting material in DCM (as a white suspension) was added 1,8-diazabicyclo(5.4.0)undec-7-ene (DBU) (1.04 mL, 6.83 mmol, 1.1 equiv., ρ = 1.018 g mL^{-1}) at ambient temperature, after which the suspension promptly clarified. Afterwards, benzenesulfonyl chloride (1.03 mL, 8.07 mmol, 1.3 equiv., ρ = 1.384 g mL^{-1}) was added dropwise. The reaction was stirred for 30 minutes at ambient temperature, after which the reaction turned bright yellow. The crude reaction mixture was partially evaporated *in vacuo*, diluted with ethyl acetate (EtOAc), and extracted with water for three times. The organic layer was dried over sodium sulfate and evaporated *in vacuo* to afford crude solid white residues with a yellow tint. The crude residue was purified by normal-phase flash chromatographic purification on a 40 g pre-packed silica column over 11 column volumes (CV) of 0–30%

EtOAc in hexanes. Titled compound **2** was eluted at ~18% EtOAc in hexanes, followed by evaporation *in vacuo* to afford a white solid residue (900 mg, 2.99 mmol, 48%).

^{1}H NMR ($CDCl_3$)	δ 8.05 (s, 1H), 7.94 – 7.87 (m, 2H), 7.77 – 7.68 (m, 1H), 7.65 – 7.56 (m, 2H), 2.25 (s, 3H).
^{13}C NMR ($CDCl_3$)	δ 137.35, 136.38, 135.29, 130.10, 127.79, 125.34, 118.50, 10.12.
TLC (UV; $KMnO_4$)	R_f: 0.30 (20% EtOAc in Hexanes)
HRMS	Calc'd for $C_{10}H_{10}BrN_2O_2S$ $[M+H]^+$: 300.9646; Found: 300.9648

D.1.2 Diethyl (5-methyl-1-(phenylsulfonyl)-imidazol-4-yl)phosphonate (Compound 3)[2]

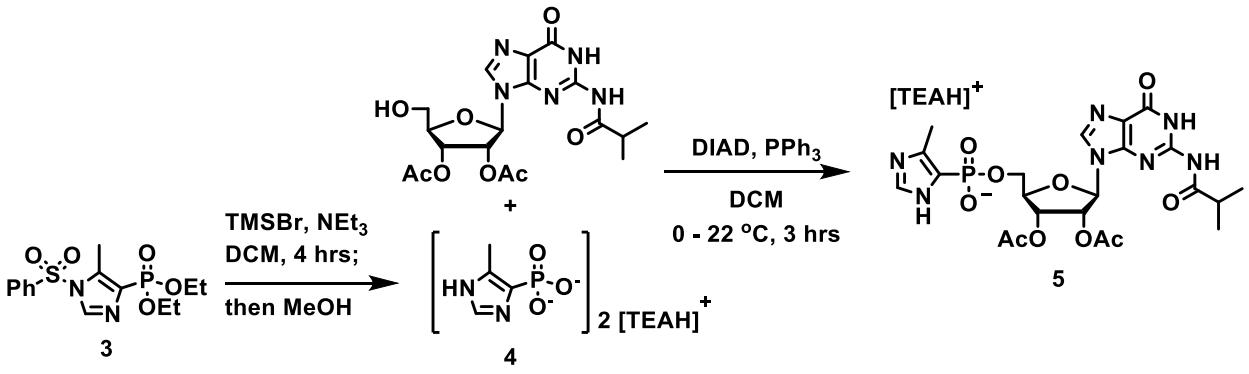


To a rigorously flame-dried round bottom flask and stir bar was added compound **2** (530 mg, 1.76 mmol), fresh tetrakis(triphenylphosphine)palladium(0) (305 mg, 0.264 mmol, 0.15 equiv.), and triphenylphosphine (PPh_3, 692 mg, 2.64 mmol, 1.5 equiv.). The flask was then evacuated and exchanged with dry argon three times. In a separate rigorously flame-dried pear-shaped flask, anhydrous dimethyl sulfoxide (18 mL, ~0.1 M), diethyl phosphite (490 μL, ~3.52 mmol, 2 equiv., ρ = 1.072 g mL^{-1}) and anhydrous triethylamine (TEA, 490 μL, ~3.52 mmol, 2 equiv., ρ = 0.726 g mL^{-1}) were sparged by a stream of dry argon for 30 min. The argon-purged reagent-solvent mixture was then cannula-transferred to the round bottom flask containing the starting material and reagents, followed by 2 h of heating at 110 °C. The reaction mixture slowly homogenized upon heating to form a golden yellow mixture. After 2 h, the reaction was

allowed to cool to ambient temperature, and diluted with 150 mL of EtOAc. The organic layer was washed with water and brine, dried over sodium sulfate and evaporated *in vacuo*. The resulting crude syrup was purified by normal-phase flash chromatography on a 40 g pre-packed silica column over 14 CV of solvents, with a linearly increasing gradient of 0–10% MeOH in CH_2Cl_2. 288 mg of starting material remained unreacted (54.3%), and the titled compound **3** was eluted from the column at ~6% MeOH in DCM, and evaporated *in vacuo* to afford a colorless syrup (224 mg, 0.625 mmol, 35%).

^{1}H NMR ($CDCl_3$)	δ 8.19 (d, *J* = 2.4 Hz, 1H), 7.96 – 7.88 (m, 2H), 7.78 – 7.71 (m, 1H), 7.63 – 7.59 (m, 2H), 4.23 – 4.02 (m, 4H), 2.56 (d, *J* = 2.3 Hz, 3H), 1.37 – 1.22 (t, *J* = 7 Hz, 6H).
^{13}C NMR ($CDCl_3$)	δ 138.24 (d, *J* = 22 Hz), 137.57 (d, *J* = 39 Hz), 137.06, 135.38, 130.06, 129.80 (d, *J* = 240 Hz), 127.88, 62.69 (d, *J* = 5.6 Hz), 16.31 (d, *J* = 7.7 Hz), 10.31.
^{31}P NMR ($CDCl_3$)	δ 10.07 (s)
TLC (UV; $KMnO_4$)	R_f : 0.34 (70% EtOAc in Hexanes)
HRMS (*m/z*)	Calc'd for $C_{14}H_{20}N_2O_5PS$ $[M+H]^+$: 359.0831; Found: 359.0841

D.1.3 *N*^{2}-isobutyryl-2′,3′-diacetylguanosine (4-methyl-1*H*-imidazol-5-yl)phosphonate, triethylammonium salt (Compound 5)



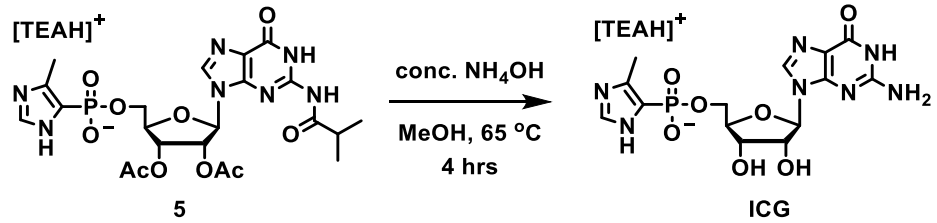
Step 1.[3] Prior to initiating this reaction, a round bottom flask containing compound **3** (370 mg, 1.03 mmol) and a stir bar was thoroughly dried under high vacuum overnight. Afterwards, anhydrous DCM (20.6 mL, 0.05 M) and TEA (360 μL, 2.58 mmol, 2.5 equiv., ρ = 0.726 g mL^{-1}) were added to the reaction vessel to dissolve **3** into a colorless homogeneous solution. Bromotrimethylsilane (TMSBr, 680 μL, ~5.17 mmol, 5 equiv., ρ = 1.16 g mL^{-1}) was then added dropwise at ambient temperature, and the reaction mixture was allowed to stir at room temperature overnight; a dense white smoke was observed when bromotrimethylsilane was added. The reaction mixture gradually changed from colorless to a deep and translucent brown color. To quench the excess bromotrimethylsilane and to decompose the resultant bis(trimethylsilyl)phosphonate esters,[4,5] an excess of methanol was added to the reaction mixture, and the pH of the mixture promptly dropped to < 1 (crudely determined by pH paper). Solvents were evaporated *in vacuo*, and the residual volatile impurities and hydrogen bromide were azeotropically removed by co-evaporation with toluene for three times. The resultant orange-red solid residue (compound **4**) was carried forward to step 2 without further purification.

Step 2.[6] Prior to initiating this reaction, the crude material (compound **4**, 100% conversion assumed; 3 equiv.) obtained from *step 1* was dried *in vacuo* under high vacuum overnight. The round bottom flask containing **4** was sequentially charged with N^2-isobutyryl-2′,3′-diacetylguanosine (151 mg, 0.344 mmol, 1 equiv.), PPh_3 (1.36 g, 5.17 mmol, 15 equiv.) and anhydrous DCM (21 mL, ~0.05 M) under an argon atmosphere. The reaction mixture was cloudy and reddish brown. Using an ice bath, the reaction was allowed to equilibrate to 0 °C for 5 min, after which diisopropyl azodicarboxylate (DIAD;

1.04 mL, 5.17 mmol, 15 equiv., ρ = 1.027 g mL^{-1}) was added dropwise to the reaction mixture over a duration of 10 min. The reaction mixture was observed to clarify into a translucent brown mixture as DIAD was added. The reaction was then allowed to stir for 3 h under ambient temperature. Subsequently, the reaction was stopped by evaporating the solvent *in vacuo* to form a brown syrup. The residue was crudely purified by normal phase flash chromatography with 12 CV of 0–20% methanol in DCM on a 40 g pre-packed silica column, followed by elution of crude compound **5** in 100% MeOH. The crude material was further purified on a prepacked 50 g C18Aq column, over 20 CV of 0–20% CH_3CN in TEAB. The titled compound **5** was eluted from the column at ~17.5% CH_3CN in TEAB, followed by lyophilization to afford **5** as a white fluffy solid. Yield was not determined at this step due to the large excess of triethylammonium bicarbonate in the sample. (*Note: To afford pure compound **5**, the eluted purified product should be promptly lyophilized. The acetyl protecting groups on the 2'- and 3'-hydroxyl groups were found to be labile under basic aqueous conditions.)*

^{1}H NMR (DMSO-d_6)	δ 8.20 (s, 1H), 7.42 (d, *J* = 1.8 Hz, 1H), 6.08 – 5.95 (m, 2H), 5.46 – 5.37 (m, 1H), 4.25 – 4.22 (m, 1H), 3.98 – 3.87 (m, 1H), 3.80 (dt, *J* = 10.7, 4.1 Hz, 1H), 2.92 (septet, *J* = 7.0 Hz, 1H), 2.72 (**TEAH$^+$**, q, *J* = 7.1 Hz), 2.23 (s, 3H), 2.06 (s, 3H), 1.93 (s, 3H), 1.08 (d, *J* = 7.0 Hz, 3H), 1.05 (d, *J* = 7.0 Hz, 3H), 1.01 (**TEAH$^+$**, t, *J* = 7.1 Hz).
^{13}C NMR (DMSO-d_6)	δ 180.94, 169.46, 169.04, 154.91, 148.63, 148.21, 139.60, 139.14 (br s), 134.96 (d, *J* = 15 Hz), 122.99 (d, *J* = 201 Hz), 121.05, 85.88, 82.64 (d, *J* = 8.5 Hz), 71.67 (d, *J* = 14 Hz), 62.97 (d, *J* = 4.1 Hz), 45.41 (**TEAH$^+$**), 34.38, 20.49, 20.16, 18.99, 18.83, 12.97, 9.77 (**TEAH$^+$**).
^{31}P NMR (CD_3OD)	δ 5.66 (s)
HRMS	Calc'd for $C_{22}H_{29}N_7O_{10}P$ $[M+H]^+$: 580.1714; Found: 582.1728

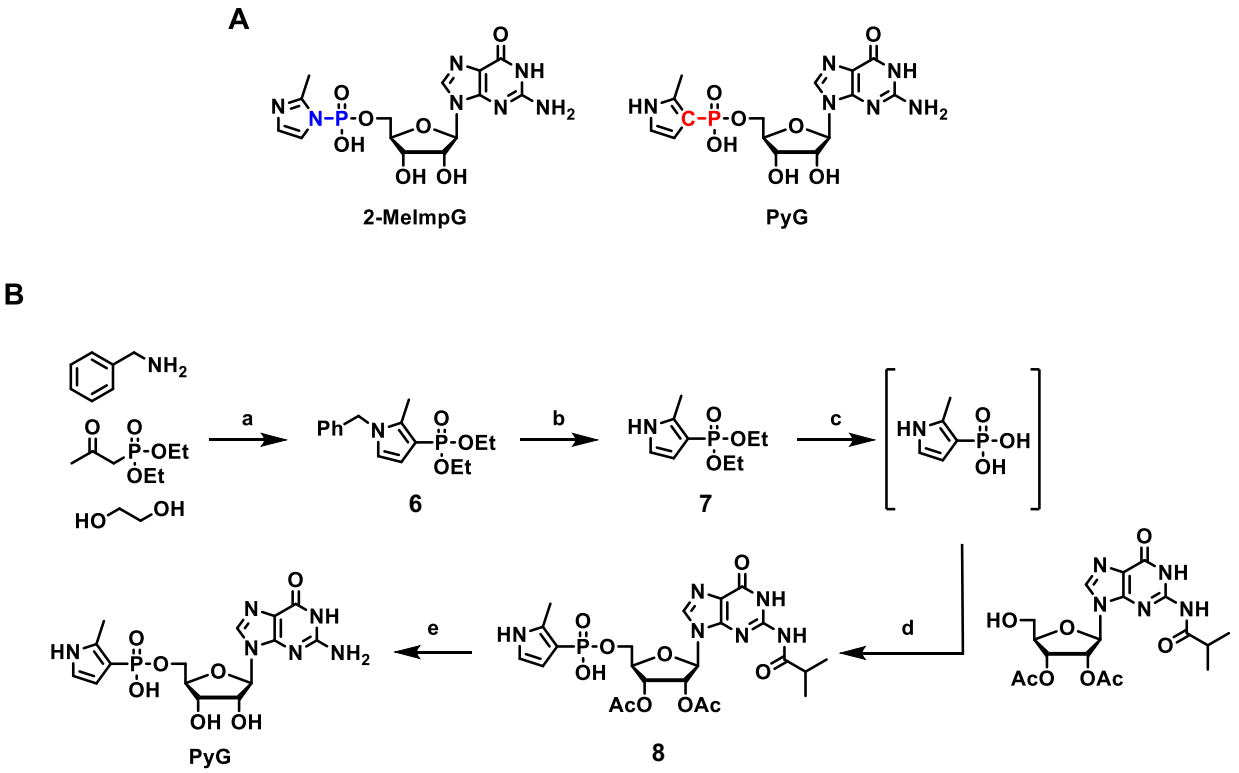
D.1.4 Guanosine (4-methyl-1*H*-imidazol-5-yl)phosphonate, sodium salt (ICG)



Compound **5** was charged to a heavy-walled high-pressure borosilicate glass reaction vessel along with a stir bar, and was then dissolved in a minimum amount of methanol. Afterwards, an excess of concentrated ammonium hydroxide (~20 mL) was added to the vessel. The vessel was sealed, heated to 65 °C with a pre-equilibrated oil bath, and stirred for 4 hours (*CAUTION: PRESSURIZED CONTENTS*) Afterwards, the reaction was allowed to equilibrate to room temperature (~0.5–1 h), followed by solvent evaporation *in vacuo* to afford a white solid deposit. The resulting crude material was resuspended in 250 mM TEAB, and the resulting suspension was filtered through a small cotton plug, and purified on a prepacked 50 g C18Aq column, over 20 CV of 0–15% CH_3CN in TEAB. ICG was eluted from the column at ~5% CH_3CN in TEAB, followed by lyophilization to afford the titled compound (in triethylammonium cation form) as a white fluffy solid. Sodium cation exchange was performed on the sample, by dissolving the sample in a minimum amount of deionized water, followed by re-precipitation in >25-fold excess (in volume) of 1:1:0.04 (v/v/v) acetone/diethyl ether/saturated $NaClO_4$ in acetone. The precipitate was centrifuged and washed with acetone twice. Final drying of the centrifuged and washed sample on high-vacuum furnished 34 mg (0.076 mmol) of ICG (Na^+ cation form) as fine white granular solids.

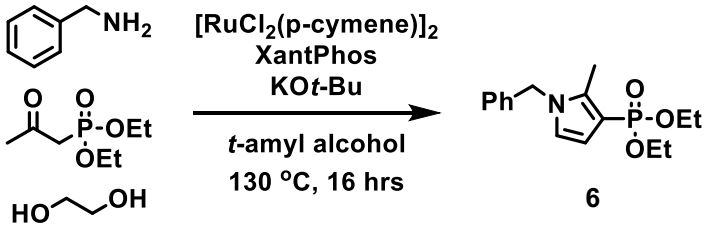
Final percentage yield of ICG was calculated by dividing the number of moles of isolated ICG•Na^+ by that of N^2-isobutyryl-2′,3′-diacetylguanosine, and was determined to be 22% over three steps.

^{1}H NMR (D_2O)	δ 7.87 (s, 1H), 7.63 (d, *J* = 2.3 Hz, 1H), 5.83 (d, *J* = 5.4 Hz, 1H), 4.73 (t, *J* = 5.3 Hz, 1H; partially overlapped with HOD peak), 4.40 (t, *J* = 4.5 Hz, 1H), 4.26 (q, *J* = 3.9 Hz, 1H), 4.00 (td, *J* = 6.3, 5.6, 3.8 Hz, 2H), 2.25 (d, *J* = 1.7 Hz, 3H)
^{13}C NMR (D_2O)	δ 154.64, 152.39, 140.70 (d, *J* = 25 Hz), 138.32, 137.30 (d, *J* = 16 Hz), 122.50 (d, *J* = 219 Hz), 117.04, 88.29, 84.75 (d, *J* = 8.7 Hz), 74.55, 71.50, 64.82 (d, *J* = 4.6 Hz), 12.30.
^{31}P NMR (D_2O)	δ 7.21 (s)
HRMS	Calc'd for $C_{20}H_{34}N_8O_7P$ $[M–H]^–$: 426.0933; Found: 426.0922



Scheme D.1. Synthesis of PyG.[7] (A) Structures of 2-MeImpG and PyG analog compared side-by-side. On PyG, the reactive nitrogen-phosphorous linkage of 2-MeImpG (labelled blue) is swapped with a chemically inert carbon-phosphorus linkage (labeled red). (B) Synthetic scheme of PyG. Reaction conditions: (a) $[RuCl_2(p\text{-cymene})]_2$, XantPhos, *t*-BuOK, *t*-amyl alcohol, 130 °C, 16 h; (b) Na^0, NH_3 (l), -78 °C, 30 min; (c) TMSBr, Et_3N, DCM, 4 h; then MeOH; (d) DIAD, PPh_3, DCM, 3 h; (e) conc. aq. ammonium hydroxide, 65 °C, 2.5 h. $[RuCl_2(p\text{-cymene})]_2$, dichloro(*p*-cymene)ruthenium(II) dimer; XantPhos, 4,5-bis(diphenylphosphino)-9,9-dimethylxanthene.

D.1.5. Diethyl (*N*-benzyl-2-methyl-3-pyrrolyl)phosphonate (Compound 6)



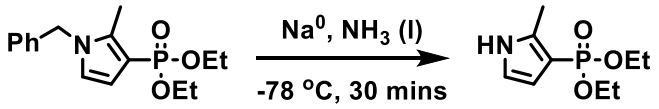
6

This procedure was adapted from Beller and coworkers.[8] Prior to reaction setup, the high-pressure glass reaction vessel was thoroughly inspected to ensure absence of any star-shaped glass cracks, followed by overnight drying in a 150 °C oven. In the septa-sealed and oven-dried glass vessel was charged a dessicator-dried magnetic stir bar, dichloro(p-cymene)ruthenium(II) dimer ($[RuCl_2(p\text{-cymene})]_2$, 91.9 mg, 0.15 mmol, 0.05 equiv.), 4,5-bis(diphenylphosphino)-9,9-dimethylxanthene (XantPhos, 173.6 mg, 0.3 mmol, 0.1 equiv.), and potassium *tert*-butoxide (67.3 mg, 0.6 mmol, 0.2 equiv.), followed by at least three cycles of argon purging. Under the argon atmosphere, *tert*-amyl alcohol (6 mL, 0.5 M), benzylamine (482 µL, 4.5 mmol, 1.5 equiv., ρ = 0.98 g/mL), ethylene glycol (370 µL, 6.6 mmol, 2.2 equiv., ρ = 1.11 g/mL) and diethyl (2-oxopropyl)phosphonate (370 µL, 3 mmol, 1 equiv., ρ = 1.01 g/mL) was successively delivered to the reaction vessel via needle injections. The reaction vessel was then sealed by a Teflon screw cap, and the reaction was allowed to stir for 16 hours at 130 °C. (*CAUTION: CONTENTS UNDER PRESSURE. Placement of a weighted safety shield in front of the reaction is recommended.*). The reaction mixture turned dark brick-red upon homogenization of solid reagents in *tert*-amyl alcohol. Afterwards, the solvent of the crude reaction mixture was evaporated *in vacuo*. The crude mixture was purified

by normal-phase flash chromatography on a 40 g pre-packed silica column over 20 column volume (CV, 960 mL) of solvents, with a linearly increasing gradient of 0–10% MeOH in CH_2Cl_2. The titled compound (224 mg, 0.729 mmol, 24.3 %.) was eluted from the column at ~5% MeOH in DCM as a dark red oil.

^{1}H NMR ($CDCl_3$)	δ 7.34 – 7.29 (m, 3H), 7.00 (d, *J* = 4 Hz, 2H), 6.64 – 6.62 (dd, *J* = 4.4, 3.0 Hz, 1H), 6.40 – 6.38 (dd, *J* = 4.4, 3.0 Hz, 1H), 5.04 (s, 2H), 4.14 – 4.01 (m, 4H), 2.37 (d, *J* = 1.8 Hz, 3H), 1.31 (t, *J* = 7.1 Hz, 6H)
^{13}C NMR ($CDCl_3$)	δ 136.9, 136.4 (d, *J* = 42 Hz, 1C), 128.8, 127.7, 126.4, 121.7 (d, *J* = 24 Hz, 1C), 111.7 (d, *J* = 19 Hz, 1C), 105.4 (d, *J* = 346 Hz, 1C), 61.3 (d, *J* = 9.7 Hz, 1C), 50.4, 16.3 (d, *J* = 9.7 Hz, 1C), 11.14
^{31}P NMR ($CDCl_3$)	δ 19.76 (s)
TLC ($KMnO_4$, I_2)	R_f : 0.53 (5% MeOH in DCM) (very weak absorption in 254 nm UV)
HRMS	Calc'd for $C_{16}H_{23}NO_3P$ $[M+H]^+$: 308.1416; Found: 308.1436

D.1.6. Diethyl (2-methyl-1*H*-pyrrol-3-yl)phosphonate (Compound 7)[9]



6 **7**

a. Prior Preparations. The Schlenk tube and stir bar used for this reaction was rigorously flame-dried. Additionally, a 0.25 M THF solution (1.3 mL) of phosphonate **1** (100 mg, 0.325 mmol, 1 equiv.) was prepared in a separate, oven-dried pear-shaped flask with standard air- and moisture free techniques. Finally, for delivery of ammonia gas into the Schlenk tube, a gas-delivery assembly was assembled immediately prior to the reaction, which comprised of a thick-walled steel cylinder containing pressurized ammonia gas, an Aldrich® corrosive lecture-bottle gas regulator (St. Louis, MO) and a

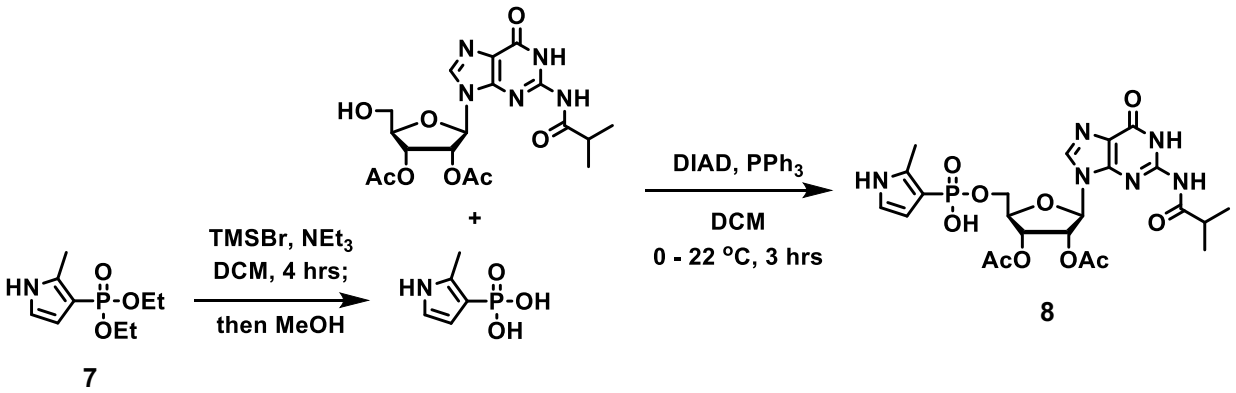
tygon tube/12-inch,16-gauge long needle assembly. All joints of the assembly were sealed with either Teflon tape or parafilm. The gas-delivery assembly was securely fastened to a stainless-steel frame at the back of a fume hood.

b. Reaction setup. The flame-dried Schlenk tube-stir bar combination was sealed with an appropriately-sized rubber septum, and was allowed to cool to room temperature while the atmosphere within was exchanged with argon for at least three times. Under an argon atmosphere, the Schlenk tube was cooled in a -78 ^{o}C acetone-dry ice bath, and dry ammonia gas was delivered directly into the Schlenk tube from a pressure-regulated lecture bottle via the tygon tube-long needle assembly. Liquid ammonia condensation was allowed to proceed until the desired volume of ammonia was obtained (~8 mL, ~0.04 M). In a separate beaker, small pieces of metallic sodium (40 mg, 1.74 mmol, 5.4 equiv.) were first washed with fresh, anhydrous hexanes to wash off the paraffin oil coating, followed by immediate transfer of the washed sodium chunks into the condensed liquid ammonia. As metallic sodium dissolves in liquid ammonia, the colorless liquid ammonia promptly turned into a deep navy blue mixture, which is characteristic of excess sodium. When the deep navy blue color persisted, the 0.25 M THF solution of compound **1** was slowly added dropwise into the sodium/ammonia mixture, which was then allowed to react at -78 ^{o}C for 30 mins. To quench the reaction, solid ammonium chloride (350 mg, 6.5 mmol, 20 equiv.) was added to the reaction mixture, and the deep naval blue color immediately dissipated upon addition. The quenched reaction mixture was then allowed to gradually warm up to room temperature over the course of four hours, while the liquid ammonia was allowed to evaporate under a gentle stream of argon. The crude residue was diluted with water and methanol, and

the resulting aqueous solution was washed with dichloromethane. The combined organic layer was dried with anhydrous sodium sulfate, followed by solvent evaporation in vacuo. The crude mixture was purified by normal-phase flash chromatography on a 12 g pre-packed silica column over 20 CV (336 mL) of solvents, with a linearly increasing gradient of 0–20 % MeOH in CH_2Cl_2. The titled compound (60 mg, 0.28 mmol, 86 %) was eluted from the column at ~10 % MeOH in DCM as a pale-yellow oil.

^{1}H NMR ($CDCl_3$)	δ 10.2 (br s, 1H), 6.60 – 6.58 (m, 1H), 6.30 – 6.27 (m, 1H), 4.08 – 3.93 (m, 4H), 2.38 (s, 3H), 1.27 (t, *J* = 7.1 Hz, 6H)
^{13}C NMR ($CDCl_3$)	δ 136.4 (d, *J* = 42 Hz, 1C), 117.4 (d, *J* = 26 Hz, 1C), 112.2 (d, *J* = 23 Hz, 1C), 102.8 (d, *J* = 348 Hz, 1C), 61.5 (d, *J* = 8.1 Hz, 1C), 16.3 (d, *J* = 11 Hz, 1C), 12.61
^{31}P NMR ($CDCl_3$)	δ 21.35 (s)
TLC ($KMnO_4$, I_2)	R_f : 0.46 (10% MeOH in CH_2Cl_2)
HRMS (*m/z*)	Calc'd for $C_9H_{17}NO_3P$ $[M+H]^+$: 218.0946; Found: 218.0943

D.1.7. *N*^{2}-isobutyryl-2′,3′-diacetylguanosine (2-methyl-1*H*-pyrrol-3-yl)phosphonate (Compound 8)



Step 1.[3] A round bottom flask containing compound **2** (112.8 mg, 0.519 mmol) and a stir bar was thoroughly dried under high vacuum overnight prior to initiating this reaction. Afterwards, anhydrous CH_2Cl_2 (5.2 mL, 0.1 M) and triethylamine (362 µL, 2.60 mmol, 5

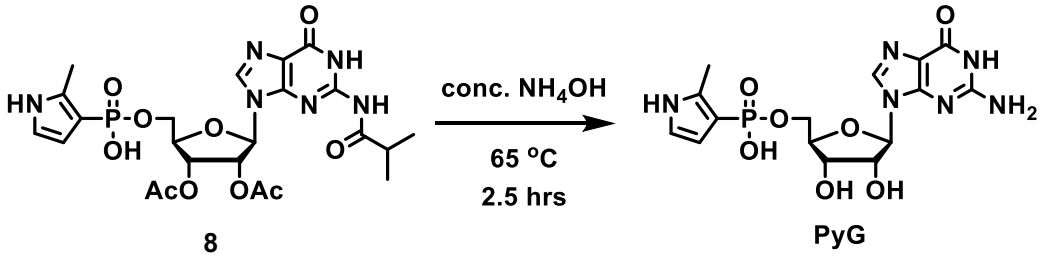
equiv., ρ = 0.726 g/mL) was added to the reaction vessel to dissolve **2** into a yellow homogeneous solution. Bromotrimethylsilane (685 µL, 5.19 mmol, 10 equiv., ρ = 1.16 g/mL) was then added to the stirring mixture dropwise under ambient temperature, and the reaction mixture was allowed to stir for 4 hrs. A dense white smoke was observed when bromotrimethylsilane was added, and the reaction mixture turned brown upon stirring. To quench the excess bromotrimethylsilane and to decompose the resultant bis(trimethylsilyl)phosphonate esters,[4,5] an excess of methanol was added to the reaction mixture. Solvents were evaporated *in vacuo*, and the residual volatile impurities and hydrogen bromide were azeotropically removed by toluene co-evaporation in three repetitions. The resultant solid residue (compound **3**) was carried forward to step 2 without further purification.

Step 2.[6] The crude material (compound **3**, 100% conversion assumed; 1.1 equiv) obtained from step 1 was dried *in vacuo* under high vacuum overnight prior to initiating this reaction. Afterwards, the round bottom flask containing **3** was back-filled with dry argon, followed by sequential addition of N^2-isobutyryl-2′,3′-diacetylguanosine (204 mg, 0.466 mmol, 1 equiv.; obtained from ChemGenes, Wilmington, MA), triphenylphosphine (681 mg, 2.60 mmol, 5.5 equiv.) and anhydrous CH_2Cl_2 (10.4 mL, ~0.05 M) under an argon atmosphere. The mixture was non-homogeneous and brown. Using an ice bath, the reaction was allowed to equilibrate to 0 °C for 5 minutes, after which diisopropyl azodicarboxylate (DIAD; 525 µL, 2.60 mmol, 5.5 equiv., ρ = 1.027 g/mL) was syringe-added to the mixture dropwise over a five-minute interval. The reaction mixture was observed to clarify into a clear brown mixture as DIAD was added. The reaction was stirred at 0 °C for 30 mins and at ambient temperature for 2.5 hrs. The reaction was

stopped by evaporating the solvent *in vacuo*, followed by normal-phase flash chromatography in a manufacturer-packed 40 g silica column pre-equilibrated with dichloromethane. The column was first flushed with 10 CV of 9:1 DCM-MeOH mixture (with 1% triethylamine) to remove non-polar impurities, followed by product elution with 5 CV of 100% MeOH (with 1% triethylamine). The dried-down crude product was further purified with reverse-phase flash chromatography by suspending the residue in 9:1 water-acetonitrile mixture, followed by direct liquid-loading into a manufacturer-packed 50 g C18Aq column pre-equilibrated with 25 mM pH 7.5 aqueous triethylammonium bicarbonate (TEAB) buffer. The resulting crude residue was purified over 25 CV of solvents (~ 1060 mL) with a linearly increasing gradient of 0–25 % CH_3CN in TEAB. The title compound was eluted from the column at ~18 % CH_3CN in TEAB, the relevant fractions were pooled, and the yield of this reaction was determined using Thermo Scientific Nanodrop 2000c Spectrophotomer (Waltham, MA) to be 36% (using the extinction coefficient of 12.08 L/mmol/cm for guanosine 5´-monophosphate[4]). The solution was then lyophilized to afford the titled compound as a fluffy white solid.

^{1}H NMR (CD_3OD)	δ 8.17 (s, 1H), 6.52 (dd, *J* = 3.5, 3.0 Hz, 1H), 6.28 (dd, *J* = 3.8, 2.9 Hz, 1H), 6.12 – 6.04 (m, 2H), 5.59 (dd, *J* = 5.3, 2.5 Hz, 1H), 4.32 – 4.30 (m, 1H), 4.07 – 4.02 (m, 1H), 3.97 – 3.92 (m, 1H), 2.83 (**TEAH**$^{+}$, q, *J* = 7.2 Hz, 6H), ~2.8 (septet, 1H, overlapped with **TEAH**$^{+}$), 2.37 (d, *J* = 1.6 Hz, 3H), 2.10 (s, 3H), 1.97 (s, 3H), 1.20 – 1.17 (m, 6H, partially overlapped with **TEAH**$^{+}$), 1.15 (**TEAH**$^{+}$, t, *J* = 7.2 Hz, 9H)
^{13}C NMR (CD_3OD)	δ 182.23, 171.3, 170.8, 157.7, 150.7, 150.0, 140.7, 134.3 (d, *J* = 37 Hz, 1C), 121.8, 116.7 (d, *J* = 24 Hz, 1C), 113.3 (d, *J* = 19 Hz, 1C), 110.4 (d, *J* = 328 Hz, 1C), 87.4, 84.1 (d, *J* = 14 Hz, 1C), 74.1, 73.3, 64.4 (d, *J* = 6.4 Hz, 1C), 47.2 (**TEAH**$^{+}$), 36.7, 20.5, 20.2, 19.5, 19.4, 13.0, 10.2 (**TEAH**$^{+}$)
^{31}P NMR (CD_3OD)	δ 16.26 (s)
HRMS	Calc'd for $C_{23}H_{30}N_6O_{10}P$ $[M-H]^+$: 581.1761; Found: 581.1786

D.1.8. Guanosine (2-methyl-1*H*-pyrrol-3-yl)phosphonate (PyG)

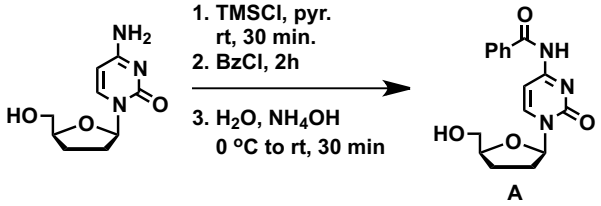


Compound **4** (33 mg, 0.057 mmol, 1 equiv.) was added to a heavy-walled and round-bottomed high-pressure glass reaction vessel along with a stir bar, and was then dissolved in concentrated ammonium hydroxide (28% NH_3 in H_2O; 1.14 mL) was added to the vessel. The vessel was sealed, heated to 65 °C with a pre-equilibrated oil bath, and stirred for 2.5 hours (*CAUTION: CONTENTS UNDER PRESSURE. The high-pressure reaction vessel should be thoroughly inspected to ensure absence of any star-shaped glass cracks*). Afterwards, the reaction was allowed to equilibrate to room

temperature, followed by solvent evaporation *in vacuo* to afford a white solid deposit. The crude material was resuspended in 100 mM triethylammonium bicarbonate buffer (pH 7.5), followed by direct liquid-loading into a manufacturer-packed 50 g C18Aq column pre-equilibrated with 25 mM pH 7.5 aqueous triethylammonium bicarbonate (TEAB) buffer. The resulting crude residue was purified over 10 CV (425 mL) of solvents, with a linearly increasing gradient of 0–15% CH_3CN in TEAB. The title compound was eluted from the column at ~6.5% CH_3CN in TEAB to afford the titled compound in triethylammonium cation form. The yield of **PyG**, based on A_{260} absorbance measurement of the pooled fractions on Nanodrop spectrophotometer, was determined to be 83 %. The fractions were then lyophilized to afford a white fluffy solid. Sodium cation exchange was performed on the sample, by dissolving the sample in a minimum amount of deionized water, followed by re-precipitation in >25-fold excess (in volume) of 1:1:0.04 (v/v/v) acetone : diethyl ether : saturated $NaClO_4$ in acetone. The precipitate was centrifuged and washed with 1:1 acetone: diethyl ether twice. Final drying of the centrifuged and washed sample on high-vacuum furnished **PyG** (Na^+ cation form) as fine white granular solids.

^{1}H NMR (D_2O)	δ 7.88 (s, 1H), 6.65 (t, 1H, J = 3:2 Hz), 6.23 (t, 1H, J = 3:2 hz), 5.87 (d, 1H, J = 5:6 Hz), 4.76 (t, 1H, J = 5:2 Hz), 4.41 (m, 1H), 4.27 (m, 1H), 3.93 (t, 2H, J = 4:0 Hz), 2.22 (s, 3H).
^{13}C NMR (D_2O)	δ 159.6, 154.5, 152.4, 138.2, 134.7 (d, J = 23:0 Hz), 117.2 (d, J = 15:0 Hz), 116.8, 112.3 (d, J = 12:9 Hz), 109.2, 87.6, 84.9 (d, J = 8:9 Hz), 74.2, 71.5, 64.1 (d, J = 3:5 Hz), 12.4.
^{31}P NMR (D_2O)	δ 16.92 (s)
HRMS	Calc'd for $C_{15}H_{18}N_6O_7P$ $[M–H]^-$: 425.0980; Found: 425.0993

D.1.9 *N*-benzoyl-2′,3′-dideoxycytosine[10]

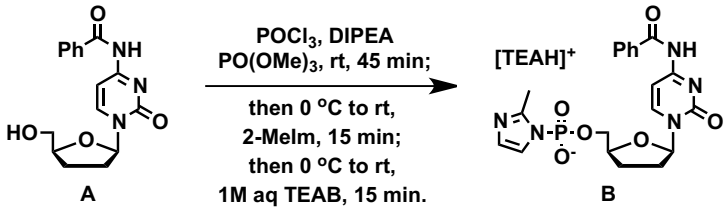


A

To an oven-dried round-bottomed flask was added zalcitabine (2′,3′-dideoxycytosine, 410 mg, 1.94 mmol) and anhydrous pyridine (pyr., 81 mL, 0.024 M). To the *non-homogeneous* mixture was then added chlorotrimethylsilane (TMSCl, 1.73 mL, 13.6 mmol, 7 equiv., ρ = 0.856 g mL^{-1}) at room temperature, followed by rigorous stirring and ultrasound sonication for 20 min. The non-homogeneous mixture clarified. Benzoyl chloride (BzCl, 1.24 mL, 10.7 mmol, 5.5 equiv., ρ = 1.211 g mL^{-1}) was then added dropwise over a period of 5 min at room temperature, and the mixture was stirred for 2 h under ambient temperature. The homogeneous reaction mixture turned light yellow. Afterwards, the reaction was cooled to 0 °C in an ice-water bath, followed by dropwise addition of water (12 mL) and concentrated aqueous ammonia (30 mL). The quenched reaction mixture was stirred at room temperature for 30 min, followed by solvent removal *via* extended rotary evaporation (> 30 min). Half of the final residue was resuspended in a minimum amount of 2:1 water:DMSO, followed by direct reverse-phase flash chromatography on a prepacked 50 g C18Aq column, over 15 CV of 0–40% CH_3CN in TEAB. The titled compound was eluted from the column at ~30% CH_3CN in TEAB. Solvent evaporation *in vacuo* of all relevant fractions afforded the titled compound **A** (400 mg; 1.27 mmol, 65.4 %) as a white fluffy solid.

^{1}H NMR (DMSO-d_6)	δ 8.55 (d, *J* = 7.4 Hz, 1H), 8.01 (d, *J* = 7.7 Hz, 2H), 7.66 – 7.56 (m, 1H), 7.50 (t, *J* = 7.6 Hz, 2H), 7.33 (d, *J* = 7.4 Hz, 1H), 5.97 (dd, *J* = 6.8, 2.4 Hz, 1H), 4.12 (ddt, *J* = 9.4, 6.4, 3.5 Hz, 1H), 3.78 (dd, *J* = 12.1, 3.4 Hz, 1H), 3.60 (dd, *J* = 12.1, 3.7 Hz, 1H), 2.40 (ddt, *J* = 14.2, 10.7, 7.4 Hz, 1H), 2.01 (ddt, *J* = 13.4, 7.3, 2.9 Hz, 1H), 1.93 – 1.70 (m, 2H).
^{13}C NMR (DMSO-d_6)	δ 167.35, 162.83, 154.42, 145.07, 133.24, 132.66, 128.43, 95.41, 87.02, 82.79, 61.56, 61.47, 32.93, 24.00.
HRMS	Calc'd for $C_{20}H_{34}N_8O_7P$ $[M+H]^+$: 316.1297; Found: 316.1295

D.1.10 *N*-benzoyl-2′,3′-dideoxycytosine 5′-phosphoro-2-methylimidazolide, triethylammonium salt[11]

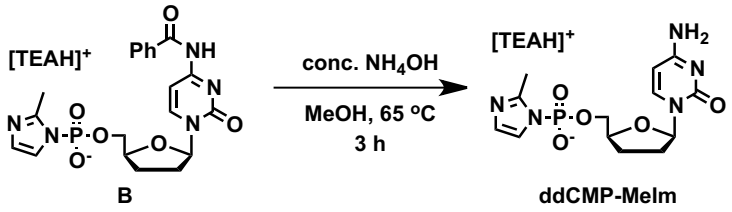


To a rigorously flame-dried and argon-exchanged round-bottomed flask was added compound **A** (100 mg, 0.317 mmol) and anhydrous trimethyl phosphate ($PO(OMe)_3$, 6.4 mL, 0.05 M; *Note: $PO(OMe)_3$ was dried over activated 4 Å molecular sieves overnight*). Compound **A** partially dissolved in $PO(OMe)_3$ to form a translucent mixture. Diisopropylethylamine (DIPEA, 85 μL, 0.48 mmol, 1.5 equiv., ρ = 0.742 g mL^{-1}) was added, followed by *slow*, *dropwise* addition of phosphorus oxychloride ($POCl_3$, 35 μL, 0.35 mmol, 1.1 equiv., ρ = 1.645 g mL^{-1}) at room temperature. A tinge of golden yellow color could be observed at the end of reagent addition. The solution was stirred at room temperature for 15 minutes, followed by two more rounds of DIPEA and $POCl_3$ additions. The reaction mixture appeared golden yellow. Afterwards, the reaction mixture was cooled to 0 °C in an ice-water bath, followed by addition of 2-methylimidazole (2-MeIm, 130.1 mg, 1.585 mmol, 5 equiv.). The reaction mixture was stirred at room temperature for 15 minutes. The reaction was cooled to 0 °C again before the addition of 500 mM aqueous TEAB (24 mL) as the final quenching reagent, after which the mixture clarified into a lightly yellowish solution. Effervescence was observed upon TEAB addition. The crude mixture was degassed by sonication, and purified by reverse-phase flash chromatography on a prepacked 50 g C18Aq column, over 25 CV of 0–25% CH_3CN in TEAB. The titled compound was eluted from the

column at ~18% CH_3CN in TEAB. Relevant fractions were pooled and lyophilized to afford the titled compound **B** as a white solid.

^{1}H NMR (D_2O)	δ 8.38 (d, *J* = 7.5 Hz, 1H), 8.03 – 7.99 (m, 2H), 7.65 – 7.60 (m, 1H), 7.54 – 7.49 (m, 3H), 7.31 (d, *J* = 7.4 Hz, 1H), 7.14 (t, *J* = 1.5 Hz, 1H), 6.75 (t, *J* = 1.5 Hz, 1H), 5.95 (dd, *J* = 6.7, 2.8 Hz, 1H), 4.15 (dt, *J* = 9.3, 5.1 Hz, 1H), 3.89 (ddd, *J* = 11.6, 5.9, 3.2 Hz, 1H), 3.76 (ddd, *J* = 11.4, 6.8, 4.3 Hz, 1H), 2.98 (**TEAH$^+$**, q, *J* = 7.3 Hz, 7H), 2.45 (s, 3H), 2.41 – 2.33 (m, 1H), 1.97 (ddt, *J* = 13.4, 7.7, 3.1 Hz, 1H), 1.87 (dddd, *J* = 12.4, 7.6, 5.8, 3.2 Hz, 1H), 1.75 – 1.62 (m, 1H), 1.14 (**TEAH$^+$**, t, *J* = 7.3 Hz, 11H)
^{13}C NMR (D_2O)	δ 167.40, 162.85, 154.32, 146.39 (d, *J* = 4.2 Hz), 144.93, 133.27, 132.64, 128.42, 124.36 (d, *J* = 9.5 Hz), 121.92 (d, *J* = 5.2 Hz), 120.94, 95.63, 87.03, 80.67 (d, *J* = 8.0 Hz), 65.35 (d, *J* = 6.1 Hz), 45.52 (**TEAH$^+$**), 32.62, 24.31, 14.57, 8.93 (**TEAH$^+$**).
^{31}P NMR (D_2O)	δ –8.85 (s)
HRMS	Calc'd for $C_{20}H_{34}N_8O_7P$ $[M–H]^-$: 458.1235; Found: 458.1245

D.1.11 2′,3′-dideoxycytosine 5′-phosphoro-2-methylimidazolide, triethylammonium salt (2-MeImpddC)



Compound **B** was charged to a heavy-walled high-pressure borosilicate glass reaction vessel eqipped with a stir bar, and was dissolved in a minimum amount of methanol. Afterwards, an excess of concentrated ammonium hydroxide (28% NH_3 in H_2O; ~10 mL) was added to the vessel. The vessel was sealed, heated to 65 °C with a pre-equilibrated oil bath, and stirred for 3 hours (*CAUTION: PRESSURIZED CONTENTS*) Afterwards, the reaction was allowed to equilibrate to room temperature (~0.5–1 h),

followed by solvent evaporation *in vacuo* to afford a white solid deposit. The resulting crude material was resuspended with 250 mM TEAB and purified on a prepacked 50 g C18Aq column, over 20 CV of 0–15% CH_3CN in TEAB. The titled compound was eluted from the column at ~5% CH_3CN in TEAB. Relevant fractions were pooled and lyophilized to afford the titled compound (in triethylammonium cation form) as a white solid.

