Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Nucleobase tautomerism in codon-anticodon decoding

Dissertation
Zur Erlangung des Doktorgrades der Naturwissenschaften am Fachbereich Chemie der Fakultät für Mathematik, Informatik und Naturwissenschaften der Universität Hamburg

vorgelegt von
Andriy Kazantsev

Fachbereich Chemie, Universität Hamburg

2020

Thesis assessors

Prof. Dr. Zoya Ignatova
Institute of Biochemistry and Molecular Biology, University of Hamburg

Prof. Karissa Sanbonmatsu
Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, USA

Examination commission

Prof. Dr. Zoya Ignatova
Institute of Biochemistry and Molecular Biology, University of Hamburg

Prof. Dr. Daniel Wilson
Institute of Biochemistry and Molecular Biology, University of Hamburg

Prof. Dr. Andrew Torda
Center for Bioinformatics, University of Hamburg

Disputation date: May 7, 2021

Approved to publish: May 7, 2021

Abstract

Accurate recognition of base pairs in the processes of Central Dogma is the basis of faithful replication and expression of genetic information. Among the possible sources of errors in this processes, G∘U mismatch recognition during codon-anticodon decoding in translation has the highest error rate. This has been linked to the occurrences of rare enol tautomers of nucleobases, which enable formation of the Watson-Crick (WC) geometry of this mismatch from the wobble (wb) geometry formed with canonical keto tautomers. The WC geometry of G∘U was observed in the environment of the closed ribosomal decoding site in equilibrium structural studies. This observation currently lacks a physicochemical explanation as well as a consistent model to reconcile it with the mechanism of decoding.

To address this problem, we studied effects of the decoding site on the wobble↔WC tautomerization reaction in G∘U (wb-WC reaction). Using quantum-mechanical/-molecular-mechanical umbrella sampling simulations, we found this reaction to be exoergic in the closed state of the decoding site, but endoergic in the open state. We also calculated the dielectric constant of the decoding site and revealed its decrease in the closed state. Together, these results provide an explanation to the structural observations.

To reconcile the stabilization of the WC geometry with the mechanism of base pair recognition in codon-anticodon decoding, we developed a new kinetic model of decoding that incorporates the wb-WC reaction parameters in the open and closed states of the decoding site. In this model, the exoergic wb-WC reaction is kinetically restricted by the decoding rates. This model explains the observations of the WC geometry at equilibrium conditions, thereby uniting structural and kinetic data. Moreover, the model reveals constraints imposed by the exoergic wb-WC reaction on the decoding accuracy: equilibration of the reaction counteracts equilibration of the open-closed transition. Our model can be a step towards a general recognition model for flexible substrates.

We applied this model, supported by additional computational studies, to provide a putative mechanism of how a specific U modification in anticodon can facilitate decoding of both A- and G-ending codons.

Zusammenfassung

Die präzise Erkennung von Basenpaaren in den Prozessen des zentralen Dogmas ist die Grundlage für eine fehlerfreie Replikation und Expression genetischer Informationen. Unter den möglichen Fehlerquellen in diesen Prozessen weist die G∘U-Basenfehlpaarung während der Codon-Anticodon-Decodierung bei der Translation die höchste Fehlerrate auf. Dies wurde mit dem Auftreten seltener Enol-Tautomere von Nukleobasen in Verbindung gebracht, die die Bildung der Watson-Crick (WC) Geometrie dieser Fehlpaarung aus der durch kanonische Keto-Tautomere gebildeten Wobble (wb) Geometrie ermöglichen. Die WC-Geometrie von G∘U wurde in Gleichgewichtsstrukturstudien in der geschlossenen ribosomalen Decodierungsstelle beobachtet. Dieser Beobachtung fehlt derzeit eine physikochemische Erklärung sowie ein konsistentes Modell, um sie mit dem Mechanismus der Decodierung in Einklang zu bringen.

Um dieses Problem anzugehen, untersuchten wir die Auswirkungen der Decodierungsstelle auf die Wobble-WC-Tautomerisierungsreaktion in G∘U (wb-WC-Reaktion). Unter Verwendung quantenmechanischer/molekularmechanischer Regenschirm-Probenahmesimulationen fanden wir, dass diese Reaktion im geschlossenen Zustand der Decodierungsstelle exoergische ist, im offenen Zustand jedoch endoergische ist. Wir haben auch die Dielektrizitätskonstante der Decodierungsstelle berechnet und eine Abnahme dieser im geschlossenen Zustand festgestellt. Zusammen liefern diese Ergebnisse eine Erklärung für die strukturellen Beobachtungen.

Um die Stabilisierung der WC-Geometrie mit dem Mechanismus der Basenpaarerkennung bei der Codon-Anticodon-Decodierung in Einklang zu bringen, haben wir ein neues kinetisches Decodierungsmodell entwickelt, das die wb-WC-Reaktionsparameter im offenen und geschlossenen Zustand der A-Stelle berücksichtigt. In diesem Modell ist die exoergische wb-WC-Reaktion durch die Decodierungsraten kinetisch begrenzt. Dieses Modell erklärt die Beobachtungen der WC-Geometrie unter Gleichgewichtsbedingungen und vereint so strukturelle und kinetische Daten. Darüber hinaus zeigt das Modell Einschränkungen in der Decodierungsgenauigkeit, die durch die exergone wb-WC-Reaktion auferlegt werden: das Gleichgewicht der Reaktion wirkt dem Gleichgewicht des offen-geschlossen-Übergangs entgegen. Unser Modell kann ein Schritt in Richtung eines allgemeinen Erkennungsmodells für flexible Substrate sein.

Wir haben dieses Modell, das durch zusätzliche Computerstudien unterstützt wird, angewendet, um einen Mechanismus dafür vorschlagen, wie eine U-Modifikation im Anticodon die Decodierung von Codons, die mit A und G enden, erleichtern kann.

List of Contents

List of Figures

List of Tables

List of Abbreviations

CV Collective variable. 38

DFT Density Functional Theory. 31

ESP Electrostatic potential. 85

MD Molecular dynamics. 37

MM Molecular mechanics. 33

N* Rare tautomer of a nucleobase N. 14

PES Potential energy surface. 26

PM3 Parametric model 3 (6,7). 31

PMF Potential of mean force. 38

PT Proton transfer. 91

SE Semiempirical methods. 30

TS Transition state. 16

US Umbrella sampling. 39

WHAM Weighted histogram analysis method. 39

Chapter 1
Introduction

1.1 Substrate recognition

Selective binding of ligands to enzymes (or ribozymes) is one of the most fundamental properties of living systems, as it ensures enzymatic catalysis and biochemical regulation in general. In the presence of multiple ligands with similar molecular structures, cellular machinery must evolve mechanisms that provide selective binding and processing of the correct (cognate) substrate and reject all other (non-cognate and near-cognate) substrates. Developing accurate models of these mechanisms is essential for understanding any biochemical process involving substrate recognition, including DNA replication and translation. Potential constraints imposed by the physical properties of substrates may hinder the intrinsic recognition capabilities of enzymes and ribozymes, and must be included in such models.

1.1.1 General models of substrate recognition

The development of substrate recognition models reflected the increasing appreciation of the flexibility in enzyme–substrate binding. In the earliest "lock-and-key" model, the specificity of enzymes was attributed to steric complementarity with their cognate substrates (Fischer, 1894). In this model, both binding partners are treated as rigid bodies (Fig. 1.1, left). Substrates bind to an enzyme with an "intrinsic" affinity, which is determined by the relative free energy of the rigid tightly bound complex.

Later, inconsistencies of the lock-and-key model in explaining the selectivity of some enzymes was noticed by Koshland (Koshland, 1959). Specifically, the lock-and-key model failed to explain the absence of ATP hydrolysis by hexokinase, the cognate substrate of which is glucose. Hydroxyl group in water can be as good nucleophile as the hydroxyl groups in glucose, but water does not hydrolyse ATP in the hexokinase active site (Koshland, 1995). Also, the mechanism of action of noncompetitive inhibitors could not be reconciled with the rigid-enzyme view in the lock-and-key model (Koshland, 1995). These inconsistencies of the lock-and-key model led Koshland to propose the "induced-fit" model (Koshland, 1959), in which an enzyme has flexibility. This model posits that the initial state of an enzyme has relatively low affinity to substrates ("open" state). A cognate substrate, but not (or to a less extent) a near-cognate substrate, induces conformational changes in a protein, leading to a catalytically competent and high-affinity (key-lock or "closed" state) arrangement in the protein's active site (Koshland, 1959) (Fig. 1.1, right). In the later "conformational selection", or "pre-equilibrium" model, the closed conformational states of an enzyme exist even in the absence of a cognate substrate, and are only stabilized by its binding (Burgen, 1981, Changeux and Edelstein, 2011). Although induced-fit and conformational selection models differ in the proposed "origin" of flexibility (Changeux and Edelstein, 2011, Gianni et al., 2014), their explanation of substrate recognition accuracy is the same, therefore here we consider these two models as the same model and call it "induced-fit" for convenience. The induced-fit model highlighted the abundance and essential role of enzyme flexibility. Although the physical mechanisms of how protein dynamics contributes to their function are yet to transition from theoretical proposals to experimentally confirmed models (Kamerlin and Warshel, 2010), enough evidence already suggests the critical role of protein dynamics in catalysis (Bhabha et al., 2011, Daniel et al., 2003), and some conformational modes seem to be evolutionary selected (Ramanathan and Agarwal, 2011).

However, in the context of substrate recognition accuracy, the function of enzyme/ribozyme flexibility within the induced-fit model remains less clear. The accuracy is defined as the ratio of probabilities of cognate vs near-cognate product formation. It has been shown theoretically that induced-fit does not provide any additional accuracy over rigid-body recognition (complementarity) at equilibrium conditions (Fersht, 1974). For a substrate recognition process with multiple reversible steps (i.e. conformational transitions of an enzyme) and a single irreversible transition (i.e. catalytic reaction), its "intrinsic" accuracy is determined by relative (cognate vs near-cognate) free energy of the pre-catalytic (closed) state (Fersht, 1974, Herschlag, 1988, Pavlov and Ehrenberg, 2018). Therefore, the induced-fit mechanism approaches its maximal

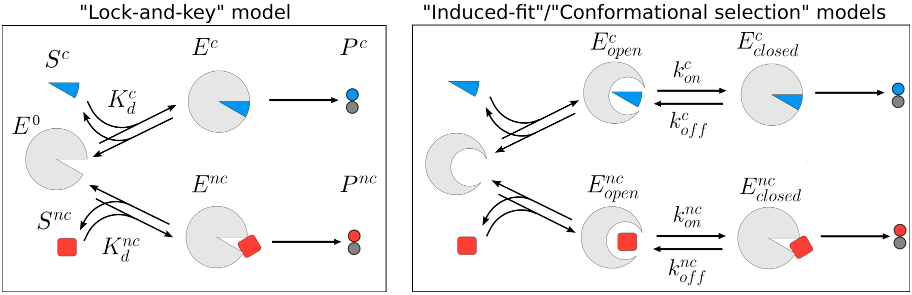


Figure 1.1.: General models of substrate recognition. **Left**, the "Lock-and-key" model. Free rigid enzyme E^0 binds to rigid substrates S^c (cognate) and S^{nc} (near-cognate) with different equilibrium dissociation constants K_d^i due to steric complementarity between the enzyme and the cognate substrate. **Right**, the "Induced-fit" or "Conformational selection" models. The free enzyme has low affinity to both S^c and S^{nc}, but the cognate substrate induces/stabilizes the high affinity state of the enzyme via a conformational (open↔closed) transition in the enzyme. The rate constants of the open↔closed transition (k^i_{on}, k^i_{off}) affect the apparent specificity of the enzyme.

accuracy at equilibrium conditions (Fersht, 1974, Herschlag, 1988, Hopefield, 1974, Pavlov and Ehrenberg, 2018), resembling the lock-and-key model, and any deviation from equilibrium reduces the accuracy. Experimental evidences in some systems support this conclusion (Johansson et al., 2012, Thompson and Karim, 1982). Any theoretical proposal on the possible advantage of out-of-equilibrium conformational transitions in substrate recognition involves additional assumptions, such as noise in substrate concentration (Herschlag, 1988, Savir and Tiusty, 2007) or trade-off between speed and accuracy (Savir and Tlusty, 2013). While the rate-accuracy trade-off is in a general a well-established model with experimental support in some systems (Johansson et al., 2012, Thompson and Karim, 1982), accuracy-limiting conformational steps can be not rate-limiting, which requires alternative explanations for their potential evolutionary advantages or constraints otherwise.

Irreversible steps during substrate recognition can increase accuracy over the equilibrium limit. This model was proposed by Hopefield and is called kinetic proof-reading (Hopefield, 1974). In the presence of kinetic proof-reading, the total substrate recognition accuracy is a product of accuracies at each step, separated by irreversible reactions (Hopefield, 1974, Mallory et al., 2019). Kinetic proof-reading only provides a mechanism of accuracy amplification in the presence of multiple intermediate irreversible steps. It cannot explain the mechanisms governing substrate recognition at each irreversible step.

The induced-fit model only includes the enzyme flexibility and continues to assume a rigid substrate. As will be discussed in the next paragraphs and demonstrated in Chapter 4, this approximation precludes the applicability of the induced-fit model to systems with intrinsically flexible substrates.

1.1.2 Base pair recognition

Among molecular recognition processes in biology, the accuracy of processes comprising the Central Dogma – DNA replication, transcription, and translation – arguably has the largest impact on living systems. Besides directly affecting the fitness of every living system (Ke et al., 2017, Santos et al., 2018, Wohlgemuth et al., 2011), recognition errors in these processes (point mutations and amino acid misincorporations) may comprise the evolutionary drive (Goldenfeld and Woese, 2011, Hershberg, 2015), thus also contributing to the rate and mechanism of evolution. Therefore, studying the biophysical mechanisms behind these errors is essential for advancing our understanding of fundamental biological processes.

The final product of gene expression is a protein, therefore the final error rate of gene expression is amino acid misincorporation rate in proteins. This error rate can be affected by various processes along the route of gene expression: DNA replication, modification and repair; mRNA synthesis (transcription) and its post-transcriptional processing; aminoacylation of tRNA by aminoacyl-tRNA synthetases; and ultimately, codon-anticodon decoding in translation. Studies employing different experimental methods conclude that codon-anticodon decoding has the highest error rate among these steps, thereby limiting the final error rate of gene expression (Garofalo et al., 2019, Kramer and Farabaugh, 2007, Mordret et al., 2019).

In translation, the ribosome selects aminoacylated tRNAs in complexes with elongation factors (EF-Tu in bacteria) – ternary complexes (TCs). Although this process may seem drastically different from NTP/dNTP selection by RNA/DNA polymerases, both processes ultimately depend on base pair recognition. In codon-anticodon decoding, given equal or cellular tRNA concentrations, the primary determinant of error rate between different codon-anticodon combinations is the type of mismatch, although other factors (e.g., tRNA identity, sequence context) also contribute to the error rate variance (Garofalo et al., 2019, Manickam et al., 2014). Various in vitro (Pernod et al., 2020, Zhang et al., 2015) and *in vivo* (Garofalo et al., 2019, Kramer and Farabaugh, 2007, Manickam et al., 2014, Mordret et al., 2019, Zhang et al., 2013) studies demonstrate G∘U/U∘G (the orientation is ignored below unless specified otherwise) mismatch in the first two codon positions as the major error-hotspot in codon-anticodon decoding, with error rate of 10^{-3}–10^{-5}. Some studies also identify U∘U as an error-hotposot (Manickam et al., 2014, Zhang et al., 2013), but G∘U mismatch is the only error hotspot consistently appearing across all studies. Therefore, G∘U mismatch in codon-anticodon decoding largely determines the final error rate of gene expression.

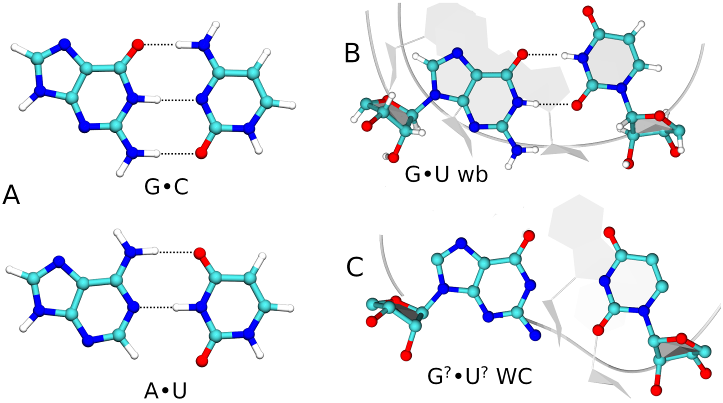


Figure 1.2.: Base pair geometries. **A** – canonical (complementary) base pairs adopt the WC geometry. **B** – G∘U mismatch in RNA duplex adopts the wb geometry (PDB ID: 4PCO). **C** – G∘U mismatch in the middle position of codon-anticodon helix in the closed decoding of site of the ribosome adopts the WC geometry (PDB ID: 6GSK). Unknown configurations of nucleobases are highlighted with $^?$.

To understand the origin of this error, one must have a model of base pair recognition in general. Complementary base pairs are characterised by a conserved base pair geometry – the Watson-Crick geometry (WC) (Fig. 1.2A). Based on this isostericity, complementary base pairs are thought to be recognized with approximately equal efficiency by binding pockets of DNA/RNA polymerases and ribosome, allowing sequence-independent selectivity – the basis of genetic information replication and expression. A model where base pairs are recognized by their shapes ("geometric selection"), instead of base identity or H-bonding, has found strong support in experiments (Goodman, 1997, Khade et al., 2013, Kool, 2002, Lee and Berdis, 2010, Moran et al., 1997, Westhof et al., 2014). In this model, the discrimination capacity of base pair recognition relies on the inability of mismatches to form the WC geometry. Instead, in their predominant forms in physiological environment, they adopt various geometries which are not isosteric to WC (Leontis et al., 2002, Westhof, 2014, Westhof et al., 2014). Particularly, G∘U(T) mismatch in DNA or RNA duplexes in water solution predominantly adopts the wobble (wb) geometry, in which U is shifted towards the major groove (Fig. 1.2-B). The wb geometry of G∘U(T) has relatively high thermodynamic stability, comparable or in particular instances even exceeding the thermodynamic stability of the canonical A∘U(T) base pair (Gu et al., 2015, Mizuno and Sundaralingam, 1978, Varani and McClain, 2000). G∘U(T) wb also has functional roles in the cell, and can be specifically recognized by proteins (Varani and McClain, 2000, Westhof et al., 2019). However, because the wb geometry is not isosteric to the WC geometry, the ribosome and DNA/RNA polymerases are able to discriminate against the G∘U(T) mismatches, and reject them with varying efficiencies during codon-anticodon decoding and replication/transcription (Westhof et al., 2014). If the G∘U(T) could change from the wb to the WC geometry, it would become effectively indistinguishable from the canonical base pairs for the ribosome and polymerases (Westhof et al., 2014).

1.1.3 Codon-anticodon decoding

Since the fundamentals of tRNA selection are shared between prokaryotic and eukaryotic translation, and most structural and kinetic studies are obtained for the prokaryotic ribosome (Rodnina et al., 2017), here we discuss tRNA selection in bacteria. In the classical scheme, tRNA selection in translation is divided into initial selection and proof-reading (Fig. 1.3). In the initial selection, TCs bind to the empty A-site leading to codon-anticodon pairing. Cognate pairing promotes GTPase activation via interaction of EF-Tu•GTP complex with sarcin-ricin loop (SRL), leading to fast irreversible GPTase hydrolysis. Near-cognate (only 1 mismatch) codon-anticodon pairs result in a greatly reduced GTPase activation rate and non-cognate (> 1 mismatch) pairs are mostly rejected at the TCs binding step (Gromadski et al., 2006). This comprises the discrimination basis of the initial selection. The irreversible GTP hydrolysis creates the necessary condition for kinetic proof-reading. In the ternary complex, tRNA remains in a strained conformation (A*/T state) (Loveland et al., 2017, Rodnina et al., 2017). During the proof-reading, the release of EF-Tu•GDP and cognate codon-anticodon pairing promote tRNA accommodation and rotation, eventually leading to A/P state of the tRNA, which is ready for irreversible peptidyl transfer reaction (Loveland et al., 2020) (Fig. 1.3). Near-cognate pairs which were not discriminated in the initial selection are thought to be more likely to be rejected than accommodated, which comprises the discrimination basis of proof-reading (Gromadski et al., 2006, Loveland et al., 2020).

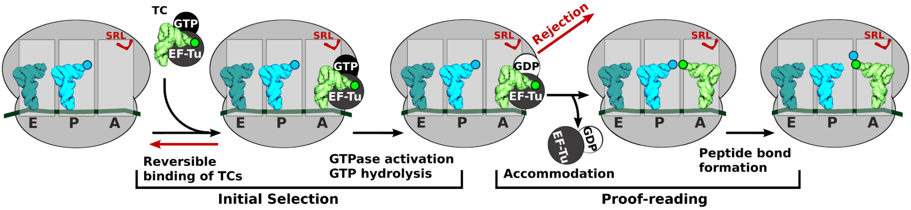


Figure 1.3.: tRNA selection in translation, consisting of the initial selection and proof-reading. See the text for explanations

In this study, we focus on the initial selection due to the following reasons: (i) recent Cryo-EM studies suggest that both proof-reading and initial selection employ similar molecular mechanisms of discrimination – the open-closed transition of the decoding site (Loveland et al., 2020) – thus implying that studying this mechanism is needed to understand both recognition steps; (ii) the molecular recognition mechanism is studied in greater detail in the initial selection compared to the proof-reading, and most importantly has much better kinetic resolution in the initial selection; this would allow us to build our model as extension to the already developed models; (iii) lastly, proof-reading is yet to be directly demonstrated for G∘U mismatches in high-fidelity conditions (see Section 4.3.3).

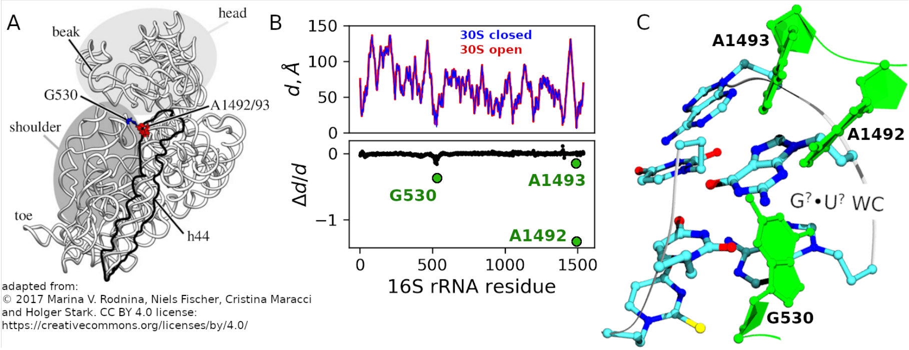
Since the early kinetic experiments (Pape et al., 1999) as well as structural studies of the ribosome fragments (Ogle et al., 2002), the initial selection has been considered to follow the induced-fit mechanism. The molecular mechanism of the induced-fit process in decoding is the conformational change in the 30S subunit – 30S closure (Ogle et al., 2002), which has been now described in great detail using structural, kinetic and computational methods. In the absence of TC in the decoding site, the 30S subunit adopts the open conformation (Loveland et al., 2017, Ogle et al., 2002). The ribosomal decoding site is comprised mostly of 16S rRNA residues of the 30S subunit (Ogle et al., 2001, 2002) (Fig. 1.4A). Ribosomal protein S12 also approaches the codon-anticodon helix, but is thought to only indirectly interact with the third position (Ogle et al., 2001). Although not involved in the codon-anticodon recognition directly, ribosomal proteins S4, S5 and S12 are important for the open-closed transition, and their mutations affect this conformational change and the consequent accuracy of decoding (Agarwal et al., 2015, Hoffer et al., 2019, Zaher and Green, 2010). The open-closed transition induced by the incoming TC involves global movements of 16S domains, but the majority of these movements are quite distant from the codon-anticodon helix (Fischer et al., 2016, Loveland et al., 2017) (Fig. 1.4B). In fact, when normalizing the closed-open distance difference to the distance in the closed state, only three rRNA residues are clearly seen to drastically change its position upon the domain closure: G530, A1492 and A1493 (Loveland et al., 2017) (Fig. 1.4B). G530 changes its conformation from *syn* to *anti* and approaches the codon-anticodon helix near the second and third position (Fig. 1.4C); A1492 and A1493 flip from h44 helix, where they are intercalated in the open 30S conformation (*off* state), to the *on* state, in which they form A-minor interactions with the first and second codon position (Demeshkina et al., 2012, Loveland et al., 2017, Ogle et al., 2002). While in the Cryo-EM structures of Loveland et al. (2017) A1493 was in the *on* state even in the open 30S conformation, previous studies showed its conformational change during the transition (Fischer et al., 2016, Rodnina et al., 2017).

rRNA residues G530, A1492, and A1493 are conserved and are believed to monitor the base pair geometry of the first two codon-anticodon base pairs (Ogle et al., 2001). Earlier studies suggested a critical role of H-bonds these rRNA residue form with the codon-anticodon helix (Ogle et al., 2001). However, later studies disproved this model (Khade et al., 2013, Schrode et al., 2017). Studies employing fluorescence spectroscopy and umbrella sampling calculations suggested that A1492 and A1493 flipping is exoergic in the presence of the cognate TC ($\Delta G = -2$ kcal/mol) and endoergic in the presence of near-cognate TC ($\Delta G = 1$ kcal/mol) (Zeng et al., 2014). It has been suggested that the origin of this specificity is desolvation of the codon-anticodon helix, which is supposed to penalize the mismatches (Ogle et al., 2002, Satpati and Åqvist, 2014). It is clear that open and closed states of the 30S correspond to the low-affinity and high-affinity states of the general induced-fit model. The binding affinity ratio of cognate and near-cognate TC to the closed decoding site corresponds to the "intrinsic" selectivity of decoding (Pavlov and Ehrenberg, 2018).

Besides the structural studies, essential insights into the mechanism of decoding are also provided by kinetic studies. Single-molecule kinetic experiments mainly by Rodnina and coworkers allow to build a kinetic model of initial selection (Gromadski et al., 2006, Gromadski and Rodnina, 2004, Rodnina et al., 2017) (Fig. 1.5). Kinetic studies reveal discrete decoding states, which at least partly correspond to the structural states observed in the Cryo-EM studies (Fislage et al., 2018, Rodnina et al., 2017). In accordance with previous studies (Fislage et al., 2018, Pavlov and Ehrenberg, 2018), here we use $C2 \to C3 \to C4$ nomenclature of the decoding states, which correspond to the empty, open and closed states of the decoding site (Fig. 1.5). Rate constants are available for near-cognate (nc) and cognate (c) fluxes, which allows calculating the error rate of initial selection as their ratio using the classical Michaelis-Menten kinetics:

$$\eta_0 = \frac{R^{nc}}{R^c} = \frac{(k_{cat}/K_m)^{nc}}{(k_{cat}/K_m)^c} = \frac{k_4^{nc}[C4_{nc}]}{k_4^c[C4_c]} \tag{1.1}$$

where R^i is the rate of decoding, k^i_{cat} is the catalytic rate constant, K^i_m is the Michaelis-Menten constant, $[C4_i]$ is the steady-state concentration of $C4$ state, and k^i_4 is the rate constant of GTPase activation, for $i = c$, nc. In accordance with the induced-fit model,



adapted from:
© 2017 Marina V. Rodnina, Niels Fischer, Cristina Maracci and Holger Stark. CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Figure 1.4.: Open↔closed transition of the decoding site. **A** – scheme of the 16S rRNA and its domains, adapted from Rodnina et al. (2017) under the Creative Commons license. The codon-anticodon helix is in the middle (not shown). **B**, top – distance of 16S rRNA residues to the middle codon nucleobase (*d*) calculated from Cryo-EM structures of the open (PDB id 5UYN) and closed decoding site (5UYN) (Loveland et al., 2017). *d* was calculated as center of mass distances between the nucleobase atoms. **B**, bottom – $\Delta d = d(closed) - d(open)$, normalized by $d(closed)$. This property allows to visualize 16S rRNA residues that change the most upon the transition to the closed state: G530, A1492 and A1493. **C** – visualization of the codon-anticodon helix surrounded by these residues in the closed 30S conformation.

kinetic measurements revealed increased forward rate constants and decreased reverse rate constants in the cognate branch compared to the near-cognate branch (Gromadski and Rodnina, 2004, Rodnina et al., 2017) (Fig. 1.5). Some rate constants, such as k_1 and q_2 (initial binding of TC), and k_3 (forward rate constant of the open↔closed transition) are the same between *c* and *nc* paths (Rodnina et al., 2017). By analyzing η_0 as a function of decoding rate constants, tRNA concentrations and other cellular properties, one can reveal potential mechanisms of the evolutionary optimization of decoding (Wohlgemuth et al., 2011).

Although the model of decoding in translation nowadays includes detailed information from multiple experimental and computational sources, it also has certain inconsistencies related to the error hotspot of decoding – G∘U mismatch. It is possible that these inconsistencies originate from the core of the decoding model – the induced-fit mechanism.

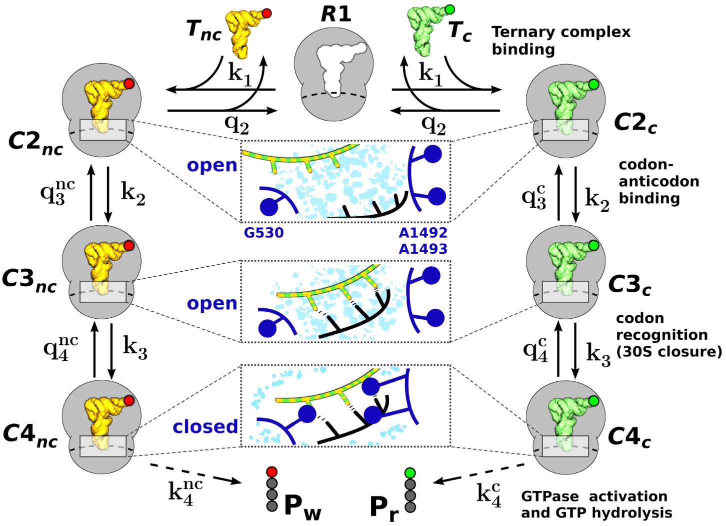


Figure 1.5.: Kinetic scheme of initial selection. The nomenclature of the states and rate constants is taken from Pavlov and Ehrenberg (2018). The collection of the decoding states is loosely supported by cryo-EM studies (Loveland et al., 2017). The schemes of the enlarged decoding site in the middle depict the state of the codon-anticodon helix and conserved 16S rRNA residues during the decoding. For visualization purposes, only the A-site tRNA is shown in the scheme. The steps following the initial selection are not considered, therefore the GTPase activation step leads directly to the peptides. $R1$ represents mRNA-programmed ribosomes with empty A-site. T_c and T_{nc} represent cognate and near-cognate ternary complexes, respectively. In $C2$ state the codon-anticodon helix is not yet formed, and the decoding site is open. In $C3$ the codon-anticodon helix forms in the open decoding site. In $C4$ the decoding site is in the closed state.

1.1.4 Challenges of the induced-fit model in base pair recognition

First indications of the inability of rigid-substrate models, such as induced-fit, to explain the mechanism of base pair recognition came from structural studies. One of the first X-ray studies of the ribosomal 30S subunit revealed the WC geometry of the G∘U mismatch at the first and third codon positions (Ogle et al., 2001, 2002). Later, G∘T and A∘C in the WC geometries were observed in the closed states of DNA polymerase active sites (Bebenek et al., 2011, Koag et al., 2014, Wang et al., 2011). At the same time, G∘U mismatches in the WC geometry were observed at the first and second codon positions in the closed decoding site of the 70S ribosomal X-ray structure (Demeshkina et al., 2012, 2013, Rozov et al., 2015, 2018) (Fig. 1.2C). Finally, recent Cryo-EM studies of the 70S ribosome also revealed WC geometries of G∘U (Loveland et al., 2017) and A∘C (Fislage et al., 2018) in the closed decoding sites.

It is worth reminding here that G∘U or G∘T in RNA or DNA duplexes in solution at physiological pH predominantly adopts the wobble geometry (Kimsey et al., 2015) (Fig. 1.2B). Therefore, the environments of the closed active/decoding sites *induce* the WC geometry of G∘U(T) mismatch, which would change substrate specificity *during* the substrate recognition. Such a process is inherently unexplainable by the induced-fit model – it only considers induced changes in the enzyme, not the substrate. No matter what are the physical origins of this phenomenon, these observations challenge the validity of the induced-fit model in base pair recognition.

Theoretical predictions based on the rigid substrate approximation challenge this model through another angle. As described in Section 1.1.1, induced-fit mechanism approaches the maximal accuracy (minimal η_0) at equilibrium conditions. Within the kinetic model of initial selection (Fig. 1.5), it means that high forward rates (k_1, k_2, k_3 and k_4^c) would increase η_0. Of particular importance is k_4^c, the irreversible[1] GTPase activation reaction, which affects deviation of the open↔closed transition from equilibrium, and thus the actual from the intrinsic accuracy (Pavlov and Ehrenberg, 2018, Wohlgemuth et al., 2011) (Fig. 1.6). The first experimental measurements of the full set of decoding rate constants and solutions of the Michaelis-Menten kinetic model based on it revealed an unexpectedly high k_4^c ($\sim 190\text{s}^{-1}$), much higher than the total rate of elongation $R_{elo} \approx 0.8$ s^{-1} (Gromadski and Rodnina, 2004) (Fig. 1.6). It indicates that k_4^c is increased by the expense of decoding accuracy (Johansson et al., 2008). To explain this, it was proposed that high k_4^c serves as a "buffer" against tRNA competition *in vivo* (Gromadski and Rodnina, 2004, Wohlgemuth et al., 2011). Later, theoretical studies employing stochastic frameworks concluded that decoding in trans-

[1] The actual irreversible reaction is GTP hydrolysis, but k_4^c is much slower than it, and thus rate-limiting

lation is optimized for higher speed at the expense of accuracy (Banerjee et al., 2017, Mallory et al., 2019). Although using sophisticated approaches from the arsenal of theoretical physics, these studies still assumed rigid substrates. Another theoretical study employed fitness functions to explain the apparent symmetry between forward cognate and reverse *nc* rates (see Fig. 1.6, where $k_4^c \approx q_4^{nc}$) (Savir and Tlusty, 2013). In all of these theoretical studies trade-off between speed and accuracy was assumed. Since k_4^c is not rate-limiting in translation elongation (Wohlgemuth et al., 2011), its high value remains unexplained.

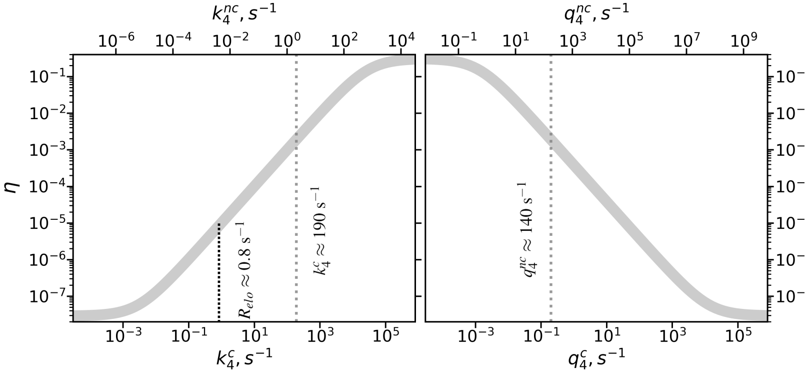


Figure 1.6.: Induced-fit predicts suboptimal initial selection mechanism. The curves are obtained as solutions to kinetic model on Fig. 1.5 using Eq. (1.1). For each k_4^i and q_4^i, η_0 was calculated by varying cognate and near-cognate rate constants simultaneously with their constant ratio. Gray vertical lines denote rate constants measured at 20 °C (Rudorf et al., 2014). R_{elo} – the total rate of translation elongation measured under the same conditions (Rudorf et al., 2014).

As will be described in the next section, structural challenges can be potentially resolved by nucleobase tautomerism. As we show in Chapter 4, addressing the problem of tautomerism can also potentially solve the problem of seemingly suboptimal decoding.

1.2 Tautomerism in base pair recognition

Base pair geometry is mostly determined by the pattern of hydrogen-bonding donor and acceptor groups, which are in turn determined by the protonation states of nucleobases. The protonation state of heterocycles can be changed via ionization (protonation/deprotonation), which is directly dependent on the pH of the solution (Kimsey et al., 2015), or via prototropic tautomerism, which is less (indirectly) dependent on pH (Kimsey et al., 2015). The ground tautomers of G and U, which enable the wobble base pairing, are keto tautomers, in which ring nitrogen atoms N1 (in guanine) and N3 (in uracil/thymine) bind imino protons, and the exocyclic oxygen atoms exist in the keto groups (Fig. 1.7) (Kimsey and Al-Hashimi, 2014, Singh et al., 2015). A proton shift from the ring imino groups to the exocyclic keto groups, forming the carbonyl groups, produces enol tautomers G* and U* (here the asterisk N* denotes rare tautomers of nucleobase N) and is defined as keto/enol tautomeism of these nucleobases. G*∘U(T) and G∘U*(T*) configurations adopt the WC geometry (Fig. 1.7). Similarly, amino protons of exocyclic amino groups of A and C can shift to ring imino groups producing imino tautomers A* and C*, which is defined as amino-imino tautomerism of these nucleobases. Another possibility for G and U to form the WC geometry is the deprotonation of U at N3. However, at neutral pH and upon unmodified nucleobases, this state has a much lower population compared to the tautomeric state (Kimsey et al., 2015, 2018). This path will be discussed only in the context of tRNA modifications (see below).

1.2.1 History of the tautomeric hypothesis

In their seminal paper following the report of DNA double helix, Watson and Crick proposed the tautomeric hypothesis, which suggests that a source of errors during base pair recognition is keto-enol and amino-imino tautomerism of nucleobases (Watson and Crick, 1953a,b). This hypothesis followed from the base pairing rules discovered by Watson and Crick, and from the suggestion that nucleobases must have predominant tautomeric forms, but can occasionally adopt alternative tautomers. The tautomeric hypothesis suggested only a general mechanism of tautomerism as a source of base pair recognition errors. It did not answer the questions of how tautomers are formed in biologically relevant processes (i.e., tautomerization reactions); how these processes are affected by enzymes (and ribozymes) that perform base pair recognition; how the processes of tautomerization reconcile with models of substrate recognition. Although alternative tautomers of nucleobases were detected experimentally (Kimsey and Al-Hashimi, 2014, Singh et al., 2015), tautomerization reactions have remained unknown and inaccessible by experiments. Therefore, the studies which attempted to develop rigorous models based on the tautomeric hypothesis employed almost exclusively computational and theoretical methods.

The common theme of previous studies addressing tautomerization is the incredibly slow rates of unimolecular tautomerization reactions in nucleobases, which implies that additional/alternative mechanisms must occur to facilitate these reactions to biologically relevant rates. The first proposal to provide such facilitation was the double proton transfer model (DPT) in canonical A∘T and G∘C base pairs by Löwdin (1963). In the DPT model, intermolecular proton transfers spontaneously happen in the canonical base pairs along the H-bonds, leading to the formation of enol and imino tautomers of nucleobases. In the original formulation the emphasis was on quantum tunneling in these reactions (explained in Section 2.1.1), but the nuclear quantum effects are not essential for the DPT model. Computational studies following Löwdin's proposal addressed the relevance of this model in different base pairs. It was shown that while the DPT mechanism in G∘C can indeed lead to stable yet short-lived G*∘C* configuration of the base pair (Cerón-Carrasco et al., 2009, Florian and Leszczynski, 1996, Zoete and Meuwly, 2004), DPT in A∘T cannot produce stable A*∘T* configuration (Florián et al., 1994). DPT mechanism was also studied in G∘T wobble base pair (Padermshoke et al., 2008). Besides the DPT model, computational studies also addressed facilitation of tautomerization by water molecules and other intermolecular interactions (Hu et al., 2004, Jacquemin et al., 2014, Li and Ai, 2009).

Even if these models of rare tautomers formation would work and produce A*, G*, C*, and T*(U*), additional models are needed to explain the processes leading to the errors of base pair recognition induced by these tautomers. In the proposal by Löwdin (1963), rare tautomers formed via DPT in the canonical base pairs have to first dissociate from the base pairs, remain in these tautomeric configurations until they enter the active site of DNA polymerases, and then, by binding to the corresponding mismatched nucleobases (A*∘C/A∘C* and G*∘T/G∘T*) will eventually lead to a point mutation. Similarly, in a model proposed by Topal and Fresco (Topal and Fresco, 1976a,b), it is assumed that tautomerization happens only in aqueous environment, and when rare tautomers of nucleobases enter water-excluded environment of DNA polymerases or ribosome, their tautomeric state is locked until the base pair recognition finishes, thus leading to errors.

Therefore, previous models of tautomerization-induced base pair recognition errors contained multiple steps and relied on unverified assumptions. Although interesting from a historical perspective, these models became irrelevant since the discovery and experimental confirmation of tautomerization reactions that happen directly in the mismatches.

1.2.2 wobble ↔ Watson-Crick tautomerization reaction

Using computational chemistry methods, Brovarets' and Hovorun predicted a tautomerization reaction in G∘U(T) base pair (Brovarets and Hovorun, 2009, 2015) (wb-WC reaction, Fig. 1.7). The wb-WC which proceeds from G∘U(T) in the wb geometry, via the ion-pair transition state (TS) to G∘U*(T*) in the WC geometry. G∘U*(T*) WC and G*∘U(T) WC exist in fast equilibrium via a DPT reaction. The wb-WC reaction does not require water or any other intermolecular interactions besides the mismatched base pair itself; the wb-WC reaction proceeds in the mismatch, thus not requiring any dissociation/association steps to cause recognition errors; any other possible tautomerization reactions (e.g., tautomerization of isolated nucleotides in water) would not affect the population of the WC geometry of the mismatch, because this population is ultimately dependent on the equilibrium and kinetic properties of the wb-WC reaction which can proceed *during* the base pair recognition. Although the wb-WC reaction is a tautomerization reaction, it is essentially a transition between the wobble and the WC geometries of the mismatch. Therefore, the wb-WC reaction shifts the focus from the relative energies of nucleobase tautomers to the relative energy of the wb and WC geometries in a direct chemical equilibrium. Considering all these advantages over the previous models, the wb-WC reaction is likely the key to unraveling the detailed mechanisms of the tautomerization-induced base pair recognition errors. The wb-WC reaction in G∘U(T) is predicted to be exoergic in gas phase, but endoergic in implicit water model (Table 1.1). This property of the wb-WC reaction would be critical in explaining its role in base pair recognition mechanisms, as demonstrated in the next chapters.

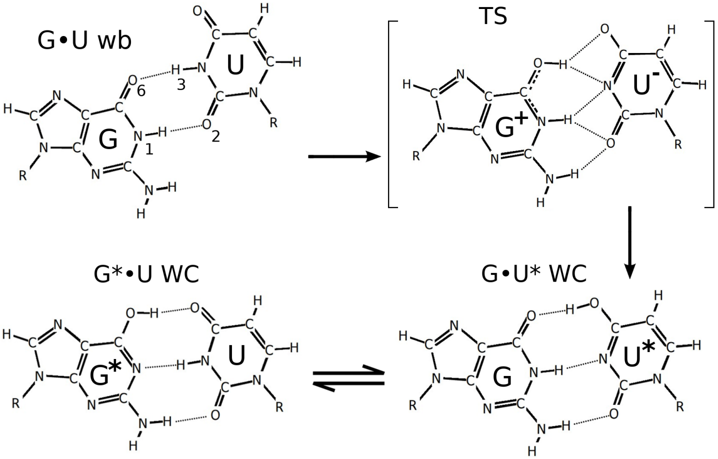


Figure 1.7.: wb-WC tautomerization reaction. Enol tautomers are denoted with asterisk (*). The wb-WC reaction consists of slow tautomerization reaction from G∘U wb to G∘U* WC via the ion-pair TS, followed by fast double proton transfer reaction between G∘U* and G*∘U.

It should be noted that Hovorun and coworkers predicted similar tautomerization reactions also in other mismatches. In A∘C mismatch, the predicted tautomerization reaction is also a wobble↔WC transition, and follows essentially the same mechanism as the wb-WC reaction in G∘U(T) (Brovarets and Hovorun, 2009, Brovarets' and Hovorun, 2015). Although the presence of highly similar tautomerization mechanisms in both R∘Y mismatches (R – purine, Y – pyrimidine) is intriguing from a theoretical viewpoint, the wb-WC reaction in A∘C may not be relevant for biological systems, as the reactant geometry in this reaction may not be the most abundant state for amino tautomers of A and C (Brovarets and Hovorun, 2009, Brovarets' and Hov-

Table 1.1.: Calculated and experimentally measured properties of the wb-WC reaction

Condition	ΔG, kcal/mol [a]	$\Delta G^{\ddagger}$, kcal/mol [b]
Calculated at M06/6-311++G(d,p) by Nomura et al. (2013):		
$\varepsilon = 1$ (vacuum)	-1	17.9
$\varepsilon = 80$ (water)	6.4	21.4
Measured with NMR by Kimsey et al. (2015):		
RNA/DNA in water	3.0 – 5.8	16.4

[a] Forward free energy change of the wb-WC reaction; [b] Activation free energy of the wb-WC reaction

orun, 2015, Leontis et al., 2002). Tautomerization reactions in Y∘Y mismatches can also lead to WC-like geometries (Brovarets' and Hovorun, 2015). The wb-WC reaction in G∘U(T) has an important advantage over similar tautomerization reactions in other mismatches – it has been experimentally confirmed, as described below. Therefore, below we discuss only the wb-WC reaction in G∘U(T) and refer to it simply as the wb-WC reaction.

The wb-WC reaction was confirmed with NMR in DNA and RNA duplexes in water solution (Kimsey et al., 2015), demonstrating agreement with the theoretically predicted properties of this reaction (Brovarets and Hovorun, 2009, 2015, Kimsey et al., 2015, Li et al., 2020, Nomura et al., 2013) (Table 1.1). Later, by incorporating the equilibrium WC G∘T populations measured with NMR in DNA duplexes in *water solution* into a numerical kinetic model of DNA replication, an excellent agreement between predicted and experimentally observed misincorporation rates was obtained, suggesting a minor role of polymerase environment on tautomerization (Kimsey et al., 2018).

1.2.3 The open question of tautomerism in base pair recognition

Observations of G∘U mismatches in the WC geometry in the closed decoding/active sites, discussed in Section 1.1.4, require explanations within the framework of nucleobase tautomerism. Authors of the structural studies reporting these observations in ribosome crystals proposed that these structures reveal high-energy states, and thus G∘U discrimination relies on "Energy expenditure for formation of tautomers" (Demeshkina et al., 2012, 2013, Rozov et al., 2015, 2018). It is hard to imagine how canonical X-ray diffraction, an equilibrium method, can capture high-energy (thus short-lived) states. Such an interpretation could still potentially work if the wb-WC reaction did not exist, i.e., if the enol tautomers were somehow selectively trapped *in vacuo* and could not convert back to keto states. However, in the presence of the wb-WC reaction, and at physiological pH, these observations can have only one reasonable interpretation: environment of the closed decoding site shifts equilibrium in the wb-WC reaction towards WC. Although seemingly obvious, this interpretation opposes the established model of tautomerism in translation (Pavlov et al., 2017, Rodnina et al., 2017, Rozov et al., 2018, Sanbonmatsu, 2014, Zeng et al., 2014). Therefore, studies pursuing the alternative interpretation must provide clear evidences in its support, in addition to reconciliation with the model of decoding. This challenge defines the direction of our study.

Furthermore, Kimsey et al. (2018) interpretation of their results contradicts experimental (Bebenek et al., 2011, Koag et al., 2014) and computational (Li et al., 2020, Maximoff et al., 2017) studies demonstrating stabilization of the WC geometry of G∘T in the closed active site of DNA polymerases. This indicates the presence of a more complicated mechanism governing base pair recognition and highlights the open question of the role of tautomerism in replication

Computational approaches provide a perfect opportunity to study tautomerization in the environment of the ribosomal decoding site. Until now, two computational studies have addressed this problem and concluded that the WC geometry of G∘U is not stabilized in the closed decoding site (Satpati and Åqvist, 2014, Zeng et al., 2014). Both studies used classical force fields. One of them used non-parameterized enol tautomers of G and U in their force field (Satpati and Åqvist, 2014) and therefore will not be discussed here. The other study obtained the enol parameters of U using automatic assignment (Zeng et al., 2014). Their approach largely overestimated aqueous keto/enol tautomerization energy in monomeric U compared to the experimental values (38 kcal/mol vs 10 kcal/mol) (Zeng et al., 2014), thus questioning the tautomerism-related conclusions of their study. These two studies highlight an issue discussed in the next chapter: classical force fields are not well suited to study chemical reactions.

1.3 tRNA modifications and ambiguous codon-anticodon decoding

1.3.1 Ambiguous codon-anticodon decoding

Interestingly, most organisms bend the codon-anticodon discrimination rules by having some tRNAs that decode multiple codons. This *ambiguous decoding* is restricted exclusively to the third codon-anticodon base pair – position 34 of tRNA, known as the wobble position. This observation led Crick to propose the wobble hypothesis, which suggests that a restricted set of noncanonical base pairs at the wobble position results in efficient translation (Crick, 1966):

> I now postulate that in the base-pairing of the third base of the codon there is a certain amount of play, or wobble, such that more than one position of pairing is possible

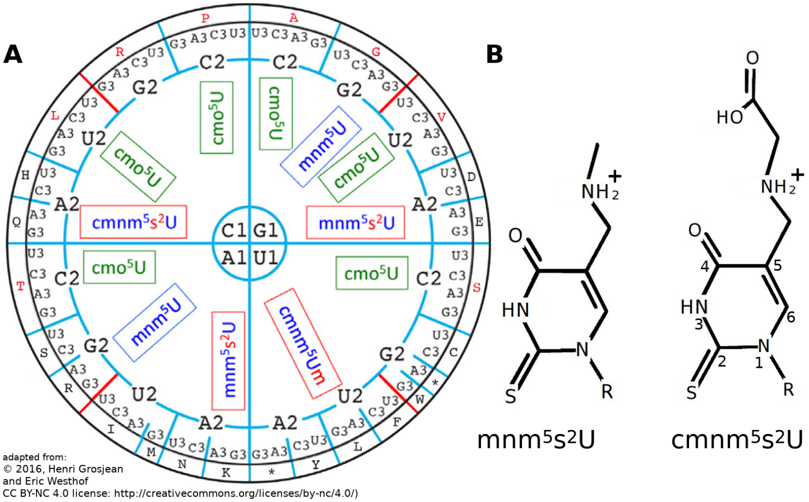
Since this postulation, the wobble hypothesis has been significantly updated, but maintained its core (Agris et al., 2018). Synonymous codons in the Standard genetic code read by the same tRNA are always[2] divided into 2-codon or 4-codon boxes (Grosjean et al., 2010). The 2-codon groups always comprise either purine-ending or pyrimidine-ending codons. Although these basic principles are maintained across all phyla, the strategies of how to divide synonymous codons between tRNAs largely differ (Grosjean et al., 2010). In bacteria, tRNAs that ambiguously decode Y-ending codons almost exclusively have G in the wobble position, and tRNAs that decode R-ending codons have U in the wobble position; tRNAs that ambiguously decode 4-codon groups have U in the wobble position exclusively (Grosjean et al., 2010). While base pair geometry in the wobble position is less strictly monitored by A1492 and A1493 compared to the first two positions, some constraints are still exerted by ribosomal residues, including G530 and C1054 of 16S rRNA (Ogle et al., 2001). An "indirect" interaction with ribosomal protein S12 via Mg^{2+} ion was also suggested

[2] Except for the initiator tRNA and tRNA^{Trp}_{CCA}, which must avoid decoding UGA stop codon

(Ogle et al., 2001). The presence of constraints at the wobble position is used to explain an interesting observation: while unmodified G34 can efficiently decode both Y-ending codons, unmodified U34 cannot decode both R-ending codons (Grosjean et al., 2010, Grosjean and Westhof, 2016). Current explanations of this phenomenon are based on the non-isostericity of the wobble base pair, which implies that G∘U34 wb and U∘G34 wb would have different stabilities in a fixed geometry of the decoding site (Grosjean and Westhof, 2016). The higher stability of the latter is supported by simplified interaction energy calculations (Grosjean and Westhof, 2016). This argument is used in support of the model where U∘G34 adopts the wobble geometry during decoding, which is also supported by *some* X-ray studies (Demeshkina et al., 2012, Ogle et al., 2001). However, the same X-ray study of the 30S subunit, but in the absence of paromomycin, revealed the WC geometry of U∘G34 (Ogle et al., 2001). Overall, the base pair geometry constraints at the third codon position, i.e. the validity of the geometric selection model at this position, is less clear compared to the first two positions, where such constraints are evident.

1.3.2 The role of wobble tRNA modifications

U34 of tRNAs involved in ambiguous decoding is almost always modified (Agris et al., 2018, Grosjean et al., 2010). Nucleobase modifications of U34 have variable chemistry and are phyla-dependent, but mostly restricted to the 2nd and 5th positions of the U34 (Grosjean et al., 2010, Machnicka et al., 2013). In *E. coli*, there are two tRNA modifications at C5 atom of U34 involved in ambiguous decoding of 2-codon groups: 5-methylaminomethyluracil (mnm^5U) and 5-carboxymethylaminomethyluracil ($cmnm^5U$). Both modifications can also coexist with thiolation at C2 atom (mnm^5s^2U, $cmnm^5s^2U$). tRNAs with these modifications ambiguously decode six 2-codon groups of R-ending codons (Grosjean and Westhof, 2016) (Fig. 1.8A). It has been shown that $(c)mnm^5U$ and s^2U modifications increase decoding of both A-ending and G-ending codons (Kurata et al., 2008, Ranjan and Rodnina, 2017, Rodriguez-Hernandez et al., 2013, Yarian et al., 2002). However, some earlier studies indicated that mnm^5U can decrease decoding efficiency of the full-cognate A-ending codons (Krüger et al., 1998). The mechanism of ambiguous decoding facilitation by these modifications is not yet well understood. It was suggested that s^2U modification increases base stacking in the codon-anticodon helix (Larsen et al., 2015). However, it does not explain why s^2U is restricted to the R-ending 2-codon groups, i.e. why it is not involved in ambiguous decoding of 4-codon groups and why it does not cause misreading of Y-ending codons (Grosjean et al., 2010).



adapted from:
© 2016, Henri Grosjean
and Eric Westhof
CC BY-NC 4.0 license: http://creativecommons.org/licenses/by-nc/4.0/)

Figure 1.8.: Ambiguous codon decoding an the role of tRNA modifications. **A** – A scheme of the ambiguous decoding in *E.coli*. Reproduced from Grosjean and Westhof (2016) under Creative Commons license. **B** – structures of mnm^5s^2U and $cmnm^5s^2U$ modifications.

The common property of s^2U and $(c)mnm^5U$ modifications is their electron-withdrawing groups, which increase the acidity of N3 of the modified U (Sochacka et al., 2017). This property is particularly strong in $(c)mnm^5U$, as its amino group is protonated in solution at physiological pH. Takai and Yokoyama (2003) suggested a model where $(c)mnm^5s^2U$ deprotonated at N3 base pairs with G via WC-like or reversed-wobble (rwb) base pair geometries. The WC geometry of G$\circ$$mnm^5U$ was observed in crystal structures of the 30S subunit (Murphy et al., 2004) and the rwb geometry of G$\circ$$mnm^5s^2U$ was observed later in a full 70S ribosome structure (Rozov et al., 2016a). pKa(N3) calculations suggest that 30 to 50 % of $(c)mnm^5s^2U$ would be deprotonated at N3 in water at neutral pH (Sochacka et al., 2017). In some tRNA species, Se^2 can be used instead of S^2. This modification has even higher acidity and is predicted to preferentially decode G-ending codons over A-ending (Leszczynska et al., 2020). Due to multiple discrepancies between *in vitro* and *in vivo* studies, different ribosomal structures and unknown ionization properties of the anticodon modifications in the ribosomal environment, understanding of ambiguous decoding mechanisms is currently lacking.

1.4 Aims of the study

The general aim of this study is to deepen understanding of the interplay between the physical properties of the wb-WC reaction, the molecular environments it occurs, and the biochemical processes it can affect. The questions we ask are:

- How this reaction is affected by different states of the molecular environments of base pair recognition systems – ribosome and DNA polymerases?
- What are the physical origins of these effects?
- How the effects exerted by these molecular machines on the wb-WC reaction influence the recognition processes they perform?
- How the interplay of these effects can be tuned by tRNA modifications?

The first three questions can be combined into a more general problem: "How to describe and study a molecular recognition system where both partners are flexible?". We address these questions with a range of computational and theoretical methods.

1.5 Structure of the dissertation

In the next Chapter we outline the theoretical background, advantages and limitations of the computational methods employed in the following chapters. In Chapter 3 we apply these methods to study effects of the molecular environments on the wb-WC reaction. In Chapter 4 we build a new kinetic model of decoding in translation which can incorporate these effects and predict their consequences for the decoding mechanism. In Chapter 5 we apply computational methods to study how the mechanisms uncovered in the previous chapters can be potentially leveraged by tRNA modification to facilitate ambiguous decoding.

Chapter 2
Theoretical background

The biological problems highlighted in the previous chapter largely arise from experimental observations that lack consistent physicochemical explanations. As any other field of natural sciences, molecular biophysics/biochemistry requires a combination of theory and experiment for progress (Bottaro and Lindorff-Larsen, 2018). It benefits from the already developed fundamental theories, mainly statistical mechanics and quantum theory, as it seeks to explain biological phenomena within the framework of physics. The variety and complexity of molecular processes in biological systems, a result of long-going evolutionary optimization, require a range of approximations to derive usable and reliable models.

2.1 Approximations and levels of theory

Quantum theory provides a complete description of matter assuming low velocities and negligible gravitational effects. In this case, all observables of a system can be obtained from its wave function derivable via the time-independent Schrödinger equation:

$$H\Psi = E\Psi \tag{2.1}$$

where H – Hamiltonian operator, Ψ – wave function and E – system energy.

Unfortunately, this equation can be solved analytically only for H atom and the corresponding isoelectronic ions. Any molecular system constitutes a many-body problem and thus requires approximate solutions. As the quantum chemistry (QM) calculations are limited by available computational resources, a hierarchy of QM methods have been developed to address a range of required precision and scale in (bio)chemical problems (Cramer, 2004, Jensen, 2007) (Fig. 2.1A).

2.1.1 Born-Oppenheimer approximation

A fully quantum treatment of a molecular system is often not required. Atomic nuclei are much heavier than electrons. While the latter can only be described by quantum mechanics, the nuclei experience quantum effects far less, and thus can be assumed as classical particles (point charges). The difference in mass results in the difference in momentum (electron "velocities" in molecules are much higher) and means that electrons virtually instantly adjust to changes in nuclear positions. This approximation allows to neglect electron-nuclear velocity coupling, and is called the Born-Oppenheimer approximation (Born and Oppenheimer, 1927). With it, the molecular wave function depends on the nuclear coordinates only parametrically, meaning that Schrödinger equation can be solved for fixed nuclear geometries. This approximation is essential in chemistry because it creates the concept of potential energy surface (PES) – the total (electronic) energy of a molecular system as a function of atomic coordinates. PES contains local minima and saddle points. These stationary points correspond to stable molecular geometries and transition states, structures and relative energies of which describe mechanisms, thermodynamics, and kinetics of chemical processes. Born-Oppenheimer approximation might be less valid for light atoms, particularly H atoms. Quantum effects in protons during chemical reactions are associated with quantum tunneling – a spatial propagation of proton wave functions between minima on PES without passing the activation energy barriers. Although commonly believed to be relevant only at very low temperatures, recent studies suggest the relevance of nuclear quantum effects at room temperature and in biochemical processes at physiological conditions (Markland and Ceriotti, 2018). Quantum tunneling results in increased rates of proton transfer reactions compared to rates predicted from the classical transition state theory (Klinman and Kohen, 2013, Markland and Ceriotti, 2018, Pusuluk et al., 2018). Computational studies predict that nuclear quantum effects can stabilize base pairs by affecting H-bonding strength via proton delocalization (Fang et al., 2016). Full-quantum treatment of molecular systems requires very computationally expensive methods, such as the path integral approach (Markland and Ceriotti, 2018). Nuclear quantum effects are less likely to dramatically affect the wb-WC reaction, as it involves movements of heavy atoms besides the proton transfers

(Fig. 1.7).

2.1.2 Calculating molecular wave functions

The Born-Oppenheimer approximation is only the first step towards a range of computationally-efficient QM methods of solving time-independent Schrödinger equation for molecular systems. For any QM method, the molecular wave function must be constructed from linear combinations of elementary (atomic) wave functions (LCAO approximation) (Cramer, 2004, Jensen, 2007). Such a combination is called a basis set and can assume various functional forms. For nonperiodic systems, it is common to derive a basis set from a combination of Gaussian functions centered on nuclei. It has proven useful to also add modified Gaussians to the basis set for a more accurate treatment of delocalized electron density: diffuse and polarized Gaussian functions. The more elementary functions are contained in the basis set, the closest the molecular wave function can be to the exact solution (i.e., the complete basis set limit). However, a larger basis set requires more computations to optimize the molecular wave function to the lowest-energy solution (see below). Therefore, in addition to a trade-off between the quality of QM methods and their computational cost, there is also a trade-off between the basis set size and the associated computational cost.

The first consistent approach to calculate molecular wave functions was Hartree-Fock method (HF; only "restricted" HF is considered here, RHF) (Cramer, 2004, Jensen, 2007). RHF is based on the mean-field approximation: each electron interacts with the average potential from all electrons in the molecular system. Quantum theory demands RHF to satisfy several essential constraints on the calculated molecular wave function: the correct treatment of electron spin, and to follow the Pauli exclusion principle. To fulfill the former, the basis set functions (i.e. one-electron orbitals) are multiplied by spin functions to produce spin-orbitals (thus now including all quantum numbers). The Pauli exclusion principle demands not only the exclusion of electrons with the same quantum numbers from the same orbital, but also that the molecular wave function changes sign whenever the coordinates of two electrons are interchanged (i.e. it should be antisymmetric). To account for this, a molecular wave function is constructed from atomic orbitals arranged in a matrix, where each electron populates one row. Determinants calculated from these matrices fulfill the antisymmetry requirement, as they change sign when the rows are interchanged. This determinant in the RHF method is called Slater determinant. The basis set enters the molecular wave function as a linear combination of one-electron orbitals ψ_i with some coefficients a_i (and multiplied by the spin functions). These coefficients are unknowns and can form a virtually infinite number of combinations resulting in different molecular orbitals and thus energies as solutions of Eq. (2.1). To find the values of these coefficients, the variational principle is used. It states that the "true" ground-state molecular wave function has the lowest energy among all other possible molecular wave functions constructed from a given basis set. The variational principle provides a condition to find the optimal combination of the coefficients a_i: such a combination must provide as lowest energy as possible. However, calculating these energies requires solving Eq. (2.1), which in turn requires knowing the electron distribution. Therefore, the molecular wave function is optimized in an iterative way using self-consistent field approach (SCF): first, a_i are guessed, and energies are calculated until convergence (some acceptable value of energy fluctuation), improving a_i towards minimal E. From the variational principle ($\delta E/\delta a_i = 0$) and Eq. (2.1), solving for a_i gives the following matrix element (for spin-orbitals μ,ν) in the matrix for Slater determinant calculation:

$$F_{\mu\nu} = \langle\mu|-\frac{1}{2}\nabla^2|\nu\rangle - \sum_{k}^{nuclei} Z_k\,\langle\mu|\frac{1}{r_k}|\nu\rangle + \sum_{\lambda\sigma} P_{\lambda\sigma}\left((\mu\nu|\lambda\sigma) - \frac{1}{2}(\mu\lambda|\nu\sigma)\right) \quad (2.2)$$

The first term in Eq. (2.2) denotes the one-electron (two-index) integral describing the electronic kinetic energy (∇^2 – Laplacian operator). The second term is the one-electron integral describing electron-nuclear attraction (Z_k – charge of a nucleus k). The last term is the difference of two-electron (four-index) integrals weighted by the density matrix $P_{\lambda\sigma}$ (where λ,σ – all other spin-orbitals) – contribution of individual basis functions to the molecular wave function. $P_{\lambda\sigma}$ is what is being optimized in the SCF procedure. The two integrals in the last term are Coulomb (classical electron repulsion) and exchange (repulsion due to Pauli principle) integrals. All integrals in Eq. (2.2) have to be calculated numerically. The 4-center integrals are particularly computationally demanding and cause RHF scaling as $\mathcal{O}(N^4)$ on the number of basis functions.

Since HF method is developed based on the "first principles" of quantum theory, HF and its more rigorous derivatives are called *ab initio* QM methods, as opposed to other methods that use experimental (empirical) corrections. Due to both low computational efficiency and low accuracy, the pure HF method is rarely used nowadays. Instead, it is a branching point for the following methods, which either simplify/parameterize some terms in HF equations for improved computational speed (and sometimes also accuracy), or add additional terms for improved accuracy towards the "exact" solution of the Schrödinger equation (Cramer, 2004, Jensen, 2007).

It is useful to denote a QM method and the basis set in a single notation coined by Pople and known as a *level of theory*: QM-method/basis-set. For example, HF/6-31+G* denotes HF method used with a basis set containing nuclear-centered Gaussians at each atom in the following combination: 6 primitive Gaussian for core electrons; the valence electron orbitals are split in two parts comprising a linear combination of 3 and 1 primitive Gaussians; polarized (+) and diffuse (*) functions are added to the atomic orbitals of heavy atoms.

2.1.3 Post-HF methods

The difference between the "true" E and E calculated from the HF method is called electron correlation. Post-HF methods aim at solving the problem of neglected electron correlation in HF method. Three main post-HF approaches are: configuration interaction (CI), coupled-cluster theory (CC), and perturbation theory (Cramer, 2004, Jensen, 2007). The common feature of these methods is the consideration of not only a single ground-state Slater determinant, as in the RHF method, but also other, excited-state determinants. Including all combinations of excitations (i.e. full-CI) is generally not possible, therefore truncation at a desired excitation is usually performed. Truncated CI is not size-extensive (E of infinitely separated molecules $\neq$ sum of individual energies), therefore is less common compared to coupled-cluster and perturbation theory. Møller–Plesset (MP) theory uses perturbation theory to include excited Slater determinants as perturbations (terms in Taylor expansion) to the RHF solution.

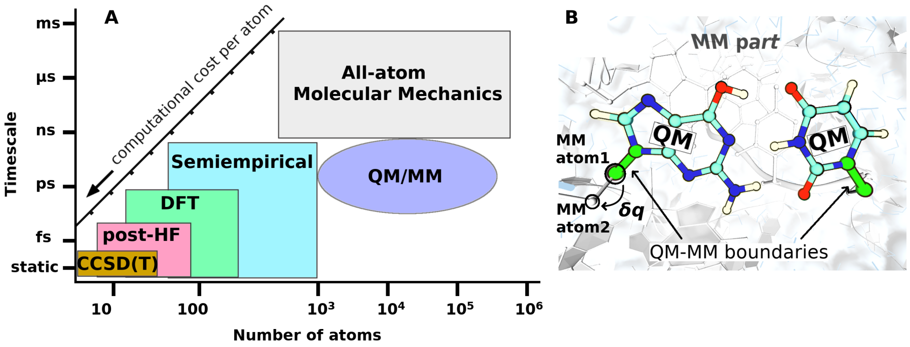


Figure 2.1.: Approximations in molecular modeling. **A** – illustration of approximate theoretical levels for describing a system-process depending on its size and timescale. **B** – a scheme of the QM/MM approach.

Truncation at the second-order perturbation (MP2) (describing pairwise electron correlation) is the most common due to its favorable cost-accuracy ratio. CC is similar to CI, but includes excitations as clusters (i.e., all coupling between excitation combinations). Including both single and double excitations results in CCSD. CCSD(T) adds excitation triplets via the MP4 approach, and is considered the "gold standard" of computational chemistry (Raghavachari et al., 1989). CCSD(T) is extremely computationally expensive and scales as $\mathcal{O}(N^7)$, thus is limited only to very small systems and commonly used as a reference (Bartlett and Musiał, 2007) (Fig. 2.1A). Recently, a much faster linear scaling DLPNO-CCSD(T) approximation was developed, which captures 99.9% of CCSD(T) energies (Guo et al., 2018)

2.1.4 Semi-empirical methods

Semiempirical methods (SE) aim at improving HF performance by neglecting some time-consuming operations and adding empirical parameters to replace others (Cramer, 2004, Jensen, 2007). A common approximation among all SE methods is the minimal basis set: only valence orbitals are included in the basis set (i.e., core electrons are added implicitly as neutralizing charge on the nuclei), and only one basis function is used per valence orbital (s,p,...). In the neglect of the diatomic differential overlap (NDDO) approximation, two-electron (4-center) integrals in Eq. (2.2) are considered only if μ and ν are centered on one atom, and λ and σ are also centered on one atom (thus only 2-centered integrals are considered). In this way, these 2-electron 2-center integrals can be reduced to evaluating two 1-electron integrals. All one-electron integrals are simplified in a way to neglect all orbital overlap from different atoms. The remaining one-electron integrals (from the same atom) are parameterised depending on the type of orbital overlap (e.g., s-s, s-p etc) based (partly) on experimental measurements of the ionization potential. These approximations allow to avoid computationally expensive numerical integration, thus making SE applicable to large systems (Fig. 2.1A). The SCF procedure is still performed, but is greatly accelerated upon this approximation. Besides the parameters derived directly from experimental data, there are also a number of free parameters that can be tuned to fit SE methods predictions to experimental data sets. The first "usable" SE method derived from the NDDO approximation was Austin model 1 (AM1) (Dewar et al., 1985). In AM1, the free parameters were fitted manually to a limited set of experimental data (atomic heat of formation). In Parametric method number 3 (PM3) (Stewart, 1989), the free parameters were fitted simultaneously using penalty functions to reproduce the heat of formation in a much larger experimental data set. The following methods (PM6 and PM7) improved the performance of PM3 by expanding the training set and fitting not only to the heat of formation, but also to the molecular geometries (Stewart, 2013). PM7 also included corrections to better reproduce the dispersion interactions and H-bonding properties (Stewart, 2013). As a result, for molecules within (or similar to) the training set, PM7 method achieved accuracy level of some density functionals (Christensen et al., 2017) (see below, and Chapter 3), while being several orders of magnitude faster and having almost linear scaling.