^{1}H NMR (D_2O)	δ 7.73 (dd, *J* = 7.4, 1.3 Hz, 1H), 7.08 (q, *J* = 1.3 Hz, 1H), 6.66 (t, *J* = 1.5 Hz, 1H), 5.92 (dd, *J* = 6.6, 3.8 Hz, 1H), 5.70 (d, *J* = 7.4 Hz, 1H), 4.03 (dq, *J* = 7.5, 3.9 Hz, 1H), 3.78 (ddd, *J* = 11.5, 5.8, 3.3 Hz, 1H), 3.68 (ddd, *J* = 11.3, 6.6, 4.6 Hz, 1H), 2.84 (**TEAH$^+$**, q, *J* = 7.2 Hz, 8H), 2.40 (d, *J* = 1.2 Hz, 2H), 2.24 – 2.15 (m, 1H), 1.82 (m, 2H), 1.66 (dt, *J* = 12.3, 8.3 Hz, 1H), 1.08 (**TEAH$^+$**, t, *J* = 7.2 Hz, 11H).
^{13}C NMR (D_2O)	δ 165.64, 155.28, 146.48 (d, *J* = 4.2 Hz), 140.74, 125.29 (d, *J* = 10 Hz), 121.75 (d, *J* = 6.1 Hz), 93.59, 85.80, 79.29 (d, *J* = 8.4 Hz), 65.89 (d, *J* = 6.0 Hz), 45.56 (**TEAH$^+$**), 32.11, 25.05, 14.86, 9.53 (**TEAH$^+$**)
^{31}P NMR (D_2O)	δ –8.72 (s)
HRMS	Calc'd for $C_{20}H_{34}N_8O_7P$ $[M–H]^-$: 354.0973; Found: 354.1001

D.2 pK_a determination of 2-MeImpG and nonhydrolyzable analogues

D.2.1 Basic principles[12]

We measured the pK_a of 2-MeImpG or its nonhydrolyzable analogue by acid titration. Protonation of the basic nitrogen atoms on 2-methylimidazole or the heteroaryl-phosphonate mimics causes the ring to be formally positively charged, shifting proton resonances downfield. The chemical shifts of these aromatic protons were plotted against the corresponding pH values to afford a sigmoidal curve. At the inflection point in the Henderson-Hasselbach equation,

$$\mathrm{pH} = \mathrm{p}K_a + \log_{10}\left(\frac{[\mathrm{A}^-]}{[\mathrm{HA}]}\right)$$

the ratio between the monomer and its protonated conjugate acid is 1:1, yielding the pK_a value.

In our NMR experiments, monomer preparations and pH adjustments were performed using D_2O and deuterated reagents, both for NMR locking and to minimize the HOD peak (4.79 ppm) on the proton spectrum. To correct the obtained pK_a values to their H_2O equivalent, we calculate:[13]

$$\mathrm{p}K_a^{\mathrm{H_2O}} = 0.929 \cdot \mathrm{p}K_a^{\mathrm{D_2O}} + 0.42$$

D.2.2 Experimental conditions

30–40 mM solutions of 2-MeImpG and its analogues in D_2O, 5 M sodium chloride in D_2O and 1 M magnesium chloride in D_2O solutions were prepared. The final sample solutions for NMR spectral acquisition, with a sample scale of 400 µL, contained 15 mM monoribonucleotide, 500 mM NaCl, and 100 mM $MgCl_2$. The pD of the final analyte solutions were adjusted by either sodium deuteroxide or deuterium chloride in D_2O. All

proton spectra were acquired with 256 scans, 1 s relaxation delay, and 45° pulse angle at 25 °C. Proton chemical shifts were referenced to residual protium in D_2O (δ = 4.79 ppm).[14]

For 2-MeImpG, ICG, and PYG, the pD of these solutions were first adjusted to ~10; for PZG, pD was adjusted to ~5. The analyte solutions were gradually acidified using DCl/D_2O solutions, followed by spectral acquisitions. The chemical shift of the aromatic protons highlighted in red (**Figures D.1–D.3**) was traced. For 2-MeImpG, ICG and PZG, sample acidification and spectral acquisition were repeated until the samples were at a pD value at which the reporter proton did not undergo further downfield shift. For PYG, acidification and spectral recording were repeated until the sample was at pD = 0.7.

D.2.3 Derivation of the pK_a fitting equation[15]

We assume that proton exchange on the leaving group moieties of the monomers or analogues is faster than the NMR time scale. Under this scenario, the observed chemical shift of the proton of interest is a weighted average between that of the unprotonated monomer (A^-) and its protonated conjugate acid (HA):

$$\delta_{\mathrm{OBS}} = \delta_{\mathrm{HA}} X_{\mathrm{HA}} + \delta_{\mathrm{A^-}} X_{\mathrm{A^-}}$$

Wherein mole fractions (X) of HA and A^- are given by:

$$X_{\mathrm{HA}} = \frac{[\mathrm{HA}]}{[\mathrm{HA}] + [\mathrm{A^-}]} \; ; \; X_{\mathrm{A^-}} = \frac{[\mathrm{A^-}]}{[\mathrm{HA}] + [\mathrm{A^-}]}$$

Thus:

$$\delta_{\mathrm{OBS}} = \frac{\delta_{\mathrm{HA}}[\mathrm{HA}] + \delta_{\mathrm{A^-}}[\mathrm{A^-}]}{[\mathrm{HA}] + [\mathrm{A^-}]} \quad (1)$$

The concentration of the unprotonated monomer, [A^-], at any pH is dictated by the Henderson-Hasselbach equation:

$$\mathrm{pH} = \mathrm{p}K_a + \log_{10}\left(\frac{[\mathrm{A}^-]}{[\mathrm{HA}]}\right)$$

$$\Rightarrow \quad \mathrm{pH} - \mathrm{p}K_a = \log_{10}\left(\frac{[\mathrm{A}^-]}{[\mathrm{HA}]}\right)$$

$$\Rightarrow \qquad \frac{[\mathrm{A}^-]}{[\mathrm{HA}]} = 10^{\,(\mathrm{pH}-\mathrm{p}K_a)}$$

$$\Rightarrow \qquad [\mathrm{A}^-] = [\mathrm{HA}] \cdot 10^{\,(\mathrm{pH}-\mathrm{p}K_a)} \qquad (2)$$

By substituting equation 2 back to equation 1, we get:

$$\delta_{\mathrm{OBS}} = \frac{\delta_{\mathrm{HA}} \cdot [\mathrm{HA}] + \delta_{\mathrm{A}^-} \cdot [\mathrm{HA}] \cdot 10^{\mathrm{pH}-\mathrm{p}K_a}}{[\mathrm{HA}] + [\mathrm{HA}] \cdot 10^{\mathrm{pH}-\mathrm{p}K_a}}$$

$$\boxed{\delta_{\mathrm{OBS}} = \frac{\delta_{\mathrm{HA}} + (\delta_{\mathrm{A}^-}) \cdot 10^{\,(\mathrm{pH}-\mathrm{p}K_a)}}{1 + 10^{\mathrm{pH}-\mathrm{p}K_a}}}$$

D.2.4 Representative data

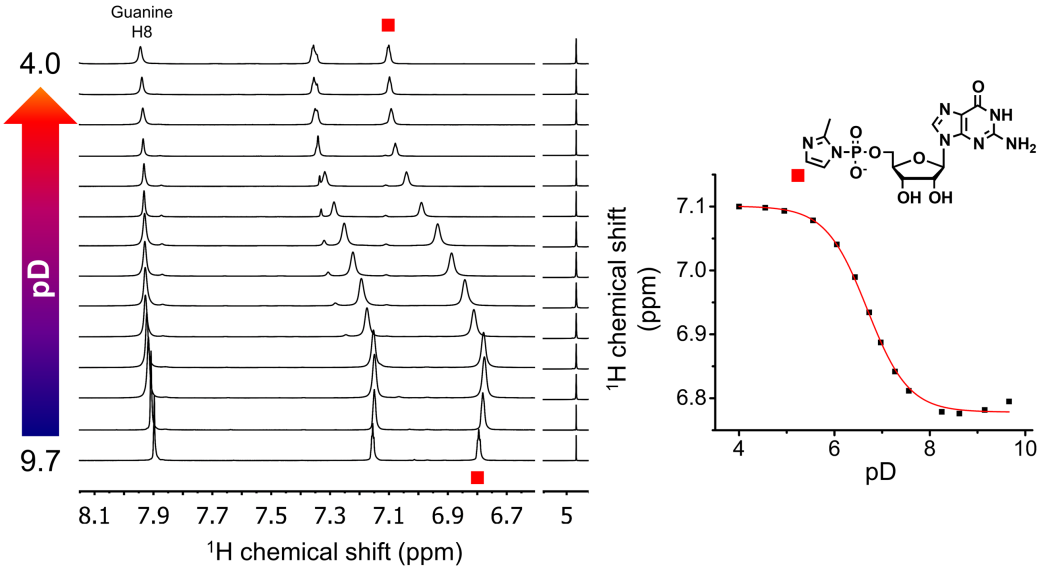


Figure D.1. Variable pD-^{1}H NMR spectroscopy of **2-MeImpG**. The heteroaromatic proton of 2-MeImpG labeled with red square (top right), with a more upfield proton chemical shift as predicted by Chemdraw 16 (PerkinElmer Informatics, Waltham, MA), was monitored as the monomer solution was gradually acidified. Experiments were performed in duplicate. **Left:** Representative stacked proton spectra of the aromatic region. The HOD peaks, referenced at 4.79 ppm, are not shown to scale. **Right:** The chemical shifts of the labeled proton were plotted against their respective monomer pD values, followed by fitting to estimate the pK_a value (6.7 ± 0.1) of the leaving group.

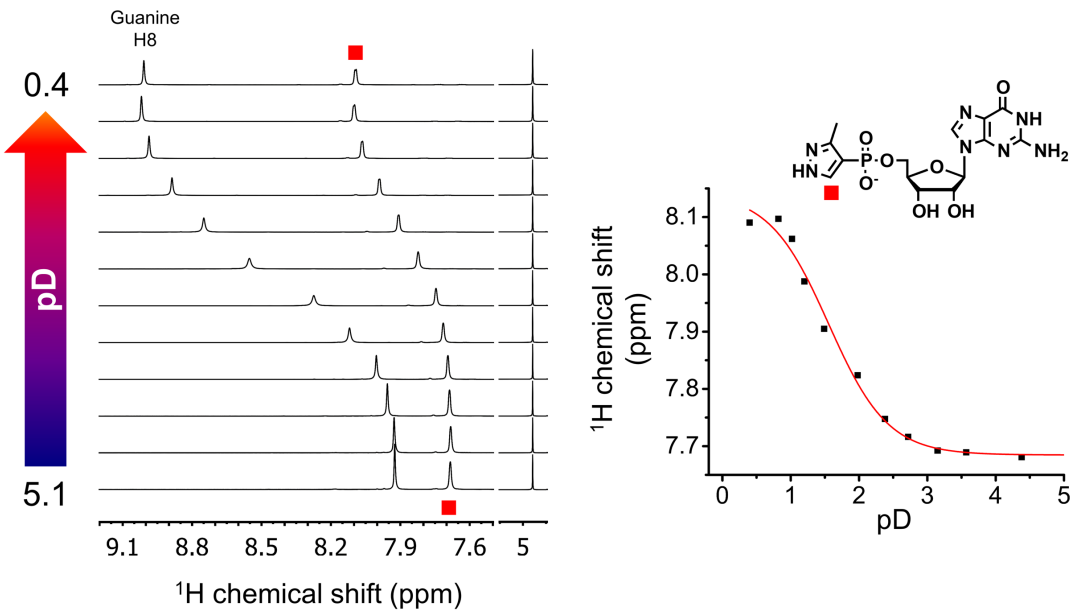


Figure D.2. Variable pD-^{1}H NMR spectroscopy of **PZG**. Experiments were performed in duplicate. **Left:** Representative stacked proton spectra of the aromatic region. The HOD peaks are not shown to scale. **Right:** The chemical shift vs. monomer pD plot. The pK_a value of the leaving group analogue is estimated to be 1.9 ± 0.1.

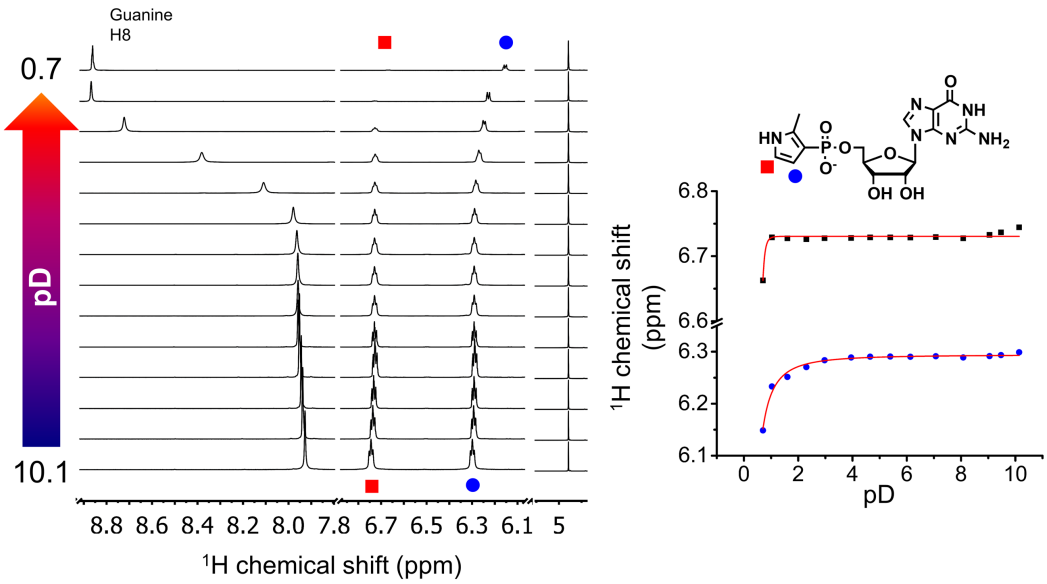


Figure D.3. Variable pD-^{1}H NMR spectroscopy of **PYG**. **Left:** Representative stacked proton spectra of the aromatic region. As the monomer solution was acidified beyond ~ pD 2.0, the proton α to the pyrrole NH (labelled by red square, top right) preferentially underwent proton-deuteron exchange prior to peak shifting,[16] leading to the disappearance of the peak. The proton β to the pyrrole NH (labelled by blue circle) underwent proton-deuteron exchange after the α proton had been significantly exchanged. The HOD peaks are not shown to scale. **Right:** The chemical shift vs. monomer pD plot.

D.3. Isothermal Titration Calorimetry (ITC)

D.3.1 Experimental conditions

ITC[17–19] was used to determine the thermodynamic parameters of binding between 2-MeImpG and its nonhydrolyzable analogues with the **P/T_S/H** sandwich duplex (Sequences: **P** (Primer): 5′-CUCAAUG-3′; **T_S** (Template): 5′-GAGUUACCCAUUGAG-3′; **H** (Helper), 5′-GUAACUC-3′). RNA strands **P, T_S** and **H** were synthesized on a MerMade Synthesizer on a 50 μmol scale in accordance to procedures laid out in **Section A.1.2.** All the ITC analyses were performed on a Nano-ITC calorimeter (TA instruments, New Castle, DE). 170 μL of the cell solution, which contained 1.5 mM each of the **P, T_S** and **H** strands, 450 mM NaCl, 50 mM Na•HEPES at pH 7, was loaded into the sample cell of the instrument. Then, 90 μL of the syringe solution, which contained the same concentration of the RNA strands, NaCl and Na•HEPES as the cell solution in addition to 15 mM guanosine nucleotides, was prepared separately and loaded into the injection syringe. The instrument was equilibrated at 20 °C for 1 hour; afterwards, 32 injections of 1.5 μL syringe solution were incrementally added to the sample cell at 300-second intervals. A plot of the resulting raw heat rate (in μcal s^{-1}) versus time (s) was recorded; the plot was then integrated to obtain the heat evolved per injection of syringe solution (in μcal). From this data, the heat evolved per change in monomer concentration (in kcal mol^{-1}) was measured. This was then plotted against the cumulative concentration of monomer throughout the titration, resulting in a sigmoidal curve, which was then fitted against the binding isotherm equation derived in **Section D.3.3.**

D.3.2 LC-HRMS analysis of oligonucleotides

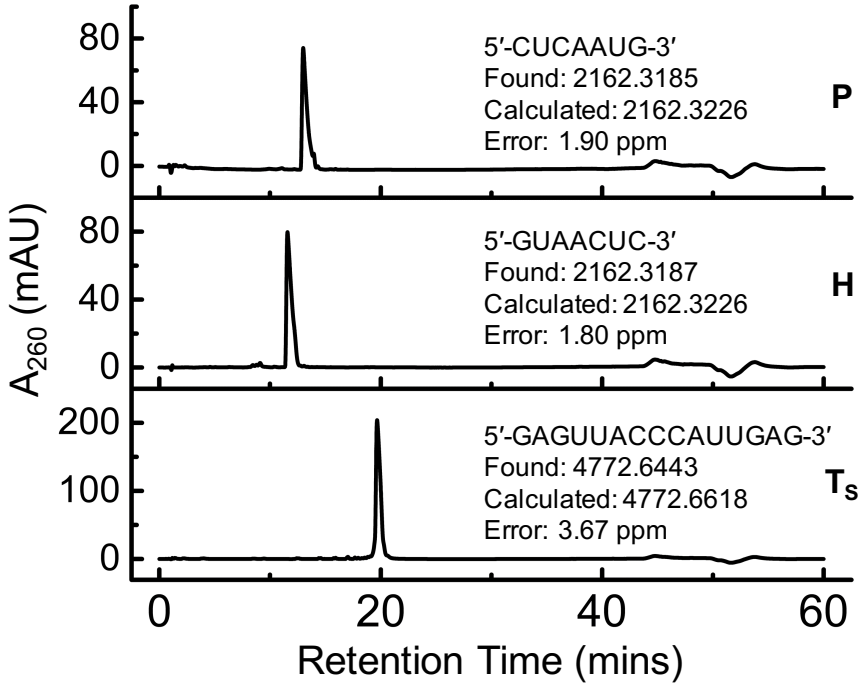


Figure D.4. UV absorbance at 260 nm, and monoisotopic monoanionic mass for the synthesized RNAs for the ITC studies. Based on the UV absorbance traces, the purity was determined to be >90% for the synthesized RNAs. Found: monoisotopic monoanionic mass; Calculated: calculated exact monoanionic mass. Y-axis: milli-absorbance response units at 260 nm.

D.3.3 Derivation of the ITC binding isotherm[17,20,21]

In our proposed mechanism, a molecule of 2-MeImpG or its analogues (M) binds to the **P/T$_S$/H** sandwich duplex (D), with a 1:1 ratio, to form DM, and the binding affinity is reflected by equilibrium association constant K_a. Throughout the ITC experiments, monomers or analogues were incrementally added to the **P/T$_S$/H** duplex. The total heat change resultant from RNA-ligand binding (Q) is a product of the molar enthalpy of duplex-monomer binding (ΔH), the volume of the ITC cell (V_0), and the concentrations of monomer-bound complex DM ([DM]). Thus:

$$Q = \Delta H \cdot V_0 \cdot [DM] \quad (1)$$

Since both ΔH and V_0 are constants, change in heat throughout the titration (dQ) must be the result of the change in concentration of DM (d[DM]):

$$dQ = \Delta H \cdot V_0 \cdot d[DM] \quad (1a)$$

Ultimately, both dQ and d[DM] should be related to the change in total concentration of the monomer ($d[M_{tot}]$). We can therefore differentiate both sides of the equation with respect to $[M_{tot}]$, and we get:

$$\frac{dQ}{dM_{tot}} = \Delta H \cdot V_0 \cdot \frac{d[DM]}{dM_{tot}} \quad (2)$$

To express [DM] in terms of $[M_{tot}]$, we need to consider the following equilibrium, which holds if and only if the binding stoichiometry between the RNA duplex and the monomer is 1:1,

$$K = \frac{[DM]}{[D_{free}] \cdot [M_{free}]} \quad (3)$$

Upon addition of monomer, some portion of total duplex ($[D_{tot}]$) will associate with a portion of added monomer to form [DM], while some will remain in the unbound state ($[D_{free}]$). The same applies to the monomer. The total concentration of duplex and monomer at any point throughout the titration is given by:

$$[D_{free}] + [DM] = [D_{tot}] \; ; \; [M_{free}] + [DM] = [M_{tot}] \quad (4)$$

Equations (3) and (4), when combined, give:

$$K = \frac{[DM]}{(D_{tot} - [DM]) \cdot (M_{tot} - [DM])}$$

$$[DM]^2 - \left(M_{tot} + D_{tot} + \frac{1}{K}\right)[DM] + M_{tot}D_{tot} = 0 \quad (5)$$

Hence, the two roots (solutions) of equation 5 are:

$$[DM] = \frac{\left(M_{tot} + D_{tot} + \frac{1}{K}\right) \pm \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}{2} \quad (6)$$

One of the two roots must be rejected to arrive at the proper isotherm. When *K* approaches positive infinity ($\lim_{K\to\infty}[DM]$), the two roots bifurcate into M_{tot} (larger root) and D_{tot} (smaller root); however, since $M_{tot} > D_{tot}$ at the end of the titration, [DM] must be limited by the total duplex concentration (D_{tot}). Hence [DM] cannot be equal to M_{tot}, and the larger root is rejected. Hence:

$$[DM] = \frac{\left(M_{tot} + D_{tot} + \frac{1}{K}\right) - \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}{2} \quad (6a)$$

From this equation, it is apparent that the change in concentration of the duplex-monomer complex (*d*[DM]) is indeed related to the change in total monomer concentration (dM_{tot}).

Subsequently:

$$\frac{d[DM]}{dM_{tot}} = \frac{d}{dM_{tot}} \frac{\left(M_{tot} + D_{tot} + \frac{1}{K}\right) - \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}{2}$$

$$\frac{d[DM]}{dM_{tot}} = \frac{1}{2}\left\{1 - \left[\frac{\left(M_{tot} - D_{tot} + \frac{1}{K}\right)}{\left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}\right]\right\} \quad (7)$$

By incorporating equation 7 into equation 2, we get the ITC binding isotherm (equation 8).

$$\frac{dQ}{dM_{tot}} = \frac{\Delta H \cdot V_0}{2} \cdot \left\{ 1 - \left[\frac{\left(M_{tot} - D_{tot} + \frac{1}{K} \right)}{\left[\left(M_{tot} + D_{tot} + \frac{1}{K} \right)^2 - 4 M_{tot} D_{tot} \right]^{\frac{1}{2}}} \right] \right\} \quad (8)$$

At this stage, the stoichiometric factor n is introduced to the term D_{tot} to reflect the binding stoichiometry between the RNA sandwich duplex and the ligand; in the ideal scenario, n = 1. To account for any heat that originates from non-idealities, e.g. heats of dilution, a constant a is added, yielding the final form:

$$\frac{dQ}{dM_{tot}} = \frac{\Delta H \cdot V_0}{2} \cdot \left\{ 1 - \left[\frac{\left(M_{tot} - n D_{tot} + \frac{1}{K} \right)}{\left[\left(M_{tot} + n D_{tot} + \frac{1}{K} \right)^2 - 4 n M_{tot} D_{tot} \right]^{\frac{1}{2}}} \right] \right\} + a$$

D.3.4 Representative data

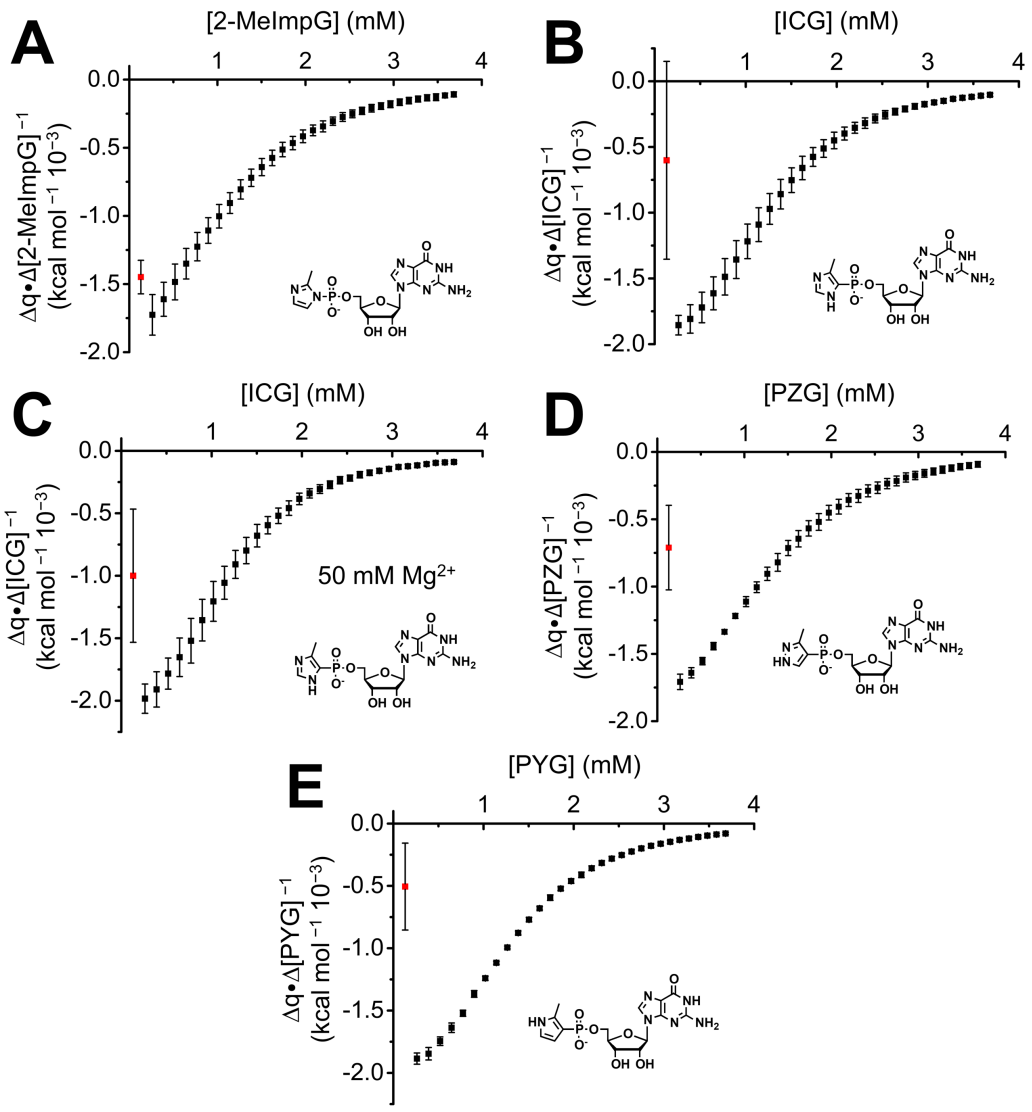


Figure D.5. (a–e) Plots of the change of heat evolved per change in monomer/analogue concentration of one injection ($\Delta q \cdot \Delta[\text{analogue}]^{-1}$), versus the cumulative change in monomer/analogue concentration, for (a) 2-MeImpG, (b) ICG, (c) ICG with 50 mM Mg^{2+}, (d) PZG and (e) PYG. During an ITC titration run, 50 µL of syringe solution, which contained 15 mM monomer/analogue, 1.5 mM of **P/T$_S$/H,** 50 mM pH 7 Na•HEPES and 450 mM NaCl, was titrated into 170 µL of cell solution, which did not contain the analogue. All titrations were run in triplicate.

Table D.1. Thermodynamic Parameters of Analogue–RNA Binding

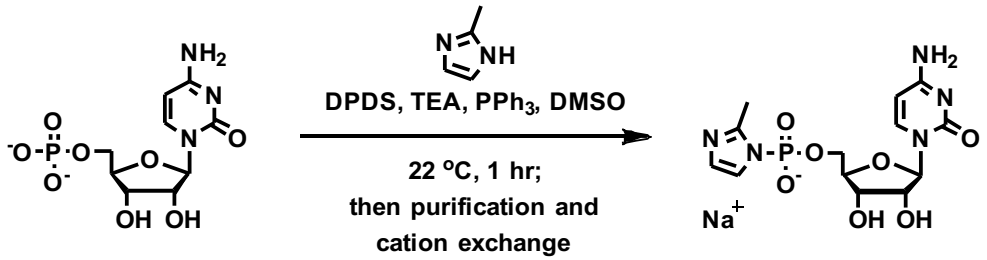
Ligand		Fitted Parameters			Derived Parameters	
		K_a (M^{-1})	ΔH (kcal mol^{-1})	n	ΔG (kcal mol^{-1})	ΔS (cal mol^{-1} K^{-1})
2-MeImpG	1	2.3×10^3	–16.7	0.75	–4.52	–41.5
	2	2.4×10^3	–14.3	0.74	–4.53	–33.4
	3	1.6×10^3	–20.3	0.66	–4.31	–54.5
	Avg	$(2.1\pm0.4)\times10^3$	–17.1±3.0	0.72±0.05	–4.45±0.12	–43.1±10.7
ICG	1	3.6×10^3	–14.0	0.81	–4.77	–31.5
	2	4.0×10^3	–14.2	0.84	–4.83	–32.0
	3	4.7×10^3	–14.0	0.93	–4.92	–30.9
	Avg	$(4.1\pm0.5)\times10^3$	–14.1±0.1	0.86±0.06	–4.84±0.07	–31.5±0.5
ICG (w/ 50 mM Mg^{2+})	1	5.6×10^3	–14.8	0.83	–5.03	–33.3
	2	4.0×10^3	–15.7	0.80	–4.83	–37.0
	3	3.7×10^3	–15.1	0.72	–4.79	–35.1
	Avg	$(4.5\pm1.0)\times10^3$	–15.2±0.5	0.78±0.06	–4.89±0.13	–35.1±1.9
PZG	1	3.1×10^3	–14.8	0.80	–4.69	–34.6
	2	2.9×10^3	–13.7	0.91	–4.64	–31.0
	3	2.5×10^3	–14.1	0.93	–4.56	–32.4
	Avg	$(2.8\pm0.3)\times10^3$	–14.2±0.6	0.88±0.07	–4.63±0.06	–32.7±1.8
PYG	1	3.8×10^3	–15.1	0.87	–4.80	–35.2
	2	3.4×10^3	–14.5	0.90	–4.74	–33.3
	3	3.8×10^3	–14.8	0.86	–4.80	–34.3
	Avg	$(3.6\pm0.2)\times10^3$	–14.8±0.3	0.88±0.02	–4.78±0.03	–34.2±0.9

Table D.1. (continued) By fitting the ITC data to the binding isotherm derived in **Section D.3.2**, we can obtain the affinity constant (K_a), change in enthalpy (ΔH) and stoichiometry factor (n) of RNA-analogue binding. Using the relationships $\Delta G = -RT\ln(K_a) = \Delta H - T\Delta S$, we can further obtain the free energy (ΔG) and entropy change (ΔS) of RNA-analogue binding. The adjusted R^2 of all fits were > 0.99.

D.4. Nonenzymatic Primer Extension

D.4.1 Preparation of cytosine 5'-phosphoro-2-methylimidazolide sodium salt (2-MeImpC•Na^+)

The protocol employed for the synthesis of 2-MeImpC was adapted from that of Joyce, Inoue and Orgel.[22]



Cytosine 5'-monophosphate (CMP) *free acid* (1 mmole, 1 equiv., 323 mg), 2-methylimidazole (5 mmoles, 5 equiv., 411 mg), triphenylphosphine (5 mmoles, 5 equiv., 1.31 g), anhydrous triethylamine (ρ = 0.726 g mL^{-1}, 5 mmoles, 5 equiv., 700 µL), and 20 mL anhydrous DMSO (50 mM) was sequentially charged into a flame-dried round-bottom flask. The content was heated and sonicated until complete homogenization (~ 10 min), followed by addition of 2,2′-dipyridyl disulfide (5 mmoles, 5 equiv., 1.1 g). The reaction was allowed to stir for 1 h. The crude product was obtained, as an off-white precipitate, by adding the reaction mixture slowly into an ice-cold solution of 200 mL acetone, 200 mL diethyl ether, 18 mL of triethylamine, and 2.5 mL of saturated sodium perchlorate ($NaClO_4$) in acetone. The precipitate was then collected by suction filtration, followed by volatile removal under high vacuum. The resulting crude residue was

purified on a 50 g C18Aq column over 20 CV of of 0–10% CH_3CN in TEAB (pH 7.5). The titled compound was eluted from the column at ~5% CH_3CN in TEAB, followed by lyophilization to afford the titled compound (in triethylammonium cation form) as a white fluffy solid. Lyophilized sample was resuspended in a minimum amount of deionized water, followed by precipitation in >25-fold excess (in volume) of 1:1:0.04 (v/v/v) acetone/diethyl ether/saturated $NaClO_4$ in acetone. The precipitate was centrifuged and washed with 1:1 acetone/ether twice. Final drying of the centrifuged and washed sample on high-vacuum furnished 2-MeImpC•Na^+ as a fine white granular solid.

D.4.2 Primer Extension: Experimental Conditions

Primer extension reactions were set up, in 10 μL scales, with the following components: 1.2 μM Cyanine 3-labeled primer (Sequence: 5′-Cy3-GACUGACUGG-3′), 2.2 μM complementary template (Sequence: 5′-AACCCGCCAGUCAGUC-3′), 200 mM Na•HEPES (pH 7 and 8) or Na•MES buffer (pH 5.5 and 6.3), 100 mM $MgCl_2$, and monomers (Set 1: 30 mM 2-MeImpC; Set 2: 30 mM 2-MeImpC & 30 mM 2MeImpG; Set 3: 30 mM 2MeImpC & 30 mM ICG•Na^+; Set 4: 30 mM 2-MeImpC & 30 mM PZG•Na^+; Set 5: 30 mM 2MeImpC & 30 mM PYG•Na^+; Set 6: 30 mM 2MeImpC & 30 mM GMP•$2Na^+$).

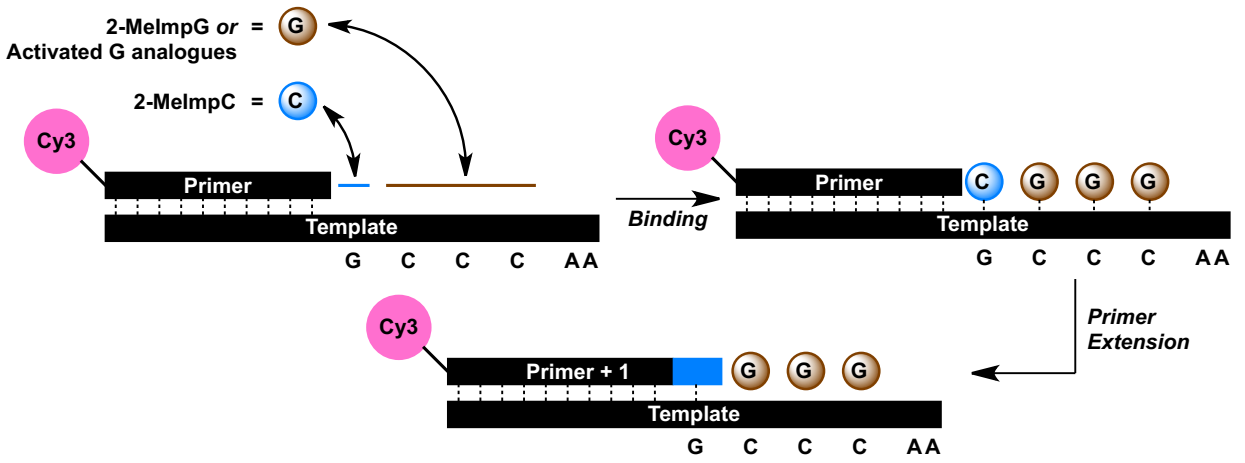


Figure D.6. Schematic diagram of the primer extension reaction. The 5′-Cy3-labeled primer was annealed with the template in the presence of 100 mM $MgCl_2$ and 200 mM buffer of desired pH at room temperature. 2-MeImpC (cyan) and stacking 2-MeImpG or nonhydrolyzable analogues (brown) was then added to the primer-template complex. The monomers first bind noncovalently to the primer-template complexes, followed by an irreversible extension step between primer and 2-MeImpC to afford the extended product.

The pH of the stock solutions of both monomers and analogues were adjusted to the pH value at which the primer extensions were run. Primer, template, aqueous $MgCl_2$ and buffer of the appropriate pH were first combined and thoroughly mixed in 200 µL thin-walled PCR tubes, followed by addition of the appropriate amount of stock solutions of 2-MeImpC and guanosine monomers or analogues (~100 mM, concentration determined by Nanodrop) to initiate the extension reactions. At appropriate time points, 0.5 µL aliquots of these reactions were quenched by addition to 15 µL of a quench buffer containing 1X TBE (89 mM Tris-borate and 2 mM EDTA, overall pH 8.3), 8.0 M urea, and 100 mM pH 8.0 EDTA•$4Na^+$. 3 µL of these quenched samples, which contained 0.12 pmole of the fluorescently-labelled primer as well as their extended products, were loaded onto a 200 mm × 200 mm × 0.4 mm denaturing polyacrylamide gel,[23] which contained 20% acrylamide monomers with a ratio of 19:1 acrylamide:*N,N'*-methylenebis(acrylamide), as well as 7.5 M urea (National Diagnostics, Atlanta, GA), followed by electrophoresis at 25 W for 1.5 – 2 h. Reaction rates were calculated by quantifying the conversion of primer to its extended products, and the log values of the fraction of unreacted primer was plotted against time. A linear regression was performed and the negative slope of the fit as plotted was reported as the initial pseudo-first order rate k_{obs}. Imaging and quantification of the fluorescently labelled primer and extended products were performed with an Amersham Biosciences Typhoon 9410 Imager, and the ImageQuant TL software package (GE Healthcare, Little Chalfont, United Kingdom). The k_{max} and K_M of saturation curves (**Figure 4.4.C**; **Figures D.13–D.14**) was estimated by GraphPad Prism 5 (GraphPad Software, San Diego CA).

D.4.3 Representative data

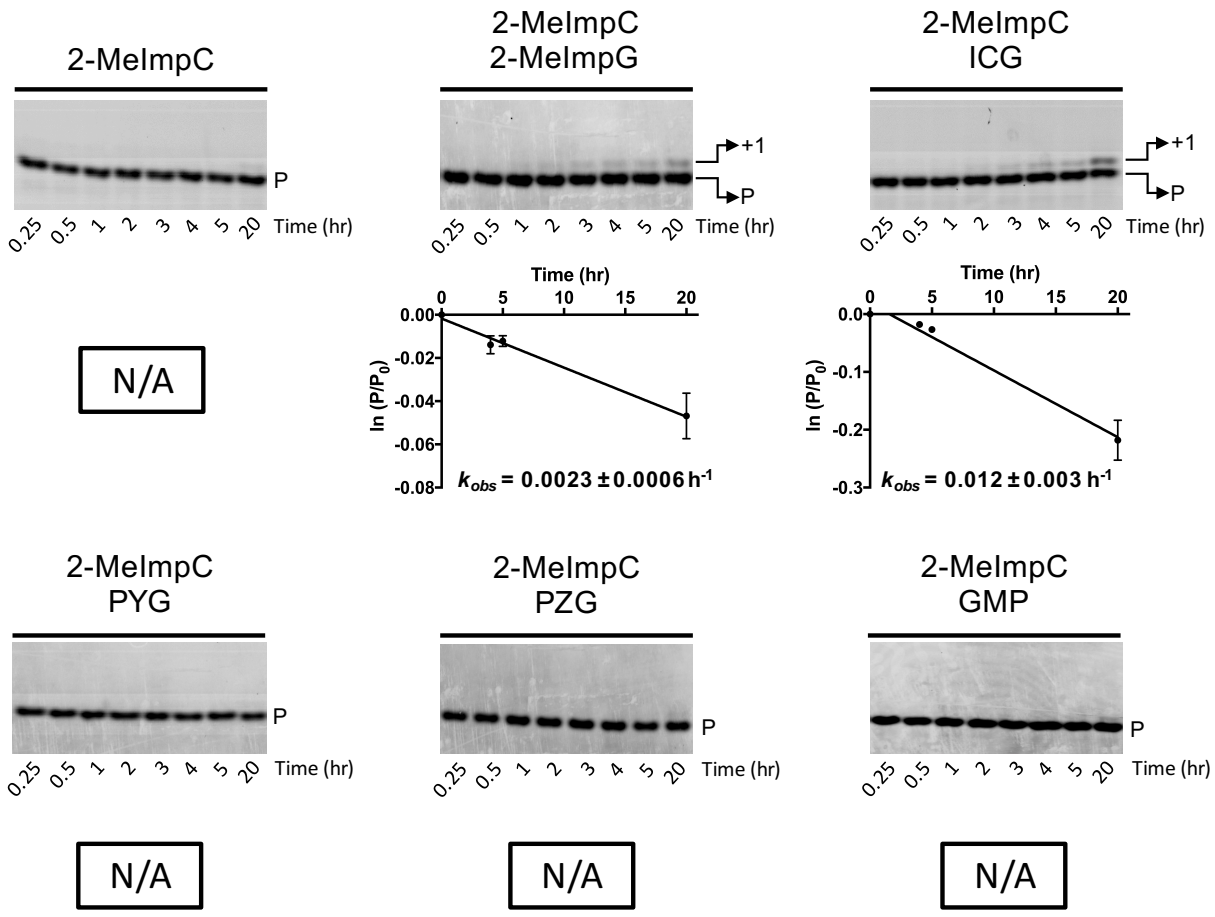


Figure D.7. Primer extension reactions at **pH 5.5**. **Top:** Representative 20% denaturing PAGE analyses of the primer extension products. P: Primer; +1: Extended primer+C product. **Bottom:** Plots of the natural log values of the fraction of unreacted primer (ln (P/P_o)) against time. The linear fits of these plots were used to determine the pseudo first-order observed rate constants (k_{obs}). All primer extension experiments presented herein were performed in triplicate, and the standard deviation was presented as error. N/A: rate too low to be reliably fitted.

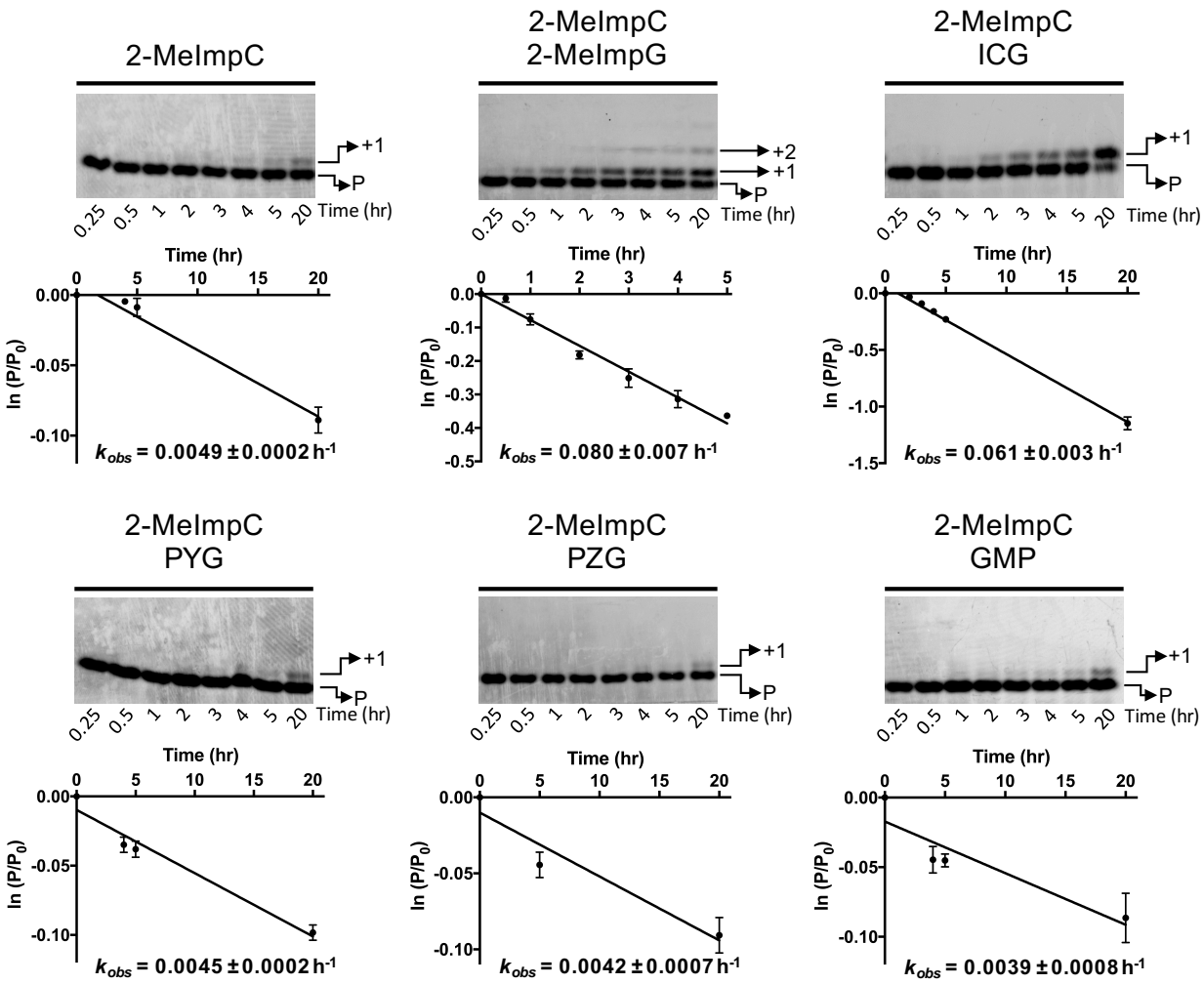


Figure D.8. Primer extension reactions at **pH 6.3**. **Top:** Representative 20% denaturing PAGE analyses of the primer extension products. P: Primer; +1, +2: Extended primer+C / primer+CG products, respectively. **Bottom:** Plots of the natural log values of the fraction of unreacted primer (ln (P/P_o)) against time. All experiments were performed in triplicate.