2.1.5 Density functional theory

Density Functional Theory (DFT) employs an alternative approach to calculate the energies of molecular systems. Instead of focusing on optimizing molecular wave functions, DFT focuses on calculating electron density (Cramer, 2004, Jensen, 2007).

Hohenberg-Kohn theorem proves that ground-state energy can be calculated *exactly* as a functional of electron density (Hohenberg and Kohn, 1964). However, this theorem does not provide such functional. Therefore, while DFT is exact, only approximate functionals can be derived. Kohn-Sham (KS) equations allow calculating energy in a manner similar to the HF method, except for exchange-correlation energy E_{xc} – the difference between the "true" energy and the energy of the noninteracting electron gas used in KS equations (Kohn and Sham, 1965). A range of DFT functionals has been proposed implementing different approaches to calculate E_{xc} (Mardirossian and Head-Gordon, 2017). Most DFT functionals use empirical corrections in the calculation of exchange-correlation terms, thus the distinction between SE and DFT methods may not always be clear. Expanding on the local density approximation (LDA), which obtains E_{xc} in KS equations for uniform electron gas, the generalized gradient approximation (GGA) introduces exchange-correlation electron density gradients as Taylor expansions of the LDA solution, including also empirical parameters fitted to the exact exchange energies of noble gases. One of the most popular GGA functionals is Becke-Lee-Yang-Par (BLYP), which has reasonable accuracy while being much less expensive compared to the following DFT functionals (Mardirossian and Head-Gordon, 2017). The next step to improve E_{xc} calculation was to add exact HF exchange to the functional, which is performed in hybrid functionals. Free parameters in this procedure can be fitted to the experimental data, resulting in a variety of hybrid functionals. One of the most popular hybrid functionals is B3LYP, which is based on BLYP and adds three fitting parameters (Becke, 1993). DFT became highly popular in computational chemistry due to its favorable computational cost (it scales as $\mathcal{O}(N^3)$) combined with accuracy far exceeding HF accuracy (Mardirossian and Head-Gordon, 2017), making at a method of choice for a range of medium-sized systems (Fig. 2.1A). The major drawback of DFT is the absence of a way for systematic improvement, as opposed to the post-HF methods. Another issue of the DFT methods is the poor description of dispersion interactions, which are crucial in biomolecular systems. To improve accuracy, empirical dispersion corrections are added to DFT functionals (such as D3 or D3BJ) (Antony et al., 2015, Kruse et al., 2012).

2.1.6 Molecular Mechanics

All methods described above were QM methods, no matter to what extent they approximate solutions of the Schrödinger equation. Quantum-mechanical treatment is topology-agnostic, meaning that the molecular topology information (bonds, angles, dihedral angles etc) is not needed in QM calculations. This generality of the QM methods comes from the explicit modeling of electrons. It allows addressing phenomena essential for (bio)chemistry: chemical reactions, in which the molecular topology changes by breaking/formation of chemical bonds. It is possible to trade this generality for increased computational speed. In the QM methods, under the Born-Oppenheimer approximation, the classical particles are only the atomic nuclei. Changing from explicit to implicit treatment of electrons would require electrons to "join" the classical nuclei to form the classical atoms. If all electrons become implicit parts of the classical atoms, then there is no electron density between nuclei, which in the QM framework "forms" chemical bonds. Therefore, the chemical bonding in a given molecular system needs to be specified explicitly, forming the molecular topology. This approach is called Molecular mechanics (MM).

Depending on the electron density distribution in a molecule, the electronic properties in the vicinity of a given nucleus can vary, which should be captured in the MM approach. Assuming that the electronic properties of an atom are influenced only by its closest neighboring atoms, it is possible to specify a finite number of atom types, which reflect such influences. The atom type is a set of parameters implicitly describing the electron density distribution around a given nucleus in a given environment: point charge and van der Waals radius. Parameters are also needed to implicitly describe the electron density between the atoms, which affects the strength of chemical bonds and other interactions, thereby affecting the molecular potential energy as a function of atomic coordinates. This is achieved by defining the potential energy function of a molecule in terms describing the classical interactions between atoms: bond stretching (2 bonded atoms), bending (3 bonded atoms), dihedral torsion (4 bonded atoms) and nonbonding pairwise interactions between atoms within a given distance cut-off: Coulomb and van der Waals interactions. Bonding terms include parameters describing the strength of these interactions (force constants) and their minimal energy values (equilibrium bond lengths/angles). A set of potential energy terms, atom types, topologies of molecules/residues, and parameters of the potential energy terms is called a force field (FF). Similarly to the SE methods, parameters in the FFs can be fitted to experimental data sets to reproduce observables of interest (geometries, densities etc). The most common force fields CHARMM, AMBER, and GROMOS use combinations of experimental fitting and fitting to high levels of QM theory (Cesari et al., 2019, Fröhlking et al., 2020, Huang and Mackerell, 2013, Kührová et al., 2019). These FFs contain parameters for many common organic molecules, including all amino acids and nucleosides. If a compound (or isomer) of interest is not available in the FF distribution, it can (should!) be parameterized using the same levels of QM theory as the rest of the FF. The MM approach is very computationally efficient and has linear scaling (assuming a fixed cut-off for nonbonded interactions), and thus commonly used for large systems and simulations thereof (see below) (Fig. 2.1A).

2.1.7 QM/MM approach

QM and MM methods correspond to different "philosophies" of computational molecular studies: accurate (and computationally expensive) topology-agnostic treatment, and fast fixed-topology treatment, respectively. However, the problem at hand often requires the virtues of both approaches simultaneously. If a chemical reaction, or any other intrinsically quantum phenomenon, is studied in a relatively large condensed phase system, one cannot use either approach independently. To address this problem, a hybrid quantum-mechanical/molecular-mechanical (QM/MM) approach was developed, which allows combining both approaches in a single system (Ahmadi et al., 2018, Himo, 2017, Janoš et al., 2016). In QM/MM approach, a small subsystem is described with QM methods, while the rest of the system is described with MM (Fig. 2.1B), allowing modeling chemical processes in large systems (Fig. 2.1A). In this approach, a package (or a module within one package) performing QM calculations provides the energies and gradients of the QM subsystem to the package which "drives" the calculations, replacing the energies and gradients of the QM subsystem calculated with MM (Melo et al., 2018). Depending on the properties exchanged between QM and MM subsystems, different embedding schemes exist. In mechanical embedding, the QM subsystem receives only the coordinates of its atoms, thus QM calculations are performed effectively *in vacuo*. In a more realistic electrostatic embedding, the QM subsystem receives also point charges from the MM subsystem, which form the electrostatic potential and polarize the QM electron density. This is the most common embedding scheme nowadays (Ahmadi et al., 2018), but it still treats polarization only one way. The most realistic scheme is polarizable embedding, in which the MM subsystem can also be polarized by the QM charge distribution. This scheme requires polarizable force fields, in which multipoles or Drude oscillators replace point charges (Jing et al., 2019, Lemkul and MacKerell, 2018). Polarizable embedding is very computationally demanding, as each QM/MM energy calculation is performed in a self-consistent way to account for the polarization of both subsystems (Loco et al., 2017, 2019). Although having superior accuracy, polarizable embedding is still too computationally expensive to be used for large systems in simulations (Bondanza et al., 2020). Therefore, in this study we used only the electrostatic embedding scheme.

The most delicate part of the QM/MM approach is the interface between QM and MM subsystems. If the QM-MM separation occurs along covalent bonds, as it usually happens in biomolecular systems, the QM subsystem would represent a radical or ion not intended to be modeled. To overcome this, different approaches have been developed, including the link atom approach (Melo et al., 2018). In this approach, an atom (usually H) is added to the QM subsystem along the QM-MM boundary bond. To solve the consequent issue of electrostatic repulsion of closely placed atoms, a charge shift scheme is used, in which a partial charge of the boundary MM atom is distributed to other MM atoms in the vicinity (Fig. 2.1B), maintaining the total charge and dipole moment distribution (Melo et al., 2018).

Besides the obvious question of the choice of QM level of theory and MM force fields in QM/MM calculations, this approach may also depend on the QM region size. While some authors argue for very large QM regions ($\sim$ 500 atoms) to reach acceptable accuracy (Kulik et al., 2016), other studies demonstrate accurate results with small QM regions ($\sim$ 30 atoms) using proton transfers in base pairs as and example (Das et al., 2018). In this study our QM region ranged from a single base pair (in Chapter 3) to three base pair with a solvation shell (Chapter 5).

2.2 Simulations and free energy calculations

Until now, computational methods were discussed in terms of calculating energy for fixed atomic coordinates. In practice, these coordinates have first to be found among all possible coordinates in $3N-6$ Cartesian space of molecule with N atoms.

2.2.1 Static calculations

In a typical (bio)chemical problem, one is interested in the relative energies of stationary points on a PES constructed from these coordinates: local minima and transition states. In theory, it is possible to probe all these coordinates (within some cut-off and increment) one by one, obtaining the full PES, from which to derive stationary points. If only the stationary points are of interest, it is much more efficient to employ geometry optimization techniques, in which some additional calculations are performed to drive the search towards these points, dramatically reducing the total computational time. QM and MM methods provide ways to calculate not only the energies, but also their derivatives on the atomic coordinates – gradients. By following these gradients (and sometimes also the second derivatives – curvatures), it is possible to reach local minima and TS in a reduced number of steps.

Calculations of relative energies on PES only provide ΔE – relative potential energy (electronic energy in case of Schrödinger equation solutions). Usually, the goal of computationally chemistry is to calculate/predict experimental observables, such as equilibrium (Boltzmann) populations and rate constants, which depend on free energy. For biological systems (constant pressure, i.e., NPT ensemble), Gibbs free energy G is of interest:

$$G = U + k_B T - TS = H - TS \tag{2.3}$$

where U – internal energy, H – enthalpy, S – entropy, T – temperature and k_B – Boltzmann's constant. Calculations of internal energy and entropy require obtaining the partition function, which includes all energy levels of a molecule (electronic, vibrational, translational, and rotational). In computational chemistry, it is most common to calculate the partition function of a molecular geometry under the rigid-rotor harmonic-oscillator approximation. This approximation allows to decouple these degrees of freedom, making it straightforward to calculate the partition function exactly. The electronic degrees of freedom are provided by the wave function, but vibrational levels require additional calculations of vibrational frequencies, which often are time-limiting, as they require calculating a second-order derivative matrix (Hessian). The energy of these vibrational frequencies (at 0 K) is called zero-point energy (ZPE). By adding electronic energy, ZPE and thermal correction, one obtains H, and by adding entropic contribution calculated from the partition function, G can be calculated.

Therefore, a typical computational chemistry study of some process (e.g., chemical reaction or conformational change) in isolated molecules (or relatively small molecular complexes) consists of geometry optimizations, calculations of vibrational frequencies and finally the calculation of relative free energies of the local minima and (if needed) TSs. This is the static approach of computational chemistry calculations. This approach works because the isolated molecule satisfies the ideal gas approximation, and the partition function can be calculated under the described above approximation. However, in the condensed phase (many interacting molecules), the exact calculation of the partition function from molecular properties is impossible due to numerous and degenerate energy levels. Moreover, PES (or free energy surface, FES) of a condensed-phase system contains a large number of local minima, which would require many computationally demanding geometry optimizations. Therefore, for condensed phases, instead of the static approach, one must sample the phase space – space of system positions and momenta.

2.2.2 Molecular dynamics simulations

Sampling of the phase space is performed via molecular dynamics (MD). In MD, the forces on atoms (nuclei) are calculated as derivatives of the potential energy functions being used. Using the forces, the coordinates are propagated by integrating the equations of motion in discrete time steps Δt. Modern integration methods often use the leapfrog or velocity Verlet algorithm, where velocities are calculated at the midpoint of Δt for positions (Jensen, 2007). By performing MD, one proceeds from studying structures to studying ensembles. The core of MD simulations is *ergodic hypothesis*, which states that the ensemble average (of some observable) is equivalent to the time average in the limit of $t \to \infty$. In real MD simulations, this limit is of course cannot be reached, and thus some *convergence* of the simulations, acceptable for a problem at hand, has to be verified. Sampling thermodynamic ensembles requires controlling temperature (for canonical, NVT ensemble) and/or pressure (for isothermal-isobaric, NPT ensemble). This is achieved by computational thermostats and barostats, which work by rescaling the velocities and forces on atoms to reach an average of the needed thermodynamic state function (Jensen, 2007). To avoid surface artifacts when sampling a molecular system, periodic boundary conditions (PBC) are employed. In PBC, one effectively simulates a "crystal" by simulating a single cell, atoms in which can interact with and visit the neighboring cells. In equilibrium MD simulations, the probability to sample a given state at the coordinate q is given by the Boltzmann distribution:

$$p(q) \propto \exp(-\frac{U(q)}{k_B T}) \tag{2.4}$$

where U – potential energy, k_B – Boltzmann constant. If one is interested in simulating a process involving energy barriers higher than thermal energy, higher energy states are very rarely sampled, precluding the application of a simple equilibrium sampling approach to most problems. In this case, enhanced sampling, or free energy methods are applied.

2.2.3 Umbrella sampling

When studying any (bio)chemical process using MD simulations, one can usually define a reaction coordinate ξ, such that the process can be described by changes of this coordinate. ξ is also known as collective variable (CV). The free energy [3] associated with CV is defined as the potential of mean force (PMF) (Roux, 1995):

[3] Here we use "PMF" and "free energy" interchangeably, although PMF is in general not equal to free energy

$$W(\xi) = -k_B T \ln\langle p(\xi)\rangle \tag{2.5}$$

where $\langle\rangle$ denotes averaging over all degrees of freedom other than ξ. As PMF includes high energy states (i.e., free energy barriers) along ξ, it is possible to add additional potential w_i at these values to overcome the barriers and sample the states of interest. This restraints the system at around the given value ξ, creating a simulation *window*. As harmonic restraints ("umbrellas") are most common, such approach and the following way to calculate PMF is known as umbrella sampling (US) (Torrie and Valleau, 1977). Adding such potential w_i to the probability distribution function results in biased probability distribution:

$$\langle p(\xi)\rangle_i^{biased} = \frac{\exp(-\frac{w_i(\xi)}{k_B T})\langle p(\xi)\rangle}{\langle \exp(-\frac{w_i(\xi)}{k_B T})\rangle} \tag{2.6}$$

Simulating a number of such windows restrained with w_i at different ξ values provides a biased probability distribution along ξ (Fig. 2.2A), from which the unbiased PMF must be constructed. Unbiased PMF at window i is calculated as:

$$W(\xi) = -k_B T \ln\langle p(\xi)\rangle - w_i + F_i \tag{2.7}$$

where F_i is an unknown free energy constant. If multiple windows are used, F_i cannot be calculated exactly. Therefore, a weighted histogram analysis method (WHAM) (Kumar et al., 1992) is used to calculate F_i iteratively until convergence, resulting in unbiased PMF (Kästner, 2011).

2.2.4 Metadynamics

Another possible way to calculate PMF in MD simulations is to add potentials at positions along ξ *during* simulations at some intervals. This approach was first applied in the local elevation method by (Huber et al., 1994), but now most commonly known as metadynamics (metd) (Bussi and Laio, 2020, Valsson et al., 2016). In MetD, positive Gaussian potentials are added to the potential energy function of ξ to improve sam-

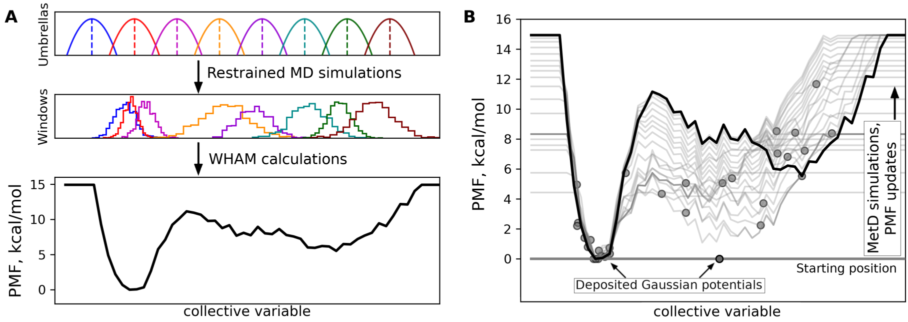


Figure 2.2.: Schemes of umbrella sampling and metadynamics. **A** – Umbrella sampling: a set of harmonic potentials ("umbrellas") is used as restraints in simulations, producing a biased ensemble which samples the dimension of interest (collective variable). Then, WHAM is used to reconstruct unbiased PMF from the biased ensembles **B** – Metadynamics: Gaussian potentials are regularly deposited to the potential function, forcing the system to sample high energy regions of the CV. Then, these potentials are summed to produce the PMF. The PMF shown on this figure is derived from our MetD calculations in Chapter 5.

pling along the coordinates by forcing the system to explore other regions Fig. 2.2. It is then straightforward to sum the deposited potentials to obtain an unbiased PMF, given that MetD simulations are converged. To improve the convergence, one can decrease the Gaussian potential height over time, which is known as well-tempered MetD (Barducci et al., 2008).

Chapter 3

Environmental effects on the tautomerization reaction

Parts of this Chapter are included in the publication (Kazantsev and Ignatova, 2020)

In this chapter, I address the question of how the molecular surroundings of a G∘U base pair affect the properties of the wb-WC reaction. To do so, I perform classical MD simulations to estimate the polarity of the environments, and QM/MM US calculations to calculate the free energy profiles of the wb-WC reactions in the studied systems.

3.1 Methods

3.1.1 Selection of the QM level of theory for QM/MM calculations

One of our goals in this study was to calculate the potential of mean force (PMF) of the wb-WC reaction in various molecular environments using the hybrid QM/MM approach. For such calculations to converge, enough conformational sampling is required, which would limit the use of computationally expensive high levels of theory, such as some density functional theory (DFT) methods. Therefore, we focused on faster levels of theory with acceptable accuracy. Semiempirical methods (SE) can provide a great speed-accuracy ratio, but can be highly inaccurate for some systems (Christensen et al., 2017, 2016, Stewart, 2017). Therefore, we first aimed to estimate the accuracy of several selected SE methods, namely, PM family methods (PM3, PM6, PM6-D3, and PM7), in describing the energy changes of the wb-WC reaction. Besides the gas phase energies, the optimal level of theory needed to accurately capture the effect of the implicit solvent. First, we compared the total energy of the WC formation (ΔE_{wc} =E(G*∘U WC) - E(G∘U wb)) calculated by several DFT and SE methods in gas phase and in the implicit water model, using geometries optimized on

Table 3.1.: Benchmark calculations of total energy of the WC geometry formation

Method	ΔE_{wc}^{gas} [a]	ΔE_{wc}^{water} [b]	$\Delta\Delta E_{wc}^{gas}$ [c]	$\Delta\Delta E_{wc}^{water}$ [d]
DLPNO-CCSD(T)/aug-cc-pVTZ	-1.98	3.17	0.0	0.0
RI-MP2/def2-TZVP	-1.93	3.0	0.05	0.17
ωB97X-D3/def2-TZVP	-1.11	4.53	0.87	1.35
ωB97X-V/def2-TZVP	-1.21	4.38	0.77	1.2
B97M-V/def2-TZVP	-2.05	3.55	0.07	0.38
B3LYP-D3BJ/def2-TZVP	-1.65	3.94	0.33	0.77
M06-2X/6-31+G(d,p)	-3.31	2.18	1.33	0.99
BLYP-D3BJ/def2-SVP	-1.15	3.73	0.83	0.56
b97-3c	-1.5	4.21	0.48	1.03
PBEh-3c	-2.55	2.72	0.57	0.45
PM3	-1.26	3.8	0.72	0.62
PM6	6.6	12.9	8.58	9.73
PM6-D3	6.6	12.9	8.58	9.73
PM7	3.74	10.11	5.72	6.93

[a] Total energy of the WC formation in gas phase (E(G*U WC) - E(GU wb)); [b] Total energy of the WC formation in implicit water model; [c] Absolute error in gas phase; [d] Absolute error in implicit water model; All energies are given in kcal/mol. All calculations were performed on base pair geometries optimized at ωB97X-D3/def2-TZVP level of theory in gas phase.

wB97X-D3/def2-TZVP level. ΔE_{wc} from DLPNO-CCSD(T)/aug-cc-pVTZ was used as a reference in the calculations of absolute errors $\Delta\Delta E_{wc}$. All tested DFT (and RI-MP2) levels demonstrated small errors compared to the reference ($<$ 1.5 kcal/mol) (Table 3.1). All levels of theory demonstrated a monotonic increase of ΔE_{wc} upon increasing relative dielectric permittivity (ε) of the implicit solvent model (Fig. 3.6). SE levels, except for PM3, demonstrated much larger errors in ΔE_{wc} (7 – 10 kcal/mol) (Table 3.1). This is not unexpected, as PM methods were parameterized to accurately reproduce the heat of formation (ΔH), not the total energy (Stewart, 2013). In order to compare ΔH_{wc} of SE and DFT methods, the latter should include zero point energy, which limits the comparison to the DFT levels with successfully optimized geometries. The sets of geometries optimized in gas phase and in the implicit water model were obtained using three levels of theory: wB97X-D3/def2-TZVP (only gas phase), B3LYP-D3BJ/def2-TZVP (no TS) and BLYP-D3BJ/def2-SVP. For each set, single-point heat of formation was calculated by the SE methods and compared to enthalpies from the DFT methods. In this case, we did not use a single reference, but instead compared SE methods to ΔH_{wc} at the DFT level from the corresponding set of geometries. PM6 and PM6 with dispersion correction (PM6-D3) again showed large errors (4 – 9 kcal/mol) (Table 3.2). PM7 and PM3 showed much lower errors (< 1 kcal/mol), comparable to the ΔH_{wc} differences between the selected DFT levels and falling within “chemical accuracy”. Although PM3 method showed relatively high accuracy, it predicted a spurious pathway of the wb-WC reaction (via ion pairs, see Fig. 3.1), therefore was not used for QM/MM simulations.

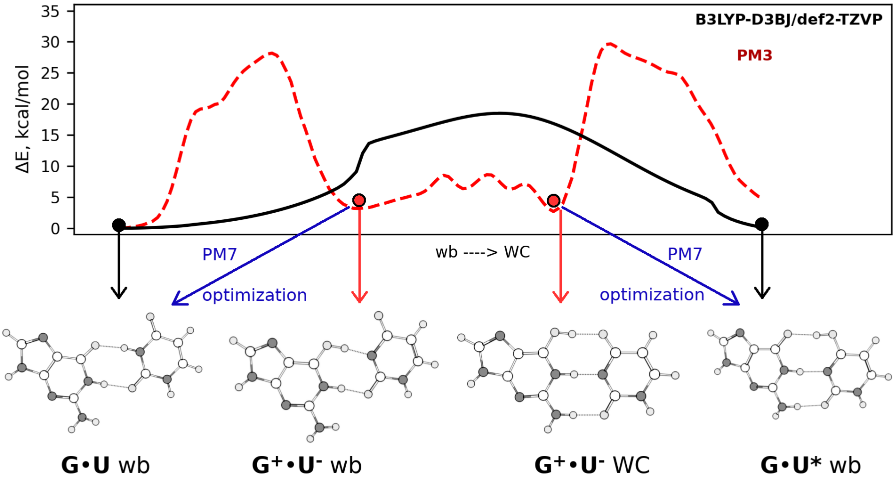


Figure 3.1.: Spurious wb-WC pathway from PM3 method. The plot shows energy profiles from nudged elastic band optimizations of the wb-WC reaction at B3LYP-D3BJ/def2-TZVP and US calculations at PM3 (red) levels of theory. The dots on the curves denote local minima. PM3 pathway contains two additional local minima: ion pairs G$^+$∘U$^-$ wb and G$^+$∘ U$^-$ WC (shown below). Geometry optimizations of these ion pair structures demonstrate that they are not stationary points on the PM7 and DFT levels, and converge to G∘U wb and G∘U* WC, respectively.

Table 3.2.: Benchmark calculations of enthalpy of the WC geometry formation and activation enthalpy

Method	ΔH_{wc}^{gas} [a]	ΔH_{wc}^{water} [b]	$\Delta H_{\ddagger}^{gas}$ [c]	$\Delta H_{\ddagger}^{water}$ [d]
ωB97X-D3/def2-TZVP	**-1.3**	–	**15.98**	–
PM7//ωB97X-D3/def2-TZVP	-1.82	–	23.57	–
PM6//ωB97X-D3/def2-TZVP	7.1	–	29.99	–
PM6-D3//ωB97X-D3/def2-TZVP	4.94	–	27.99	–
PM3//ωB97X-D3/def2-TZVP	-1.26	–	29.06	–
B3LYP-D3BJ/def2-TZVP	**-2.51**	**3.75**	–	–
PM7//B3LYP-D3BJ/def2-TZVP	-3.01	2.75	–	–
PM6//B3LYP-D3BJ/def2-TZVP	7.0	12.72	–	–
PM6-D3//B3LYP-D3BJ/def2-TZVP	4.76	10.58	–	–
PM3//B3LYP-D3BJ/def2-TZVP	-1.54	4.15	–	–
BLYP-D3BJ/def2-SVP	**-1.41**	**3.48**	**17.72**	**20.54**
PM7//BLYP-D3BJ/def2-SVP	-2.53	3.29	20.66	22.27
PM6//BLYP-D3BJ/def2-SVP	5.79	11.94	27.29	28.18
PM6-D3//BLYP-D3BJ/def2-SVP	3.39	9.59	25.16	26.42
PM3//BLYP-D3BJ/def2-SVP	-1.58	4.06	28.85	28.28
PM7(opt)//BLYP-D3BJ/def2-SVP	-0.85	5.27	22.4	23.57
PM7(opt)//B3LYP-D3BJ/def2-TZVP	-1.37	4.31	–	–

[a] Enthalpy of the WC geometry formation in gas phase; [b] Enthalpy of the WC geometry formation in implicit water model; [c] Activation enthalpy in gas phase; [d] Activation enthalpy in implicit water model; All enthalpies are given in kcal/mol. The reference DFT entries are highlighted with a bold font.

We also assessed ΔH_{wc} from PM7-optimized geometries. For this, we used two sets of DFT-optimized reference geometries (BLYP-D3BJ/def2-SVP and B3LYP-D3BJ/def2-TZVP), introduced normally distributed noise (mean = 0, SD = 0.5 Å) to the Cartesian coordinates, and subjected them to re-optimization using PM7. ΔH_{wc} from the re-optimized PM7 geometries demonstrated slightly higher, but still reasonable errors ($<$ 2 kcal/mol) (Table 3.2). This indicates that not only single-point energies, but also gradients (used in geometry optimizations, and later – in QM/MM MD simulations) are reasonably accurate on the PM7 level for local minima geometries. Next, we compared SE and DFT methods based on $\Delta H_{\ddagger}$ – activation enthalpy of the wb-WC reaction. In $\Delta H_{\ddagger}$, SE methods demonstrated much higher errors (2 – 14 kcal/mol) (Table 3.2). PM7 performance was the best among the SE methods (2 – 8 kcal/mol errors), but still far from reaching acceptable accuracy.

Based on these benchmark calculations, we selected PM7 as a QM level of theory for all QM/MM calculations. Large error in $\Delta H_{\ddagger}$ does not allow us to use the activation free energy barriers obtained from these calculations. Since a single-point and gradient calculation using PM7 takes only a fraction of a second for a base-pair-sized system, this choice of a method allowed us to obtain a high cumulative length of the QM/MM MD trajectories in umbrella sampling calculations ($\sim$ 120 ns), which in turn allowed estimating the errors of conformational sampling and convergence. All DFT, RI-MP2, and DLPNO-CCSD(T) calculations were performed in Orca 4.2.1 (Neese, 2018). Tight convergence criteria were set for SCF calculations (10^{-8} a.u.), as well as for geometry optimization. Quasi-Newton optimiser using the BFGS update was used for the local minima optimization, while Berny algorithm was used for TS optimization. Conductor-like polarizable continuum model (CPCM) method was used for the implicit solvation model (Barone and Cossi, 1998). MOPAC-2016 was used for all SE calculations (Stewart, 2016).

3.1.2 System setup

Benchmark systems. Solution NMR structure of DNA dodecamer containing two wobble G∘T base pairs (PDB ID: 1BJD) was used as the initial structure for the DNA system. A heptamer centered on one of the G∘T base pairs was selected (5'-CGTGACG-3', 5'-CGTTACG-3'). The heptamer was solvated in 50 Å x 50 Å x 50 Å box of TIP3P water (Jorgensen et al., 1983). Na^{+} and Cl^{-} ions were added to neutralize the system and reach NaCl concentration of 0.15 M. To build the benzene system, the single G∘T base pair from the DNA system was taken after the equlibration. Deoxyribose was retained, but phosphate groups were removed to obtain the neutral system. The base pair was solvated in 50 Å x 50 Å x 50 Å box of benzene.

DNA-polymerases. Although a crystal structure of pol-β with G∘T base in WC geometry in the closed active site is available (PDB ID: 4PGX), Mn^{2+} ions were used to stabilize the closed active site in this structure (Koag et al., 2014). Our goal was to model the effect of the native, Mg^{2+}-bound active site on the wb-WC reaction. Therefore, we used the crystal structure with a cognate G∘C base pair and Mg^{2+} ions in the active site (PDB ID: 4KLF) (Freudenthal et al., 2013) as the initial structure for our pol-β model. dCTP was mutated to dTTP. X-ray structure of T7 DNA polymerase with a cognate G∘C base pair and Mg^{2+} ions in the active site and bound to thioredoxin (PDB ID: 1T7P) (Doublié et al., 1998) was used as the initial structure for our pol-T7 model. dC was mutated to dT and thioredoxin was excluded in the model. Missing protein residues in the DNA-pol models were added using Psfgen plugin in the VMD suite (Humphrey et al., 1996) and were partially optimized before the equilibration protocol. Both models were solvated in TIP3P water boxes (Jorgensen et al., 1983), maintaining at least 16 Åfrom the box edges to the solute.

A-site models. X-ray structure of *T.thermophilus* 70S ribosome bound to $tRNA^{Thr}$ in the A-site (PDB ID: 6GSK) was used as the initial structure for all A-site models. In this structure, U∘G base pair in the second codon (AUC) - anticodon (GGU) position is solved in the WC geometry (Rozov et al., 2018). The “closed” A-site model contained no manual changes in the decoding center, and the “ribosomal fingers” (rRNA residues A1492, A1493 and G530) were in *out* conformation, surrounding the codon-anticodon helix. As a proxy of the “open” state of the decoding center we created “abasic” model, in which we deleted nucleobases of A1492, A1493, and G530 residues, leaving only their sugar-phosphate backbone. Abasic model, in contrast to a more realistic “open” model, alleviated the need to use harmonic restraints on these residues, which would affect dipole moment fluctuations, important in one of our analyses. Harmonic restraints to create the open state of the decoding center were used only in QM/MM umbrella sampling simulations, as described in the corresponding paragraph. Preparation of the closed and abasic models of the decoding center was performed similarly to the other studies (Zeng et al., 2014). All residues within 35 Å radius of the center of mass (c.o.m.) of the second codon-anticodon base pair were selected for the model of the decoding site. Obtained spheres of 35 Å radius were solvated in 120 Å x 120 Å x 120 Å box of TIP3P water (Jorgensen et al., 1983). Na^+ and Cl^- ions were used to neutralize the system and create NaCl concentration of 0.15 M. The outer 7 Å shell of the solute in the A-site models was restrained in all subsequent simulations with a force constant of $30\,\text{kcal mol}^{-1}\text{Å}^{-2}$. The inner $28\,\text{Å}$ sphere was not restrained in the final production simulations, but was also restrained during the equilibration protocol, as described below.

All Mg^{2+} ions from the ribosome X-ray structure were excluded for two reasons: 1) Mg^{2+} have been shown to create artifacts when classical force fields parameters were used, thus usually requiring specialized or polarizable force fields for accurate modeling (Casalino et al., 2017, Sponer et al., 2018) and 2) in X-ray crystallography experiments, electron density peaks labeled as Mg^{2+} ions may often be in fact water molecules or Na^{+} ions (Zheng et al., 2015). While Mg^{2+} ions are undoubtedly important for ribosome structure and function (Nierhaus, 2014), their accurate modeling is out of scope in this study. Their complete exclusion may even improve the reliability of the models given that classical force fields are used, and the MD trajectories lengths are relatively short (Robbins and Wang, 2013). In contrast to the ribosome structures with a large number of modeled Mg^{2+} ions, DNA polymerases contain 1-3 well-defined and tightly bound Mg^{2+} ions, known to be essential for dNTP binding (Beard and Wilson, 2006). Therefore, Mg^{2+} were retained in all models of DNA polymerases.

The open A-site model in the US simulations (see below) was prepared as following. We used torsional harmonic restraints on A1492 and A1493 as in the study by Zeng et al. (2014). By moving these harmonic restraints with a force constant of $0.04\,\text{kcal mol}^{-1}\text{Å}^{-2}$, A1492 and A1493 changed their conformation from *out* to *in* and intercalated into h44 rRNA helix, characteristic of the open A-site state (Ogle et al., 2002, Zeng et al., 2014). As the initial coordinates for the open A-site model we used frames from the end part of the US trajectories of the closed A-site model. Visual inspection of the trajectories revealed that the closed→open transition happened on the timescale of appr. 10 ps. In all models, during equilibration and all classical MD simulations, the studied G∘U(T) base pair was maintained in the WC geometry by using G* parameters from CHARMM36 force field (Xu et al., 2016). These parameters did not affect the energies obtained in the QM/MM calculations, as the base pair was modeled with PM7 level instead.

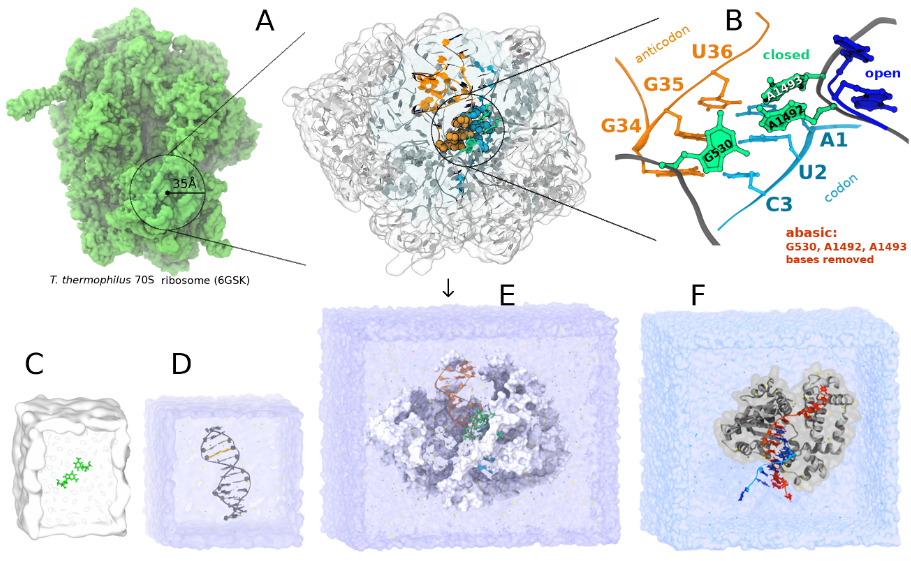


Figure 3.2.: Molecular systems used in MD simulations. **A** – approximate size and position of the A-site model (right) in the 70S ribosome structure (left). **B** – A-site models differed by the presence and conformation of the conserved rRNA residues G530, A1492 and A1493. The closed (native) model contained no manual changes in these residues (green); in the open model (used only in US calculations) A1492 and A1493 were in the *in* (open) conformation (blue); in the abasic model nucleobases of all three rRNA residues were deleted. **C-D** – visualization of the final systems used in MD simulations and QM/MM US calculations: single G∘T base pair in benzene **C**; DNA duplex in water (dodecamer is shown, but heptamer was used for QM/MM US calculations) **D**; A-site model **E** and DNA polymerase (only pol-β is shown) **F**. Scale in this depiction only approximate.

3.1.3 MD simulations

NAMD 2.12 package was used for all MD simulations (Phillips et al., 2020). CHARMM36 force field was used for MM part in all MD simulations (Best et al., 2012, Denning et al., 2011, Huang and Mackerell, 2013). Periodic boundary conditions were used in all MD simulations. Particle Mesh Ewald method (Darden et al., 1993) was applied to treat electrostatic interactions and a cutoff of 12 Å was used for the van der Waals interactions. Solvated and neutralized models were subjected to 1,000 steps of steepest-descent optimization of water and ions coordinates while the rest of the structure was restrained with a force constant of $50\,\mathrm{kcal/mol/\mathring{A}^2}$. Solvent and ions, as well as cell volume were equilibrated with 1-2 ns of NPT simulations with 1 fs time step at standard conditions using Langevin thermostat and barostat, maintaining the same restraints on the solute. The obtained coordinates were used for 10,000 steps of steepest-descent optimization without any restraints apart from the outer shell in the A-site models. The optimized coordinates were used for 400 ps of gradual heating of the systems to 298 K with 1 K increment every 400 fs. Classical production simulations were conducted in NVT ensemble using Langevin thermostat at standard conditions with 2 fs time step. SETTLE algorithm (Miyamoto and Kollman, 1992) was used for rigid bonds in water while SHAKE/RATTLE algorithm (Andersen, 1983) was used for rigid H-containing bonds in other molecules.

3.1.4 Estimation of dielectric constant

We employed Kirkwood-Fröhlich formula (KFF) to calculate the static dielectric constant ε of the molecular environments surrounding G∘U(T) base pairs. KFF relates ε to dipole moment M fluctuations in a given volume (Kolafa and Viererblová, 2014, Pitera et al., 2001). We used KFF for the case of surrounding permittivity $\varepsilon_{RF} = \varepsilon$ (Yang et al., 1995):

$$\varepsilon = \frac{3\alpha + 1 + \sqrt{9\alpha^2 + 6\alpha + 9}}{4} \tag{3.1}$$

where

$$\alpha = \frac{\langle M^2 \rangle - \langle M \rangle^2}{3\varepsilon_0 V k_B T} \tag{3.2}$$

where ε_0 – dielectric constant of vacuum, k_BT – thermal energy, and V is the volume of the probed region.

For each studied system, at least two replicas of at least 75 ns of classical MD trajectories were collected. ε was measured in spheres of 5, 7, 9, and 12 Å radii centered at the base pairs of interest. For the DNA polymerases, we only focused on the G∘T base pair in the active site. For the A-site models, we calculated ε in spheres surrounding each codon-anticodon base pair separately. The same replicas of MD trajectories were used for the calculations at four different radii at the three codon-anticodon positions. First 20 ns of each trajectory were excluded from the analysis. The general algorithm consisted of the following steps:

1. The center of the probed region was selected as C2 atom of the codon nucleobase at the given base pair (the template nucleobase for the DNA polymerases);

2. For a given cutoff radius, a selected trajectory was analyzed for the number of protein and RNA (DNA) residues present in the cutoff sphere at each MD frame; the largest set of residues was selected for the calculations, where this set was kept constant;

3. For a given cutoff radius, a selected trajectory was analyzed for the number of water molecules in the probed region. The mean number was selected for the calculations; this number was kept constant during the calculations by slight changes of the cutoff radius for water at each MD frame until the water selection converged to the needed number of water molecules; variation of the cutoff was below 1.5 Å therefore, the cutoff radii presented on the figures reflect only the mean radii of water selection spheres;

4. The total selection at each frame consisted of a set of protein and RNA (DNA) residues, and a converged set of selected water molecules. The selection did not include ions, as it has been revealed previously that the dynamic contribution from ions to the static dielectric constant is relatively small and can be neglected (Chandra, 2000). For this total selection, two properties were calculated: volume and dipole moment. Volume was calculated using VMD (Humphrey et al., 1996) as a density map with 1.0 Å resolution; due to computational cost limitations, the density grids were calculated every 4 ns. Dipole moment was calculated by VMD (Humphrey et al.,

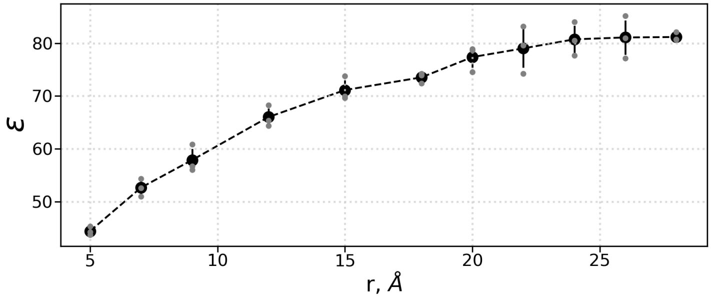


Figure 3.3.: Dependence of ε calculated in our KFF-based approach applied to the box of 42,000 TIP3P water molecules. Black curve shows the mean values from 3 replicas, while gray circles denote individual replicas.

1996) using the c.o.m. of the selection as the reference point.

5. To obtain the scalar volume values, the density grids were integrated using 1.0 isovalue. Low fluctuations of the volume values of a given trajectory were confirmed, and the mean volume was used for ε calculations using Eq. (3.1).

To assess the validity of our approach, we performed a benchmark calculation using a box of 42,000 TIP3P water molecules. Three replicas of approximately 30 ns were collected using classical MD in NVT ensemble at standard conditions. The described above approach was applied to this benchmark system. The center of mass of the box was used as the center of the probe spheres. The result of this benchmark calculation is shown on Fig. 3.3.

As the figure demonstrates, ε calculated in our approach displays a size-dependence, similarly to the previous studies (Gereben and Pusztai, 2011): ε of TIP3P water converges to its experimental value of $\sim$80 only at approximately 20 Å radius of the probe sphere. Therefore, our approach does not allow for quantitative ε calculations of the relatively small regions of the decoding and active sites around the studied base pairs. In this study, we restrict to qualitative comparison of the closed and abasic models of the A-site, and to pol-β and T7-pol DNA polymerases. Fig. 3.4 shows the cumulative mean square dipole moment fluctuation $\langle M^2\rangle - \langle M\rangle^2$ in all MD trajectories. As the figure demonstrates, most trajectories converged to relatively constant fluctuation levels, justifying the use of KFF.

3.1.5 Umbrella sampling

Selection of the collective variables. wb-WC reaction involves slow motions of heavy atoms (geometry change) and fast proton transfers (PT) (Fig. 1.7), making it impossible to describe with a 1D collective variable (CV). Therefore, the geometry change and PTs were described with two separate CVs. To describe the geometry change, we used the path collective variable (pathCV). PathCV requires a set of structures (images) describing the process and used as a reference (Branduardi et al., 2007). In pathCV, the position of a given coordinate frame on the path (s) is calculated as:

$$s = \frac{\sum_{i=1}^{N} i \exp(-\lambda R[X - X_i])}{\sum_{i=1}^{N} \exp(-\lambda R[X - X_i])} \tag{3.3}$$

and the distance from the path (z):

$$z = -\frac{1}{\lambda} \ln \left[\sum_{i=1}^{N} \exp(-\lambda R[X - X_i]) \right] \tag{3.4}$$

where $R[X - X_i]$ is a distance metric, describing the distance from a given frame to the path image i. As a distance metric, we used RMSD as implemented in colvars module of NAMD 2.12 (Fiorin et al., 2013). λ is a parameter that can be tuned for optimal performance of pathCV. We used λ value of 300 throughout all pathCV calculations.
To obtain the reference path, we optimized the full wb-WC reaction in G∘U *in vacuo* using nudged elastic band method (NEB). We used NEB-TS implementation in Orca 4.2.1 (Neese, 2018), which is a combination of climbing-image NEB (CI-NEB) (Henkelman et al., 2000) and eigenvalue-following optimization of a TS guess. NEB-TS was performed on B3LYP-D3/def2-TZVP level of theory using the default spring force constant of 0.1 Eh Bohr^{-2} and tight criteria for SCF convergence. The optimized path contained 34 frames. It's potential energy profile is shown on Fig. 3.5.
To create the set of images for pathCV, the double-proton-transfer (DPT) part of the NEB path (frames 26 to 34) was excluded, as it did not contain the geometry changes.

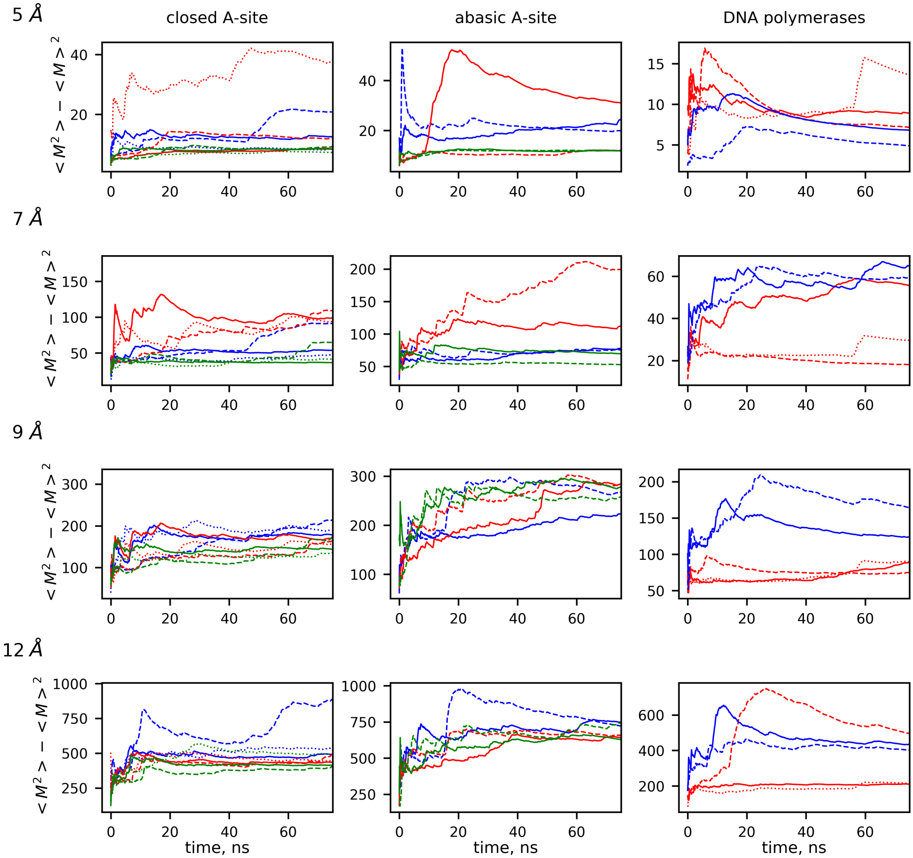


Figure 3.4.: Cumulative mean square dipole moment fluctuation in all analyzed MD trajectories. Each row represents a distance cutoff, while each column represents a studied system. For the A-site models, fluctuations of the dipole moment around three codon-anticodon positions are shown on each plot as blue, red and green curves denoting the first, second and third codon-anticodon position, respectively. Solid, dashed and dotted lines denote different replicas. For the DNA polymerases, red and blue curves denote pol-β and T7-pol, respectively.

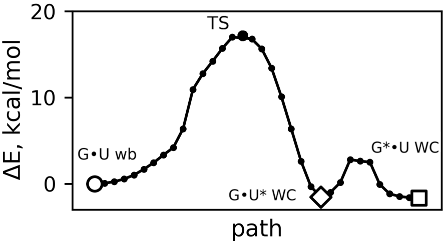


Figure 3.5.: Reference energy profile of the wb-WC reaction from NEB calculations.

Only ring atoms of the nucleobases were included into pathCV calculations. PTs in the wb-WC reaction were described as a distance difference hb = d(O6-H3) – d(N1-H1). Characterization of the wb-WC reaction in 2D CV space (s; hb) allowed to clearly distinguish three minima and TS on the reference NEB path (Fig. 3.8A), thus these two CV were selected for the US simulations. Although the previous computational studies revealed only G∘U wb → G∘U* WC path in the wb-WC reaction (Brovarets and Hovorun, 2015), the possibility of the alternative path G∘U wb → G*∘U WC, or a bifurcation leading to both products could not be excluded. To verify the TS from the NEB calculations, and to exclude at least the post-TS bifurcation, we performed committor analysis in gas phase on BLYP-D3/def2-SVP. Approximately 50 Born-Oppenheimer MD simulations were started from the TS with randomly initialized velocities matching 298 K using Berendsen thermostat with 2 fs period. Simulations were performed for 200 fs with 1 fs time step. Committor analysis revealed roughly equal partition between reactant (wb) and product (G∘U* WC), suggesting validity of the TS (Fig. 3.8A). No trajectories led to G∘U* WC, which allows to exclude the possibility of a post-TS bifurcation in this reaction (Fig. 3.8A). Therefore, the selected reference path accurately describes the wb-WC reaction.

Setup of the US calculations. To calculate potential of mean force (PMF) of the wb-WC reaction in the selected models, we applied umbrella sampling (US) simulations. We used 34 frames from the NEB path as the initial base pair coordinates of the US windows. For each studied system, a frame from the end part of the corresponding classical MD trajectory was selected as the initial system coordinates for US simulations. To prepare the initial US windows, coordinates of the studied G∘U(T) base pair in the initial system were changed to the pre-aligned coordinates of each NEB frame. Thus, 34 initial US windows in each simulated system differed only by the base pair coordinates. z was not used as a CV for PMF calculation. Instead, in all US simulations, a "half-harmonic" boundary potential was added at z value of 0.15 with a force constant of $100\,\mathrm{kcal/mol/Å}^2$. This prevented simulations from visiting largely out-of-plane or shifted base pair conformations of the base pair, where s would not be well-defined. At the same time, using the boundary potential instead of restraining z at 0 allowed us to ignore z in WHAM calculations. To improve sampling in the vicinity of minima while still covering the full wb-WC path, we used two layers of US windows with different force constants applied to the chosen CVs s and hb. The more "rigid" layer with the force constant of $50\,\mathrm{kcal/mol/Å}^2$ applied to both s and hb covered all US frames in the TS region and selected frames in the vicinity of minima. The more "flexible" layer with the force constant of $5\,\mathrm{kcal/mol/Å}^2$ only covered frames in the vicinity of the minima. The two layers together contained 48 US windows per each system.

QM/MM US simulations. All US simulations were performed in hybrid quantum-mechanical/ molecular-mechanical (QM/MM) scheme. PM7 (Stewart, 2013) in MOPAC-2016 (Stewart, 2016) was used for the QM region and CHARMM36 force field (Best et al., 2012, Denning et al., 2011, Huang and Mackerell, 2013) in NAMD2.12 (Melo et al., 2018) was used for the MM region. In all systems, the QM region comprised only the nucleobases in the base pair of interest (26-29 atoms), with the QM/MM interface placed at the glycosidic bonds. The hydrogen link-atom approach with charge shift scheme was used to treat the link atoms. Tight SCF convergence criteria ($10^{-8}\,\mathrm{kcal/mol}$) were used in PM7 calculations in MOPAC. It is worth emphasizing that in the MOPAC2016/ NAMD inteface, NAMD reads heats of formation from MOPAC (Melo et al., 2018). Each US window was minimized with 100 steps of steepest-descent before the start of the production simulations. All US simulations were performed in NVT ensemble using Langevin thermostat at standard conditions. Integration step was 0.2 fs, and trajectories were collected every 40 fs. First 12 ps of the US trajectories were discarded from the analysis. Each US window was simulated for at least 100 ps up to 400 ps, resulting in the cumulative 120 ns of QM/MM simulations.

PMF calculations. wham-2D code by Alan Grossfield was used for all WHAM calculations. US trajectories were divided into batches of approximately 20 ps, and PMF was calculated from each batch separately to monitor the convergence. The grid for WHAM calculations was 0:1 with 0.025 step for s and -1.8:1.8 with 0.05 Å step for hb. Empty bins were filtered out. The grid was divided into regions corresponding to wb, G∘U* WC and G*∘U WC basins. ΔG of each of the three states was calculated as local PMF minima in each basin. For each batch, ΔG_{wc} was calculated as ΔG(G*∘U WC) - ΔG(G∘U wb). Visual inspection of the ΔG_{wc} convergence allowed to arbitrary assign converged regions of the trajectories (Fig. 3.9). These regions were then treated as a single batch in each trajectory, PMFs from which are shown on Fig. 3.8. MEPSA was used to calculate a minimal free energy paths from PMF (Marcos-Alcalde et al., 2015).

3.2 Results and Discussion

3.2.1 Dielectric constant calculations

It was speculated previously that closing of the A-site desolvates the codon-anticodon helix, increasing base-pairing selectivity by energy penalty from the lost H-bonds with water in mismatches (Ogle et al., 2002, Satpati and Åqvist, 2014). We hypothesized that desolvation might also bring an opposite contribution to the accuracy of G∘U mismatch recognition. Previous computational studies of the wb-WC reaction (Brovarets and Hovorun, 2009, 2015, Li et al., 2020, Nomura et al., 2013) as well as our benchmark calculations (Fig. 3.6) demonstrate that this reaction is exoergic in gas phase and very nonpolar implicit solvents. We reasoned that a potentially decreased polarity of the closed ribosomal A-site may be responsible for the stabilization of the WC geometry of G∘U observed in the structural studies.

We applied Kirkwood-Fröhlich formula (KFF) to qualitatively compare dielectric constant ε of the decoding site environments between the closed and abasic models. We observed a consistent ε decrease in the closed model for all codon-anticodon positions

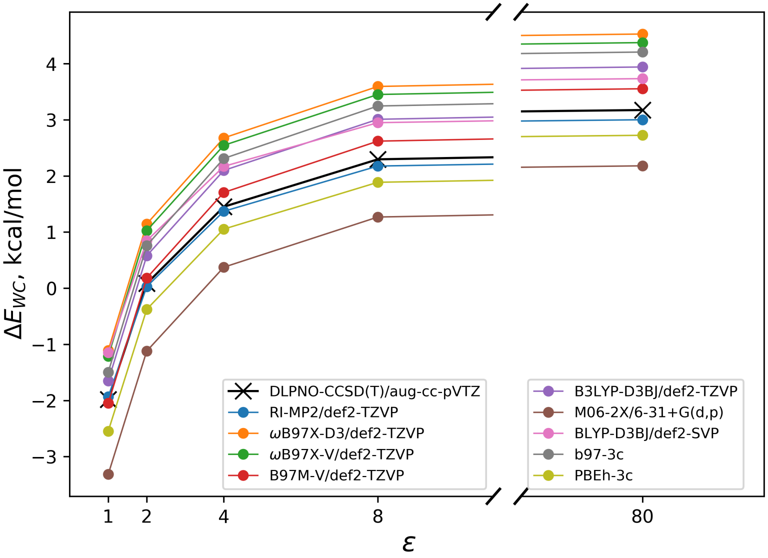


Figure 3.6.: ΔE_{wc} as a function of dielectric constant ε of the implicit solvent model. Each level of theory is denoted by different colors. The reference level of theory (DLPNO-CCSD(T)/aug-cc-pVTZ) is shown in black.

(Fig. 3.7). However, only at some distance cutoffs and positions the difference was statistically significant, which can also be explained by the limited length and replica numbers of the MD trajectories Fig. 3.4. From our KFF calculations we conclude that the closing of the A-site indeed decreases ε of the environment of all three codon-anticodon positions. A more detailed approach would be needed to obtain quantitative information on the dielectric environment of the codon-anticodon helix.

We also compared the active sites of pol-β and T7-pol using the same approach. The active site environment in pol-β was generally less polar compared to T7-pol, but the difference was statistically significant only at 12 Å (Fig. 3.7). In sum, our KFF calculations revealed decreased polarity in the closed compared to the abasic A-site model, and in pol-β compared to T7-pol. Although these effects might indicate a putative environmental contribution on the energetics of the wb-WC reaction, we cannot estimate the extent of this contribution on the wb-WC equilibrium. Therefore, free energy calculations of the wb-WC reaction with modeled environments of the decoding/active sites are needed.

3.2.2 PMF of the wb-WC reaction from QM/MM US calculations

We calculated potential of mean force (PMF) of the wb-WC reaction in the studied molecular environments using QM/MM umbrella sampling (US) calculations. The PMF from the converged part (see below) of the US trajectories are shown on Fig. 3.8. The only quantitative property we derived from the PMFs was the total free energy change of the wb-WC reaction $\Delta G_{wc} = \Delta G$(G*∘U WC) - ΔG(G∘U wb). ΔG_{wc} was used to evaluate the convergence of the US simulations by calculating PMF separately for consecutive batches of $\sim$ 20 ps.

Benchmark systems. In order to verify the validity of our approach, we first applied

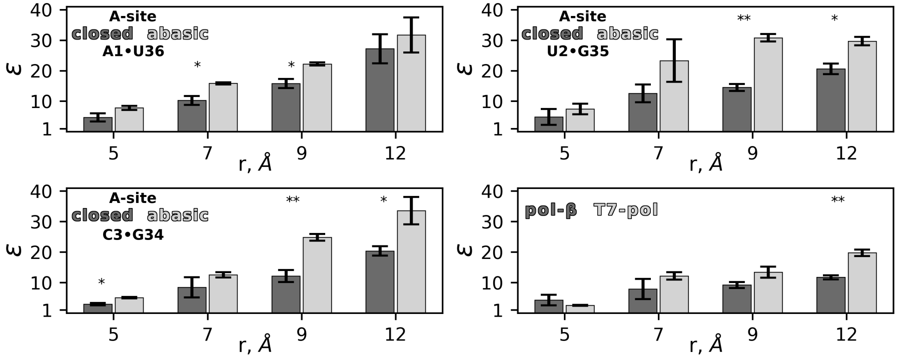


Figure 3.7.: Dielectric constant ε of the base pair surroundings. **A**–**C** – ε difference between closed and abasic A-site models around each codon-anticodon base pair (A1-U36, U2-G35 and C3-G34) measured in spheres of radius r. **D** – ε difference between pol-β and T7-pol around the G∘T base pair in the active site. The bars denote mean and standard deviation of at least three replica per model. Single and double asterisk denotes significant differences with $P < 0.05$ and $P < 0.01$ (Student's t-test), respectively.

Table 3.3.: ΔG_{wc} values from the US simulations

System	ΔG_{wc} [a]
DNA	7.2 ± 1.1
benzene	-0.5 ± 0.6
A-site closed	-0.7 ± 0.6
A-site abasic	11.2 ± 0.9
A-site open	6.6 ± 0.3
pol-β	-4.4 ± 1.6
T7-pol	7.2 ± 0.4

[a] Mean ΔG_{wc} ± standard deviation from the regions selected for convergence estimation on Fig. 3.9, kcal/mol;

it to benchmark systems: G∘T base pair in the DNA heptamer duplex in water and a single G∘T base pair in benzene. Fig. 3.9B shows ΔG_{wc} trajectory for the benchmark systems, and Table 3.3 shows the converged values of ΔG_{wc}. After initial fluctuations from 4 to 10 kcal/mol, ΔG_{wc} of the DNA model converged to 7.2 ± 1.1 kcal/mol. Our calculations overestimate experimental ΔG_{wc} range of 3.3-4.9 kcal/mol (Kimsey et al., 2018), and are consistent with the recent computational result of $\sim$ 6 kcal/mol (Li et al., 2020), although in the study by Li et al. (2020) G∘U* configuration had lower energy than G*∘U. To assess the ability of our QM/MM setup to reproduce ΔG_{wc} dependence on ε observed in implicit solvent models, we applied it to a single G∘T base pair in benzene (ε = 2.3). ΔG_{wc} in benzene converged to -0.5 ± 0.6 kcal/mol, which is consistent with the QM calculations in implicit solvent model of similar ε (Fig. 3.6). In both benchmark systems the positions of the minima on the PMF were not qualitatively altered compared to the reference reaction path (Fig. 3.8). From our benchmark calculations we conclude that our setup is able to faithfully estimate environmental effects on the wb-WC reaction, revealing satisfactory agreement with experiments and higher levels of QM theory.

Ribosome A-site. We started the QM/MM US simulations of the A-site effects on the wb-WC reaction in the second codon-anticodon UG base pair in the models of closed and abasic A-site. ΔG_{wc} in the closed A-site converged to -0.7 ± 0.6 kcal/mol, while in the abasic A-site it converged to 11.2 ± 0.9 kcal/mol (Fig. 3.9C, Table 3.3).

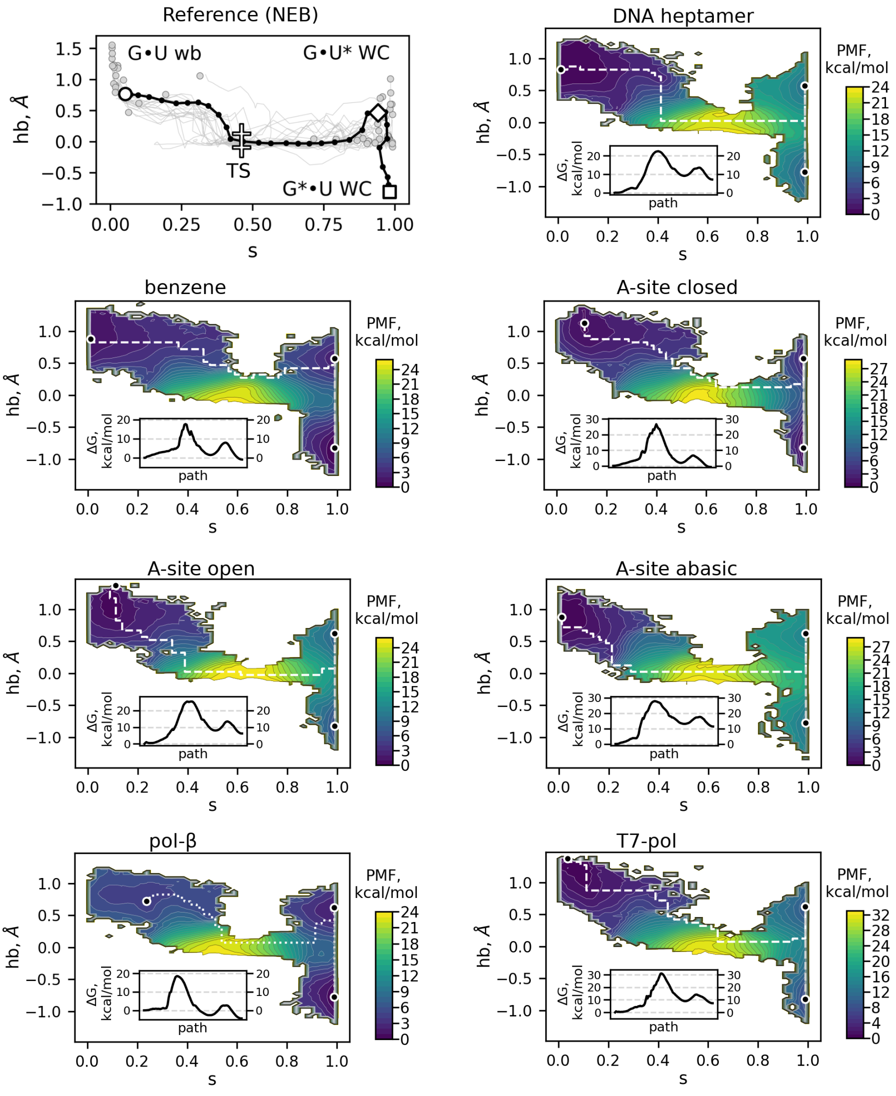


Figure 3.8.: Converged PMF from the US calculations. **A** – The reference wb-WC path derived from NEB calculations. **B–D** – US-derived PMF of the wb-WC reaction in all systems. Black circles denote the local minima of the wb-WC reaction. White dashed line denotes the minimal free energy path of the wb-WC reaction, from G○T wb to G*○T WC. The inset shows the free energy profile along the minimal free energy path.

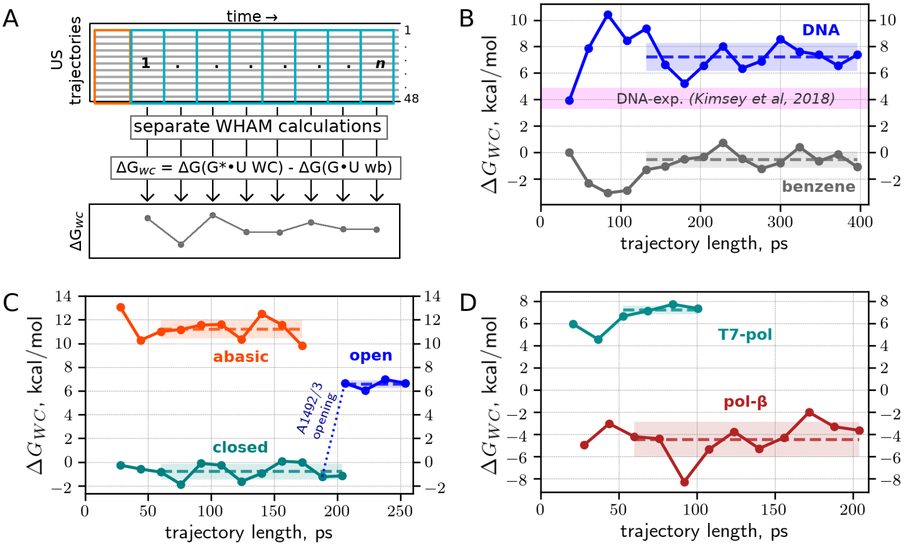


Figure 3.9.: ΔG_{wc} convergence in the US calculations. **A** – Scheme of the convergence estimation in the US calculations. **B–D** – convergence of the ΔG_{wc} in benchmark systems **B**, A-site models with G∘U in the middle position **C** and DNA polymerases **D**. For each ΔG_{wc} trajectory, the shaded area and the dashed line of the corresponding color denote the standard deviation and mean of the set of batches selected for convergence evaluation. Magenta-shaded area in **B** denotes the range of ΔG_{wc} in DNA duplexes in solution measured with NMR by Kimsey et al. (2018).

However, the abasic model is only a rough approximation of the open state of the A-site. To estimate effects of the ribosomal fingers A1492 and A1493 more accurately, we created the "open" A-site model. In this model, harmonic restraints used in the previous studies (Zeng et al., 2014) were applied to change A1492 and A1493 conformations from *out* to *in* (Fig. 3.2B). G530 remained in the closed-like conformation in the open model. These restraints were applied to the coordinates from the end part of the closed state US trajectories and maintained for approximately 70 ps in each US window. ΔG_{wc} in the open A-site converged to 6.6 ± 0.3 kcal/mol (Fig. 3.9C, Table 3.3). While in the abasic model the position of the wb minimum was not qualitatively altered compared to the reference reaction path ($s = 0$), in the closed and open models it shifted to higher values of the s variable ($s \approx 0.1$), suggesting a slight destabilization effect on the wobble geometry, likely exerted by G530 (Fig. 3.8). This effect was not enough to shift the wb-WC equilibrium towards WC in the open model, indicating a critical role of A1492 and A1493 residues.

The model of the ribosomal A-site used in our simulations is of course a simplification of the real conditions that may affect the wb-WC properties during the decoding. The major simplifications are: the model is only a part of the ribosome structure, with an outer shell that was restrained during all simulations; absent Mg^{2+} ions in the model; the "open" model differed from the "closed" model only by conformations of A1492 and A1493 nucleotides. However, we argue that this model is able to capture the major effects of the open$\rightarrow$ closed transition on the wb-WC equilibrium. Distant structural elements and their change during the 30S domain closure are unlikely to dramatically affect the wb-WC reaction, and no Mg^{2+} ions have been reported to directly interact with the second codon-anticodon position to the best of our knowledge. The role of G530, which was not affected in the open model, is however not explored in our study. It is remained to be investigated how the wb-WC reaction is affected in the first and third codon-anticodon positions. It would also be interesting to study how WC-leading tautomerization reactions in other mismatches, predicted by Hovorun in coworkers (Brovarets and Hovorun, 2015, Brovarets' and Hovorun, 2015, Brovarets and Hovorun, 2015), are affected by the ribosome.

DNA polymerases. The previous QM/MM study have addressed effects of the DNA polymerase λ environment on the wb-WC reaction (Li et al., 2020). Here, we analyze effects of two other DNA polymerases: low fidelity human DNA polymerase β (pol-β) and high-fidelity DNA polymerase from T7 virus (T7-pol). In both models, the wb-WC reaction was simulated in the G∘T base pair in the closed state of the polymerase active site. ΔG_{wc} in pol-β converged to -4.4 ± 1.6 kcal/mol, indicating largely exoergic wb-WC reaction. This result corroborates previous structural studies, revealing WC-like G∘T base pair in the closed active site of pol-β (Koag et al., 2014). At the same time, ΔG_{wc} in T7-pol converged to 7.2 ± 0.4 kcal/mol, resembling ΔG_{wc} value in the DNA duplex (Fig. 3.9B,D, Table 3.3). To the best of our knowledge, WC-like G∘T base pairs were never observed in the active site of T7-pol. Our results are in line with the absence of such findings. Comparing PMFs from pol-β and T7-pol revealed a striking shift in the wb minimum position towards higher ($s \approx 0.2$) values of path variable s in pol-β (Fig. 3.8). We attribute this shift to steric effects in pol-β, that constrain a base pair geometry in the closed active site.

In sum, using QM/MM US calculations, we demonstrated the exoergic wb-WC reaction in G○U at the middle codon-anticodon position in the closed A-site model, while it was endoergic in the open and abasic models. Decreased ε in the closed state of the decoding site can be one of the contributions to this effect. Shifts in the wb minimum position on the PMF in the closed and open models might indicate steric constraints on the base pair geometry. A similar effect was observed in pol-β active site, in which the wb-WC was highly exoergic, in contrast to T7-pol, where the reaction equilibrium was not affected compared to the DNA duplex. More detailed studies are needed to delineate the environmental contributions on the properties of the wb-WC reaction.