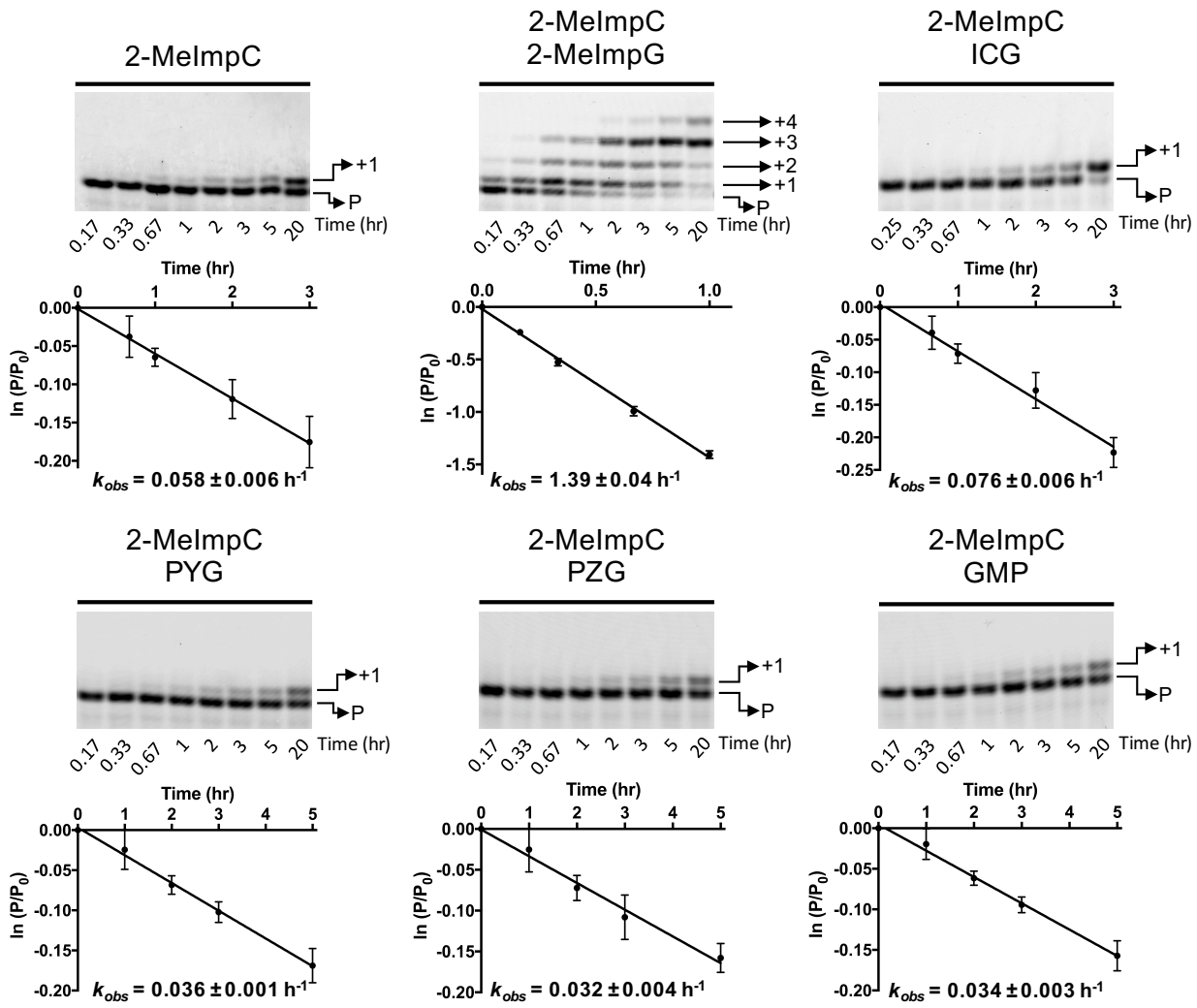


Figure D.9. Primer extension reactions at **pH 7.0**. **Top:** Representative 20% denaturing PAGE analyses of the primer extension products. P: Primer; +1, +2, +3, +4: Extended primer+C / primer+CG / primer+CGG / primer+CGGG products, respectively. **Bottom:** Plots of the natural log values of the fraction of unreacted primer (ln (P/P_o)) against time. All experiments were performed in triplicate.

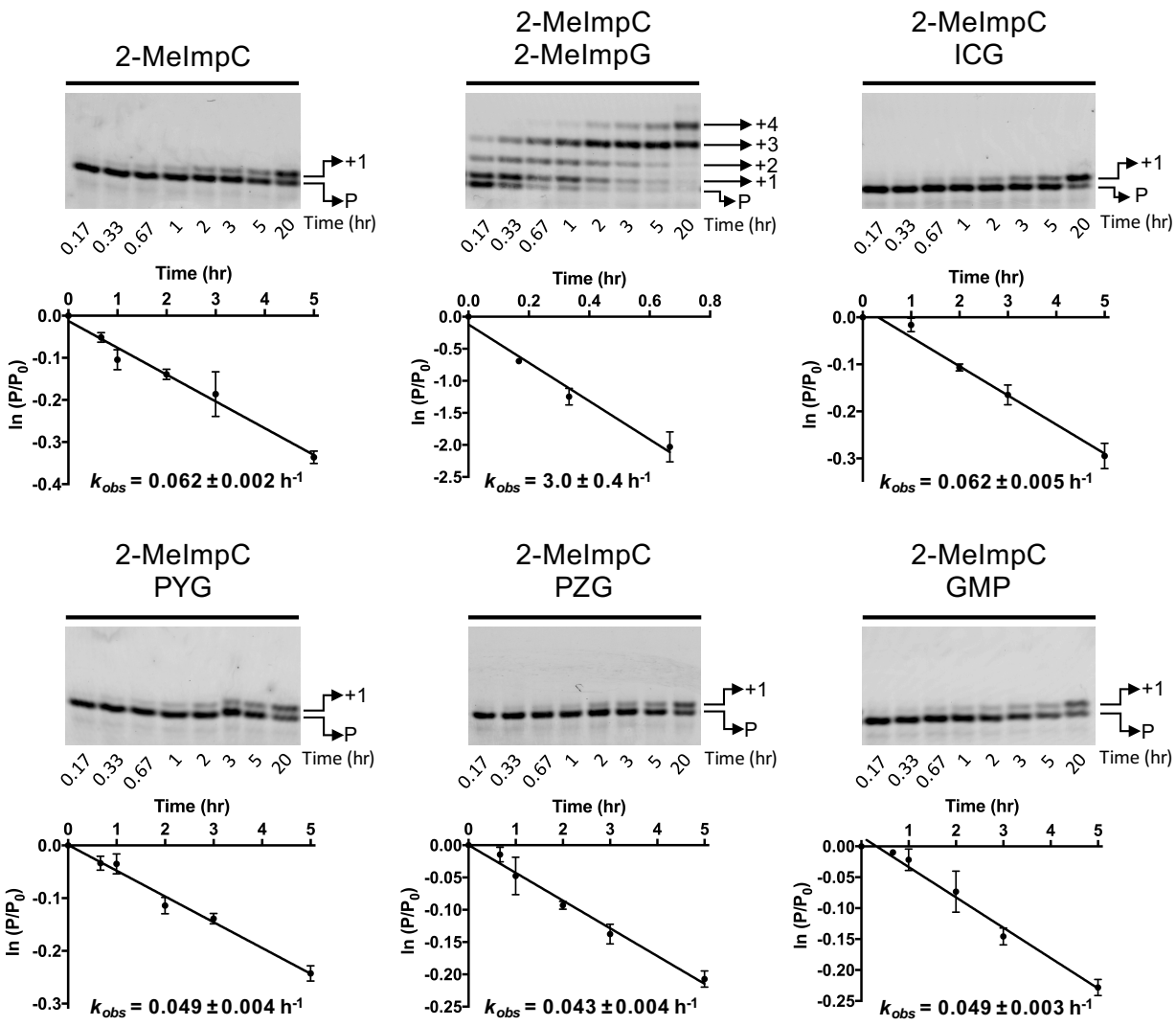


Figure D.10. Primer extension reactions at **pH 8.0**. **Top:** Representative 20% denaturing PAGE analyses of the primer extension products. P: Primer; +1, +2, +3, +4: Extended primer+C / primer+CG / primer+CGG / primer+CGGG products, respectively. **Bottom:** Plots of the natural log values of the fraction of unreacted primer (ln (P/P_o)) against time. All experiments were performed in triplicate. We also performed primer extension reactions with only 2-MeImpG, in the absence of 2-MeImpC, at pH 5.5, 6.3, 7.0 and 8.0. No extended product was observed after 20 hours (data not shown).

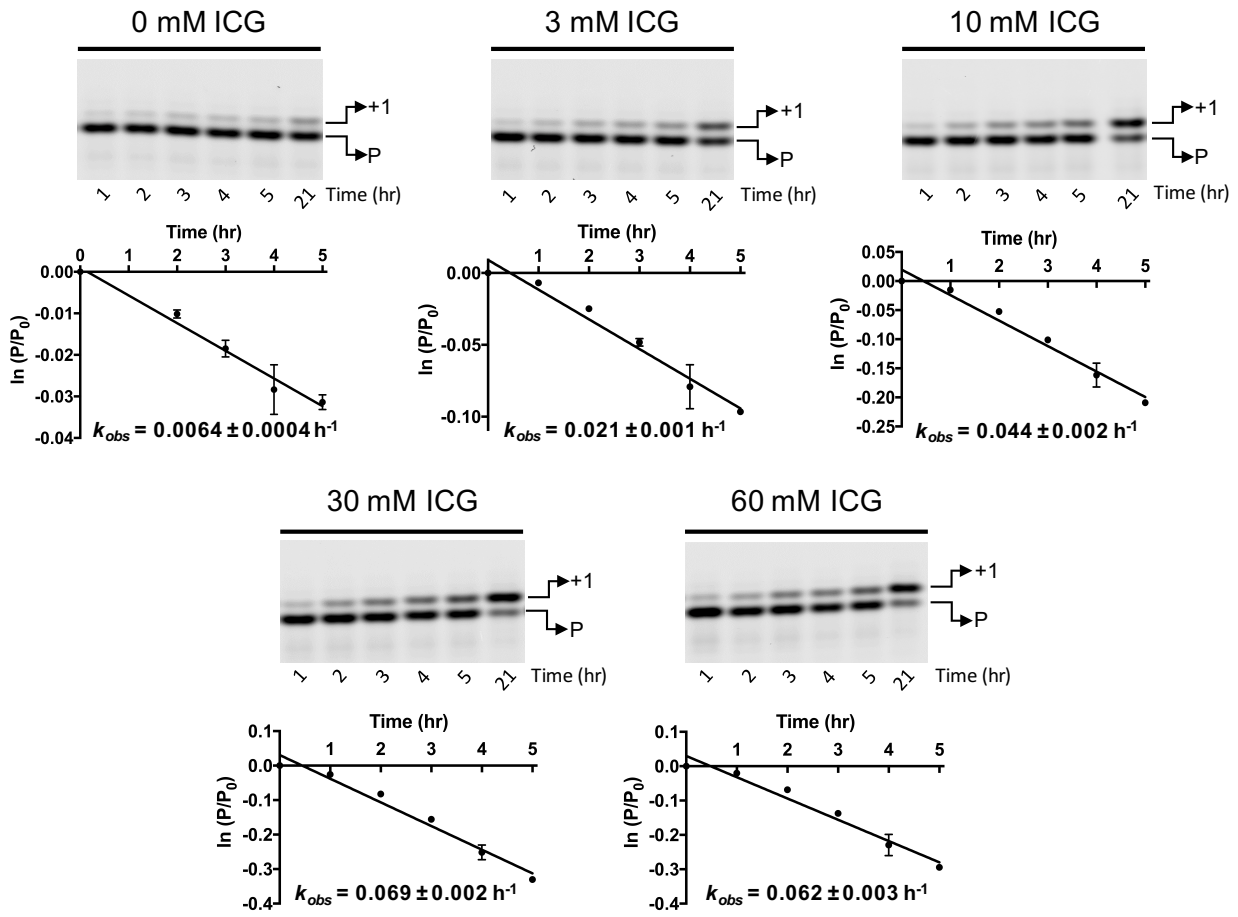


Figure D.11. **ICG**-stacked primer extension reactions at **pH 6.3**. **Top:** Representative 20% denaturing PAGE analyses of the primer extension products. P: Primer; +1: Extended primer+C product. **Bottom:** Plots of the natural log values of the fraction of unreacted primer (ln (P/P_o)) against time. All experiments were performed in triplicate. The Michaelis-Menten plot of ICG-stacked extensions at pH 6.3 was shown in **Figure 4.4.C.**

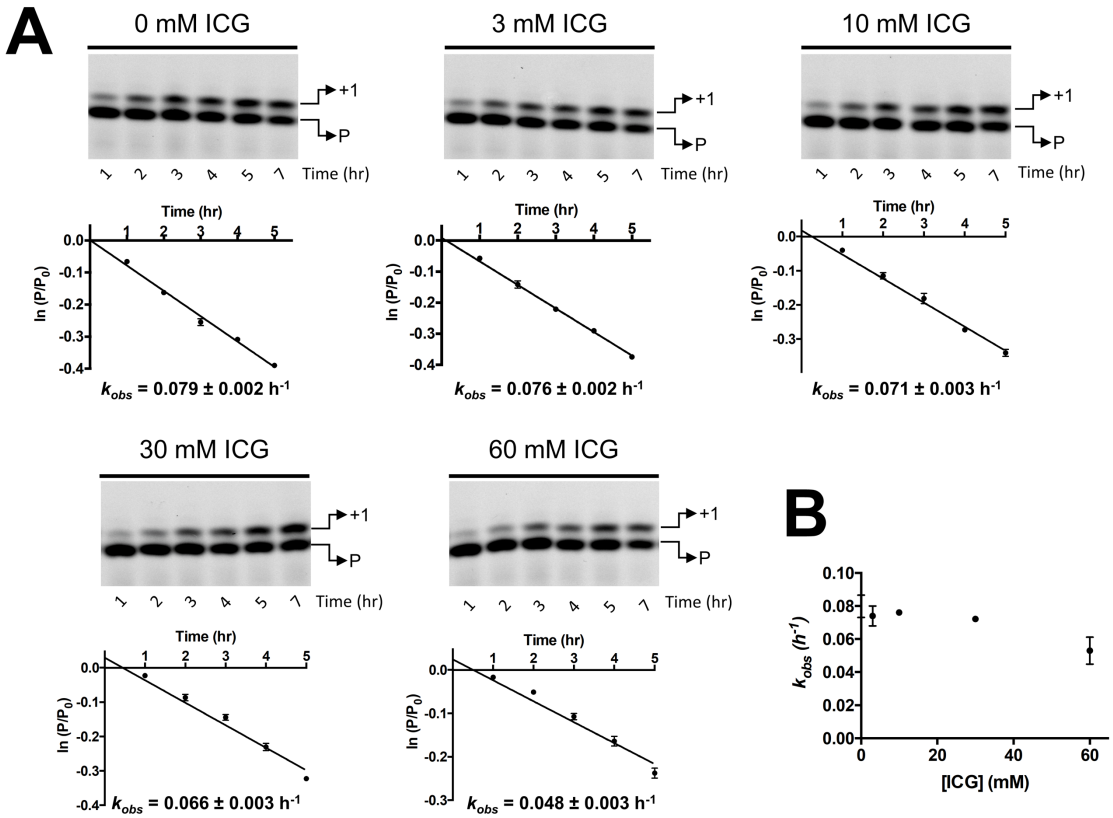


Figure D.12. Primer extension reactions in the presence of **ICG** at **pH 8.0**. (a) **Top:** Representative 20% denaturing PAGE analyses of the primer extension products. P: Primer; +1: Extended primer+C product. **Bottom:** Plots of the natural log values of the fraction of unreacted primer (ln (P/P_o)) against time. All experiments were performed in triplicate. (b) Saturation curve of the ICG-stacked primer extensions at pH 8.0. At this pH, increasing concentrations of ICG do not enhance the rate of primer extension. At rates of ~0.07 h^{-1}, the rates of ICG-stacked extensions are comparable with that of the unstacked (0 mM ICG) extension. Hence, when ICG is used as a stacking monomer at pH 8.0, it offers minimal catalysis to primer extension. Significant precipitation was observed for the primer extension with 60 mM ICG, possibly explaining the slightly lower extension rate.

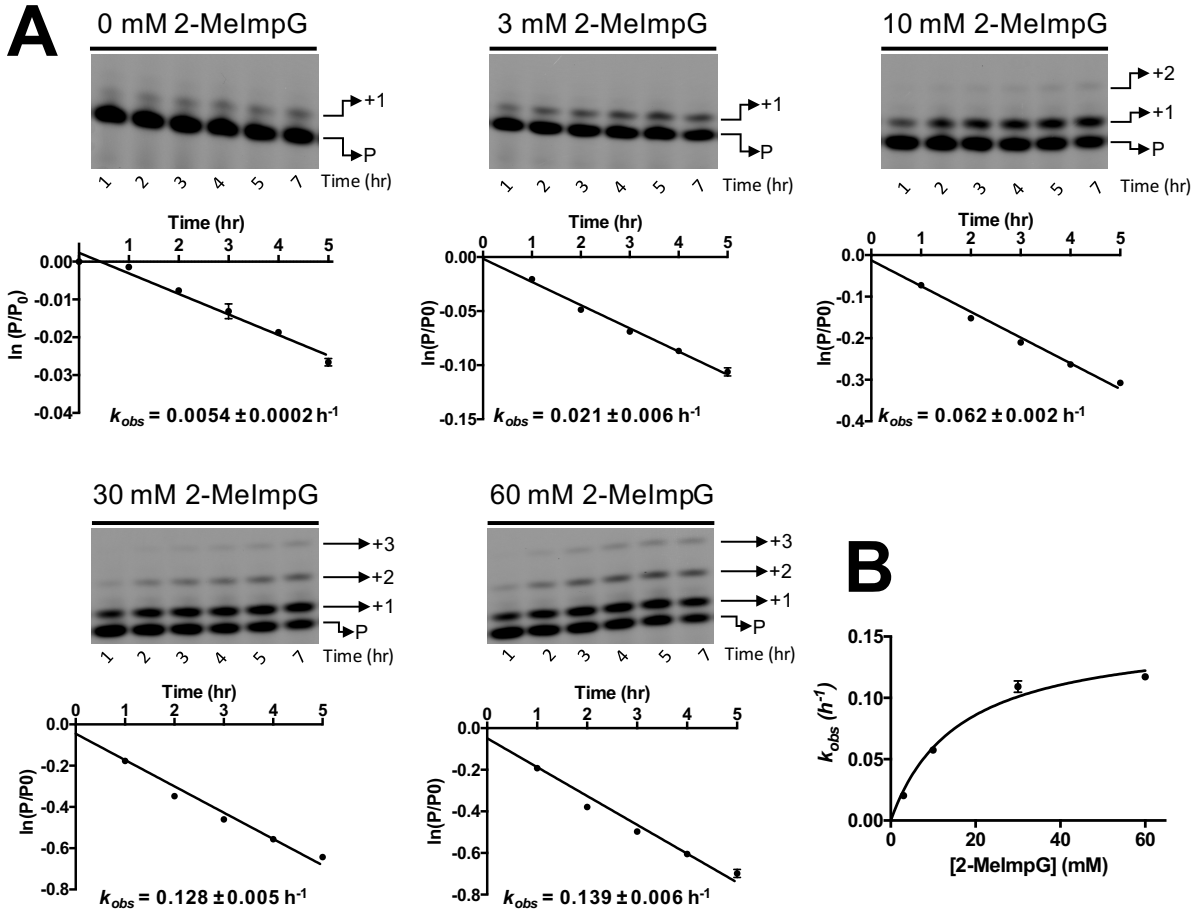


Figure D.13. 2-MeImpG-catalyzed primer extension reactions at **pH 6.3**. (a) **Top:** Representative 20% denaturing PAGE analyses of the primer extension products. P: Primer; +1, +2, +3: Extended primer+C / primer+CG / primer+CGG products, respectively. **Bottom:** Plots of the natural log values of the fraction of unreacted primer (ln (P/P_o)) against time. All experiments were performed in triplicate. (b) Saturation curve of the 2-MeImpG-stacked primer extensions at pH 6.3. The estimated k_{max} and K_M was 0.155 ± 0.008 h^{-1} and 16 ± 2 mM, respectively (R^2 = 0.983). At this pH, the rate enhancement afforded by 2-MeImpG approaches the maximum when the concentration of 2-MeImpG is at, or above 30 mM.

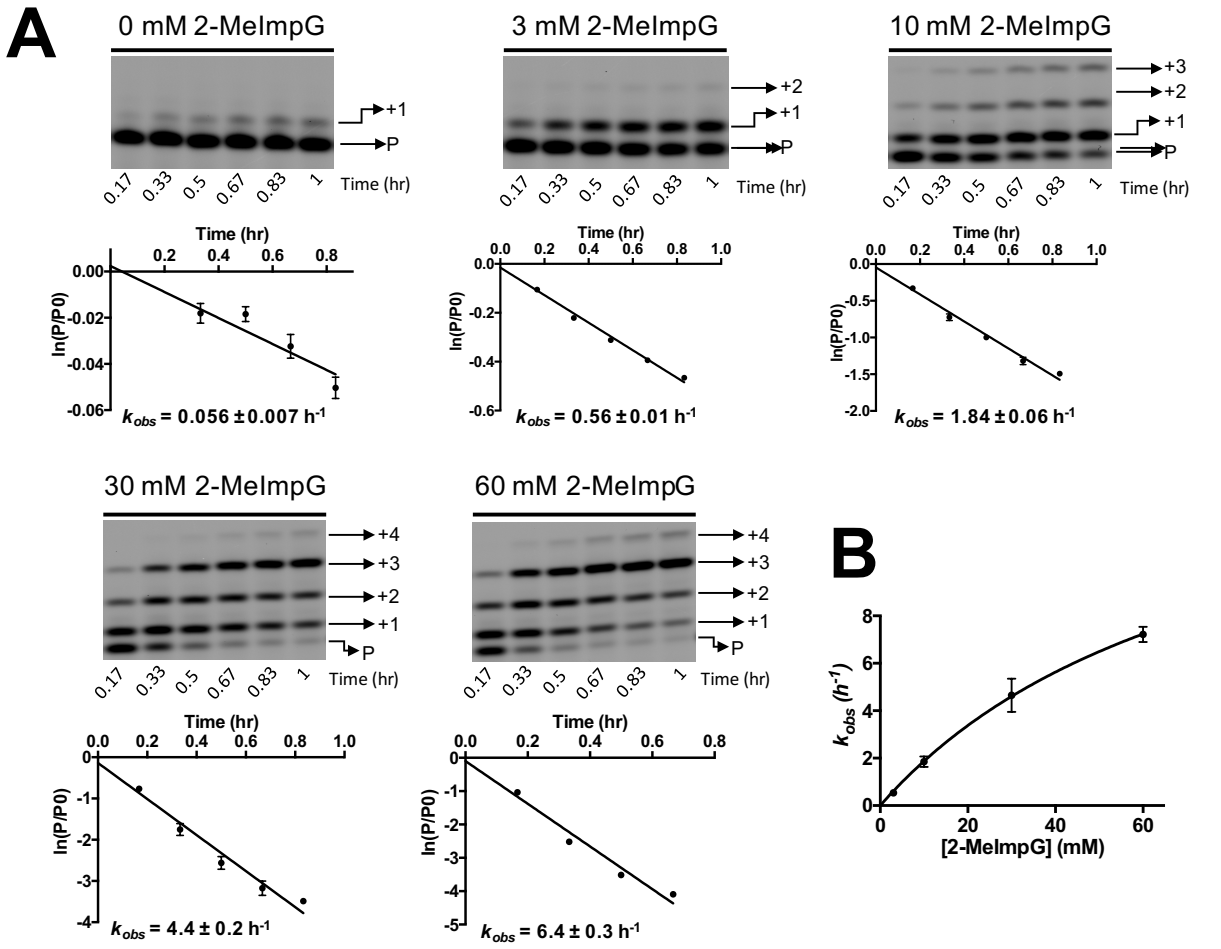


Figure D.14. **2-MeImpG**-catalyzed primer extension reactions at **pH 8.0**. (a) **Top:** Representative 20% denaturing PAGE analyses of the primer extension products. P: Primer; +1, +2, +3, +4: Extended primer+C / primer+CG / primer+CGG / primer+CGGG products, respectively. **Bottom:** Plots of the natural log values of the fraction of unreacted primer (ln (P/P_o)) against time. All experiments were performed in triplicate. (b) Saturation curve of the 2-MeImpG-stacked primer extensions at pH 8.0. The estimated k_{max} and K_M was 17 ± 2 h^{-1} and 80 ± 18 mM, respectively (R^2 = 0.988). Therefore, the 60-fold rate increase conferred by 30 mM 2-MeImpG at primer extension pH of 8.0, as shown in **Figure 4.4.B**, does not represent the maximal catalytic capability conferred by 2-MeImpG; the true rate enhancement afforded by downstream stacking 2-MeImpG would be greater than 60-fold at pH 8.0. The much higher observed level of extension rate catalysis displayed by 2-MeImpG at pH 8.0 is possibly related to the enhanced formation of the imidazolium-bridged dinucleotide intermediate.

D.5. CpICG: Detection and Structure Elucidation

D.5.1 Coincubation of 2-MeImpddC with nonhydrolyzable analogues

D.5.1.1. Experimental conditions: liquid chromatography, mass spectrometric and tandem MS/MS analyses

2-MeImpddC-analogue coincubation experiments were set up, in 20 μL scales, with the following components: 30 mM nonhydrolyzable analogues, 30 mM 2-MeImpddC, and 200 mM buffer (pH 6.3 Na•MES or pH 8.0 Na•HEPES), followed by *prompt* pH adjustments with aqueous hydrochloric acid or sodium hydroxide. At appropriate time points, 1 μL aliquots of these reactions were diluted 100-fold with 99 μL of water, followed by prompt injection of 1−2 μL of the diluted samples (300−600 pmole of 2-MeImpddC and analogue) into the liquid-chromatography-coupled mass spectrometer.

The formation of dideoxycytidine-analogue conjugate was assessed using an Agilent 1100 high-performance liquid chromatography system coupled with a ThermoFisher LTQ XL linear ion trap mass spectrometer. The LC-MS setup was equipped with a solvent degasser, auto sampler, diode array detector, and an electrospray ionization source. Chromatographic resolution of 2-MeImpddC, analogue, as well as the ddC-G dinucleotide conjugates and degradation products was achieved under reverse-phase conditions across an Agilent Polaris Amide-C18 column (150 mm × 2.0 mm i.d., 3 μm particle size), with aqueous mobile phase being 5 mM neutral triethylammonium acetate buffer and organic mobile phase being acetonitrile. Organic mobile phase was ramped linearly from 2% to 3.2% over 0−6 min and from 3.2% to 7.0% over 6−20 min. The flow rate was maintained at 0.2 mL min^{-1} and the column was kept under ambient temperature. Concomitantly, the *m/z* ratio of the chromatographed

eluent was continuously monitored in negative ion mode from 150–1000 *m/z* with a scan rate of 16,667 Da s^{-1}, with the following parameters: drying gas flow, 25 L/min; capillary temperature, 350 °C; spray voltage, 4000 V; spray current, 100 μA; capillary voltage, –0.5 V; and tube lens voltage, –52 V. For the tandem MS/MS analysis, the source parameters were identical, and the parent ions were fragmented using collision induced dissociation (CID) with nitrogen, and normalized collision energy was set at 25.

See **Figures D.15−D.17**, **D.19** and **D.20** for representative data.

D.5.1.2. Experimental conditions: ^{31}Phosphorus NMR Time Courses

2-MeImpddC and ICG were coincubated in 100 μL scales, with the following components: 30 mM ICG, 30 mM 2-MeImpddC, and 200 mM pH 6.3 Na•MES, and 20% D_2O. This sample solution was *promptly* adjusted to pH 6.3 with *minimal* amounts of aqueous hydrochloric acid or sodium hydroxide, transferred to a 3mm NMR tube (Wilmad-LabGlass, SP Scienceware, Warminster, PA), followed by ^{31}P NMR time course analysis over 12 h on a Varian Inova 400 MHz spectrometer equipped with a broadband PFG (z-gradient) probe (161 MHz for ^{31}P). Proton-decoupled phosphorus chemical shifts were reported in parts per million (ppm) values on the δ scale; each ^{31}P spectrum was acquired with 512 scans, 1 s relaxation delay, spectral window of –26 to 32 ppm, and 45° pulse angle at 25 °C (Acquisition time per spectrum was approximately 30 min), and acquired at the beginning of each full hour of incubation. At the end of the time course analysis, deuterated Na_2HPO_4 was introduced to the 2-MeImpddC-ICG sample as an internal reference (final Na_2HPO_4 concentration ~30 mM). The sample was adjusted to pH 6.3 with minimal aq. HCl and NaOH, and the ^{31}P spectrum was reacquired to afford a Na_2HPO_4-referenced spectrum for the last time point. The ^{31}P

resonance of the added Na_2HPO_4 at pH 6.3 was determined to be 0.87 ppm, using an external reference of 85% H_3PO_4 (0 ppm by definition) in D_2O contained in a coaxial NMR tube. All other ^{31}P spectra (spectra corresponding to time points 0−11h) were referenced to this Na_2HPO_4-referenced spectrum.

See **Figure D.18** for representative data.

Notes on sample transfer and NMR tube cleaning: Samples were transferred into the 3 mm NMR tube by a Hamilton 500 µL gas-tight glass syringe and 7-inch, 22s-gauged needle assembly (needle inner diameter: 0.152 mm; Reno, NV); dead volume within the needle was estimated to be 3.2 µL. The syringe-needle assembly was thoroughly washed by Decon Contrad 70 detergent (King of Prussia, PA), 18 MΩ water and deuterium oxide prior to use. 3 mm NMR tubes were cleaned by soaking in conc. HCl for >1 h, followed by thorough 18 MΩ water and acetone rinsing. Residual acetone was removed by vacuum evaporation inside a house-vacuumed chamber overnight. 3mm NMR tubes should not be dried inside an oven.

D.5.1.3. Representative data

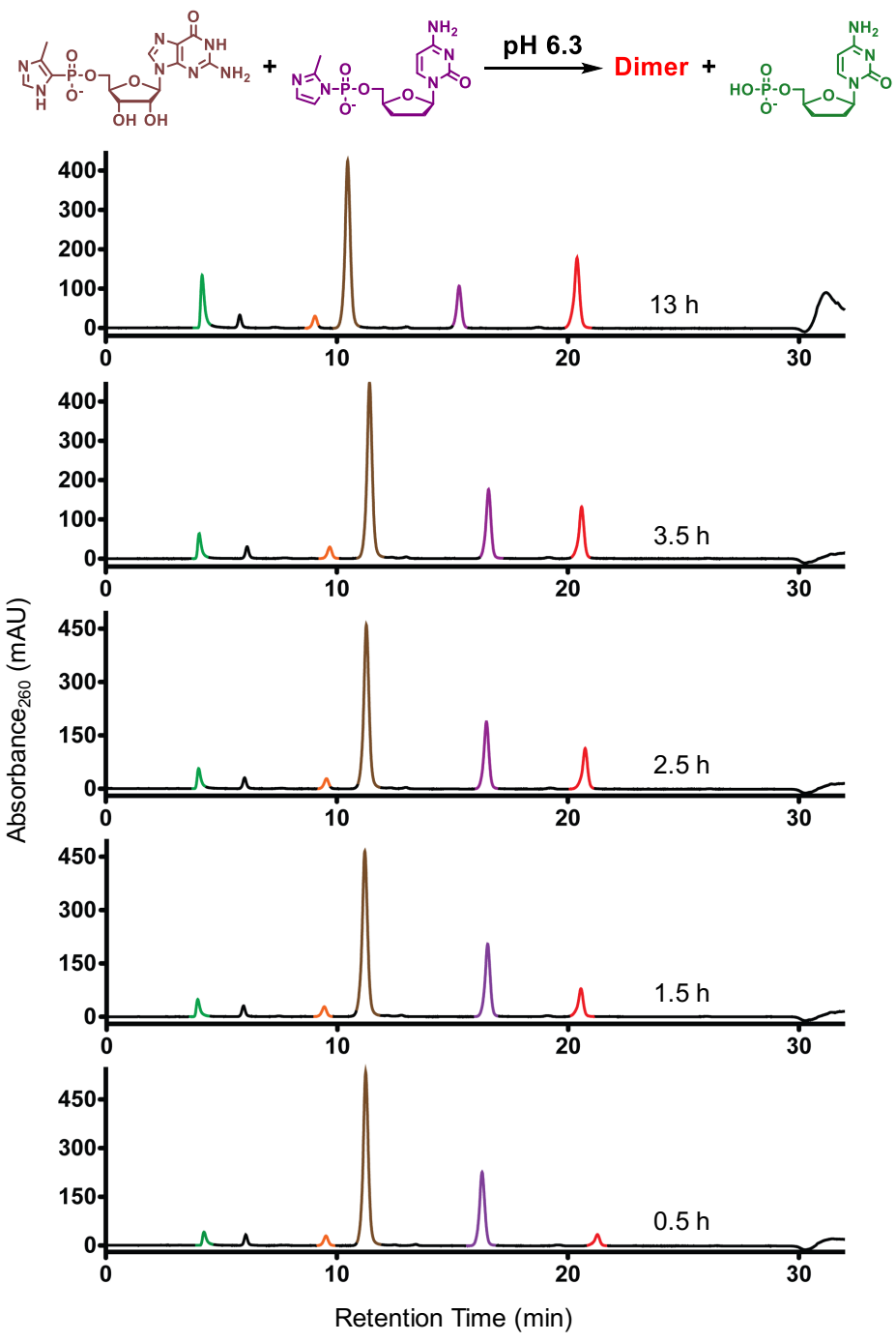


Figure D.15. 2-MeImpddC was coincubated with **ICG** at **pH 6.3**. The formation of ddC-ICG dinucleotide was assessed over **13 h**. The red, purple, brown, orange and green peaks correspond to ddC-ICG dinucleotide, 2-MeImpddC, ICG, guanosine (trace impurity from ICG synthesis) and dideoxycytidine monophosphate (ddCMP; a byproduct of 2-MeImpddC hydrolysis), respectively. All peak assignments were performed by comparing the monoanionic *m/z* of the LC peaks with the expected monoanionic *m/z* of the proposed species; individual mass spectra that aided the assignments of the LC peaks are omitted for clarity. The ddC-ICG and CMP peaks grow in intensity over the duration of the coincubation experiment, while the 2-MeImpddC peak intensity decreases. X axis: retention time in min; Y-axis: 260 nm UV absorbance in units of mAU (milli-absorbance units).

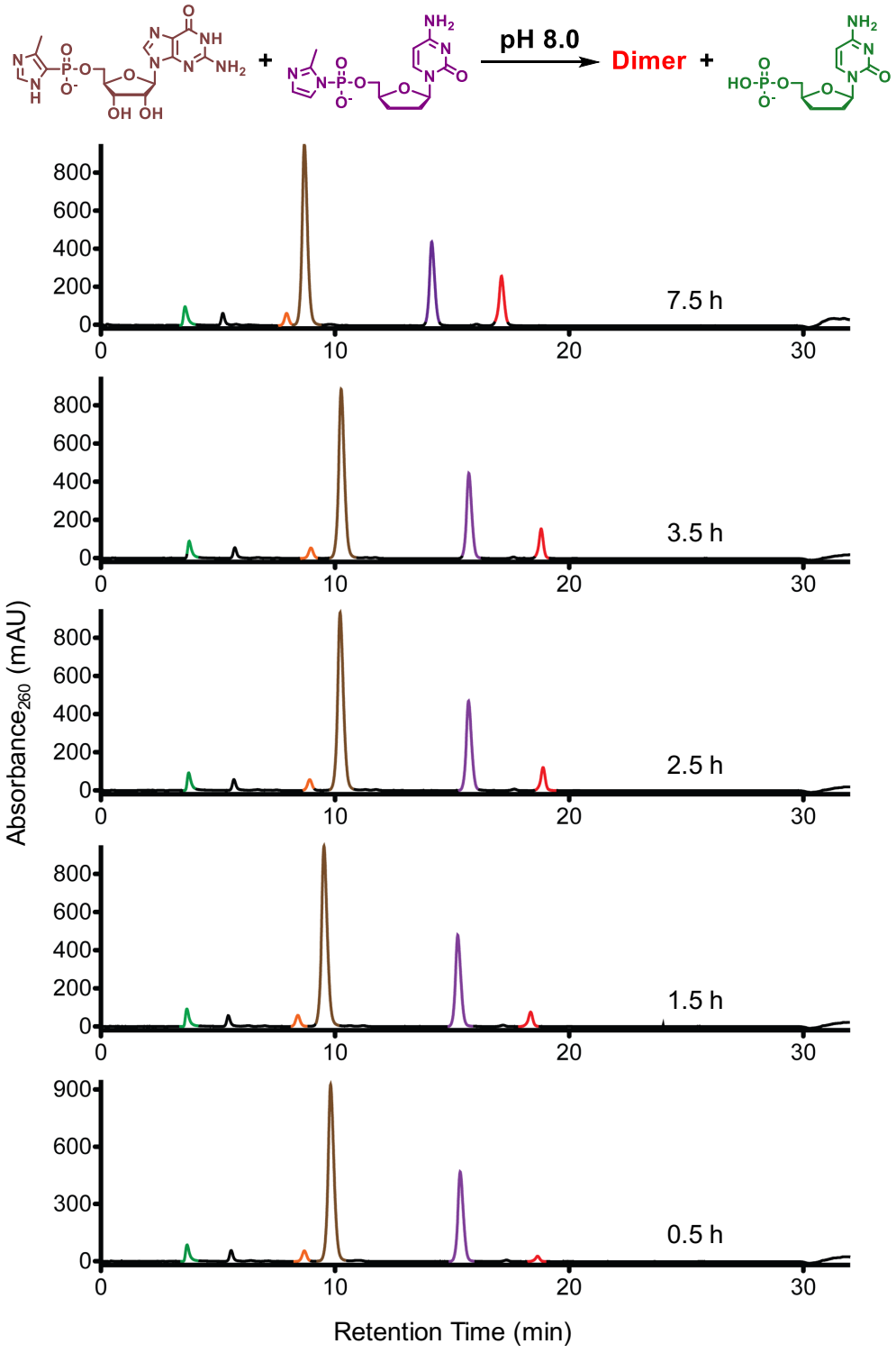


Figure D.16. **2-MeImpddC** was coincubated with **ICG** at **pH 8.0**. The formation of ddC-ICG dinucleotide was assessed over **7.5 h**. The color coding of the LC peaks, and the X- and Y-axes of the stacked liquid chromatograms are the same as **Figure D.15.** All peak assignments were performed by means of mass spectrometry.

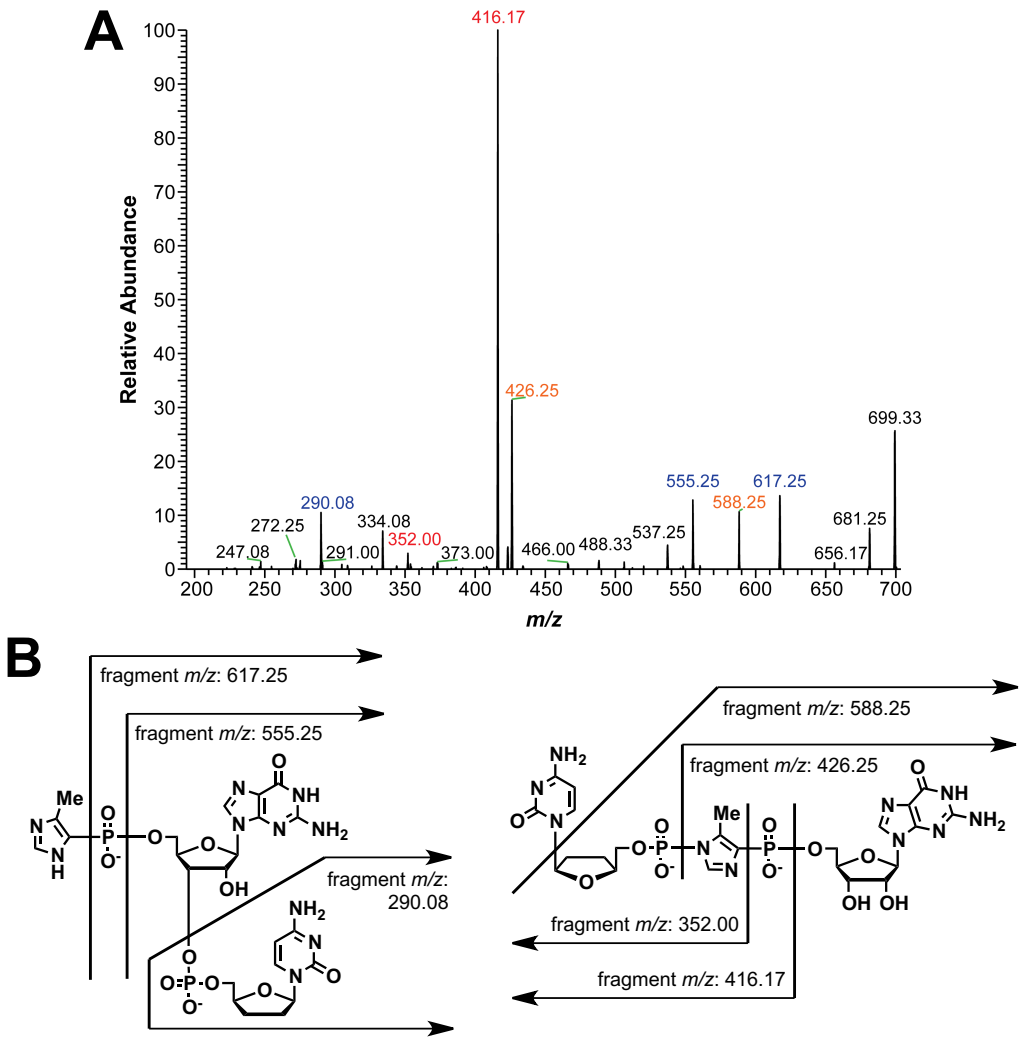


Figure D.17. (a) MS^2 spectrum of ddC-ICG dinucleotide. 2-MeImpddC and ICG were coincubated at pH 6.3 for 2 h. The ddC-ICG dinucleotide peak, with a retention time of 19.1 min (**Figure 4.6.B**), has a major *m/z* peak of 699.33. This peak was then isolated as the parent ion for the subsequent MS^2 analysis and partially fragmented. The *m/z* peaks labeled red arise exclusively from fragments of the 5′-5′-imidazole-bridged ddC-ICG dinucleotide; the *m/z* peaks labeled blue arise from fragments of the 3′-5′- or 2′-5′-phosphodiester linked ddC-ICG dinucleotide, and the *m/z* peaks labeled orange could originate from either regioisomer. (b) Possible structures of the ddC-ICG dinucleotide. **Left**: One of the possible regioisomers of ddC-ICG dinucleotide, with the guanosine and dideoxycytidine nucleotides linked *via* 3′-5′ or 2′-5′ phosphodiester linkages. **Right**: The desired regioisomer of the ddC-ICG dinucleotide, with the guanosine and dideoxycytidine nucleotides linked *via* 5′-5′-imidazole-diphosphate bridge. The bolded arrows indicate the fragmented bond and the associated fragment *m/z*.

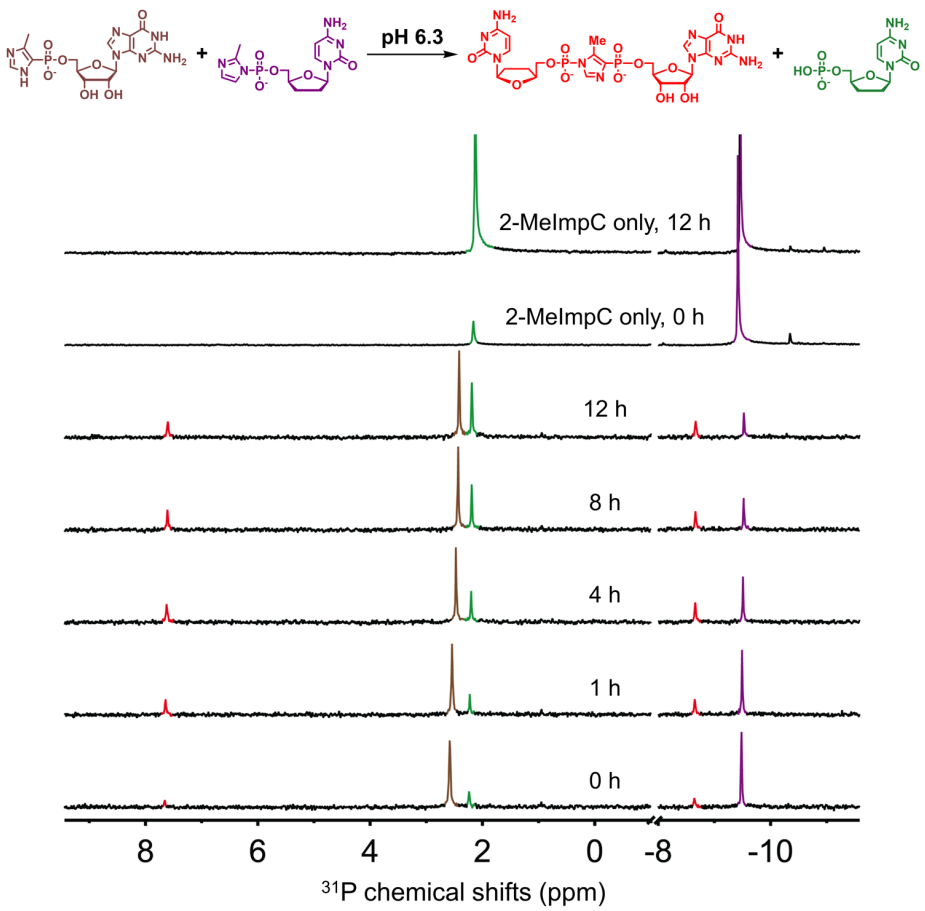


Figure D.18. ^{31}P NMR time course analysis of the **2-MeImpddC-ICG** coincubation experiment performed at **pH 6.3**. ^{31}P NMR resonances colored red, purple, brown, and green correspond to the resonances of ddC-ICG dinucleotide, 2-MeImpddC, ICG and ddCMP respectively. The formation of ddC-ICG dinucleotide was assessed over **12 h**. The 32 to 9 ppm, –1 to –8 ppm and –12 to –26 ppm regions of the ^{31}P spectra, which have no appreciable peaks, are omitted for clarity. The bottom five traces show the ^{31}P spectra of the 2-MeImpddC-ICG coincubation experiment at their indicated time points; the top two traces show the result of the 2-MeImpddC degradation experiment (negative control), in which ICG was omitted. All spectra were indirectly referenced to Na_2HPO_4 in accordance to methods outlined in **Section D.5.1.2**. The negative control experiment clearly showed that the increasing peak at 2.2 ppm was ddCMP, the hydrolysis product of 2-MeImpddC. As expected, the 2-MeImpddC peak at –9.5 ppm decreases over time, due to hydrolysis and reaction with ICG. Throughout the time course, the two new observed peaks at –8.7 and 7.6 ppm are roughly comparable to one another in terms of both peak height and peak integration; the peak intensities of the two new resonances increase from 0 to 4 h, after which they stagnate. Judging from the facts that 1) ddC-ICG dinucleotide is shown to form by means of mass spectrometry, 2) only two new commensurate resonances are observed, and 3) no new resonance is observed in the ppm range wherein phosphodiester resonances typically resides (1 to –5 ppm), we preliminarily conclude that the regiochemistry of the ddC-ICG linkage is *via* the proposed 5′-5′-imidazole bridge.

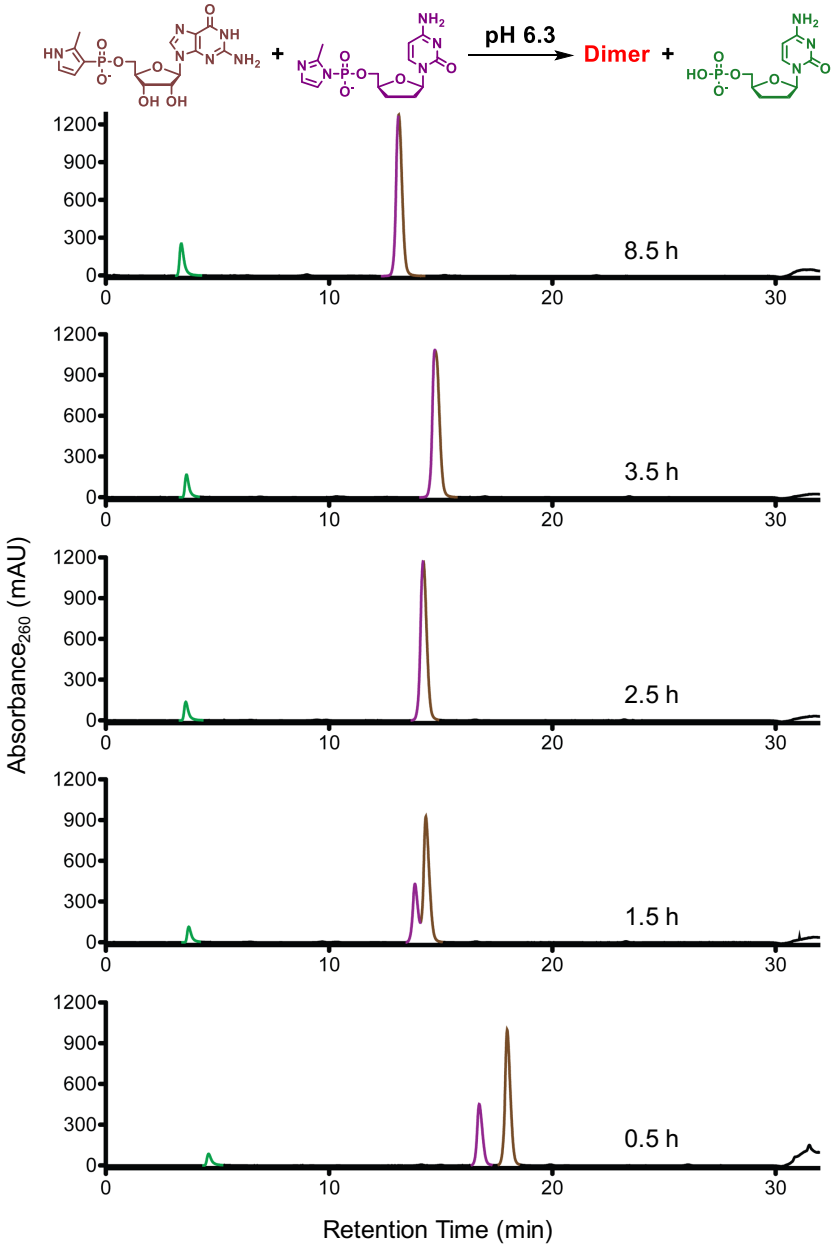


Figure D.19. 2-MeImpddC was coincubated with **PYG** at **pH 6.3**. Possible formation of ddC-PYG conjugate was assessed over a duration of **8.5 h**. The purple, brown, and green peaks correspond to 2-MeImpddC, PYG and ddCMP, respectively. All peak assignments were performed by means of mass spectrometry. The elution times of the initially well-resolved PYG and 2-MeImpddC peaks decrease as the coincubation progresses, and the two peaks coelute from the 2.5 h time point and beyond. Regardless of the observed drifts in peak elution time, with the exception of ddCMP, no other peaks are observed to grow in intensity. To confirm that 2-MeImpddC does not react with PYG, the expected monoanionic *m/z* of the ddCMP-PYG dinucleotide (calculated *m/z*: 697.14; mass window: ±1 dalton) was used to search the total ion count (TIC) spectra accompanying the shown liquid chromatographs for any trace dinucleotide formation. The focused *m/z* search reveals that 2-MeImpddC does not react with PYG to any appreciable extent. This indicates that 2-MeImpddC hydrolysis is the only major reaction taking place throughout the 2-MeImpddC-PYG coincubation, and that the nucleophilic 3′-OH, 2′-OH and guanine exocyclic amine groups of PYG does not react appreciably with the activated phosphate of 2-MeImpddC. Thus this experiment provides the first clues that dinucleotide formation between ICG with 2MeImpddN occurs primarily *via* a 5′-5′-imidazole linkage, and that the regioselectivity for dinucleotide formation is high.

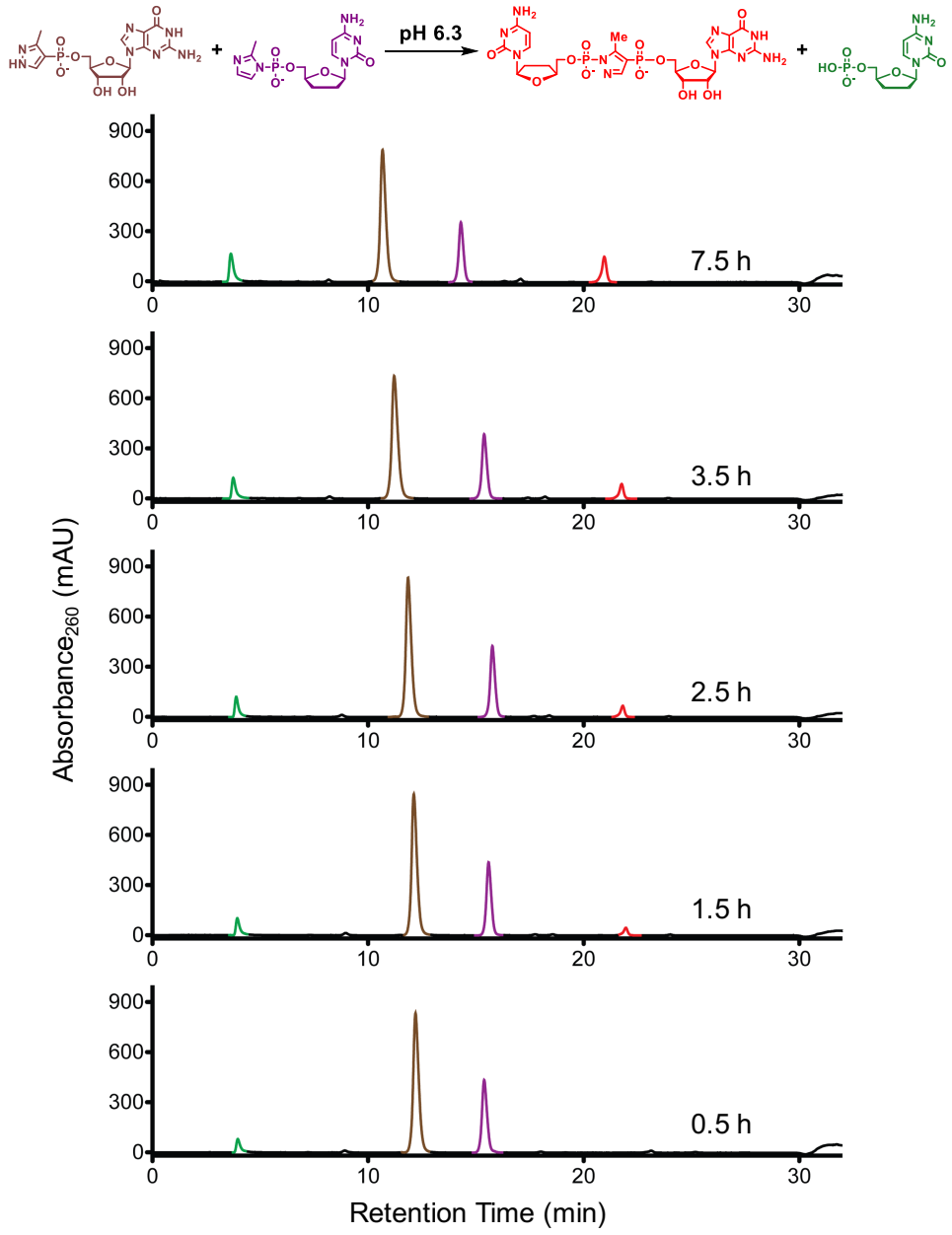


Figure D.20. 2-MeImpddC was coincubated with **PZG** at **pH 6.3**. The formation of ddC-PZG conjugate was assessed over **7.5 h**. The red, purple, brown and green peaks correspond to ddC-PZG dinucleotide, 2-MeImpddC, PZG and ddCMP respectively. All peak assignments were performed by means of mass spectrometry. The ddC-PZG dinucleotide peak grows in intensity over the course of the coincubation experiment. Given the results shown in **Figure D.19**, we hypothesize the pyrazolyl group of PZG reacts with 2-MeImpddC to form the 5′-5′-pyrazole-bridged ddC-PZG dinucleotide.

D.5.2 ^{31}P NMR time course for coincubation experiments with 2-MeImpC & nonhydrolyzable analogues

Dinucleotide formation between 2-MeImpC and various nonhydrolyzable 2-MeImpG analogues were monitored by ^{31}P NMR for 9 or 10 h. Experimental conditions and methods were identical to those outlined in **Section D.5.1.2**, except that 2MeImpddC was replaced by 2-MeImpC. The ^{31}P resonance of the added Na_2HPO_4 at pH 6.3 and pH 8.0 was determined to be 0.87 and 2.62 ppm, respectively, using an external reference of 85% H_3PO_4 (0 ppm by definition) in D_2O contained in a coaxial NMR tube. See **Figures D.21–D.26** for representative data.

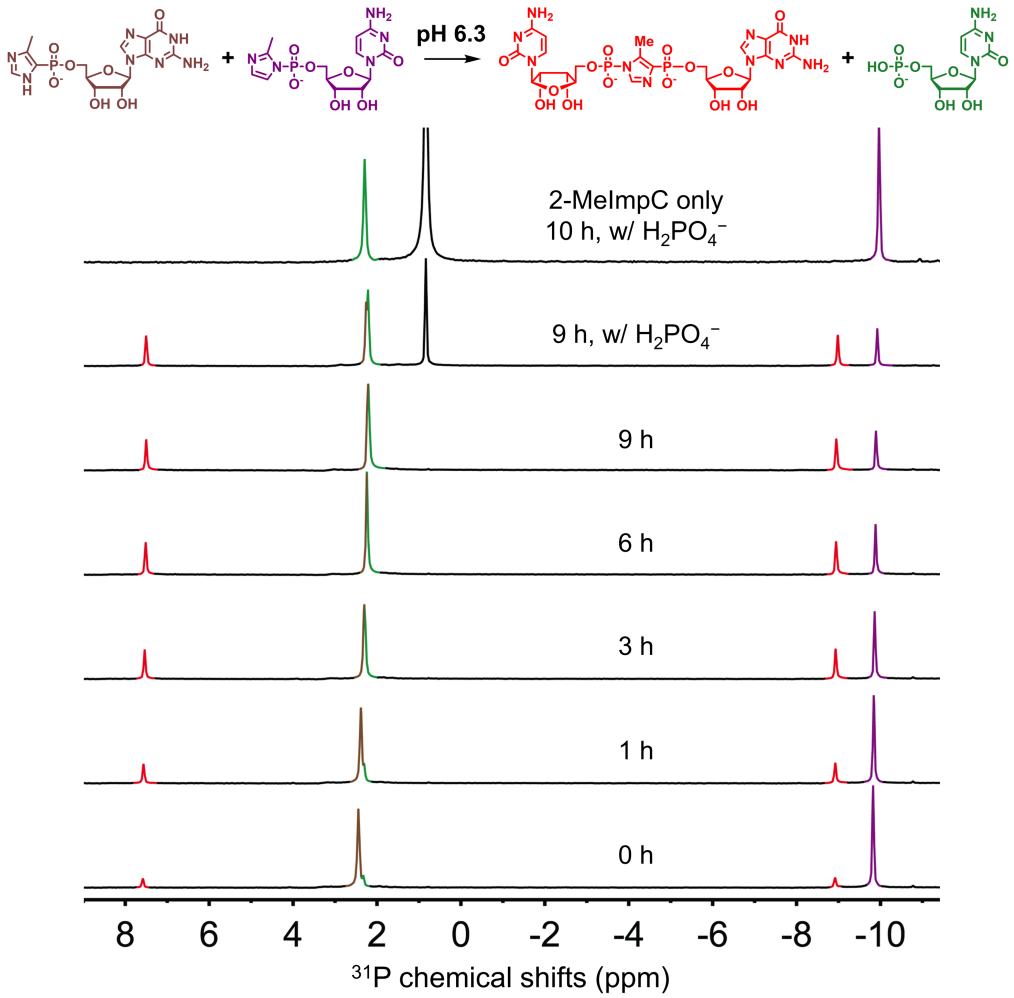


Figure D.21. ^{31}P NMR time course analysis of the 2-MeImpC-**ICG** coincubation experiment performed at **pH 6.3**. ^{31}P NMR resonances colored red, purple, brown, and green correspond to the resonances of rC-ICG dinucleotide, 2-MeImpC, ICG and CMP respectively. The formation of rC-ICG dinucleotide was assessed over **9 h**. The ^{31}P spectral region between 32 to –26 ppm was monitored; the 32 to 9 ppm and –9 to –26 ppm regions of the spectra, which have no appreciable peaks throughout the time course analysis, are omitted for clarity. The bottom five traces show the ^{31}P spectra of the 2-MeImpC-ICG coincubation experiment at their indicated time points. The second top trace shows the reacquired Na_2HPO_4-referenced ^{31}P spectra of the 2-MeImpC-ICG sample at the 9 h time point (Na_2HPO_4 peak was referenced to 0.87 ppm). The top trace shows the Na_2HPO_4-referenced ^{31}P trace of the negative control experiment, in which 2-MeImpC and MES buffer were coincubated for 10 h to monitor 2-MeImpC hydrolysis. The top trace clearly shows that the increasing peak at 2.25 ppm is CMP, the hydrolysis product of 2-MeImpC. At pH 6.3, the phosphorus resonance of ICG overlaps

Figure D.21. (continued) significantly with that of CMP. The 2-MeImpC peak at –9.9 ppm decreases over time, due to hydrolysis and reaction with ICG. Throughout the time course, the two new observed peaks at -8.9 and 7.5 ppm are roughly comparable to one another in terms of both peak height and peak integration; their peak intensities increase appreciably from 0 to 3 h, after which they plateau. Judging from the fact that 1) ddC-ICG dinucleotide was shown to form by means of mass spectrometry (See **Figure D.37**), and 2) only two new commensurate ^{31}P resonances are observed, we preliminarily conclude that the formation of the new rC-ICG dinucleotide proceeds with good regioselectivity. As a negative control experiment, PYG was coincubated with 2-MeImpC; no dinucleotide formation is observed (**Figure D.23–D.24**), suggesting that the nucleophilic 3′-OH, 2′-OH and guanine exocyclic amine groups of ICG should not react appreciably with 2-MeImpC. We therefore conclude preliminarily that the regiochemistry of the rC-ICG linkage is *via* the proposed 5′-5′-imidazole bridge.

The rates of dinucleotide formation and degradation at pH 6.3 were assessed by LCMS analysis. Associated data are shown in **Figures D.37–D.41** (**Section D.5.4**)

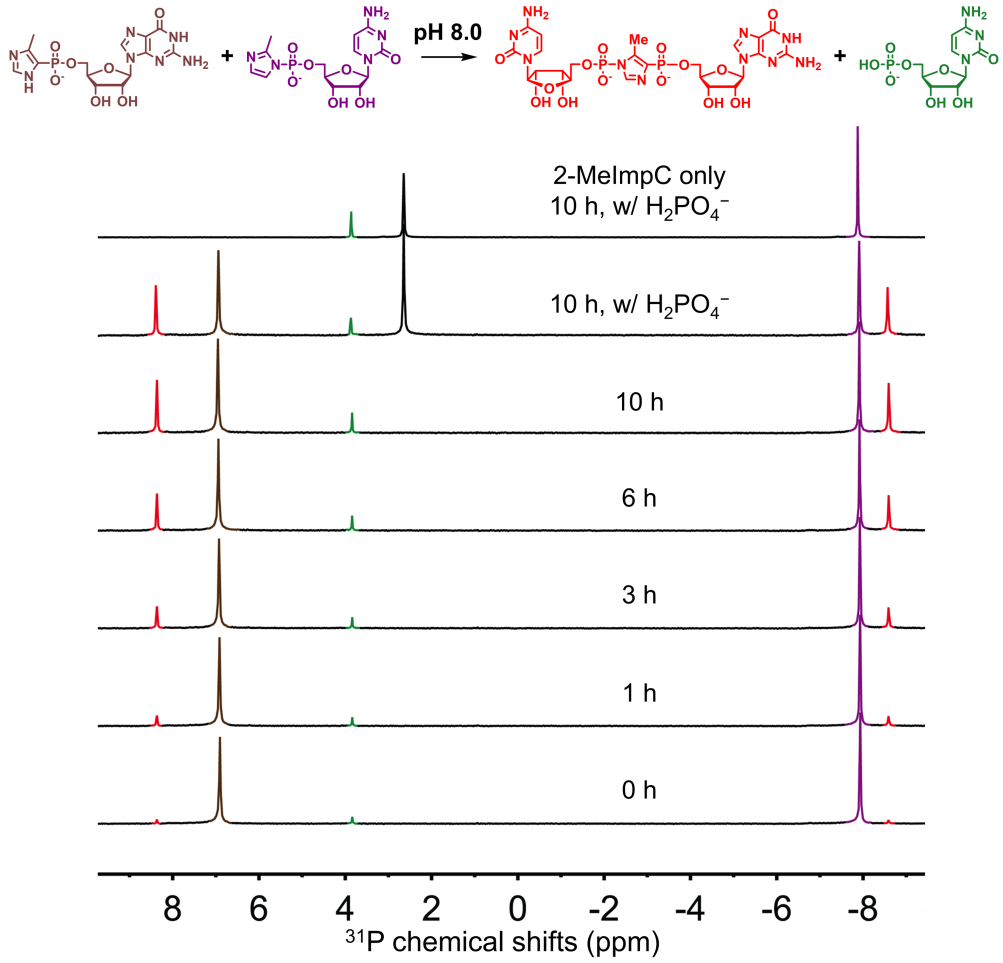


Figure D.22. ^{31}P NMR time course analysis (**10 h**) of the 2-MeImpC-**ICG** coincubation experiment performed at **pH 8.0**. ^{31}P NMR resonances colored red, purple, brown, and green correspond to the resonances of rC-ICG dinucleotide, 2-MeImpC, ICG and CMP respectively. At pH 8.0, the resonances of all phosphorus-containing species are baseline-resolved. The spectral regions between 30 to 9 ppm and –9 to –26 ppm are omitted for clarity. The bottom five traces show the ^{31}P spectra of the 2-MeImpC-ICG coincubation at their indicated time points. The second top trace shows the reacquired, Na_2HPO_4-referenced spectrum of the coincubated sample at the 10 h time point. The top trace shows the Na_2HPO_4-referenced trace of the 2-MeImpC degradation experiment (negative control) at the 10 h time point (Na_2HPO_4 peak was determined to be 2.62 ppm). The top trace clearly shows that the increasing peak at 3.8 ppm was CMP. Throughout the time course, the two new observed peaks at –8.7 and 8.4 ppm are roughly comparable to one another both in terms of peak height and peak integration; their peak intensities increase throughout the time course. These two new resonances are attributed to the formation of the 5′-5′-imidazole-bridged CpICG dinucleotide. The rates of CpICG formation and degradation at pH 8.0 are shown in **Figures D.37–D.41** (**Section D.5.4**).

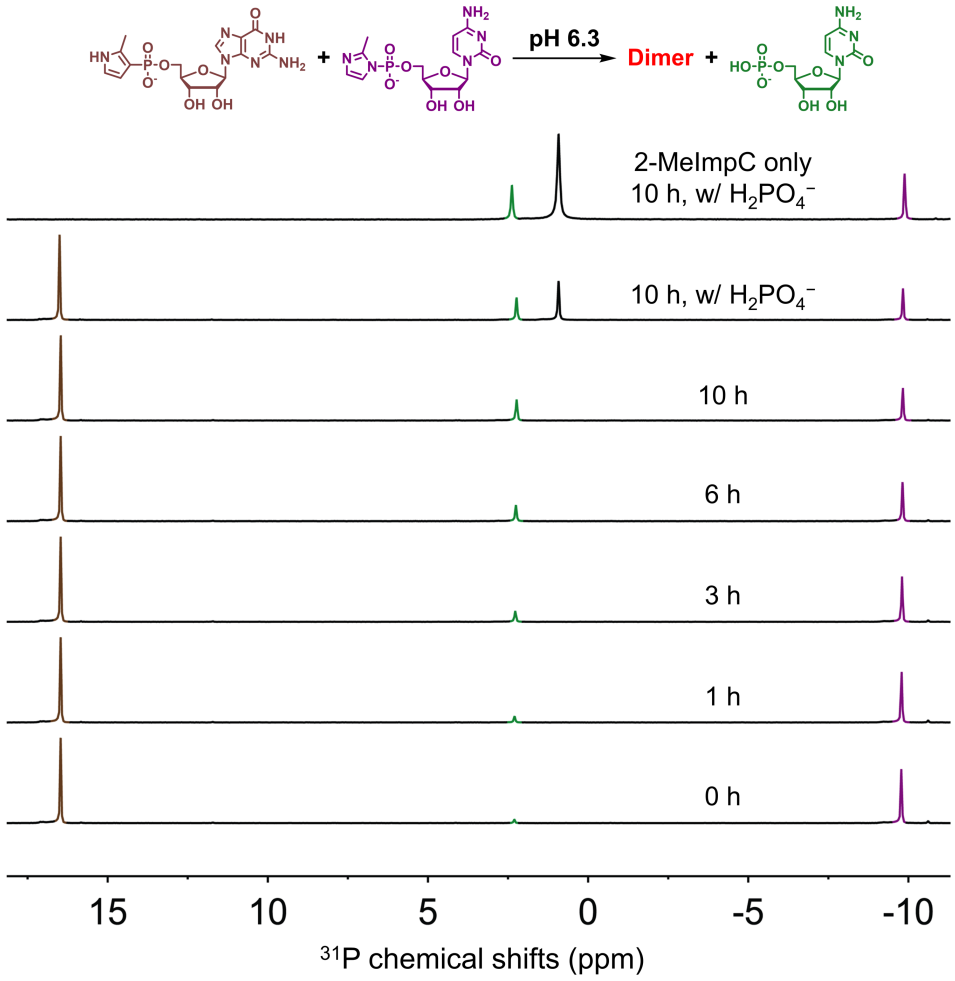


Figure D.23. ^{31}P NMR time course analysis (**10 h**) of the 2-MeImpC-**PYG** coincubation experiment performed at **pH 6.3**. ^{31}P NMR resonances colored purple, brown, and green are assigned to 2-MeImpC, PYG and CMP respectively. The spectral regions between 32 to 18 ppm and –11 to –26 ppm are omitted for clarity. The bottom five traces show the ^{31}P spectra of the 2-MeImpC-PYG coincubation at their indicated time points. The second top trace shows the reacquired, Na_2HPO_4-referenced spectrum of the coincubated sample at the 10 h time point. The top trace shows the Na_2HPO_4-referenced ^{31}P trace of the 2-MeImpC degradation experiment (negative control) at the 10 h time point. The increasing peak at 2.2 ppm is CMP; no other substantial phosphorus resonance other that of CMP can be detected throughout the time course. This suggests that PYG does not react with 2-MeImpC to form dinucleotides, and that the potentially nucleophilic groups of PYG, such as the ribose 2′-OH and 3′-OH, and the guanine exocyclic amine moieties, do not react with 2-MeImpC.

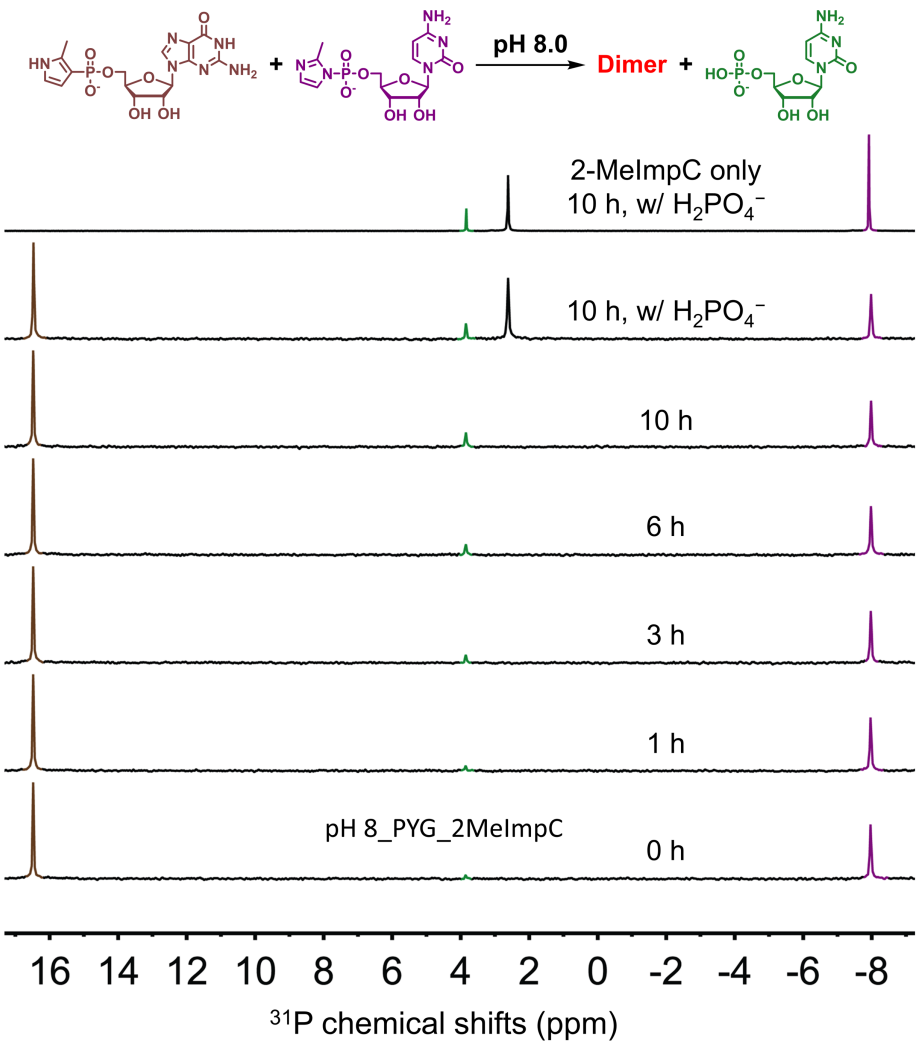


Figure D.24. ^{31}P NMR time course analysis (**10 h**) of the 2-MeImpC-**PYG** coincubation experiment performed at **pH 8.0**. ^{31}P NMR resonances colored purple, brown, and green are assigned to 2-MeImpC, PYG and CMP respectively. The spectral regions between 30 to 17 ppm and –9 to –28 ppm are omitted for clarity. The bottom five traces show the ^{31}P spectra of the 2-MeImpC-PYG coincubation at their indicated time points. The second top trace shows the reacquired Na_2HPO_4-referenced spectrum of the coincubated sample at the 10 h time point. The top trace shows the Na_2HPO_4-referenced ^{31}P trace of the 2-MeImpC degradation experiment (negative control) at the 10 h time point. The increasing peak at 3.9 ppm is CMP, and no other phosphorus resonance other that of CMP can be detected throughout the time course. This suggests that PYG does not react with 2-MeImpC to form dinucleotides at pH 8.0.

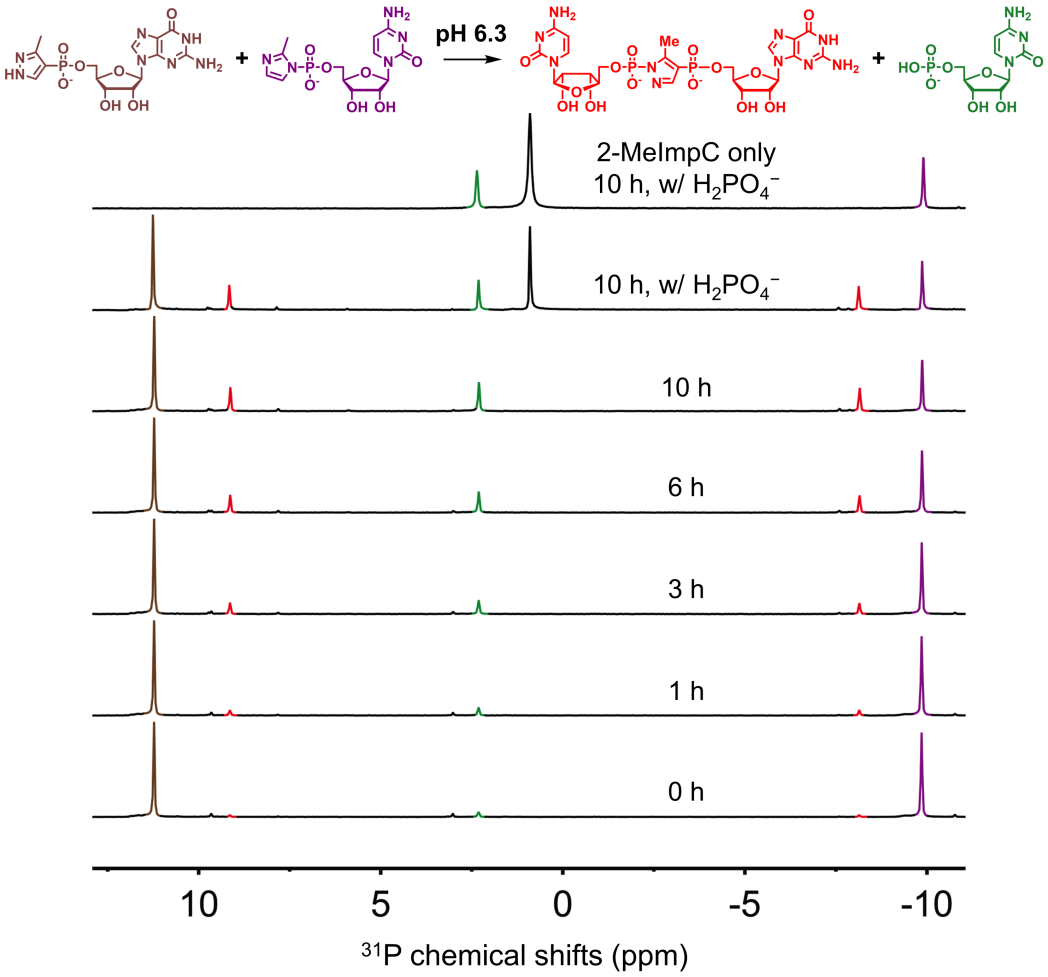


Figure D.25. ^{31}P NMR time course analysis (**10 h**) of the 2-MeImpC-**PZG** coincubation experiment performed at **pH 6.3**. ^{31}P NMR resonances colored purple, brown, and green are assigned to 2-MeImpC, PZG and CMP respectively. The spectral regions between 32 to 13 ppm and –11 to –26 ppm are omitted for clarity. The bottom five traces show the ^{31}P spectra of the 2-MeImpC-PZG coincubation at their indicated time points. The second top trace shows the reacquired, Na_2HPO_4-referenced spectrum of the coincubated sample at the 10 h time point. The top trace shows the Na_2HPO_4-referenced trace of the 2-MeImpC degradation experiment (negative control) at the 10 h time point. The increasing peak at 2.3 ppm is CMP. Two new peaks are observed at –8.2 and 9.1 ppm, and their peak intensities increase throughout the time course. Given the results shown in **Figures D.23−D.24** that the ribose and nucleobase of PZG are unlikely to react with 2-MeImpC, we hypothesized that these two peaks originate from the 5′-5′-pyrazole-bridged CpICG dinucleotide. Despite the fact that PZG can react with 2-MeImpC to yield the 5′-5′-pyrazole bridged dinucleotide, in primer extension assays

Figure D.25. (continued) utilizing PZG as the downstream mononucleotides (**Figure 4.4**), PZG does not catalyze primer extension to any significant extent. We hypothesize that the 5′-5′-pyrazole-bridge, which has a protonation pK_a similar to or lower than 1.94 (**Figure 4.1**; **Figure D.2**), would not be protonated under the examined reaction pH. Thus, nucleophilic attack of the primer 3′-OH group on the pyrazole-bridge would result in the displacement of an anionic pyrazolo-phosphonate group, which is unfavored.

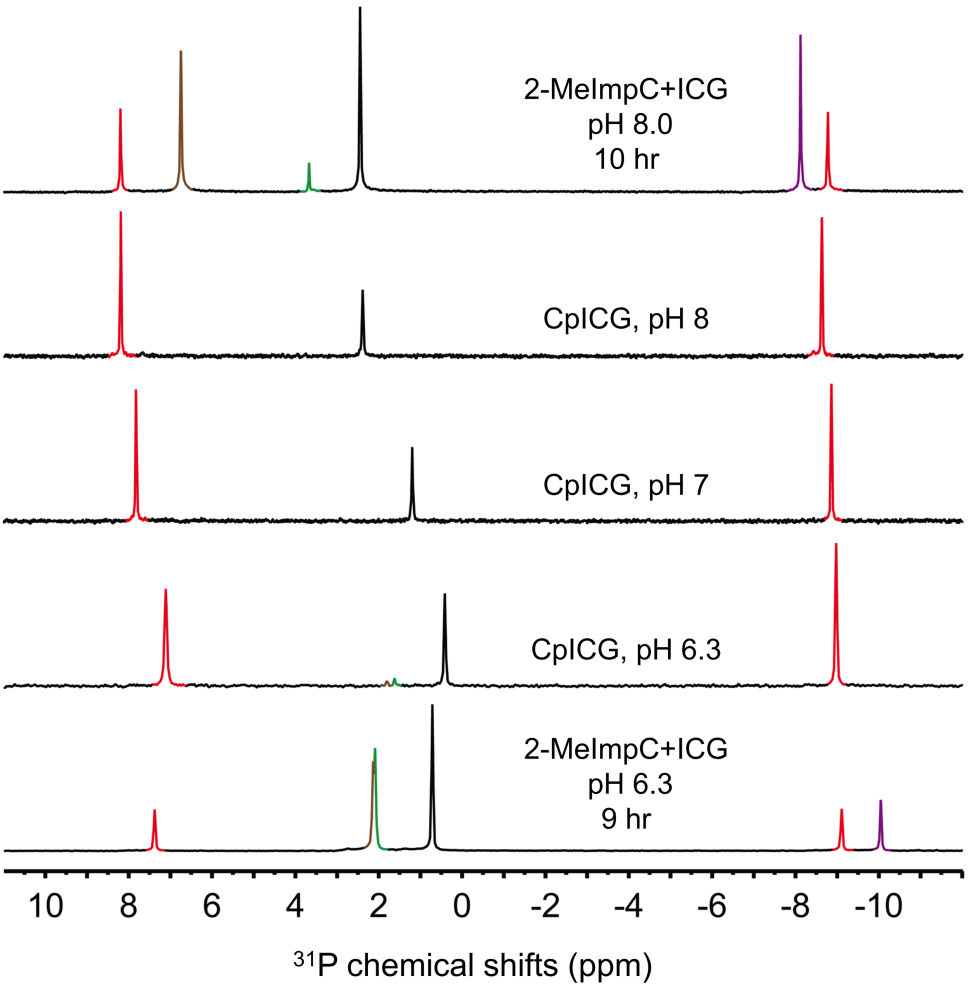


Figure D.26. The purified CpICG dinucleotide was subjected to variable pH-^{31}P NMR analysis to ascertain the possible effect of 5′-5′-imidazole bridge protonation state on the ^{31}P chemical shifts of the two phosphates of CpICG. 100 µL of 10 mM CpICG in D_2O was prepared, and was referenced with deuterated Na_2HPO_4; the ^{31}P resonances of Na_2HPO_4 in this deuterated sample at pH 6.3, 7.0 and 8.0 were determined to be 0.63, 1.38 and 2.61 ppm, respectively, using an external reference of 85% H_3PO_4 in D_2O. The deuterated CpICG sample was first adjusted to pD 8, followed by progressive acidification to pD 6.3 with DCl in D_2O. Throughout the course of sample acidification, the two phosphorus peaks of CpICG shifted downfield (the middle three traces): the upfield ^{31}P resonance of CpICG shifted 0.3 ppm downfield, while the downfield ^{31}P resonance of CpICG shifted 0.8 ppm downfield. The phosphorus resonances of the purified CpICG at pD 8.0 and pD 6.3 are compared with the resonances of CpICG in the corresponding 2-MeImpC-ICG coincubation experiments (**Figure D.21−D.22**; top and bottom traces). The resonance values of CpICG obtained from either source are comparable to one another, with a maximal resonance shift of 0.4 ppm.

D.5.3 Synthesis, purification and NMR analysis of CpICG

2-MeImpC and ICG were coincubated, in 770 μL scales, with the following components: 75 mM nonhydrolyzable analogues, 75 mM 2-MeImpC, and 200 mM Na•HEPES (pH 8.0). The reaction mixture was adjusted to pH 8.0 using minimal amounts of aqueous HCl or NaOH, followed by incubation at room temperature for 24 h. The reaction mixture was then split into five equal portions and purified by preparative-scale HPLC on a Varian Prostar 210 HPLC system across an Agilent Eclipse XDB C18 column (250 mm × 21.2 mm i.d., 7 μm particle size). The aqueous mobile phase was 25 mM aqueous triethylammonium bicarbonate buffer at ~pH 8.5, and organic mobile phase was acetonitrile. Organic mobile phase was ramped linearly from 2% to 4% over 0−3 min and from 4% to 10% over 3−20 min. The flow rate was maintained at 10 mL min^{-1}. CpICG was eluted from the column at 7.5% CH_3CN in TEAB, which was promptly lyophilized at sub-zero temperature on a VirTis AdVantage 2.0 benchtop lyophilizer (SP Scientific, Warminster, PA) at <50 mTorr over 48 h to afford the titled compound (in triethylammonium cation form) as a white fluffy solid.

^{1}H and proton-decoupled ^{13}C NMR spectroscopies were performed on a Varian Inova 500 MHz spectrometer equipped with a Varian HCN triple resonance cold probe (500 MHz for ^{1}H, 125 MHz for ^{13}C). ^{1}H-^{1}H gCOSY, ^{1}H-^{13}C gc2HSQCse and ^{1}H-^{13}C gHMBCAD correlation spectroscopies were performed on an Agilent Inova 700 MHz spectrometer equipped with an Agilent triple resonance helium cold probe (700 MHz for ^{1}H, 175 MHz for ^{13}C). Proton-decoupled ^{31}P and ^{1}H-^{31}P gHMBCAD correlation spectroscopies were recorded on a Varian Inova 400 MHz spectrometer equipped with a broadband PFG (z-gradient) probe (400 MHz for ^{1}H, 161 MHz for ^{31}P). Proton and

carbon chemical shifts were reported in parts per million (ppm) values on the δ scale. The proton spectrum was referenced to residual protium in D_2O (DHO, δ = 4.79 ppm).[14] The carbon spectrum was referenced to the β carbons of triethylammonium cations (δ = 9.07 ppm).[14] The phosphorus spectrum was referenced to sodium hydrogen phosphate (δ = 0 ppm). All NMR spectra were recorded at 25 °C. Data were reported as follows: chemical shift, multiplicity (s = singlet, d = doublet, q = quartet, m = multiplet, br = broad), and integration.

^{1}H NMR (D_2O)	δ 7.90 (s, 1H), 7.87 (d, *J* = 2.2 Hz, 1H), 7.48 (d, *J* = 7.6 Hz, 1H), 5.77 (d, *J* = 6.0 Hz, 1H), 5.75 (d, *J* = 2.6 Hz, 1H), 5.71 (d, *J* = 7.6 Hz, 1H), 4.63 (t, *J* = 5.5 Hz, 1H), 4.34 – 4.31 (m, 1H), 4.19 – 4.15 (m, 1H), 4.13 – 4.03 (m, 4H), 3.99 – 3.94 (m, 3H), 3.11 (**TEAH$^+$**, q, *J* = 7.3 Hz), 2.46 – 2.42 (d, *J* = 1.2 Hz, 3H), 1.20 (**TEAH$^+$**, t, J = 7.3 Hz).
^{13}C NMR (D_2O)	δ 166.73, 161.38, 160.44, 158.16, 155.39, 152.46, 141.93 (dd, *J* = 19, 6.4 Hz), 141.24, 137.21 (dd, *J* = 34, 5.5 Hz), 132.48 (dd, *J* = 225, 9.3 Hz), 117.03, 96.72, 90.54, 87.56, 84.68 (d, *J* = 8.7 Hz), 82.57 (d, *J* = 9.2 Hz), 75.03, 74.66, 71.42, 69.79, 65.26 (d, *J* = 5.1 Hz), 64.57 (d, *J* = 4.8 Hz), 47.45 (**TEAH$^+$**), 11.31, 9.07 (**TEAH$^+$**).
^{31}P NMR (D_2O)	δ 5.80, −11.02 (pD = 8.0)
HRMS	Calc'd for $C_{20}H_{34}N_8O_7P$ $[M–H]^-$: 731.1345; Found: 731.1365

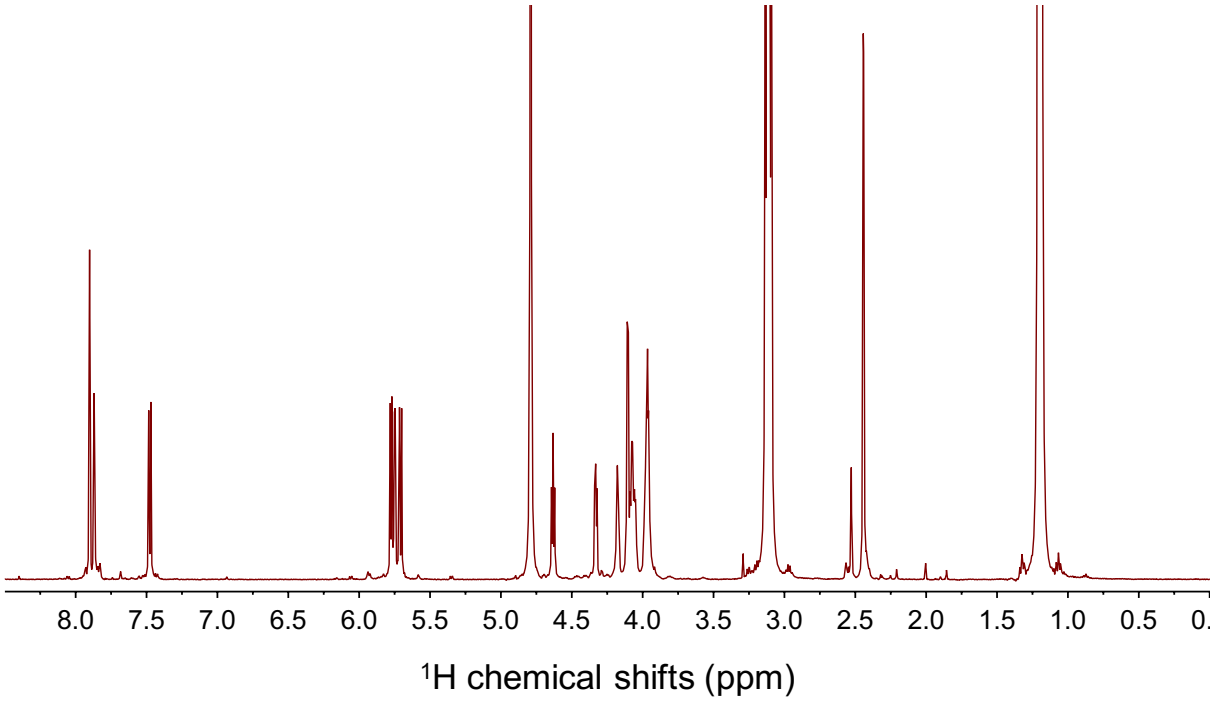


Figure D.27. Full ^{1}H NMR spectrum for CpICG**.**

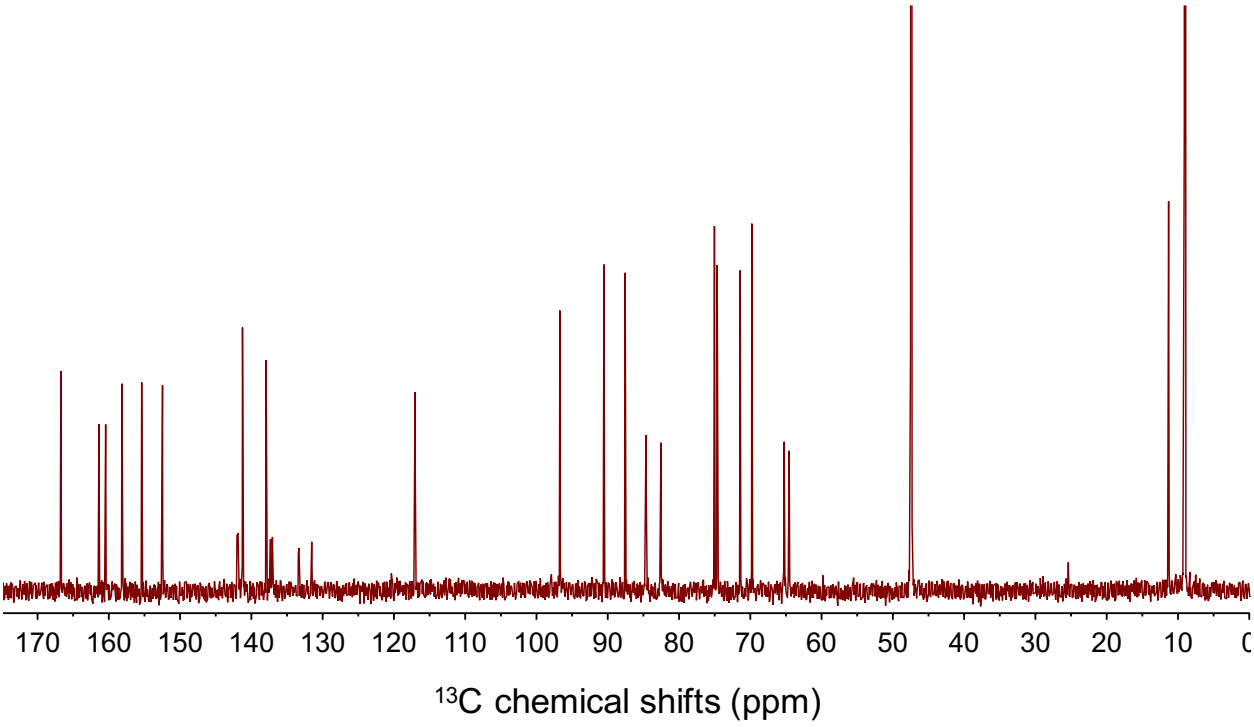


Figure D.28. Full ^{13}C NMR spectrum for CpICG**.**

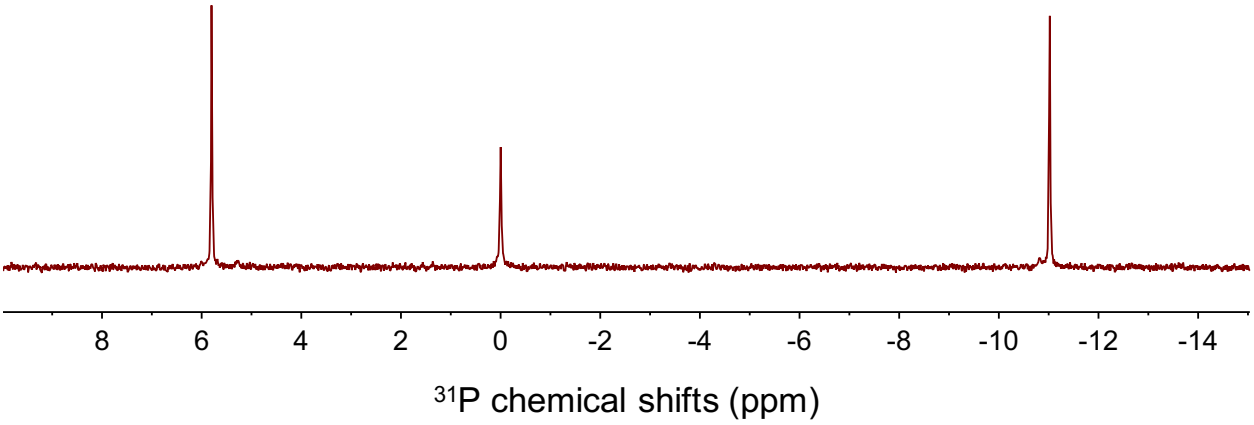


Figure D.29. Full ^{31}P NMR spectrum for CpICG (pH 8.0). The ^{31}P resonance at 0 ppm is the internal standard, Na_2HPO_4.