Chapter 4

Effects of the tautomerization reaction on decoding

Parts of this Chapter are included in the publication (Kazantsev and Ignatova, 2020)

Results in the previous chapter revealed the exoergic wb-WC reaction in the closed A-site at the middle codon-anticodon position, in agreement with the structural studies (Demeshkina et al., 2012, 2013, Loveland et al., 2017, Rozov et al., 2015, 2018). Interpretation of these structural studies by their authors, and the model of codon-anticodon selection they proposed (Demeshkina et al., 2012, 2013, Rozov et al., 2015, 2018), are not consistent with the exoergic wb-WC reaction (and with the structural observations themselves), and should be reconsidered (see also Section 1.2.3). Instead of the proposed "tautomerization energy penalty", the model should include the tautomerization reaction explicitly, as it apparently proceeds *during* the codon-anticodon decoding.

In this chapter I develop a model of initial selection in codon-anticodon decoding which incorporates the wb-WC reaction. I derive the analytical solutions of this new model, compare them to numerical solutions, and demonstrate how the wb-WC reactions affects the error rate of decoding. I discuss relation of the new model to the classical induced-fit model, and show the generality of the new model.

4.1 Methods

4.1.1 Selecting the rate constants for the kinetic model

To the best of our knowledge, no set of experimental rate constants of decoding in translation is available for a codon-anticodon combination with a G◦U mismatch. Therefore, we selected the set of rate constants corresponding to an A◦C mismatch in the first codon-anticodon positions from Rudorf et al. (2014). The rate constants are shown in Table 4.1. C2↔C3 transition (codon reading) is very rapid, which prevents estimation of its rate constants (Rodnina et al., 2017). Therefore, we used arbitrary rate constants for this transition, namely k_2, q_3^c and q_3^{nc}. To maintain at least some consistency with the previous studies, we calculated these rate constants from a free energy diagram in Pavlov and Ehrenberg (2018). However, the low free energy barriers in this diagram results in extremely high rate constants ($\sim 10^9$ s^{-1}) that would prohibit the use of numerical calculations. Therefore, we uniformly scaled these rate constants down to values acceptable for numerical calculations, while keeping the values high enough to not have significant rate-limiting effects (Table 4.1).

All equilibrium populations were calculated from the free energy change ΔG according to Boltzmann population at standard conditions:

$$P_{eq} = \frac{\exp \frac{-\Delta G}{RT}}{\exp \frac{-\Delta G}{RT} + 1} \tag{4.1}$$

where RT is the thermal energy.

Rate constants were calculated from the free energy of activation $\Delta G_{\ddagger}$ according to Eyring equation with transmission coefficient 1:

$$k = \frac{k_B T}{h} \exp \frac{-\Delta G_{\ddagger}}{RT} \tag{4.2}$$

where k_B is Boltzmann's constant and h is Planck's constant.

4.1.2 Numerical calculations

Numerical solutions of the kinetic systems of decoding were obtained by numerical integration of ordinary differential equations (ODE) in Python 3.6. The same rate constants were used for both analytical solutions and numerical calculations. As the initial conditions, we used $[R_1] = 50\mu M$, $[T_c] = 1\mu M$, $[T_{nc}] = 1\mu M$, and zero concentrations for the rest of the states. A close to constant concentrations of T_c and T_{nc} were maintained by using rapid zero-order formation reactions and first-order degradation reactions, rate constants of which were set to result in the desired concentrations of T_c and T_{nc}. Rapid equilibrium approximation for the wb-WC reaction in states C2 and C3 was maintained by using very low $\Delta G_{\ddagger}$ in these states (6 – 8 kcal/mol). Numerical

Table 4.1.: Values of the decoding rate constants used for kinetic modeling

Designation [a]	value	unit	condition [b]	reference
k_1	140	μM^{-1} s^{-1}	20 °C, HiFi	Rudorf et al. (2014)
q_2	85	s^{-1}	20 °C, HiFi	Rudorf et al. (2014)
k_2	720 [c]	s^{-1}	arbitrary [c]	Pavlov and Ehrenberg (2018)
q_3^c	25 [c]	s^{-1}	arbitrary [c]	Pavlov and Ehrenberg (2018)
q_3^{nc}	3900 [c]	s^{-1}	arbitrary [c]	Pavlov and Ehrenberg (2018)
k_3	180	s^{-1}	20 °C, HiFi	Rudorf et al. (2014)
q_4^c	0.2	s^{-1}	20 °C, HiFi	Rudorf et al. (2014)
q_4^{nc}	140	s^{-1}	20 °C, HiFi	Rudorf et al. (2014)
k_4^c	190	s^{-1}	20 °C, HiFi	Rudorf et al. (2014)
k_4^{nc}	0.6	s^{-1}	20 °C, HiFi	Rudorf et al. (2014)

[a] Designations for the rate constants used in our study, as well as in Pavlov and Ehrenberg (2018). Designations in Rudorf et al. (2014) can be different;
[b] Experimental conditions at which the rate constants were measured. HiFi – high-fidelity conditions (3.5 mM Mg^{2+}, 0.5 mM spermidine, and 8 mM putrescine) (Gromadski et al., 2006);
[c] The values were obtained from the free energy diagram in Pavlov and Ehrenberg (2018), and uniformly rescaled to obtain values eligible for numerical calculations using ODE.

integration at each point in the space of rate constants was performed for 800 s with 2 μs step ($4 \cdot 10^8$ steps). The concentrations from the last step were used as steady-state concentrations. $\eta(ODE)$ was calculated as $\frac{[P_W]}{[P_R]}$. $P_{wc}(ODE)$ was calculated as $\frac{[C4_{nc}^{WC}]}{[C4_{nc}^{WC}]+[C4_{nc}^{wb}]}$.

4.2 Results

4.2.1 Approximations and analytical solutions

Let us revisit the classical kinetic model of initial selection (Fig. 1.5) and the error rate in this model:

$$\eta_0 = \frac{R^{nc}}{R^c} = \frac{(k_{cat}/K_m)^{nc}}{(k_{cat}/K_m)^c} = \frac{k_4^{nc}[C4_{nc}]}{k_4^c[C4_c]} \qquad \text{(1.1 revisited)}$$

For the case of G∘U mismatch recognition, the near-cognate (nc) branch includes both wobble and WC geometries of the mismatch, which interconvert via the wb-WC reaction. From the consideration of geometric selection, these two geometries would be associated with different decoding rate constants. To introduce the wb-WC reaction into the kinetic model of decoding, the nc branch must be separated into wb and WC branches. We performed this separation, leaving the cognate (c) branch unchanged (Fig. 4.1A). Rate constants between nc-wb and nc-WC states correspond to the rate constants of the wb-WC reaction in a given environment of the A-site. To simplify the model for the sake of deriving analytical solutions, we assumed an instant wb-shifted wb-WC equilibrium in the states $C2$ (initial binding) and $C3$ (codon recognition in the open A-site). This approximation is well justified if, as suggested by our US calculations (Table 3.3), the open state of the decoding site does not shift the equilibrium towards WC, and thus the equilibration time is negligible. In this approximation, the wb-WC equilibrium in $C2$ does not affect the model and needed only for model completeness. The wb-WC reaction in $C4$ (codon recognition in the closed A-site) was modeled explicitly via its forward and reverse rate constants k_f, k_r. Inspired by the previously applied numerical kinetic model of selection in replication by Kimsey et al. (2018), we made the following assumptions: (i) nc-WC states are characterised with the same decoding rate constants as the cognate states, since the WC geometries are assumed to be indistinguishable for the decoding site; (ii) GTPase activation rate constant (k_4) in $C4_{nc}^{wb}$ state is set to zero, assuming that the major contribution to this reaction rate comes from the nc-WC state, thus the contribution from nc-wb can be ignored; (iii) escape rate constants of the nc-wb states are taken from the classical nc states (q_i^{nc}), assuming that the cumulative escape rate is dominated by the nc-wb states, thus the nc→nc-wb rescaling can be neglected. Given the assumption (ii), the apparent near-cognate k_4:

$$k_4^{nc} = P_{wc} k_4^c \qquad (4.3)$$

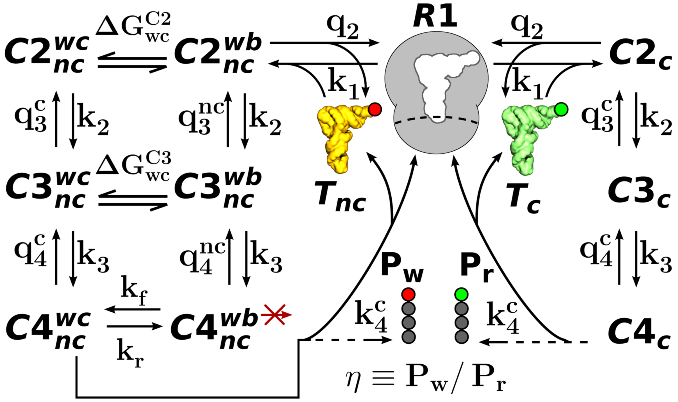


Figure 4.1.: Kinetic model of initial selection with the wb-WC reaction. See the text and Fig. 1.5 for the explanation of the states and transitions.

where P_{wc} is the population of the $C4^{wc}_{nc}$ state – a function of the equilibrium WC populations in $C3$ and $C4$ states, and constraints imposed by the decoding rates on the reaction kinetics in $C4$. Using the equation for product concentration in a first-order reversible reaction, we derived (see Derivation of Eq. (4.4)) the equation for the error rate induced by the wb-WC reaction:

$$\eta = \frac{[C4^{wc}_{nc}]+[C4^{wb}_{nc}]}{[C4_c]}\left(P^{eq}_{WC} + (P^{C3}_{WC}\frac{q^{nc}_4}{k^c_4+q^c_4} - P^{eq}_{WC})\exp(-\frac{k_f+k_r}{k^c_4+q^c_4})\right) \tag{4.4}$$

where P^{eq}_{WC} is the equilibrium WC population in $C4$ for a given (k_f, k_r), and P^{C3}_{WC} is the equilibrium WC population in $C3$. For convenience, the variables P^{eq}_{WC}, P^{C3}_{WC} and k_f are expressed below in terms of free energy differences ΔG^{C4}_{wc}, ΔG^{C3}_{wc} and $\Delta G_{\ddagger}$ (activation free energy), respectively, using Eq. (4.1) and Eq. (4.2) at standard conditions.

To analyze the kinetic model, some numerical values of the decoding rate constants must be used. This set of values must correspond to experimental measurements to obtain meaningful solutions of the model. The set of experimental decoding rate constants of the near-cognate branch used in the kinetic modeling below does not correspond to a G∘U mismatch, but to an A∘C mismatch instead (see Section 4.1.1). This

limitation and other approximations preclude quantitative predictions of the error rate from our model. Instead, we address more general effects of an out-of-equilibrium reaction in a substrate on the decoding process, with consideration of some known experimental parameters of the decoding and wb-WC kinetics.
First, we analyzed how ΔG_{wc}^{C3} and $\Delta G_{\ddagger}$ of the exoergic ($\Delta G_{wc}^{C4} = -1$ kcal/mol) wb-WC reaction affect η predicted from Eq. (4.4) (Fig. 4.2A). For the range of ΔG_{wc} and $\Delta G_{\ddagger}$ in RNA duplexes in solution reported in (Kimsey et al., 2018), η overlaps with the *in vitro* error range of G$\circ$U mismatches ($\sim 10^{-2}$ – 10^{-4}) (Garofalo et al., 2019, Manickam et al., 2014, Mordret et al., 2019, Pernod et al., 2020, Zhang et al., 2013). However, this result clearly cannot be interpreted as a validation of Eq. (4.4). To test Eq. (4.4), the decoding rate constants for G$\circ$U mismatches should be measured, and $\Delta G_{\ddagger}$ in the decoding site accurately estimated. We also analyzed individual contributions of ΔG_{wc}^{C3} and $\Delta G_{\ddagger}$ to η. In the absence of another corresponding contribution, $\Delta G_{\ddagger}$ in the vicinity of the experimental range had higher individual contribution than ΔG_{wc}^{C3}, reaching the maximal difference of almost three orders of magnitude ((Fig. 4.2B). However, when each contribution was evaluated at the mean experimental value of the other contribution (i.e. ΔG_{wc}^{C3} contribution was evaluated at $\Delta G_{\ddagger} = \langle \Delta G_{\ddagger} \rangle_{exp}$), the maximal difference was less than three-fold ((Fig. 4.2A). Although this result cannot be interpreted quantitatively, it could reveal a general principle, which is discussed in Section 4.3.4.

4.2.2 η dependence on decoding rate constants

Next, we addressed the dependence of η on the decoding rate constants k_4^c (GTPase activation) and $q_4^{c/nc}$ (cognate and near-cognate escape rate constants of the $C4$ state). These rate constants determine the deviation of the open↔closed transition in decoding from equilibrium conditions, thereby limiting the accuracy of decoding in the classical model (Savir and Tlusty, 2013, Wohlgemuth et al., 2011). Eq. (4.4) suggests two possible kinetic regimes of the wb-WC reaction for the case of positive ΔG_{wc}^{C3}. In the "fast" regime ($(k_f + k_r) >> (k_4^c + q_4^c)$), kinetics of the wb-WC reaction and P_{WC}^{C3} are irrelevant, and $P_{wc} = P_{WC}^{eq}$. For $P_{WC}^{eq} = k_4^{nc}/k_4^c$ (corresponds to $\Delta G_{wc}^{C4} \approx 3.4$ kcal/mol), the fast regime is equivalent to the classical error η_0 (Eq. (1.1)) in both k_4^c

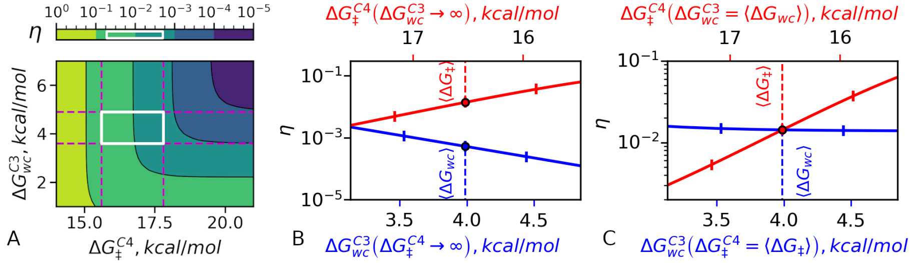


Figure 4.2.: η as a function of wb-WC reaction energies. **A** – η as a 2D function of $\Delta G_{\ddagger}$ and ΔG_{wc}^{C3}, calculated using Eq. (4.4) for $\Delta G_{wc}^{C4} = -1$ kcal/mol. The highlighted region corresponds to the experimental range of $\Delta G_{\ddagger}$ and ΔG_{wc} in RNA duplexes (Kimsey et al., 2018). **B** – individual contributions of $\Delta G_{\ddagger}$ and ΔG_{wc}^{C3} to η. Each contribution is evaluated in the absence of another contribution (associated $\Delta G \rightarrow \infty$). Circles and dashed vertical lines denote the mean values and standard deviations of the corresponding experimental range ($\langle \Delta G \rangle$) (Kimsey et al., 2018); the vertical markers on the curves denote standard deviations of the corresponding experimental range (Kimsey et al., 2018). **C** – individual contributions of $\Delta G_{\ddagger}$ and ΔG_{wc}^{C3} to η, evaluated at mean experimental values of another contributions.

and $q_4^{c/nc}$ dependencies, which is confirmed by numerical calculations (Fig. 4.3A-B). Therefore, our model satisfies the correspondence principle (Bohr, 1920): it reduces to the classical induced-fit model when the wb-WC reaction is at equilibrium in the pre-chemistry step of decoding.

A more intriguing and relevant given the predicted wb-WC parameters is the "slow" regime, where the wb-WC equilibrium in $C4$ is shifted to the WC geometry ($k_f > k_r$), but the kinetics of the reaction is restricted by the decoding rates. To visualize the slow regime, we chose wb-WC parameters ($\Delta G_{wc}^{C4} = -1$ kcal/mol, $\Delta G_{\ddagger} = 17.8$ kcal/mol), which approximately correspond to the predicted ΔG_{wc}^{C4} (Table 3.3) and experimentally observed $\Delta G_{\ddagger}$ in RNA duplexes (Kimsey et al., 2018). For $k_f < k_4^c$, Eq. (4.4) predicts a virtually flat $\eta(k_4^c)$ curve, as confirmed by numerical solutions, and ΔG_{wc}^{C3} affects the offset of the curves (Fig. 4.3A). The explanation for the flat $\eta(k_4^c)$ curve in the slow regime is derived in Appendix-B and visualized on Fig. 4.4: the linear approximations of P_{wc} and $\frac{[C4_{nc}]}{[C4_c]}$ have inverse dependency on k_4^c, which cancels in η. Equilibration of the exoergic wb-WC reaction in $C4$ with unfavorable contribu-

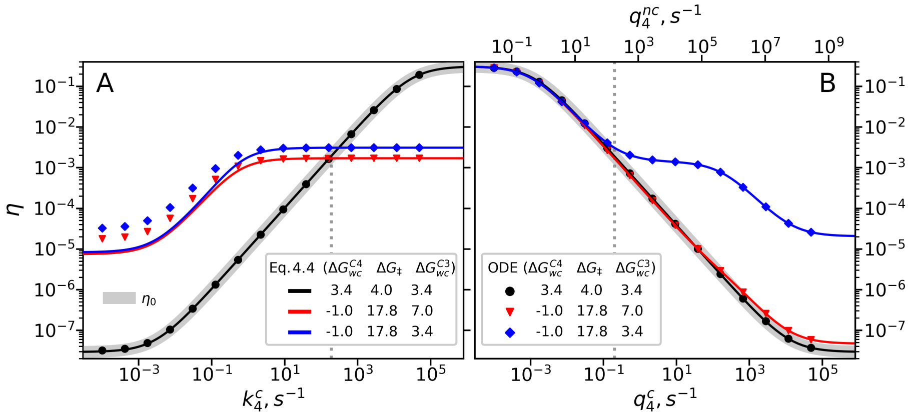


Figure 4.3.: η as a function of rate constants of the open-closed transition. η as a function of k_4^c (**A**) and $q_4^{c/nc}$ (**B**), calculated by (1.1) for the canonical system (η_0), and, for the specified wb-WC parameters, from numerical simulations (ODE) and by (4.4). $\eta(q_4^{c/nc})$ was obtained by varying q_4^c and q_4^{nc} simultaneously at the constant q_4^c/q_4^{nc} ratio. All ΔG values in the plot legend are in kcal/mol. The gray dotted vertical lines denote experimental values of k_4^c and $q_4^{c/nc}$.

tion to the accuracy counteracts the equilibration of the open→closed transition with favorable contribution, thereby constraining the decoding accuracy.

In the linear approximation, for $k_r << k_f < k_4^c$ we can derive an equality (Appendix-B):

$$k_4^{nc} = k_f + q_4^{nc} P_{WC}^{C3} \tag{4.5}$$

Eq. (4.5) suggests that the near-cognate rate constant of GTPase activation k_4^{nc} is defined by the parameters of the wb-WC reaction for a biologically-relevant range of k_4^c. k_4^{nc} of a codon-anticodon combination with a G∘U mismatch is yet to be measured in high-fidelity conditions. In DNA replication, the similarity between $k_{pol}^{incorrect}$ and k_f values was already noted by Kimsey et al. (2018), but remained unexplained within their numerical kinetic modeling approach.

In $\eta(q_4^{c/nc})$, the slow kinetic regime with a very low P_{WC}^{C3} ($\Delta G_{wc}^{C3} = 7.0$ kcal/mol) closely follows η_0. However, the same kinetic regime with $\Delta G_{wc}^{C3} = 3.4$ kcal/mol results in an almost flat curve for $q_4^{nc} > k_4^c > q_4^c$ (Fig. 4.3B). This effect is explained

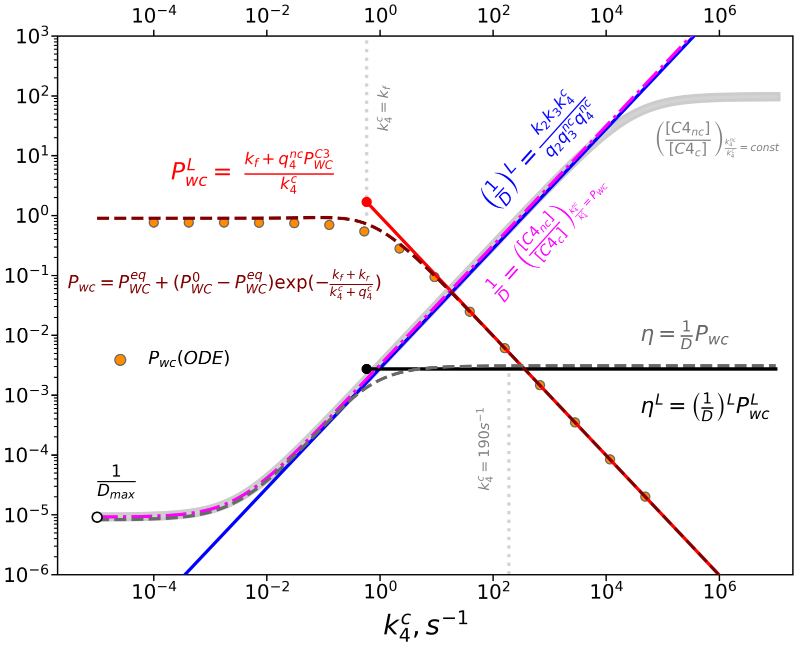


Figure 4.4.: Visualization of the terms contributing to the flat $\eta(k_4^c)$. Equations on the plot label curves with the corresponding colors. $D^L, P^L_{wc} and \eta_L$ represent the linear approximations of the corresponding equations (see Appendix). The linear approximations are valid only for $k_4^c > k_f$ and serve to visualize cancellation of k_4^c in η_L – the reason behind the flat $\eta(k_4^c)$ trade-off. The curves on the plot are calculated for the wb-WC parameters ($\Delta G^{C4}_{wc} = -1$ kcal/mol, $\Delta G^{C3}_{wc} = 3.4$, $\Delta G_{\ddagger} = 17.8$ kcal/mol)

by the kinetic partitioning term in $P^{C3}_{WC} \frac{q_4^{nc}}{k_4^c + q_4^c}$, which grows proportionally with q_4^{nc} and cancels the classical equilibration process. Intuitively, it can be understood as following: increasing "rejection" of the wobble geometry from $C3$ into $C4$ results in the proportional shift toward WC in the wb-WC equilibrium which is transferred from $C3$ in $C4$. When this contribution exceeds the contribution of the wb-WC kinetics in $C4$, $\eta(q_4^{c/nc})$ becomes flat.

It is also worth noting that our model does not contradict the experiments with varying Mg^{2+}. The experiments with Mg^{2+} (affects q_2) revealed linear rate-accuracy trade-offs (Zhang et al., 2018). For all considered wb-WC parameters, the rate-accuracy trade-offs obtained by varying q_2 are linear in η_0, as well as in η (Fig. 4.5).

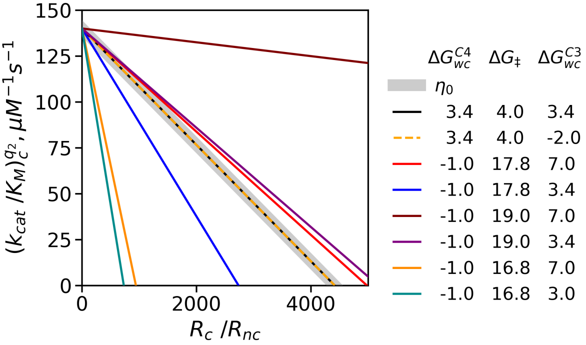


Figure 4.5.: Linear rate-accuracy trade-offs obtained by varying q_2 (modeling the effects of Mg^{2+} concentration). Instead of visualizing $\eta(q_2)$, the plot is designed to be directly comparable to the plots in (Zhang et al., 2018). As can be seen, the linear trade-offs at q_2 do not necesserily imply fast tautomeric equilibration throughout all initial selection process, contrary to what suggested by (Zhang et al., 2018). Instead, kinetic and thermodynamic properties of the wb-WC reaction in C3 and C4 states affect the slope of the stright trade-off lines, similarly to the effects of type and position of a mismatch observed in (Zhang et al., 2018). wb-WC parameter combinations are shown on the right.

In sum, Eq. (4.4), confirmed by the numerical calculations, predicts that a high $\Delta G_{\ddagger}$ in $C4$ causes a flat $\eta(k_4^c)$ curve for $k_f < k_4^c$, while high P_{WC}^{C3} causes an almost flat $\eta(q_4^{c/nc})$ for $q_4^{nc} > k_4^c > q_4^c$. Two potentially independent parameters of the wb-WC reaction constrain the decoding accuracy of the G∘U mismatch recognition in translation. This might have significant implications for the evolutionary optimization of decoding.

4.2.3 Tautomerization as a potential evolutionary constraint

To explore this further, we considered the cumulative error rate of translation, η_c, as a sum of η contributions from mismatches with slow, WC-shifted wb-WC transitions (e.g., G∘U mismatches) and with "classical" η_0-like behavior. In terms of our model (Fig. 4.1), the latter can be interpreted as mismatches having a fast endoergic transition to the WC geometry in $C4$, or mismatches whose GTPase activation rate is governed by low WC-independent rate constant (i.e. $k_4^{wb} \neq 0$, $\frac{k_4^{wb}}{k_4^c} = const$), that is ignored when considering G∘U mismatches. Recent studies show that the error hotspots (primarily G∘U mismatches) have error rate of $\sim 10^{-2} - 10^{-4}$ *in vitro*, and the error rate of other mismatches span a range of $10^{-4} - 10^{-7}$ (Garofalo et al., 2019, Manickam et al., 2014, Mordret et al., 2019, Pernod et al., 2020, Zhang et al., 2013). Based on these findings, we approximated η_c as sum of η of the slow kinetic regime ($\Delta G_{wc}^{C4} = -1$ kcal/mol, $\Delta G_{\ddagger} = 17.8$ kcal/mol, $\Delta G_{wc}^{C3} = 3.4$ kcal/mol) and a range of η_0 curves with rescaled k_4^{nc}/k_4^c ratio that results in η_0 range of $10^{-3} - 10^{-8}$ at the experimental values of the decoding rate constants. We calculated η_c as a function of k_4^c and $q_4^{c/nc}$. The resulting η_c curves are shown as dashed lines on Fig. 4.6. η_c is governed by η of the G∘U mismatch for all values of $q_4^{c/nc}$, and for decreasing values of k_4^c.

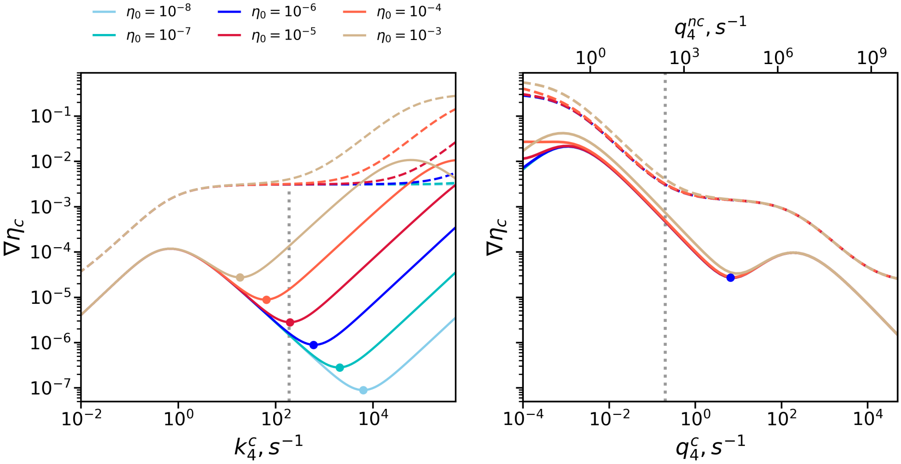


Figure 4.6.: Cumulative error rate η_c and its gradient $\nabla\eta_c$. η_c and $\nabla\eta_c$ were calculated for a range of scaled down η_0 curves as functions of k_4^c and $q_4^{c/nc}$. The colors denote the η_0 values at $k_4^c = 190s^{-1}$ and $q_4^{nc} = 140s^{-1}$ as contributions to η_c. The dashed and solid lines denote η_c and $\nabla\eta_c$, respectively. Local minima in $\nabla\eta_c$ are shown as circle of corresponding colors. The gray vertical lines denote the experimental values of k_4^c and $q_4^{c/nc}$.

The evolutionary optimization of the values of k_4^c and $q_4^{c/nc}$ cannot be explained solely on η or η_c, which are monotonic functions of these rate constants. To explore the possible reasons for the evolutionary optimization of these rate constants to the values observed in the experiments, we looked for the extrema in the gradient of η_c, $\nabla\eta_c$. $\nabla\eta_c$ determines the noise (i.e. standard deviation) of error rate η_c, given that η_c is monotonic. However, calculating the noise of $\eta_c(k_4^c)$ and $\eta_c(q_4^{c/nc})$ would require knowing the variance of k_4^c and $q_4^{c/nc}$. Therefore, for this illustrative purpose, we restricted to the analysis of $\nabla\eta_c$. Numerical gradients $\nabla\eta_c$ were calculated for each η_c curve and shown on Fig. 4.6 as solid lines. At η_0 contribution of 10^{-5}, the experimental value of k_4^c is at the local minimum of $\nabla\eta_c$ (Fig. 4.6). We selected this η_0 parameter to visualize the cognate rate of decoding R_{cog}, η_c and $\nabla\eta_c$ as 2D functions of k_4^c and $q_4^{c/nc}$ (Fig. 4.7). Both rate constants for a wide range do not significantly contribute to R_{cog} (Fig. 4.7). The exoergic wb-WC reaction constraints η_c at a plateau (Fig. 4.7B): for the ribosome to significantly improve the accuracy, the decoding rate needs to be reduced. In contrast, for a classical mismatch, the ribosome could reduce the error rate at least 1000-fold without significantly affecting the decoding rate (Fig. 4.7A).

In the cumulative approximation, the ribosome (the experimental values of decoding rate constants) is positioned in a region of local minimum of $\nabla\eta_c$ (Fig. 4.7C). It illustrates the possibility of the evolutionary optimization for the minimal gradient, and thus for the minimal noise of error rate. For the case of η of a G∘U mismatch only, this position is shifted from the minimum (Fig. 4.7B). Consideration of the cumulative error of decoding would likely be important to understand the evolutionary optimization of the ribosome, but requires more experimental information and more realistic approximations.

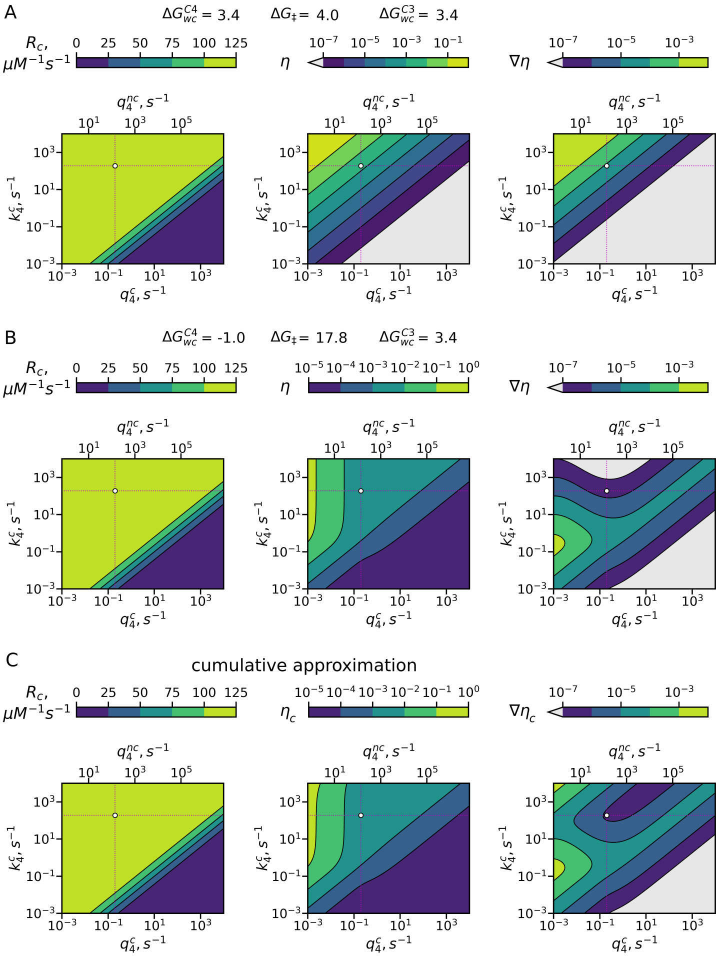


Figure 4.7.: Constraints on η and $\nabla\eta$ from the tautomerization. Fast regime (**A**), slow regime (**B**) and the cumulative approximation (**B**) are compared. The cognate rate of decoding R_c, η, and $\nabla\eta$ were calculated as functions of k_4^c and $q_4^{c/nc}$. For the cumulative approximation, each function was calculated as a sum of corresponding contributions from the slow regime of the wb-WC reaction ($\Delta G_{wc}^{C4} = -1$ kcal/mol, $\Delta G_{\ddagger} = 17.8$ kcal/mol, $\Delta G_{wc}^{C3} = 3.4$ kcal/mol) and a "classical" mismatch with rescaled k_4^{nc}/k_4^c ratio that results in $\eta_0 = 10^{-5}$ at the experimental values of the decoding rate constants. The dotted lines and a white circle denote the experimental values of k_4^c and $q_4^{c/nc}$.

4.3 Discussion

4.3.1 Unification of the structural and kinetic data

The main motivation for development of the new kinetic model of decoding was inconsistencies in the current model of G∘U mismatch recognition, which attributed the discrimination capacity to the "energy penalty" of tautomerization (Demeshkina et al., 2012, 2013, Rozov et al., 2015, 2018), failing to explain the observations of the WC geometry of this mismatch in the structural studies. Our new model (Eq. (4.4)), supported by the results of our QM/MM US calculations (Table 3.3), predicts that the WC geometry would be predominant at equilibrium conditions, i.e. $P_{wc} \to 1$ if $(k_f + k_r) > (k_4^c + q_4^c)$. Since q_4^c is small (Rudorf et al., 2014), and $k_r < k_f$ for the exoergic wb-WC reaction, decreasing k_4^c below k_f would be enough to stabilize the WC geometry in the closed A-site. The structural studies reporting the G∘U in the WC geometry were conducted in equilibrium conditions: in X-ray diffraction studies EF-Tu was absent (Demeshkina et al., 2012, 2013, Rozov et al., 2015, 2018), implying $k_4^c = 0$; similarly, in the Cryo-EM studies a non-hydrolysable GTP analogue (GDPCP) was used (Loveland et al., 2017), also implying $k_4^c = 0$. Therefore, our model unites structural and kinetic data in the field of codon-anticodon decoding: it explains how the WC geometry of G∘U can be stabilized in the closed A-site, while still being discriminated by the ribosome. It also explains the seemingly suboptimal position of the ribosome in the space of decoding rate constants: wb-WC parameters constrain the error rate of decoding to a plateau, precluding a significant increase in accuracy without dramatic losses in the decoding rate. Under such constraints, the decoding in translation might be evolutionarily optimized to minimize other properties, such as the error rate gradient, which is intrinsically related to noise. However, this speculation requires more sophisticated studies.

4.3.2 Historical perspective

Our new model builds on the induced-fit model, but allows to consider flexible substrates that change *during* decoding. On Fig. 4.8 we illustrate the evolution of the substrate recognition models using simplified schemes of the three discussed models. The scheme also illustrates the way the models can be reduced to the preceding ones. In accordance with the correspondence principle by Bohr (1920), our model reduces to the induced-fit/conformational-selection model upon equilibrium in the substrate at the last pre-chemistry step. Similarly, the induced-fit/conformational-selection model reduces to the lock-and-key model upon equilibrium in ribo-/enzyme.

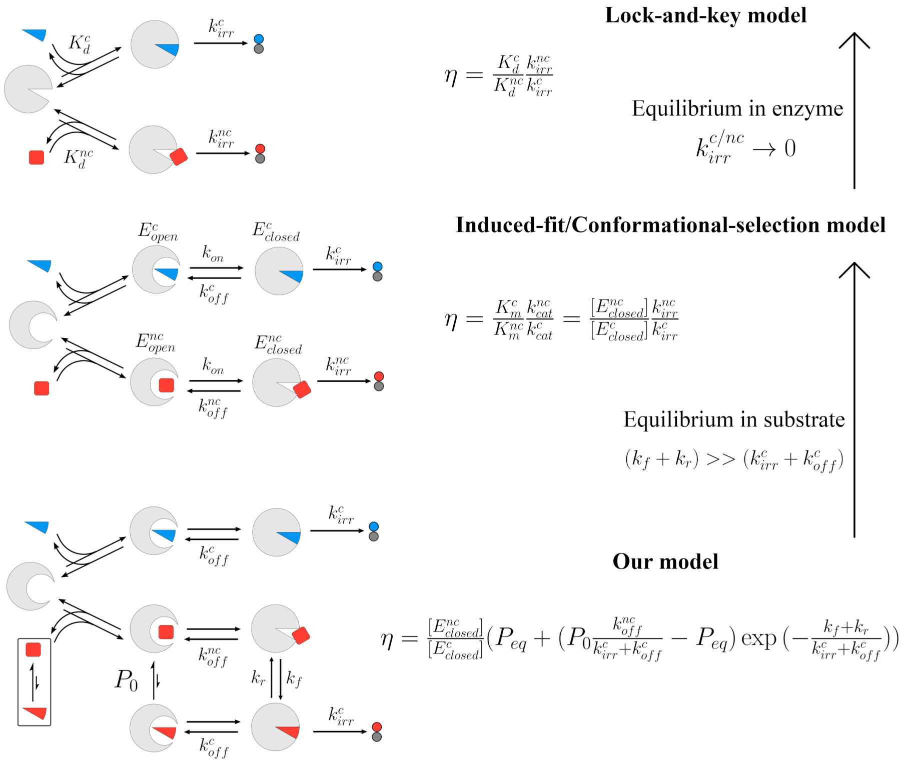

Lock-and-key model
K_d^c
k_{irr}^c
K_d^{nc}
k_{irr}^{nc}
$\eta = \frac{K_d^c}{K_d^{nc}} \frac{k_{irr}^{nc}}{k_{irr}^c}$
Equilibrium in enzyme
$k_{irr}^{c/nc} \to 0$
Induced-fit/Conformational-selection model
E_{open}^c
E_{closed}^c
k_{on}
k_{off}^c
E_{open}^{nc}
E_{closed}^{nc}
k_{off}^{nc}
$\eta = \frac{K_m^c}{K_m^{nc}} \frac{k_{cat}^{nc}}{k_{cat}^c} = \frac{[E_{closed}^{nc}]}{[E_{closed}^c]} \frac{k_{irr}^{nc}}{k_{irr}^c}$
Equilibrium in substrate
$(k_f + k_r) >> (k_{irr}^c + k_{off}^c)$
Our model
P_0
k_r
k_f
$\eta = \frac{[E_{closed}^{nc}]}{[E_{closed}^c]}(P_{eq} + (P_0 \frac{k_{off}^{nc}}{k_{irr}^c + k_{off}^c} - P_{eq}) \exp(-\frac{k_f + k_r}{k_{irr}^c + k_{off}^c}))$

Figure 4.8.: Correspondence between the substrate recognition models. Our new model reduces to the induced-fit model upon the equilibrium in substrate, and the induced-fit model reduces to the lock-and-key model upon equilibrium in enzyme.

In our model, the near-cognate GTPase activation rate constant k_4^{nc} is an apparent rate constant, which is defined as a product of the cognate k_4^{nc} and the tautomeric population in $C4$ (Eq. (4.3)). This is in line with the earlier suggestion of the GTPase activation reaction as an "internal standard" in translation, i.e., its independence on the nature of the tRNA (Thompson, 1988), although this suggestion was mainly based on outdated experiments, which are now challenged by the modern data (Gromadski et al., 2006). However, more recent theoretical studies also argue that the "true" k_4 must be independent of the codon-anticodon combination, but proposed different origin of the decreased apparent k_4^{nc} (Pavlov and Ehrenberg, 2018).

4.3.3 Potential refutations

The main prediction of the new model is the vastly different dependence of the error rate η on $q_4^{c/nc}$ and k_4^c compared to the classical model (see Section 4.2.2). This prediction can be tested experimentally. To the best of our knowledge, currently available studies of rate-accuracy trade-offs do not confirm or disprove this prediction of our model.

Thompson and Karim (1982) used a slowly hydrolyzable GTP analog GTP[γS] to study the binding of cognate (tRNA^{Phe}) and non-cognate (tRNA$^{Leu}_{CAG}$) ternary complexes to the poly(U)-programmed ribosomes. They concluded that the observed K_d^c/K_d^{nc} ratio, lower than would have been expected in experiments with GTP, was due to the reduced rate of GTP[γS] hydrolysis, and therefore the decoding was close to equilibrium conditions (Thompson and Karim, 1982). One of the features of this study that makes it not a refutation of our prediction is the use of non-cognate tRNA with two mismatches (U∘G in the first and U∘C in the third codon-anticodon positions).

A study by Zeidler et al. (1995) concluded that H85Q variant of EF-Tu decreased the error of Leu vs Phe misincorporation from bulk tRNA on poly(U)-programmed *T. thermophilus* ribosomes at 65 °C due to a decreased rate of GTP hydrolysis by this variant. However, in their study the effect of the H85Q variant on tRNA binding was higher than on the GTP hydrolysis (Zeidler et al., 1995), therefore the source of the decreased error could not be clearly determined. Moreover, the contribution of individual mismatches could not be estimated.

Although not available yet, testing $\eta(k_4^c)$ dependence seems feasible by e.g., GTP analogs with lower or higher hydrolysis rate by EF-Tu. Testing the prediction of η dependence on increasing $(q_4^{c/nc})$ might be more complicated, as it would require selective destabilization of the closed A-site state. The error-restrictive mutations in ribosomal protein S12 in the study by Zaher and Green (2010) did not significantly affect $(q_4^{c/nc})$, but rather the steps following the initial selection.

It is important to emphasize that for $q_4^{nc} < k_4^c$ and for a reasonable range of $\Delta G_{\ddagger}$, our model predicts similar behavior of $\eta(q_4^{c/nc})$ as in the classical model (Fig. 4.3B), thus not contradicting the experiments with aminoglycosides (Zhang et al., 2018) and ribosome ambiguity mutations (Hoffer et al., 2019).

Our model of the wb-WC reaction effects on decoding accuracy was restricted to the initial selection, but it does not contradict the concept of proof-reading in translation. Recent cryo-EM studies demonstrated that in the post-GTP-hydrolysis states the decoding site resets back to the open state (Loveland et al., 2020), which could potentially reset the equilibrium in the wb-WC reaction back to the wobble geometry. However, proof-reading of G∘U mismatches in translation has yet to be directly demonstrated. Recent single-molecule studies revealed a possibility of multiple rounds of EF-Tu•GTP rebinding and GTP hydrolysis to the same aa-tRNA in the A-site (Morse et al., 2020). Such possibility, although yet to be shown for near-cognate complexes, challenges the common (and the only one) method to calculate proof-reading contribution to the accuracy as the excess of GTP hydrolysis compared to peptide bond formation (Ehrenberg et al., 1986, Zhang et al., 2016).

4.3.4 Alternative explanation to the "optimal decoding" by the ribosome

A theoretical study by Savir and Tlusty (2013) attributed the "optimal" values of decoding rate constants $q_4^{c/nc}$ and $k_4^{c/nc}$, in which q_4^{nc} and k_4^c are similar, to a symmetry in a fitness solution that optimizes both rate and accuracy. Since neither of these rate constants for a broad range significantly contributes to the cognate rate of even the initial selection (given the *in vitro* rate constants values) (Wohlgemuth et al., 2011), it remains unclear how exactly such solution optimizes fitness. Our model provides an alternative explanation. In our model the apparent similarity between q_4^{nc} and k_4^c can be explained with the kinetic partitioning term $\frac{q_4^{nc}}{k_4^c+q_4^c}$ in P_{wc}^0. For $q_4^{nc} > k_4^c$, and assuming $q_4^c << k_4^c$, contributions of P_{wc}^0 to the error rate will exceed contributions from the slow wb-WC reaction in C4 (i.e. k_f), and mostly cancel the classical equilibration, thus having virtually no advantage to the accuracy of decoding. For $q_4^{nc} < k_4^c$, the error is dominated by k_f in C4 and increases similarly to the classical model. It is possible to suggest that decoding in translation was optimized under the constraints of two independent wb-WC parameters P_{wc}^{C3} and k_f, resulting in an "optimal" solution where both parameters have almost equal contribution to the error rate (see Fig. 4.2). Although our model can justify the ribosome position at $q_4^{nc} \approx k_4^c$, a rigorous explanation would still require a consideration of additional variables, since $q_4^{nc} \approx k_4^c$ is not a local minimum of η.

It is worth mentioning here another explanation to the high k_4^c in decoding. Rodnina and coworkers suggested that high k_4^c is needed to "buffer" the detrimental effect of tRNA competition *in vivo*, i.e. the apparent k_4^c *in vivo* would be reduced, and the high intrinsic k_4^c is needed to not reduce it to rate-limiting value (Gromadski et al., 2006, Gromadski and Rodnina, 2004, Wohlgemuth et al., 2011). This suggestion does not contradict our model: $\eta(k_4^c)$ is mostly flat under the tautomerization constraints, and thus k_4^c can indeed be tuned to optimize other decoding properties without losing accuracy.

4.3.5 Potential application to DNA replication

DNA polymerases and the ribosome have similar recognition mechanisms, i.e., the presence of the open→closed transition of the active/decoding site (Tsai and Johnson, 2006). Therefore, it is reasonable to suggest that some DNA polymerases can be plagued by the wb-WC reaction in G∘T in a similar way as our model predicts it for the ribosome. We speculate that the dependence of the error rate on the rate constants of the open→closed transition in pol-β would also display flat regions. It is also intriguing to speculate that the potentially reduced steric constraints on the WC geometry in the active site of T7-pol (which is in line with the absence of structural studies showing the WC geometry of G∘T in the active site of this DNA polymerase), discussed in

Chapter 3, could be an adaptation to allow T7-pol to operate in close to equilibrium conditions of G∘T recognition with respect to the wb-WC kinetics. Previous theoretical studies have suggested that this DNA polymerase operates in the energetic regime of recognition, as opposed to the kinetic regime in another DNA polymerase, pol-γ (Sartori and Pigolotti, 2013).

The recent study by Kimsey et al. (2018) may seem as a refutation of our model applied to DNA polymerases. Since Al-Hashimi and coauthors (Kimsey et al., 2018) could predict misincorporation rates of several DNA polymerases from the equilibrium WC geometry populations in solution, it may indicate that no stabilization of the WC geometry happens in the active sites of any of the DNA polymerases they studied, or such stabilization does not affect the error rate.

However, this experimental observation does not contradict our model. In the numerical kinetic modeling approach employed by Kimsey et al. (2018), the wb-WC reaction was modeled only in the open state of the active site, using the experimentally measured ΔG_{wc} in water. Continuing our analogy with the ribosome, it is reasonable to suggest that the environment of the open active site of DNA polymerases does not significantly perturb the WC population compared to water solution. The linearity of the model by Kimsey et al. (2018) results in $P_{wc}^{closed} = P_{wc}^{open}$ (i.e. the wb-WC reaction in the closed state is absent, and the open→closed kinetic partition term equals 1). In our model, in the vicinity of experimental $\Delta G_{\ddagger}$ and ΔG_{wc} values, and for the used values of the decoding rate constants, Eq. (4.4) predicts comparable contributions to η from P_{wc}^{0} and from the slow wb-WC reaction in C4 (k_f) (Fig. 4.2C). In fact, as discussed above, we speculate that the similarity of the contributions from P_{wc}^{0} and k_f could be a result of the evolutionary optimization under the constraints of these two independent parameters. Therefore, the numerical kinetic model by Kimsey et al. (2018) (implicitly) predicted P_{wc}^{open} from the experimental data, which can still be a good approximation to P_{wc}^{closed} given that $P_{wc}^{open} \approx \frac{k_f}{q_{closed \to open}^{nc}}$. If in some DNA polymerases the equilibrium of the wb-WC reaction is not significantly affected by the closed active site environment, like our QM/MM calculations predict it for T7-pol (Table 3.3), $P_{wc}^{closed} \approx P_{wc}^{open}$ would be an even more realistic approximation. Therefore, the study by Kimsey et al. (2018) is neither a confirmation nor a refutation of our model. Since ΔG_{wc} and $\Delta G_{\ddagger}$

of the wb-WC reaction measured by Kimsey et al. (2018) in duplexes in solution were correlated (Fig. 4.9), it can be challenging to experimentally distinguish predictions of our model from the interpretations by Kimsey et al. (2018).

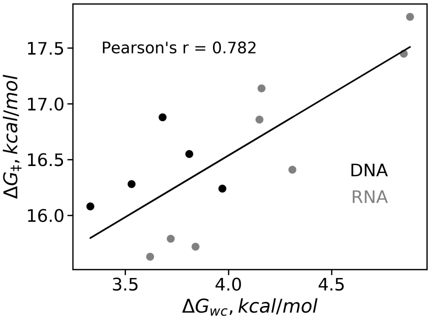


Figure 4.9.: Linear correlation between ΔG_{WC} and $\Delta G_{\ddagger}$ of the wb-WC reaction in G○U(T) of RNA and DNA duplexes in solution, obtained with NMR by Kimsey et al. (2018). Only neutral pH conditions and unmodified nucleobases were considered. Measurements in DNA and RNA duplexes are shown in black and grey, respectively. The black line denotes the linear fit of all measurements. Pearson's correlation coefficient is shown on the plot.

Chapter 5

Effects of tRNA modifications on proton transfers in base pairs

In this chapter I focus on how mnm^{5}s^{2}U tRNA modification involved in ambiguous codon decoding affects tautomerization and other proton transfer reactions in base pairs and propose an approach for experimental verification of the theoretical predictions.

5.1 Methods

5.1.1 Static QM calculations

In static approach we focused on isolated base pairs containing A, G, U and mnm^{5}s^{2}U in two protonation state of its amino group. We will denote mnm^{5}s^{2}U as Ux (Ux$^+$ denotes the protonated state of the modification). In some calculations A was replaced with 2-aminopurine, a prospective fluorophore for potential experimental verification. We also optimized all considered monomers in all protonation/isomeric states involved in the studied base pairs. All nucleobases contained methyl group at their glycosidic positions to model the glycosidic bond. This approach provides more realistic modeling of the electronic structure of the nucleobases compared to the NH group. We performed geometry optimizations on B3LYP-D3BJ/def2-TZVP and PM7 levels of theory in gas phase and implicit solvation models (COSMO). Orca 4.1 (Neese, 2018) was used for all DFT calculations, while MOPAC-2016 (Stewart, 2016) was used for all SE calculations. Optimization algorithms and convergence criteria were the same as described in Section 3.1.1. Basis set superposition error (BSSE) correction was included in all DFT calculations for correct estimation of interaction free energies. All final optimized geometries contained none (local minima) or exactly one (TS) imaginary frequencies. ΔG was calculated under harmonic approximation, as described in Chapter 2. Multiwfn (Lu and Chen, 2012) was used to calculate electrostatic potential

(ESP) of the base pairs at electron density (ρ) isovalue of 0.001 au. VMD (Humphrey et al., 1996) was used to visualize the ESP maps. Gas phase interaction free energy was defined as

$$\Delta G_{int} = G(base\ pair) - \sum_i G(monomer_i) \tag{5.1}$$

where monomer *i* is optimized in the same configuration it exists in a given base pair. NEB optimization was performed to find TSs and calculate the potential energy barriers of the wb-WC reaction. As described in Section 3.1.5, we used NEB-TS implementation, the final step of which is TS optimization using Berny algorithm. For Ux, we performed NEB between the wobble and G∘U*x configurations. For Ux$^+$, the product in NEB calculations was G$^+$∘Ux. All NEB calculations were performed in gas phase.

5.1.2 QM/MM metadynamics

All QM/MM simulations were performed in A-site models. We used structures of the *T.thermophilus* 70S ribosome 5E7K and 5IB8 as initial coordinates. In 5E7K, tRNA$_{LYS}$ anticodon UxUU binds cognate codon AAA (Rozov et al., 2016a). In 5IB8, tRNA$_{LYS}$ anticodon UxUU binds near-cognate codon codon GAA, forming the WC geometry of G∘U base pair at the first codon position (Rozov et al., 2016b). Except for this differences, the models were prepared and equilibrated exactly as described in Section 3.1.2.

In contrast to the Chapter 3, where all QM regions included only nucleobases of a single base pair, in all QM/MM simulations in this chapter we included all six nucleobases of the codon-anticodon helix into the QM region, described using PM7 method. In this chapter we employed metadynamics (MetD) instead of umbrella sampling. All MetD calculations were addressing proton transfer (PT) mechanism in A∘U and A∘Ux$^+$ base pairs.

1D well-tempered MetD simulations were performed to calculate PMF of PT reaction in the unmodified A∘U base pair in the second codon-anticodon position, and in the A∘Ux$^+$ base pair in the wobble position. These simulations were performed in models initiated from the PDB ID 5IB8. In the case of this PT reaction, the selection of a suitable CV is straightforward – N1-H3 distance is an obvious descriptor of the H3 transfer to N1 (see Fig. 5.2A), and thus was selected as CV in all MetD simulations. The grid for d(N1-H3) sampling was 0.8 Å : 2.5 Å with 0.05 Å step. Half-harmonic potentials were used to restrain sampling outside this grid (the lower boundary is restrained just by atomic repulsion). The initial Gaussian height was 0.25 kcal/mol, bias "temperature" 4200 K, and the hill deposition rate 1000 step^{-1}.

2D non-tempered MetD simulations were used to probe NH2 of mnm^5 group protonation/deprotonation equilibrium (i.e. measure its pKa) simultaneously with N1-H3 proton transfer. We ignored a possibility of NH2 deprotonation by interactions with other RNA residues, simply due to distance constraints. We only considered water molecules as proton donors/acceptors. To allow for proton transfers between water molecules and the NH2 group, we included a first solvation shell (13 water molecules) around the mnm^5 modification into the QM region, to the total of 123 atoms in the QM region. We did not use dynamic QM region (i.e. updating molecules into the QM region based on distance cut-offs), therefore the QM solvation shell became slightly diluted with MM water molecules over the course of MetD simulations. However, due to relatively short simulation time ($\sim$ 50 ps), this effect was small compared to the lack of sufficient sampling. In the 2D MetD simulations the CV describing the PT between the nucleobases was the same as in the 1D MetD simulations above. To probe pKa of the NH2 group, we used coordination number (CN), as implemented in NAMD colvar module (Fiorin et al., 2013). CN defines coordination (i.e. number of contacts) between two sets of atoms Z and Y as

$$CN(Z,Y) = \sum_{i \in Z} \sum_{j \in Y} \frac{1 - (|x_i - x_j|/d_0)^n}{1 - (|x_i - x_j|/d_0)^m} \tag{5.2}$$

where d_0 is the distance cut-off, x_k is coordinates of atom i or j, and n and m are exponents that control long/short range behaviour of Eq. (5.2). In our simulations, the first set included only N atom of the NH2 group of the mnm^5 modification, and the second set included its two protons as well as all (26) protons of the QM water molecules. The distance cut-off was 1.6 Å, and the exponents n and m were 6 and 12, respectively.

5.1.3 Fluorescence spectroscopy and multivariate curve resolution

Fluorescence spectroscopy was applied to measure pKa(N1) of 2-aminopurine triphosphate (2APTP) via pH titration for an illustrative purpose. Tecan Spark was used to measure UV emission spectra of 2APTP in water solution. 2APTP (TriLink Biotech) was dissolved in 150 mM NaCl MiliQ water to final concentration of 10 μM. For each titration step, small volumes of concentrated HCl were added, pH measured and 100 μl of sample were moved to a separate tube. 38 samples were collected and placed on 96 plate for fluorescence measurements in Tecan Spark. Emission spectra were recorded at fixed excitation wavelength of 253 $\pm$ 2 nm, which corresponds to 2AP excitation maximum at neutral pH (Gargallo et al., 2001).

Multivariate curve resolution (MCR) with alternative regression (AR) (Camp, 2019) was applied to calculate pKa of 2APTP following the approach described in (Gargallo et al., 2001). MCR-AR solves the problem of matrix factorization to find the optimal species concentration matrix **C** and their corresponding signature matrix **S**, that best explain the input data matrix **D**:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \tag{5.3}$$

where **E** – is the error matrix. In the case of emission spectra taken at multiple samples, MCR finds the optimal concentration profiles of the species in the samples, and individual (pure) spectra of the species. Non-negativity and constant total concentration constraints were imposed. The number of the species N should be estimated prior to MCR. In this illustrative experiment, only two species were expected: neutral and N1-protonated 2APTP, therefore we used N = 2. In potential future multicomponent experiments with unknown number of species (e.g. not only protonation, but also free/base paired 2AP) N should be calculated, e.g. using approaches described in (Wentzell et al., 2006). Since MCR is iterative optimization algorithm, it requires an initial approximation to either **C** or **S**. We performed 250 MCR runs using 38 raw emission spectra as **D**, and a matrix of (new) random numbers with dimensions 2, 38 as initial **C**. This approach reveals MCR dependence on initial **C** approximations. pKa values were calculated for the set of sigmoid solutions by fitting concentration profiles to sigmoid function and calculating their midpoints.

5.2 Results and Discussion

5.2.1 Acceleration of the wb-WC reaction in G∘Ux base pairs

In the previous chapters we proposed a model of tautomerization-induced base pair recognition errors upon G∘U mismatch in the first two codon-anticodon positions. If the geometric selection (i.e. selection by WC base pair geometries) is also valid at the wobble position, the error rate of G∘U recognition at this position becomes the efficiency of ambiguous decoding. Therefore, Eq. (4.4) can be applied to understand possible mechanisms of ambiguous decoding facilitation by Ux. If this anticodon modification increases P_{wc}^{C3} or k_f, or both, it will increase P_{wc}, thereby facilitating decoding of G-ending codons. Increasing P_{wc}^{C3} would mean increasing the equilibrium population of the WC base pairs in solution. For WC configurations formed with enol tautomers, this is hardly achievable, especially by Ux, which contains electron-withdrawing group, that can only destabilize enol tautomers (Hartono et al., 2018). A more realistic mechanism would be formation of G∘U^{-}x WC base pairs, as described in Section 1.3.2. Disadvantages of such model are discussed in Section 5.2.4. Here we propose an alternative (non-exclusive) model of Ux mechanism: acceleration of the wb-WC reaction, i.e. increasing k_f. Acceleration of the wb-WC reaction by halogen derivatives of U was already proposed by Hovorun and coworkers as an explanation of their mutagenicity (Brovarets and Hovorun, 2015). Although recent experimental studies suggest a connection between U$_{hal}$ mutagenicity and formation of G∘U$_{hal}^{-}$ WC base pairs (Kimsey et al., 2018), wb-WC reaction acceleration could not be excluded either. Therefore, we analyzed how the Ux modification decreases $E_{\ddagger}$ of the wb-WC reaction.

Starting with the neutral Ux, we optimized base pair geometries along the wb-WC reaction and performed NEB calculations between G∘Ux wb and G∘U*x WC. Neutral

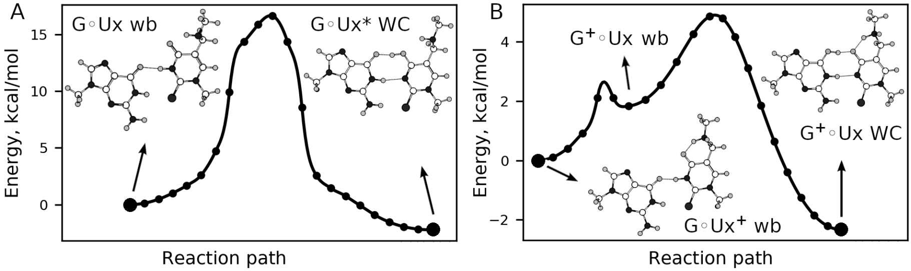


Figure 5.1.: NEB energy profiles of the wb-WC reaction in G∘Ux: with neutral Ux (**A**) and with Ux protonated at mnm^5 (**B**). Product and reactant structures are shown as insets in the plots. All NEB calculations were performed in gas phase.

Ux did not dramatically affect the wb-WC pathway and its relative energies (Fig. 3.5, Fig. 5.1A). The reaction proceeded from the wobble geometry to the G∘U*x WC configuration via a similar TS, which had similar $E_{\ddagger}$ ($\sim$ 17 kcal/mol) compared to the unmodified base pair (Fig. 3.5, Fig. 5.1A). Therefore, the neutral Ux modification does not notably accelerate the wb-WC reaction.

Next, we focused on the NH2$^+$ form of the mnm^{5}s^{2}U modification (Ux$^+$), which is its predominant form in water solution (Sochacka et al., 2017). When optimizing base pair geometries, we observed G∘U*x$^+$ WC converging to G$^+$∘Ux WC upon optimization. We then calculated the reaction path between G∘Ux$^+$ wb and G$^+$∘Ux WC using NEB. We found that Ux$^+$ dramatically changed the wb-WC reaction path and energy barriers (Fig. 5.1B). The product, G$^+$∘Ux WC configuration, was approximately 2 kcal/mol more stable than the wobble geometry. The reaction proceeded via two transition states: the first step ($E_{\ddagger}^1 \approx 2.5$ kcal/mol) involved single proton transfer from Ux$^+$ to G, forming the G$^+$∘Ux wb intermediate (Fig. 5.1B). In the next step ($E_{\ddagger}^2 \approx 3$ kcal/mol), the base pair geometry changed from wb to WC, forming the G$^+$∘Ux WC product. Therefore, Ux$^+$ modification transformed the wb-WC reaction from tautomerization to ionization. The very low total energy barrier ($\sim$ 5 kcal/mol) implies that this reaction will be virtually in instant equilibrium during the codon-anticodon decoding process, given the decoding rate constants used in Chapter 4. This would result in $\eta \approx 1$, meaning almost equally efficient decoding of G-ending codons, assuming the geometric selection in place at the wobble position.

However, there are major obstacles in extrapolating the result on Fig. 5.1B to the mechanism of ambiguous decoding facilitation by Ux^+. First, it is not clear if the base pair geometry is constrained in the wobble codon position, i.e. if the WC geometry in this position is *required* for efficient decoding (see Section 1.3.1). Second, the relevancy of this mechanism is questionable, as some portion of Ux^+ in water solution would be deprotonated at N3, and cannot (does not need to) participate in the wb-WC reaction, instead forming the WC geometry spontaneously. Sochacka et al. (2017) estimate the N3-ionized fraction of mnm^{5}s^{2}U and cmnm^{5}s^{2}U modifications to be 30% to 50% in water. The ionized fraction in the less polar ribosome environment is yet to be measured. Third, the NEB calculations were performed in gas phase on isolated base pairs. More realistic simulations are needed, such as QM/MM US calculations performed in Chapter 3. We attempted to apply this approach to Ux^+, but the resulting PMFs were plagued by sampling far away from the reference reaction path, indicating the inadequacy of the latter. Free energy calculations of the wb-WC reaction in G∘Ux^+ base pair in the decoding site environment might require another, path-independent approaches.

In sum, we showed how the Ux^+ modification dramatically changes the wb-WC reaction mechanism, reducing it to instant equilibrium. This might provide explanation to its role in decoding of G-ending codons, but more sophisticated future simulations and experiments are needed to study this mechanism in realistic environments.

5.2.2 Proton-transfer-mediated stabilization of A∘Ux base pairs

Most studies of U34 tRNA modifications address their role in decoding of G-ending codons, as this is arguably their most distinct contribution. However, some tRNA modifications, including (c)mnm^{5}s^{2}U, also facilitate decoding of their full-cognate, A-ending codons (Hagervall et al., 1998, Kurata et al., 2008, Yarian et al., 2002). This property is mostly unexplained and requires a physicochemical explanation.

While simulating the wb-WC reaction in unmodified G∘U at the first position, but including all three codon-anticodon base pairs into the QM region, we observed an interesting occurrence in the A∘Ux^+ base pair at the third position. In multiple independent trajectories, the H3 proton of Ux^{+} spontaneously transferred to the N1 atom of A, thus ionizing adenine to A^{+}. To the best of our knowledge, this reaction has never been addressed before. We set to study this proton transfer (PT) mechanism in more details using static QM calculations and QM/MM MetD simulations.

The scheme of the PT reaction is shown on Fig. 5.2A. The likely reason this reaction has not been studied earlier is because it does not happen in the unmodified A∘U base pair – geometry optimizations on both DFT and PM7 levels of theory in gas phase revealed the absence of a stationary point at $\mathrm{A}^{+}\circ\mathrm{U}^{-}$, as it converged to A∘U upon the optimization (Table 5.1). In the neutral state of the Ux modification (i.e. NH group in the mnm^{5} moiety), the PT product exists, but is 5.7 kcal/mol higher in energy than the A∘Ux reactant (Table 5.1). Finally, upon the charged $\mathrm{NH2}^{+}$ group in the mnm^{5} moiety, the reaction is exoergic with $\Delta G_{f} = -5.3$ kcal/mol at B3LYP-D3BJ/def2-TZVP level (Table 5.1). Geometry optimization at PM7 level revealed a similar trend, but with a substantial absolute error of 1.9 – 3.7 kcal/mol (Table 5.1). We also analyzed 2-aminopurine (2AP), which can be potentially used as a fluorophore to test this mechanism (see below). 2AP did not significantly affect the PT energies, thus allowing its use for experimental verification (Table 5.1).

Proton transfers rearrange the charge distribution in base pairs, altering electrostatic potential (ESP), which contains important information about intermolecular interactions. We calculated and visualized ESP to better understand the effect of the PT mechanism on the base pairs. It is clear that in the case of neutral Ux modification, the PT reaction results in the formation of ion pair $\mathrm{A}^{+}\circ\mathrm{Ux}^{-}$, where the WC edges of nucleobases bear opposite charges (Fig. 5.2B,top). This results in extremely high gas phase interaction energy ΔG_{int} of approximately -100 kcal/mol. However, the ionization energy counteracts this stabilization, making the PT reaction endoergic (Table 5.1). Upon the charged modification Ux^{+}, the ESP also demonstrates charge separation at the WC edges in the product, resulting in ΔG_{int} increase from -4.8 kcal/-mol in the reactant to -26.4 kcal/mol in the product (Fig. 5.2B,bottom). In this case, the reaction is exoergic and thus can contribute to stabilization of the $\mathrm{A}\circ\mathrm{Ux}^{+}$ base pair, potentially leading to increased codon-anticodon binding and facilitated decoding of A-ending codons observed experimentally (Hagervall et al., 1998, Kurata et al.,

Table 5.1.: PT mechanism energies

configuration	ΔG_f^{DFT} [a]	ΔG_f^{PM7} [b]	$\Delta\Delta G_{int}$ [c]
A∘U	– [d]	– [d]	-1.8 [e]
A∘Ux	5.7	9.4	-105
A∘Ux$^+$	-5.3	-3.4	-21
2AP∘Ux	6.6	–	–
2AP∘Ux$^+$	-4.8	–	–

[a] forward free energy change of the reaction on Fig. 5.2A, calculated with B3LYP-D3BJ/def2-TZVP, kcal/mol; [b] forward free energy change of the reaction on Fig. 5.2A, calculated with PM7, kcal/mol; [c] difference in interaction free energy in the forward reaction on Fig. 5.2A, kcal/mol; [d] the product of the reaction is not a stationary point; [e] ΔG_{int} of the base pair configuration;

2008, Yarian et al., 2002). However, these calculations were performed only in a gas phase. Including the solvent and the decoding site environment is essential to assess the relevancy of this mechanism for codon-anticodon decoding.

As we observed a qualitative consistency between DFT and PM7 energies of the PT reaction (Table 5.1), we used PM7 optimizations of base pairs containing Ux$^+$ to obtain dependence of ΔG_f on the dielectric constant ε of the implicit solvent model. As could be expected from the electrostatic nature of the interactions in the PT mechanism, the reaction switched from exoergic in gas phase and very non-polar environments to endoergic in water (Fig. 5.3A). We were also intrigued to observe the presence of the non-WC geometry of the A$^+$∘Ux base pair – the reversed wobble (rwb) geometry. In the non-polar ε, optimizations of this structure converge to the normal WC geometry of A$^+$∘Ux (Fig. 5.3A). However, at higher ε it becomes a distinct local minimum with relative energy lower than the normal A$^+$∘Ux WC base pair, but still higher than A∘Ux$^+$ WC (Fig. 5.3A). Interestingly, a very similar geometry of this base pair was already observed in X-ray structures of the ribosome (Rozov et al., 2016a) (Fig. 5.3B), but its nature remained unexplained. Calculations on higher levels of theory, and including ribosomal environment, would be necessary to exclude possible computational artifacts and understand a potential role of this unusual base pair geometry in codon-anticodon interactions.

Finally, we aimed to evaluate contribution of the decoding site environment on the PT

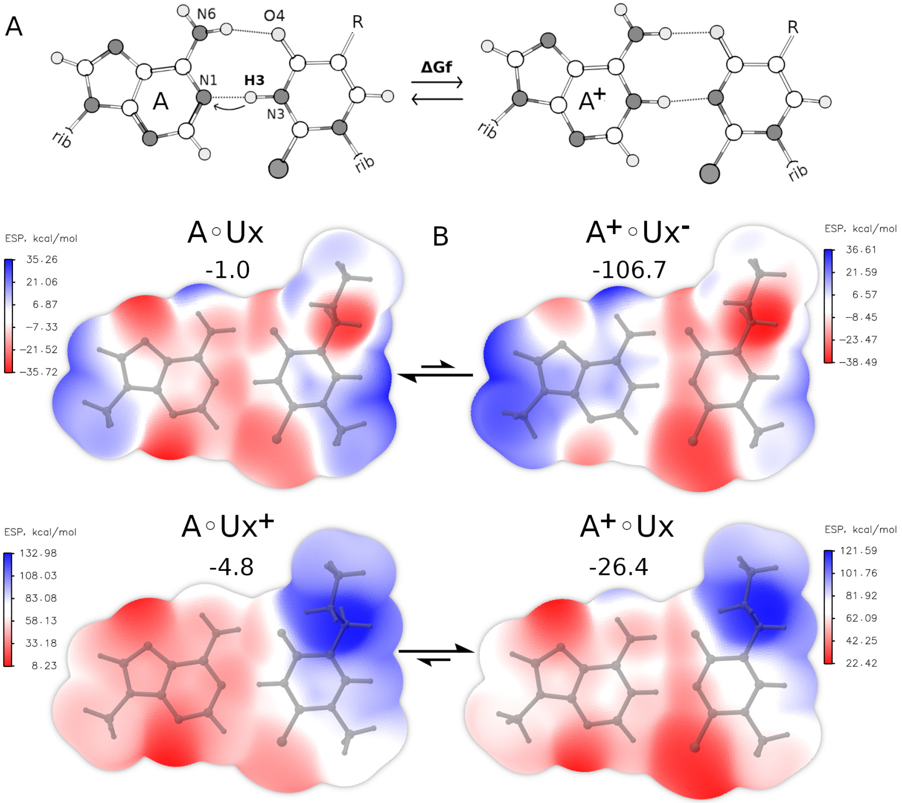


Figure 5.2.: PT-mediated mechanism of A∘Ux base pair stabilization. **A** – scheme of the PT reaction; **B** – ESP of base pairs for neutral (top) and protonated (bottom) configurations of the mnm^5 modification. ESP was calculated ar $\rho = 0.001$ a.u. isosurface. Numbers below the base pair notations represent gas phase interaction free energy ΔG_{int}, kcal/mol. Shifted equilibrium arrows depict the relative direction of the equilibrium from Table 5.1.