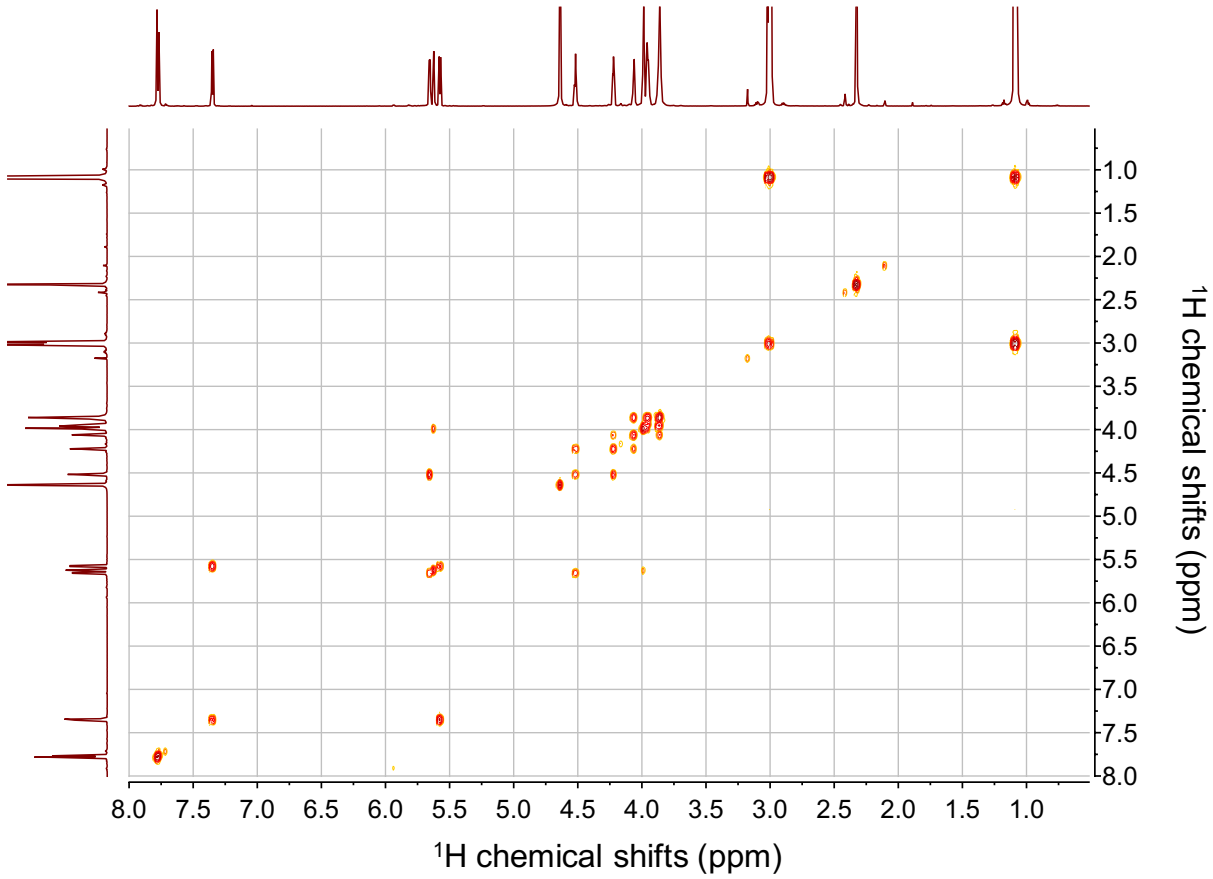


Figure D.30. Full ^{1}H–^{1}H gCOSY spectrum for CpICG.

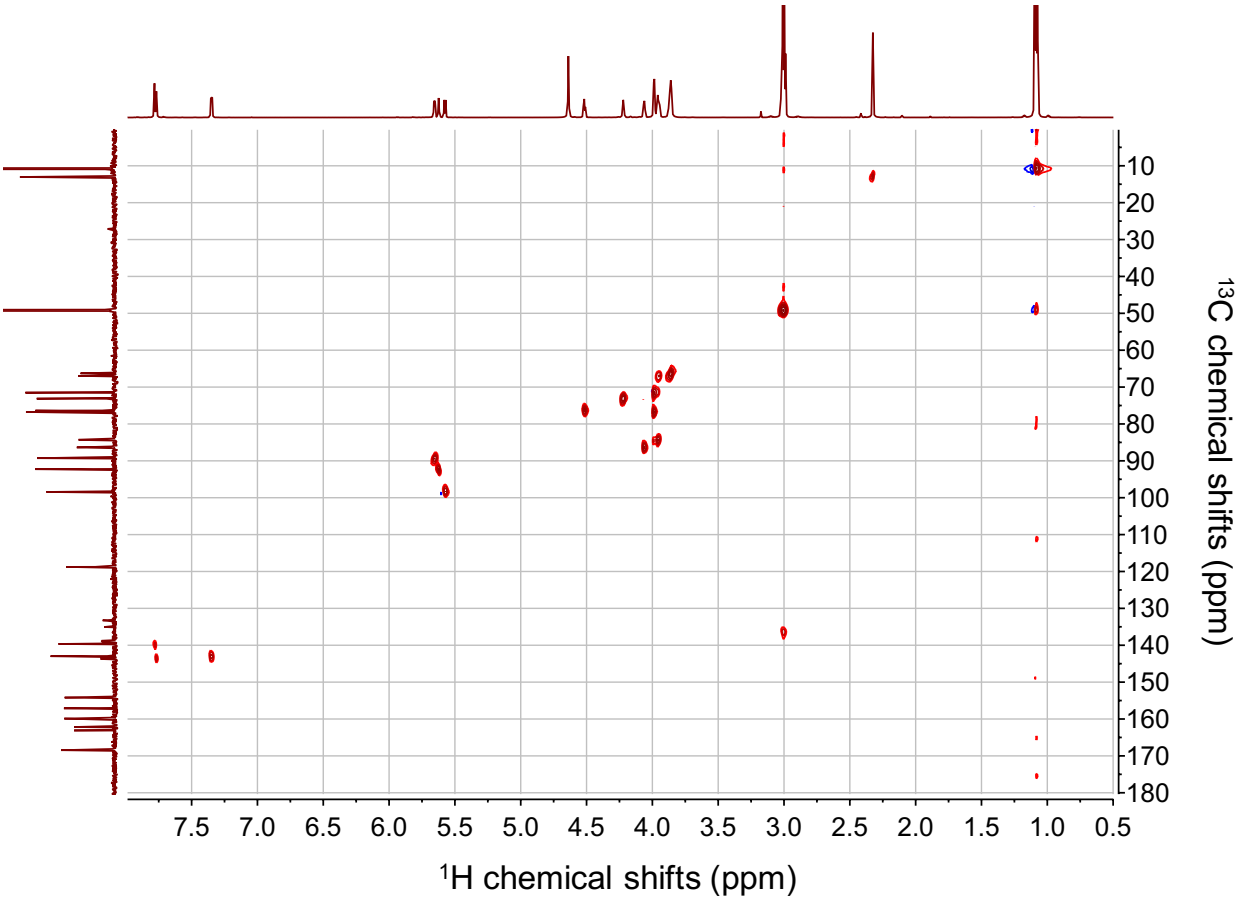


Figure D.31. Full ^{1}H–^{13}C gc2HSQCse spectrum for CpICG.

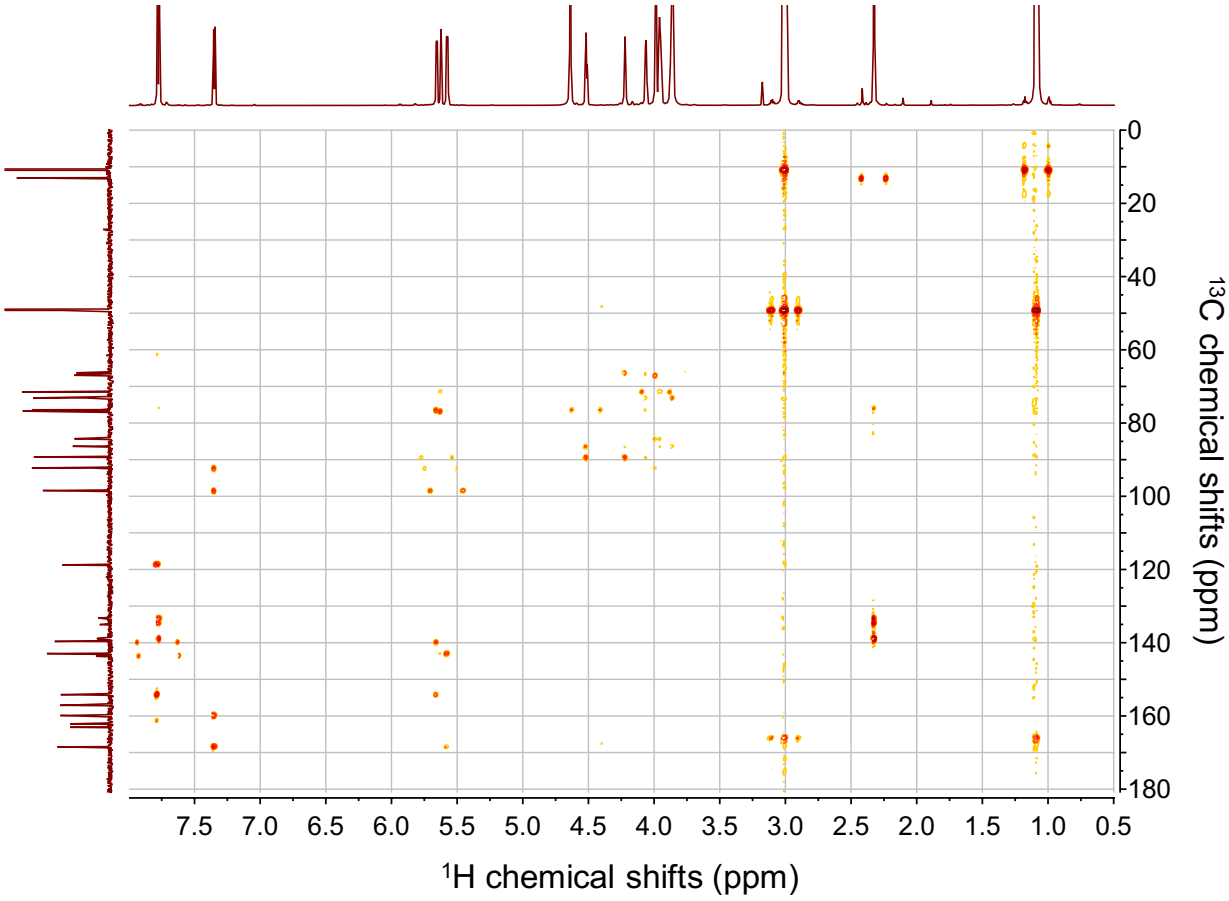


Figure D.32. Full ^{1}H–^{13}C gHMBCAD spectrum for CpICG.

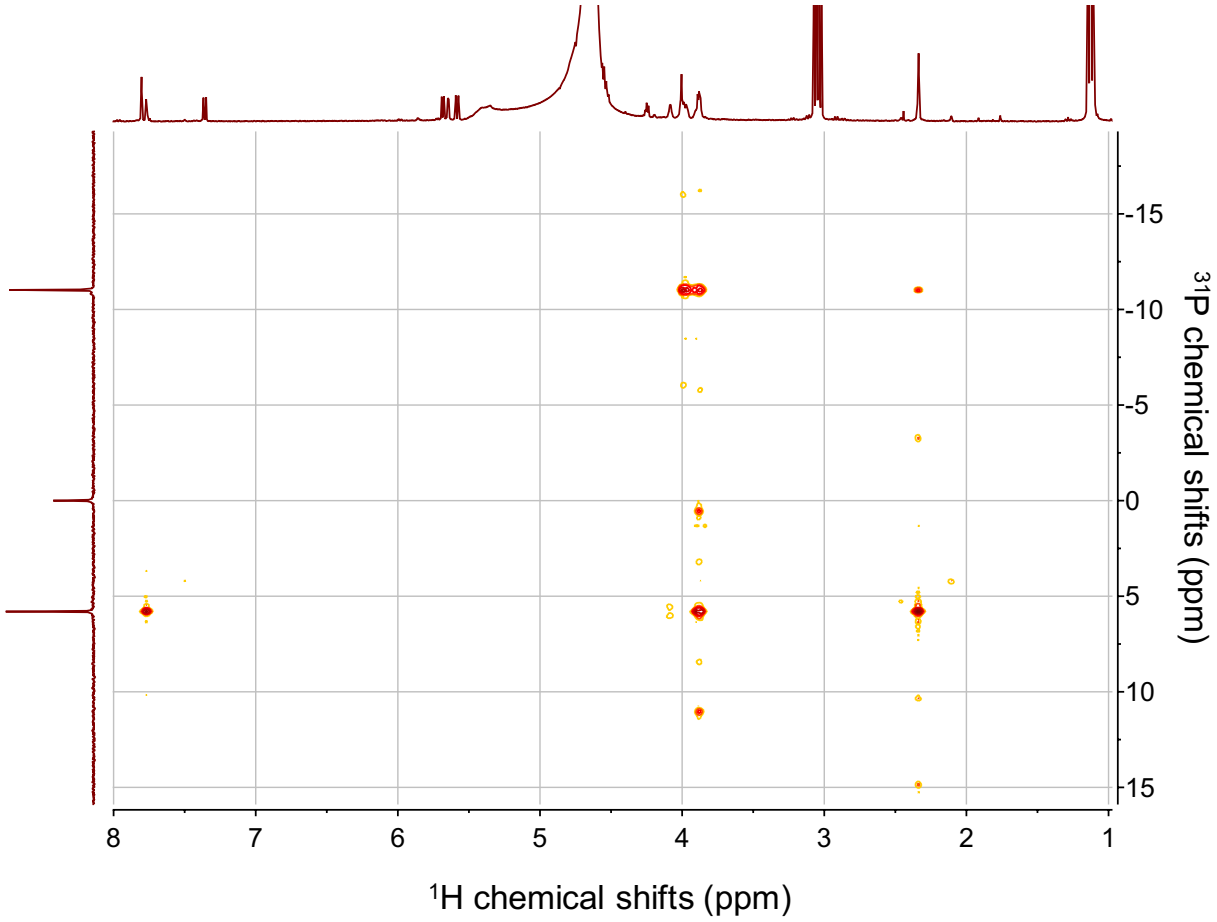


Figure D.33. Full ^{1}H–^{31}P gHMBCAD spectrum for CpICG.

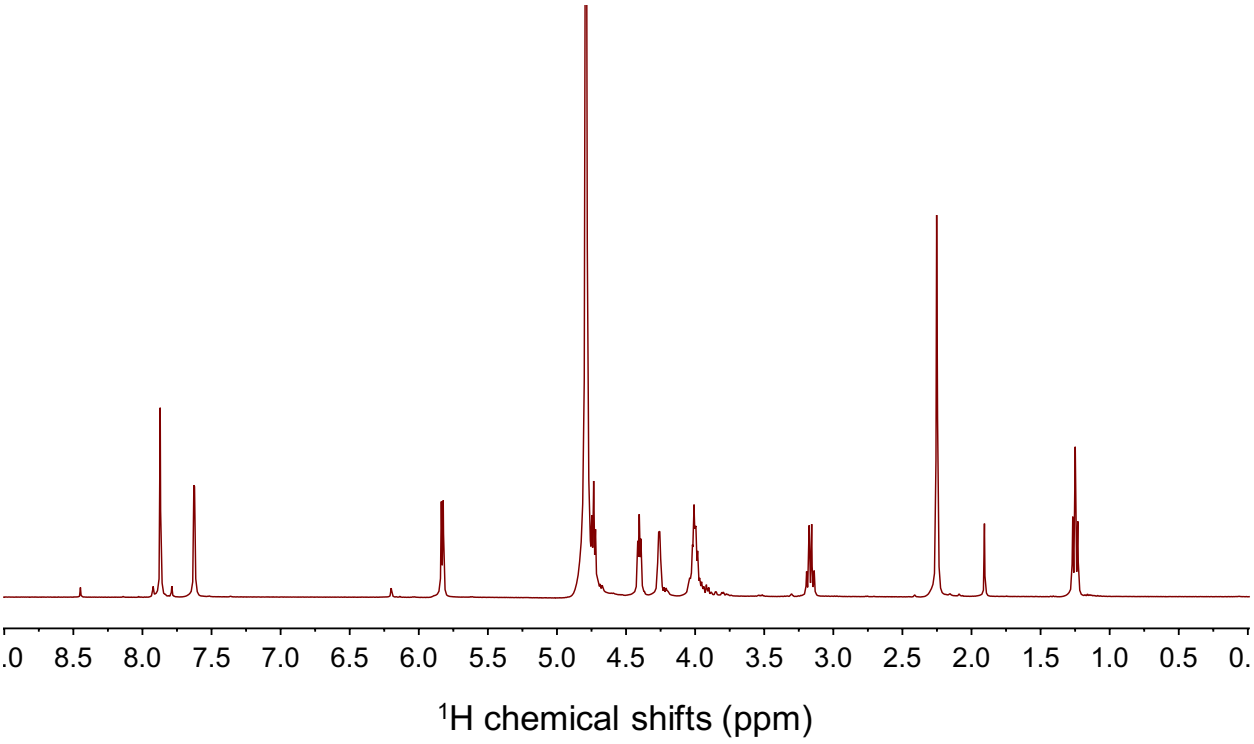


Figure D.34. Full ^{1}H NMR spectrum for ICG**.**

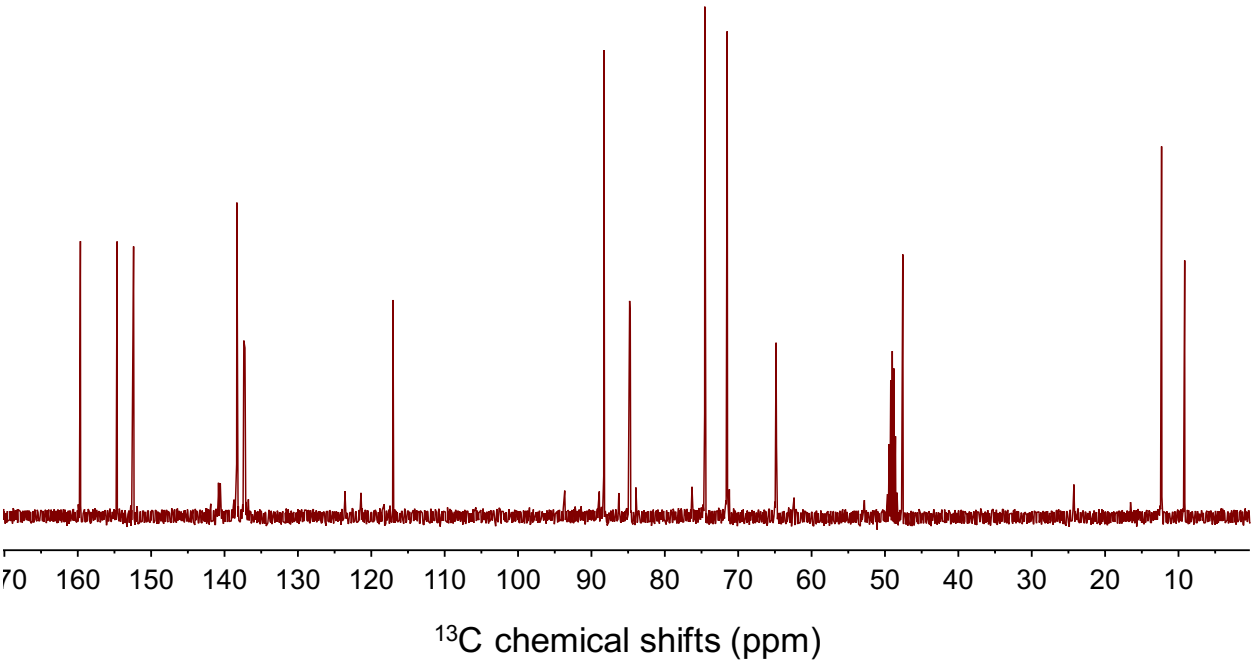


Figure D.35. Full ^{13}C NMR spectrum for ICG**.**

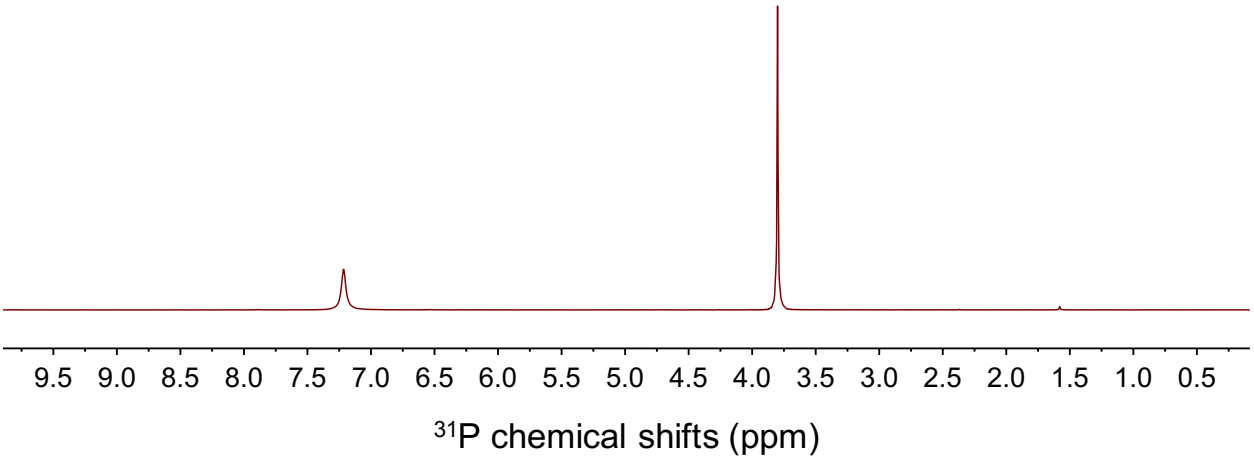


Figure D.36. Full ^{31}P NMR spectrum for ICG**.** The ^{31}P resonance at 3.8 ppm is the internal standard, trimethyl phosphate.

D.5.4 Formation and Degradation of purified CpICG

D.5.4.1 Experimental conditions: liquid chromatography coupled mass spectrometry analyses

CpICG dinucleotide formation from 2-MeImpC and ICG was monitored by the Agilent-ThermoFisher LCMS system (see **Section D.5.1.1**) for 10 h. Experimental conditions and methods were identical to those outlined in **Section D.5.1.1**, except that 2MeImpddC was replaced by 2-MeImpC. See **Figures D.37–D.38** for representative data.

Degradation of CpICG at pH 6.3 and 8.0 was monitored by LCMS. Degradation experiments were set up, in 20 µL scales, with 20 mM CpICG and 200 mM buffer (pH 6.3 Na•MES or pH 8.0 Na•HEPES). The sample pH was then promptly adjusted with aqueous hydrochloric acid or sodium hydroxide. At appropriate time points, 2 µL of the

100-fold-diluted samples (400 pmole of CpICG) were injected into the LCMS system. See **Figures D.39−D.40** for representative data.

The second-order rate of CpICG formation and first-order rate of CpICG degradation was obtained by monitoring the decay of ICG and CpICG absorbance over time, respectively. See **Figure D.41** for detailed protocol and data.

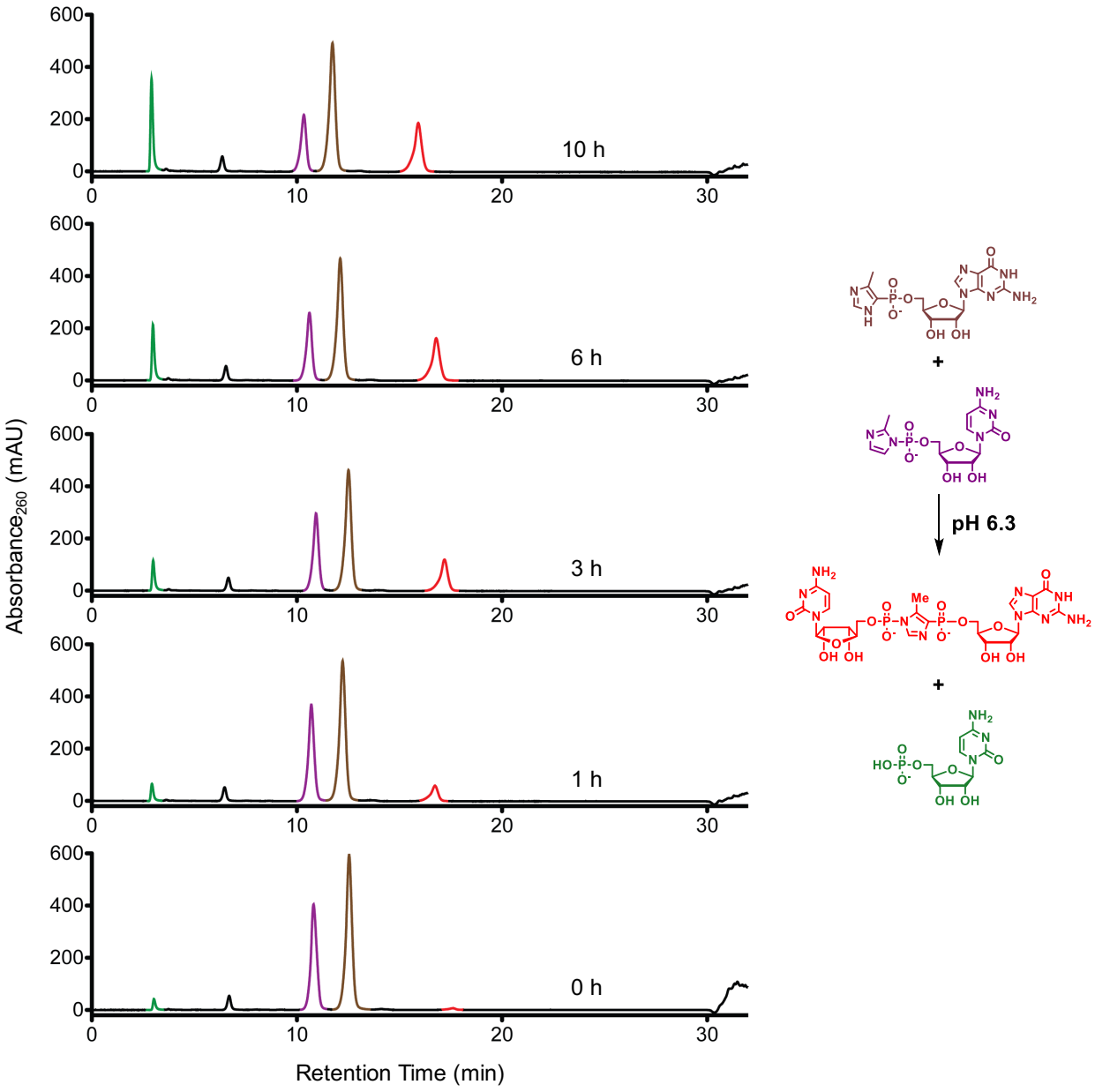


Figure D.37. **2-MeImpC** was coincubated with **ICG** at **pH 6.3**. The formation of the CpICG dinucleotide was assessed by LCMS over a duration of **10 h**. The red, purple, brown and green LC peaks correspond to CpICG dinucleotide, 2-MeImpddC, ICG and CMP, respectively. All peak assignments were performed by means of mass spectrometry. The CpICG and CMP peaks grow in intensity over the duration of the coincubation experiment, while the 2-MeImpC peak intensity decreases. The decrease in ICG peak area was used to calculate the second order rate constant for CpICG formation, and relevant data are shown in **Figure D.41**. X axis: retention time in min; Y-axis: 260 nm UV absorbance in units of mAU (milli-absorbance units).

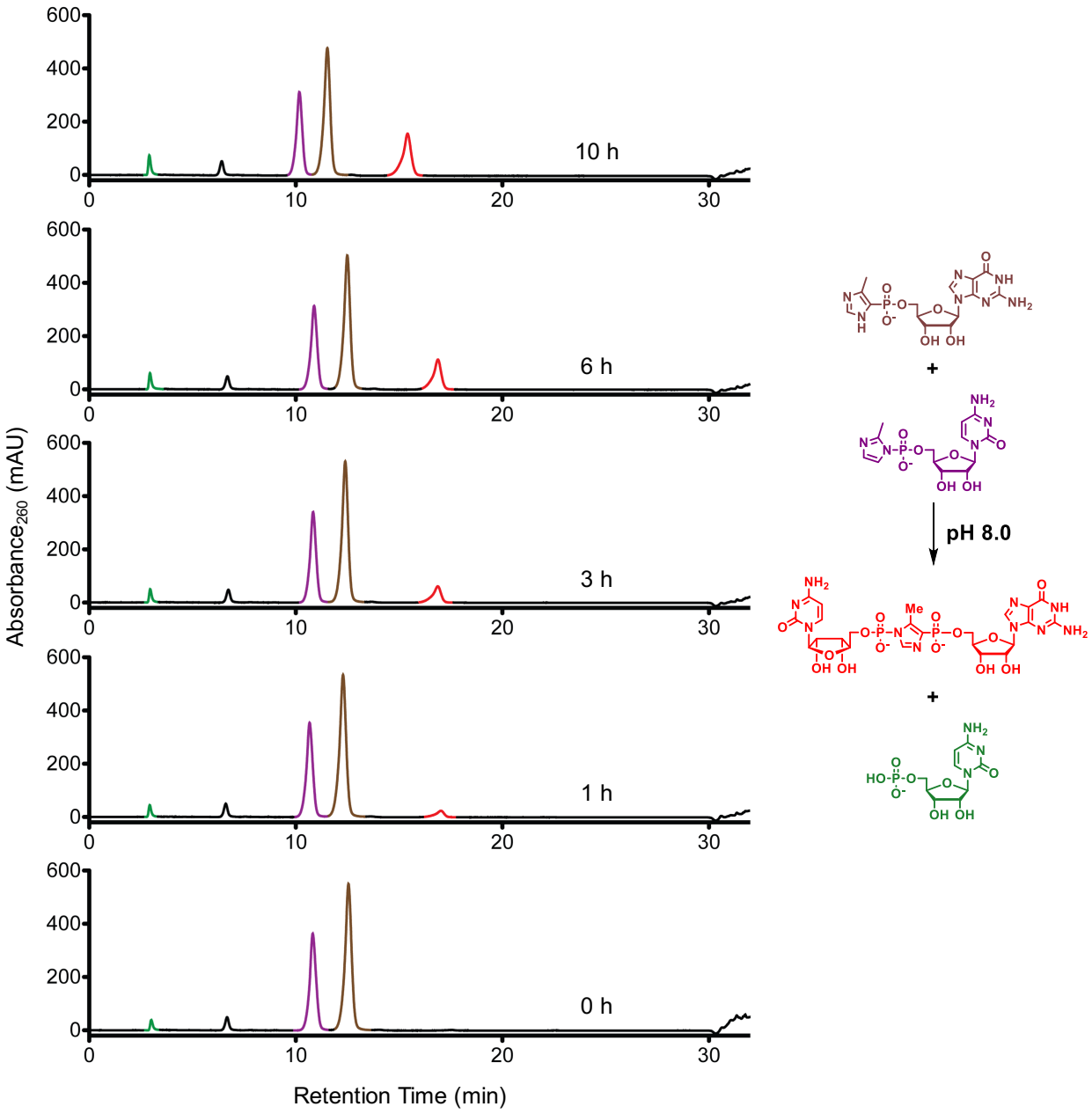


Figure D.38. **2-MeImpC** was coincubated with **ICG** at **pH 8.0**. The formation of the CpICG dinucleotide was assessed with LCMS over a duration of **10 h**. The red, purple, brown and green LC peaks correspond to CpICG dinucleotide, 2-MeImpddC, ICG and CMP respectively. All peak assignments were performed by means of mass spectrometry. At pH 8.0, the CMP peak at the 10 h time point is appreciably smaller than that of pH 6.3, implying that hydrolysis of 2-MeImpC and CpICG is slower at pH 8.0. The decrease in ICG peak area was used to calculate the second order rate constant for CpICG formation, and relevant data are shown in **Figure D.41**.

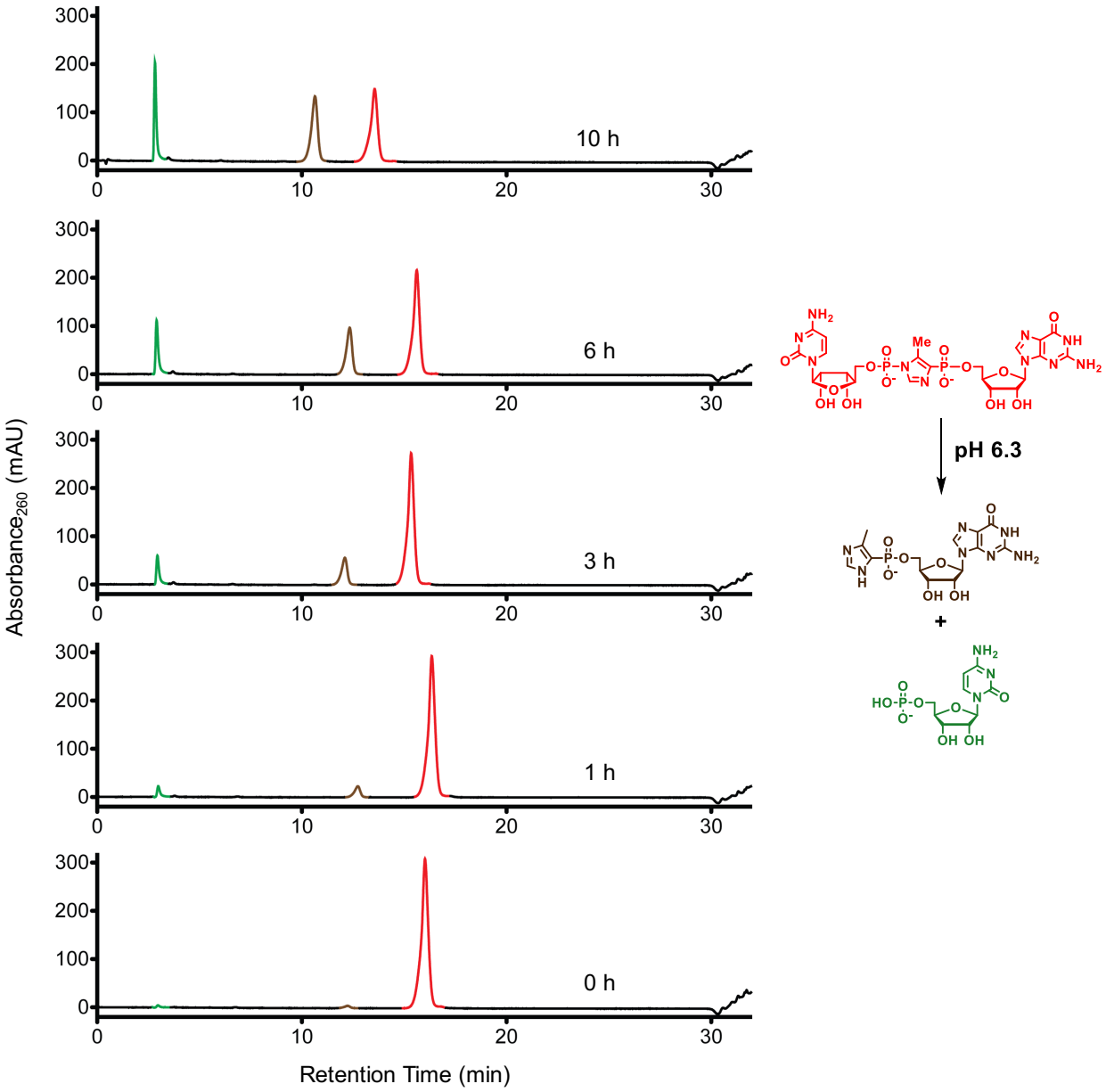


Figure D.39. **CpICG** was incubated at **pH 6.3** under ambient temperature over a duration of **10 h**. The degradation of CpICG was monitored by LCMS. The red, brown and green LC peaks correspond to CpICG, ICG and CMP respectively. All peak assignments were performed by means of mass spectrometry. The decrease in CpICG peak area was used to calculate the pseudo-first order rate constant for CpICG hydrolysis, and relevant data are shown in **Figure D.41**.

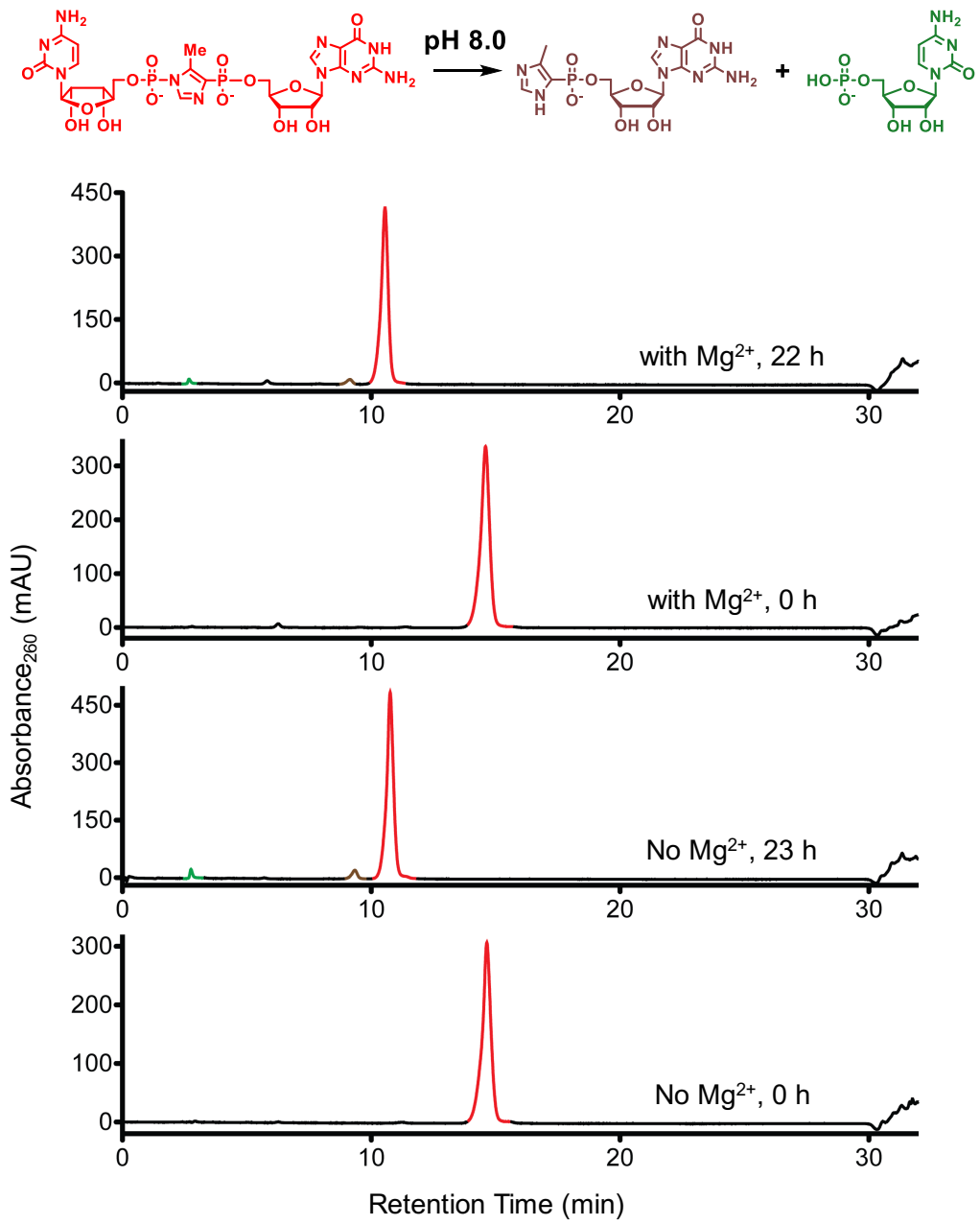


Figure D.40. **CpICG** was incubated at **pH 8.0** under ambient temperature over a duration of **22 h**, both in the absence and presence of 50 mM Mg^{2+}. The degradation of CpICG was monitored by LCMS. All peak assignments were performed by means of mass spectrometry. The red LC peak correspond to CpICG. After 22 h, CpICG degrades only minimally, regardless of whether Mg^{2+} is present (the top and third traces). The brown and green LC peaks correspond to ICG and CMP respectively, and their respective peak areas are less than 5% of that of CpICG. Given the minimal degradation, the pseudo-first order rate of CpICG hydrolysis at pH 8.0 cannot be reliably calculated.

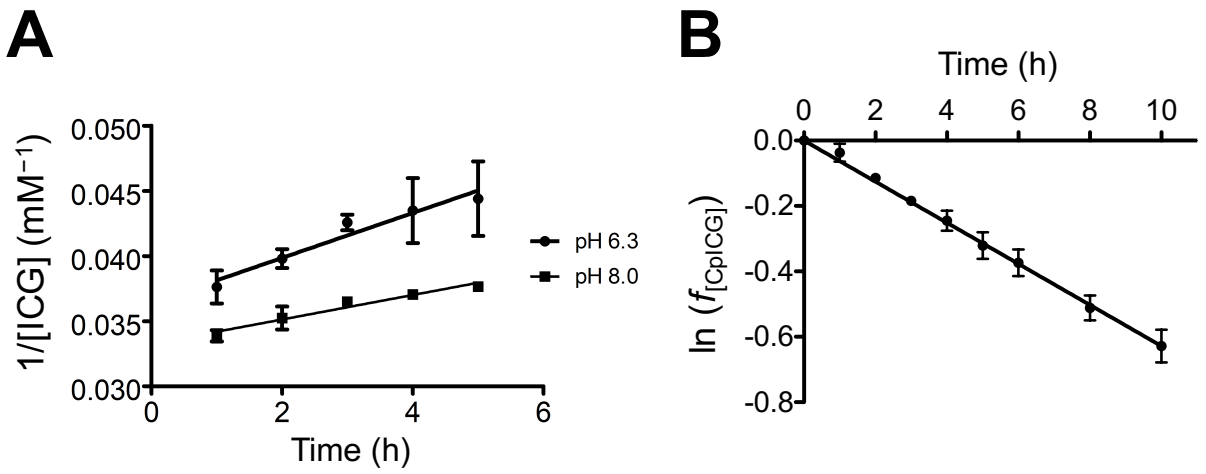


Figure D.41. (a) Second order rate plot showing the rate of CpICG **formation** at **pH 6.3** and **8.0**, using the LCMS data shown in **Figures D.37−D.38**. For all the time points taken (from 0 h to 10 h), their respective ICG LC peaks were integrated to afford total absorbance values. Fraction of remaining ICG at each time point was obtained by dividing the integrated ICG absorbance values by that obtained at the 0 h time point, and was used to compute the reciprocal of ICG concentration at each time point (1/[ICG]). These values were plotted against time and linear regression was then performed. The slope of the fit was divided by 2 to obtain the second order rate k_{obs} of CpICG formation. Experiments were performed in duplicate. The k_{obs} values of CpICG formation at pH 6.3 and pH 8.0 were estimated to be (0.9 ± 0.2) × 10^{-3} h^{-1} mM^{-1} and (0.47 ± 0.07) × 10^{-3} h^{-1} mM^{-1}, respectively. (b) Pseudo-first order rate plot showing the rate of CpICG **degradation** at **pH 6.3**, using the LCMS data shown in **Figure D.39.** The fraction of remaining CpICG ($f_{[CpICG]}$)) at each time point was calculated, and their natural log values (ln ($f_{[CpICG]}$)) were plotted against time. Experiments were performed in duplicate. The estimated k_{obs} was 0.063 ± 0.002 h^{-1}.

D.5.5. Primer extensions with CpICG as activated dinucleotide substrate.

Primer extension reactions were set up, in 10 μL scales, with the following components: 1.2 μM Cyanine 3-labeled primer (Sequence: 5′-Cy3-GACUGACUGG-3′), 2.2 μM complementary template (Sequence: 5′-**AACCCG**CCAGUCAGUC-3′; bolded nucleotides were not duplexed with primer), 200 mM pH 6.3 Na•MES or pH 8.0 Na•HEPES buffer, 100 mM $MgCl_2$, and various concentrations of CpICG dinucleotide (Set 1: 1 mM; Set 2: 3 mM; Set 3: 10 mM; Set 4: 20 mM). With this primer-template duplex, CpICG binds to the primer +1 and +2 positions with the activated N-P bond of the 5′-5′-imidazole bridge pointing towards the primer 3′-OH group. For the negative control experiments (**Figure 4.8**), the template sequence was modified (Sequence: 5′-**AAGC**CCAGUCAGUC-3′) such that CpICG would be binding to the primer +1 and +2 positions with the unreactive C-P bond of the 5′-5′-imidazole bridge pointing towards the primer 3′-OH group. The detailed procedures for primer extension reactions, as well as subsequent reaction quenching, gel electrophoresis, and data analysis procedures were identical to that outlined in **Section D.4.2.** Time points were taken at 0.25, 0.5, 0.75, 1, 2, 3, 4 and 5 h.