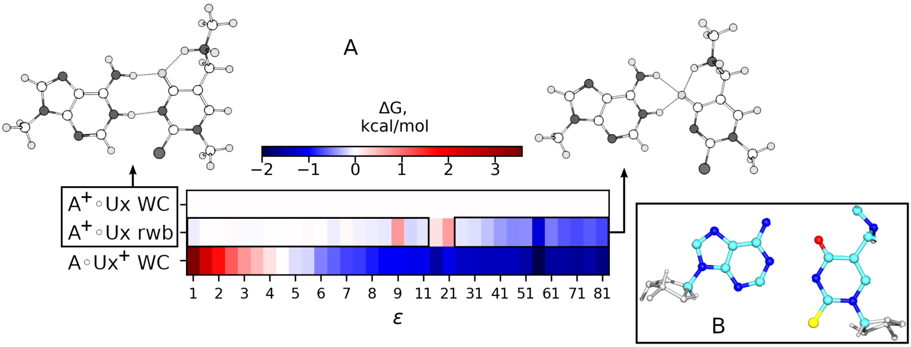


Figure 5.3.: ΔG_f of the PT reaction as a function of ε. **A** – relative free energies of A∘Ux base pair configuration in the PT reaction. Besides reactant (bottom row) and product (top row) of the PT reaction in Fig. 5.2A, we also observed a shifted conformation of the product configuration (middle row). At low ε this structure is not a stationary point and converges to the WC geometry upon optimization. At higher ε it is more stable than the WC geometry. **B** – a similar geometry of the A∘Ux base pair observed in the ribosome X-ray structure in the decoding site (Rozov et al., 2016a).

mechanism by performing QM/MM MetD simulations of the PT mechanism in the A-site models. First, we ran two well-tempered 1D MetD simulations – in A∘Ux$^+$ at the third position, and in the unmodified A∘Ux at the second position. N1-H3 distance was used as CV. The simulations did not reach full convergence (Fig. 5.4B), but can be used for qualitative assessment. As anticipated from the static calculations, PT in the unmodified base pair was highly endoergic, with only a slight presence of local minimum at the product region with $\Delta G_f = 12$ kcal/mol (Fig. 5.4A). PT in the A∘Ux$^+$ was very exoergic with $\Delta G_f = -8.9$ kcal/mol (Fig. 5.4A). This exoergicity can be explained by the decreased polarity of the closed decoding site, revealed in Chapter 3. Next, we wanted to perform an independent MetD simulation to compare ΔG_f, and also to estimate the pKa of the amino group int the mnm^5 modification, which is essential for the PT mechanism. We ran 2D non-tempered MetD with the QM region including water molecules surrounding the modification (Fig. 5.4C). In these MetD simulations one of the CV was d(N1-H3) as above, and another was the coordination number (CN) between the N atom of mnm^5, and the protons. Although

2D MetD simulations sampled CN values in the deprotonated (NH) region (Fig. 5.4D), the sampling in this CV was not sufficient, and thus the pKa cannot be accurately estimated. From 2D MetD, we extracted a 1D PMF in the well-sampled NH2$^+$ region (Fig. 5.4A, blue), which was very similar to the PMF from 1D MetD, with $\Delta G_f = -8.6$ kcal/mol. We conclude that the PT reaction is exoergic in A∘Ux$^+$ at the wobble position in the closed decoding site, but more rigorous simulations would be needed for quantitative information.

5.2.3 A proposed experimental test of the predicted mechanism

The PT-mediated base pair stabilization model can be experimentally tested. In the PT mechanism, adenine gets protonated at N1, which can be observed as a pKa shift at this atom (Fig. 5.5). If an experiment is to be performed in the biologically relevant environment of the ribosomal decoding site, this pKa shift would be inaccessible for common methods, such as potentiometric titration and NMR. Fortunately, it would be potentially accessible for fluorescence spectroscopy using fluorescent adenine analog 2AP. 2AP has a pKa(N1) close to adenine, and its fluorescence spectra, both emission and excitation, are pH sensitive (Gargallo et al., 2001). Neutral 2AP is very bright, but 2AP$^+$ has greatly reduced fluorescence intensity, and a slight excitation maximum shift (Gargallo et al., 2001). This properties allow to measure pKa(N1) of 2AP using pH titration combined with fluorescence spectroscopy. We illustrate this approach, already well established in the literature (Gargallo et al., 2001), by collecting emission spectra of 2AP triphosphate (2APTP) for a pH range and using multivariate curve resolution (MCR-AR) to calculate its pKa Fig. 5.6. The obtained pKa of 3.87 is in good agreement with the previously measured value of 3.30 in 2AP nucleoside (Gargallo et al., 2001), given that the negatively charged phosphate moiety usually shifts pKa by about 0.4 pH units (Sochacka et al., 2017). The advantage of MCR-AR is that it can potentially calculate pKa in more complicated experiments with multiple 2AP species (Gargallo et al., 2001).

However, the proposed experiment contains several major limitations. First, the base-paired neutral 2AP already has very low fluorescence intensity (Jean and Hall, 2001). It can still be used to measure pKa shifts in base pairs in small RNA/DNA duplexes

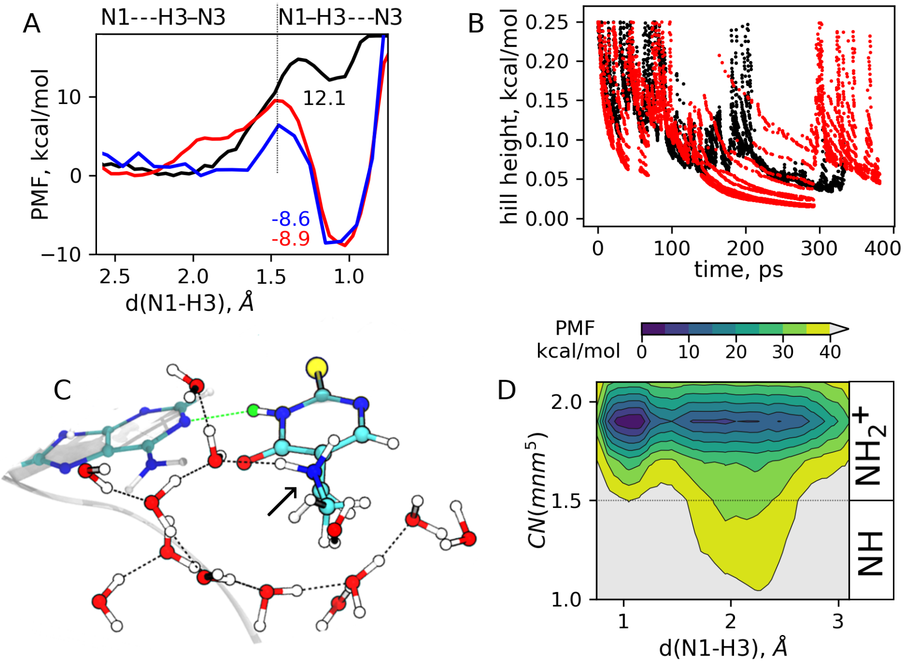


Figure 5.4.: PMF of the PT reaction in A∘Ux from QM/MM MetD simulations. **A** – PMF from well-tempered 1D MetD simulations of PT in A∘U at the second codon-anticodon position (black) and A∘Ux at the third codon-anticodon position (red), as well as 1D PMF derived from non-tempered 2D MetD simulations of PT in A∘Ux at the third codon-anticodon position with extended QM region (blue). d(N1-H3) values in the vicinity of 1 Å correspond to protonated adenine at N1 atom. **B** – convergence of well-tempered MetD simulations measured as Guassian hill height decrease. Color code corresponds to **A**. Both simulations approached convergence, but did not reach it completely. C – structure of a frame from 2D MetD simulations visualizing the first solvation shell around the mnm^5 modification (highlighted by an arrow), included in the QM region. H3 proton is highlighted in green. D – full PMF from 2D MetD simulations. Coordination number of the amino group in the mnm^5 modification (CN(mnm^5)) was used measure protonation state of the amino group. CN ≈ 2 corresponds to the protonated amino group, CN ≈ 1 corresponds to the neutral amino group.

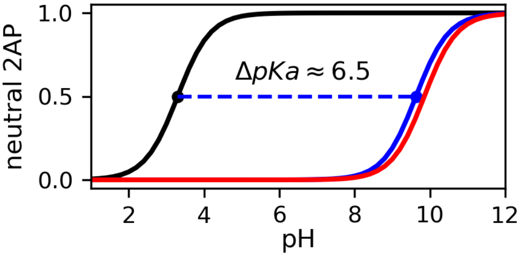


Figure 5.5.: PT mechanism results in pKa shift at N1 of A (2AP). Curves were calculated using Henderson-Hasselbalch equation using pKa(2AP) = 3.3 (Gargallo et al., 2001) and ΔpKa calculated from ΔG_f showed on Fig. 5.4A. Colors of the curves correspond to colors on Fig. 5.4A.

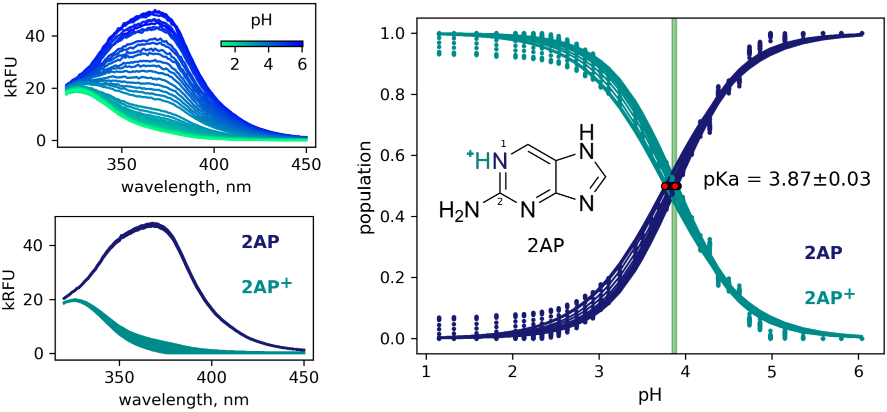


Figure 5.6.: Illustration of the MCR method to measure pKa of 2APTP. **A** – raw emission UV spectra of 2APTP in water upon pH titration (color coded). kRFU – $10^3\cdot$ relative fluorescence units **B** – optimized concentration profiles of the species. Blue and cyan curves represent sigmoid fit functions to the converged concentration matrices. Dotes of the corresponding colors denote the elements of the concentration matrices. Red dots denote midpoints of the sigmoid functions, thus representing pKa. Green rectangle is centered on the mean pKa, and its width is equal to standard deviation of pKa values. The standard deviation of 0.03 only reflects statistical error due to initial approximations.**C** – spectral components of the species from the optimized solutions

(Gargallo et al., 2001), but it could probably be completely invisible having ribosome as a background, which contains Trp residues that overlap with 2AP emission spectrum (Holz et al., 1998). This challenge can probably be overcome using fluorescently-silent ribosomes, in which Trp was substituted by 4-fluoroTrp (Bronskill and Wong, 1988). Second, the pKa shift from the PT mechanism will interfere with deprotonation of the NH^2 group somewhere in high pH range, which can make it impossible to obtain the titration curve. Obviously, the ribosome (and other components in the potential experiment) itself constitutes an obstacle for pH titration, as ribosomal proteins would denature at pH far from neutral. However, the ribosome is stable in the vicinity of neutral pH and can be titrated at least up to pH 8.5 (Johansson et al., 2011).

5.2.4 Advantages of the proposed model

As discussed in Section 1.3.2, the current model of the mechanism of ambiguous decoding facilitation by (c)mnm^{5}s^{2}U modifications is deprotonation of the modified uracil at N3 in solution, leading to G∘U$^-$x WC (or rwb) base pairs (Sochacka et al., 2017, Takai and Yokoyama, 2003). In this model, decoding rate of the wobble G-ending codon (R_G) would be proportional to the ionized fraction of the total Ux concentration:

$$R_G = R_0 K_a [Ux]_{tot} \tag{5.4}$$

where R_0 is concentration-independent cognate rate of decoding and K_a is acidity constant of Ux at N3. We can approximate the ambiguous decoding efficiency as $\lambda = R_A + R_G$, where R_A is the rate of decoding of fully-cognate A-ending codon. Ionized U$^-$x cannot form WC base pair with neutral A (Leszczynska et al., 2020). Base pairing with A$^+$, protonated at N1, would result in the same base pair stabilization as occurs upon intermolecular PT described in Section 5.2.2. However, the fraction of A$^+$ is minor at physiological pH (Gargallo et al., 2001). Therefore, ionization of Ux would decrease R_A, as it decreases the fraction of Ux that can base pair with A:

$$R_A = R_0(1-K_a)[Ux]_{tot} \tag{5.5}$$

It implies that ionization does not increase the ambiguous decoding efficiency, as R_A and R_G contributions cancel each other (Fig. 5.7):

$$\lambda = R_A + R_G = R_0[Ux]_{tot} = const \tag{5.6}$$

In this model, the increased fraction of U^{-}x facilitates decoding of G-ending codons by the expense of A-ending codons. While such model might be correct for modifications with high intrinsic acidity, such as Se-including U modifications (Leszczynska et al., 2020), it contradicts some experimental observations showing increased decoding of *both* R-ending codons by (c)mnm^{5}s^{2}U-modified anticodons (Hagervall et al., 1998, Kurata et al., 2008, Yarian et al., 2002).

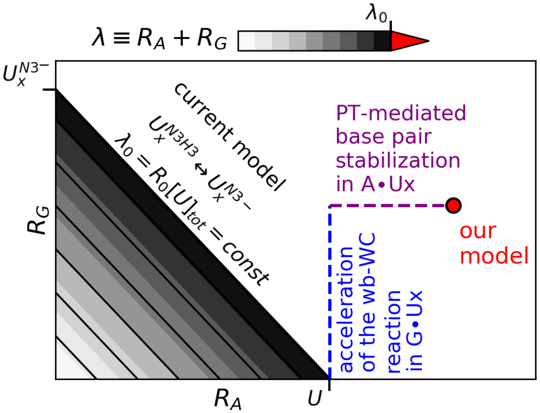

$\lambda \equiv R_A + R_G$
λ_0
U_x^{N3-}
current model
$U_x^{N3H3} \leftrightarrow U_x^{N3-}$
$\lambda_0 = R_0[U]_{tot} = const$
PT-mediated base pair stabilization in A•Ux
our model
acceleration of the wb-WC reaction in G•Ux
R_G
R_A
U

Figure 5.7.: Schematic representation of the advantages of our model. Our model is compared to the commonly accepted model (Sochacka et al., 2017, Takai and Yokoyama, 2003) based on their potential to explain increased efficiency of ambiguous decoding (λ) upon anticodon modification.

In contrast, our model does not introduce trade-off between R_A and R_G. It is based on contributions to R_A and R_G that are not mutually exclusive: PT-mediated base pair stabilization increases R_A over unmodified U, and wb-WC reaction acceleration increases R_G (Fig. 5.7). Future experiments to test these two models would need to measure Ux ionized fraction inside the ribosomal decoding site.

Conclusions

In this study we addressed a problem related to the model of base pair recognition in general: how the wb-WC tautomerization reaction in G∘U contributes to the errors of base pair recognition. We employed a computational setup based on QM/MM US simulations to calculate PMF of this reaction in various molecular environments including the open and closed states of the ribosomal decoding site. Tested by benchmark calculations, this approach revealed that the wb-WC reaction was endoergic in the open state, but exoergic in the closed state of the decoding site. We suggested a contribution to this energy difference from the decreased ε in the closed state. The same approach revealed differences between two DNA polymerases: active site of pol-β stabilized the WC geometry of G∘T, while T7-pol did not, which is consistent with available structural studies.

Our studies of the wb-WC reaction in the decoding site provided a realistic physicochemical explanation to the observations of G∘U in the WC geometry in the closed decoding site. However, a reconciliation of this model with the kinetic model of decoding was still lacking. To address this, we developed a new model, into which we explicitly incorporated the wb-WC reaction. As kinetics of the exoergic wb-WC reaction in this model is restricted by the decoding rates, this model at last unites structural and kinetic data in the field of codon-anticodon decoding: it explains the structural observations at equilibrium conditions, while allowing G∘U to be discriminated by a working ribosome. Moreover, the model predicted that the equilibration of the wb-WC reactions constraints the error rate of decoding by counteracting the equilibration of the open-closed transition of the decoding site. This prediction provides a plausible explanation to the seemingly suboptimal mechanism of decoding pictured by models where a substrate is rigid. This model illustrates the necessity to involve substrate flexibility into models of substrate recognition, similarly as enzyme flexibility was appreciated 60 years ago to produce the induce-fit model. We hope that our study provided a step towards more general models of substrate recognition.

Bibliography

Agarwal, D., Kamath, D., Gregory, S. T., and O'Connor, M. (2015). Modulation of decoding fidelity by ribosomal proteins S4 and S5. *Journal of Bacteriology*, 197(6):1017–1025.

Agris, P. F., Eruysal, E. R., Narendran, A., Väre, V. Y., Vangaveti, S., and Ranganathan, S. V. (2018). Celebrating wobble decoding: Half a century and still much is new. *RNA Biology*, 15(4-5):537–553.

Ahmadi, S., Barrios Herrera, L., Chehelamirani, M., Hostaš, J., Jalife, S., and Salahub, D. R. (2018). Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review. *International Journal of Quantum Chemistry*, 118(9):e25558.

Andersen, H. C. (1983). Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics*, 52(1):24–34.

Antony, J., Sure, R., and Grimme, S. (2015). Using dispersion-corrected density functional theory to understand supramolecular binding thermodynamics. *Chemical Communications*, 51(10):1764–1774.

Banerjee, K., Kolomeisky, A. B., and Igoshin, O. A. (2017). Elucidating interplay of speed and accuracy in biological error correction. *Proceedings of the National Academy of Sciences of the United States of America*, 114(20):5183–5188.

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical Review Letters*, 100(2):20603.

Barone, V. and Cossi, M. (1998). Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *Journal of Physical Chemistry A*, 102(11):1995–2001.

Bartlett, R. J. and Musiał, M. (2007). Coupled-cluster theory in quantum chemistry. *Reviews of Modern Physics*, 79(1):291–352.

Beard, W. A. and Wilson, S. H. (2006). Structure and mechanism of DNA polymerase β. *Chemical Reviews*, 106(2):361–382.

Bebenek, K., Pedersen, L. C., and Kunkel, T. A. (2011). Replication infidelity via a mismatch with Watson-Crick geometry. *Proceedings of the National Academy of Sciences of the United States of America*, 108(5):1862–1867.

Becke, A. D. (1993). Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652.

Best, R. B., Zhu, X., Shim, J., Lopes, P. E., Mittal, J., Feig, M., and MacKerell, A. D. (2012). Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ Dihedral Angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273.

Bhabha, G., Lee, J., Ekiert, D. C., Gam, J., Wilson, I. A., Dyson, H. J., Benkovic, S. J., and Wright, P. E. (2011). A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science*, 332(6026):234–238.

Bohr, N. (1920). Über die Serienspektra der Elemente. *Zeitschrift für Physik*, 2(5):423–469.

Bondanza, M., Nottoli, M., Cupellini, L., Lipparini, F., and Mennucci, B. (2020). Polarizable embedding QM/MM: The future gold standard for complex (bio)systems? *Physical Chemistry Chemical Physics*, 22(26):14433–14448.

Born, M. and Oppenheimer, R. (1927). Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389(20):457–484.

Bottaro, S. and Lindorff-Larsen, K. (2018). Biophysical experiments and biomolecular simulations: A perfect match? *Science*, 361(6400):355–360.

Branduardi, D., Gervasio, F. L., and Parrinello, M. (2007). From A to B in free energy space. *Journal of Chemical Physics*, 126(5):054103.

Bronskill, P. M. and Wong, J. T. (1988). Suppression of fluorescence of tryptophan residues in proteins by replacement with 4-fluorotryptophan. *The Biochemical journal*, 249(1):305–308.

Brovarets, O. O. and Hovorun, D. M. (2009). Physicochemical mechanism of the wobble DNA base pairs Gu·Thy and Ade·Cyt transition into the mismatched base pairs Gua*·Thy and Ade·Cyt* formed by the mutagenic tautomers. *Ukrainica Bioorganica Acta 2*, 7(2):12–18.

Brovarets', O. O. and Hovorun, D. M. (2015). A novel conception for spontaneous transversions caused by homo-pyrimidine DNA mismatches: A QM/QTAIM highlight. *Physical Chemistry Chemical Physics*, 17(33):21381–21388.

Brovarets, O. O. and Hovorun, D. M. (2015). How many tautomerization pathways connect Watson-Crick-like G·T DNA base mispair and wobble mismatches? *Journal of Biomolecular Structure and Dynamics*, 33(11):2297–2315.

Brovarets', O. O. and Hovorun, D. M. (2015). Tautomeric transition between wobble A·C DNA base mispair and Watson-Crick-like A·C mismatch: Microstructural mechanism and biological significance. *Physical Chemistry Chemical Physics*, 17(23):15103–15110.

Brovarets, O. O. and Hovorun, D. M. (2015). Wobble-Watson-Crick tautomeric transitions in the homo-purine DNA mismatches: A key to the intimate mechanisms of the spontaneous transversions. *Journal of Biomolecular Structure and Dynamics*, 33(12):2710–2715.

Burgen, A. S. (1981). Conformational changes and drug action. *Federation Proceedings*, 40(13):2723–2728.

Bussi, G. and Laio, A. (2020). Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2(4):200–212.

Camp, C. H. (2019). PyMCR: A python library for multivariatecurve resolution analysis with alternating regression (MCR-AR). *Journal of Research of the National Institute of Standards and Technology*, 124.

Casalino, L., Palermo, G., Abdurakhmonova, N., Rothlisberger, U., and Magistrato, A. (2017). Development of site-specific Mg2+-RNA force field parameters: A dream or reality? Guidelines from combined molecular dynamics and quantum mechanics simulations. *Journal of Chemical Theory and Computation*, 13(1):340–352.

Cerón-Carrasco, J. P., Requena, A., Zúñiga, J., Michaux, C., Perpète, E. A., and Jacquemine, D. (2009). Intermolecular proton transfer in microhydrated guanine-cytosine base Pairs: A new mechanism for spontaneous mutation in DNA. *Journal of Physical Chemistry A*, 113(39):10549–10556.

Cesari, A., Bottaro, S., Lindorff-Larsen, K., Banáš, P., Šponer, J., and Bussi, G. (2019). Fitting Corrections to an RNA Force Field Using Experimental Data. *Journal of Chemical Theory and Computation*, 15(6):3425–3431.

Chandra, A. (2000). Static dielectric constant of aqueous electrolyte solutions: is there any dynamic contribution? *Journal of Chemical Physics*, 113(3):903–905.

Changeux, J. P. and Edelstein, S. (2011). Conformational selection or induced fit? 50 Years of debate resolved. *F1000 Biology Reports*, 3(1).

Christensen, A. S., Kromann, J. C., Jensen, J. H., and Cui, Q. (2017). Intermolecular interactions in the condensed phase: Evaluation of semi-empirical quantum mechanical methods. *Journal of Chemical Physics*, 147(16):161704.

Christensen, A. S., Kubař, T., Cui, Q., and Elstner, M. (2016). Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chemical Reviews*, 116(9):5301–5337.

Cramer, C. J. (2004). *Essentials of Computational Chemistry Theories and Models Second Edition*. Wiley.

Crick, F. H. (1966). Codon—anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology*, 19(2):548–555.

Daniel, R. M., Dunn, R. V., Finney, J. L., and Smith, J. C. (2003). The role of dynamics in enzyme activity. *Annual Review of Biophysics and Biomolecular Structure*, 32(1):69–92.

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092.

Das, S., Nam, K., and Major, D. T. (2018). Rapid Convergence of Energy and Free Energy Profiles with Quantum Mechanical Size in Quantum Mechanical-Molecular Mechanical Simulations of Proton Transfer in DNA. *Journal of Chemical Theory and Computation*, 14(3):1695–1705.

Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M., and Yusupova, G. (2012). A new understanding of the decoding principle on the ribosome. *Nature*, 484(7393):256–259.

Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M., and Yusupova, G. (2013). New structural insights into the decoding mechanism: Translation infidelity via a G·U pair with Watson-Crick geometry. *FEBS Letters*, 587(13):1848–1857.

Denning, E. J., Priyakumar, U. D., Nilsson, L., and MacKerell, A. D. (2011). Impact of 2-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA. *Journal of Computational Chemistry*, 32(9):1929–1943.

Dewar, M. J., Zoebisch, E. G., Healy, E. F., and Stewart, J. J. (1985). AM1: A New General Purpose Quantum Mechanical Molecular Model1. *Journal of the American Chemical Society*, 107(13):3902–3909.

Doublié, S., Tabor, S., Long, A. M., Richardson, C. C., and Ellenberger, T. (1998). Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution. *Nature*, 391(6664):251–258.

Ehrenberg, M., Kurland, C. G., and Ruusala, T. (1986). Counting cycles of EF-Tu to measure proofreading in translation. *Biochimie*, 68(2):261–273.

Fang, W., Chen, J., Rossi, M., Feng, Y., Li, X. Z., and Michaelides, A. (2016). Inverse Temperature Dependence of Nuclear Quantum Effects in DNA Base Pairs. *Journal of Physical Chemistry Letters*, 7(11):2125–2131.

Fersht, A. R. (1974). Catalysis, binding and enzyme substrate complementarity. *Proceedings of the Royal Society of London - Biological Sciences*, 187(1089):397–407.

Fiorin, G., Klein, M. L., and Hénin, J. (2013). Using collective variables to drive molecular dynamics simulations. *Molecular Physics*, 111(22-23):3345–3362.

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993.

Fischer, N., Neumann, P., Bock, L. V., Maracci, C., Wang, Z., Paleskava, A., Konevega, A. L., Schröder, G. F., Grubmüller, H., Ficner, R., Rodnina, M. V., and Stark, H. (2016). The pathway to GTPase activation of elongation factor SelB on the ribosome. *Nature*, 540(7631):80–85.

Fislage, M., Zhang, J., Brown, Z. P., Mandava, C. S., Sanyal, S., Ehrenberg, M., and Frank, J. (2018). Cryo-EM shows stages of initial codon selection on the ribosome by aa-tRNA in ternary complex with GTP and the GTPase-deficient EF-TuH84A. *Nucleic Acids Research*, 46(11):5861–5874.

Florián, J., Hrouda, V., and Hobza, P. (1994). Proton Transfer in the Adenine-Thymine Base Pair. *Journal of the American Chemical Society*, 116(4):1457–1460.

Florian, J. and Leszczynski, J. (1996). Spontaneous DNA mutations induced by proton transfer in the guanine.cytosine base pairs: An energetic perspective. *Journal of the American Chemical Society*, 118(12):3010–3017.

Freudenthal, B. D., Beard, W. A., Shock, D. D., and Wilson, S. H. (2013). Observing a DNA polymerase choose right from wrong. *Cell*, 154(1):157.

Fröhlking, T., Bernetti, M., Calonaci, N., and Bussi, G. (2020). Toward empirical force fields that match experimental observables. *Journal of Chemical Physics*, 152(23):230902.

Gargallo, R., Vives, M., Tauler, R., and Eritja, R. (2001). Protonation studies and multivariate curve resolution on oligodeoxynucleotides carrying the mutagenic base 2-aminopurine. *Biophysical Journal*, 81(5):2886–2896.

Garofalo, R., Wohlgemuth, I., Pearson, M., Lenz, C., Urlaub, H., and Rodnina, M. V. (2019). Broad range of missense error frequencies in cellular proteins. *Nucleic Acids Research*, 47(6):2932–2945.

Gereben, O. and Pusztai, L. (2011). On the accurate calculation of the dielectric constant from molecular dynamics simulations: The case of SPC/E and SWM4-DP water. *Chemical Physics Letters*, 507(1-3):80–83.

Gianni, S., Dogan, J., and Jemth, P. (2014). Distinguishing induced fit from conformational selection. *Biophysical Chemistry*, 189:33–39.

Goldenfeld, N. and Woese, C. (2011). Life is Physics: Evolution as a Collective Phenomenon Far From Equilibrium. *Annual Review of Condensed Matter Physics*, 2(1):375–399.

Goodman, M. F. (1997). Hydrogen bonding revisited: Geometric selection as a principal determinant of DNA replication fidelity. *Proceedings of the National Academy of Sciences of the United States of America*, 94(20):10493–10495.

Gromadski, K. B., Daviter, T., and Rodnina, M. V. (2006). A uniform response to mismatches in codon-anticodon complexes ensures ribosomal fidelity. *Molecular Cell*, 21(3):369–377.

Gromadski, K. B. and Rodnina, M. V. (2004). Kinetic Determinants of High-Fidelity tRNA Discrimination on the Ribosome. *Molecular Cell*, 13(2):191–200.

Grosjean, H., de Crécy-Lagard, V., and Marck, C. (2010). Deciphering synonymous codons in the three domains of life: Co-evolution with specific tRNA modification enzymes. *FEBS Letters*, 584(2):252–264.

Grosjean, H. and Westhof, E. (2016). An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Research*, 44(17):8020–8040.

Gu, X., Mooers, B. H., Thomas, L. M., Malone, J., Harris, S., and Schroeder, S. J. (2015). Structures and Energetics of Four Adjacent G·U Pairs That Stabilize an RNA Helix. *Journal of Physical Chemistry B*, 119(42):13252–13261.

Guo, Y., Riplinger, C., Becker, U., Liakos, D. G., Minenkov, Y., Cavallo, L., and Neese, F. (2018). Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [DLPNO-CCSD(T)]. *Journal of Chemical Physics*, 148(1):011101.

Hagervall, T. G., Pomerantz, S. C., and McCloskey, J. A. (1998). Reduced misreading of asparagine codons by Escherichia coli tRNA(Lys) with hypomodified derivatives of 5-methylaminomethyl-2-thiouridine in the wobble position. *Journal of Molecular Biology*, 284(1):33–42.

Hartono, Y. D., Ito, M., Villa, A., and Nilsson, L. (2018). Computational Study of Uracil Tautomeric Forms in the Ribosome: The Case of Uracil and 5-Oxyacetic Acid Uracil in the First Anticodon Position of tRNA. *Journal of Physical Chemistry B*, 122(3):1152–1160.

Henkelman, G., Uberuaga, B. P., and Jónsson, H. (2000). Climbing image nudged elastic band method for finding saddle points and minimum energy paths. *Journal of Chemical Physics*, 113(22):9901–9904.

Herschlag, D. (1988). The role of induced fit and conformational changes of enzymes in specificity and catalysis. *Bioorganic Chemistry*, 16(1):62–96.

Hershberg, R. (2015). Mutation—the engine of evolution: Studying mutation and its role in the evolution of bacteria. *Cold Spring Harbor Perspectives in Biology*, 7(9):a018077.

Himo, F. (2017). Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions. *Journal of the American Chemical Society*, 139(20):6780–6786.

Hoffer, E. D., Maehigashi, T., Fredrick, K., and Dunham, C. M. (2019). Ribosomal ambiguity (ram) mutations promote the open (off) to closed (on) transition and thereby increase miscoding. *Nucleic Acids Research*, 47(3):1557–1563.

Hohenberg, P. and Kohn, W. (1964). Inhomogeneous electron gas. *Physical Review*, 136(3B):B864.

Holz, B., Klimasauskas, S., Serva, S., and Weinhold, E. (1998). 2-Aminopurine as a fluorescent probe for DNA base flipping by methyltransferases. *Nucleic Acids Research*, 26(4):1076–1083.

Hopefield, J. J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proceedings of the National Academy of Sciences of the United States of America*, 71(10):4135–4139.

Hu, X., Li, H., Liang, W., and Han, S. (2004). Theoretical study of the proton transfer of uracil and (water)n (n = 0-4): Water stabilization and mutagenicity for uracil. *Journal of Physical Chemistry B*, 108(34):12999–13007.

Huang, J. and Mackerell, A. D. (2013). CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25):2135–2145.

Huber, T., Torda, A. E., and van Gunsteren, W. F. (1994). Local elevation: A method for improving the searching properties of molecular dynamics simulation. *Journal of Computer-Aided Molecular Design*, 8(6):695–708.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38.

Jacquemin, D., Zúñiga, J., Requena, A., and Céron-Carrasco, J. P. (2014). Assessing the importance of proton transfer reactions in DNA. *Accounts of Chemical Research*, 47(8):2467–2474.

Janoš, P., Trnka, T., Kozmon, S., Tvaroška, I., and Koča, J. (2016). Different QM/MM Approaches To Elucidate Enzymatic Reactions: Case Study on ppGalNAcT2. *Journal of Chemical Theory and Computation*, 12(12):6062–6076.

Jean, J. M. and Hall, K. B. (2001). 2-Aminopurine fluorescence quenching and lifetimes: Role of base stacking. *Proceedings of the National Academy of Sciences of the United States of America*, 98(1):37–41.

Jensen, F. (2007). *Introduction to Computational Chemistry*. Wiley.

Jing, Z., Liu, C., Cheng, S. Y., Qi, R., Walker, B. D., Piquemal, J. P., and Ren, P. (2019). Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annual Review of Biophysics*, 48(1):371–394.

Johansson, M., Ieong, K. W., Trobro, S., Strazewski, P., Åqvist, J., Pavlov, M. Y., and Ehrenberg, M. (2011). pH-sensitivity of the ribosomal peptidyl transfer reaction dependent on the identity of the A-site aminoacyl-tRNA. *Proceedings of the National Academy of Sciences of the United States of America*, 108(1):79–84.

Johansson, M., Lovmar, M., and Ehrenberg, M. (2008). Rate and accuracy of bacterial protein synthesis revisited. *Current Opinion in Microbiology*, 11(2):141–147.

Johansson, M., Zhang, J., and Ehrenberg, M. (2012). Genetic code translation displays a linear trade-off between efficiency and accuracy of tRNA selection. *Proceedings of the National Academy of Sciences of the United States of America*, 109(1):131–136.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935.