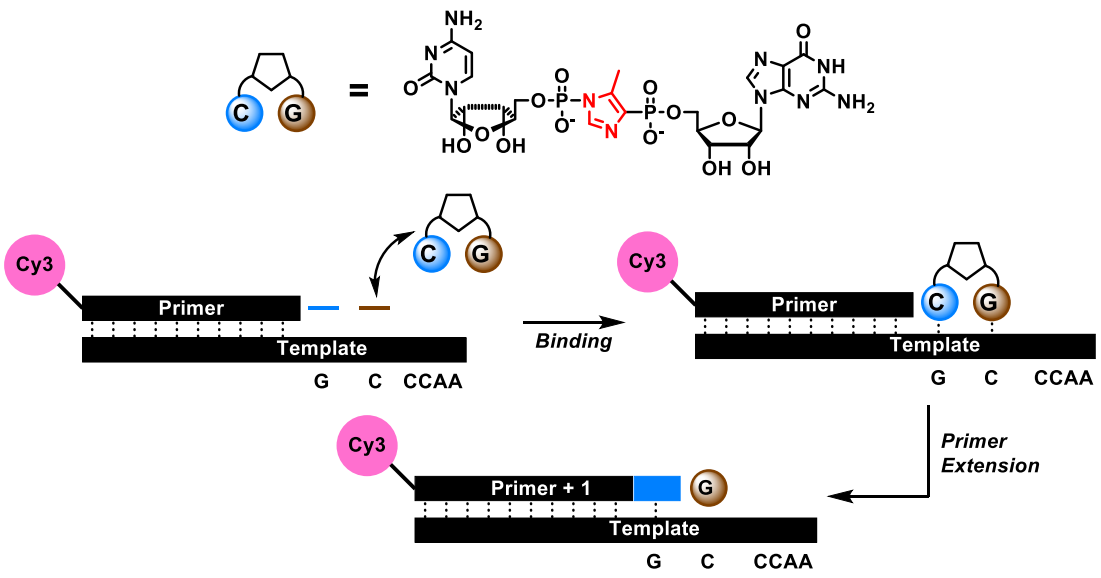


Figure D.42. Schematic diagram of the primer extension reaction. The 5′-Cy3-labeled primer was annealed with the template in the presence of 100 mM $MgCl_2$ and 200 mM buffer of desired pH at room temperature. CpICG was then added to the primer-template complex. The dinucleotide first binds noncovalently to the duplex, followed by an irreversible extension step to afford the primer+1 product.

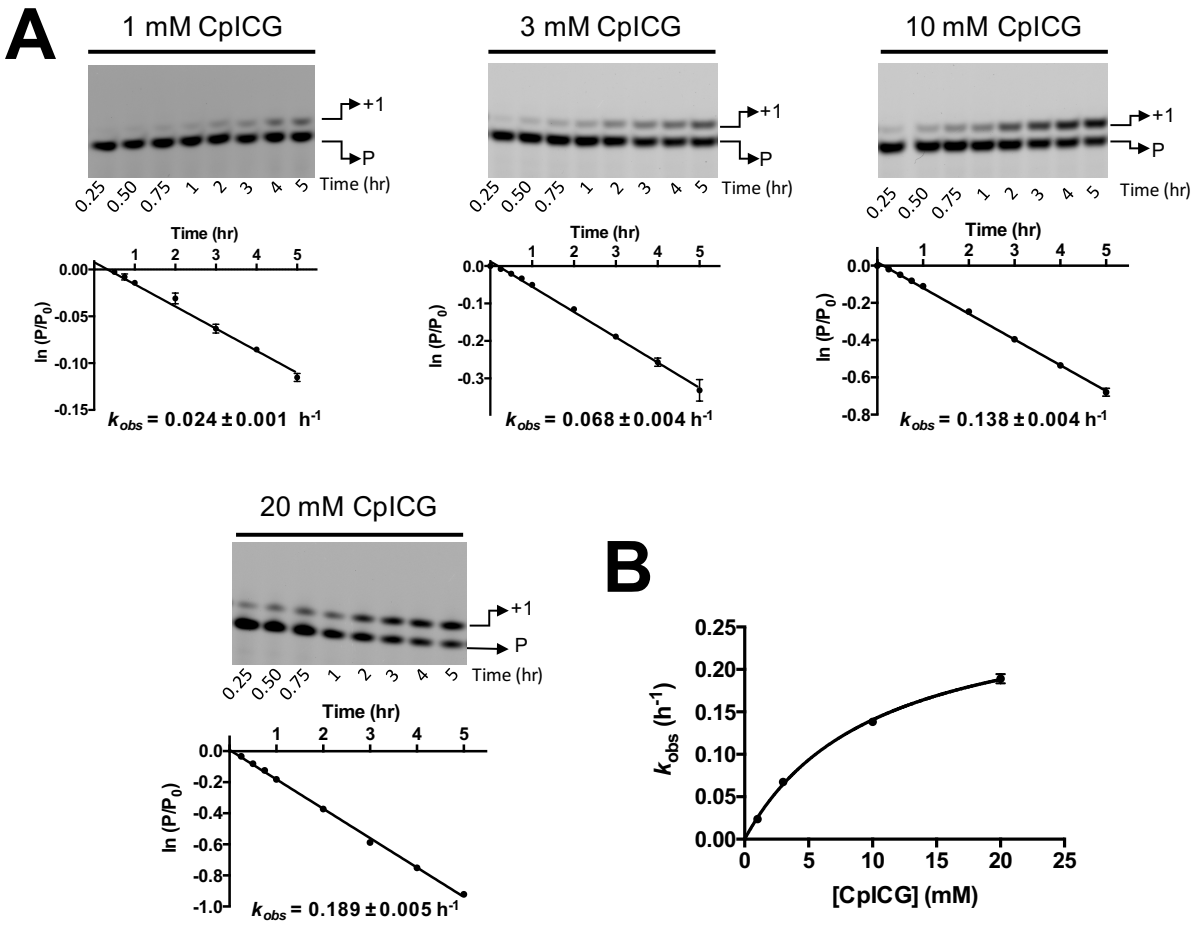


Figure D.43. **CpICG**-driven primer extension reactions at **pH 6.3**. (a) **Top:** Representative 20% denaturing PAGE analyses of the primer extension products. P: Primer; +1: Extended primer+C product. **Bottom:** Plots of the natural log values of the fraction of unreacted primer (ln (P/P_o)) against time. All experiments were performed in triplicate. (b) Saturation curve of the CpICG-driven primer extensions at pH 6.3. The estimated k_{max} and K_M were 0.28 ± 0.01 h^{-1} and 10.6 ± 0.8 mM, respectively (R^2 = 0.996).

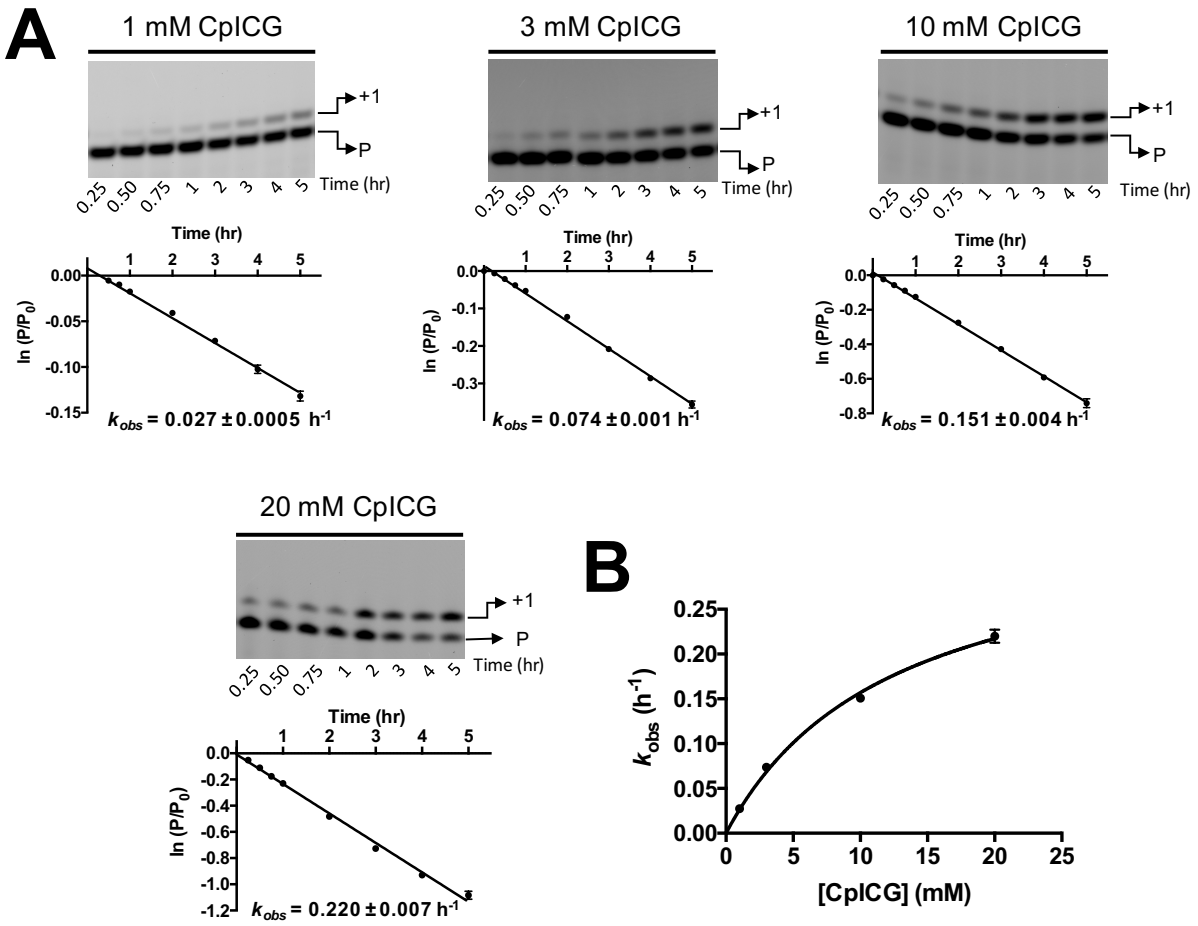


Figure D.44. **CpICG**-driven primer extension reactions at **pH 8.0**. (a) **Top:** Representative 20% denaturing PAGE analyses of the primer extension products. P: Primer; +1: Extended primer+C product. **Bottom:** Plots of the natural log values of the fraction of unreacted primer (ln (P/P_o)) against time. All experiments were performed in triplicate. (b) Saturation curve of the CpICG-driven primer extensions at pH 8.0. The estimated k_{max} and K_M were 0.35 ± 0.02 h^{-1} and 12 ± 1 mM, respectively (R^2 = 0.994). Thus, pH changes do not appear to alter the k_{max} and K_M of CpICG-driven primer extension to any significant extent (**Figure D.43**). See the main text for further discussions on the possible mechanistic origin of this experimental observation.

D.5.6 pK_a determination of CpICG.

A 120 µL solution of CpICG in D_2O, with 7.77 mM CpICG•2TEAH$^+$ and 500 mM NaCl, was prepared. The pD of this solution was measured to be 8.66. The acidity of the analyte solution was adjusted by DCl in D_2O, followed by proton spectrum acquisition with 128 scans, 1 s relaxation delay, and 45° pulse angle at 25 °C. The chemical shift of the imidazole methine proton of the CpICG 5′-5′-imidazole bridge (highlighted in red, **Figure D.45.A**) was traced. Sample acidification and spectral acquisition was repeated until the analyte solution was at pD ≈ 1.5. All proton spectra were referenced to residual protium in D_2O (δ = 4.79 ppm).[14] Experiments were performed in duplicate.

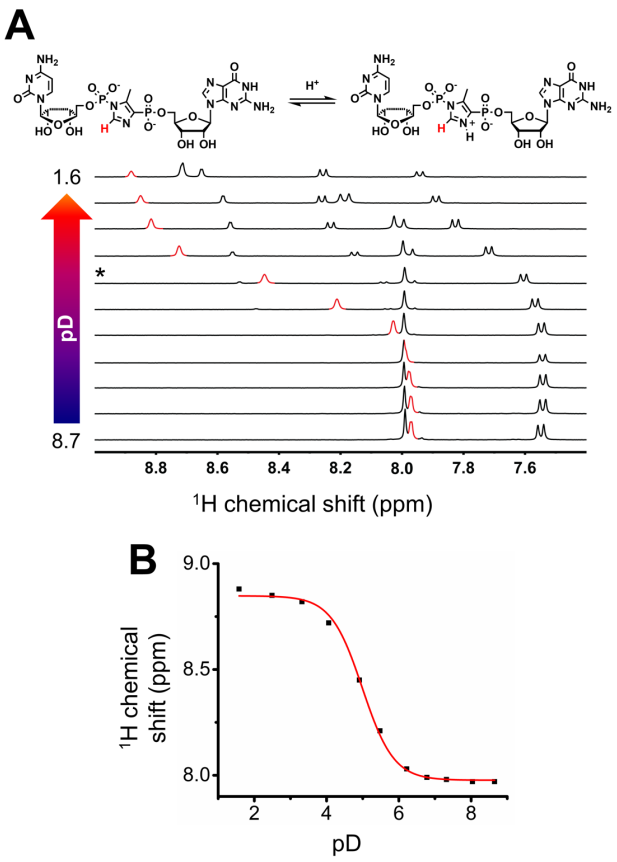


Figure D.45. (a) Representative stacked spectra for the variable pH-proton spectroscopy of CpICG. Only the aromatic region (7.5–9 ppm) was shown. The methine proton of the 5′-5′-bridged imidazole (labeled red) was traced. As CpICG was acidified beyond pD = 5 (asterisked spectrum), CpICG began to hydrolyze to become CMP and ICG, resulting in additional peaks in the aromatic region. The cytidine base of CpICG also become protonated over the course of pD titration, resulting in the observed downfield shift of the cytidine doublet signal at ~7.6 ppm. (b) The chemical shift values of the imidazole bridge methine proton were plotted against their respective solution pD values, followed by data fitting (**Section D.2.3**) and value correction (**Section D.2.1**) to estimate the protonation pK_a of the 5′-5′-imidazole bridge of CpICG. The corrected protonation pK_a of the CpICG imidazole bridge under H_2O environment is ~5.1.

D.7. References

(1) Van Den Berge, E.; Robiette, R. *J. Org. Chem.* **2013**, *78*, 12220–12223.

(2) Kalek, M.; Ziadi, A.; Stawinski, J. *Org. Lett.* **2008**, *10*, 4637–4640.

(3) Matulic-Adamic, J.; Haeberli, P.; Usman, N. *J. Org. Chem.* **1995**, *60*, 2563–2569.

(4) McKenna, C. E.; Higa, M. T.; Cheung, N. H.; McKenna, M.-C. *Tetrahedron Lett.* **1977**, *18*, 155–158.

(5) Błażewska, K. M. *J. Org. Chem.* **2013**, *79*, 408–412.

(6) Taylor, S. D.; Mirzaei, F.; Bearne, S. L. *J. Org. Chem.* **2008**, *73*, 1403–1412.

(7) Li, L.; Prywes, N.; Tam, C. P.; O'Flaherty, D. K.; Lelyveld, V. S.; Izgu, E. C.; Pal, A.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, ASAP.

(8) Zhang, M.; Fang, X.; Neumann, H.; Beller, M. *J. Am. Chem. Soc.* **2013**, *135*, 11384–11388.

(9) Kuehne, M. E.; Lambert, B. F. *Org. Synth.* **1963**, *43*, 22.

(10) Siddiqui, M. A.; Hughes, S. H.; Boyer, P. L.; Mitsuya, H.; Van, Q. N.; George, C.; Sarafinanos, S. G.; Marquez, V. E. *J. Med. Chem.* **2004**, *47*, 5041–5048.

(11) Izgu, E. C.; Oh, S. S.; Szostak, J. W. *Chem. Commun.* **2016**, *52*, 3684–3686.

(12) Bezençon, J.; Wittwer, M. B.; Cutting, B.; Smieško, M.; Wagner, B.; Kansy, M.; Ernst, B. *J. Pharm. Biomed. Anal.* **2014**, *93*, 147–155.

(13) Krezel, A.; Bal, W. *J. Inorg. Biochem.* **2004**, *98*, 161–166.

(14) Gottlieb, H. E.; Kotlyar, V.; Nudelman, A. *J. Org. Chem.* **1997**, *62*, 7512–7515.

(15) Rabenstein, D. L.; Sayer, T. L. *Anal. Chem.* **1976**, *48*, 1141–1146.

(16) Gilow, H. M.; Hong, Y. H.; Millirons, P. L.; Snyder, R. C.; Casteel, W. J. *J. Heterocycl. Chem.* **1986**, *23*, 1475–1480.

(17) Wiseman, T.; Williston, S.; Brandts, J. F.; Lin, L. N. *Anal. Biochem.* **1989**, *179*, 131–137.

(18) Connors, K. A. *Binding Constants : The Measurement of Molecular Complex Stability*; Wiley: New York, 1987.

(19) Schmidtchen, F. P. In *Analytical Methods in Supramolecular Chemistry*; Schalley, C., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Germany, 2007; pp 55–78.

(20) Tam, C. P.; Fahrenbach, A. C.; Björkbom, A.; Prywes, N.; Izgu, E. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2017**, *139*, 571–574.

(21) Thordarson, P. *Chem. Soc. Rev.* **2011**, *40*, 1305–1323.

(22) Joyce, G. F.; Inoue, T.; Orgel, L. E. *J. Mol. Biol.* **1984**, *176*, 279–306.

(23) Rio, D. C.; Ares Jr., M.; Hannon, G. J.; Nilsen, T. W. *Cold Spring Harb. Protoc.* **2010**, *2010*, pdb.prot5444.

Appendix E

Supplemental Materials for Chapter 5

E.1. General Methods: Titration Studies with NMR.

The starting "duplex" solutions all contained 1.5 mM of RNA duplex, 500 mM sodium chloride, and 10% D_2O; the monomer solutions contained the appropriate concentration of guanosine 5′-monophosphate (disodium salt form), as well as the same concentration of the RNA duplex (1.5 mM), D_2O (10%), and sodium cation (500 mM) as that of the starting duplex solution, in order to maintain a constant duplex concentration and ionic strength throughout the titration experiments. Thus, for example, a monomer solution containing 200 mM of GMP·2Na^+ was assumed to contain 400 mM of Na^+, and 100 mM of sodium chloride was added to the monomer solution to bring the concentration of Na^+ of the monomer solution to 500 mM. The pH of both duplex and monomer solutions was adjusted to 7.0 (±0.1) using trace amounts of either aqueous sodium hydroxide or hydrochloric acid, and the NMR spectra were acquired on a Varian INOVA 400 MHz NMR spectrometer equipped with a broadband PFG (z-gradient) probe. Suppression of the bulk water resonance was achieved using a Watergate pulse sequence.[1,2] Each spectrum was recorded after 128–256 scans, with an optimized delay period on pulse sequence (d1) of 1.0 s and a pulse width (pw) of 15 μs. Initial concentrations of the duplex and monomer solutions were determined by UV-Vis measurements, in which the UV absorption of the samples at 260 nm were analyzed on an Amersham Bioscience Ultrospec 3100 UV-Vis spectrophotometer (GE Healthcare, Pittsburgh, PA) in a quartz cell with 1 cm pathlength. The theoretical molar extinction coefficients at 260 nm of all the RNA strands used in this manuscript were calculated with Integrated DNA Technologies' OligoAnalyzer 3.1

(Coralville, IA).[3,4] The chemicals shifts were referenced externally using a co-axial NMR tube containing a solution of pentafluorobenzaldehyde (δ = 10.285 ppm) in $CDCl_3$.

E.2. LC-MS Analysis of Oligonucleotides

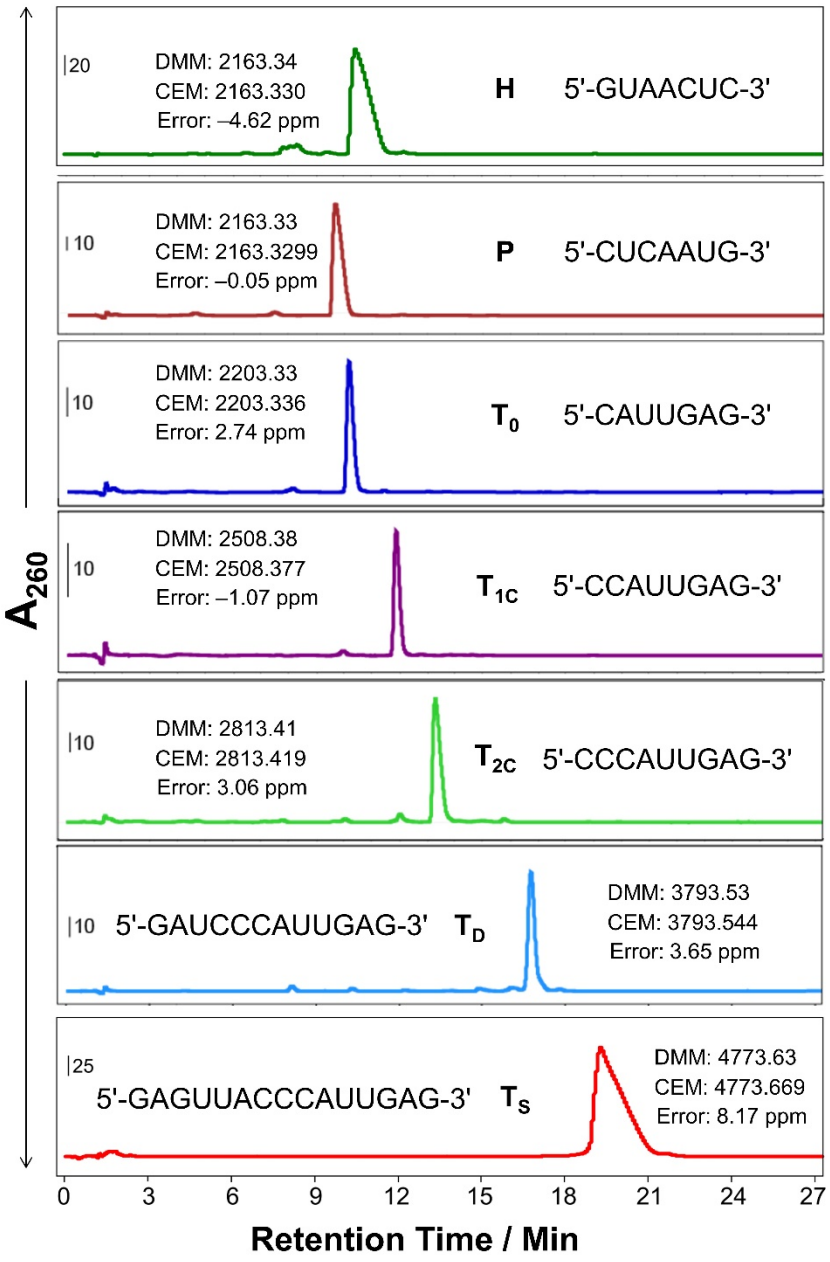


Figure E.1. Absorbance at 260 nm and deconvoluted neutral masses for the synthesized RNAs. Based on the UV traces the purity was determined to be >90% for the synthesized RNAs. DMM: deconvoluted mono-isotopic mass, and CEM: calculated exact mass. Y-axis: Absorbance response units at 260.4 nm. The numbered scale bars denote milli-absorbance units.

E.3. 1- and 2D ^{1}H NMR Spectroscopy

E.3.1. Sequence Design Strategies for Primer/Template/Helper Strands

The primer-template duplexes used in this study were chosen to satisfy a few important criteria. The primer, helper, and template strands are long enough such that they anneal to form stable A-form double helices under the conditions of the titration experiments (500 mM Na^{+}, pH 7, 12 ^{o}C or 20 ^{o}C). Secondly, the primer-template complexes give rise to imino proton resonances that are clearly defined and well-resolved in the 10–15 ppm region of the ^{1}H NMR spectrum, such that any spectral changes upon incremental GMP addition can be clearly traced. After extensive screening, we found that the primer sequence 5'-CUCAAUG-3' satisfies the above criteria, as shown in the stacked proton NMR spectra in sections to follow.

For all primer-template duplexes studied herein, we flanked the GMP binding site with a 3'-G in the primer and a 5'-G in the helper oligonucleotide guanidylate nucleotides such that the consecutive run of Gs (opposite Cs on the template) will mimic model primer-extension conditions in which activated G monomers polymerize on oligo- or poly-C templates. In this well-studied model reaction,[5–7] the noncovalent binding of activated G monomers is presumed to be supported by both upstream and downstream G nucleotides, yet direct thermodynamic evidence to support this hypothesis remains elusive. By measuring the binding affinity of GMP with both upstream and downstream nucleotides being G, we have been able to quantify the magnitude of the contribution of the flanking G residues to GMP binding.

E.3.2. Assignment of the Imino Proton Resonances

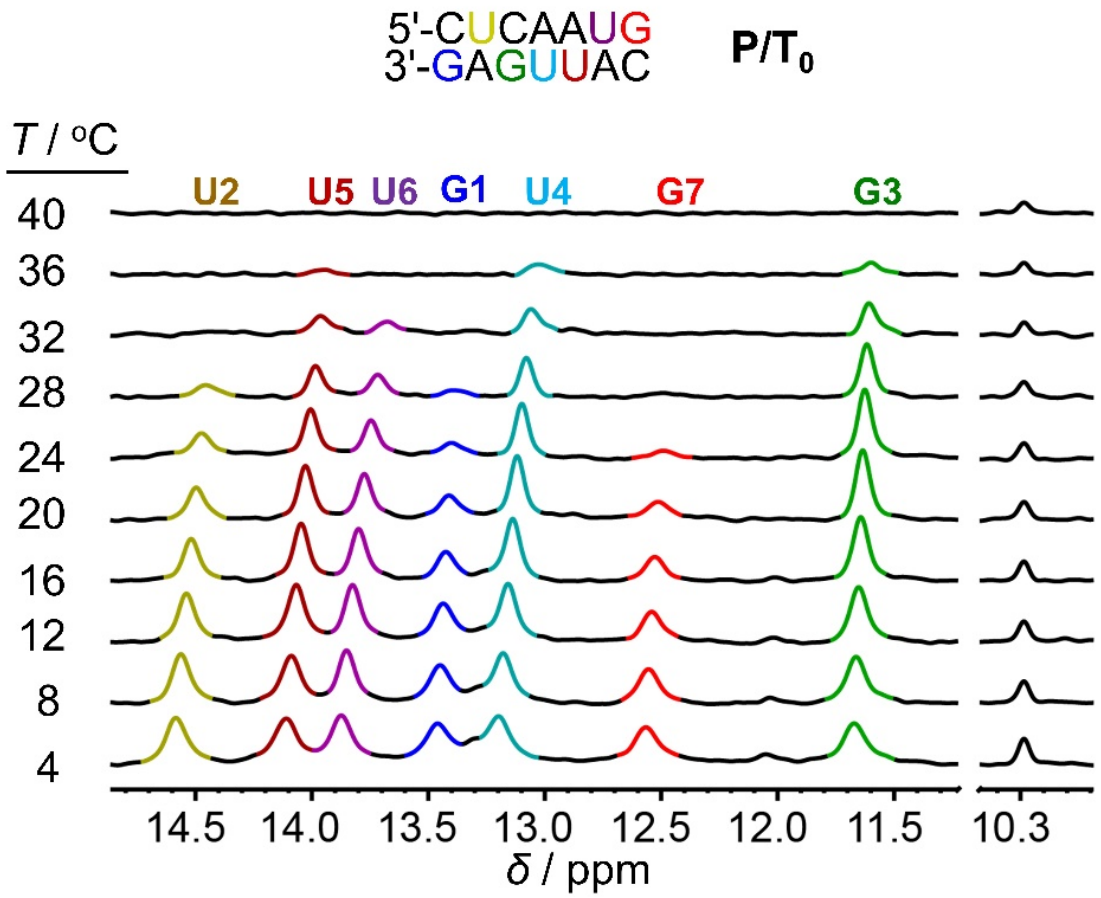


Figure E.2. Variable temperature proton NMR experiments of the **P/T$_0$** duplex. Each spectrum was recorded in the presence of 500 mM NaCl and 10% D_2O at pH 7 with a total strand concentration of 3 mM. The pentafluoro-benzaldehyde reference resonates at 10.285 ppm.

Variable temperature and 2D ^{1}H-^{1}H NOESY experiments were carried out in order to assign the imino proton resonances. The **P/T$_0$** duplex, blunt-ended on both ends of the duplex and no binding sites for GMP, is the focus of this discussion since it contains the same core sequence as all the duplexes studied herein. The variable temperature spectra are shown in **Figure E.2**. At 4 °C, seven resonances are observed in the region of 11.5 to 15 ppm that arise from the imino protons of the Watson-Crick pairs of the duplex.[8,9] The exchange rate of the imino protons is dependent on temperature, and the imino

protons of base pairs closer to the termini of the duplex tend to undergo faster exchange compared to those closer to the center of the duplex. As a consequence, the imino protons at the edges of the duplex are expected to broaden out into the baseline at lower temperatures relative to those in the center.[10] As the temperature was raised past 4 ^{o}C, the first signals observed to undergo broadening and decreases in peak height are G1 and G7, since these resonances arise from the imino protons of the terminal base pairs. G1 can be differentiated from G7 from the titration data, since monomer binding to a 5´-overhang slows down the exchange rate of the imino proton on the adjacent base pair.[11] The remaining G resonance, G3, can be assigned to the most upfield imino signal observed, because G imino resonances tend to be observed in more upfield region than U imino resonances.[11] The next two resonances that begin to broaden out and decrease in height as the temperature is increased are U2 and U6, followed by U5, and then the resonances arising from the most internal imino protons, G3 and U4. Differentiating between the resonances for U2 and U6 was accomplished using 2D ^{1}H-^{1}H NOESY. By 40 ^{o}C, all resonances have broadened into the baseline.

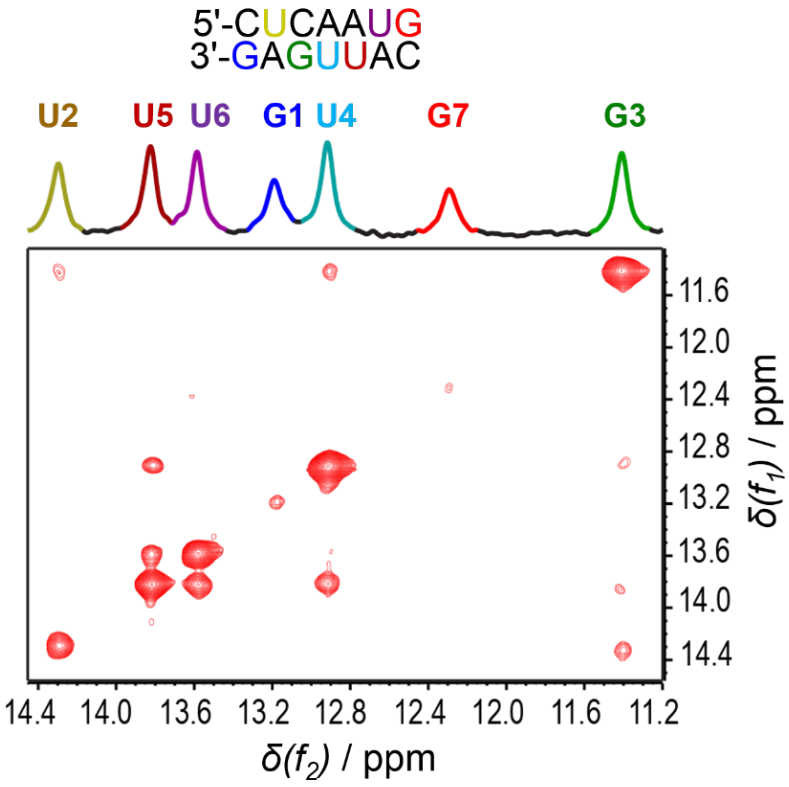


Figure E.3. ^{1}H-^{1}H NOESY experiments of the **P/T$_0$** duplex. This spectrum was recorded at 4 °C with 200 ms mixing time, in the presence of 500 mM NaCl and 10% D_2O at pH 7 with a total strand concentration of 3 mM. The cross-peaks (off-diagonal red signals on the spectrum) represent interactions between neighboring imino protons.

The ^{1}H-^{1}H NOESY spectrum of the **P/T$_0$** duplex was carried out at 4 °C with a 200 ms mixing time using a WATERGATE pulse sequence[1,2] in order to suppress the bulk water signal at δ ≈ 5.01 ppm,[12] and is shown in **Figure E.3**. The U resonance which shares no crosspeaks with any other U resonance is confirmed to be U2; it also possesses crosspeaks with G3. No crosspeaks are observed with G1 likely because this proton being on a terminal base pair is exchanging too rapidly with H_2O. The U5 assignment is corroborated by the two sets of crosspeaks it shares with two other U resonances. The resonance at ~13.6 ppm is assigned to U6, which shares a set of crosspeaks with U5. No crosspeaks between U6 and G7 are observed due to the G7 imino proton being in fast exchange with H_2O. U4 shares a set of crosspeaks with G3 and U5. ^{1}H-^{1}H NOESY experiments were also carried out on the **P/T$_{1C}$** and **P/T$_{2C}$** duplexes (**Figure E.4**). In these cases, faint NOESY crosspeaks are also observed between U2 and G1.

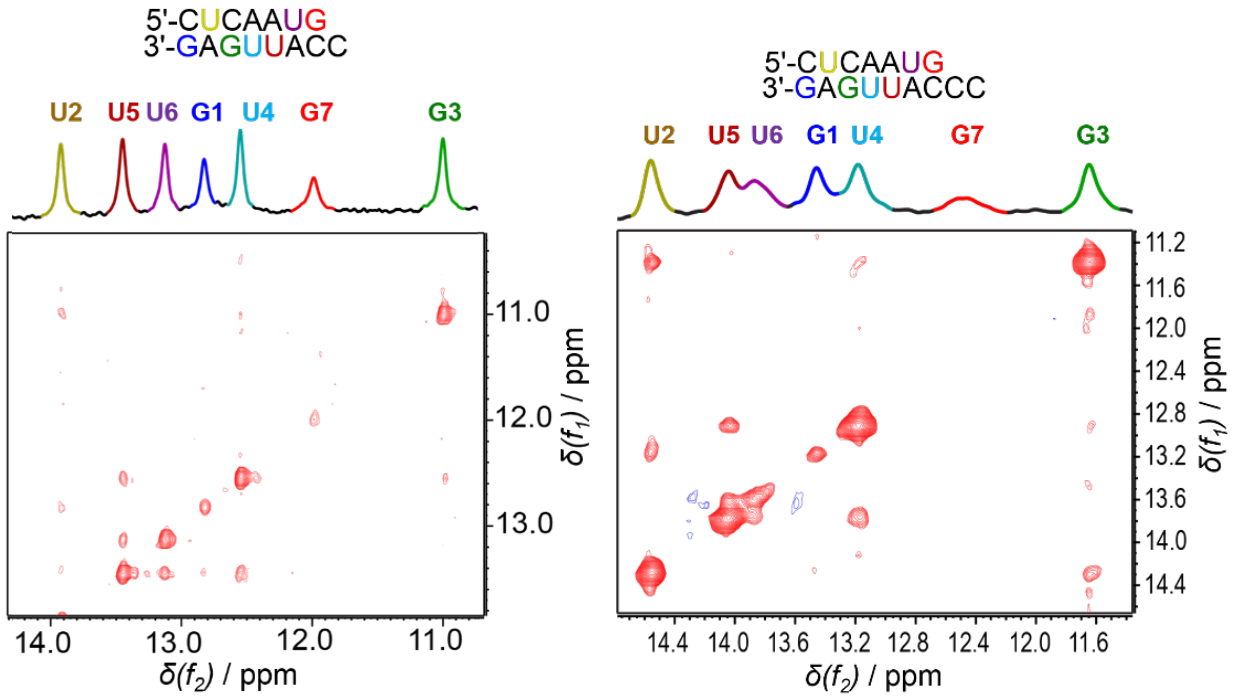


Figure E.4. ^{1}H-^{1}H NOESY experiments for **P/T$_{1C}$** and **P/T$_{2C}$** duplexes were performed in order to assign the imino proton resonances. Left: ^{1}H-^{1}H NOESY experiment of the **P/T$_{1C}$** duplex. This spectrum was recorded at 4 °C with 200 ms mixing time, in the presence of 500 mM NaCl and 10% D_2O at pH 7 with a total strand concentration of 3 mM. Right: ^{1}H -^{1}H NOESY 2D spectrum of the **P/T$_{2C}$** duplex recorded at 8 °C, 200 ms mixing time, 10% D_2O, pH 7 and a total strand concentration of 6 mM. The cross-peaks (off-diagonal red signals on the spectrum) represent interactions between neighboring imino protons.

E.3.3. Titration of GMP into P/T$_0$

In order to evaluate the ability of GMP to interact with a duplex containing no 5´-overhang binding site, we carried out a titration of GMP from 0 to 250 mM into a solution of the **P/T$_0$** duplex. In a previous study,[11] we found that a titration of CMP into a solution containing a similarly blunt-ended palindromic duplex under identical conditions led to only very minor changes in the chemical shifts of all the imino resonances of the duplex, and the resonances arising from the imino protons of the terminal base pairs underwent significant line broadening. In the case of the GMP titration here, however, different behavior is

observed (**Figure E.5.a.)**. The most notable changes that occur as the concentration of GMP is increased take place with the G1 and G7 resonances arising from the imino protons on the terminal base pairs. As was the case in the CMP titration reported previously,[11] these signals begin to broaden out significantly as the concentration of the monomer is increased. At 250 mM of GMP, the G1 and G7 signals are a fraction of their original heights. Unlike the previous CMP titration, however, these signals both undergo a significant upfield shift.

The plot of the change in chemical shift of all the imino resonances is shown in **Figure E.5.b**. The degree of change in chemical shift is related to the location of the respective imino protons in the duplex. G1 and G7 shift the most followed by U2 and U6. The most internal protons, which give rise to resonances G3, U4 and U5, undergo almost no changes in chemical shift. These observations are consistent with a mechanism in which the GMP associates with the termini of the duplex. We hypothesize that GMP stacks in a side-on fashion to the base pairs on the duplex termini, leading to the change in chemical shifts observed, while the excess GMP catalyzes the exchange of these imino protons, causing them to broaden out. Another possibility is that GMP is able to displace the 3´-G on either side of the duplex, forming a Watson-Crick pair with the 5´-C nucleobases on both of the $\mathbf{T_0}$ or **P** strands. In either case, the fact that the most internal signals G3, U4 and U5 are so little affected indicates that changes in chemical shift caused by this binding event on the termini do not propagate past two base pairs. In other words, the changes in chemical shift of these terminal protons represent two independent binding events. We assessed the intensity of such nonspecific side-on stacking of GMP to the two termini of $\mathbf{P/T_0}$ by fitting the change in chemical shifts of G1 and G7 imino

protons of **P/T$_0$** to a standard single-site binding model, assuming that the ends of the duplex are not binding with a higher-order aggregate of GMP. The nonspecific GMP stacking to the G1 and G7 termini is weak indeed, with fitted binding constants of 7(1) and 13(1) M^{-1}, respectively (**Figure E.5c**).

For the cases of the **P/T$_{1C}$** and **P/T$_{2C}$** titrations with GMP, in order to quantify the binding of GMP to the 5′-overhang, the changes in chemical shift of G1 and U2 must be excluded from any best-fit calculations, because these shifts arise from an unrelated binding event.

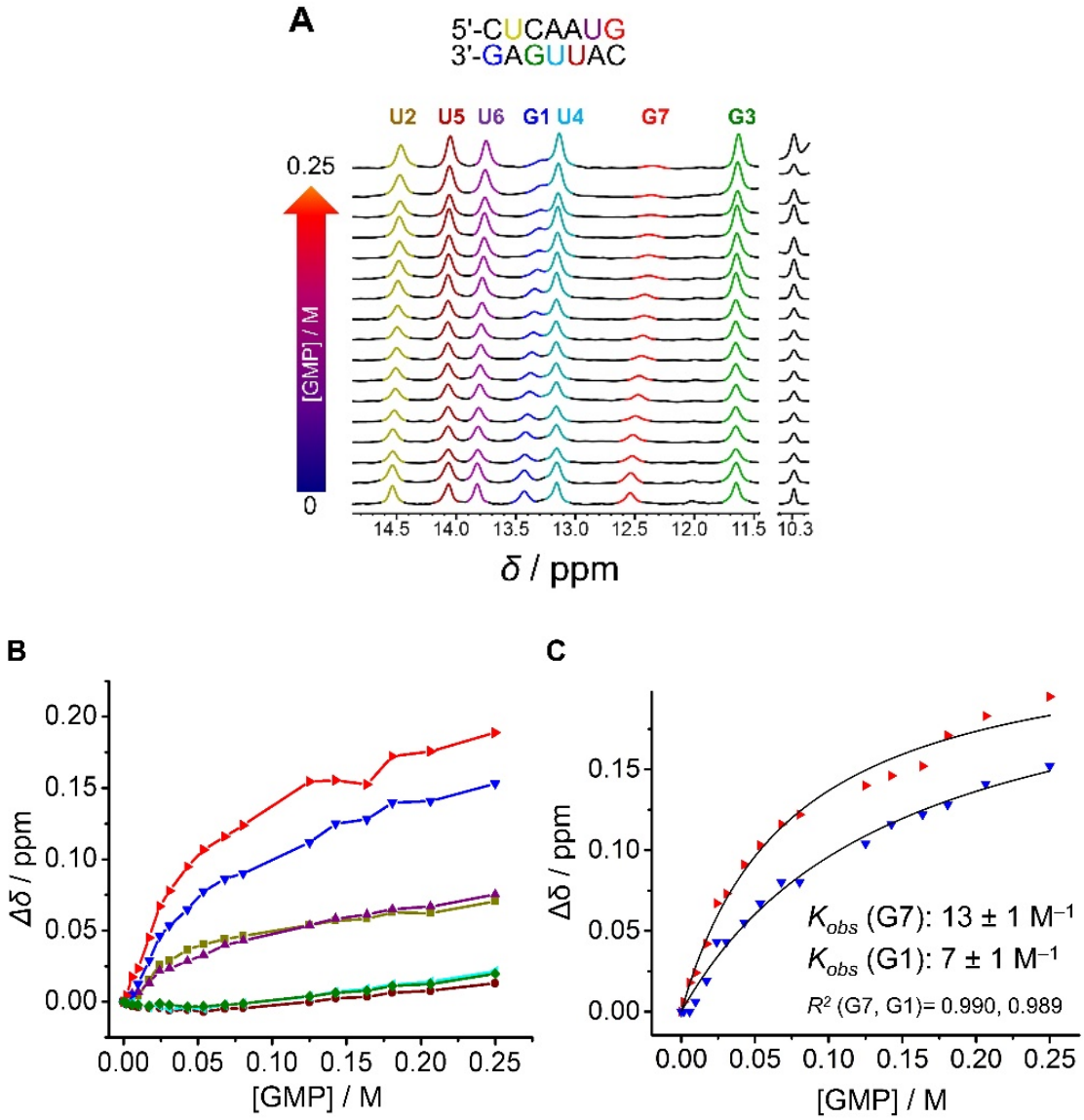


Figure E.5. Titration of the **P/T$_0$** duplex with GMP, monitored by ^{1}H NMR. **a.** The 10 to 15 ppm region of the ^{1}H NMR spectra, where all the imino protons of the duplex resonate, were monitored throughout the titration, and stacked for easy comparison. Each spectrum was recorded at 12 °C in the presence of 500 mM Na^+ and 10% D_2O at pH 7 with a total strand concentration of 3 mM. The signal resonating around 10.3 ppm is that of the pentafluorobenzaldehyde reference. **b.** Plot of the change in chemical shift (in ppm) of the imino proton resonances shown in **a**, with respect to GMP concentration. The two terminal guanosine imino protons (G1 and G7, shown in blue and red data points, respectively) shifted the most, followed by the other imino protons. ►:G7; ▼:G1; ▲:U6; ■: U2; ●:U5; ◄: U4; ♦: G3. **c.** Nonspecific stacking of GMP onto both G1 and G7 termini of **P/T$_0$** was modeled by the single-site binding isotherm, with fitted binding constants of 7(1) and 13(1) M^{-1}, respectively.

E.3.4. Titration of GMP into P/T$_{1C}$

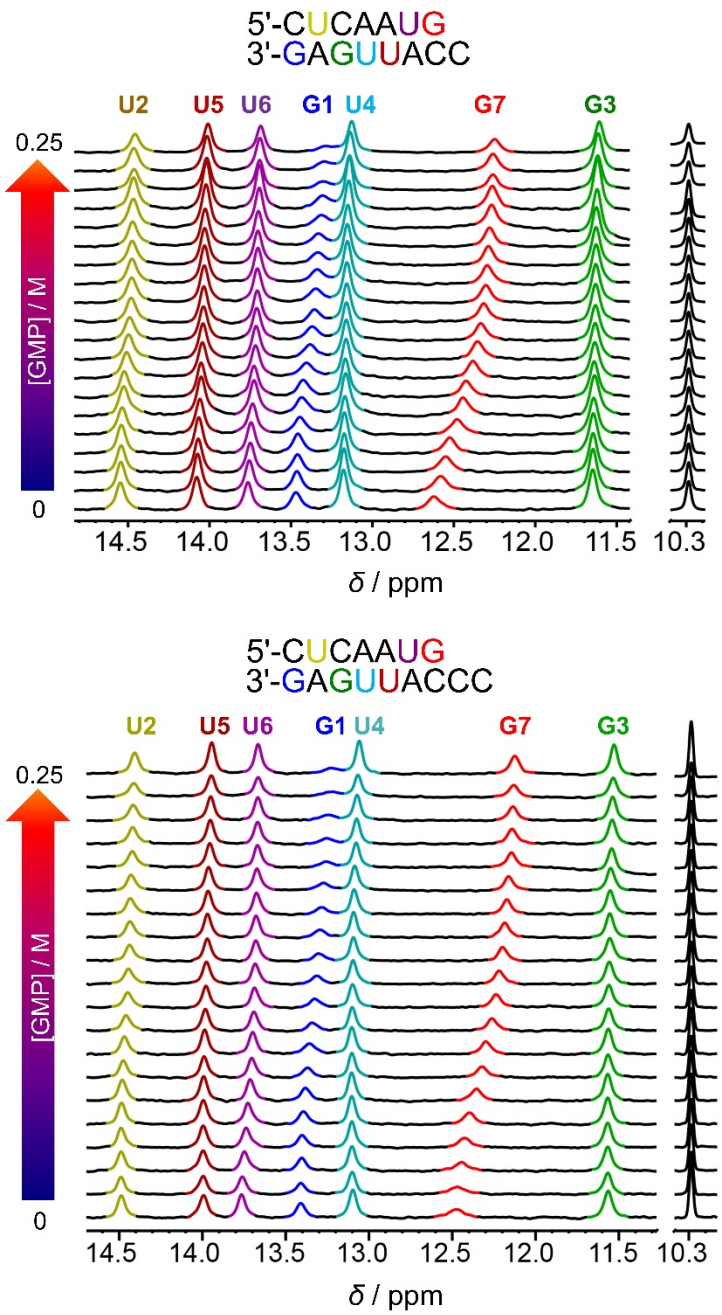


Figure E.6. Titration of the **P/T$_{1C}$** and **P/T$_{2C}$** duplexes with GMP while monitoring the 10 to 15 ppm region of the ^{1}H NMR spectra. The component spectra for both GMP–**P/T$_{1C}$** (top) and GMP–**P/T$_{2C}$** (bottom) titrations were stacked for convenient visualization. The 5´-C overhang of **P/T$_{1C}$**, and the 5´-CC overhang of **P/T$_{2C}$**, engages in binding with GMP. All spectra were recorded at 12 °C in the presence of 500 mM Na^{+} and 10% D_2O at pH 7 with a total strand concentration of 3 mM. The signal resonating at 10.285 ppm is that of the pentafluorobenzaldehyde reference.

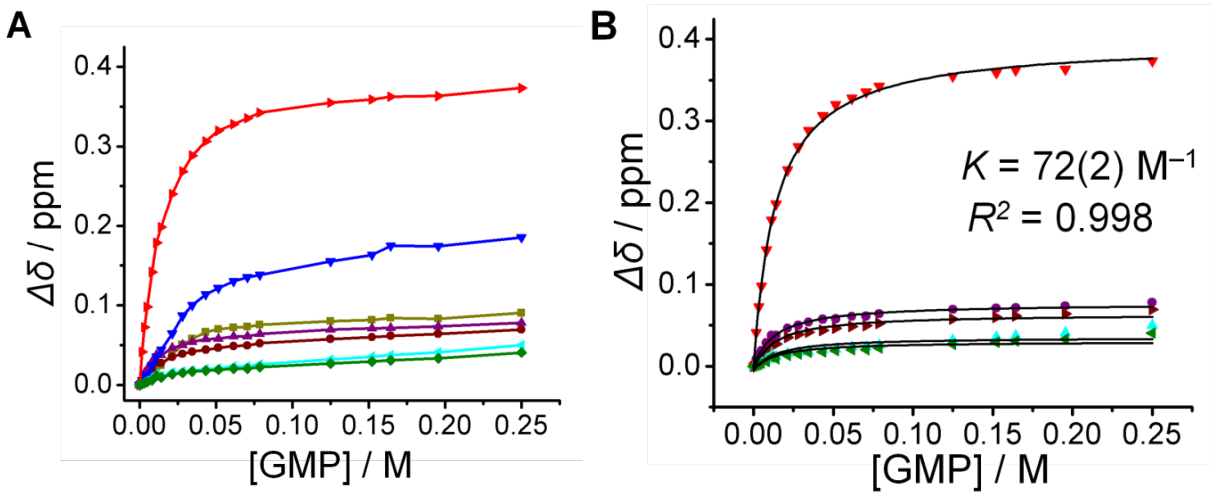


Figure E.7. Using the spectral data shown in **Figure E.6** (top), the changes in chemical shifts of all imino proton resonances of the **P/T$_{1C}$** duplex were plotted against GMP concentration. **a.** The imino proton signal of G7 (shown in red data points) shifted the most, followed by that of G1 (blue data points). The other imino protons (U2, G3, U4, U5, U6) shifted to a lesser degree (yellow, green, cyan, crimson, and purple, respectively). **b.** The data obtained from **S7a** was fitted against the single-binding-site model with the G7, U6, U5, U4 and G3 resonances, which afforded the binding constant of GMP relative to the RNA duplex. If only G7 was used in the fit, K was found to be 75(2) M^{-1} with an R^2 value of 0.999.

In order to evaluate the strength of binding of GMP to a duplex with only a single 5′-C overhang, we carried out a titration from 0 to 250 mM GMP into the **P/T$_{1C}$** duplex solution (**Figure E.6**., top). For blunt-ended duplexes like **P/T$_{1C}$**, the NMR titration data is fitted to the well-studied single-site binding isotherm,[13] which offers an unambiguous description of the binding behavior of GMP with the cognate template binding site. As the concentration is increased, the most notable changes occur to the G7 resonance. Like the **P/T$_0$** titration, the G1 signal significantly broadens while shifting upfield. A plot of the change in chemical shifts of all the proton resonances versus the concentration of GMP is shown in **Figure E.7a.** All resonances are observed to undergo some upfield shift. It is likely that most of the chemical shift for G1 and U2 arises from side-on association of

GMP to this terminus, a binding event which is independent of the binding of GMP to the 5´-overhang. The relatively small changes in chemical shift of the internal proton resonances U6, U5, U4 and G3 likely are a result of small changes in conformation that occur to the duplex upon binding of GMP to the 5´-C overhang. Only the resonances of G7, U6, U5, U4 and G3 were used when fitting the chemical shift data to a standard single-site binding isotherm (**Figure E.7b**). When only the G7 resonance was used, the fitted value of *K* differed only slightly.

We found that the K_a of GMP binding to the overhang position of **P/T$_{1C}$** (72 M^{-1}) is roughly 40% higher than the value we have previously reported for GMP binding with a self-complementary palindromic duplex, with rotationally-symmetric binding sites on both of the duplex ends (52 M^{-1}).[11] However, following translation of the binding constants to the corresponding free energy of GMP-duplex binding (ΔG), we found that the difference in ΔG between the two scenarios is small, with ΔG for GMP–palindromic duplex and GMP–**P/T$_{1C}$** binding calculated to be –2.24 and –2.42 kcal mol^{-1} respectively. The difference of ~0.18 kcal mol^{-1} in ΔG (or a calculated 7.6% percentage difference) is actually much less than that of a single hydrogen bond. We are reluctant to speculate on the experimental factors or nonidealities that give rise to these small differences, but given the fact that **P/T$_{1C}$** is a blunt-ended double helix while the duplexes we reported in 2015 are palindromic, subtle and under-appreciated structural differences in the binding pockets between the **P/T$_{1C}$** and the palindromic duplexes likely exist. Such small dissimilarities could easily account for the observed small difference in binding free energies, by some small, and hard to predict, entropic or enthalpic effects.

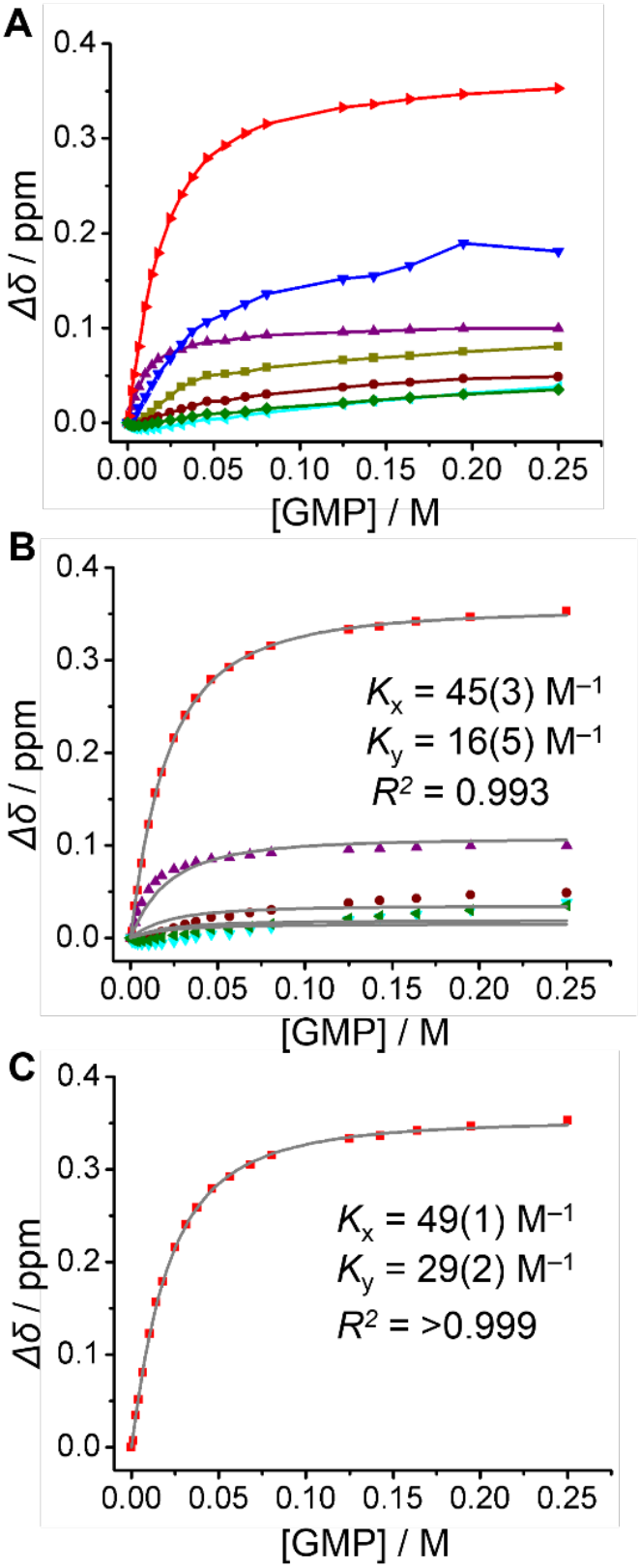


Figure E.8. Using the spectral data shown in **Figure E.6** (bottom), the changes in chemical shifts of all imino proton resonances of the **P/T$_{2C}$** duplex were plotted against GMP concentration. **a.** The extent of cumulative imino proton shift follows the following order: G7 > G1> U6 > U2 > U5 > U4 ≈ G3. The plot was fitted against the two separate two-site binding isotherms derived in **Section S5**. **b.** The imino proton resonances of G7, U6, U5, U4 and G3 globally fitted against equation 6a. **c.** The imino proton resonances of G7 fitted against equation 6b. For more discussion, see the main text and **Section S5**.

E.3.5. Titration of GMP into P/T_{2C}

A titration from 0 to 250 mM GMP into a solution of the **P/T_{2C}** duplex was carried out in order to evaluate the strength of binding of two GMPs to an RNA duplex with a 5´-CC overhang. **Figure E.6 (bottom)** shows the stacked proton NMR spectra for this titration, monitoring the 10 to 15 ppm region of the ^{1}H NMR spectra. All imino proton signals are well resolved, and a plot of the change in chemical shift of all seven imino protons is shown in **Figure E.8a**. The changes in chemical shifts of the G7, U6, U5, U4 and G3 imino resonances were plotted and globally fitted against the sequential two-site binding isotherm (equation 6a derived in **Section E.5** of supplemental information, which assumes $\Delta\delta_{\mathbf{P/TC2}:2GMP} = \Delta\delta_{\mathbf{P/TC2}:GMP}$ for all resonances). The two binding constants, K_x and K_y reflect the affinity of GMP for the first and second binding sites, respectively, and were found to be 45 M^{-1} and 16 M^{-1} with an R^2 factor of 0.993 (**Figure E.8b**). Note that, since the chemical shift for G1 and U2 arises from association of GMP to this terminus, a binding event independent of the binding of GMP to the 5´-overhang, the changes in chemical shifts of G1 and U2 resonances are excluded from the best fit calculations. When the NMR shift data for G7 imino proton alone is fitted against another equation 6b in **Section E.5**, which assumes $\Delta\delta_{\mathbf{P/TC2}:2GMP} = 1.25\Delta\delta_{\mathbf{P/TC2}:GMP}$ for this resonance, the two binding constants were found to be 49 M^{-1} and 29 M^{-1}, and the R^2 factor for this fit is greater than 0.999 (**Figure E.8c**).

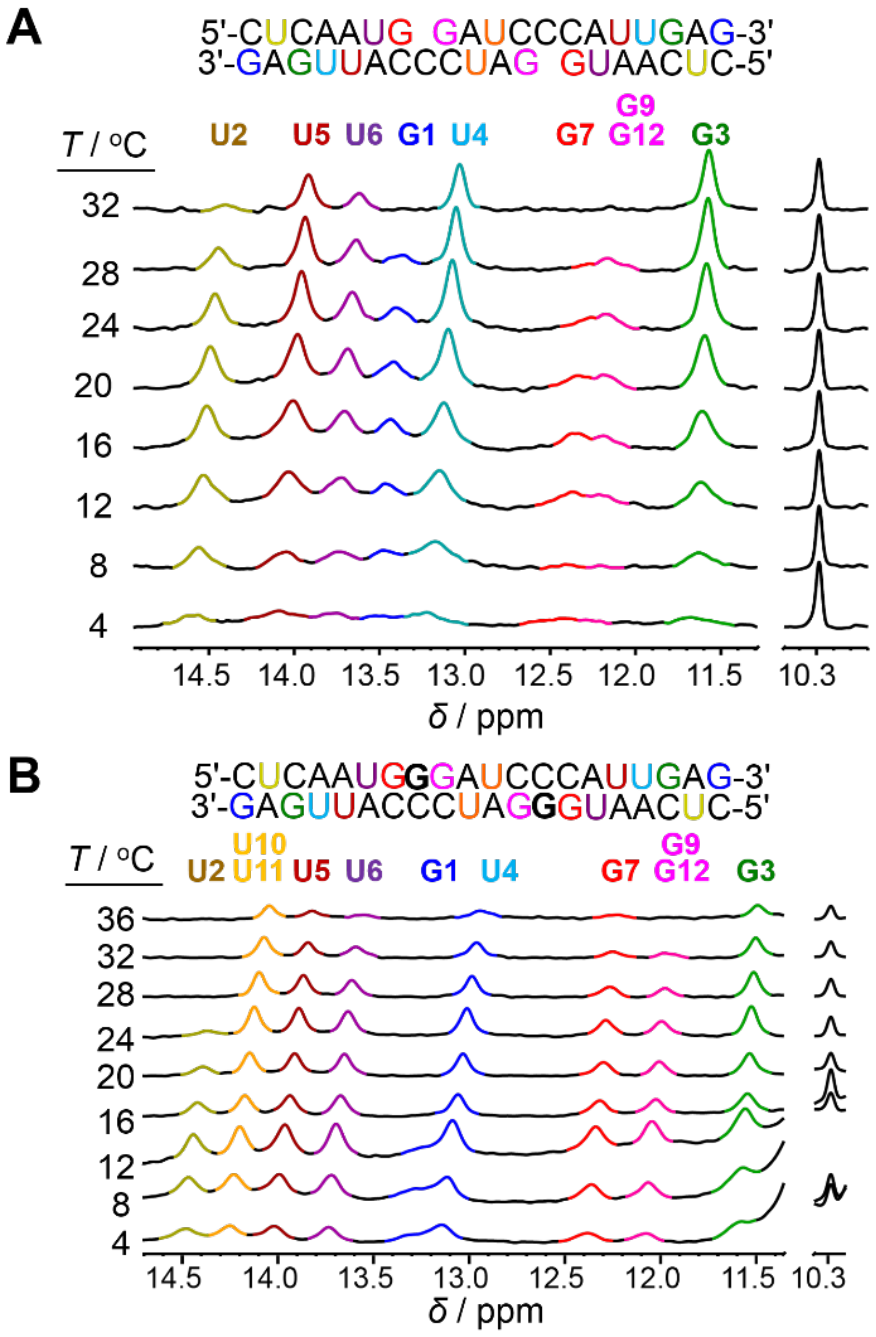


Figure E.9. Comparison of the variable-temperature ^{1}H NMR analysis of the $(\mathbf{P/T_D})_2$ duplex, in the absence or presence of GMP monomer. **a.** Variable-temperature ^{1}H NMR analysis of the $(\mathbf{P/T_D})_2$ duplex *without* GMP monomer, performed from 4–32 °C. **b.** Variable-temperature (4–36 °C) ^{1}H NMR of the $(\mathbf{P/T_D})_2$ duplex solution *with* 250 mM of GMP. All duplex NMR spectra were recorded in the presence of 500 mM NaCl, 10% D_2O, at pH 7, with a total strand concentration of 3 mM.

E.3.6. Variable Temperature Analysis of $(P/T_D)_2$

Variable temperature ^{1}H NMR spectroscopy of a solution of $(\mathbf{P/T_D})_2$ before and after addition of GMP provides some insight into the structural dynamics of the duplex before and after monomer binding. The variable temperature ^{1}H NMR spectra of the $(\mathbf{P/T_D})_2$ duplex solution from 4 to 32 ^{o}C are shown in **Figure E.9a**. At 4 ^{o}C, the imino proton signals are broader than those at 12 ^{o}C. We hypothesize that this broadening is a consequence of the fact that a majority of the $\mathbf{P/T_D}$ duplex has dimerized, forming a relatively large complex that undergoes slow tumbling on the NMR timescale. The variable temperature ^{1}H NMR data obtained on the $\mathbf{P/T_0}$ duplex supports this hypothesis in that the imino proton peaks of this truncated duplex are comparatively sharper at 4 ^{o}C. Raising the temperature of the $(\mathbf{P/T_D})_2$ duplex solution results in sharpening of the resonances, followed by broadening and decreases in peak intensities of the G1, G7 and G12 signals. Above 32 ^{o}C, these peaks, having decreased in height while broadening into the baseline, are no longer observable. The situation differs when GMP is present. The variable temperature spectra of the $(\mathbf{P/T_D})_2$ duplex solution after adding 250 mM of GMP are shown in **Figure E.9b**. Even at 32 ^{o}C, the G7 and G9/G12 resonances remain observable, as well as the U10/U11 resonance. These observations are consistent with the specific binding of GMP into the binding pockets formed by the dimerized duplex, such that the stably bound GMPs significantly slow down the rate of exchange of the imino protons closest to the binding site.

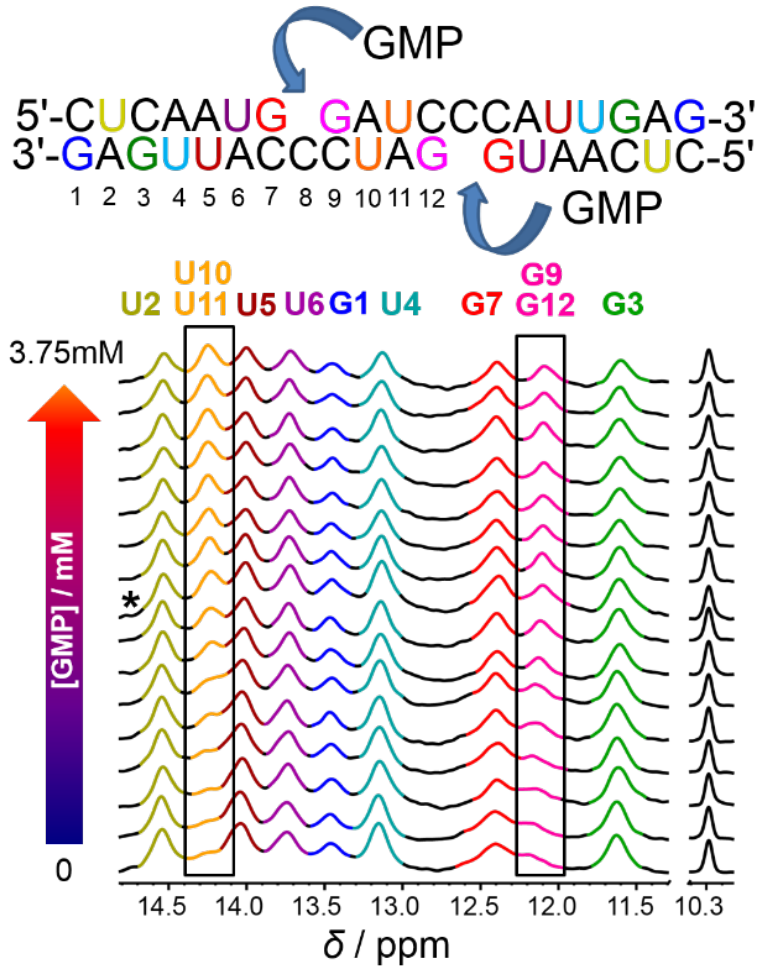


Figure E.10. ^{1}H NMR titration of GMP into a solution containing 0.75 mM of the $(\mathbf{P/T_D})_2$ duplex, recorded at 12 °C, pH 7 and 500 mM Na^+. The titration was carried out from 0 to 3.75 mM of GMP. The resonances assigned to the U10/U11 and G9/G12 imino protons (boxed), become more intense as the concentration of GMP is increased. The asterisked spectrum marks a total GMP concentration of 1.43 mM (~2 equivalents). By using the fact that ~75% of the binding sites are occupied after the addition of approximately two equivalents of GMP relative to $(\mathbf{P/T_D})_2$, a single point calculation of the binding constant can be carried out, and reveals a K_{OBS} of 8000 M^{-1}.

E.3.7. Titration of GMP into $(\mathbf{P/T_D})_2$

We designed the palindromic $(\mathbf{P/T_D})_2$ duplex, in which the $\mathbf{T_D}$ strand possesses a 5′-overhang that consists of a cytidine binding site for GMP, followed by a self-complementary CUAG region (**Figure E.10**). The $(\mathbf{P/T_D})_2$ duplex contains two identical binding pockets that are supported by both upstream and downstream oligonucleotides. This duplex allows us to determine the degree to which a stably bound downstream element enhances the association of GMP. The ^{1}H NMR titration of GMP with $(\mathbf{P/T_D})_2$ is shown in **Figure E.10**. Prior to the addition of GMP (bottom trace), seven signals arising from the imino protons of the core sequence (U1 to G7) are observed. Two additional

weak resonances observed at ~14.2 and ~12.1 ppm are assigned to the U10/U11 and G9/G12 imino protons of the 5′-overhang, respectively. These signals confirm that a significant fraction of the **P/T$_D$** duplex is in its dimerized form. As the concentration of GMP is increased to 3.75 mM (5 equivalents with respect to (**P/T$_D$**)$_2$, these U10/U11 and G9/G12 signals become more intense. In the absence of GMP, these respective protons are in fast exchange with H_2O.[10] The binding of GMP, however, stabilizes their base pairs, slows down the exchange and leads to the observed increases in intensity. Using these resonances as indicators, ~75% of the binding sites of the (**P/T$_D$**)$_2$ are occupied after the addition of roughly two equivalents of GMP (1.5 mM) with respect to (**P/T$_D$**)$_2$. This observation indicates that binding is in the moderate-to-tight regime,[14] and stands in contrast to the much weaker associations observed in the case of **P/T$_{1C}$** and **P/T$_{2C}$**.

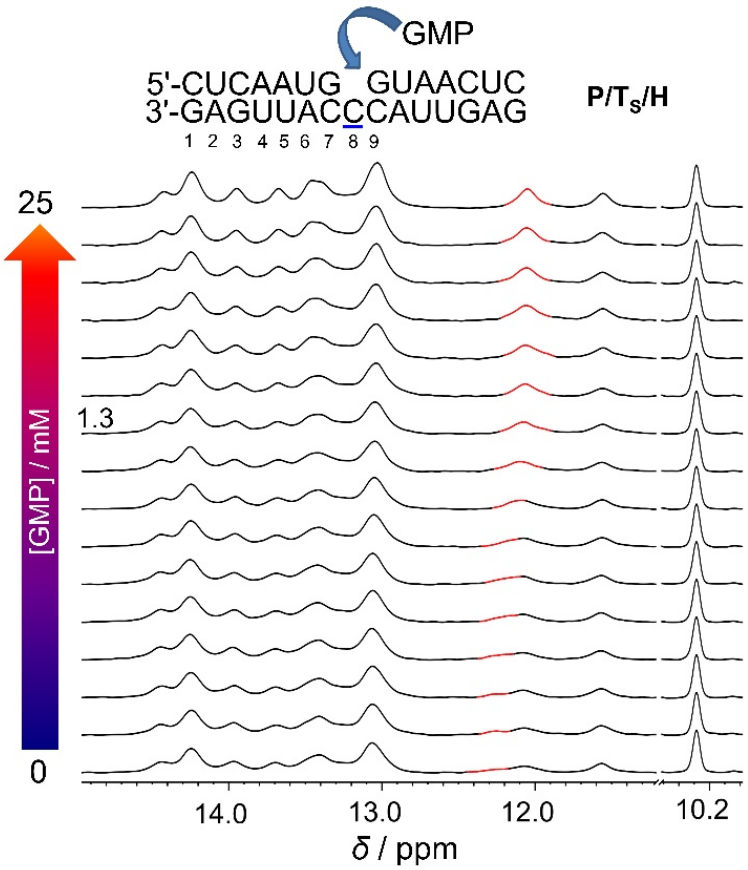


Figure E.11. ^{1}H NMR titration of GMP into a solution containing 1.5 mM of the **P/T$_S$/H** duplex, recorded at 12 ^{o}C, pH 7 and 500 mM Na^{+}. The titration was carried out from 0 to 25 mM of GMP. The resonance highlighted in red likely arise from either the G7 or G9 imino proton.

E.3.8. Titration of GMP into P/T$_S$/H

We carried out a titration of GMP into a solution of **P/T$_S$/H** while monitoring the imino proton region of the ^{1}H NMR spectrum. Due to the complexity of this duplex, many of the imino signals overlapped, and could not be unambiguously assigned. Nonetheless, as we titrated GMP into the solution of **P/T$_S$/H**, we observed what is most likely either the G7 or G9 imino signal around 12.3 ppm becoming sharper, shifting upfield and eventually overlapping with another signal (**Figure E.11**, highlighted in red). After approximately 1 eq of GMP (1.3 mM), no other significant changes in the spectrum could be observed. These observations are consistent with tight binding of GMP into the pocket of the **P/T$_S$/H** duplex.

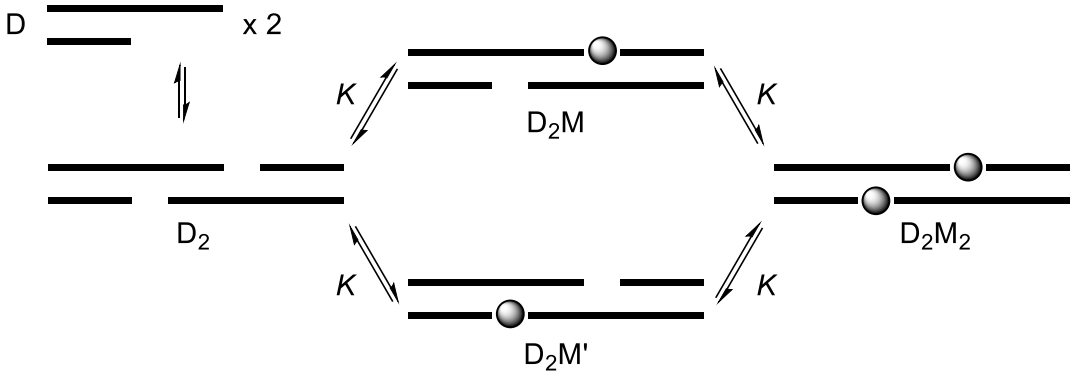
E.4. Isothermal Titration Calorimetry (ITC)

E.4.1. General Methods.

ITC[15,16] was carried out in order to accurately determine the thermodynamic parameters of binding between GMP and the $(\mathbf{P/T_D})_2$ duplex as well as the $\mathbf{P/T_S/H}$ sandwich duplex. All the ITC analyses were done with a TA instruments Nano-ITC calorimeter (New Castle, DE). To setup the experiment, 170 µL of the cell solution, which contains 2 mM each of the $\mathbf{T_D}$ and **P** strands (total RNA strand concentration 4 mM) and 500 mM NaCl at pH 7, was first transferred to the instrument cell. Then, 90 µL of the syringe solution, which contains the same concentration of the $\mathbf{T_D}$ and **P** strands and NaCl as the cell solution in addition to 20 mM GMP, was prepared separately and loaded into the syringe. The entire experimental setup was allowed to equilibrate to 12 ^{o}C for over 1 hour; afterwards, 32 injections of 1.5 µL each were incrementally added to the cell at 200-second intervals. A plot of the resulting raw heat rate (in µcal s^{-1}) versus time (s) was recorded; the plot was then integrated to obtain the heat evolved per injection of syringe solution (in µcal). From this data, the cumulative heat evolved per change in GMP concentration (in kcal mol^{-1}) was measured. This data was then plotted against the cumulative concentration of GMP throughout the titration, resulting in a sigmoidal curve, which was then fitted against the binding isotherm equation derived below. The same protocol was also applied to the $\mathbf{P/T_S/H}$ duplex except at a concentration of 1.5 mM for each strand and at a temperature of 20 ^{o}C, and the data was fit to a single-site binding isotherm.

E.4.2. Derivation of the ITC Binding Isotherm.

Consider the following binding scheme:



Scheme E.1. The proposed binding scheme between GMP and the $(\mathbf{P/T_D})_2$ dimer. To simplify the notation, $\mathbf{P/T_D}$ is referred to as **D** and GMP as **M**.

Our proposed mechanism[13]

begins with the dimerization of the $\mathbf{P/T_D}$ duplex (**D**) to become a dimer ($\mathbf{D_2}$) with two identical binding sites for GMP (**M**) to bind (**Scheme S1**). A GMP monomer first binds to one of the two identical binding sites to form either $\mathbf{D_2M}$ or $\mathbf{D_2M'}$, followed by binding of another GMP to the remaining site to form $\mathbf{D_2M_2}$. Since the two GMP binding sites are symmetric by C_2 rotational symmetry, they are considered equivalent to one another, and hence the binding of GMP to the $(\mathbf{P/T_D})_2$ duplex is likely to be statistical, wherein all four microscopic binding constants K and their associated enthalpy change (ΔH) are assumed to be equal. In this study, the ITC binding isotherm used to fit the measured change in heat per change in GMP concentration as a function of concentration of GMP, is derived using the statistical binding assumption.

Throughout the titration experiment, GMP is incrementally added to the $(\mathbf{P/T_D})_2$ duplex. The heat evolved by the duplex-monomer binding (Q) is a product of: 1) the molar

enthalpy of duplex-monomer binding, 2) the volume of the ITC cell, and 3) the concentrations of monomer-bound complexes **D_2M**, **D_2M'**, **and D_2M_2**. Thus:

$$Q = \Delta H \cdot V_0 \cdot ([D_2M] + [D_2M'] + 2[D_2M_2]) \quad (1)$$

Since the concentration of monomer-bound duplexes can be expressed as the product of total complex concentration $D_{2^{tot}}$ and the mole fraction of the monomer-bound duplexes f_{D_2M} , $f_{D_2M'}$, and $f_{D_2M_2}$, we get:

$$Q = \Delta H \cdot V_0 \cdot D_{2^{tot}} \cdot \left(f_{D_2M} + f_{D_2M'} + 2f_{D_2M_2}\right) \quad (2a)$$

Note that the formation of the **D_2M_2** complex has twice the molar enthalpy of the **D_2M** and **D_2M'** complex. Equation 2a can also be rewritten as:

$$Q = \Delta H \cdot V_0 \cdot D_{2^{tot}} \cdot \left(\frac{[D_2M] + [D_2M'] + 2[D_2M_2]}{[D_2] + [D_2M] + [D_2M'] + [D_2M_2]}\right) \quad (2b)$$

Given that $[D_2M] = [D_2M'] = K[M][D_2]$, and that $[D_2M_2] = K^2[M]^2[D_2]$, equation (2b) now becomes:

$$Q = \Delta H \cdot V_0 \cdot D_{2^{tot}} \cdot \left(\frac{2K[M][D_2] + 2K^2[M]^2[D_2]}{[D_2] + 2K[M][D_2] + K^2[M]^2[D_2]}\right)$$

$$= \Delta H \cdot V_0 \cdot D_{2^{tot}} \cdot \left(\frac{2[D_2]K[M](1 + K[M])}{[D_2](1 + K[M])^2}\right)$$

$$= \Delta H \cdot V_0 \cdot 2D_{2^{tot}} \cdot \left(\frac{K[M]}{1 + K[M]}\right) \quad (2c)$$

Also, since two **P/T_D** duplexes (**D**) engage in self-dimerization to become **D_2**, $2D_{2^{tot}} = D_{tot}$. Later in the derivation process, we will introduce a stoichiometric factor, *n*, that can account for incomplete dimerization of **P/T_D** duplexes. Therefore, in the ideal scenario, *n* = 2. Assuming complete self-dimerization, we have:

$$Q = \Delta H \cdot V_0 \cdot D_{tot} \cdot \left(\frac{K[M]}{1 + K[M]}\right) \quad (3)$$

The next step is to take the derivative of Q with respect to M_{tot}, the total monomer concentration in order to arrive at the final form of the ITC binding isotherm. This isotherm (equation 3) is mathematically equivalent to the non-dimerized **P/T$_D$** duplexes **D** binding a single monomer **M**, as shown by the following equilibrium:

$$[D] + [M] \overset{K}{\rightleftharpoons} [DM]$$

where *K* is equal to:

$$K = \frac{[DM]}{[D] \cdot [M]} \tag{4}$$

We will use this simpler scenario to arrive at the final ITC binding isotherm. Upon addition of monomer, some duplex will bind to the monomer to form [DM] while some will remain in the unbound state. The same logic applies to the monomer as well. The total concentration of duplex and dimer at any point in the titration is given by:

$$[D] + [DM] = [D_{tot}]\,;\; [M] + [DM] = [M_{tot}] \tag{5}$$

Combining equations (4) and (5) gives:

$$K = \frac{[DM]}{(D_{tot} - [DM]) \cdot (M_{tot} - [DM])} \tag{6}$$

$$[DM]^2 - \left(M_{tot} + D_{tot} + \frac{1}{K}\right)[DM] + M_{tot}D_{tot} = 0 \tag{6a}$$

Hence, the two roots (solutions) of equation 6a are:

$$[DM] = \frac{\left(M_{tot} + D_{tot} + \frac{1}{K}\right) \pm \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}{2} \tag{7}$$

One of the two roots must be rejected in order to develop the proper ITC binding isotherm. In order to gain some insight, we consider the limit of [DM] when *K* approaches positive infinity ($\lim_{K\to\infty}[DM]$). The two roots bifurcate into M_{tot} (larger root) and D_{tot} (smaller root);

however, since M_{tot} > D_{tot} at the end of the titration, [DM] must be limited by the total duplex concentration (D_{tot}). Hence [DM] cannot be equal to M_{tot}, and the larger root is rejected. Hence:

$$[DM] = \frac{\left(M_{tot} + D_{tot} + \frac{1}{K}\right) - \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}{2} \qquad (7a)$$

Throughout the titration experiment, monomer (GMP) is incrementally added to the cell containing the duplex solution, and then engages in monomer-duplex binding. Therefore, the change in concentration of the duplex-monomer complex (*d*[DM]) is necessarily related to the change in total monomer concentration (*d*M_{tot}). Hence:

$$\frac{d[DM]}{dM_{tot}} = \frac{d}{dM_{tot}} \frac{\left(M_{tot} + D_{tot} + \frac{1}{K}\right) - \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}{2}$$

$$\frac{d[DM]}{dM_{tot}} = \frac{1}{2}\left\{1 - \left[\frac{\left(M_{tot} - D_{tot} + \frac{1}{K}\right)}{\left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}\right]\right\} \qquad (8)$$

The total heat evolved by duplex-monomer binding (Q) is a product of: 1) the molar concentration of the duplex-monomer complex, 2) the volume of the cell, and 3) the enthalpic change of duplex-monomer binding. Thus:

$$Q = \Delta H \cdot V_0 \cdot [DM] \qquad (9)$$

By differentiating equation 9 with respect to M_{tot}, we get:

$$\frac{dQ}{dM_{tot}} = \Delta H \cdot V_0 \cdot \frac{d[DM]}{dM_{tot}} \qquad (10)$$

By incorporating equation 8 into equation 10, we get the ITC binding isotherm (equation 11).

$$\frac{dQ}{dM_{tot}} = \frac{\Delta H \cdot V_0}{2} \cdot \left\{ 1 - \left[\frac{\left(M_{tot} - D_{tot} + \frac{1}{K} \right)}{\left[\left(M_{tot} + D_{tot} + \frac{1}{K} \right)^2 - 4 M_{tot} D_{tot} \right]^{\frac{1}{2}}} \right] \right\} \tag{11}$$

At this stage, D_{tot} is substituted with $nD_{2^{tot}}$ to reflect the fact that GMP is binding with the self-dimerized (**P/T$_D$**)$_2$ duplex during the ITC experiment. As elaborated before, the stoichiometric factor *n* reflects the degree of dimerization **P/T$_D$** duplex, and hence should be equal to 2 when fit to the data. To account for any heat that originates from non-idealities, e.g. heats of dilution, a constant *a* is added, yielding the final form:

$$\boxed{\frac{dQ}{dM_{tot}} = \frac{\Delta H \cdot V_0}{2} \cdot \left\{ 1 - \left[\frac{\left(M_{tot} - nD_{2^{tot}} + \frac{1}{K} \right)}{\left[\left(M_{tot} + nD_{2^{tot}} + \frac{1}{K} \right)^2 - 4 n M_{tot} D_{2^{tot}} \right]^{\frac{1}{2}}} \right] \right\} + a}$$

E.4.3. Titration of GMP into (P/T$_D$)$_2$: Derivation of Binding Constants by NMR.

The prior section established that the binding of GMP to the (**P/T$_D$**)$_2$ duplex can be satisfactorily explained by the statistical binding model. This finding implies that, at any given point of the ^{1}H NMR titration of GMP with (**P/T$_D$**)$_2$, the equation below can be used to estimate the association constant *K* for this binding event:

$$K = \frac{[DM]}{(D_{tot} - [DM]) \cdot (M_{tot} - [DM])} \tag{6}$$

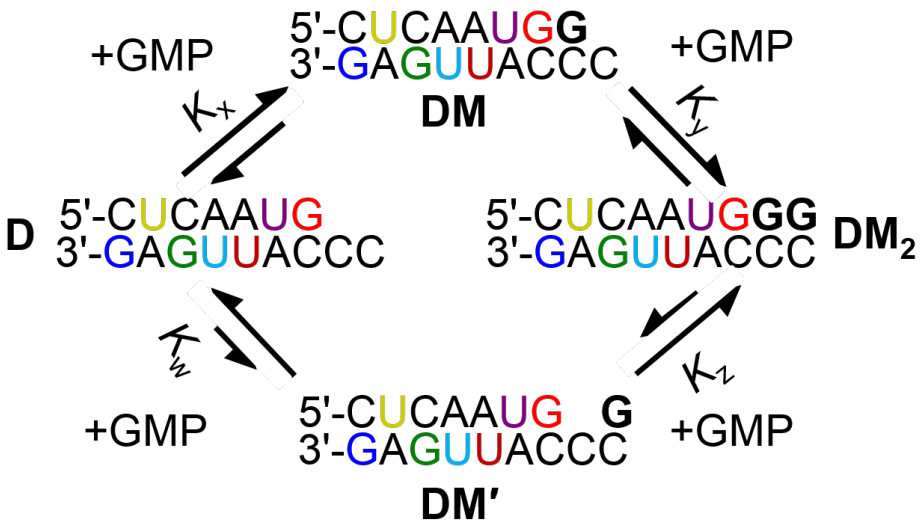
Proton NMR titration of GMP into (**P/T$_D$**)$_2$ was carried out in accordance to procedures laid out in **Section S1d** of this Supporting Information. GMP was titrated into a duplex solution containing 1.5 mM of **P/T$_D$** (0.75 mM of (**P/T$_D$**)$_2$). Binding of GMP to (**P/T$_D$**)$_2$ was monitored by the intensity increase of the U10/U11 and G9/G12 imino proton signals of (**P/T$_D$**)$_2$.

Within the main text, we suggest that ~75% of the binding sites offered by $(\mathbf{P/T_D})_2$ are occupied based on the intensities of the U10/U11 and G9/G12 signals after the addition of 1.5 mM of GMP, or two equivalents of GMP relative to the maximum theoretical concentration of $(\mathbf{P/T_D})_2$ (0.75 mM). Therefore, M_{tot} = 1.5 mM. D_{tot} corresponds to the maximum concentration of $\mathbf{P/T_D}$ duplex, which is also 1.5 mM. With ~75% of the binding sites occupied, the concentration of duplex-GMP complex [DM] necessarily equals to 0.75 times the maximum theoretical concentration of the $\mathbf{P/T_D}$ duplex. Therefore, *K* equals:

$$K = \frac{0.75 \times 0.0015\ M}{(0.0015M - 0.75 \times 0.0015\ M)^2} = 8000\ M^{-1}$$

E.5. Derivation of the Sequential Two-Site Binding Isotherms

The sequential two-site binding isotherms[13,17] used to measure the first and second binding constants for the ^{1}H NMR titration data are based on the following binding scheme:



Scheme E.2. The proposed binding scheme between the 5′-CC overhang of the **P/T$_{2C}$** RNA duplex and GMP. **D** stands for duplex and **M** stands for the GMP monomer.

Where K_x, K_y, K_w and K_z are all the microscopic binding constants governing the specific binding of GMP to either of the 5′-CC overhangs of the template. When binding of the GMP monomer is in fast equilibrium on the timescale of the ^{1}H NMR experiment, the observed NMR signal of interest can be written in terms of a weighted average based on the mole fractions of each equilibrating species in solution together with their respective resonance frequencies:

$$\delta_{\mathrm{OBS}} = \delta_{\mathrm{D}} f_{\mathrm{D}} + \delta_{\mathrm{DM}} f_{\mathrm{DM}} + \delta_{\mathrm{DM'}} f_{\mathrm{DM'}} + \delta_{\mathrm{DM_2}} f_{\mathrm{DM_2}} \qquad (1)$$

where δ_D, δ_{DM}, $\delta_{DM'}$ and δ_{DM2} are the chemical shifts of the respective species labeled in **Scheme S2**, and f_D, f_{DM}, $f_{DM'}$ and f_{DM2} are the corresponding mole fractions. In order to simplify the notation, D stands for the **P/T$_{2C}$** duplex and M stand for the GMP monomer.

Since the sum of the mole fractions must be equal to one, equation (1) can be rewritten as:

$$\delta_{\mathrm{OBS}} - \delta_{\mathrm{D}} = (\delta_{\mathrm{DM}} - \delta_{\mathrm{D}})f_{\mathrm{DM}} + (\delta_{\mathrm{DM'}} - \delta_{\mathrm{D}})f_{\mathrm{DM'}} + (\delta_{\mathrm{DM_2}} - \delta_{\mathrm{D}})f_{\mathrm{DM_2}}$$

Defining $\delta - \delta_{\mathrm{D}}$ as $\Delta\delta$ leads to:

$$\Delta\delta_{\mathrm{OBS}} = \Delta\delta_{\mathrm{DM}}f_{\mathrm{DM}} + \Delta\delta_{\mathrm{DM'}}f_{\mathrm{DM'}} + \Delta\delta_{\mathrm{DM_2}}f_{\mathrm{DM_2}} \quad (2)$$

In the main text $\Delta\delta_{\mathrm{DM}}$ and $\Delta\delta_{\mathrm{DM2}}$ are referred to as $\Delta\delta_{\mathbf{P/T2C}:\mathrm{GMP}}$ and $\Delta\delta_{\mathbf{P/T2C}:\mathrm{2GMP}}$, respectively. Rewriting the mole fractions in terms of concentrations yield:

$$\Delta\delta_{\mathrm{OBS}} = \frac{\Delta\delta_{\mathrm{DM}}[\mathrm{DM}] + \Delta\delta_{\mathrm{DM'}}[\mathrm{DM'}] + \Delta\delta_{\mathrm{DM_2}}[\mathrm{DM_2}]}{[\mathrm{D}] + [\mathrm{DM}] + [\mathrm{DM'}] + [\mathrm{DM_2}]} \quad (3)$$

Expressing concentrations in terms of their microscopic equilibrium constants result in:

$$\Delta\delta_{\mathrm{OBS}} = \frac{\Delta\delta_{\mathrm{DM}}K_x[\mathrm{M}] + \Delta\delta_{\mathrm{DM'}}K_w[\mathrm{M}] + \Delta\delta_{\mathrm{DM_2}}K_xK_y[\mathrm{M}]^2}{1 + K_x[\mathrm{M}] + K_w[\mathrm{M}] + K_xK_y[\mathrm{M}]^2} \quad (4)$$

This equation is the form of the isotherm before any assumptions about the values of the microscopic equilibrium constants and Δδ have been made. By making different assumptions about the relative values of K_x, K_w, and $\Delta\delta_{\mathrm{DM}}$, $\Delta\delta_{\mathrm{DM'}}$ and $\Delta\delta_{\mathrm{DM2}}$, different models can be obtained. The first model we investigate makes these assumptions: (*i*) the binding is sequential, *i.e.*, that K_x is much greater than K_w, (*ii*) $\Delta\delta_{\mathrm{DM'}}$ is less than $\Delta\delta_{\mathrm{DM}}$ and (*iii*) $\Delta\delta_{\mathrm{DM2}}$ is equal to $\Delta\delta_{\mathrm{DM}}$. The assumption that $K_x >> K_w$ is reasonable from the point of view that GMP binding to the terminal C of the 5'-CC overhang would be without any base-stacking interactions[6] with the 3'-G terminus of the **G** strand. The assumptions about $\Delta\delta_{\mathrm{DM'}}$ and $\Delta\delta_{\mathrm{DM2}}$ are based on the fact that ring current effects brought on by bound GMP are distance dependent, and so presumably the GMP bound in the second position is too far away to exert a large effect on the G7 or other resonances. With these assumptions, equation (4) reduces to:

$$\Delta\delta_{\mathrm{OBS}} = \frac{\Delta\delta_{\mathrm{MAX}}(K_x[\mathrm{M}] + K_xK_y[\mathrm{M}]^2)}{1 + K_x[\mathrm{M}] + K_xK_y[\mathrm{M}]^2} \quad (5a)$$

where $\Delta\delta_{MAX}$ is equal to $\Delta\delta_{DM2}$. There are other observations with respect to the observed chemical shifts of the imino protons, however, which are not totally consistent with this model, *i.e.*, that $\Delta\delta_{DM2}$ is equal to $\Delta\delta_{DM}$ for the case of the G7 resonance, and for this reason, we tested a range of other plausible values. Using the **P/T$_{1C}$** titration as a control, the most affected chemical shift is that for the G7 proton closest to the binding site, followed by the U6 proton, the U5, and then the U4. (The resonances for the G1 and U2 protons also undergo significant changes in chemical shifts, although for a different process. See sections on the ^{1}H NMR of **P/T$_0$** and **P/T$_{1C}$** duplexes, and the explanations therein, pg. S11–S15). These observations indicate that the degree of change in chemical shift is related to the proximity of the respective proton to the 5'-binding site. From these data, we can gauge that an imino proton two bases away from the binding site undergoes a chemical shift change which is approximately 25% of the change for an imino proton immediately next to the binding site. From these observations, we can reason in the case of our two-binding site model for **P/T$_{2C}$**, that the binding of the second GMP causes an upfield change in chemical shift of the G7 proton that is 25% that of the first, *i.e.*, $\Delta\delta_{DM2} = 1.25\Delta\delta_{DM}$.

$$\Delta\delta_{\mathrm{OBS}} = \frac{\Delta\delta_{\mathrm{DM}}(K_x[\mathrm{M}] + 1.25K_xK_y[\mathrm{M}]^2)}{1 + K_x[\mathrm{M}] + K_xK_y[\mathrm{M}]^2} \quad (5b)$$

We next reasoned the largest conceivable upfield chemical shift produced by the GMP bound in the second position would be that it yielded as much of a change in chemical shift as that of the first monomer, *i.e.*, $\Delta\delta_{DM2} = 2\Delta\delta_{DM}$. On the other hand, looking to the **P/T$_{1C}$** titration again, the maximum change in the G7 resonance is about 10% more than

that observed for the **P/T$_{2C}$** titration. Therefore, it is also possible that the GMP bound in the second position actually induces a slight *downfield* shift of the G7 resonance, *i.e.*, $\Delta\delta_{DM2} = 0.9\Delta\delta_{DM}$. We fit each of these scenarios as well as many other ratios of $\Delta\delta_{DM2}$ to $\Delta\delta_{DM}$ in between, and found no matter what the value of $\Delta\delta_{DM2}/\Delta\delta_{DM}$ employed during the best fit of the G7 titration data, K_x is always greater than K_y (**Figure E.12**).