Kamerlin, S. C. and Warshel, A. (2010). At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis. *Proteins: Structure, Function and Bioinformatics*, 78(6):1339–1375.

Kästner, J. (2011). Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6):932–942.

Kazantsev, A. and Ignatova, Z. (2020). Tautomerization constraints the accuracy of codon-anticodon decoding. *bioRxiv*, page 2020.10.19.344408.

Ke, Z., Mallik, P., Johnson, A. B., Luna, F., Nevo, E., Zhang, Z. D., Gladyshev, V. N., Seluanov, A., and Gorbunova, V. (2017). Translation fidelity coevolves with longevity. *Aging Cell*, 16(5):988–993.

Khade, P. K., Shi, X., and Joseph, S. (2013). Steric complementarity in the decoding center is important for tRNA selection by the ribosome. *Journal of Molecular Biology*, 425(20):3778–3789.

Kimsey, I. and Al-Hashimi, H. M. (2014). Increasing occurrences and functional roles for high energy purine-pyrimidine base-pairs in nucleic acids. *Current Opinion in Structural Biology*, 24(1):72–80.

Kimsey, I. J., Petzold, K., Sathyamoorthy, B., Stein, Z. W., and Al-Hashimi, H. M. (2015). Visualizing transient Watson-Crick-like mispairs in DNA and RNA duplexes. *Nature*, 519(7543):315–320.

Kimsey, I. J., Szymanski, E. S., Zahurancik, W. J., Shakya, A., Xue, Y., Chu, C. C., Sathyamoorthy, B., Suo, Z., and Al-Hashimi, H. M. (2018). Dynamic basis for dG· dT misincorporation via tautomerization and ionization. *Nature*, 554(7691):195–201.

Klinman, J. P. and Kohen, A. (2013). Hydrogen tunneling links protein dynamics to enzyme catalysis. *Annual Review of Biochemistry*, 82(1):471–496.

Koag, M. C., Nam, K., and Lee, S. (2014). The spontaneous replication error and the mismatch discrimination mechanisms of human DNA polymerase β. *Nucleic Acids Research*, 42(17):11233–11245.

Kohn, W. and Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133.

Kolafa, J. and Viererblová, L. (2014). Static dielectric constant from simulations revisited: Fluctuations or external field? *Journal of Chemical Theory and Computation*, 10(4):1468–1476.

Kool, E. T. (2002). Active site tightness and substrate fit in DNA replication. *Annual Review of Biochemistry*, 71(1):191–219.

Koshland, D. E. (1959). Enzyme flexibility and enzyme action. *Journal of cellular and comparative physiology*, 54(S1):245–258.

Koshland, D. E. (1995). The Key–Lock Theory and the Induced Fit Theory. *Angewandte Chemie International Edition in English*, 33(23-24):2375–2378.

Kramer, E. B. and Farabaugh, P. J. (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *Rna*, 13(1):87–96.

Krüger, M. K., Pedersen, S., Hagervall, T. G., and Sørensen, M. A. (1998). The modification of the wobble base of tRNA(Glu) modulates the translation rate of glutamic acid codons in vivo. *Journal of Molecular Biology*, 284(3):621–631.

Kruse, H., Goerigk, L., and Grimme, S. (2012). Why the standard B3LYP/6-31G* model chemistry should not be used in DFT calculations of molecular thermochemistry: Understanding and correcting the problem. *Journal of Organic Chemistry*, 77(23):10824–10834.

Kührová, P., Mlýnský, V., Zgarbová, M., Krepl, M., Bussi, G., Best, R. B., Otyepka, M., Šponer, J., and Banáš, P. (2019). Improving the Performance of the Amber RNA Force Field by Tuning the Hydrogen-Bonding Interactions. *Journal of Chemical Theory and Computation*, 15(5):3288–3305.

Kulik, H. J., Zhang, J., Klinman, J. P., and Martínez, T. J. (2016). How large should the QM region be in QM/MM calculations? the case of catechol O-methyltransferase. *Journal of Physical Chemistry B*, 120(44):11381–11394.

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992). THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021.

Kurata, S., Weixlbaumer, A., Ohtsuki, T., Shimazaki, T., Wada, T., Kirino, Y., Takai, K., Watanabe, K., Ramakrishnan, V., and Suzuki, T. (2008). Modified uridines with C5-methylene substituents at the first position of the tRNA anticodon stabilize U·G wobble pairing during decoding. *Journal of Biological Chemistry*, 283(27):18801–18811.

Larsen, A. T., Fahrenbach, A. C., Sheng, J., Pian, J., and Szostak, J. W. (2015). Thermodynamic insights into 2-thiouridine-enhanced RNA hybridization. *Nucleic Acids Research*, 43(16):7675–7687.

Lee, I. and Berdis, A. J. (2010). Non-natural nucleotides as probes for the mechanism and fidelity of DNA polymerases. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1804(5):1064–1080.

Lemkul, J. A. and MacKerell, A. D. (2018). Polarizable force field for RNA based on the classical drude oscillator. *Journal of Computational Chemistry*, 39(32):2624–2646.

Leontis, N. B., Stombaugh, J., and Westhof, E. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Research*, 30(16):3497–3531.

Leszczynska, G., Cypryk, M., Gostynski, B., Sadowska, K., Herman, P., Bujacz, G., Lodyga-Chruscinska, E., Sochacka, E., and Nawrot, B. (2020). C5-substituted 2-selenouridines ensure efficient base pairing with guanosine; consequences for reading the NNG-3 synonymous mRNA codons. *International Journal of Molecular Sciences*, 21(8):2882.

Li, D. and Ai, H. (2009). Catalysis effects of water molecules and of charge on intramolecular proton transfer of uracil. *Journal of Physical Chemistry B*, 113(34):11732–11742.

Li, P., Rangadurai, A., Al-Hashimi, H. M., and Hammes-Schiffer, S. (2020). Environmental Effects on Guanine-Thymine Mispair Tautomerization Explored with Quantum Mechanical/Molecular Mechanical Free Energy Simulations. *Journal of the American Chemical Society*, 142(25):11183–11191.

Loco, D., Lagardère, L., Caprasecca, S., Lipparini, F., Mennucci, B., and Piquemal, J. P. (2017). Hybrid QM/MM Molecular Dynamics with AMOEBA Polarizable Embedding. *Journal of Chemical Theory and Computation*, 13(9):4025–4033.

Loco, D., Lagardère, L., Cisneros, G. A., Scalmani, G., Frisch, M., Lipparini, F., Mennucci, B., and Piquemal, J. P. (2019). Towards large scale hybrid QM/MM dynamics of complex systems with advanced point dipole polarizable embeddings. *Chemical Science*, 10(30):7200–7211.

Loveland, A. B., Demo, G., Grigorieff, N., and Korostelev, A. A. (2017). Ensemble cryo-EM elucidates the mechanism of translation fidelity. *Nature*, 546(7656):113–117.

Loveland, A. B., Demo, G., and Korostelev, A. A. (2020). Cryo-EM of elongating ribosome with EF-Tu•GTP elucidates tRNA proofreading. *Nature*, 584(7822):640–645.

Löwdin, P. O. (1963). Proton tunneling in DNA and its biological implications. *Reviews of Modern Physics*, 35(3):724–732.

Lu, T. and Chen, F. (2012). Multiwfn: A multifunctional wavefunction analyzer. *Journal of Computational Chemistry*, 33(5):580–592.

Machnicka, M. A., Milanowska, K., Oglou, O. O., Purta, E., Kurkowska, M., Olchowik, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K. M., Helm, M., Bujnicki, J. M., and Grosjean, H. (2013). MODOMICS: A database of RNA modification pathways - 2013 update. *Nucleic Acids Research*, 41(D1):D262–D267.

Mallory, J. D., Kolomeisky, A. B., and Igoshin, O. A. (2019). Trade-Offs between Error, Speed, Noise, and Energy Dissipation in Biological Processes with Proofreading. *Journal of Physical Chemistry B*, 123(22):4718–4725.

Manickam, N., Nag, N., Abbasi, A., Patel, K., and Farabaugh, P. J. (2014). Studies of translational misreading in vivo show that the ribosome very efficiently discriminates against most potential errors. *Rna*, 20(1):9–15.

Marcos-Alcalde, I., Setoain, J., Mendieta-Moreno, J. I., Mendieta, J., and Gómez-Puertas, P. (2015). MEPSA: Minimum energy pathway analysis for energy landscapes. *Bioinformatics*, 31(23):3853–3855.

Mardirossian, N. and Head-Gordon, M. (2017). Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. *Molecular Physics*, 115(19):2315–2372.

Markland, T. E. and Ceriotti, M. (2018). Nuclear quantum effects enter the mainstream. *arXiv*, 2(3):109.

Maximoff, S. N., Kamerlin, S. C. L., and Florián, J. (2017). DNA Polymerase λ Active Site Favors a Mutagenic Mispair between the Enol Form of Deoxyguanosine Triphosphate Substrate and the Keto Form of Thymidine Template: A Free Energy Perturbation Study. *Journal of Physical Chemistry B*, 121(33):7813–7822.

Melo, M. C., Bernardi, R. C., Rudack, T., Scheurer, M., Riplinger, C., Phillips, J. C., Maia, J. D., Rocha, G. B., Ribeiro, J. V., Stone, J. E., Neese, F., Schulten, K., and Luthey-Schulten, Z. (2018). NAMD goes quantum: An integrative suite for hybrid simulations. *Nature Methods*, 15(5):351–354.

Miyamoto, S. and Kollman, P. A. (1992). Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):952–962.

Mizuno, H. and Sundaralingam, M. (1978). Stacking of crick wobble pair and watson-crick pair: Stability rules of G-U pairs at ends of helical stems in tRNAs and the relation to codon-anticodon wobble interaction. *Nucleic Acids Research*, 5(11):4451–4462.

Moran, S., Ren, R. X., and Kool, E. T. (1997). A thymidine triphosphate shape analog lacking Watson-Crick pairing ability is replicated with high sequence selectivity. *Proceedings of the National Academy of Sciences of the United States of America*, 94(20):10506–10511.

Mordret, E., Dahan, O., Asraf, O., Rak, R., Yehonadav, A., Barnabas, G. D., Cox, J., Geiger, T., Lindner, A. B., and Pilpel, Y. (2019). Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. *Molecular Cell*, 75(3):427–441.

Morse, J. C., Girodat, D., Burnett, B. J., Holm, M., Altman, R. B., Sanbonmatsu, K. Y., Wieden, H. J., and Blanchard, S. C. (2020). Elongation factor-Tu can repetitively engage aminoacyl-tRNA within the ribosome during the proofreading stage of tRNA selection. *Proceedings of the National Academy of Sciences of the United States of America*, 117(7):3610–3620.

Murphy, F. V., Ramakrishnan, V., Malkiewicz, A., and Agris, P. F. (2004). The role of modifications in codon discrimination by tRNALysUUU. *Nature Structural and Molecular Biology*, 11(12):1186–1191.

Neese, F. (2018). Software update: the ORCA program system, version 4.0. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1).

Nierhaus, K. H. (2014). Mg2+, K+, and the ribosome. *Journal of Bacteriology*, 196(22):3817–3819.

Nomura, K., Hoshino, R., Shimizu, E., Hoshiba, Y., Danilov, V. I., and Kurita, N. (2013). DFT Calculations on the Effect of Solvation on the Tautomeric Reactions for Wobble Gua-Thy and Canonical Gua-Cyt Base-Pairs. *Journal of Modern Physics*, 04(03):422–431.

Ogle, J. M., Brodersen, D. E., Clemons, J., Tarry, M. J., Carter, A. P., and Ramakrishnan, V. (2001). Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science*, 292(5518):897–902.

Ogle, J. M., Murphy IV, F. V., Tarry, M. J., and Ramakrishnan, V. (2002). Selection of tRNA by the ribosome requires a transition from an open to a closed form. *Cell*, 111(5):721–732.

Padermshoke, A., Katsumoto, Y., Masaki, R., and Aida, M. (2008). Thermally induced double proton transfer in GG and wobble GT base pairs: A possible origin of the mutagenic guanine. *Chemical Physics Letters*, 457(1-3):232–236.

Pape, T., Wintermeyer, W., and Rodnina, M. (1999). Induced fit in initial selection and proofreading of aminoacyl-tRNA on the ribosome. *EMBO Journal*, 18(13):3800–3807.

Pavlov, M. Y. and Ehrenberg, M. (2018). Substrate-Induced Formation of Ribosomal Decoding Center for Accurate and Rapid Genetic Code Translation. *Annual Review of Biophysics*, 47(1):525–548.

Pavlov, M. Y., Liljas, A., and Ehrenberg, M. (2017). A recent intermezzo at the Ribosome Club. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1716).

Pernod, K., Schaeffer, L., Chicher, J., Hok, E., Rick, C., Geslain, R., Eriani, G., Westhof, E., Ryckelynck, M., and Martin, F. (2020). The nature of the purine at position 34 in tRNAs of 4-codon boxes is correlated with nucleotides at positions 32 and 38 to maintain decoding fidelity. *Nucleic acids research*, 48(11):6170–6183.

Phillips, J. C., Hardy, D. J., Maia, J. D., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., McGreevy, R., Melo, M. C., Radak, B. K., Skeel, R. D., Singharoy, A., Wang, Y., Roux, B., Aksimentiev, A., Luthey-Schulten, Z., Kalé, L. V., Schulten, K., Chipot, C., and Tajkhorshid, E. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. *Journal of Chemical Physics*, 153(4):044130.

Pitera, J. W., Falta, M., and Van Gunsteren, W. F. (2001). Dielectric properties of proteins from simulation: The effects of solvent, ligands, pH, and temperature. *Biophysical Journal*, 80(6):2546–2555.

Pusuluk, O., Farrow, T., Deliduman, C., Burnett, K., and Vedral, V. (2018). Proton tunnelling in hydrogen bonds and its implications in an induced-fit model of enzyme catalysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2218):20180037.

Raghavachari, K., Trucks, G. W., Pople, J. A., and Head-Gordon, M. (1989). A fifth-order perturbation comparison of electron correlation theories. *Chemical Physics Letters*, 157(6):479–483.

Ramanathan, A. and Agarwal, P. K. (2011). Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. *PLoS Biology*, 9(11):e1001193.

Ranjan, N. and Rodnina, M. V. (2017). Thio-Modification of tRNA at the Wobble Position as Regulator of the Kinetics of Decoding and Translocation on the Ribosome. *Journal of the American Chemical Society*, 139(16):5857–5864.

Robbins, T. J. and Wang, Y. (2013). Effect of initial ion positions on the interactions of monovalent and divalent ions with a DNA duplex as revealed with atomistic molecular dynamics simulations. *Journal of Biomolecular Structure and Dynamics*, 31(11):1311–1323.

Rodnina, M. V., Fischer, N., Maracci, C., and Stark, H. (2017). Ribosome dynamics during decoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1716).

Rodriguez-Hernandez, A., Spears, J. L., Gaston, K. W., Limbach, P. A., Gamper, H., Hou, Y. M., Kaiser, R., Agris, P. F., and Perona, J. J. (2013). Structural and mechanistic basis for enhanced translational efficiency by 2-thiouridine at the tRNA anticodon wobble position. *Journal of Molecular Biology*, 425(20):3888–3906.

Roux, B. (1995). The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1-3):275–282.

Rozov, A., Demeshkina, N., Khusainov, I., Westhof, E., Yusupov, M., and Yusupova, G. (2016a). Novel base-pairing interactions at the tRNA wobble position crucial for accurate reading of the genetic code. *Nature Communications*, 7.

Rozov, A., Demeshkina, N., Westhof, E., Yusupov, M., and Yusupova, G. (2015). Structural insights into the translational infidelity mechanism. *Nature Communications*, 6:7251.

Rozov, A., Westhof, E., Yusupov, M., and Yusupova, G. (2016b). The ribosome prohibits the G•U wobble geometry at the first position of the codon-anticodon helix. *Nucleic Acids Research*, 44(13):6434–6441.

Rozov, A., Wolff, P., Grosjean, H., Yusupov, M., Yusupova, G., and Westhof, E. (2018). Tautomeric G•U pairs within the molecular ribosomal grip and fidelity of decoding in bacteria. *Nucleic Acids Research*, 46(14):7425–7435.

Rudorf, S., Thommen, M., Rodnina, M. V., and Lipowsky, R. (2014). Deducing the Kinetics of Protein Synthesis In Vivo from the Transition Rates Measured In Vitro. *PLoS Computational Biology*, 10(10):e1003909.

Sanbonmatsu, K. Y. (2014). Flipping through the Genetic Code: New Developments in Discrimination between Cognate and Near-Cognate tRNAs and the Effect of Antibiotics. *Journal of Molecular Biology*, 426(19):3197–3200.

Santos, M., Pereira, P. M., Varanda, A. S., Carvalho, J., Azevedo, M., Mateus, D. D., Mendes, N., Oliveira, P., Trindade, F., Pinto, M. T., Bordeira-Carriço, R., Carneiro, F., Vitorino, R., Oliveira, C., and Santos, M. A. (2018). Codon misreading tRNAs promote tumor growth in mice. *RNA Biology*, 15(6):773–786.

Sartori, P. and Pigolotti, S. (2013). Kinetic versus energetic discrimination in biological copying. *Physical Review Letters*, 110(18):188101.

Satpati, P. and Åqvist, J. (2014). Why base tautomerization does not cause errors in mRNA decoding on the ribosome. *Nucleic acids research*, 42(20):12876–12884.

Savir, Y. and Tiusty, T. (2007). Conformational proofreading: The impact of conformational changes on the specificity of molecular recognition. *PLoS ONE*, 2(5):e468.

Savir, Y. and Tlusty, T. (2013). The ribosome as an optimal decoder: A lesson in molecular recognition. *Cell*, 153(2):471–479.

Schrode, P., Huter, P., Clementi, N., and Erlacher, M. (2017). Atomic mutagenesis at the ribosomal decoding site. *RNA Biology*, 14(1):104–112.

Singh, V., Fedeles, B. I., and Essigmann, J. M. (2015). Role of tautomerism in RNA biochemistry. *Rna*, 21(1):1–13.

Sochacka, E., Lodyga-Chruscinska, E., Pawlak, J., Cypryk, M., Bartos, P., Ebenryter-Olbinska, K., Leszczynska, G., and Nawrot, B. (2017). C5-substituents of uridines and 2-thiouridines present at the wobble position of tRNA determine the formation of their keto-enol or zwitterionic forms - A factor important for accuracy of reading of guanosine at the 3'-end of the mRNA codons. *Nucleic Acids Research*, 45(8):4825–4836.

Sponer, J., Bussi, G., Krepl, M., Banas, P., Bottaro, S., Cunha, R. A., Gil-Ley, A., Pinamonti, G., Poblete, S., Jurečka, P., Walter, N. G., and Otyepka, M. (2018). RNA structural dynamics as captured by molecular simulations: A comprehensive overview. *Chemical Reviews*, 118(8):4177–4338.

Stewart, J. J. (1989). Optimization of parameters for semiempirical methods I. Method. *Journal of Computational Chemistry*, 10(2):209–220.

Stewart, J. J. (2013). Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters. *Journal of Molecular Modeling*, 19(1):1–32.

Stewart, J. J. (2016). Mopac2016.

Stewart, J. J. (2017). An investigation into the applicability of the semiempirical method PM7 for modeling the catalytic mechanism in the enzyme chymotrypsin. *Journal of Molecular Modeling*, 23(5):154.

Takai, K. and Yokoyama, S. (2003). Roles of 5-substituents of tRNA wobble uridines in the recognition of purine-ending codons. *Nucleic Acids Research*, 31(22):6383–6391.

Thompson, R. C. (1988). EFTu provides an internal kinetic standard for translational accuracy. *Trends in Biochemical Sciences*, 13(3):91–93.

Thompson, R. C. and Karim, A. M. (1982). The accuracy of protein biosynthesis is limited by its speed: High fidelity selection by ribosomes of aminoacyl-tRNA ternary complexes containing GTP[γS]. *Proceedings of the National Academy of Sciences of the United States of America*, 79(16 I):4922–4926.

Topal, M. D. and Fresco, J. R. (1976a). Base pairing and fidelity in codon-anticodon interaction. *Nature*, 263(5575):289–293.

Topal, M. D. and Fresco, J. R. (1976b). Complementary base pairing and the origin of substitution mutations. *Nature*, 263(5575):285–289.

Torrie, G. M. and Valleau, J. P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199.

Tsai, Y. C. and Johnson, K. A. (2006). A new paradigm for DNA polymerase specificity. *Biochemistry*, 45(32):9675–9687.

Valsson, O., Tiwary, P., and Parrinello, M. (2016). Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annual Review of Physical Chemistry*, 67(1):159–184.

Varani, G. and McClain, W. H. (2000). The G·U wobble base pair: A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Reports*, 1(1):18–23.

Wang, W., Hellinga, H. W., and Beese, L. S. (2011). Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 108(43):17644–17648.

Watson, J. D. and Crick, F. H. (1953a). Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–967.

Watson, J. D. and Crick, F. H. (1953b). The structure of DNA. *Cold Spring Harbor symposia on quantitative biology*, 18:123–131.

Wentzell, P. D., Karakach, T. K., Roy, S., Juanita, M. J., Allen, C. P., and Werner-Washburne, M. (2006). Multivariate curve resolution of time course microarray data. *BMC Bioinformatics*, 7(1):343.

Westhof, E. (2014). Isostericity and tautomerism of base pairs in nucleic acids. *FEBS Letters*, 588(15):2464–2469.

Westhof, E., Yusupov, M., and Yusupova, G. (2014). Recognition of Watson-Crick base pairs: Constraints and limits due to geometric selection and tautomerism. *F1000Prime Reports*, 6:19.

Westhof, E., Yusupov, M., and Yusupova, G. (2019). The multiple flavors of GoU pairs in RNA. *Journal of Molecular Recognition*, 32(8):e2782.

Wohlgemuth, I., Pohl, C., Mittelstaet, J., Konevega, A. L., and Rodnina, M. V. (2011). Evolutionary optimization of speed and accuracy of decoding on the ribosome. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1580):2979–2986.

Xu, Y., Vanommeslaeghe, K., Aleksandrov, A., MacKerell, A. D., and Nilsson, L. (2016). Additive CHARMM force field for naturally occurring modified ribonucleotides. *Journal of Computational Chemistry*, 37(10):896–912.

Yang, L., Weerasinghe, S., Smith, P. E., and Pettitt, B. M. (1995). Dielectric response of triplex DNA in ionic solution from simulations. *Biophysical Journal*, 69(4):1519–1527.

Yarian, C., Townsend, H., Czestkowski, W., Sochacka, E., Malkiewicz, A. J., Guenther, R., Miskiewicz, A., and Agris, P. F. (2002). Accurate translation of the genetic code depends on tRNA modified nucleosides. *Journal of Biological Chemistry*, 277(19):16391–16395.

Zaher, H. S. and Green, R. (2010). Hyperaccurate and Error-Prone Ribosomes Exploit Distinct Mechanisms during tRNA Selection. *Molecular Cell*, 39(1):110–120.

Zeidler, W., Egle, C., Ribeiro, S., Wagner, A., Katunin, V., Kreutzer, R., Rodnina, M., Wintermeyer, W., and Sprinzl, M. (1995). Site-Directed Mutagenesis of Thermus thermophilus Elongation Factor Tu: Replacement of His85, Asp81 and Arg300. *European Journal of Biochemistry*, 229(3):596–604.

Zeng, X., Chugh, J., Casiano-Negroni, A., Al-Hashimi, H. M., and Brooks, C. L. (2014). Flipping of the ribosomal A-site adenines provides a basis for tRNA selection. *Journal of molecular biology*, 426(19):3201–3213.

Zhang, J., Ieong, K. W., Johansson, M., and Ehrenberg, M. (2015). Accuracy of initial codon selection by aminoacyl-tRNAs on the mRNA-programmed bacterial ribosome. *Proceedings of the National Academy of Sciences of the United States of America*, 112(31):9602–9607.

Zhang, J., Ieong, K. W., Mellenius, H., and Ehrenberg, M. (2016). Proofreading neutralizes potential error hotspots in genetic code translation by transfer RNAs. *Rna*, 22(6):896–904.

Zhang, J., Pavlov, M. Y., and Ehrenberg, M. (2018). Accuracy of genetic code translation and its orthogonal corruption by aminoglycosides and Mg2+ ions. *Nucleic Acids Research*, 46(3):1362–1374.

Zhang, Z., Shah, B., and Bondarenko, P. V. (2013). G/U and certain wobble position mismatches as possible main causes of amino acid misincorporations. *Biochemistry*, 52(45):8165–8176.

Zheng, H., Shabalin, I. G., Handing, K. B., Bujnicki, J. M., and Minor, W. (2015). Magnesium-binding architectures in RNA crystal structures: Validation, binding preferences, classification and motif detection. *Nucleic Acids Research*, 43(7):3789–3801.

Zoete, V. and Meuwly, M. (2004). Double proton transfer in the isolated and DMA-embedded guanine-cytosine base pair. *Journal of Chemical Physics*, 121(9):4377–4388.

Appendix A

Derivation of Eq. (4.4)

Let us again revisit the error rate of the classical system:

$$\eta_0 = \frac{R^{nc}}{R^c} = \frac{(k_{cat}/K_m)^{nc}}{(k_{cat}/K_m)^c} = \frac{k_4^{nc}[C4_{nc}]}{k_4^c[C4_c]} \qquad \text{(1.1 revisited)}$$

where R^i is the rate of decoding, k^i_{cat} is the catalytic rate constant, K^i_m is the Michaelis-Menten constant, $[C4_i]$ is the steady-state concentration of $C4$ state, and k^i_4 is the rate constant of GTPase activation, for $i = c$ (cognate), nc (near-cognate).
According to Pavlov and Ehrenberg (2018), R^i can be expressed in terms of rate constants as following:

$$R^i = \frac{k_1}{1 + a^i_2(1 + a^i_3(1 + a^i_4))} \qquad \text{(A.1)}$$

where $a^i_2 = q_2/k_2$, $a^i_3 = q^i_3/k_3$, $a^i_4 = q^i_4/k^i_4$.
Now let us consider the decoding scheme on Fig. 4.1. R^{nc} can be expressed as following :

$$k_4^{nc}[C4_{nc}] = k_4^{wb}[C4^{wb}_{nc}] + k_4^{wc}[C4^{wc}_{nc}] \qquad \text{(A.2)}$$

Given $k_4^{wb} = 0$ *and* $k_4^{wc} = k_4^c$ (see main text) and from Eq. (A.2), we obtain Eq. (4.3). From Eq. (1.1) and Eq. (4.3), the error from the wb-WC reaction in the $C4$ state η:

$$\eta = \frac{[C4_{nc}]}{[C4_c]} P_{wc} \qquad \text{(A.3)}$$

Now, we can write P_{wc} (WC population in state $C4_{nc}$) in terms of forward and reverse rate constants of the tautomerization reaction (k_f, k_r) using equation for product concentration in reversible first order chemical reaction at time τ:

$$P_{WC} = P^{eq}_{wc} + (P^0_{wc} - P^{eq}_{wc}) \exp\left(-(k_f + k_r)\tau\right) \qquad \text{(A.4)}$$

where P_{wc}^{eq} is the equilibrium WC population in $C4_{nc}$ for a given (k_f, k_r), and P_{wc}^{0} is the initial WC population in $C4_{nc}$. The meaning of P_{wc}^{0} is the contribution of P_{wc}^{C3} to P_{wc}. Such contribution is affected by the relative forward and reverse $C3 \leftrightarrow C4$ rates between wb and WC "branches". Therefore, P_{wc}^{0} is the WC population in $C3$ state, kinetically partitioned into $C4$ state:

$$P_{wc}^{0} = P_{wc}^{C3} \frac{K_{C3\to C4}^{wc}}{K_{C3\to C4}^{wb}} = P_{wc}^{C3} \frac{q_4^{nc}}{k_4^c + q_4^c} \tag{A.5}$$

where $K_{C3\to C4}^{wc} = \frac{k_3}{k_4^c + q_4^c}$ and $K_{C3\to C4}^{wb} = \frac{k_3}{q_4^{nc}}$.
τ has a meaning of a lifetime over which the product can form. The residence time of $C4_{nc}^{wc}$ state (see Fig. 4.1):

$$\tau = \frac{1}{k_4^c + q_4^c} \tag{A.6}$$

From Eq. (A.3), Eq. (A.4) and Eq. (A.6) we obtain Eq. (4.4):

$$\eta = \frac{[C4_{nc}]}{[C4_c]} \left(P_{WC}^{eq} + (P_{WC}^{C3} \frac{q_4^{nc}}{k_4^c + q_4^c} - P_{WC}^{eq}) \exp{(-\frac{k_f + k_r}{k_4^c + q_4^c})} \right) \tag{4.4 revisited}$$

When calculating η by equation Eq. (4.4) using Eq. (A.1), k_4^{nc} should be substituted with $P_{wc} k_4^c$ according to Eq. (4.3).

Appendix B

Linear approximation of Eq. (4.4)

Linear approximations were used to visualize the cancelling of $\eta(k_4^c)$ dependence in the slow kinetic regime of the wb-WC reaction, and to derive Eq. (4.5). For the linear approximation, we assumed the slow kinetic regime of the reaction. With this assumption, $P_{wc}^{eq} = 1$ since ΔG_{wc} is negative. Furthermore, k_r in the numerator of the exponential term in Eq. (4.4) can be omitted since $k_f >> k_r$. To simplify the solution even further, we assumed $q_4^c << k_4^c$, and thus we can approximate $\tau = 1/k_4^c$ and $P_{wc}^0 = P_{wc}^{C3} \frac{q_4^{nc}}{k_4^c}$. Under these assumptions we can write:

$$P_{wc} = 1 + (P_{wc}^0 - 1)\exp\left(-\frac{k_f}{k_4^c}\right) \tag{A.1}$$

To derive the linear approximation from Eq. (A.1) we also assume $k_f < k_4^c$. This allows to neglect higher-order terms in the Taylor expansion of the exponential term in Eq. (A.1). Thus, the linear approximation of Eq. (A.4):

$$P_{wc}^L = 1 + (P_{wc}^0 - 1)\left(1 - \frac{k_f}{k_4^c}\right) = \frac{k_f + q_4^{nc} P_{wc}^{C3}}{k_4 c} + \frac{k_f q_4^{nc} P_{wc}^{C3}}{(k_4 c)^2} \approx \frac{k_f + q_4^{nc} P_{wc}^{C3}}{k_4 c} \tag{A.2}$$

Fig. 4.4 demonstrates the validity of the linear approximation Eq. (A.2), as it matches non-approximated Eq. (A.4) and numerical calculations for the biologically relevant region of k_4^c. By expressing P_{wc}^L in terms of k_4^c and k_4^{nc} according to Eq. (4.3), we obtain Eq. (4.5).

With Eq. (A.2) we expressed P_{wc}^L as a linear function of k_4^c. However, the cancelling of k_4^c in η of the slow kinetic regime of the wb-WC reaction is still not immediately evident. To clearly observe this, we also needed to simplify $\frac{[C_4^{nc}]}{[C_4^c]}$ (designated as $\frac{1}{D}$ below). To do this, we first considered equilibration of D as a function of k_4^c. At $k_4^c \to 0$, D approaches its equilibrium value D_{max}. D_{max} is the "intrinsic selectivity" from Pavlov and Ehrenberg (2018), multiplied by $\frac{k_4^{nc}}{k_4^c}$:

$$D_{max} = \frac{a_3^{nc} a_4^{nc}}{a_3^c a_4^c} \frac{k_4^{nc}}{k_4^c} = \frac{q_3^{nc} q_4^{nc}}{q_3^c q_4^c} \tag{A.3}$$

At $k_4^c \to \infty$, D approaches its minimal value D_{min}. The exact form of D_{min} is irrelevant below, as $D_{min} << D_{max}$. Using these two variables as initial and equilibrium concentrations, we can describe the process of D equilibration using the equation for a first-order reversible reaction:

$$D = D_{max} + (D_{min} - D_{max}) \exp\left(-\frac{Q}{K}\right) \approx D_{max} - D_{max} \exp\left(-\frac{Q}{K}\right) \tag{A.4}$$

where Q and K are total reverse and forward rates of decoding. The equilibration of D is limited by the cognate decoding rates (Pavlov and Ehrenberg, 2018). Therefore, we can approximate Q and K from the cognate rate constants only:

$$Q = \prod q_i^c = q_2 q_3^c q_4^c \tag{A.5}$$

and

$$K = \prod k_i^c = k_2 k_3 k_4^c \tag{A.6}$$

Since we are interested in the non-equilibrium region where $K > Q$, the ratio in the exponential term in Eq. (A.4) allows to apply the linear approximation to the Taylor expansion of the exponential, similarly as performed above to obtain Eq. (A.2):

$$D^L = D_{max} - D_{max}\left(1 - \frac{Q}{K}\right) = \frac{QD_{max}}{K} \tag{A.7}$$

Using Eq. (A.3), Eq. (A.5) and Eq. (A.6), from Eq. (A.7) we obtain the linear approximation of $\frac{1}{D}$:

$$\left(\frac{1}{D}\right)^L = \frac{K}{QD_{max}} = \frac{k_2 k_3 k_4^c}{q_2 q_3^{nc} q_4^{nc}} \tag{A.8}$$

Eq. (A.8) is a good approximation to $\frac{[C_4^{nc}]}{[C_4^c]}$ in the non-equilibrium region, as demonstrated on Fig. 4.4.

From Eq. (A.8) and Eq. (A.2) we obtain the linear approximation of η for the slow kinetic regime of the wb-WC reaction at non-equilibrium conditions of both decoding and the wb-WC reaction:

$$\eta^{L} = \left(\frac{1}{D}\right)^{L} P_{wc}^{L} = \frac{k_2 k_3}{q_2 q_3^{nc}} \left(\frac{k_f}{q_4^{nc}} + P_{wc}^{C3}\right) \tag{A.9}$$

which is indeed independent of k_4^c. Fig. 4.4 demonstrates the validity of Eq. (A.9).

Appendix C

List of hazardous substances used according to the GHS

Substance	GHS hazard	Hazard statement	Precautionary statement
2-aminopurine	GHS07	H302	P264, P301+P312, P270, P330, P501
Hydrochloric acid 37 %	GHS05, GHS07	H290, H314, H335	P260, P280, P390, P303+P361+P353, P305+P351+P338, P403+P233, P501

Acknowledgements

The work was supported by a fellowship from Deutscher Akademischer Austauschdienst (DAAD) for graduate studies. I am grateful for computational support by Norddeutscher Verbund für Hoch- und Höchstleistungsrechnen (HLRN) and by Hummel cluster at the University of Hamburg. I thank Prof. Andrew Torda for helpful comments during my graduate studies.
I thank Prof. Dmytro Hovorun, my MSc diploma supervisor, who inspired me to pursue the problem of nucleobase tautomerism.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, die vorliegende Dissertation selbst verfasst und keine anderen als die angegebenen Hilfsmittel benutzt zu haben. Die eingereichte schriftliche Fassung entspricht der auf dem elektronischen Speichermedium. Ich versichere, dass diese Dissertation nicht in einem früheren Promotionsverfahren eingereicht wurde.

Aachen, den ____________________ Unterschrift: ________________________