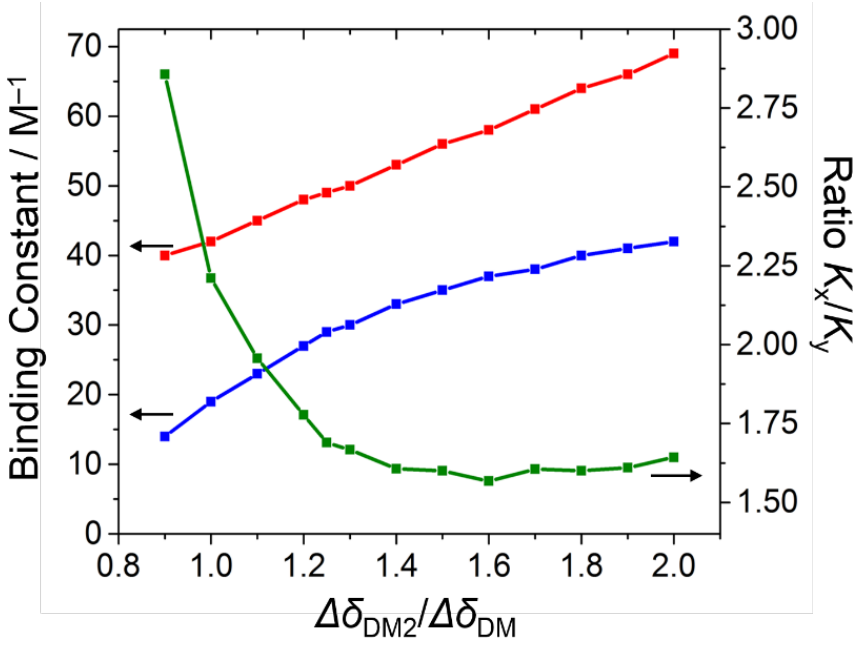
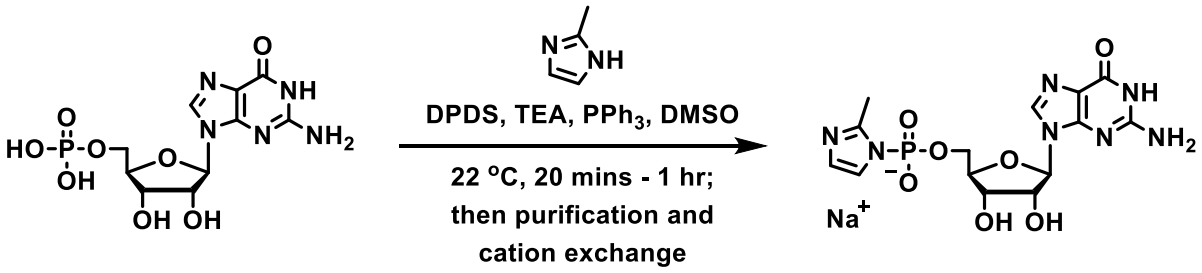


Figure E.12. The effect of varying the ratio of $\Delta\delta_{DM2}/\Delta\delta_{DM}$ on the fitted values of K_x (red) and K_y (blue) and their ratio (green). Each point on the graph represents binding constants determined from the best fit to the model described above using the specified ratio of $\Delta\delta_{DM2}/\Delta\delta_{DM}$. R^2 values for each fit were greater than 0.999. Note that, in the main text, $\Delta\delta_{DM}$ and $\Delta\delta_{DM2}$ are referred to as $\Delta\delta_{\mathbf{P/T2C}:GMP}$ and $\Delta\delta_{\mathbf{P/T2C}:2GMP}$, respectively.

E.6. Nonenzymatic Primer Extension

E.6.1. Preparation of guanosine 5′-phosphoro-2-methylimidazolide (2-MeImpG) sodium salt. The protocol employed for the synthesis of 2-MeImpG was adapted from that of Joyce, Inoue and Orgel.[18]



<u>*Pre-reaction lyophilization.*</u> Guanosine 5'-monophosphate (GMP) *free acid* (1 mmole, 1 equiv., 363 mg) and 2-methylimidazole (10 mmoles, 10 equiv., 821 mg) was dissolved in 20 mL of deionized water in a 250 mL round bottom flask, followed by pH adjustment, with 32% aqueous hydrochloric acid, to a pH of 5.5. The resulting solution was lyophilized at < 30 mtorr and room temperature for 2 days to remove water. *This step was performed to facilitate the subsequent dissolution of GMP starting material in dimethyl sulfoxide (DMSO).*

<u>*Dipyridyl disulfide reaction and crude product isolation.*</u> The dried mixture was dissolved in 75 mL anhydrous DMSO (13.3 mM), followed by addition of triphenylphosphine (10 mmoles, 10 equiv., 2.623 g) and anhydrous triethylamine (10 mmoles, 10 equiv., 1.4 mL). The mixture was heated, sonicated and stirred <u>*until complete homogenization*</u> (typically took >10 mins), followed by the addition of 2,2′-dipyridyl disulfide (DPDS, 10 mmoles, 10 equiv., 2.2 g). Upon addition of DPDS, the reaction mixture turned golden yellow. Reaction progress was monitored by ^{31}P NMR spectroscopy; NMR samples were prepared by transferring ca. 250 µL of the reaction mixture into an NMR tube containing

ca. 250 µL of DMSO-d_6. Using *in-situ* generated triphenylphosphine oxide as an internal standard (δ = 27 ppm), reaction progress was monitored by the disappearance of the starting GMP in the 0–1 ppm region. *This reaction typically took 20 mins to 1 hour to complete*. The crude product was obtained, as an off-white precipitate, by adding the reaction mixture slowly into an ice-cold solution of 200 mL acetone, 125 mL diethyl ether, 15 mL of triethylamine, and 2 mL of saturated sodium perchlorate ($NaClO_4$) in acetone. The resulting mixture was chilled in a –20 °C freezer overnight to ensure complete product precipitation. The precipitate was then collected by suction filtration, followed by the removal of volatiles under high vacuum.

Reverse-phase purification of crude product. The dried crude material was dissolved in 5 mL of triethylammonium bicarbonate buffer (500 mM, pH 7.5), and the resulting solution was subjected to reverse-phase purification on Teledyne Isco's CombiFlash Rf system (Lincoln, NE) on a prepacked 50g C18-Aq column, on a gradient of 0–12.5% acetonitrile (Buffer A: 20 mM aqueous triethylammonium bicarbonate at pH 7.5) over 15 column volume (~640 mL). Desired product was typically eluted from the column at ~5% acetonitrile, which was immediately frozen in liquid nitrogen, followed by low-temperature lyophilization (<–10 °C, < 20 mTorr) on SP Scientific's VirTis AdVantage 2.0 benchtop lyophilizer for 1 week.

Cation exchange of final product. The residual triethylammonium cation of the purified 2-MeImpG product was exchanged with sodium cation, by dissolving the sample in a minimum amount of deionized water, followed by re-precipitation in >25-fold excess (in volume) of 1:1:0.04 (v/v/v) acetone : diethyl ether : saturated $NaClO_4$ in acetone. The precipitate was centrifuged and washed with 1:1 acetone: diethyl ether twice. Final drying

of the centrifuged and washed sample on high-vacuum afforded 2-MeImpG (Na^+ form) as fine white granular solids.

Note: Usage of GMP *disodium salt* in lieu of free acid form diminishes the isolated product yield, presumably due to its poor solubility in anhydrous DMSO. Likewise, failure to completely dissolve GMP-free acid into DMSO *prior* to the DPDS treatment diminishes product yield.

E.6.2. Primer Extension: Experimental Conditions. Three sets of primer extension reactions were set up, in 20 μL-scales, with the following components: 200 μM fluorescein-labelled **P** primer, 220 μM template (set 1, $\mathbf{T_{1C}}$; set 2, $\mathbf{T_{2C}}$; set 3, $\mathbf{T_D}$), 250 mM pH 7.5 Na•HEPES buffer, 200 mM $MgCl_2$, and 50 mM 2-MeImpG. Primer, template, pH 7.5 HEPES buffer and aqueous $MgCl_2$ solutions were first combined and thoroughly mixed in 200 μL thin-walled PCR tubes, followed by adding an appropriate amount of 2-MeImpG stock solutions (≥200 mM, determined by Nanodrop) onto the *lids* of the PCR tubes. To *simultaneously* initiate the extension reactions in all three sets of conditions, the 2-MeImpG droplets on the lids of the PCR tubes were spun-down into their respective primer extension reaction mixtures using a table-top centrifuge. At appropriate time points, 2 μL aliquots of these reactions were quenched by addition to 38 μL of quench buffer containing 1X TBE (89 mM Tris-borate and 2 mM EDTA, overall pH 8.3), 8.0 M urea, and 100 mM pH 8.0 Na•EDTA. These quenched samples were further diluted 25 fold into a gel loading buffer solution containing 1X TBE and 8.0M urea (500-fold dilution of the original primer extension reaction mixture in total). Finally, 4 μL of these diluted samples, which contained 0.4 μM of the fluorescently-labelled **P** primer as well as their extended products, were loaded onto a 400 mm × 300 mm × 0.4 mm polyacrylamide denaturing

gel (gel mixture contains 20% acrylamide monomers with a ratio of 19:1 acrylamide : *N,N´*-methylenebis-acrylamide, as well as 7.5 M urea), followed by electrophoresis at 50 W for 4 hours. Imaging and quantification of the fluorescently labelled **P** and extended products were performed with an Amersham Biosciences Typhoon 9410 Imager, and the ImageQuant TL software package (GE Healthcare, Little Chalfont, United Kingdom).

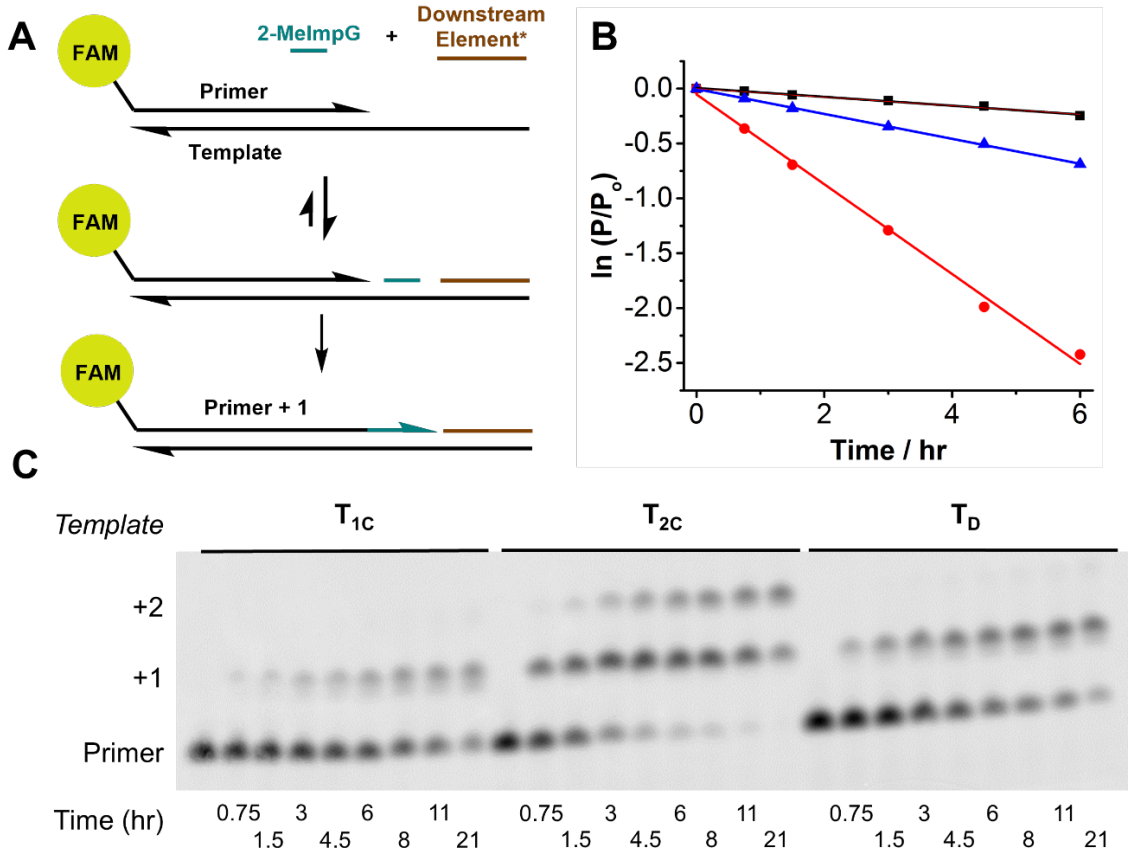


Figure E.13. Non-enzymatic primer extension reactions, templated by $\mathbf{T_{1C}}$, $\mathbf{T_{2C}}$, and $\mathbf{T_D}$, were performed to assess how different downstream binding elements affect the rate of extension of **P** strand. **a.** Schematic diagram of the assay. The 5′-FAM-labeled **P** primer was annealed with templates $\mathbf{T_{1C}}$, $\mathbf{T_{2C}}$, and $\mathbf{T_D}$, and incubated with 50 mM 2-MeImpG monomer (cyan) at room temperature and at pH 8.0, with 200 mM $MgCl_2$ and 250 mM Na·HEPES. The 2-MeImpG and downstream elements (brown) first bind noncovalently to the primer-template complexes, followed by an irreversible extension step to afford the primer + 1 product. For $\mathbf{P/T_{2C}}$, the downstream element is another molecule of 2-MeImpG, while that for the case of $(\mathbf{P/T_D})_2$ is the self-complementary region of another $\mathbf{T_D}$ template oligomer. **b.** Plot of the natural log value of the fraction of unreacted primer (ln (**P**/P_o)) against time. The linear fits of these plots were used to determine the pseudo first-order observed rate constants for $\mathbf{P/T_{1C}}$ (■), $\mathbf{P/T_{2C}}$ (•), and $(\mathbf{P/T_D})_2$ (▲), which are 0.039(1), 0.42(1), and 0.11(1) hr^{-1} respectively. **c.** A representative 20% denaturing PAGE analysis of the products of the primer extension reactions.

E.7. Structural Characterizations of RNA Duplexes

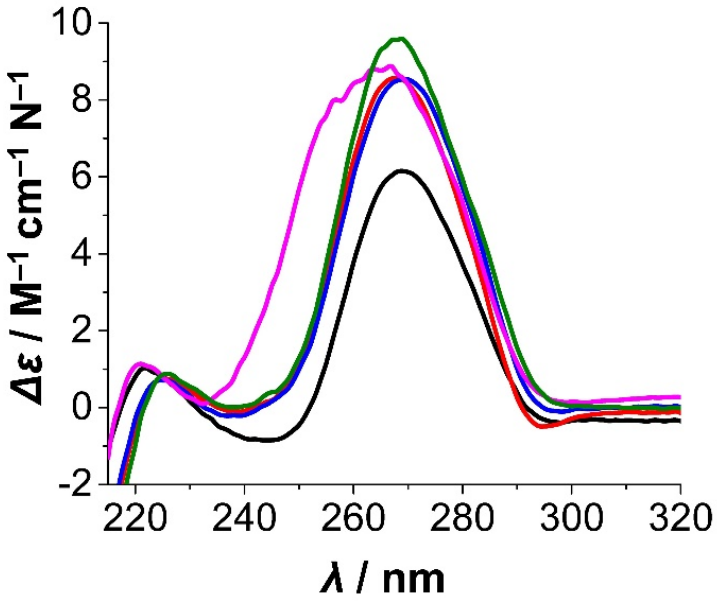


Figure E.14. Overlaid CD spectra of all RNA duplexes studied herein (**P/T$_0$**, black; **P/T$_{1C}$**, red; **P/T$_{2C}$**, blue; **P/T$_D$**, green; **P/T$_s$/H**, pink). All CD spectra were recorded in the presence of 500 mM NaCl, 5 mM Na_2HPO_4, 0.5 mM EDTA at pH 7 and 12°C, with a duplex concentration of 0.1 mM. The solution is contained in a quartz spectrophotometer cell with 1 mm pathlength. All spectra were baselined with the CD spectrum of the buffer solution without the duplexes (500 mM NaCl, 5 mM Na_2HPO_4, 0.5 mM EDTA, and pH 7).

Analysis by circular dichroism (CD) spectroscopy of all the duplexes studied herein was carried out to ascertain the conformation of the double helices. All CD experiments were performed on an Aviv Biomedical Model 202 CD Spectrometer (Lakewood, NJ). The CD spectra for **P/T$_0$**, **P/T$_{1C}$**, **P/T$_{2C}$**, **P/T$_D$** and **P/T$_s$/H** duplexes are shown in **Figure E.14**. All spectra were obtained under the same conditions. For each duplex, the CD spectrum displays one large positive peak around 270 nm in addition to a very small negative dip between 230 and 250 nm. These spectral characteristics indicate that the RNA double helices are in their A-form conformation.[19]

E.8. Nearest neighbor analysis of P/T_S/H duplex

We used nearest-neighbor analysis[20] to approximate the degree to which helper oligonucleotide **H** influences the binding affinity of GMP. In essence, we first calculated the free energy of formation of the fully complementary **P/T_S/H** duplex at 20 °C, assuming that the bound GMP can be treated as if it were covalently attached with both **P** and **H** to form a perfectly complementary strand with **T_S**. We then repeat the same calculation on the original duplex taking into account the one-nucleotide discontinuity in the middle, subtracting this energy from the former. In the absence of the bound GMP, this discontinuity is modeled as two partial duplexes, in which one has a 5′-dangling end and the other a 3′-dangling end. Since the only difference between the two scenarios is the one-nucleotide discontinuity, the enthalpic and entropic parameters of the nucleotide sequences flanking the middle one-nucleotide discontinuity will eventually cancel out. In practice, we can conveniently approximate the GMP binding with **P/T_S/H** duplex by calculating the free energy difference between the following two scenarios (bolded and underlined nucleotide represents the GMP monomer):

Scenario 1: 5′–G**<u>G</u>**G–3′ / 3′–CCC–5′ Scenario 2: 5′–G G–3′ / 3′–CCC–5′

Enthalpic difference [$\Delta(\Delta H)$] between scenario 1 and 2 = –12.2 – 12.2 – (–4.1 + 3.3) = –23.6 kcal mol^{-1}

Entropic difference [$\Delta(\Delta S)$] between scenario 1 and 2 = –29.7 – 29.7 – (–10.7 + 11.8) = –60.5 cal K^{-1} mol^{-1}

Difference in free energy of binding [$\Delta(\Delta G)$] between scenario 1 and scenario 2 at 20 °C (293.15 K)= $\Delta(\Delta H) - T\Delta(\Delta S)$ = –5.86 kcal mol^{-1}

Hence, the estimated K_a of GMP to **P/T$_S$/H** was:

$$e^{-\Delta G/RT} = e^{5.86/0.00198\times 293.15} = 2.35 \times 10^4$$

It is important to note that this analysis should be expected to overestimate the binding energy on account of the fact that the GMP is only noncovalently bound, which it does.

E.9. References

(1) Piotto, M.; Saudek, V.; Sklenár, V. *J. Biomol. NMR* **1992**, *2*, 661–665.

(2) Lippens, G.; Dhalluin, C.; Wieruszeski, J. M. *J. Biomol. NMR* **1995**, *5*, 327–331.

(3) Cantor, C. R.; Warshaw, M. M.; Shapiro, H. *Biopolymers* **1970**, *9*, 1059–1077.

(4) Cavaluzzi, M. J.; Borer, P. N. *Nucleic Acids Res.* **2004**, *32*, e13.

(5) Inoue, T.; Orgel, L. E. *J. Mol. Biol.* **1982**, *162*, 201–217.

(6) Kanavarioti, A.; Hurley, T. B.; Baird, E. E. *J. Mol. Evol.* **1995**, *41*, 161–168.

(7) Wu, T.; Orgel, L. E. *J. Am. Chem. Soc.* **1992**, *114*, 5496–5501.

(8) Conte, M. R.; Conn, G. L.; Brown, T.; Lane, A. N. *Nucleic Acids Res.* **1996**, *24*, 3693–3699.

(9) Lane, A. N.; Jenkins, T. C.; Frenkiel, T. A. *Biochim. Biophys. Acta.* **1997**, *1350*, 205–220.

(10) Gueron, M.; Leroy, J. L. *Methods Enzym.* **1995**, *261*, 383–413.

(11) Izgu, E. C.; Fahrenbach, A. C.; Zhang, N.; Li, L.; Zhang, W.; Larsen, A. T.; Blain, J. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2015**, *137*, 6373–6382.

(12) Gottlieb, H. E.; Kotlyar, V.; Nudelman, A. *J. Org. Chem.* **1997**, *62*, 7512–7515.

(13) Connors, K. A. *Binding Constants : The Measurement of Molecular Complex Stability*; Wiley: New York, 1987.

(14) Li, Y.; Griend, D. A. V.; Flood, A. H. *Supramol. Chem.* **2009**, *21*, 111–117.

(15) Wiseman, T.; Williston, S.; Brandts, J. F.; Lin, L. N. *Anal. Biochem.* **1989**, *179*, 131–137.

(16) Schmidtchen, F. P. In *Analytical Methods in Supramolecular Chemistry*; Schalley, C., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Germany, 2007; pp 55–78.

(17) Hirose, K. In *Analytical Methods in Supramolecular Chemistry*; Schalley, C., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Germany, 2007; pp 17–54.

(18) Joyce, G. F.; Inoue, T.; Orgel, L. E. *J. Mol. Biol.* **1984**, *176*, 279–306.

(19) Gratzer, W. B.; Richards, E. G. *Biopolymers* **1971**, *10*, 2607–2614.

(20) Turner, D. H.; Sugimoto, N.; Freier, S. M. *Annu. Rev. Biophys. Biophys. Chem.* **1988**, *17*, 167–192.

Appendix F

Supplemental Materials for Chapter 6

F.1. Preparation of Oligonucleotides for Primer Extension and Crystallography

We cocrystallized ICG with locked-nucleic-acid (LNA)-rigidified RNA primer-template complexes that contained either one or two 5-methylcytidine LNA overhangs at their 5′-termini to favor analog binding to the template. LNA locks the backbone into the A-type helical conformation and facilitates crystallization.[1] The chemical synthesis of the nonhydrolyzable monomer ICG (**Figure 6.1.A**) is reported elsewhere.[2] All the native RNAs for primer extension experiments and structure determination were synthesized in-house by standard solid-phase techniques. Cyanine 3 labeled RNA was purchased from Integrated DNA Technologies (Coralville, IA). The 5′-imidazolyl-phosphonate-capped helper RNA oligomer (sequence: 5′-ICG-CACCUCA-3′) and 5′-GpppG-capped RNA (sequence: 5′-GpppG-CACCUCA-3′) was prepared by *in vitro* transcription using T7 RNA polymerase (**Figure 6.1.B**). The DNA template (5′-***TG*****AGGTGC**TATAGTGAGTCGTA-TTAACG-3′), and its DNA complement containing the T7 promoter sequence (5′-CGTTAATACGACTCACTATA**GCACCTCA**-3′), were prepared by solid-phase DNA synthesis. The bolded 8 nucleotides correspond to the template region of the double-stranded DNA. The two nucleotides at the 5′ end of the DNA template (underlined and italicized T and G) were 2′-*O*-methylated to reduce the heterogeneity of RNA products.[3] HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs, Ipswich, MA) was used for the *in vitro* transcription of oligonucleotides; the DNA template (0.1 mM) was

transcribed in a polymerase-containing cocktail with 10 mM ATP, CTP, UTP, and ICG monomers or GpppG dinucleotide. The reaction was initiated by addition of the polymerase solution (provided within the kit; 60 µL of polymerase solution per 1 mL of transcription reaction), and the mixture was incubated at 37 °C for 16 h in a BioRad T100 thermocycler (Hercules, CA). The transcribed product was extracted by DNase I digestion, RNA ethanol precipitation, and 25% polyacrylamide denaturing gel electrophoresis purification. The purified oligonucleotide was analyzed by LC–MS on an Agilent 1200 HPLC coupled to an Agilent 6230 TOF-MS equipped with an auto sampler and diode array detector. The sample was separated by IP-RP-HPLC on a 100 × 1.0 mm (length × i.d.) Xbridge C18 column using gradient elution between (A) aqueous 200 mM 1,1,1,3,3,3-hexafluoro-2- propanol with 1.25 mM triethylamine, pH 7.0,[4] and (B) methanol, where the sample was eluted between 2.5% and 15% B over 16 min with a flow rate of 0.1 mL/min at 50 °C. Samples were analyzed in negative mode from 239 m/z to 3200 m/z with a scan rate of 1 spectrum/s (**Figures F.1** and **F.2**).

F.2. Primer Extension Assays

We confirmed that our synthetic nonhydrolyzable guanosine 4-methylimidazolylphosphonate, located at the 5′-end of a downstream helper oligonucleotide, could catalyze primer extension using a hairpin primer-template substrate. Nonenzymatic RNA polymerization rates were tested with different downstream helper oligomers (**Figure 6.3.A**). Primer extension reactions contained: 2 µM Cy3-labeled RNA primer-template hairpin (Sequence: 5′-Cy3-UGAGGUGCGGAGCG CGAAAGCGCUC-3′), 250 mM HEPES buffer (pH 8.0), 50 mM $MgCl_2$, 10 mM 2-MeImpC monomer, and 5 µM downstream helper oligonucleotides (Set 1: no helper; Set 2: 5′-OH-

GCACCUCA-3′; Set 3: 5′-PO_3^{2-}-GCACCUCA-3′; Set 4: 5′-ICG-CACCUCA-3′). In this context, the RNA primer component of the hairpin would be extended by one cytidine nucleotide, and the activated cytidine monomer should not bind at the downstream (primer+2) position on the template. Hairpin primer-template duplex, aqueous $MgCl_2$, HEPES buffer and downstream helper oligomer were first combined and thoroughly mixed in 200 µL thin-walled PCR tubes, followed by addition of 2-MeImpC to initiate the primer extension reactions. At appropriate time points, 0.5 µL aliquots of these reactions were quenched by addition to 20 µL of a quench buffer containing 1X TBE (89 mM Tris-borate and 2 mM EDTA, overall pH 8.3), 8.0 M urea, and 100 mM EDTA•$4Na^+$, pH 8.0. 4 µL of these quenched samples, which contained 0.12 pmol of the fluorescently-labeled primer as well as their extended products, were loaded onto a 350 mm × 400 mm × 5 mm 20% denaturing polyacrylamide gel[5] (National Diagnostics, Atlanta, GA), followed by electrophoresis at 50 W for 5 hrs. Reaction rates were calculated by quantifying the conversion of primer to its extended products, and the negative log of the fraction of unreacted primer was plotted against time. A linear regression was per-formed and the slope of the fit as plotted was reported as the pseudo-first order rate k_{obs}. Imaging and quantification of the fluorescently labeled primer and extended products were performed with an Amersham Biosciences Typhoon 9410 Imager, and the ImageJ software package.[6]

F.3. RNA Crystallization, Data Collection, and Structure Determination

We cocrystallized the nonhydrolyzable ICG monomer and the imidazolium-bridged dinucleotide analog GpppG with different RNA primer-template oligonucleotides and solved the structures. The sequences we used included self-complementary RNAs, which

contained one or two binding sites for ICG (**Figure 6.2**), and primer-template hairpin RNAs, which could assemble with either ICG or GpppG and downstream helper oligonucleotides (**Figures 6.4–6.8**), and the hairpin RNA structure containing GpppG-capped downstream oligo (**Figure 6.7**). The detailed crystallization procedures are described *vide infra*, and the optimized crystallization conditions for the RNA-ligand complexes are listed in **Table F.2** (*vide infra*). Either 35% v/v (+/-)-2-methyl-2, 4-pentanediol or the mother liquor containing 50% glycerol were used as cryoprotectants during crystal mounting. The data sets were collected at the SIBYLS beamlines 821 and 822 at the Advanced Light Source, Lawrence Berkeley National Laboratory. The structures were solved by molecular replacement using our previous RNA-PZG complex structure[1] or using the partial or entire structure 4FNJ[7] from the PDB data bank as the search model. The detailed procedures, together with data collection, phasing, and refinement statistics for the 11 determined structures are listed in **Tables F.3** and **F.4**.

Table F.1. Crystallographic and structural features of RNA-ICG complexes.

	RNA 15mer-monomer	RNA 14mer-monomer	H-34-Im	H-34-PO	H-34-OH
PDB code	5V0H	5V0J	5UX3	5V9Z	5V0O
space group	P321	R32	$P4_12_12$	$P4_32_12$	$P4_32_12$
RNA structural features	duplex	duplex	tetraloop /stem	tetraloop /stem	tetraloop /stem
“leaving group” of bound monomer	disordered imidazolyl	disordered imidazolyl	ordered imidazolyl	disordered imidazolyl	disordered imidazolyl
downstream helper	none	ICG monomer	5'-ICG 8mer	5'-phospho-rylated 8mer	5'-hydroxyl 8mer
3'-O--P distance, Å/ 3'-O--P-C angle, °	6/ND	4.6/ND	3.8/100.8	~3.9/~111.1	~3.8/~106.1
2'-O--P distance, Å/ 2'-O--P-C angle, °	7.8/ND	6.8/ND	3.7/120.6	~4.2/~142.4	~4.1/~120.6

Table F.1. (continued) Crystallographic and structural features of RNA-ICG complexes.

	H-35-OH	H-35-PO	H-43-GpppG	H-34-GpppG-oligo	H-34-L	H-35-L
PDB code	5VCF	5VCI	6AZ4	6BMD	5UZ6	5VGW
space group	R3	R3	$P4_12_12$	C2	C_12_1	R3
RNA structural features	tetraloop/ receptor/stem	tetraloop/ receptor/stem	tetraloop/ receptor/stem	tetraloop/stem	tetraloop /stem	tetraloop/ receptor/stem
“leaving group” of bound monomer	disordered imidazolyl	disordered imidazolyl	ordered triphosphate	ordered triphosphate	disordered imidazolyl	none
downstream helper	5'-hydroxyl 13mer	5'-phospho-rylated 13mer	5'-hydroxyl 9mer	none	none	none
3'-O--P distance, Å/ 3'-O--P-C angle, °	~4.4/ND	~4.6/ND	3.7/138.8	~3.2/139.1	~4.4/ND	~4.1/ND
2'-O--P distance, Å/ 2'-O--P-C angle, °	~6.0/ND	~6.4/ND	5.6/126.6	~4.8/149.7	~6.5/ND	~5.7/ND

F.4. LC-MS analysis.

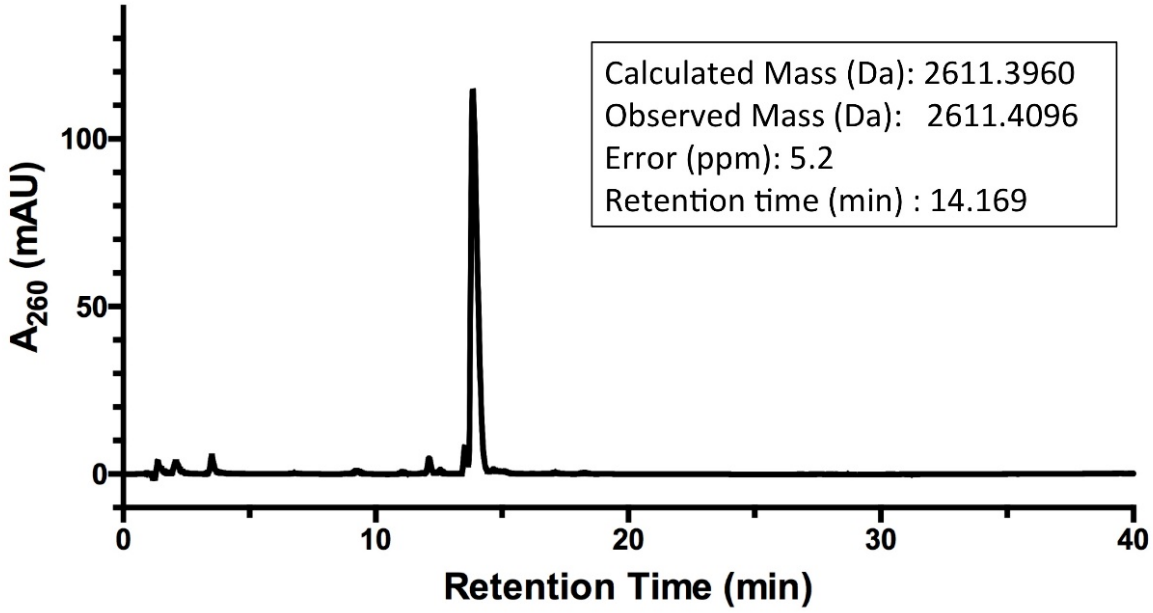


Figure F.1. HPLC-TOF analysis of transcript RNA 8-mer 5′-ICG-CACCUCA-3′ after purification by PAGE-gel. HPLC-TOF analysis was performed as described in **Section F.1**. The RNA retention time was 14.2 min.

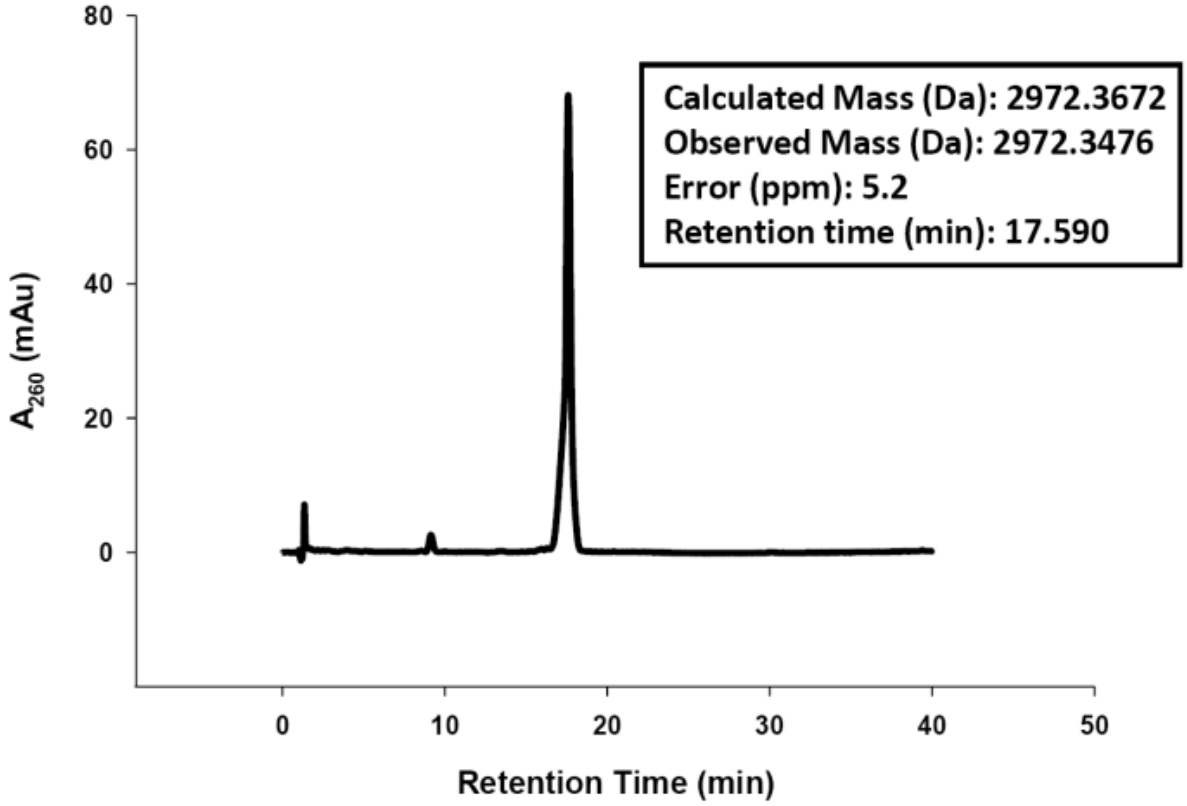


Figure S2. HPLC-TOF analysis of transcript RNA 9-mer 5′-GpppG-CACCUCA-3′ after purification by PAGE-gel. The analysis was performed as described in **Section F.1**, using a linear gradient of 2.5% to 20% buffer B (Methanol) over 20 min. The RNA retention time was 17.6 min.

F.5. X-ray crystallography.

F.5.1. Crystallization. Nucleic Acid Mini Screen Kit, Natrix High Throughput Kit and Index High Throughput Kit (Hampton Research, Aliso Viejo, CA), and JBS Classic HTS II (Jena Biosciences, Jena, Germany) were used for screening crystallization conditions by the sitting drop vapor diffusion method. For RNA duplex-monomer crystallization, RNA samples (1 mM) were mixed with ICG (50 mM) at a 1:1 ratio. For RNA hairpin-monomer/RNA hairpin-GpppG crystallization, RNA template (1 mM), RNA downstream helper oligomer (1 mM) and ICG monomer/GpppG ligand (75 mM) were mixed at a 1:1:1 ratio. For crystallization of the RNA hairpin complex, the RNA template-loop-primer strand (0.5 mM) and the downstream GpppG-helper oligomer strand (0.5 mM) were mixed at a 1:1 ratio. All the components were heated to 80 °C for 2 min, and then cooled down slowly to room temperature. All crystals grew at 18 °C. For RNA duplex-monomer complexes, 35% MPD was used as cryoprotectant during crystal mounting. For all other complexes, the mother liquor containing 50% glycerol was used as cryoprotectant during crystal mounting. The optimized crystallization conditions are listed below.

Table F.2. Optimized conditions for crystallization

	Optimized crystallization conditions
RNA 15mer-ICG (PDB: 5V0H)	50 mM $MgCl_2$, 0.05 M ammonium acetate, 0.05 M tris hydrochloride pH 7.5, 10% v/v (+/-)-2-methyl-2,4-pentanediol
RNA 14mer-ICG (PDB: 5V0J)	50 mM $MgCl_2$, 0.1 M potassium chloride, 0.05 M sodium chloride, 0.1 M HEPES pH 7.5, 15% v/v (+/-)-2-methyl-2,4-pentanediol
RNA H-34-IM (PDB: 5UX3)	50 mM $MgCl_2$, 3 M ammonia sulfate, 100 mM HEPES 7.5
RNA H-34-PO (PDB: 5V9Z)	50 mM $MgCl_2$, 3 M ammonia sulfate, 100 mM HEPES 7.5
RNA H-34-OH (PDB: 5V0O)	50 mM $MgCl_2$, 3 M ammonia sulfate, 100 mM HEPES 7.5
RNA H-35-OH (PDB: 5VCF)	50 mM $MgCl_2$, 0.2 M lithium sulfate monohydrate, 0.1 M BIS-TRIS pH 7.0, 25% w/v Polyethylene glycol 3,350
RNA H-35-PO (PDB: 5VCI)	50 mM $MgCl_2$, 0.2 M lithium sulfate monohydrate, 0.1 M BIS-TRIS pH 7.0, 25% w/v Polyethylene glycol 3,350
RNA H-43-GpppG (PDB: 6AZ4)	50 mM $MgCl_2$, 0.1 M Sodium cacodylate trihydrate pH 6.5, 1.4 M Sodium acetate trihydrate
RNA H-34-GpppG-oligo (PDB: 6BMD)	50 mM $MgCl_2$, 3 M ammonia sulfate, 100 mM HEPES 7.5
RNA H-34-L (PDB: 5UZ6)	50 mM $MgCl_2$, 3 M ammonia sulfate, 100 mM HEPES 7.5
RNA H-35-L (PDB: 5VGW)	50 mM $MgCl_2$, 0.2 M lithium sulfate monohydrate, 0.1 M BIS-TRIS pH 7.0, 25% w/v Polyethylene glycol 3,350

F.5.2. Data collection and structure refinement. All data collection was performed under a stream of nitrogen at 99 K. The data sets were collected at the SIBYLS beamline 821 and 822 at the Advanced Light Source, Lawrence Berkeley National Laboratory. The distances between detector and the crystal were set to 200-300 mm and the collecting wavelength was set to 1 Å. The crystals were exposed for 1 second per image with one degree oscillation, and 180 images were taken for each data set. The data was processed using HKL2000.[8] All the structures were solved by molecular replacement.[9] The RNA models were from our previously reported structures, or based on RNA hairpin structure in the literature. All the structures were refined using Refmac.[10] The refinement protocol included simulated annealing, refinement, restrained B-factor refinement, and bulk solvent correction.[10] During refinement, the topologies and parameters for locked nucleic acids and for the ligand ICG and GpppG (LCC, LCA, LCG, rIMG, GP3) were constructed and applied. After several cycles of refinement, the water and magnesium atoms were added. Data collection, phasing, and refinement statistics of the determined structures are listed in **Tables F.3** and **F.4**.

Table F.3. Data collection statistics.

Structure	RNA 15mer-ICG	RNA 14mer-ICG	H-34-Im complex	H-34-PO complex	H-34-OH complex
PDB code	5V0H	5V0J	5UX3	5V9Z	5V0O
Space group	P321	R32	$P4_12_12$	$P4_32_12$	$P4_32_12$
Unit cell parameters, Å (°)	42.85, 42.85, 83.84 (90, 90, 120)	43.39, 43.39, 253.85 (90, 90, 120)	46.53, 46.53, 156.94 (90, 90, 90)	45.28, 45.28, 154.57 (90, 90, 90)	44.75, 44.75, 154.20 (90, 90, 90)
Resolution range, Å (last shell)	50-1.90 (1.97-1.90)	100-1.50 (1.53-1.50)	50-2.50 (2.59-2.50)	50-2.50 (2.59-2.50)	50-2.70 (2.80-2.70)
Unique reflections	7272 (702)	15290 (744)	6436 (604)	5769 (418)	4657 (437)
Complete-ness, %	97.4 (95.4)	99.4 (99.9)	98.7 (94.1)	95.4 (73.1)	97.3 (94.4)
R_{merge}**, %**	13.7 (46.8)	4.1 (57.5)	8.0 (69.5)	10.0 (71.0)	10.0 (77.7)
<I/σ(I)>	10 (2.9)	35.4 (1.8)	20.2 (1.0)	16.3 (1.7)	18.2 (1.6)
Redundancy	7.3 (4.2)	6.3 (5.6)	5.1 (3.9)	10.6 (7.2)	9.1 (6.4)

Table F.3 (continued). Data collection statistics.

Structure	H-35-OH complex	H-35-PO complex	H-43-GpppG complex	H-34-GpppG-oligo complex	H-34-L complex	H-35-L complex
PDB code	5VCF	5VCI	6AZ4	6BMD	5UZ6	5VGW
Space group	R3	R3	$P4_12_12$	C2	C_12_1	R3
Unit cell parameters, Å (°)	71.07, 71.07, 71.27 (90, 90, 120)	69.61, 69.61, 70.85 (90, 90, 120)	79.39, 79.39, 73.65 (90, 90, 90)	107.39, 63.76, 65.84 (90, 122, 90)	106.69, 63.07, 65.92 (90, 120, 90)	71.06, 71.06, 69.78 (90, 90, 120)
Resolution range, Å (last shell)	50-2.80 (2.90-2.80)	50-2.60 (2.69-2.60)	50-2.98 (3.09-2.98)	50-3.00 (3.11-3.00)	50-2.10 (2.18-2.10)	50-2.40 (2.49-2.40)
Unique reflections	3296 (316)	3864 (338)	4713 (476)	7315 (557)	20855 (1912)	4799 (508)
Completeness, %	99.7 (100)	97.8 (87.8)	91.6 (95.2)	95.9 (74.5)	96.9 (89.0)	96.1 (98.8)
R_{merge}, %	7.2 (49.5)	5.4 (52.6)	13.3 (69.8)	19.0 (37.7)	5.2 (32.3)	7.3 (63.9)
$\langle I/\sigma(I) \rangle$	17.4 (1.5)	18.4 (1.8)	9.2 (1.6)	5.8 (2.0)	15.5 (1.8)	15.0 (2.3)
Redundancy	3.3 (3.3)	3.2 (2.4)	4.5 (4.4)	4.8 (3.2)	2.7 (2.0)	3.7 (3.1)

Table F.4. Structure refinement statistics.

Structure	RNA 15mer-ICG	RNA 14mer-ICG	H-34-Im complex	H-34-PO complex	H-34-OH complex
PDB code	5V0H	5V0J	5UX3	5V9Z	5V0O
Molecules per asymmetric unit	1	1	1	1	1
Resolution range, Å (last shell)	37.11-1.90	84.62-1.50	44.61-2.50	43.45-2.51	42.97-2.70
R_{work}**, %**	22.7	23.7	21.8	21.8	21.8
R_{free}**, %**	27.2	27.0	26.1	28.1	27.7
Number of reflections	6924	14527	6154	5466	4446
Bond length R.M.S., Å	0.026	0.027	0.015	0.013	0.012
Bond angle R.M.S.	2.686	2.963	2.564	2.438	1.781
Average B-factors, Å^2	43.86	36.47	77.74	73.17	86.79

Table F.4 (continued). Structure refinement statistics.

Structure	H-35-OH complex	H-35-PO complex	H-43-GpppG complex	H-34-GpppG-oligo complex	H-34-L complex	H-35-L complex
PDB code	5VCF	5VCI	6AZ4	6BMD	5UZ6	5VGW
Molecules per asymmetric unit	1	1	1	3	3	1
Resolution range, Å (last shell)	46.58-2.80	45.91-2.60	50.00-2.98	50.00-3.00	50.00-2.10	46.15-2.42
R_{work}, %	16.3	16.2	18.1	18.3	22.1	20.73
R_{free}, %	21.3	21.2	20.7	23.0	26.7	21.29
Number of reflections	3154	3649	4464	6922	19827	4546
Bond length R.M.S., Å	0.009	0.010	0.009	0.012	0.014	0.008
Bond angle R.M.S.	1.890	1.787	2.185	2.060	2.452	1.755
Average B-factors, Å^2	64.45	75.43	73.20	26.41	47.84	68.15

F.6. References.

(1) Zhang, W.; Tam, C. P.; Wang, J.; Szostak, J. W. *ACS Cent. Sci.* **2016**, *2*, 916–926.

(2) Tam, C. P.; Zhou, L.; Fahrenbach, A. C.; Zhang, W.; Walton, T.; Szostak, J. W. *J. Am. Chem. Soc.* **2018**, *140*, 783–792

(3) Kao, C.; Zheng, M.; Rüdisser, S. *RNA* **1999**, *5*, 1268–1272.

(4) Apffel, A.; Chakel, J. A.; Fischer, S.; Lichtenwalter, K.; Hancock, W. S. *Anal. Chem.* **1997**, *69*, 1320–1325.

(5) Rio, D. C.; Ares Jr., M.; Hannon, G. J.; Nilsen, T. W. *Cold Spring Harb. Protoc.* **2010**, *2010*, pdb.prot5444.

(6) Abràmoff, M. D.; Magalhães, P. J.; Ram, S. J. *Biophotonics International*. 2004, pp 36–41.

(7) Coonrod, L. A.; Lohman, J. R.; Berglund, J. A. *Biochemistry* **2012**, *51*, 8330–8337.

(8) Otwinowski, Z.; Minor, W. *Methods Enzym.* **1997**, *276*, 307–326.

(9) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. *J. Appl. Crystallogr.* **2007**, *40*, 658–674.

(10) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. *Acta Crystallogr. D Biol. Crystallogr.* **1997**, *53*, 240–255.